UNIVERSIDADE FEDERAL DE UBERLÂNDIA – UFU

FACULDADE DE ENGENHARIA ELÉTRICA - FEELT

ENGENHARIA DE COMPUTAÇÃO

JOÃO PAULO CÂNDIDO NASCIMENTO

OS DESAFIOS EM LIDAR COM DADOS PROBLEMÁTICOS: UM ESTUDO EM CIÊNCIA DE DADOS SOBRE A DENGUE EM BRASÍLIA / DF

Uberlândia

2019

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

JOÃO PAULO CÂNDIDO NASCIMENTO

2019

JOÃO PAULO CÂNDIDO NASCIMENTO

OS DESAFIOS EM LIDAR COM DADOS PROBLEMÁTICOS: UM ESTUDO EM CIÊNCIA DE DADOS SOBRE A DENGUE EM BRASÍLIA / DF

Trabalho de Conclusão de Curso apresentado à Faculdade de Engenharia Elétrica – FEELT, com requisito para a obtenção do Título de bacharel em Engenharia de Computação.

Orientador: Prof. Igor Santos Peretta

Uberlândia

2019

JOÃO PAULO CÂNDIDO NASCIMENTO

OS DESAFIOS EM LIDAR COM DADOS PROBLEMÁTICOS: UM ESTUDO EM CIÊNCIA DE DADOS SOBRE A DENGUE EM BRASÍLIA / DF

Trabalho de Conclusão de Curso apresentado à faculdade de Engenharia Elétrica – FEELT, como requisito para a obtenção do Título de Bacharel em Engenharia de Computação.

COMISSÃO JULGADORA:

Prof. Dr. Marcelo Rodrigues de Sousa

Faculdade de Engenharia Elétrica

Prof. Dr. Márcio José da Cunha

Faculdade de Engenharia Elétrica

Prof. Dr. Igor Santos Peretta

Faculdade de Engenharia Elétrica

Professor Orientador – Presidente da Banca Examinadora

Uberlândia, cinco de julho de dois mil e dezenove

RESUMO

O presente trabalho consiste em um estudo de caso da dengue em Brasília, realizado com a finalidade de explorar ferramentas de ciência de dados para obter *insights* e discutir as dificuldades que existem durante um processo de manipulação de dados. Dessa maneira, foram extraídos dados de fontes oficiais de entidades públicas do Brasil sobre casos de dengue, sobre condições climáticas, sobre censos demográficos e sobre repasses de verbas destinadas ao combate de epidemiologias, todos a respeito da cidade de Brasília e para o período de janeiro de 2001 a dezembro de 2017. Após a extração dos dados, foi necessário um processo de preparação dos mesmos, em virtude de inconsistências encontradas com relação a dados faltantes ou problemas de granularidade de tempo (alguns dados são registros diários, outros mensais, outros anuais). Em seguida, as variáveis foram submetidas a testes de normalidade, e transformações foram feitas buscando aproximar aquelas que não passaram nos testes da distribuição normal. Com isso, foram construídos modelos de regressão linear múltipla separados por estação, que envolveram os casos de dengue e as variáveis climáticas com suas interações, onde, para o verão e o outono, os modelos resultantes foram significantes para explicar as variações nos casos de dengue, enquanto que, para a primavera e o inverno, os modelos encontrados não são suficientes para explicar tais variações. Foram feitas também análises envolvendo o repasse de verbas e o total de casos de dengue, onde se concluiu que, em alguns momentos, o repasse de verbas do ano corrente apresentou uma correlação positiva com os casos de dengue do ano anterior. Por fim, avaliou-se o efeito de características gerais da população (faixa etária, sexo e zona de residência), buscando determinar se tais características influenciam em uma maior ou menor probabilidade de se contrair dengue. Ao longo de todo o trabalho, foram discutidas dificuldades encontradas durante a condução de uma análise de ciência de dados.

Palavras-chave: Ciência de Dados, Modelos, Regressão Linear Múltipla, Probabilidade, Dengue, Clima.

ABSTRACT

The present work consists of a case study of dengue fever in Brasilia, carried out to explore data science tools to gain insights and discuss the difficulties that exist during a data manipulation process. In this way, data were extracted from official sources of public entities in Brazil on cases of dengue, climatic conditions, demographic censuses and on lending of funds destined to fight epidemiology, all regarding the city of Brasília and for the period of January from 2001 to December 2017. After the data extraction, a process of preparation of the data was necessary, due to inconsistencies found with respect to missing data or problems of granularity of time (some data are daily, other monthly, other annual). Afterwards, the variables were submitted to normality tests, and transformations were made in order to approximate those that did not pass the normal distribution tests. Thus, multiple linear regression models separated by season were constructed, which involved the cases of dengue and the climatic variables with their interactions, where, for summer and autumn, the resulting models were significant to explain the variations in dengue cases, whereas for spring and winter the models found are not sufficient to explain such variations. Analyzes were also carried out involving the transfer of funds and the total number of dengue cases, where it was concluded that, at some moments, the transfer of funds from the current year showed a positive correlation with the cases of dengue of the previous year. Finally, the effect of general characteristics of the population (age, sex and area of residence) was evaluated, trying to determine if these characteristics influence in a greater or lesser probability of contracting dengue. Throughout the work, difficulties encountered during the conduction of a data science analysis were discussed.

Keywords: Data Science, Models, Multiple Linear Regression, Probability, Dengue Fever, Weather.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE SIGLAS

ANOVA Análise de Variância

SQT Soma dos Quadrados Totais

SQR Soma dos Quadrados Residuais

SQE Soma dos Quadrados Explicada

SUS Sistema Único de Saúde

DATASUS Departamento de Informática do Sistema Único de Saúde

BDMEP Banco de Dados Meteorológicos para Ensino e Pesquisa

FNS Fundo Nacional de Saúde

TFVS Teto Financeiro De Vigilância Em Saúde

PFVS Piso Fixo De Vigilância Em Saúde

DAGVS Departamento de Apoio à Gestão da Vigilância em Saúde

IBGE Instituto Brasileiro de Geografia e Estatística

SUMÁRIO

1. INTRODUÇÃO

A área de *data science* tem conquistado seu lugar no mercado pelo potencial de aplicação de suas ferramentas – visualização de dados (gráficos, diagramas e outros), estudos com algoritmos complexos, tendo recursos atuais de pesquisa -, atualmente a abrangência do uso de ciência de dados tem acrescido cada vez mais nas suas dimensões de atuação. Tem-se visto como recorrente busca de seu uso a oportunidade de trazer, através de sua implementação *insights* valiosos para as entidades, sejam elas públicas ou privadas, de maneira objetiva e eficaz a fim de solucionar possíveis adversidades, independentemente de sua complexidade.

A ciência de dados se caracteriza por ser uma área de estudos multidisciplinares que busca obter conhecimentos de grande valor a partir da análise de dados, empregando-se, para este fim, métodos científicos e técnicas avançadas de pesquisa de dados, *machine learning* e inteligência artificial.

Pode-se perceber um inter-relacionamento advindo da ciência de dados ente o universo acadêmico e o universo dos negócios. Nas universidades, passa se o conhecimento sobre o método científico e a formalização dos estudos e pesquisas. Já para os negócios, a importância advém da rápida resolução de seus conflitos internos, dentro de suas setorizações e ramificações de cadeia.

Deste modo, avalia-se que a ciência de dados busca unir os dois universos, elaborando formalizações de soluções rápidas e eficazes tanto para as entidades governamentais, quanto para as não governamentais. Contudo, para isto, é preciso um conhecimento dos dados de composição da análise, visto que, suas características irão determinar a ferramenta adequada para sua manipulação.

Tendo em vista o uso da ciência de dados para processamento dos dados e entendimento de suas áreas abrangentes de conhecimento, o presente trabalho tem por objetivo geral explorar ferramentas de ciências de dados para manipular dados problemáticos, discutindo assim, sobre as dificuldades deste processo.

Junto aos objetivos específicos que visam o estudo de como dados de variáveis climáticas, dados de características gerais da população e dados de repasse de verbas destinadas ao combate de epidemiologias influenciam na quantidade de casos de dengue de janeiro de 2001 a dezembro de 2017.

Tal pesquisa foi iniciada levando em consideração a popularidade do uso de ciência de dados, com enfoque da relevância social da epidemiologia da dengue, visando compreender o processo de manipulação dos dados, juntamente às dificuldades que o rodeiam.

Por fim, o trabalho divide-se em uma estrutura introdutória inicial, com orientação quanto à metodologia usada para a realização da pesquisa; seguida do desenvolvimento com base no referencial teórico de autores que consideram os conceitos relativos à Ciência de dados, Distribuições de probabilidade, Teste de hipótese, Variáveis dependentes e independentes, Correlação e Regressão; e, por fim, em conclusão, serão apresentados os resultados úteis do trabalho face as observações nas considerações finais.

2. REFERENCIAL TEÓRICO

Este capítulo aborda a fundamentação teórica dos principais conceitos utilizados para o desenvolvimento deste trabalho, trazendo definições, equações e imagens para discutir cada um deles.

2.1. Ciência de Dados

Todos os estudos que envolvem a preparação de dados e as técnicas de manipulação e modelagem dos mesmos, podem ser chamados de ciência de dados. Em outras palavras, um profissional dessa área, denominado cientista de dados, é capaz de obter *insights* (informações valiosas) a partir de dados inicialmente desorganizados [1].

Dessa maneira, a ciência de dados é uma área de estudo capaz de auxiliar na tomada de decisões e no planejamento de estratégias nos mais diversos segmentos. Joel Grus apresenta em seu livro [1] diversos exemplos de empresas que utilizam ferramentas de ciência de dados para melhorar a experiência de clientes e impulsionar vendas e lucros. Além disso, Grus ainda defende que essa área de estudo pode ser usada para auxiliar o Governo com estratégias mais eficazes de moradia ou saúde pública.

2.2. Probabilidade

Em qualquer experimento aleatório, sempre existe uma incerteza a respeito da ocorrência ou não de um determinado evento. Como uma medida dessa chance com a qual podemos esperar que o evento ocorra, surge o conceito de probabilidade. Dessa maneira, é conveniente atribuir um número entre 0 e 1 para tal medida, uma vez que, se existe certeza de que o evento ocorrerá, pode-se dizer que sua probabilidade é 100% (ou 1), mas se a certeza for de que o evento não ocorrerá, sua probabilidade é zero.

2.3. Distribuições de Probabilidade

Uma distribuição de probabilidade pode ser definida como uma função que busca estudar a frequência com que determinado evento ocorre [4]. Dessa forma, uma distribuição de probabilidade pode ser também chamada de distribuição de frequências, e, para um evento qualquer, de acordo com sua distribuição, pode-se estimar a probabilidade de o mesmo ocorrer.

Dessa maneira, diversos são os tipos de distribuição de probabilidade que existem, a fim de descrever o comportamento tanto de variáveis discreta quanto contínuas. Para dados discretos, existem as distribuições: Bernoulli, binomial, hipergeométrica, geométrica e Poisson.

Por sua vez, quando se trata de dados contínuos, pode-se citar as distribuições: normal, gama, exponencial, beta, qui-quadrado, t de Student, F de Snedecor, entre várias outras [3].

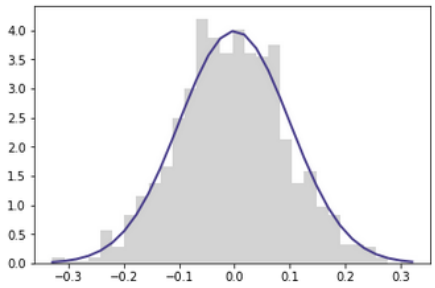
2.3.1. Distribuição Normal

A distribuição normal, que pode também ser chamada de curva de Gauss ou distribuição gaussiana, é uma das mais utilizadas distribuições de probabilidade, em virtude de diversos métodos estatísticos poderem ser aplicados apenas a distribuições normais. A função densidade de probabilidade de uma distribuição normal é dada por:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)} \quad (1)$$

Sendo que o símbolo σ se refere ao desvio-padrão e o símbolo μ à média [2].

FIGURA 1 – GRÁFICO DE UMA DISTRIBUIÇÃO NORMAL DE DADOS



A Fig. 1 mostra um exemplo de uma distribuição normal de dados, gerada através de um script Python. A amostra de dados foi gerada a partir da função *normal* da biblioteca *random*, que recebe como argumento a média e o desvio padrão desejados. Os argumentos passados foram σ = 0,1 (desvio-padrão) e μ = 0.0 (média). Dessa forma, observando o gráfico, a distribuição normal é simétrica em relação à média 0.0, e seus valores estão contidos majoritariamente entre -3σ e +3σ.

2.4. Testes de Hipótese

Durante uma análise estatística, em diversas situações, a tomada de decisões sobre as populações se faz necessária com base em informações de amostra, de forma que tais decisões são denominadas decisões estatísticas (Schiller, Srinivasan e Spiegel). Na tentativa de chegar a essas decisões, é necessário levantar hipóteses, as quais são chamadas de hipóteses estatísticas e, normalmente, são declarações a respeito de distribuições de probabilidades dos dados [5].

Ainda segundo os autores, ao levantar hipóteses, normalmente é definida uma hipótese nula denominada H_0 e hipóteses alternativas denominadas H_1. No contexto de determinar se um procedimento é melhor que outro, uma hipótese nula poderia ser a de que não há diferença entre os procedimentos testados, ou seja, quaisquer diferenças observadas entre eles podem ser atribuídas a flutuações na amostragem da população [5].

2.4.1. Nível de Significância

Schiller, Srinivasan e Spiegel ainda apresentam o conceito de significância estatística como um importante fator ao testar uma determinada hipótese. Entende-se por nível de significância do teste a probabilidade máxima com a qual se deseja arriscar a rejeição de uma hipótese, quando na verdade ela é verdadeira [5].

Um nível de significância de 0,05 é o mais adotado pela comunidade, embora outros valores possam ser usados livremente. Se tal nível de 0,05 (ou 5%) for escolhido, pode-se estimar que existem 5 chances em 100 de rejeitar a hipótese quando deveria ser aceita. Em outras palavras, sempre que as hipóteses nulas forem verdadeiras, existe cerca de 95% de confiança de a decisão estar correta [5].

2.4.2. P-Valor

Geralmente, durante um teste de hipótese, um parâmetro denominado p-valor é calculado, este parâmetro é utilizado para se comparar com o nível de significância estabelecido e tomar decisões acerca da aceitação ou rejeição de hipóteses. Dessa forma, p-valores menores que o nível de significância fornecem evidências para rejeitar H_0 em favor de H_1, enquanto p-valor superior ao nível estabelecido fornece evidência para não rejeitar H_0 em favor de H_1 [5].

2.4.3. Teste de Normalidade

Segundo as definições de Schiller, Srinivasan e Spiegel acerca de decisões e hipóteses, os testes de normalidade podem ser entendidos como testes de hipóteses, onde se compara a distribuição de uma amostra de dados qualquer com a distribuição normal. Logo, um teste de normalidade possui uma hipótese nula (H_0) de que não há diferenças entre a distribuição dos dados em questão e a distribuição normal e uma hipótese alternativa (H_1) de que a distribuição dos dados se difere da normal.

Dessa forma, adotando-se o nível de significância de 5%, quando o p-valor resultante de um teste de normalidade for menor que 5%, pode-se rejeitar H_0 e inferir que a distribuição dos dados difere-se de uma distribuição gaussiana, e, quando o mesmo for maior que 5%, não

existem evidências suficientes para rejeitar H_0 e pode ser feita a inferência de que a distribuição dos dados se aproxima da normal.

Dentre os testes de normalidade, destacam-se os testes de D'Agostino e Pearson [6][7], de Shapiro-Wilk [8] e de Kolmogorov-Smirnov [9][10].

2.5. Transformação de Box-Cox

A transformação de Box-Cox é uma transformação de dados que pode ser aplicada buscando aproximar a distribuição de uma amostra a uma distribuição de dados normal. Tal transformação se dá por encontrar um valor λ tal que:

$$y' = \begin{cases} \frac{y^{\lambda}-1}{\lambda}, se\ \lambda > 0 \\ \ln(y), se\ \lambda = 0 \end{cases} \quad (2)$$

onde y é o valor original da variável e y' é o valor atribuído a ela após a transformação [22].

2.6. Variáveis Dependentes e Independentes

Em uma análise estatística, se uma variável envolvida no conjunto de dados que está sendo analisado é utilizada para explicar outra, ela é chamada de variável independente ou explicativa. A variável que se deseja explicar a partir de outras é chamada de variável dependente ou resposta [11].

2.7. Correlação

Correlação pode ser entendida como uma relação entre duas variáveis, de forma que, com a variação de uma variável independente, pode-se observar uma variação correspondente na variável resposta. Uma correlação pode ser positiva (onde pequenos valores de uma variável X correspondem a pequenos valores de uma variável Y e grandes valores de X correspondem a grandes valores de Y) ou negativa (onde pequenos valores de X correspondem a grandes valores de Y, e grandes valores de X correspondem a pequenos valores de Y). Além disso, existem situações em que não há correlação entre as variáveis, isto é, não se pode observar um padrão na relação entre as variáveis X e Y [11].

2.8. Regressão

Normalmente, ao encontrar correlações entre variáveis, pode-se desejar quantificar matematicamente tal relação, por meio de uma expressão que conecta uma variável resposta a uma ou mais variáveis explicativas. Esse processo é usualmente referido como regressão, e o

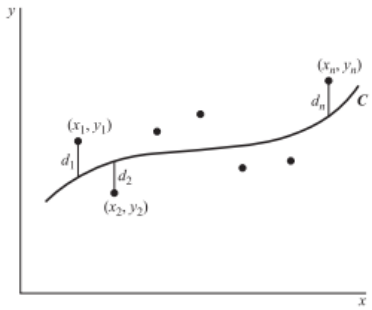
resultado produzido é uma curva de ajuste aos dados utilizados. Quando tal curva produzida é uma reta, ela é denominada uma regressão linear. Porém, diversos outros tipos de curva como parábolas podem ser produzidos [5].

Uma regressão pode ser ainda classificada de acordo com a quantidade de variáveis independentes que ela inclui, de forma que, quando se utiliza apenas uma variável, trata-se de uma regressão simples. Por sua vez, quando duas ou mais variáveis são incluídas na expressão para explicar determinada variável resposta, tem-se uma regressão múltipla [12].

2.8.1. Método dos Mínimos Quadrados

Em um processo de regressão, busca-se encontrar a curva mais adequada para descrever uma determinada variável resposta. Nesse processo, de acordo com a Fig. 2 (extraída de [5]), para um dado valor x_1, haverá uma diferença d_1 entre o real valor y_1 e o valor ajustado pelo modelo, conforme determinado a partir da curva C. Tal diferença é chamada de desvio, erro ou resíduo, e pode ser tanto positiva quanto negativa ou nula. De forma análoga, os desvios d_2, ..., d_n podem ser obtidos.

FIGURA 2 – AJUSTE DE CURVA PARA UMA VARIÁVEL RESPOSTA Y EM FUNÇÃO DE UMA VARIÁVEL EXPLICATIVA X



Uma medida de qualidade do ajuste da curva C ao conjunto de dados é fornecida pela soma das diferenças, sendo que, se ela é pequena, o ajuste é bom, e, se for grande, o ajuste é ruim. Assim, a curva que minimiza a soma das diferenças é denominada de curva de melhor ajuste, e o método de obtenção dessa curva é denominado método dos mínimos quadrados. A equação a seguir ilustra tal processo de minimização [5].

$$d_1^2 + d_2^2 + \cdots + d_n^2 = mínimo \quad (3)$$

No presente trabalho, a Regressão pelo Método dos mínimos quadrados foi implementada através de funções presentes na biblioteca *stats* da linguagem Python. Mais

informações dos cálculos matemáticos envolvidos no processo de busca do melhor ajuste de curva podem ser encontradas em [5].

2.8.2. Coeficiente de Determinação e Coeficiente de Determinação Ajustado

O coeficiente de determinação, denominado R^2, pode ser entendido como a fração da variação total da resposta, explicada pela regressão por mínimos quadrados, ou seja, o mesmo corresponde a porcentagem da variável dependente que o modelo consegue explicar [5]. Tal coeficiente se faz utilizado como indicador do grau de ajustamento de um modelo de regressão linear simples [12].

Porém, quando se trata da regressão múltipla, tal indicador não é confiável, uma vez que ao adicionar novas variáveis ao modelo, seu valor cresce gradativamente independentemente de quais ou quantas são as variáveis introduzidas. Neste sentido, o coeficiente de determinação ajustado (R^2 ajustado), é um indicador que melhor expressa o grau de ajustamento de um modelo de regressão linear múltipla [12].

2.8.3. Análise de variância

A análise de variância (ANOVA) é uma ferramenta que objetiva evidenciar se existem diferenças significativas entre as medias de grupos diferentes de uma mesma amostra, para isso, uma hipótese H_0 de que não existem diferenças entre as medias dos grupos é formulada, e a análise de variância, por meio de um teste denominado teste *F*, objetiva aceitar ou rejeitar tal hipótese [5].

No contexto da regressão linear múltipla, a análise de variância visa estimar quais variáveis utilizadas na construção do modelo, são mais significativas para a variável resposta. Deste modo, define-se H_0 como a hipótese de que nenhuma das variáveis explicativas contribui significativamente para o modelo. Logo, ao conduzir tal análise são geradas duas estatísticas importantes, sendo elas, o valor de F e o p-valor associado a ela. Se p-valor menor que o nível de significância estabelecido, a hipótese H_0 é rejeitada, representando que pelo menos uma das variáveis independentes contribuam para explicar a variável resposta [12].

Além disso, ao conduzir a análise de variância para cada variável explicativa independente, H_0 constitui-se como a hipótese de que a variável em questão é significante para o modelo. Da mesma forma, se o p-valor resultante for menor que o nível de significância, tal variável contribui de maneira relevante para explicar a variável resposta [13].

Vale ressaltar que a ANOVA supõe uma distribuição normal das variáveis envolvidas no processo, sejam elas dependentes ou independentes. Além disso, tal ferramenta estatística também tem como pressuposto que as observações sejam independentes, ou seja, a observação de uma variável não pode interferir em outra [23].

2.8.3.1. Teste de F para a hipótese nula da ANOVA

Em termos matemáticos, a análise de variância é composta pela decomposição da soma dos quadrados totais (SQT) em soma dos quadrados residuais (SQR) e soma dos quadrados explicada (SQE). Uma soma de quadrados é uma medida de variação de uma variável qualquer, e, como os nomes sugerem, SQT corresponde à variação total da variável resposta, SQE representa a variação da variável resposta que é explicada pelo modelo e SQR é a variação da variável resposta que o modelo não consegue explicar [12].

$$SQT = SQR + SQE \quad (4)$$

Dessa forma, a fim de testar a hipótese H_0 de uma ANOVA, utiliza-se a estatística F, dada por:

$$F = \frac{\frac{SQR}{p}}{\frac{SQE}{n-p-1}} \quad (5)$$

onde n é a quantidade de amostras (registros) e p a quantidade de variáveis [12]. Após o cálculo do valor de F observado na amostra (F_{obs}), ele é então comparado a um valor de F tabelado (F_{tab}), dado a partir da quantidade de amostras n e da quantidade de variáveis p. Assim, se $F_{obs} > F_{tab}$ com um nível de confiança maior que a significância estatística estabelecida, pode-se rejeitar H_0. Tal nível de confiança é expresso por um p-valor que representa a probabilidade da sentença $F_{obs} > F_{tab}$ ser por razões aleatórias. Em outras palavras, se p-valor menor que um nível de significância de, por exemplo, 5%, significa que existe menos de 5% de chance de F_{obs} ser maior que F_{tab} por uma razão aleatória. Então, rejeita-se H_0 em favor da hipótese alternativa de que existem diferenças significativas entre as médias dos grupos [5].

3. DESENVOLVIMENTO

Na seguinte seção, será apresentado o estudo de caso realizado no presente trabalho, explorando as ferramentas utilizadas, os métodos implementados e os resultados obtidos.

3.1. Ferramentas Utilizadas

Para o desenvolvimento do estudo de caso, foi utilizada como ferramenta a linguagem de programação Python. Tal linguagem é de alto nível e multiparadigma, de modo que pode ser orientada a objetos, funcional e procedural. Sua sintaxe é bem clara e concisa, e possui muitos recursos que podem ser utilizados por meio de suas bibliotecas. [24]

Python possui uma tipagem dinâmica, e uma de suas principais vantagens é poder escrever um programa em poucas linhas de código quando comparado a outras linguagens. Devido aos grandes benefícios que ela traz, é amplamente utilizada para análise e processamento de dados.

Dessa forma, para implementar todos os algoritmos de análise dos casos de dengue em Brasília, foram utilizadas diversas bibliotecas do Python destinadas à manipulação de dados. Todas essas bibliotecas, bem como sua utilização direta no presente trabalho, estão listadas na tabela abaixo.

TABELA 1: BIBLIOTECAS UTILIZADAS NO DESENVOLVIMENTO DESTE TRABALHO

Biblioteca	Utilização	Documentação
numpy	Utilizada para realizar operações matemáticas com *arrays* de forma mais rápida e eficaz.	https://www.numpy.org/
pandas	Utilizada para criar tabelas (*dataframes*) que facilitam o carregamento de dados e sua posterior manipulação em um *script* Python.	https://pandas.pydata.org/pandas-docs/stable/
scipy.stats	Utilizada para realizar os testes de normalidade, as transformações de Box-Cox e a implementação do modelo de ANOVA.	https://docs.scipy.org/doc/scipy/reference/stats.html
seaborn	Utilizada para gerar gráficos de distribuições e resíduos juntamente com a *matplotlib*.	https://seaborn.pydata.org/
matplotlib	Utilizada para gerar gráficos de distribuições e resíduos juntamente com a *seaborn*.	https://matplotlib.org/3.1.0/contents.html
statsmodels	Utilizada para gerar o modelo de regressão linear pelo método dos mínimos quadrados.	https://www.statsmodels.org/stable/index.html

3.2. Estudo sobre a Dengue em Brasília/DF

Para o estudo deste trabalho, foram utilizados dados coletados na cidade de Brasília sobre registros de casos de dengue. Vale ressaltar que todos os conjuntos de dados, bem como todos os *scripts* construídos na linguagem Python, que serão discutidos nas sessões seguintes, encontram-se no repositório do autor (https://github.com/joaopcandido/tcc-estudo-dengue-brasilia).

Segundo o Ministério da Saúde [14], o vírus de transmissão da dengue é classificado como arbovírus, cujo qual se caracteriza pela transmissão através de picadas de insetos, em destaque mosquitos. Classificam-se em quatro tipos (sorotipos 1, 2, 3 e 4), onde cada pessoa pode ser contaminada pelos quatro sorotipos, sendo a mesma imunizado ao tipo já adquirido. Tem-se que a transmissão da dengue se dá por meio da picada do mosquito *Aedes Aegypt,* que se reproduz principalmente em locais de água parada. Além disso, é importante evidenciar que sua proliferação ocorre em regiões tropicais e subtropicais.

Vê-se que a dengue é uma das doenças virais mais recorrentes no Brasil atualmente, notada como um dos maiores transtornos para a saúde pública do mundo [16]. Deste modo, trata-se em destaque os registros de casos de dengue da cidade de Brasília, escolhida como alvo de discussão deste trabalho por ser a capital do país. Foram extraídos também, além dos dados referentes à doença, dados meteorológicos da cidade, com o intuito de relacionar os números de casos da doença com o comportamento do clima da região. Além desses, foram reunidos ainda dados referentes ao repasse de verbas para vigilância em saúde e combate a epidemiologias, bem como dados dos censos demográficos de 2000 e 2010 [19].

3.2.1. Levantamento dos dados

Para os registros de casos de dengue, foram extraídos dados do portal do Departamento de Informática do Sistema Único de Saúde (DATASUS) [17], de 2001 a 2017, sendo todos os disponíveis até o momento da realização deste trabalho. A opção em se utilizar apenas a cidade de Brasília como alvo de estudo se deu devido à dificuldade em consultar os dados no portal, visto que cada cidade e cada ano precisam ser consultados separadamente. Além disso, os registros foram consultados por faixa etária, sexo e zona de residência.

Cada consulta foi realizada gerando-se uma página contendo os valores separados por vírgulas. A partir disso, os dados eram copiados para um editor de texto e salvados no formato ‘.csv’. Durante o processo de extração, foram gerados 51 arquivos, sendo 17 referente a cada tipo de consulta (sexo, faixa etária, zona de residência).

Já para os dados referentes ao clima de Brasília, a extração foi realizada a partir do Banco de Dados Meteorológicos para Ensino e Pesquisa (BDMEP) [18], pertencente ao Instituto Nacional de Meteorologia. Neste processo, foram extraídos dados coletados em três períodos diferentes do dia (manhã, tarde e noite), dispostos em temperatura bulbo seco, temperatura bulbo úmido, umidade relativa, pressão atmosférica e direção do vento, desde 01 de janeiro de 2001 a 31 de dezembro de 2017, período escolhido em função do período dos registros de casos de dengue coletados.

É importante ressaltar aqui que a diferença entre as variáveis de temperatura (bulbo seco e bulbo úmido) consiste no fato de que a medição feita com bulbo seco objetiva medir a temperatura em si do ambiente, enquanto o bulbo úmido é um indicativo de evaporação da água, já que mede a temperatura da água durante a evaporação. Dessa maneira, quanto mais seco estiver o ar atmosférico, maior será tal temperatura e, quanto maior a umidade, menor ela será. Idealmente, se a umidade relativa do ar atingir um nível de 100%, a temperatura medida pelo bulbo seco será nula [21].

Por sua vez, os dados relacionados ao repasse de verbas para Vigilância em Saúde e Combate a Epidemiologias foram coletados a partir do portal *online* do Fundo Nacional de Saúde (FNS) [20], consultados ano a ano para o período de 2001 a 2017 e salvos em um arquivo '.csv'. Vale ressaltar que os dados especificamente relacionados ao combate à Dengue consistem nos fundos Teto Financeiro De Vigilância Em Saúde (TFVS) e Piso Fixo De Vigilância Em Saúde (PFVS), segundo orientação via e-mail do Departamento de Apoio à Gestão da Vigilância em Saúde (DAGVS).

Por fim, os dados dos censos demográficos, realizados pelo Instituto Brasileiro de Geografia e Estatística (IBGE) a cada 10 anos, foram extraídos diretamente do portal do IBGE [19]. Conforme encontrado no *site*, o censo demográfico "constitui a principal fonte de referência para o conhecimento das condições de vida da população em todos os municípios do País e em seus recortes territoriais internos, tendo como unidade de coleta a pessoa residente, na data de referência, em domicílio do Território Nacional". Dessa forma, tais dados foram utilizados para se ter informações gerais da população de Brasília, separada por faixa etária, sexo e zona de residência.

3.3. Necessidade de Tomada de Decisões

Durante a condução do estudo, a tomada de algumas decisões se fez necessária, em função de características dos dados e de inconsistências que os mesmos apresentaram. A primeira decisão tomada foi o nível de significância estatística adotado. Conforme mencionado anteriormente, esse nível é normalmente adotado em 5% (ou 0,05). Por isso, este foi o valor também adotado neste trabalho.

Devido a inconsistências encontradas nos dados, outra decisão que precisou ser tomada está relacionada aos dados climáticos de Brasília, já que a coluna "TempBulboUmido" apresentou registros ausentes referentes a outubro, novembro e dezembro de 2017. Logo, com o intuito de completá-los para não comprometer a análise deste período, foi feita a média entre todos os meses de outubro de 2001 a 2016, entre todos os meses de novembro de 2001 a 2016 e entre todos os meses de dezembro de 2001 a 2016 separadamente, para serem inseridos nos registros ausentes.

Além disso, ainda sobre os dados meteorológicos, a coluna "DirecaoVento" foi desconsiderada de todas as análises, uma vez que, ainda que os registros sejam numéricos, eles representam apenas direções diferentes para o vento. Dessa maneira, tal variável se comporta como uma variável categórica, e o cálculo de média não se aplica a ela.

TABELA 2: REGISTROS DISTINTOS DA COLUNA "DIRECAOVENTO" COM A QUANTIDADE DE VEZES EM QUE CADA UM SE REPETE

Registro encontrado	Quantidade de vezes que se repete
0	6214
5	4197
14	3822
32	1711
9	1331
23	526
36	484
18	133
27	101
15	30
10	10
35	7
19	6
4	6
30	6
7	4
24	3
2	3

Registro encontrado	Quantidade de vezes que se repete
16	3
33	3
6	2
8	2
20	1
1	1
12	1
3	1
11	1
17	1

É importante mencionar ainda que, em alguns momentos ao longo do desenvolvimento do trabalho, foi necessário segregar os dados em estações do ano (verão, outono, inverno e primavera). Porém, como os registros de dengue são mensais, as estações foram definidas, grosso modo, como: verão compreendendo os meses de janeiro a março, outono compreendendo de abril a junho, inverno de julho a setembro e primavera de outubro a dezembro.

Por último, para se analisar o impacto de características gerais da população sobre os casos de dengue, foram levantados dados dos censos demográficos. Porém, conforme mencionado na seção anterior, o IBGE realiza o censo a cada 10 anos. Dessa maneira, dados dos censos de 2000 e 2010 foram extraídos e, a partir dos mesmos, foi feita uma estimativa da população para os demais anos (de 2001 a 2009 e de 2011 a 2017), separada por faixa etária, sexo e zona de residência.

3.4. Preparação dos dados

3.4.1. Preparação dos Arquivos

Como discutido anteriormente, a consulta dos dados relacionados a casos de dengue gerou 51 arquivos diferentes. Dessa forma, para facilitar o carregamento dos dados e posterior análise, foi desenvolvido um *script* escrito na linguagem *Python* para unificar todos os arquivos de cada tipo de consulta em um só. Para exemplificar, os 17 arquivos referentes aos casos de dengue consultados por faixa etária foram condensados em um único arquivo. O mesmo foi feito para os arquivos de sexo e zona de residência.

Já os dados dos censos demográficos de 2000 e 2010, conforme mencionado anteriormente, foram extraídos diretamente do *site* do IBGE. Porém, ao baixar os dados, todas as planilhas relacionadas às características da população de Brasília foram extraídas, como religião, empregabilidade, entre outras. Por isso, foi feita uma busca manual pelos dados

referentes a faixa etária, sexo e zona de residência, os quais foram diretamente inseridos em um *script Python*. Em seguida, foi calculado um percentual de cada categoria dentro de cada uma das três características com relação à população geral, a fim de estimar a população dos anos de 2001 a 2009 e 2011 a 2017. Os resultados obtidos com tal procedimento são mostrados nas tabelas a seguir.

TABELA 3: ESTIMATIVA DA POPULAÇÃO TOTAL DE BRASÍLIA PARA OS ANOS DE 2001 A 2017, SEPARADA POR SEXO

Ano	Total	Masculino	Feminino
2001	2103047	1005538	1097509
2002	2154948	1030353	1124595
2003	2206849	1055169	1151680
2004	2258750	1079984	1178766
2005	2310651	1104800	1205851
2006	2362552	1129616	1232936
2007	2414453	1154431	1260022
2008	2466354	1179247	1287107
2009	2518255	1204062	1314193
2010	2570160	1228880	1341280
2011	2622057	1253694	1368363
2012	2673958	1278509	1395449
2013	2725859	1303325	1422534
2014	2777760	1328141	1449619
2015	2829661	1352956	1476705
2016	2881562	1377772	1503790
2017	2933463	1402587	1530876

TABELA 4: ESTIMATIVA DA POPULAÇÃO TOTAL DE BRASÍLIA PARA OS ANOS DE 2001 A 2017, SEPARADA POR ZONAS DE RESIDÊNCIA

Ano	Total	Urbana	Rural
2001	2103047	2013050	89997
2002	2154948	2064695	90253
2003	2206849	2116436	90413
2004	2258750	2168270	90480
2005	2310651	2220200	90451
2006	2362552	2272224	90328
2007	2414453	2324343	90110
2008	2466354	2376557	89797
2009	2518255	2428865	89390
2010	2570160	2481272	88888
2011	2622057	2533766	88291
2012	2673958	2586358	87600
2013	2725859	2639045	86814

Ano	Total	Urbana	Rural
2014	2777760	2691827	85933
2015	2829661	2744703	84958
2016	2881562	2797674	83888
2017	2933463	2850740	82723

TABELA 5: ESTIMATIVA DA POPULAÇÃO TOTAL DE BRASÍLIA PARA OS ANOS DE 2001 A 2017, SEPARADA POR FAIXA ETÁRIA

Ano	Total	<1 Ano	01-04	05-09	10-14	15-19	20-39	40-59	60-64	65-69	70-79	80 e +
2001	2103047	41846	159589	192569	193846	225997	795203	376788	44823	28591	32311	11483
2002	2154948	41628	159459	194095	196935	226409	815565	395691	47344	30520	34735	12566
2003	2206849	41351	159132	195465	199942	226572	835962	415056	49933	32508	37237	13689
2004	2258750	41013	158610	196679	202868	226486	856396	434884	52591	34555	39818	14849
2005	2310651	40615	157891	197739	205712	226151	876864	455174	55316	36661	42477	16048
2006	2362552	40156	156977	198643	208475	225568	897368	475928	58110	38826	45215	17286
2007	2414453	39638	155866	199391	211156	224735	917908	497143	60972	41050	48031	18562
2008	2466354	39059	154560	199984	213755	223654	938483	518822	63902	43333	50925	19877
2009	2518255	38420	153057	200422	216273	222324	959094	540963	66900	45674	53898	21230
2010	2570160	37721	151359	200704	218709	220745	979742	563567	69967	48075	56949	22622
2011	2622057	36962	149464	200830	221063	218917	1000422	586632	73101	50534	60078	24052
2012	2673958	36142	147374	200802	223336	216840	1021140	610161	76304	53053	63286	25521
2013	2725859	35262	145087	200617	225527	214515	1041892	634153	79575	55630	66572	27029
2014	2777760	34322	142605	200278	227636	211941	1062681	658607	82914	58267	69937	28575
2015	2829661	33322	139926	199783	229664	209117	1083505	683523	86321	60962	73379	30159
2016	2881562	32261	137052	199132	231610	206045	1104364	708902	89797	63716	76900	31782
2017	2933463	31140	133981	198326	233475	202725	1125259	734744	93340	66529	80500	33444

Vale ressaltar que a soma de cada coluna exceto a ‘Total’, para cada uma das separações, consiste em um valor próximo ao apresentado na coluna ‘Total’. O motivo de não ser o valor exato se dá por conta de arredondamentos realizados. Porém, para os cálculos de probabilidades, os arredondamentos foram desconsiderados, e os dados foram utilizados com casas decimais, conforme será mostrado nas seções seguintes.

3.4.2. Testes de Normalidade

3.4.2.1. Casos de Dengue

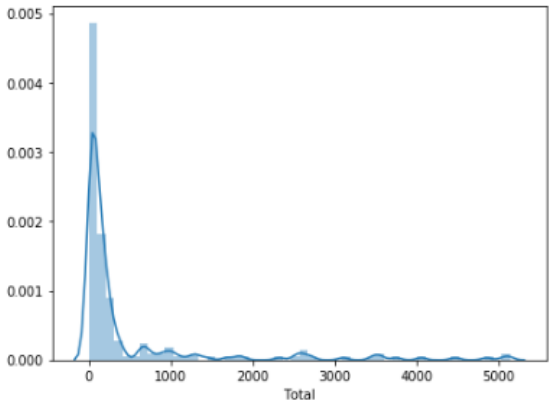
Antes de elaborar análises estatísticas e criar modelos de descrição ou previsão embasados em um conjunto de dados, é necessário realizar testes de normalidade, a fim de verificar se o mesmo obedece à uma distribuição gaussiana (normal). Tal verificação se faz necessária, em virtude de algumas análises poderem ser aplicadas apenas em distribuições normais, como é o caso da ANOVA.

Como mencionado, a biblioteca stats foi utilizada para realizar os testes de normalidade nos dados. Para isso, três testes presentes em tal biblioteca foram implementados: *normaltest* (teste que combina os testes de D'Agostino e Pearson), *kstest* (teste de Kolmogorov-Smirnov) e *shapiro* (teste de Shapiro-Wilk). Assim, foi realizado o teste de normalidade sobre os casos totais de dengue, como mostrado nas imagens a seguir.

TABELA 6: RESULTADOS DOS TESTES DE NORMALIDADE PARA O TOTAL DE CASOS DE DENGUE

Teste	P-Valor
Normal	3,32E-33
Shapiro	1,38E-23
Kolmogorov–Smirnov	8,05E-148

FIGURA 3 – DISTRIBUIÇÃO DE PROBABILIDADES PARA O TOTAL DE CASOS DE DENGUE



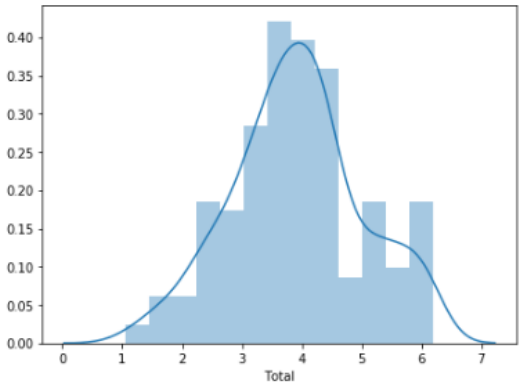
Como pode ser observado graficamente, a distribuição dos dados não possui as características de uma distribuição gaussiana e, como mostrado nos resultados dos testes, a hipótese H_0, a qual define que a distribuição dos dados vem de uma normal, pode ser rejeitada (todos os p-valores foram inferiores ao nível de significância estabelecido).

Na tentativa de encontrar uma distribuição que se aproximasse da normal, os dados foram separados por estação e submetidos separadamente aos testes (a descrição detalhada do procedimento realizado encontra-se no Apêndice A). Ainda assim, não foi possível encontrar uma distribuição que se aproximasse da normal. Porém, existem transformações que podem ser realizadas a fim de alterar a distribuição de um conjunto de dados, com o intuito de aproximá-lo de uma distribuição normal. Uma dessas transformações é a de *Box-Cox*, que foi realizada sobre o conjunto contendo o total de casos de dengue, por meio da função *boxcox* da biblioteca *scipy.stats*.

TABELA 7: RESULTADOS DOS TESTES DE NORMALIDADE PARA O TOTAL DE CASOS DE DENGUE APÓS A TRANSFORMAÇÃO DE BOX-COX

Teste	P-Valor
Normal	0,838098863
Shapiro	0,067225918
Kolmogorov–Smirnov	0,564131066

FIGURA 4 – DISTRIBUIÇÃO DE PROBABILIDADE DO TOTAL DE CASOS DE DENGUE APÓS A TRANSFORMAÇÃO DE BOX-COX



Como mostrado na Tabela 7, a transformação de *Box-Cox* aproximou a distribuição de casos de uma distribuição normal, já que os p-valores dos testes foram todos maiores que o nível de significância estatística adotado. Logo, não existem evidências estatísticas suficientes para rejeitar a hipótese de que os dados transformados vêm de uma distribuição normal.

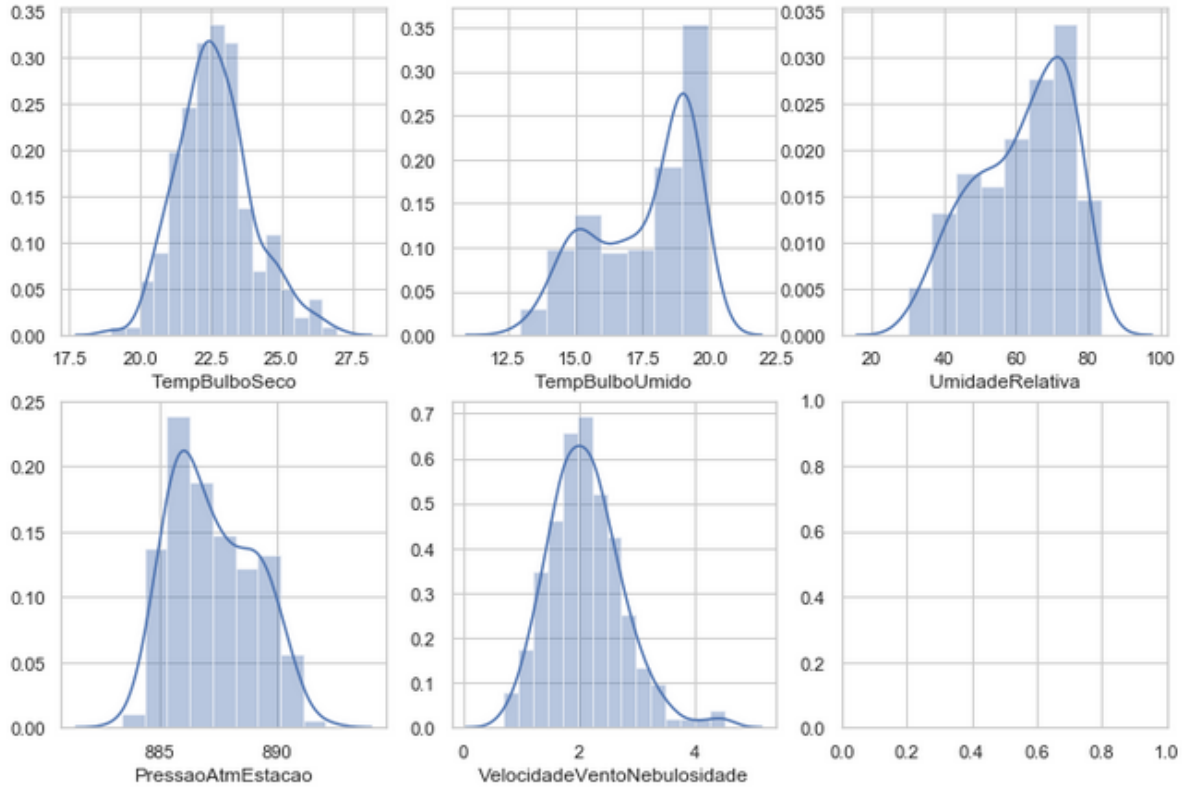
3.4.2.2. Meteorologia

Como mencionado anteriormente, os dados de meteorologia são registros diários, coletados nos três períodos do dia. Dessa forma, antes de realizar os testes de normalidade, os dados foram agregados mensalmente, fazendo-se a média de cada coluna. Em seguida, os dados foram submetidos aos testes de normalidade.

TABELA 8: TESTES DE NORMALIDADE DOS DADOS DE METEOROLOGIA AGREGADOS MENSALMENTE

Coluna	Normal	Shapiro	Kolmogorov–Smirnov
TempBulboSeco	0,012811015	0,016633386	0,139866093
TempBulboUmido	1,0359E-10	3,16415E-11	2,92E-18
UmidadeRelativa	3,91951E-06	1,09437E-06	6,51E-48
PressaoAtmEstacao	4,64764E-05	0,00012414	5,87E-10
VelocidadeVentoNebulosidade	1,46606E-05	0,000444887	9,83E-04

FIGURA 5 – DISTRIBUIÇÕES DOS DADOS DE METEOROLOGIA AGREGADOS MENSALMENTE



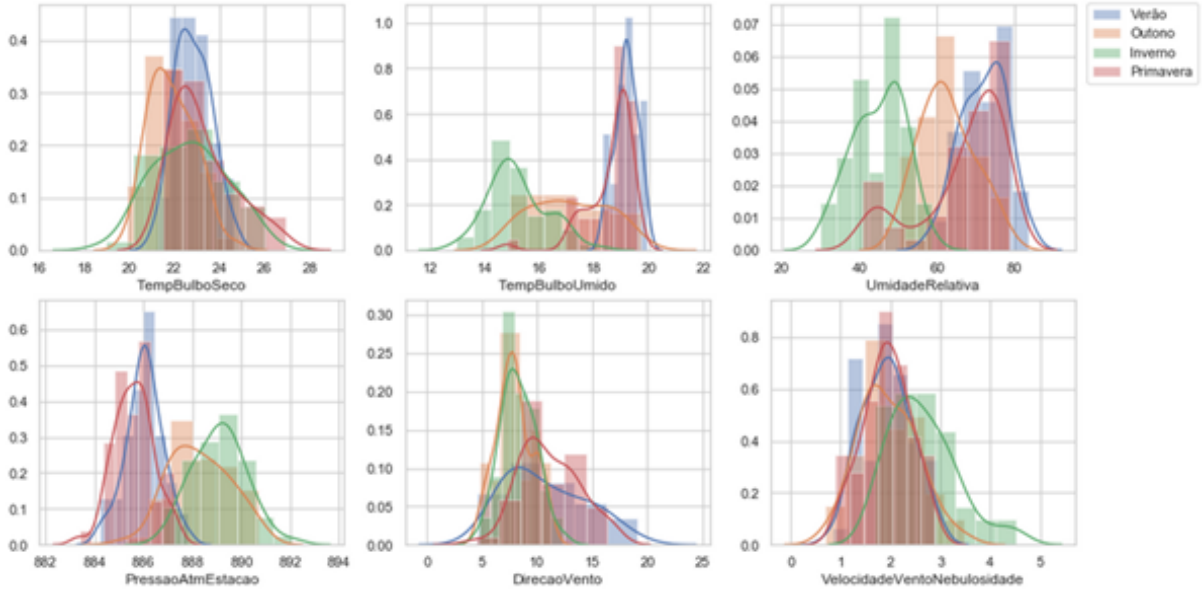
Como mostrado na Tabela 8, apenas a coluna ‘TempBulboSeco’ pelo teste de Komolgorov-Smirnov possui distribuição próxima da normal. Assim, em uma nova tentativa de encontrar uma distribuição normal para os registros de Meteorologia, os dados originais (sem transformação) foram separados por estação (verão, outono, inverno e primavera) e submetidos a novos testes. Na Fig. 6, as colunas que possuem distribuições próximas da normal estão destacadas em negrito, e os testes que resultaram em p-valores maiores que o nível de significância adotado estão destacados em cinza.

FIGURA 6 – TESTES DE NORMALIDADE DOS DADOS DE METEOROLOGIA SEPARADOS POR ESTAÇÃO

Verão	Outono
***TempBulboSeco**	***TempBulboSeco**
- Normal: 0.6241267860977351	- Normal: 0.4285458136804614
- Shapiro: 0.661307156085968	- Shapiro: 0.3587791621685028
- Kolmogorov–Smirnov: 0.6263169406602667	- Kolmogorov–Smirnov: 0.8262506200836308
***TempBulboUmido**	***TempBulboUmido**
- Normal: 0.4947490055410101	- Normal: 0.0014687090751567644
- Shapiro: 0.3036584258079529	- Shapiro: 0.025821365416049957
- Kolmogorov–Smirnov: 0.006408408669017842	- Kolmogorov–Smirnov: 0.08322851458574458
***UmidadeRelativa**	***UmidadeRelativa**
- Normal: 0.5317257699767945	- Normal: 0.3752807507284129
- Shapiro: 0.4571700692176819	- Shapiro: 0.38942471146583557
- Kolmogorov–Smirnov: 1.1020275484137807e-07	- Kolmogorov–Smirnov: 5.926421155384419e-08
***PressaoAtmEstacao**	***PressaoAtmEstacao**
- Normal: 0.7347160238979067	- Normal: 0.6698580766423166
- Shapiro: 0.9666841626167297	- Shapiro: 0.39508405327796936
- Kolmogorov–Smirnov: 0.3787193365734262	- Kolmogorov–Smirnov: 0.21569421400818042
***VelocidadeVentoNebulosidade**	***VelocidadeVentoNebulosidade**
- Normal: 0.6277672181769303	- Normal: 0.645080285966642
- Shapiro: 0.6940707564353943	- Shapiro: 0.8423389196395874
- Kolmogorov–Smirnov: 0.02846498220390272	- Kolmogorov–Smirnov: 0.13927841670653646

Inverno	Primavera
***TempBulboSeco**	*TempBulboSeco
- Normal: 0.5003098135520574	- Normal: 0.025861553531967653
- Shapiro: 0.6369422078132629	- Shapiro: 0.0006836451939307153
- Kolmogorov–Smirnov: 0.04309511297235862	- Kolmogorov–Smirnov: 0.01247573625133762
***TempBulboUmido**	*TempBulboUmido
- Normal: 0.31870019482710865	- Normal: 1.062858288172069e-08
- Shapiro: 0.18121540546417236	- Shapiro: 6.401793939403433e-07
- Kolmogorov–Smirnov: 0.48966041245479014	- Kolmogorov–Smirnov: 0.023325148825244293
***UmidadeRelativa**	*UmidadeRelativa
- Normal: 0.2973110327281065	- Normal: 0.008334852022157226
- Shapiro: 0.27637779712677	- Shapiro: 6.15112639934523e-06
- Kolmogorov–Smirnov: 2.4569500220402207e-11	- Kolmogorov–Smirnov: 8.622356577958037e-16
***PressaoAtmEstacao**	***PressaoAtmEstacao**
- Normal: 0.7328733092695257	- Normal: 0.8528365455646644
- Shapiro: 0.940138578414917	- Shapiro: 0.5031830668449402
- Kolmogorov–Smirnov: 0.8062971653087325	- Kolmogorov–Smirnov: 0.4325358193116822
***VelocidadeVentoNebulosidade**	***VelocidadeVentoNebulosidade**
- Normal: 0.02716763129553984	- Normal: 0.7366476122582339
- Shapiro: 0.013667384162545204	- Shapiro: 0.807231068611145
- Kolmogorov–Smirnov: 0.09588789921746768	- Kolmogorov–Smirnov: 0.04304719472731229

FIGURA 7 – DISTRIBUIÇÕES DOS DADOS DE METEOROLOGIA SEPARADOS POR ESTAÇÃO



Após a realização de todos os testes de cada coluna de cada conjunto, os dados foram agregados em uma única tabela e salvos em um único arquivo para facilitar a implementação de cada algoritmo de análise que será discutido na próxima seção.

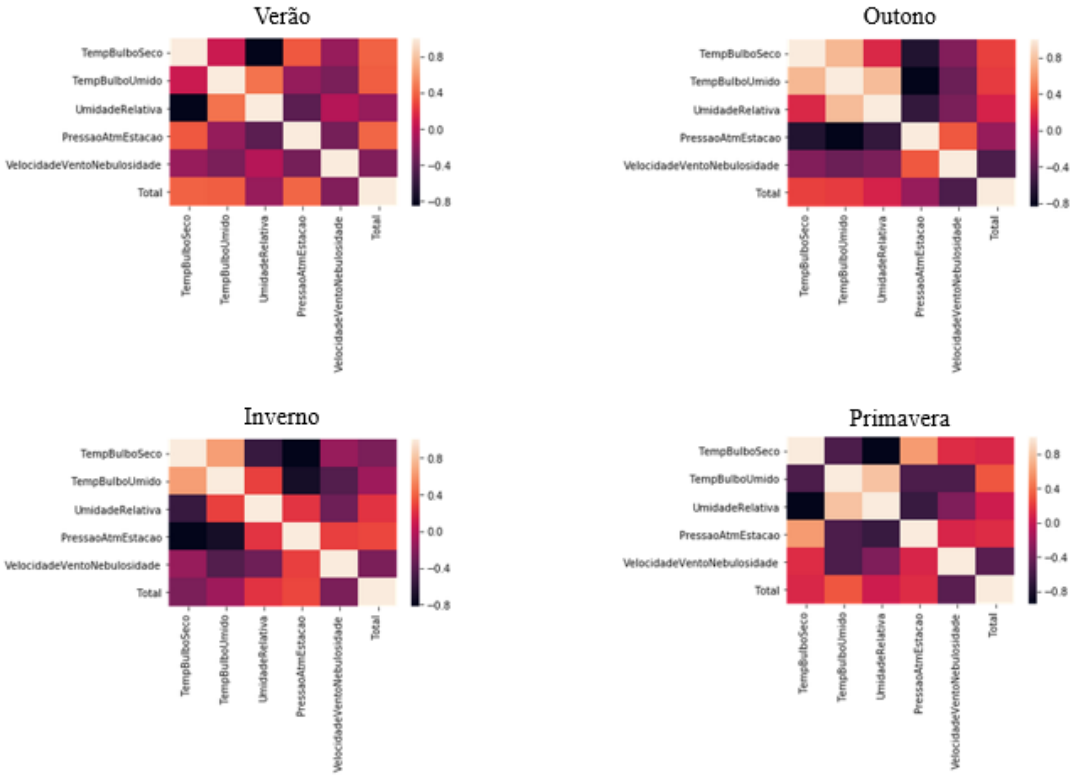
3.5. Análise dos Dados

3.5.1. Correlações entre as variáveis climáticas e o total de casos de dengue, separados por estação

Para entender o comportamento do total de casos de dengue de acordo com variações climáticas, uma matriz de correlações foi montada na forma de mapa de calor.

Dessa maneira, foram estudadas as correlações entre todas as variáveis climáticas, incluindo as interações de segunda e terceira ordem, juntamente com o total de casos de dengue. Os mapas de calor resultantes são mostrados na Fig. 9. É importante evidenciar que um mapa separado para cada estação foi gerado devido ao comportamento das variáveis climáticas, que, separadas por estação, possuem distribuições normais.

FIGURA 8 – MAPAS DE CALOR DE CORRELAÇÕES ENTRE AS VARIÁVEIS CLIMÁTICAS E O TOTAL DE CASOS DE DENGUE SEPARADOS POR ESTAÇÃO



Os mapas de calor são gerados combinando todas as variáveis duas a duas, dificultando assim sua leitura para um melhor entendimento. Dessa forma, novos gráficos foram gerados, comparando as correlações de cada variável climática somente com o total de casos de dengue, conforme mostra as Fig. 9, 10, 11 e 12. Além disso, uma tabela com o valor numérico de cada correlação foi gerada (Tabela 9).

FIGURA 9 – CORRELAÇÕES DE CADA VARIÁVEL CLIMÁTICA COM O TOTAL DE CASOS DE DENGUE - VERÃO

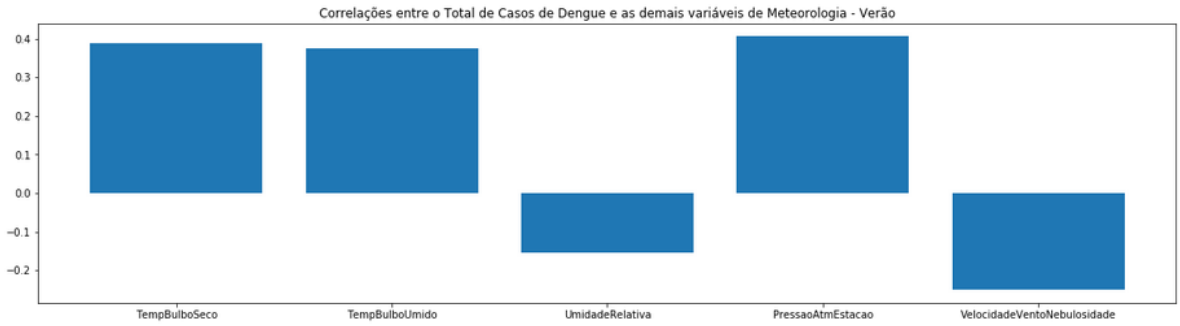


FIGURA 10 – CORRELAÇÕES DE CADA VARIÁVEL CLIMÁTICA COM O TOTAL DE CASOS DE DENGUE - OUTONO

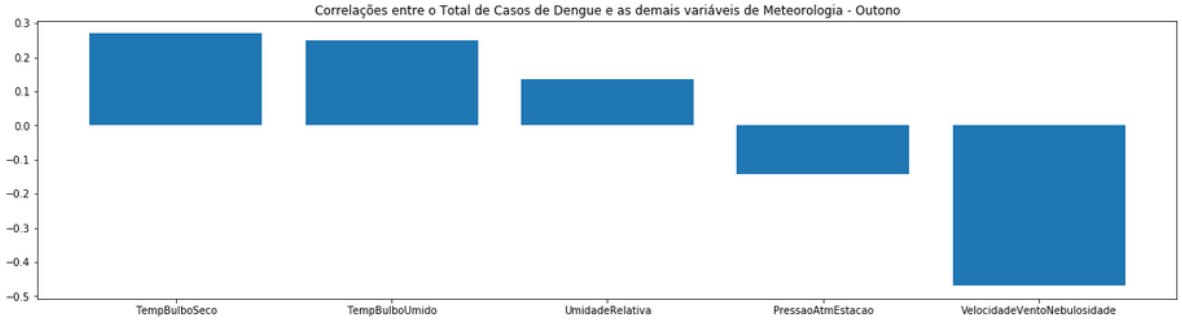


FIGURA 11 – CORRELAÇÕES DE CADA VARIÁVEL CLIMÁTICA COM O TOTAL DE CASOS DE DENGUE - INVERNO

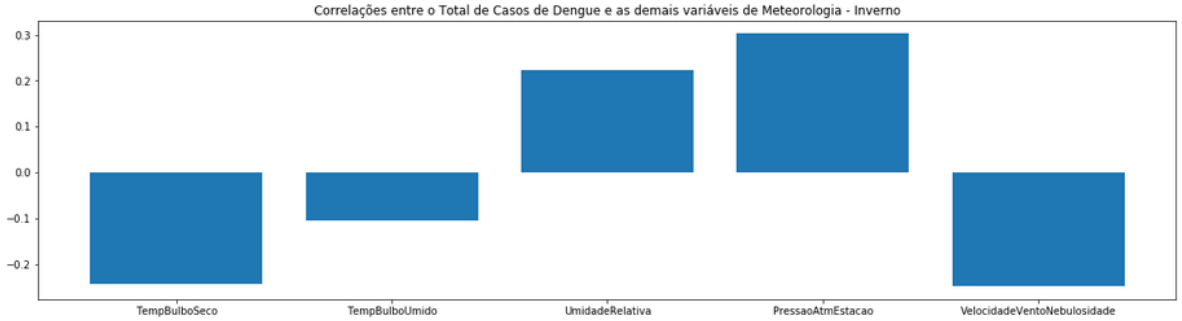


FIGURA 12 – CORRELAÇÕES DE CADA VARIÁVEL CLIMÁTICA COM O TOTAL DE CASOS DE DENGUE - PRIMAVERA

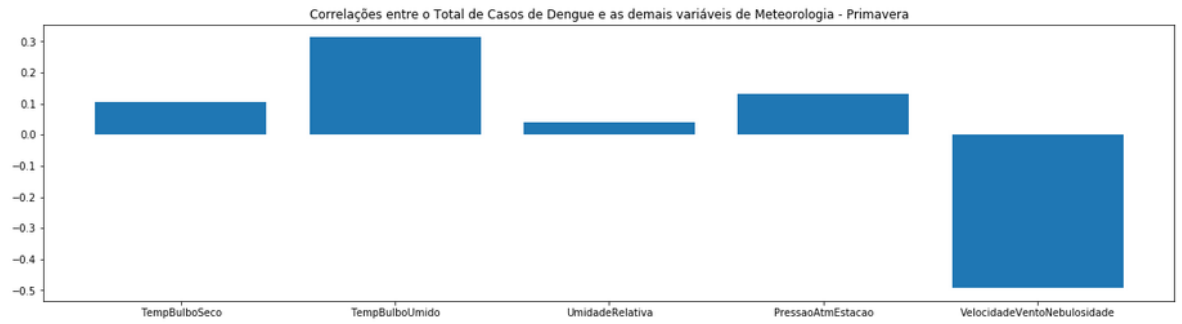


TABELA 9: CORRELAÇÕES ENTRE AS VARIÁVEIS CLIMÁTICAS E O TOTAL DE CASOS DE DENGUE, POR ESTAÇÃO

Variável Climática	Verão	Outono	Inverno	Primavera
TempBulboSeco	0,387	0,269	-0,244	0,105
TempBulboUmido	0,374	0,251	-0,105	0,313
UmidadeRelativa	-0,154	0,135	0,224	0,041
PressaoAtmEstacao	0,406	-0,143	0,303	0,133
VelocidadeVentoNebulosidade	-0,251	-0,470	-0,249	-0,494

É importante notar que, durante o verão, as variáveis que mais se correlacionam com o total de casos de Dengue são ‘PressaoAtmEstacao’, ‘TempBulboSeco’ e ‘TempBulboUmido’, de forma que, dentro do período analisado, valores elevados dessas três variáveis correspondem a quantidades elevadas de casos de dengue. Por sua vez, durante o outono, as correlações mais fortes foram entre os casos de dengue e as variáveis ‘VelocidadeVentoNebulosidade’, ‘TempBulboSeco’ e ‘TempBulboUmido’, sendo a primeira uma correlação negativa e as outras duas positivas. Isso significa que valores elevados dessas variáveis corresponderam a quantidades menores de casos de dengue.

Além disso, para o inverno, as correlações mais fortes são entre os casos de dengue e ‘PressaoAtmEstacao’, ‘UmidadeRelativa’, ‘TempBulboSeco’ e ‘VelocidadeVentoNebulosidade’, sendo que, para as duas primeiras, a correlação é positiva, permitindo que a mesma inferência feita para a estação de verão seja feita aqui e, para as duas últimas, a correlação é negativa, semelhante ao que aconteceu para o período de outono. Por último, para a primavera, as variáveis que apresentaram correlações mais fortes com o total de casos de dengue foram ‘VelocidadeVentoNebulosidade’ (negativa) e ‘TempBulboUmido’ (positiva).

Dessa maneira, tais correlações podem fornecer *insights* importantes em se tratando de ações para combater surtos de dengue, uma vez que, quando, principalmente, a variável ‘VelocidadeVentoNebulosidade’ sofrer alguma alteração negativa, o efeito esperado é um aumento da quantidade total de casos de dengue.

3.5.2. Regressão Linear e ANOVA

Como apresentado anteriormente, a ANOVA possui como suposição a normalidade dos dados, isto é, todas as variáveis envolvidas precisam possuir distribuições que se assemelham à normal. Além disso, os dados precisam ser independentes, ou seja, nenhuma variável pode ter influência direta na outra.

Logo, foram desenvolvidos algoritmos utilizando os casos totais de dengue após a transformação de *Box-Cox*, bem como os dados de meteorologia separados por estação, que, segundo a Fig. 6, possuem distribuições próximas à distribuição normal. Dessa forma, nas próximas subseções, onde serão apresentados os modelos de cada estação, os dados foram submetidos a uma inversa da transformação de *Box-Cox* a fim de construir gráficos de linha temporal com os valores reais de casos de dengue.

É importante notar que, para a primavera, as colunas ‘TempBulboSeco’, ‘TempBulboUmido’ e ‘UmidadeRelativa’ não passaram nos testes de normalidade. Além disso, as variáveis ‘TempBulboUmido’ no outono e ‘VelocidadeVentoNebulosidade’ no inverno passaram em apenas um dos três testes. Ainda assim, todas as variáveis foram incluídas, e um modelo separado para cada estação foi construído.

Conforme mostrado anteriormente, um modelo de regressão linear múltipla pode ser composto, além das variáveis independentes, de interações entre tais variáveis. Dessa forma, após a tentativa de criar modelos apenas com as variáveis independentes (Apêndice B), foram definidos modelos com interações entre todas as variáveis, combinadas em todas as ordens possíveis (como se trata de 5 variáveis, a interação de maior ordem é a quinta). É importante evidenciar aqui que, para facilitar a criação dessas interações e a implementação da regressão linear, as variáveis foram renomeadas segundo a Tabela 10.

TABELA 10: ALTERAÇÃO DO NOME DAS COLUNAS REFERENTES ÀS VARIÁVEIS CLIMÁTICAS

Nome Original	Novo Nome
TempBulboSeco	A
TempBulboUmido	B
UmidadeRelativa	C
PressaoAtmEstacao	D
VelocidadeVentoNebulosidade	E

Além disso, para os dados de temperatura, foi adotada a escala *Kelvin* e, portanto, os dados foram transformados adicionando-se o valor 273,15 a cada registro de temperatura. Então, após a mudança de escala de temperatura, foram gerados modelos constituídos por uma regressão linear múltipla (pelo método dos mínimos quadrados) bem como uma tabela de análise de variância (ANOVA).

Os resultados dos modelos criados serão discutidos nas sessões seguintes. Vale evidenciar que os coeficientes encontrados pelo algoritmo de regressão linear para cada modelo estão detalhados no Apêndice C.

3.5.2.1. Modelo para estação verão

Os resultados do modelo de regressão linear de quinta ordem implementado para a estação “verão” estão mostrados na Fig. 13, enquanto a Fig. 14 mostra o resultado gráfico do modelo, bem como a distribuição dos resíduos (gráfico da direita). O valor de R^2 ajustado (Adj.

R-squared) foi de 0.502 (ou 50,2%), ou seja, o modelo é capaz de explicar 50,2% dos casos de dengue baseado nas variáveis climáticas. Além disso, o p-valor da estatística F do modelo (Prob (F-statistic)) foi inferior ao nível de significância estabelecido para este trabalho (0.0152 < 0.05). Dessa maneira, o modelo, de modo geral, é significante para explicar o comportamento dos casos de dengue durante o verão.

FIGURA 13 – REGRESSÃO LINEAR MÚLTIPLA PARA OS DADOS PERTENCENTES AO PERÍODO DE VERÃO

OLS Regression Results

Dep. Variable:	Transformed	**R-squared:**	0.811
Model:	OLS	**Adj. R-squared:**	0.502
Method:	Least Squares	**F-statistic:**	2.624
Date:	Thu, 27 Jun 2019	**Prob (F-statistic):**	0.0152
Time:	20:17:50	**Log-Likelihood:**	-18.977
No. Observations:	51	**AIC:**	102.0
Df Residuals:	19	**BIC:**	163.8
Df Model:	31		
Covariance Type:	nonrobust		

FIGURA 14 – RESULTADOS DO AJUSTE DO MODELO DE REGRESSÃO LINEAR PARA A ESTAÇÃO DE VERÃO

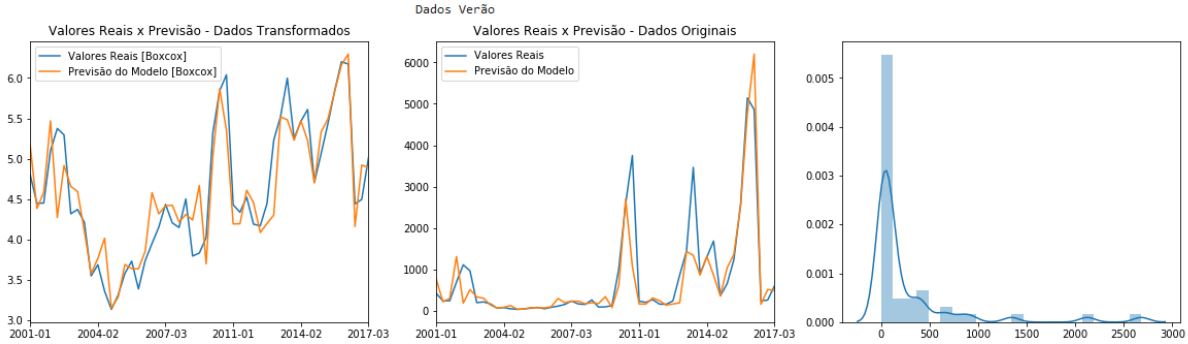


TABELA 11: ANÁLISE DE VARIÂNCIA DOS COEFICIENTES DO MODELO PARA A ESTAÇÃO VERÃO

Variáveis	F	PR(>F)
B:C:D:E	14,618507	0,001147
A:D:E	9,577509	0,005964
C:D:E	6,452338	0,019972
A:B:D	5,682064	0,027728
B:C:E	4,423926	0,048992

A Tabela 11 foi construída com as variáveis que apresentaram p-valores da estatística F (PR(>F)) menores que o nível de significância. Fica evidente, assim, que as variáveis mais significativas para explicar os casos de dengue durante o período de verão são, na verdade, interações entre as variáveis independentes envolvidas no modelo, sendo a mais significante uma combinação de quatro variáveis ('TempBulboUmido' – B, 'UmidadeRelativa' – C, 'PressaoAtmEstacao' – D, 'VelocidadeNebulosidadeVento' – E). O fato de as interações serem as variáveis que mais explicam os casos de dengue durante o verão dificultam sua análise, uma vez que é preciso estudar o comportamento de todas as variáveis individuais que compõem a interação.

3.5.2.2. Modelo para estação Outono

A Fig. 15 mostra os resultados obtidos com o modelo de regressão linear de quinta ordem implementado para a estação "outono", e a Fig. 16 mostra o resultado gráfico do modelo, bem como a distribuição dos resíduos (gráfico da direita). De acordo com a Fig. 15, o valor de R^2 ajustado (Adj. R-squared) foi de 0.597 (ou 59,7%), ou seja, o modelo é capaz de explicar 59,7% dos casos de dengue baseado nas variáveis climáticas. Além disso, o p-valor da estatística F do modelo (Prob (F-statistic)) foi inferior ao nível de significância estabelecido para este trabalho (0.00347). Dessa maneira, o modelo, de modo geral, é significante para explicar o comportamento dos casos de dengue durante o outono.

FIGURA 15 – REGRESSÃO LINEAR MÚLTIPLA PARA OS DADOS PERTENCENTES AO PERÍODO DE OUTONO

OLS Regression Results

Dep. Variable:	Transformed	**R-squared:**	0.847
Model:	OLS	**Adj. R-squared:**	0.597
Method:	Least Squares	**F-statistic:**	3.393
Date:	Fri, 28 Jun 2019	**Prob (F-statistic):**	0.00347
Time:	00:08:34	**Log-Likelihood:**	-27.585
No. Observations:	51	**AIC:**	119.2
Df Residuals:	19	**BIC:**	181.0
Df Model:	31		
Covariance Type:	nonrobust		

FIGURA 16 – RESULTADOS DO AJUSTE DO MODELO DE REGRESSÃO LINEAR PARA A ESTAÇÃO DE OUTONO

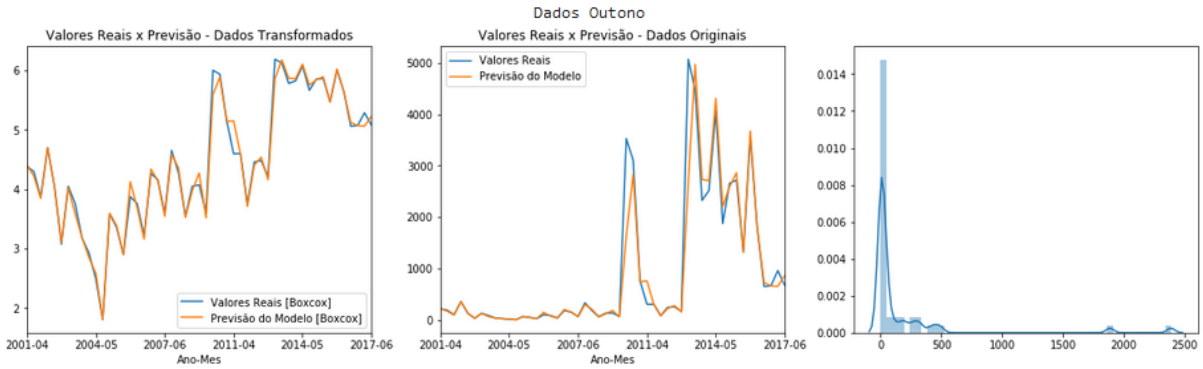


TABELA 12: ANÁLISE DE VARIÂNCIA DOS COEFICIENTES DO MODELO PARA A ESTAÇÃO OUTONO

Variáveis	F	PR(>F)
E	27,555823	0,000046
D	7,818934	0,011518
A:B	5,791666	0,026440
A:B:D:E	5,740223	0,027036

A Tabela 12 foi construída com as variáveis que apresentaram p-valores da estatística F (PR(>F)) menores que o nível de significância, evidenciando assim as variáveis mais significativas para explicar os casos de dengue durante o período do outono. Dentre tais variáveis, destacam-se ‘PressaoAtmEstacao’ (D) e ‘VelocidadeVentoNebulosidade’ (E), que são as variáveis independentes sem interação que se mostraram significativas para explicar o comportamento da dengue no período analisado. Além delas, a interação entre ‘TempBulboSeco’ e ‘TempBulboUmido’ também é significante para o período de outono.

Buscando uma justificativa para as variáveis climáticas que foram apontadas como significativas, foi feito um levantamento bibliográfico a respeito das características do mosquito transmissor da doença, *Aedes aegypt*. Foi descoberto, então, que tal mosquito normalmente se reproduz durante o voo [15]. Além disso, os ovos do mosquito eclodem em ambientes úmidos (com água parada) [15].

Dessa forma, o fato de o *Aedes aegypt* se reproduzir durante o voo pode explicar o motivo de a variável ‘VelocidadeVentoNebulosidade’ ser significativa. Além disso, a interação significativa entre ‘TempBulboSeco’ e ‘TempBulboUmido’ pode estar relacionada com as condições de eclosão dos ovos do mosquito (ocorre em ambientes úmidos, com água parada),

já que a interação entre tais variáveis de temperatura pode estar relacionada à taxa de evaporação da água.

3.5.2.3. Modelo para estação Inverno

A Fig. 17 mostra os resultados obtidos com o modelo de regressão linear de quinta ordem implementado para a estação "Inverno". Por sua vez, a Fig. 18 mostra o resultado gráfico do modelo, bem como a distribuição dos resíduos (gráfico da direita). De acordo com a Fig. 17, o valor de R^2 ajustado (Adj. R-squared) foi de 0.223 (ou 22,3%), ou seja, o modelo é capaz de explicar 22,3% dos casos de dengue baseado nas variáveis climáticas. Além disso, o p-valor da estatística F do modelo (Prob (F-statistic)) foi inferior ao nível de significância estabelecido para este trabalho (0.0380). Dessa maneira, o modelo, de modo geral, não é significante para explicar o comportamento dos casos de dengue durante o inverno.

FIGURA 17 – REGRESSÃO LINEAR MÚLTIPLA PARA OS DADOS PERTENCENTES AO PERÍODO DE INVERNO

OLS Regression Results

Dep. Variable:	Transformed	**R-squared:**	0.705
Model:	OLS	**Adj. R-squared:**	0.223
Method:	Least Squares	**F-statistic:**	1.463
Date:	Fri, 28 Jun 2019	**Prob (F-statistic):**	0.194
Time:	00:43:10	**Log-Likelihood:**	-37.381
No. Observations:	51	**AIC:**	138.8
Df Residuals:	19	**BIC:**	200.6
Df Model:	31		
Covariance Type:	nonrobust		

FIGURA 18 – RESULTADOS DO AJUSTE DO MODELO DE REGRESSÃO LINEAR PARA A ESTAÇÃO DE INVERNO

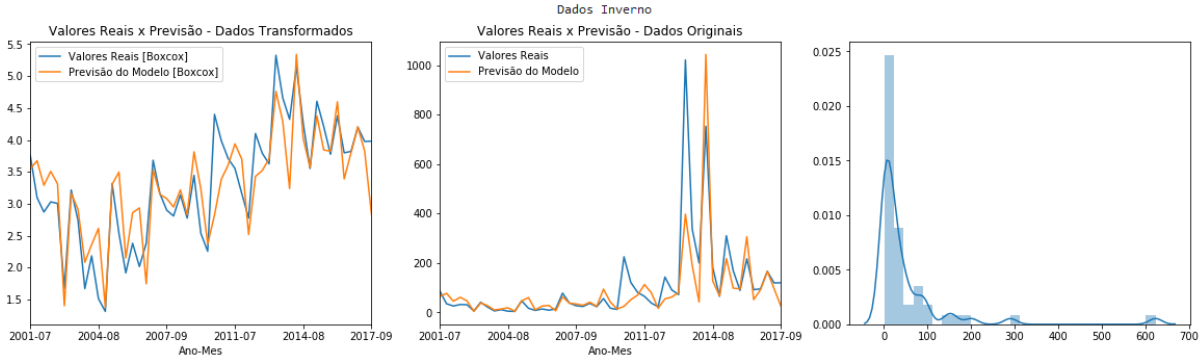


TABELA 13: ANÁLISE DE VARIÂNCIA DOS COEFICIENTES DO MODELO PARA A ESTAÇÃO INVERNO

Variáveis	F	PR(>F)
B:C:E	5,874659	0,02551

Como pode ser visto na Tabela 13, a única variável que apresentou p-valor da estatística F (PR(>F)) menor que o nível de significância foi a interação entre 'TempBulboUmido' (B), 'UmidadeRelativa' (C) e 'VelocidadeNebulosidadeVento' (E). Porém, como o modelo não é significativo, qualquer afirmação a respeito de tal interação pode estar equivocada.

3.5.2.4. Modelo para estação Primavera

A Fig. 19 mostra os resultados obtidos com o modelo de regressão linear de quinta ordem implementado para a estação "primavera", enquanto a Fig. 20 mostra o resultado gráfico do modelo, bem como a distribuição dos resíduos (gráfico da direita). De acordo com a Fig. 19, o valor de R^2 ajustado (Adj. R-squared) foi de 0.301 (ou 30,1%), ou seja, o modelo é capaz de explicar 30,1% dos casos de dengue baseado nas variáveis climáticas. Além disso, o p-valor da estatística F do modelo (Prob (F-statistic)) foi superior ao nível de significância estabelecido para este trabalho (0.141). Dessa maneira, o modelo, de modo geral, não é significante para explicar o comportamento dos casos de dengue durante a primavera.

FIGURA 19 – REGRESSÃO LINEAR MÚLTIPLA PARA OS DADOS PERTENCENTES AO PERÍODO DE PRIMAVERA

OLS Regression Results

Dep. Variable:	Transformed	**R-squared:**	0.734
Model:	OLS	**Adj. R-squared:**	0.301
Method:	Least Squares	**F-statistic:**	1.695
Date:	Fri, 28 Jun 2019	**Prob (F-statistic):**	0.115
Time:	00:49:03	**Log-Likelihood:**	-27.161
No. Observations:	51	**AIC:**	118.3
Df Residuals:	19	**BIC:**	180.1
Df Model:	31		
Covariance Type:	nonrobust		

FIGURA 20 – RESULTADOS DO AJUSTE DO MODELO DE REGRESSÃO LINEAR PARA A ESTAÇÃO DE PRIMAVERA

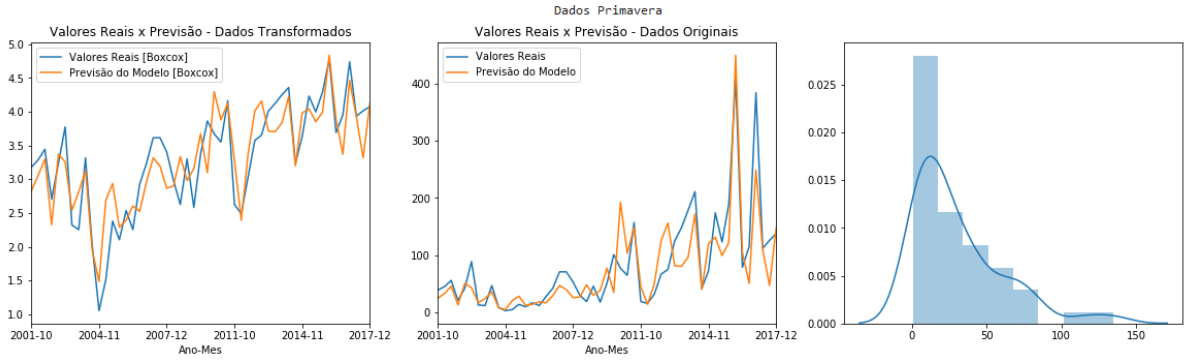


TABELA 14: ANÁLISE DE VARIÂNCIA DOS COEFICIENTES DO MODELO PARA A ESTAÇÃO PRIMAVERA

Variáveis	F	PR(>F)
E	6,828818	0,017096
A:D:E	6,778969	0,017449
C:D:E	6,173999	0,022450
B:D:E	5,944422	0,024757
A:C:E	4,882502	0,039603

A Tabela 14 foi construída com as variáveis que apresentaram p-valores da estatística F (PR(>F)) menores que o nível de significância, evidenciando assim as variáveis mais significativas para explicar os casos de dengue durante o período de primavera. Destaca-se aqui a 'VelocidadeVentoNebulosidade'. Porém, assim como para o modelo referente ao inverno, qualquer afirmação sobre a relação entre os casos de dengue e tal variável, ou entre os casos de dengue e as demais variáveis de interação, pode estar equivocada, já que o modelo não é significante para explicar a variável resposta.

3.5.2.5. Considerações sobre os Modelos

Diante dos resultados apresentados nas seções anteriores, é importante discutir o modelo criado para as estações inverno e primavera. Diferentemente das outras duas estações, a regressão linear implementada gerou um modelo que não é significante para explicar os casos de dengue, já que o p-valor da estatística F é superior ao nível de significância estatística. Logo, não existem evidências estatísticas suficientes para rejeitar a hipótese H_0 da ANOVA, ou seja, as variações entre os dados climáticos e os registros de casos de dengue durante a primavera podem ser aleatórias.

É importante ressaltar que, para os modelos referentes ao verão e ao outono, ainda que os valores de R^2 ajustado tenham sido superiores aos modelos do inverno e da primavera e os p-valores das estatísticas F dos modelos tenham sido inferiores à significância, rejeitando assim H_0, tal análise é válida apenas para o período de janeiro de 2001 a dezembro de 2017 e para a cidade de Brasília, ou seja, qualquer suposição a respeito das relações encontradas por esses dois modelos (verão e outono) fora deste período pode estar equivocada. Além disso, as variáveis significantes encontradas precisam ser exploradas com mais detalhes, bem como as características do mosquito *Aedes aegypt* precisam ser mais estudadas a fim de se buscar justificativas para os resultados encontrados e validar os modelos.

3.5.3. Análise do Repasse de Verbas destinadas ao combate de Epidemiologias diante do comportamento dos casos de Dengue

Conforme mostrado anteriormente, foram extraídos dados de repasse de verbas destinados ao combate de epidemiologias. Tais dados foram coletados a fim de se analisar o comportamento do investimento público diante da evolução dos casos de dengue, e avaliar o impacto dessas verbas na quantidade de casos de dengue. Vale notar aqui que a verba analisada é destinada ao combate de epidemiologias de um modo geral, não necessariamente à dengue. Logo, qualquer suposição feita pode ser equivocada.

Como o repasse de verbas é anual, os dados de casos de dengue foram agregados anualmente. Além disso, para conduzir a análise, os dados de repasse de verbas foram associados aos casos de dengue do ano anterior (por exemplo, os repasses de 2002 foram associados aos casos de 2001, e assim sucessivamente).

Tal decisão foi tomada devido ao intuito de analisar o comportamento do investimento público, considerando que, em 2010, por exemplo, a decisão da quantidade de verba destinada ao combate de epidemiologias que foi repassada se deu devido à quantidade de casos de dengue registrados em 2009. Dessa maneira, foi implementado um mapa de calor para avaliar a correlação entre tais variáveis, como mostrado na Fig. 21 e na Tabela 15.

FIGURA 21 – MAPA DE CALOR DAS CORRELAÇÕES ENVOLVENDO A QUANTIDADE DE CASOS DE DENGUE DO ANO ANTERIOR E O REPASSE DE VERBAS DESTINADAS AO COMBATE DE EPIDEMIOLOGIAS DO ANO CORRENTE

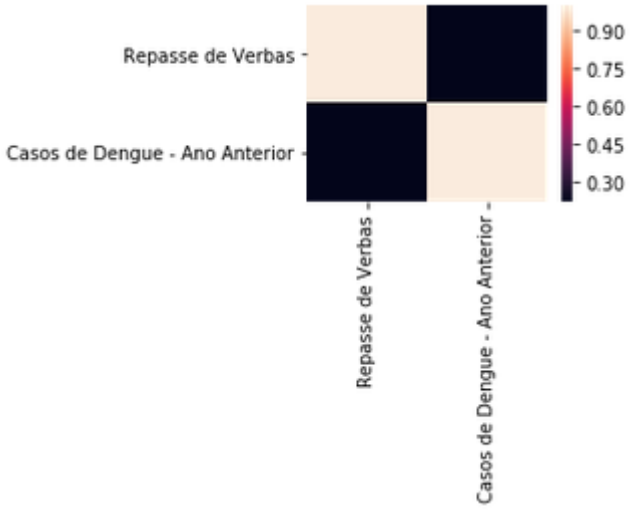
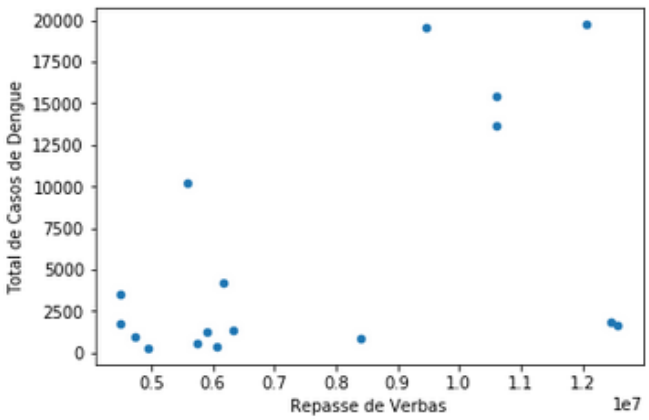


TABELA 15: QUANTIFICAÇÃO DAS CORRELAÇÕES ENVOLVENDO A QUANTIDADE DE CASOS DE DENGUE DO ANO ANTERIOR E O REPASSE DE VERBAS DESTINADAS AO COMBATE DE EPIDEMIOLOGIAS DO ANO CORRENTE

Matriz de Correlações	Repasse de Verbas	Casos de Dengue - Ano Anterior
Repasse de Verbas	1.000000	0.223461
Casos de Dengue - Ano Anterior	0.223461	1.000000

FIGURA 22 – GRÁFICO DE DISPERSÃO ENTRE A QUANTIDADE DE CASOS DE DENGUE DO ANO ANTERIOR E O REPASSE DE VERBAS DESTINADAS AO COMBATE DE EPIDEMIOLOGIAS DO ANO CORRENTE



O gráfico de dispersão da Fig. 22 sugere que, para alguns anos, quando a quantidade de casos de dengue do ano anterior foi pequena, o repasse de verbas foi menor e, quando um maior número de casos foi registrado no ano anterior, o repasse de verbas também foi maior. Porém, conforme já discutido, afirmar que a quantidade de verba repassada em um ano corrente está diretamente ligada aos casos de dengue registrados no ano anterior pode ser equivocada, uma vez que diversos outros fatores desconhecidos podem influenciar na decisão de repasse de verbas, além de tal repasse ser destinado ao combate de epidemiologias de modo geral.

Ainda assim, tal análise de verbas se faz importante para discutir o investimento em políticas públicas de combate a epidemiologias, já que, ao combinar com modelos de predição, como os apresentados na seção anterior, pode-se otimizar o uso de tal investimento. Essa otimização pode ser alcançada com base no comportamento do clima em cada estação, uma vez que, ao identificar alterações nas variáveis climáticas, é possível estimar o reflexo disso na quantidade de casos de dengue para, assim, tomar ações preventivas utilizando a verba pública destinada ao combate de epidemiologias.

3.5.4. Análise da influência de Fatores Demográficos na quantidade de casos de Dengue

Os dados referentes à dengue, quando extraídos, foram consultados separando-os por características gerais da população, como faixa etária, sexo e zona de residência. Como apresentado na Introdução (1), um dos objetivos do presente trabalho é determinar se tais características demográficas são fatores determinantes para a quantidade de casos de dengue. Em outras palavras, busca-se analisar se alguma categoria dentro, por exemplo, de faixa etária (como pessoas de 60 a 69 anos), apresenta uma quantidade de registros de casos de dengue significativa.

Deste modo, foram conduzidas análises de variância (ANOVA) envolvendo os casos de dengue e os dados demográficos, ambos separados por cada característica da população. Para tal, os dados de dengue foram primeiramente agrupados anualmente somando-se a quantidade de casos, já que as estimativas de dados populacionais feitas a partir dos censos de 2000 e 2010 são anuais. Além disso, novas faixas etárias foram definidas combinando as já existentes (mostradas na Tabela 5), a saber: 0 a 19 anos, 20 a 39 anos, 40 a 59 anos, 60 a 79 anos e 80 anos e acima.

Em seguida, foram calculadas probabilidades dentro de cada conjunto de dados. Para exemplificar, no ano de 2010, a quantidade total de casos de dengue foi de 15471, sendo 14852

de pessoas que residem na zona urbana. Assim, a probabilidade de residir em uma área urbana quando se tem dengue é de aproximadamente 96%.

Após o cálculo das probabilidades, foi definida uma coluna denominada ‘Grupo’ em cada conjunto de dados, sendo que para os registros de dengue, foi atribuída a *string* ‘Dengue’ e, para os do censo, foi atribuído ‘Censo’. Em seguida, os dados foram unificados em um único conjunto de dados, e análises de variância foram conduzidas comparando cada coluna individualmente com a coluna ‘Grupo’, a fim de testar a hipótese H_0 da ANOVA de que não existem diferenças entre as médias dos grupos, objetivando verificar se as variações entre as probabilidades da população de um modo geral e as probabilidades dos casos de dengue são significativas ou são simplesmente aleatórias.

Dessa maneira, conforme mostrado na Tabela 16, a hipótese H_0 foi aceita para as variáveis ‘Masculino’, ‘Feminino’ e ’70 - 79’ (já que o p-valor foi maior que o nível de significância estatística), o que significa que não existem diferenças entre as médias de probabilidade de cada um dos grupos (‘Dengue’ e ‘Censo’). Em outras palavras, ter dengue ou não é uma questão aleatória analisando-se o sexo do indivíduo ou o fato de ele pertencer à faixa etária de 60 a 79 anos.

TABELA 17: RESULTADOS DA ANOVA ENTRE OS GRUPOS DE DENGUE E CENSO ONDE H_0 FOI ACEITA

Característica	Estatística F	P-valor
Masculino	0,313977	0,579152
Feminino	0,673754	0,417819
60 - 79	0,243995	0,624709

Por outro lado, as demais variáveis apresentadas na Tabela 18 tiveram a hipótese H_0 rejeitada, implicando que existe diferença entre as médias de probabilidade dos dois grupos (dengue e censo), ou seja, pertencer a alguma categoria dentre as variáveis demográficas em questão implica em uma probabilidade maior ou menor de se contrair dengue.

TABELA 18: RESULTADOS DA ANOVA ENTRE OS GRUPOS DE DENGUE E CENSO ONDE H_0 FOI REJEITADA

Característica	Estatística F	P-valor
Urbana	21,265523	6,124585e-05
Rural	7,196841	1,145961e-02
0 - 19	52,193474	3,313014e-08
20 - 39	28,908419	6,640000e-06
40 - 59	19,972772	9,226730e-05
80 acima	31,123821	3,685699e-06

Assim, para analisar se as características mostradas na Tabela 18 contribuem para uma maior ou menor probabilidade de se ter dengue, foi feita uma média geral para as probabilidades de cada grupo, e os resultados foram colocados na Tabela 19.

Como pode ser observado, tanto o fato de morar em uma área urbana quanto em uma rural são significativos para uma menor probabilidade de se ter dengue. Porém, tais números alcançados podem ser inconsistentes, uma vez que o esperado é que morar em uma das regiões seja favorável à probabilidade de se ter dengue, e morar em outra não. Tal inconsistência pode ser advinda dos casos registrados em região periurbana, que foram combinados com a categoria urbana. Entretanto, investigações mais detalhadas precisam ser feitas antes que tal suspeita seja afirmada.

Por sua vez, as faixas etárias de 0 a 19 e 80 acima são menos propícias a se ter dengue. Já indivíduos pertencentes às faixas etárias de 20 a 39 e 40 a 59 anos possuem maior probabilidade de contrair dengue.

TABELA 19: COMPARAÇÃO ENTRE AS PROBABILIDADES DE SE PERTENCER A UMA CATEGORIA DE MODO GERAL E TENDO DENGUE

Grupo	Urbana	Rural	0 - 19	20 - 39	40 - 59	80 acima
Censo	96,45%	3,55%	32,98%	38,09%	21,48%	0,84%
Dengue	88,13%	2,64%	23,29%	44,67%	25,21%	0,40%

Tal *insight* obtido quanto às categorias mais propícias de contraírem dengue pode ser útil em se tratando de otimizar os gastos de repasses destinados ao combate de epidemiologias, uma vez que podem ser feitos investimentos específicos para combater a dengue para indivíduos que estão nas faixas etárias 20 a 39 anos e 40 a 59 anos.

4. CONSIDERAÇÕES FINAIS

Com o desenvolvimento deste trabalho, foi possível explorar diversas ferramentas de análise de dados, buscando um entendimento do processo como um todo, desde a extração dos dados até a obtenção de modelos e resultados, passando pelas transformações necessárias para melhor ajustá-los aos modelos. Durante todo esse processo, diversas dificuldades com relação aos dados foram encontradas.

A primeira delas foi relacionada a limites de dados disponíveis, já que os dados de casos de dengue disponibilizados pelo SUS compreendem apenas o período de janeiro de 2001 a dezembro de 2017. Além disso, questões relacionadas à granularidade dos dados precisaram ser trabalhadas, já que os dados de meteorologia são registros coletados três vezes ao dia, enquanto os registros de casos de dengue são mensais. Por sua vez, os dados de repasse de verba destinados ao combate de epidemiologias são registros anuais e, deste modo, para realizar as análises com tais dados, os registros de casos de dengue precisaram ser agrupados anualmente.

Outra dificuldade que foi enfrentada refere-se à ausência de dados. Os dados referentes às variáveis climáticas apresentaram registros ausentes para a coluna 'TempBulboUmido' para os meses de outubro, novembro e dezembro de 2017. Por sua vez, os dados dos censos demográficos oferecem estatísticas da população apenas de 10 em 10 anos, de forma que foi necessário fazer estimativas sobre a população de Brasília para os demais anos além de 2000 e 2010.

Ademais, outra questão importante é referente às variáveis de cada conjunto de dados. Um desafio encontrado nesse sentido está relacionado à zona de residência. Para os registros de dengue, existem três tipos, a saber: urbana, rural e periurbana. Porém, os dados do censo contabilizaram como zonas de residência dos indivíduos apenas urbana e rural. Por isso, foi necessário agregar os casos da região periurbana aos da região urbana. Vale ressaltar que tal decisão foi feita como pressuposto pelo autor, podendo levar a inconsistências como a apresentada na seção 3.5.4 para as probabilidades referentes às zonas de residência.

Uma vez contornadas todas as dificuldades expostas, foram construídos os modelos de regressão linear múltipla de cada estação. Vale lembrar que os modelos construídos para a primavera e para o inverno não são significantes para explicar os casos que foram registrados durante a estação. Dessa maneira, qualquer variação nas variáveis climáticas e suas interações para esses dois períodos com relação às variações nos casos de dengue podem ser simplesmente aleatórias.

É importante notar que, além das dificuldades encontradas envolvendo os dados durante a condução de todas as análises, é preciso estar ciente de que os números dos casos de dengue de Brasília podem não necessariamente refletir a realidade, já que, como os registros de dengue são coletados em todas as unidades de saúde de Brasília, indivíduos que ficaram doentes durante o período analisado podem não procurar um posto de saúde, deixando de ser contabilizados, ou podem ocorrer erros na contagem em cada localidade. Ainda assim, como os dados disponibilizados pelo SUS são os dados a que o autor teve acesso, as análises e resultados obtidos foram construídos a partir deles.

Dessa maneira, após todas as análises feitas com ferramentas estatísticas e os respectivos desafios encontrados durante o desenvolvimento, o autor conclui que a ciência de dados é uma área de grande potencial para se obter *insights* a partir de dados com a construção de modelos estatísticos, porém, é importante evidenciar que muitas dificuldades e limitações existem durante esse processo quando lidando com dados problemáticos.

Por fim, vale mostrar que os resultados e conclusões atingidos durante o desenvolvimento desta pesquisa se aplicam apenas a Brasília e ao período de 2001 a 2017. Qualquer inferência feita para outra cidade ou para a cidade de Brasília, mas fora do período em questão, é equivocada, uma vez que o comportamento dos casos de dengue pode ser completamente diferente. Ainda assim, este trabalho constitui uma prova de conceito, ou seja, os algoritmos aqui implementados podem ser replicados utilizando os dados de outras cidades, além das análises poderem ser feitas novamente para Brasília, mas acrescentando dados de períodos que não foram trabalhados durante este estudo.

Portanto, a pesquisa em questão propõe que trabalhos futuros possam ser feitos empregando os mesmos algoritmos para cidades diferentes, a fim de buscar relações entre os casos de dengue e as condições climáticas de diferentes localidades. Além disso, questões relacionadas aos fundamentos das variáveis climáticas e ao comportamento do ciclo de vida do mosquito *Aedes aegypt* podem ser exploradas com mais detalhes, a fim de buscar um entendimento correto a respeito de como tais variáveis e suas interações influenciam nos casos de dengue.

Pode-se, ainda, continuar atualizando os modelos com dados mais recentes, com o intuito de construir modelos capazes de entender o presente e fazer projeções para o futuro. Por último, novos algoritmos podem ser implementados com novas variáveis para compreender, com mais precisão, o comportamento dos casos de dengue, objetivando adotar medidas preventivas para o controle da doença.

REFERÊNCIAS

[1] GRUS, Joel. Data Science from Scratch. 1005 Gravenstein Highway North, Sebastopol, CA: O'Reilly Media, 2015. E-book.

[2] HASTINGS, N.A.J.; PEACOCK, J.B. Statistical distributions: a handbook for students and practitioners. New York: J. Wiley, 1975. 130p

[3] SPIEGEL, R.A.; SCHILLER, J.; SRINIVASAN, R.A. Probabilidade e estatística. 2.ed. Porto Alegre: Bookman, 2004. 398p

[4] Callegari-Jacques, Sídia M. (2003) Bioestatística: princípios e aplicações. Artmed. Porto Alegre

[5] SPIEGEL, Murray R.; SCHILLER, John J.; SNIRIVASAN, R. Alu. Probability and Statistics. United States: The McGraw-Hill Companies, 2013. E-book.

[6] D'Agostino, R. B. (1971), "An omnibus test of normality for moderate and large sample size", Biometrika, 58, 341-348

[7] D'Agostino, R. and Pearson, E. S. (1973), "Tests for departure from normality", Biometrika, 60, 613-622

[8] Shapiro, S. S. & Wilk, M.B (1965). An analysis of variance test for normality (complete samples), Biometrika, Vol. 52, pp. 591-611.

[9] Kolmogorov A (1933). "Sulla determinazione empirica di una legge di distribuzione". G. Ist. Ital. Attuari. 4: 83–91.

[10] Smirnov N (1948). "Table for estimating the goodness of fit of empirical distributions". Annals of Mathematical Statistics. 19 (2): 279
281. doi:10.1214/aoms/1177730256.

[11] GRIFFITHS, Dawn. Head First Statistics. 1005 Gravenstein Highway North, Sebastopol, CA 95472: O'Reilly Media, 2009. E-book.

[12] RODRIGUES, Sandra Cristina Antunes. Modelo de Regressão Linear e suas Aplicações. 2012. Relatório de Estágio para obtenção do Grau de Mestre (Mestrado) - Covilhã, 2012. E-book.

[13] ANOVA (2-way, N-way). Disponível em: https://pythonfordatascience.org/anova-2-way-n-way/. Acesso em: 16 maio 2019.

[14] DENGUE: causas, sintomas, tratamento e prevenção. Disponível em: http://www.saude.gov.br/saude-de-a-z/dengue. Acesso em: 8 abr. 2019.

[15] SANTOS, Vanessa Sardinha dos. "Ciclo de vida do Aedes aegypti"; Brasil Escola. Disponível em: https://brasilescola.uol.com.br/animais/ciclo-vida-aedes-aegypti.htm. Acesso em 28 de junho de 2019.

[16] DENGUE: um dos principais problemas de saúde pública no Brasil e no mundo. 24 abr. 2014. Disponível em: https://iis.org.br/farol-da-ilha/dengue-um-dos-principais-problemas-de-saude-publica-no-brasil-e-no-mundo/. Acesso em: 8 abr. 2019.

[17] INFORMAÇÕES de Saúde. Disponível em: http://tabnet.datasus.gov.br/. Acesso em: 22 mar. 2019.

[18] BDMEP - Banco de Dados Meteorológicos para Ensino e Pesquisa. Disponível em: http://www.inmet.gov.br/portal/index.php?r=bdmep/bdmep. Acesso em: 26 mar. 2019.

[19] CENSO Demográfico - IBGE. Disponível em: https://www.ibge.gov.br/estatisticas/sociais/populacao/9662-censo-demografico-2010.html?edicao=9749&t=o-que-e. Acesso em: 26 mar. 2019.

[20] FUNDO Nacional de Saúde - Consulta. [S. l.], 11 jun. 2019. Disponível em: https://consultafns.saude.gov.br/#/consolidada. Acesso em: 7 maio 2019.

[21] Francisco Mendonça e Inês Moresco Danni-Oliveira, Climatologia - Noções Básicas e Climas do Brasil, página 58, 2007, Editora Oficina de Textos

[22] G.E.P. Box and D.R. Cox, “An Analysis of Transformations”, Journal of the Royal Statistical Society B, 26, 211-252 (1964).

[23] Çetinkaya-Rundel, 2015 - Statistical Analysis and Inference Course / Diez et al. 2015, OpenIntro Statistics

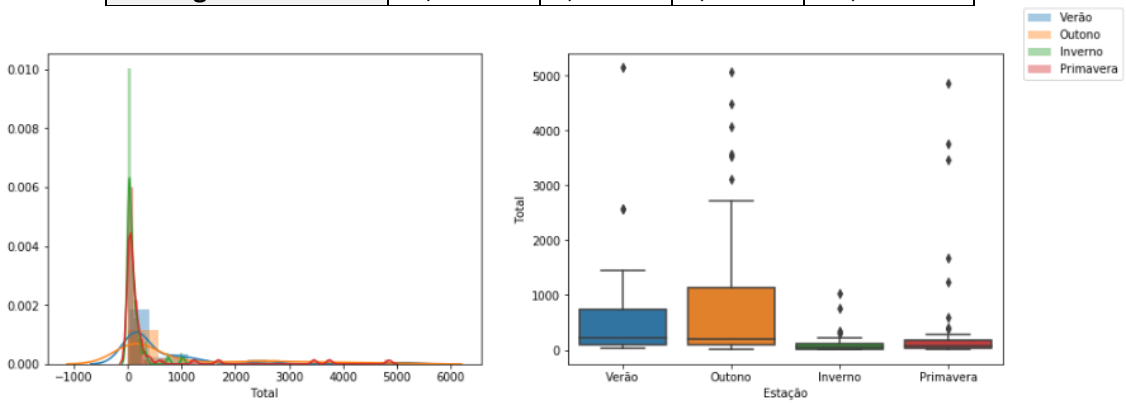
[24] VAZ, Welton. Saiba mais como o python surgiu e qual o seu cenário atual. [S. l.], 1 fev. 2018. Disponível em: https://www.eusoudev.com.br/python-como-surgiu/. Acesso em: 3 abr. 2019.

APÊNDICE A

A.1 Testes de Normalidade sobre os casos de Dengue separados por Estação

Como se trata de registros mensais, as estações foram separadas, grosso modo, como: janeiro, fevereiro e março (verão), abril, maio e junho (outono), julho, agosto e setembro (inverno), outubro, novembro e dezembro (primavera). Os resultados de tal separação são mostrados abaixo.

Teste	Verão	Outono	Inverno	Primavera
Normal	4,07E-12	4,95E-05	1,33E-15	6,42E-18
Shapiro	1,80E-09	5,47E-09	2,39E-11	9,95E-15
Kolmogorov–Smirnov	3,59E-20	7,73E-28	7,73E-28	1,64E-55



A.2 Testes de Normalidade sobre os casos de Dengue separados por Faixa Etária, por Sexo e por Zona de Residência

Além dos casos totais de dengue, foram realizados testes de normalidade nos dados separados por faixa etária, por sexo e por zona de residência. Porém, nenhuma das distribuições se aproximou da normal. Assim, a transformação de Box-Cox, aplicada sobre os casos totais de dengue, também foi aplicada aos outros cenários, e os resultados são mostrados a seguir.

TABELA A.1: TESTES DE NORMALIDADE REALIZADOS SOBRE OS CASOS DE DENGUE SEPARADOS POR FAIXA ETÁRIA

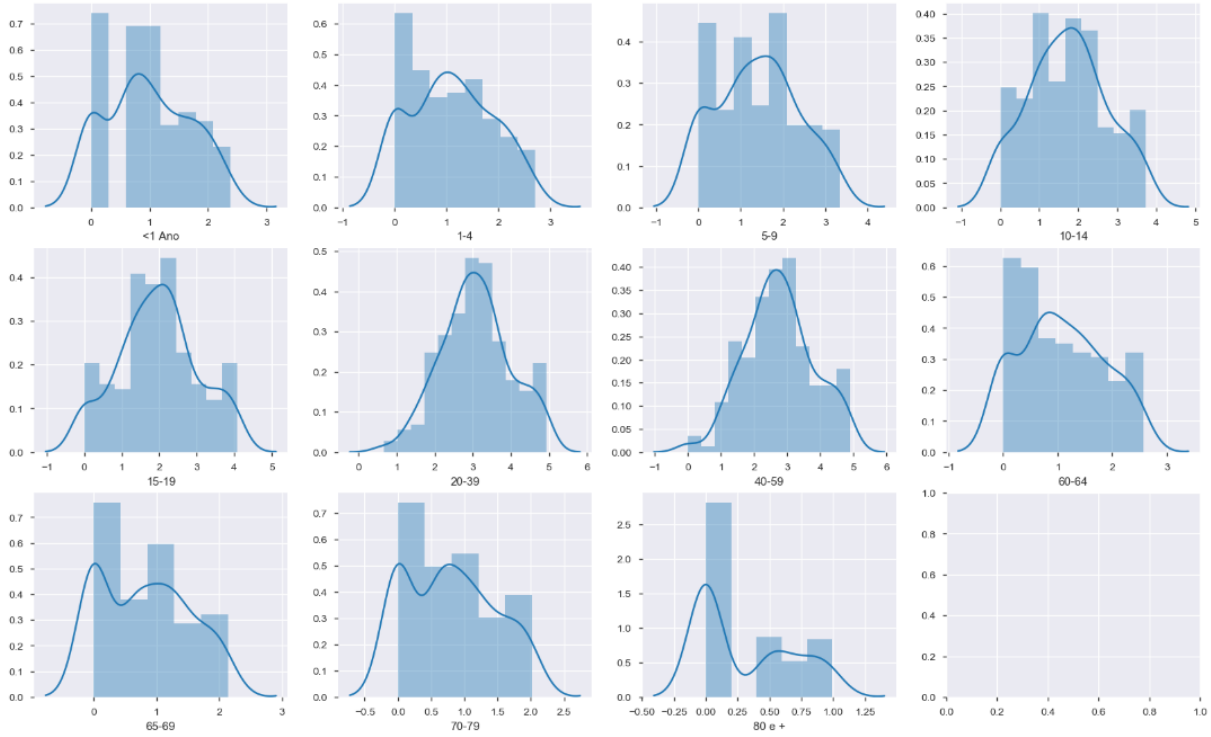
Teste	Normal	Shapiro	Kolmogorov–Smirnov
<1 Ano	2,30E-36	1,81E-22	3,41E-100
1-4	5,03E-44	3,62E-24	1,02E-122
5-9	1,02E-38	3,68E-24	2,43E-145
10-14	5,51E-35	5,02E-24	8,05E-148
15-19	3,58E-32	1,17E-23	2,35E-145
20-39	5,65E-32	1,95E-23	2,59E-145

Teste	Normal	Shapiro	Kolmogorov–Smirnov
40-59	1,97E-34	1,32E-23	2,35E-145
60-64	1,09E-34	2,10E-23	1,67E-130
65-69	9,00E-35	1,70E-23	5,19E-133
70-79	2,88E-32	1,25E-23	2,17E-129
80 e +	1,13E-44	1,05E-24	1,90E-101

TABELA A.2: TESTES DE NORMALIDADE REALIZADOS SOBRE OS CASOS DE DENGUE SEPARADOS POR FAIXA ETÁRIA APÓS A TRANSFORMAÇÃO DOS DADOS.

Teste	Normal	Shapiro	Kolmogorov–Smirnov
<1 Ano	8,04E-10	2,11E-08	2,86E-05
1-4	4,81E-08	1,00E-07	1,13E-03
5-9	2,25E-05	1,12E-06	1,78E-02
10-14	2,19E-02	2,36E-04	4,24E-01
15-19	2,11E-01	4,98E-04	4,67E-01
20-39	7,05E-01	5,77E-02	4,26E-01
40-59	8,48E-01	3,92E-02	8,08E-01
60-64	2,27E-08	7,04E-08	6,78E-04
65-69	3,99E-21	6,21E-11	5,15E-08
70-79	7,73E-15	1,65E-10	1,87E-08
80 e +	6,31E-58	2,90E-17	1,03E-26

FIGURA A.1 – DISTRIBUIÇÃO DOS DADOS SEPARADOS POR FAIXA ETÁRIA, APÓS A TRANSFORMAÇÃO



Como pode ser evidenciado na Fig. A.1, apenas algumas distribuições se aproximaram da normal, sendo elas: ’10-14’ (apenas pelo teste de Komolgorov-Smirnov), ’15-19’ (pelos testes Normal e Komolgorov-Smirnov), ’20-39’ (pelos três testes), ’40-59’ (pelos testes Normal e Komolgorov-Smirnov). As demais colunas não possuem distribuições próximas da normal.

Já para os casos separados por sexo, as distribuições se aproximaram da normal, para o sexo masculino, pelos três testes, e, para o sexo feminino, pelo teste ‘Normal’, conforme mostrado na tabela A.4.

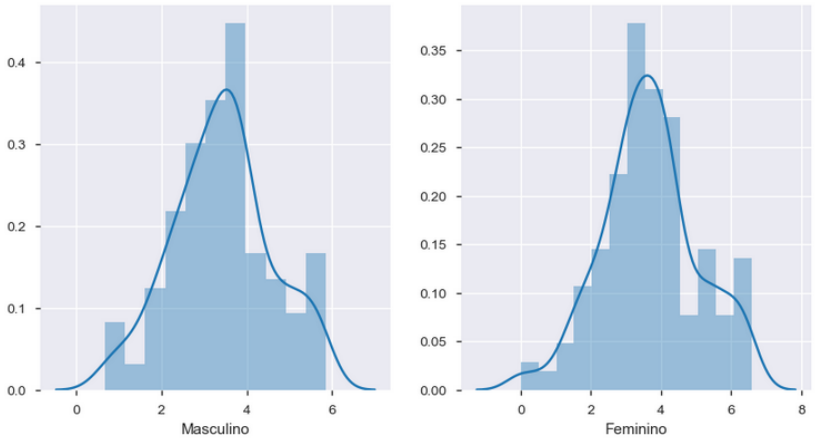
TABELA A.3: TESTES DE NORMALIDADE REALIZADOS SOBRE OS CASOS DE DENGUE SEPARADOS POR SEXO

Teste	Normal	Shapiro	Kolmogorov–Smirnov
Masculino	1,87E-33	1,55E-23	8,05E-148
Feminino	4,39E-33	1,37E-23	8,05E-148

TABELA A.4: TESTES DE NORMALIDADE REALIZADOS SOBRE OS CASOS SEPARADOS POR SEXO APÓS A TRANSFORMAÇÃO

Teste	Normal	Shapiro	Kolmogorov–Smirnov
Masculino	0,78	0,05	0,39
Feminino	0,98	0,03	0,02

FIGURA A.2 – DISTRIBUIÇÃO DOS DADOS SEPARADOS POR SEXO APÓS A TRANSFORMAÇÃO



Por fim, os dados separados por zona de residência não apresentaram distribuição que se aproximasse da normal, mesmo após a realização da transformação. Os resultados dos testes realizados são mostrados a seguir.

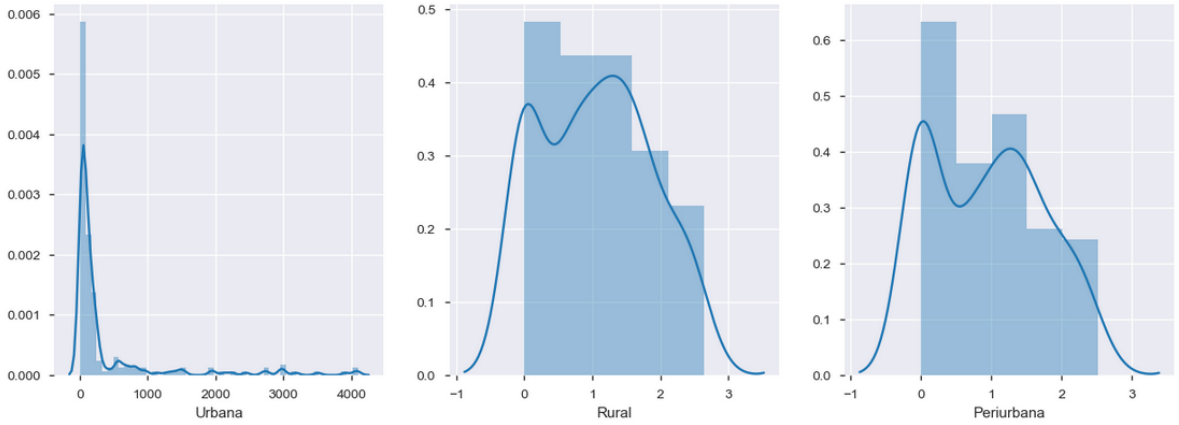
TABELA A.5: TESTES DE NORMALIDADE REALIZADOS SOBRE CASOS DE DENGUE SEPARADOS POR ZONA DE RESIDÊNCIA

Teste	Normal	Shapiro	Kolmogorov–Smirnov
Urbana	4,72E-33	1,49E-23	2,35E-145
Rural	3,97E-35	3,56E-23	6,71E-131
Periurbana	8,70E-45	7,62E-25	1,84E-145

TABELA A.6: TESTES DE NORMALIDADE REALIZADOS SOBRE CASOS DE DENGUE SEPARADOS POR ZONA DE RESIDÊNCIA APÓS A TRANSFORMAÇÃO

Teste	Normal	Shapiro	Kolmogorov–Smirnov
***Urbana**	4,72E-33	1,49E-23	2,35E-145
***Rural**	1,19E-11	6,70E-09	2,87E-04
***Periurbana**	5,51E-21	1,37E-10	1,03E-05

FIGURA A.3 – DISTRIBUIÇÃO DOS DADOS SEPARADOS POR ZONA DE RESIDÊNCIA APÓS A TRANSFORMAÇÃO



APÊNDICE B

B.1 Regressão linear e análise de variância dos dados climáticos relacionados com os casos de dengue em Brasília

Na tentativa de buscar um modelo de regressão linear que descrevesse o comportamento do total de casos de dengue em função das variáveis climáticas, foram implementados quatro algoritmos de regressão pelo método dos mínimos quadrados, um para cada estação do ano. Foram incluídas no modelo as variáveis climáticas "TempBulboSeco", "TempBulboUmido", "UmidadeRelativa", "PressaoAtmEstacao" e "VelocidadeVentoNebulosidade", e a variável dependente (resposta) foi o total de casos de dengue após a transformação de Box-Cox.

B.1.1 Resultados para a estação verão

Como pode ser observado na Fig. B.1, o p-valor da estatística F (Prob F-statistic) do modelo de regressão para os dados referentes ao verão é menor do que o nível de significância adotado. Isso quer dizer que o modelo é significante. Ao analisar o valor de R^2 ajustado (Adj. R-squared), é possível notar que o modelo é suficiente para explicar 15% dos casos de dengue em Brasília durante o verão.

Além disso, de acordo com o resultado da ANOVA (Tabela B.1), apenas a variável 'PressaoAtmEstacao' (D) possui p-valor (PR(>F)) menor que o nível de significância, ou seja, apenas tal variável é significante para explicar os casos de dengue de acordo com esse modelo.

FIGURA B.1 – REGRESSÃO LINEAR MÚLTIPLA COM OS DADOS PERTENCENTES AO PERÍODO DE VERÃO

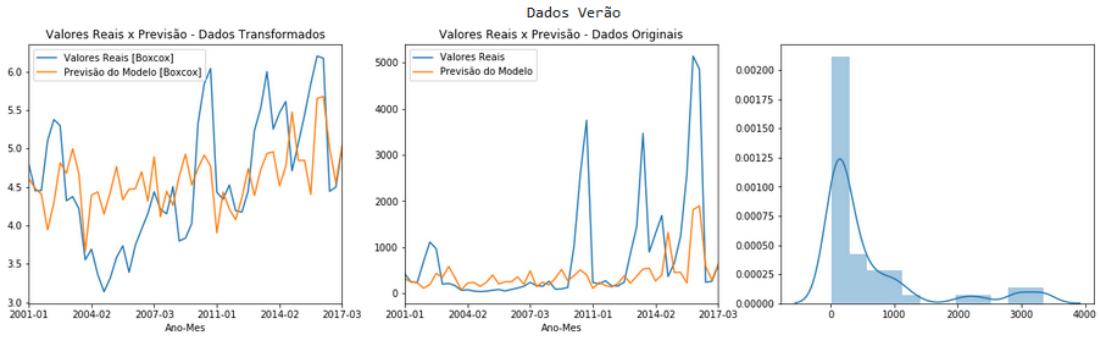
OLS Regression Results

Dep. Variable:	Transformed	**R-squared:**	0.235
Model:	OLS	**Adj. R-squared:**	0.150
Method:	Least Squares	**F-statistic:**	2.769
Date:	Fri, 28 Jun 2019	**Prob (F-statistic):**	0.0290
Time:	14:50:43	**Log-Likelihood:**	-54.574
No. Observations:	51	**AIC:**	121.1
Df Residuals:	45	**BIC:**	132.7
Df Model:	5		
Covariance Type:	nonrobust		

TABELA B.1: RESULTADOS DA ANOVA PARA OS DADOS PERTENCENTES AO PERÍODO DE VERÃO

ANOVA	F	PR(>F)
A	0,205713	0,652328
B	0,041459	0,839572
C	0,095960	0,758162
D	5,067023	0,029317
E	0,120155	0,730482

FIGURA B.2 - RESULTADOS DO MODELO DE REGRESSÃO IMPLEMENTADO PARA OS DADOS PERTENCENTES AO PERÍODO DE VERÃO



B.1.2 Resultados para a Estação Outono

Da mesma forma, para os dados referentes ao outono, o p-valor da estatística F também foi inferior ao nível de significância adotado (Fig. B.3). Por outro lado, o R^2 ajustado é de 0,397, o que significa que o modelo explica apenas 39,7% dos resultados. Além disso, o resultado da ANOVA (Tabela B.2**)** mostra que apenas a variável 'VelocidadeVentoNebulosidade' (E) é significante para explicar os casos de dengue de acordo com esse modelo, já que apresentou p-valor (PR(>F)) menor que 5% (ou 0,05).

FIGURA B.3 – REGRESSÃO LINEAR MÚLTIPLA COM OS DADOS PERTENCENTES AO PERÍODO DE OUTONO

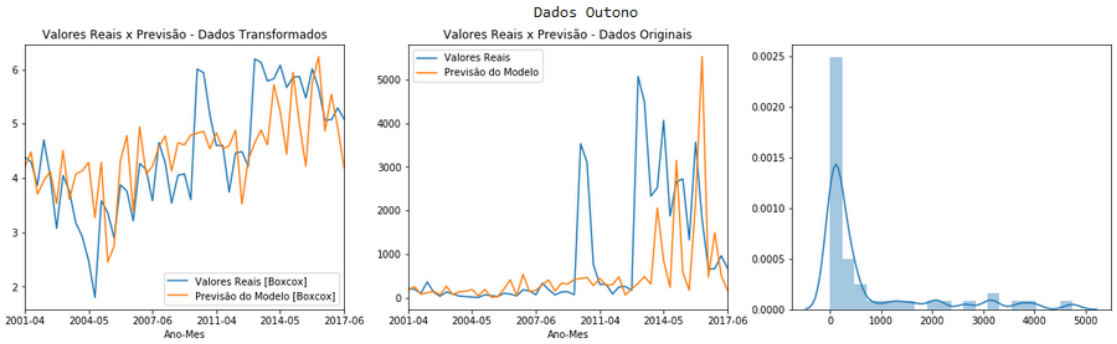
OLS Regression Results

Dep. Variable:	Transformed	**R-squared:**	0.457
Model:	OLS	**Adj. R-squared:**	0.397
Method:	Least Squares	**F-statistic:**	7.586
Date:	Fri, 28 Jun 2019	**Prob (F-statistic):**	3.08e-05
Time:	14:57:01	**Log-Likelihood:**	-59.868
No. Observations:	51	**AIC:**	131.7
Df Residuals:	45	**BIC:**	143.3
Df Model:	5		
Covariance Type:	nonrobust		

TABELA B.2: RESULTADOS DA ANOVA PARA OS DADOS PERTENCENTES AO PERÍODO DE OUTONO

ANOVA	F	PR(>F)
A	0,076855	0,782876
B	0,618846	0,435600
C	0,484535	0,489955
D	3,288958	0,076419
E	23,173304	0,000017

FIGURA B.4 - RESULTADOS DO MODELO DE REGRESSÃO IMPLEMENTADO PARA OS DADOS PERTENCENTES AO PERÍODO DE OUTONO



B.1.3 Resultados para a estação inverno

Por sua vez, analisando os dados do inverno, o p-valor da estatística F foi inferior ao nível de significância adotado (Fig. B.5), significando que o modelo de modo geral é significante para explicar o total de casos de dengue. Por outro lado, o R^2 ajustado foi de 0,133, o que significa que o modelo explica 13,3% dos resultados. Por fim, o resultado da ANOVA (Tabela B.3**)** mostra que apenas a variável "VelocidadeVentoNebulosidade" (E) é significante para explicar os casos de dengue de acordo com esse modelo, já que possui p-valor (PR(>F)) menor que 5% (ou 0,05).

FIGURA B.5 – REGRESSÃO LINEAR MÚLTIPLA COM OS DADOS PERTENCENTES AO PERÍODO DE INVERNO

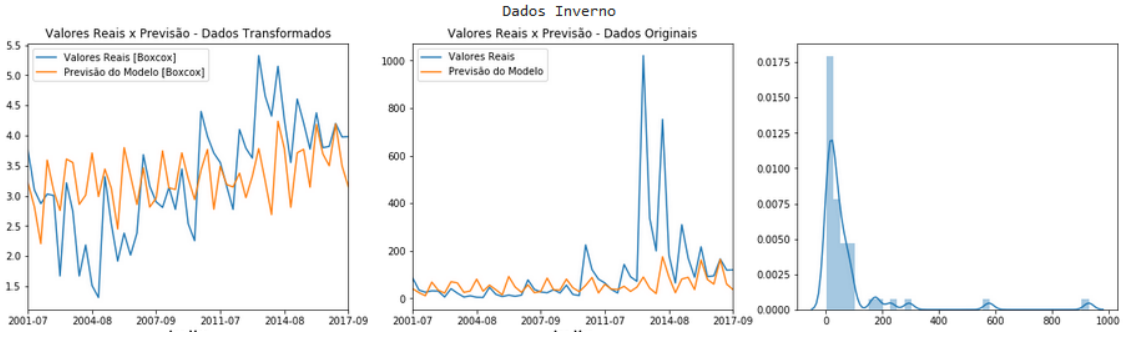
OLS Regression Results

Dep. Variable:	Transformed	**R-squared:**	0.220
Model:	OLS	**Adj. R-squared:**	0.133
Method:	Least Squares	**F-statistic:**	2.540
Date:	Fri, 28 Jun 2019	**Prob (F-statistic):**	0.0416
Time:	15:03:35	**Log-Likelihood:**	-62.147
No. Observations:	51	**AIC:**	136.3
Df Residuals:	45	**BIC:**	147.9
Df Model:	5		
Covariance Type:	nonrobust		

TABELA B.3: RESULTADOS DA ANOVA PARA OS DADOS PERTENCENTES AO PERÍODO DE INVERNO

ANOVA	F	PR(>F)
A	0,130115	0,720000
B	0,134412	0,715619
C	0,007638	0,930743
D	3,367467	0,073111
E	6,561058	0,013845

FIGURA B.6 - RESULTADOS DO MODELO DE REGRESSÃO IMPLEMENTADO PARA OS DADOS PERTENCENTES AO PERÍODO DE INVERNO



B.1.4 Resultados para a estação primavera

Por último, o modelo construído para a primavera apresentou um p-valor da estatística F inferior ao nível de significância adotado (Fig. B.7**)**. Por sua vez, o R^2 ajustado é de 0.168, o que significa que o modelo explica 16,8% dos casos de dengue durante a primavera. Além disso, o resultado da ANOVA (Tabela B.4**)** mostra que, assim como no inverno, apenas a variável "VelocidadeVentoNebulosidade" (E) é significante para explicar os casos de dengue de acordo com esse modelo, já que o p-valor (PR(>F)) resultante é inferior a 5% (ou 0,05).

FIGURA B.7 – REGRESSÃO LINEAR MÚLTIPLA COM OS DADOS PERTENCENTES AO PERÍODO DE PRIMAVERA

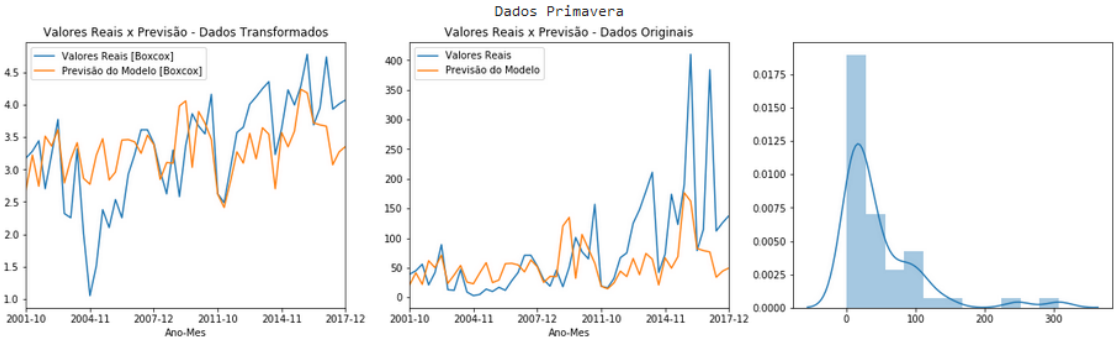
OLS Regression Results

Dep. Variable:	Transformed	**R-squared:**	0.252
Model:	OLS	**Adj. R-squared:**	0.168
Method:	Least Squares	**F-statistic:**	3.026
Date:	Fri, 28 Jun 2019	**Prob (F-statistic):**	0.0194
Time:	15:08:08	**Log-Likelihood:**	-53.579
No. Observations:	51	**AIC:**	119.2
Df Residuals:	45	**BIC:**	130.7
Df Model:	5		
Covariance Type:	nonrobust		

TABELA B.4: RESULTADOS DA ANOVA PARA OS DADOS PERTENCENTES AO PERÍODO DE PRIMAVERA

ANOVA	F	PR(>F)
A	0,026722	0,870882
B	0,404021	0,528239
C	0,156318	0,694437
D	0,048623	0,826475
E	5,811812	0,020065

FIGURA B.8 - RESULTADOS DO MODELO DE REGRESSÃO IMPLEMENTADO PARA OS DADOS PERTENCENTES AO PERÍODO DE PRIMAVERA



APÊNDICE C

A seguir são apresentados os coeficientes de cada variável envolvendo cada modelo de regressão linear múltipla implementado. Vale ressaltar que as variáveis significativas para explicar o modelo, conforme apresentado no desenvolvimento, estão destacadas em cinza e negrito.

TABELA C.1: COEFICIENTES DA REGRESSÃO MÚLTIPLA PARA A ESTAÇÃO VERÃO

Variáveis	Coeficientes
Intercept	5656,52
A	-5911,42
B	-7232,95
A:B	7132,04
C	-5346,92
A:C	5477,58
B:C	6642,13
A:B:C	-5778,08
D	-19315,59
A:D	22174,53
B:D	24905,22
A:B:D	**-26951,88**
C:D	19305,95
A:C:D	-23350,84
B:C:D	-24080,44
A:B:C:D	25373,44
E	-14394,05
A:E	14331,35
B:E	18478,83
A:B:E	-17174,83
C:E	13508,28
A:C:E	-13129,48
B:C:E	**-16727,79**
A:B:C:E	13269,58
D:E	51439,95
A:D:E	**-58449,97**
B:D:E	-66268,16
A:B:D:E	70441,21
C:D:E	**-51664,93**
A:C:D:E	62506,83
B:C:D:E	**64002,42**
A:B:C:D:E	-66361,14
B:C:D:E	64002,42
A:B:C:D:E	-66361,14

TABELA C.2: COEFICIENTES DA REGRESSÃO MÚLTIPLA PARA A ESTAÇÃO OUTONO

Variáveis	Coeficientes
Intercept	15,07
A	187,89
B	649,95
A:B	**-954,12**
C	-317,10
A:C	-472,34
B:C	-270,74
A:B:C	1021,39
D	**62,18**
A:D	-643,77
B:D	-1376,49
A:B:D	2119,37
C:D	356,61
A:C:D	1836,88
B:C:D	844,74
A:B:C:D	-2942,89
E	**186,53**
A:E	-784,16
B:E	-1903,70
A:B:E	**3365,35**
C:E	415,76
A:C:E	1017,76
B:C:E	1365,14
A:B:C:E	-3627,77
D:E	-656,78
A:D:E	2492,66
B:D:E	4025,45
A:B:D:E	**-7387,80**
C:D:E	25,74
A:C:D:E	-4818,38
B:C:D:E	-3623,90
A:B:C:D:E	9968,85
B:C:D:E	64002,42
A:B:C:D:E	-66361,14

TABELA C.3: COEFICIENTES DA REGRESSÃO MÚLTIPLA PARA A ESTAÇÃO INVERNO

Variáveis	Coeficientes
Intercept	-4,75
A	-194,70
B	710,53
A:B	-417,83
C	337,71
A:C	-947,32
B:C	-2229,68
A:B:C	3538,54
D	-35,72
A:D	400,09
B:D	-1070,93
A:B:D	529,01
C:D	-393,98
A:C:D	1277,82
B:C:D	3367,91
A:B:C:D	-5314,57
E	-67,56
A:E	484,35
B:E	-746,57
A:B:E	-105,97
C:E	-324,22
A:C:E	1385,19
B:C:E	**2226,82**
A:B:C:E	-3560,55
D:E	174,94
A:D:E	-919,75
B:D:E	1170,41
A:B:D:E	331,58
C:D:E	318,34
A:C:D:E	-1720,46
B:C:D:E	-3566,82
A:B:C:D:E	5202,54
B:C:D:E	64002,42
A:B:C:D:E	-66361,14

TABELA C.4: COEFICIENTES DA REGRESSÃO MÚLTIPLA PARA A ESTAÇÃO PRIMAVERA

Variáveis	Coeficientes
Intercept	-1082,38
A	1401,28
B	444,12
A:B	-1093,59
C	1383,63
A:C	-746,18
B:C	-966,56
A:B:C	1050,67
D	3000,75
A:D	-3619,49
B:D	-544,22
A:B:D	2103,89
C:D	-3297,36
A:C:D	356,46
B:C:D	1449,13
A:B:C:D	-656,01
E	**4186,48**
A:E	-5727,66
B:E	-2659,40
A:B:E	5333,29
C:E	-5366,86
A:C:E	**4517,28**
B:C:E	4530,12
A:B:C:E	-5954,88
D:E	-10456,59
A:D:E	**14140,54**
B:D:E	**4135,33**
A:B:D:E	-10974,62
C:D:E	**12552,06**
A:C:D:E	-7966,37
B:C:D:E	-8129,64
A:B:C:D:E	10322,41
B:C:D:E	64002,42
A:B:C:D:E	-66361,14