UNIVERSIDADE FEDERAL DE UBERLÂNDIA
INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS DO PONTAL
CURSO DE MATEMÁTICA

ANA LÍVIA RODRIGUES NASCIMENTO

REGRESSÃO LINEAR MÚLTIPLA NA IDENTIFICAÇÃO DE FATORES RELACIONADOS À DEMANDA POR SERVIÇOS DE COMPARTILHAMENTO DE BICICLETAS

ITUIUTABA - MG
2023

ANA LÍVIA RODRIGUES NASCIMENTO

REGRESSÃO LINEAR MÚLTIPLA NA IDENTIFICAÇÃO DE FATORES RELACIONADOS À DEMANDA POR SERVIÇOS DE COMPARTILHAMENTO DE BICICLETAS

Trabalho de Conclusão de Curso apresentado ao Instituto de Ciências Exatas e Naturais do Pontal da Universidade Federal de Uberlândia como requisito parcial para obtenção do título de bacharel em Matemática.

Orientadora: Profª. Dra. Gabriella de Freitas Alves

ITUIUTABA - MG

2023

ANA LÍVIA RODRIGUES NASCIMENTO

REGRESSÃO LINEAR MÚLTIPLA NA IDENTIFICAÇÃO DE FATORES RELACIONADOS À DEMANDA POR SERVIÇOS DE COMPARTILHAMENTO DE BICICLETAS

Trabalho de Conclusão de Curso apresentado ao Instituto de Ciências Exatas e Naturais do Pontal da Universidade Federal de Uberlândia como requisito parcial para obtenção do título de bacharel em Matemática.

Ituiutaba - MG, 01/02/2023

Banca Examinadora:

_____

Profª. Dra. Gabriella de Freitas Alves

_____

Profª. Dra. Franciella Marques da Costa

_____

Profª. Dra. Kátia Gomes Facure Giaretta

Dedico este trabalho a minha mãe Eliana Pereira do Nascimento e ao meu pai Omar Rodrigues do Nascimento, pois foi graças aos seus esforços que cheguei até aqui. Não existem palavras para expressar toda a minha gratidão por tudo que fizeram por mim durante toda a minha vida.

AGRADECIMENTOS

Agradeço a minha orientadora Gabriella de Freitas Alves, por sua disposição, dedicação e paciência durante a realização do trabalho.

Agradeço as professoras Franciella Marques da Costa e Kátia Gomes Facure Giaretta, pelos comentários, sugestões e avaliação.

Agradeço a minha amiga e colega de curso Amanda Vitória de Jesus Mendes, pela companhia, ajuda e motivação.

Agradeço aos professores do curso de Matemática, por todos os ensinamentos.

Agradeço a todas as pessoas que contribuíram de forma direta ou indireta para a realização deste trabalho.

“Pega a bicicleta e vai!”

(Janaísa Cardoso)

RESUMO

Os sistemas de compartilhamento de bicicletas proporcionam diversos benefícios econômicos, climáticos e na saúde e por isso tem se tornado cada vez mais comuns. O objetivo deste trabalho foi aplicar métodos de análise de regressão linear múltipla, para identificação dos principais fatores de influência na demanda por este tipo de serviço. Para atingir o objetivo considerou-se como variável dependente a contagem diária de bicicletas alugadas no provedor de compartilhamento BikeIndia, dos Estados Unidos da América (EUA), nos anos de 2018 e 2019. As variáveis independentes para o estudo foram estação do ano, ano, mês, feriado, dia da semana, dia útil, situação climática, temperatura, sensação térmica, umidade e velocidade do vento. O modelo ajustado mostrou uma relação linear significativa entre a contagem de bicicletas alugadas e as variáveis ano, feriado, temperatura, umidade e velocidade do vento. Concluiu-se que no ano de 2019 houve um aumento significativo nos aluguéis de bicicletas em comparação ao ano de 2018 e nos feriados verificou-se uma redução nesses aluguéis. Além disso, verificou-se uma tendência de aumento nos aluguéis de bicicletas com o aumento na temperatura, enquanto o aumento na umidade e na velocidade do vento tende a reduzir a quantidade de aluguéis.

Palavras-chave: sistema de compartilhamento de bicicletas; desenvolvimento sustentável; multicolinearidade; variáveis binárias; método stepwise.

ABSTRACT

Bike sharing systems provide several economic, climate and health benefits and therefore have become increasingly common. The objective of this study was to apply multiple linear regression analysis methods to identify the main factors influencing the demand for this type of service. In order to achieve the objective, the daily count of bicycles rented at the BikeIndia sharing provider, from the United States of America (USA), in the years 2018 and 2019 was considered as the dependent variable. The independent variables for the study were the season of the year, year, holiday period, working day, weekend, climate situation, temperature, wind chill, humidity and wind speed. The adjusted model showed a significant linear relationship between the rented bike count and the year, holiday, temperature, humidity, and wind speed variables. It was concluded that in 2019 there was a significant increase in bicycle rentals compared to 2018 and on holidays there was a reduction in rentals. In addition, there was a trend towards an increase in bicycle rentals with an increase in temperature, while the increase in humidity and wind speed tends to reduce the number of rentals.

Keywords: bike sharing system; sustainable development; multicollinearity; binary variables; *stepwise* method.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A implantação e utilização de sistemas de compartilhamento de bicicletas tem sido cada vez mais comuns na última década. Atualmente estes sistemas podem ser encontrados em várias cidades do Brasil e do mundo. De acordo com o Instituto de Políticas de Transporte e Desenvolvimento (ITDP), o compartilhamento pode ocorrer por meio de um sistema gratuito ou pago e sua essência consiste na retirada de uma bicicleta em um local, geralmente em uma estação, e na devolução da mesma em outro local (ou outra estação) (ITDP, 2014).

Nos sistemas gratuitos um conjunto de bicicletas é disponibilizado sem custos e as estações se localizam próximas a equipamentos públicos, possibilitando a identificação dos usuários por meio de uma equipe responsável (FRADE; RIBEIRO, 2014). Já nos sistemas pagos, os usuários pagam em função do tempo de uso e geralmente se utiliza de tecnologia como cartões magnéticos, que possibilitam a identificação automática do usuário, além do desbloqueio das bicicletas nas estações (FRADE; RIBEIRO, 2014).

Segundo o ITDP (2014), o primeiro sistema de compartilhamento foi criado em 1965 por Luud Schimmelpennink, na época vereador de Amsterdam, com o objetivo de reduzir o fluxo de automóveis no centro da cidade. Embora a proposta de Luud tenha sido recusada pela assembleia municipal, algumas pessoas que apoiavam sua ideia distribuíram bicicletas gratuitamente para utilização em toda cidade, porém, logo depois elas foram apreendidas pela polícia local sob alegação de estarem incitando o roubo, por ficarem destrancadas (ITDP, 2014).

Embora um sistema de menor escala tenha sido criado em três cidades de Oregon, nos Estados Unidos da América, foi em La Rochelle na França no ano de 1993, que ocorreu uma nova tentativa de implementar um sistema de bicicletas compartilhadas gratuito e mais regulamentado, e nesse mesmo ano, um sistema parecido foi implementado em Cambridge, Inglaterra (ITDP, 2014).

Apesar das tentativas citadas anteriormente, a grande adesão a esses sistemas foi percebida apenas na última década, e conforme se evidencia seus benefícios econômicos, climáticos e na saúde, mais sistemas vão sendo criados (ITDP, 2018). O ciclismo possui baixo custo em relação a outros meios de locomoção, não contribui com as mudanças climáticas, poluição sonora e atmosférica e, se praticado regularmente ajuda na prevenção

de doenças, sendo assim um meio de transporte que auxilia no desenvolvimento sustentável (CARVALHO; FREITAS, 2012).

Embora o sistema de compartilhamento de bicicletas tenha se expandido pelo mundo, muitas empresas sucumbiram por não conseguirem atender as necessidades dos usuários e das cidades onde se estabeleciam (FELIPE, 2018). Além disso, a pandemia de Covid-19 afetou diversos setores da economia, inclusive estes sistemas. A mudança de hábitos e o isolamento social afetaram a demanda de muitos produtos, gerando dificuldades no faturamento das empresas. Para contornar essa situação foi necessário pensar em novas estratégias de venda como o uso da internet para divulgação e/ou comercialização, porém nem todas as empresas conseguiram comercializar remotamente seus produtos de forma satisfatória, sendo essas as mais afetadas no pós-pandemia (REZENDE; MARCELINO; MIYAJI, 2020).

De acordo com Heydari, Konstantinoudis, Behsoodi (2021), em alguns períodos da pandemia os sistemas de compartilhamento de bicicletas foram influenciados de maneira negativa e em outros de maneira positiva, sendo essas influências tanto devido as alterações na demanda pelo serviço quanto devido as mudanças que ocorreram com os transportes públicos. Os autores mostraram que no sistema de compartilhamento de bicicletas Santander Cycles, localizado em Londres, durante a pandemia de Covid-19 ocorreu uma redução nos aluguéis de bicicletas nos meses de março e abril de 2020, porém no mês de maio de 2020 recuperou-se a demanda e a mesma conservou-se da maneira como era esperada caso não ocorresse a pandemia. Foi observado também que em maio e junho de 2020 houve um pequeno aumento no número de aluguéis e nos meses de abril, maio e junho de 2020 ocorreu um aumento considerável no tempo médio de contratação.

A empresa Citi Bike, localizada em Nova York, teve um aumento de 67% na demanda pelos seus serviços de compartilhamento de bicicletas em março de 2020 devido as pessoas estarem evitando o uso do metrô (IBOLD et al., 2020). Ao contrário do que ocorreu com a empresa Citi Bike, Lopes (2021) concluiu por meio de históricos de viagens dos anos 2019 e 2020 da organização BIXI Montreal, responsável pelo gerenciamento do sistema de compartilhamento de bicicletas da cidade de Montreal, que houve uma redução significativa na demanda por viagens de bicicleta devido a pandemia de Covid-19.

Segundo o ITDP (2014), cada cidade ajusta o seu compartilhamento de bicicletas considerando a densidade, topografia, clima, infraestrutura e cultura local, mostrando assim que cada localidade tem seu modelo de compartilhamento. Assim, torna-se importante identificar para cada local, quais fatores estão mais relacionados a demanda por esse tipo serviço. Essa identificação poderá ser extremamente útil para alavancar os negócios e auxiliar os empresários do ramo de compartilhamento. Neste contexto, uma técnica estatística que poderia ser extremamente útil é a análise de regressão linear. Esta análise possibilita investigar a relação entre variáveis e tem sido aplicada em diversas áreas (MONTGOMERY; PECK; VINING, 2021).

De acordo com Gujarati e Porter (2011), a regressão linear pode ser vista como o estudo da dependência de uma variável denominada variável dependente a uma ou mais variáveis independentes, sendo neste último caso denominada análise de regressão linear múltipla. Dessa forma, a regressão linear múltipla pode ser utilizada para investigar e modelar a relação entre a demanda por aluguéis de bicicletas (variável dependente) e diversos fatores (variáveis independentes) que possam estar relacionados a essa demanda.

Alguns autores já utilizaram a análise de regressão linear para identificar fatores relacionados à sistemas de compartilhamento de bicicletas. De Arruda et al. (2016), analisou a relação entre o consumo colaborativo[1] de bicicletas compartilhadas e os valores pessoais dos consumidores e conseguiu identificar variáveis associadas à valores pessoais que influenciam o consumo colaborativo de bicicletas.

Bittencourt (2020) investigou por meio da análise de regressão a relação entre o número de viagens/estação e diversas variáveis independentes. Nesta análise o autor identificou que a utilização de bicicletas no período da manhã relacionou-se fortemente com o trabalho ou atividades de compras; no período da tarde as bicicletas foram utilizadas como forma complementar de transporte; as estações mais eficientes são as localizadas em áreas com diversos usos de solo (uso comercial e residencial).

Considerando as constantes mudanças que ocorrem nos sistemas de compartilhamento de bicicletas ao longo dos anos e as diferentes peculiaridades de cada

---

[1] "O consumo colaborativo é uma tendência global e em crescimento alavancado por motivações individuais e questões ambientais, sociais e econômicas presentes no cotidiano das pessoas, como uma forma de fomentar um consumo mais consciente e sustentável. Dentre as práticas de consumo colaborativo, o compartilhamento de bicicletas é o mais praticado mundialmente." (DE ARRUDA et al., 2016)

local para o sucesso destes sistemas, este trabalho teve como objetivo aplicar métodos de análise de regressão linear múltipla a um conjunto de dados proveniente de uma empresa privada de compartilhamento de bicicletas, para identificação dos principais fatores de influência na demanda pelo serviço.

2 REFERENCIAL TEÓRICO

2.1 Análise de Regressão Linear

A análise de regressão linear é uma técnica estatística amplamente utilizada em diversas áreas da ciência, possibilitando a investigação e modelagem da relação entre variáveis (MONTGOMERY; PECK; VINING, 2021). Ao longo da história da estatística muitos nomes contribuíram para o desenvolvimento da regressão linear atual. Dentre eles, os mais notáveis foram: Francis Galton (1822-1911) e Karl Pearson (1857-1936) (CASTRO et al., 2012).

Em 1885, Galton usou pela primeira vez o termo regressão no estudo comparativo das alturas entre pais e filhos, referindo-se a sua observação de que ocorria a regressão das alturas observadas à altura média da população (MEMÓRIA, 2004).

O nome de Karl Pearson está muito relacionado ao coeficiente mais utilizado para estimar a possível correlação existente entre duas variáveis, denominado coeficiente de correlação de Pearson. Porém, embora Pearson tenha desenvolvido a matemática rigorosa por traz deste coeficiente, foi Galton que possibilitou o desenvolvimento dos conceitos de correlação e regressão modernos (CASTRO et al., 2012). O motivo do desenvolvimento do coeficiente de correlação se deu pela curiosidade de Galton em entender o quão fortemente as características de uma geração seriam herdadas pela geração seguinte (STANTON, 2001). Os estudos posteriores de Pearson levaram ao desenvolvimento da regressão e correlação múltiplas e basearam diversos artigos escritos por ele e seus colaboradores (MEMÓRIA, 2004).

Atualmente a análise de regressão linear é utilizada em áreas como estatística, matemática, biologia, física, engenharia entre outras, principalmente as que tem o objetivo de relacionar variáveis (SAMPAIO, 2015).

Ao investigar a relação entre duas variáveis, a análise é denominada de regressão linear simples. Neste caso, o modelo de regressão é dado pela seguinte equação:

$$y = \beta_0 + \beta_1 x + \mathcal{E} \tag{1}$$

onde y é a variável dependente, x é a variável independente, β_0 é o intercepto, β_1 é a inclinação da reta e ε é o erro (SCHMIDT; FINAN, 2018).

O termo “linear” na análise de regressão diz respeito a linearidade dos parâmetros do modelo, ou seja, os parâmetros são sempre elevados à primeira potência (GUJARATI; PORTER, 2011).

Em muitas aplicações da análise de regressão a variável dependente está relacionada a duas ou mais variáveis independentes e nesses casos a análise é denominada de regressão linear múltipla (MONTGOMERY; RUNGER, 2013).

2.2 Análise de Regressão Linear Múltipla

Na análise de regressão linear múltipla, a relação entre a variável dependente e k variáveis independentes pode ser modelada pela seguinte equação (MONTGOMERY; RUNGER, 2013):

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \mathcal{E}_i \tag{2}$$

onde y_i é a variável dependente, x_{ki} são as variáveis independentes, $\beta_0, \beta_1, \ldots, \beta_k$ são os parâmetros e ε_i é o erro, com $\mathrm{i} = 1, \ldots, \mathrm{n} \mid \mathrm{n} \in \mathbb{N}$.

Na forma matricial, o modelo será dado por (MONTGOMERY; RUNGER, 2013):

$$Y = \beta \cdot X + \varepsilon \tag{3}$$

onde

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}; \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix};$$

$$X = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix} \qquad e \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_i \end{bmatrix}$$

Ao estabelecer o modelo de regressão linear múltipla, pressupomos que: (HOFFMAN, 2016)

i. A variável dependente é função linear das variáveis independentes;
ii. Os valores das variáveis independentes são fixos;
iii. $E(\varepsilon_i) = 0$, ou seja, $E(\varepsilon) = \vec{0}$, onde $\vec{0}$ representa um vetor de zeros;
iv. Os erros são homocedásticos, isto é, $E(\varepsilon_i^2) = \sigma^2$;
v. Os erros são não correlacionados entre si;
vi. Os erros têm distribuição normal.

2.2.1 Equação Estimada (Método dos Mínimos Quadrados)

O método mais utilizado para a estimação dos parâmetros é o Método dos Mínimos Quadrados (MMQ). Segundo Hoffmann (2016), tal método consiste em determinar um valor para os parâmetros a fim de minimizar a soma dos quadrados dos desvios entre os valores amostrais e estimados. Os parâmetros estimados pelo Método dos Mínimos Quadrados podem ser obtidos por meio da equação (4):

$$\hat{\beta} = (X'X)^{-1}(X'Y) \tag{4}$$

onde

$$X'X = \begin{bmatrix} n & \sum x_{1i} & \sum x_{2i} & \cdots & \sum x_{ki} \\ \sum x_{1i} & \sum x_{1i}^2 & \sum x_{1i}x_{2i} & \cdots & \sum x_{1i}x_{ki} \\ \sum x_{2i} & \sum x_{1i}x_{2i} & \sum x_{2i}^2 & \cdots & \sum x_{2i}x_{ki} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum x_{ki} & \sum x_{1i}x_{ki} & \sum x_{2i}x_{ki} & \cdots & \sum x_{ki}^2 \end{bmatrix};$$

$$X'Y = \begin{bmatrix} \sum y_i \\ \sum x_{1i}y_i \\ \vdots \\ \sum x_{ki}y_i \end{bmatrix}$$

e $(X'X)^{-1}$ é a inversa da matriz $(X'X)$.

Para chegar na equação (4), de acordo com Montgomery, Peck, Vining (2021), deve-se encontrar o vetor dos estimadores de mínimos quadrados, $\hat{\beta}$, de forma a minimizar a soma dos quadrados dos erros $\left(S(\beta)\right)$:

$$\begin{aligned} S(\beta) &= \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon'\varepsilon = (Y - X\beta)'(Y - X\beta) \\ &= Y'Y - \beta'X'Y - Y'X\beta + \beta'X'X\beta \\ &= Y'Y - 2\beta'X'Y + \beta'X'X\beta \end{aligned}$$

Tem-se que $\beta'X'Y$ é um escalar e sua transposta $Y'X\beta$ também é um escalar. Os estimadores de mínimos quadrados devem satisfazer:

$$\left.\frac{\partial S}{\partial \beta}\right|_{\hat{\beta}} = -2X'Y + 2X'X\hat{\beta} = 0$$

Simplificando, obtém-se a equação (5):

$$X'Y = X'X\hat{\beta} \tag{5}$$

que são as equações normais dos mínimos quadrados.

Para solucionar a equação (5), multiplica-se ambos os lados por $(X'X)^{-1}$ e assim chega-se na equação (4):

$$\hat{\beta} = (X'X)^{-1}X'Y$$

que só é válida se $(X'X)^{-1}$ existir.

Após a estimação dos parâmetros, o modelo de regressão linear múltipla estimado, é dado pela seguinte equação:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_k x_{ki} \tag{6}$$

onde $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$, são os parâmetros estimados.

De acordo com Gujarati e Porter (2011) os estimadores de mínimos quadrados possuem algumas propriedades ideais. Um estimador, $\hat{\beta}$, é considerado o melhor estimador linear não tendencioso de β se cumprir as condições:

i. É linear, ou seja, é função linear de uma variável aleatória.
ii. É não tendencioso, ou seja, seu valor esperado $E(\hat{\beta})$ é igual ao verdadeiro valor β.
iii. Possui variância mínima na classe de todos os estimadores lineares não tendenciosos, sendo o estimador com menor variância e não tendencioso conhecido como estimador eficiente.

2.2.2 Análise de Variância (Teste F)

Na análise de regressão linear múltipla o teste F, também denominado de análise de variância, é visto como um teste de significância global e é utilizado para verificar se existe uma relação global significativa entre a variável dependente e o conjunto de todas as variáveis independentes.

As hipóteses do teste são dadas por (MONTGOMERY; PECK; VINING, 2021):

$$\begin{cases} H_0: \ \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ H_1: \beta_j \neq 0 \text{ , para pelo menos um } j = 1, \cdots, k \text{ ; com k } \in \mathbb{N} \end{cases}$$

A rejeição da hipótese H_0 implica que pelo menos uma das variáveis independentes $X_1, X_2, \ldots, X_k$ contribui significativamente para o modelo.

O teste geralmente é resumido em uma tabela de análise de variância, conforme tabela 1 (MONTGOMERY; RUNGER, 2013):

Tabela 1 – Análise de variância para regressão linear múltipla

FV	***GL***	***SQ***	***QM***	F_c
Regressão	$k = p - 1$	$SQRegressão$	$QMRegressão$	$\frac{QMRegressão}{QMResíduo}$
Resíduo	$n - p$	$SQResíduo$	$QMResíduo$	-
Total	$n - 1$	$SQTotal$	-	-

FV: Fontes de Variação; *GL*: Graus de Liberdade; *SQ*: Soma de Quadrados; *QM*: Quadrado Médio; F_c: Valor de F calculado. Fonte: Montgomery e Runger (2013)

na qual,

$$SQTotal = (Y'Y) - \frac{\left(\sum_{n}^{i=1} y_i\right)^2}{n};$$

$$SQRegressão = \hat{\beta}'(X'Y) - \frac{\left(\sum_{n}^{i=1} y_i\right)^2}{n};$$

$$SQResíduo = SQTotal - SQRegressão;$$

$$QMResíduo = \frac{SQResíduos}{n - p};$$

$$QMRegressão = \frac{SQRegressão}{p - 1}$$

com n sendo o tamanho da amostra e p a quantidade de parâmetros.

A hipótese H_0 é rejeitada se $F_c \geq F_{tab}$, onde F_{tab} se baseia em uma distribuição F com $v_1 = p - 1$ e $v_2 = n - p$ graus de liberdade, respectivamente (ANDERSON; SWEENEY; WILLIANS, 2007).

2.2.3 Teste t

O teste t é utilizado para analisar se cada uma das variáveis independentes individuais é significativa, sendo que para cada variável do modelo realiza-se um teste t individual (ANDERSON; SWEENEY; WILLIANS, 2007).

Suas hipóteses são dadas por:

$$\begin{cases} H_0: \beta_j = 0 \\ H_1: \beta_j \neq 0 \end{cases}$$

Caso H_0 não seja rejeitada, a variável independente x_j pode ser excluída do modelo, e para a realização do teste utiliza-se a seguinte estatística: (MONTGOMERY; PECK; VINING, 2021)

$$t_c = \frac{\hat{\beta}_j - \beta_j}{S(\beta_j)} \tag{7}$$

onde, $S(\beta_j) = \sqrt{QMResíduo \times (C_{jj})}$ e C_{jj} é o elemento da diagonal da matriz inversa $(\mathrm{X}'\mathrm{X})^{-1}$ correspondente ao parâmetro $\hat{\beta}_j$.

Rejeita-se H_0 se $t_c \leq -t_{\left(\frac{a}{2}, n-p\right)}$ ou $t_c \geq t_{\left(\frac{a}{2}, n-p\right)}$, onde $t_{\left(\frac{a}{2}, n-p\right)}$ baseia-se em uma distribuição $t-$Student com $n-p$ graus de liberdade (ANDERSON; SWEENEY; WILLIANS, 2007).

2.2.4 Coeficiente de Determinação Múltiplo e Coeficiente de Determinação Múltiplo Ajustado

O coeficiente de determinação múltiplo, denominado R^2, é uma estatística global utilizada para avaliar o ajuste do modelo, e é obtido pela seguinte equação: (MONTGOMERY; RUNGER, 2013)

$$R^2 = \frac{SQRegressão}{SQTotal} \tag{8}$$

Uma desvantagem da estatística R^2 é o fato de seu valor sempre aumentar quando se adiciona uma variável independente no modelo e por esse motivo é preferível utilizar o coeficiente de determinação ajustado R_a^2, dado por: (ANDERSON; SWEENEY; WILLIANS, 2007)

$$R_a^2 = 1 - \left[(1 - R^2) \cdot \frac{(n-1)}{n-k-1}\right] \tag{9}$$

onde, n é o tamanho da amostra e k o número de variáveis independentes.

O coeficiente de determinação ajustado, R_a^2, aponta a eficiência de ajuste da equação de regressão linear múltipla estimada assim como R^2, porém evitando-se uma superestimação do impacto da adição de uma variável independente. Ao multiplicar R^2 por 100, pode-se interpretar o valor resultante como a porcentagem da variação da

variável dependente y que pode ser explicada pela equação de regressão linear múltipla estimada (ANDERSON; SWEENEY; WILLIANS, 2007).

2.2.5 Multicolinearidade

Ao trabalhar com análise de regressão múltipla, é comum se deparar com variáveis independentes correlacionadas, sendo essa correlação denominada multicolinearidade. Nos casos em que as variáveis independentes estão altamente correlacionadas, a multicolinearidade pode afetar a análise, como alterar o sinal das estimativas dos parâmetros obtidos por mínimos quadrados e impossibilitar a determinação do efeito individual de qualquer variável independente sobre a variável dependente (ANDERSON; SWEENEY; WILLIANS, 2007).

Uma das maneiras de detectar esse problema é por meio do Fator de Inflação da Variância (FIV) para cada parâmetro estimado, medido por:

$$FIV(\beta_j) = \frac{1}{(1 - R_j^2)} \tag{10}$$

com $j = 1, \ldots, k$. Se $FIV(\beta_j) > 10$ constata-se a presença de multicolinearidade (MONTGOMERY; RUNGER, 2013). Esta estatística mostra como a variância de um estimador é inflada pela presença de multicolinearidade (GUJARATI; PORTER, 2011).

Além da estatística FIV existem outras medidas e testes utilizados para verificar se a multicolinearidade existente causará problemas na análise. De acordo com Anderson, Sweeney e Willians (2007) se o coeficiente de correlação amostral entre duas variáveis independentes for maior que 0,70 ou menor que − 0,70 a multicolinearidade constitui um problema potencial e deve-se evitar incluir variáveis altamente correlacionadas no modelo.

2.2.6 Variável Binária (Variável “Dummy”)

Embora grande parte das análises de regressão envolvam variáveis independentes quantitativas, em muitas situações é necessário analisar variáveis independentes

qualitativas, como sexo (masculino, feminino), método de pagamento (dinheiro, cartão de crédito), dentre outras (ANDERSON; SWEENEY; WILLIANS, 2007).

Para que se possa realizar a análise é necessário criar uma ou mais variáveis fictícias assumindo valores numéricos, de forma que representem todas as categorias da variável qualitativa considerada (CHARNET et al., 2008). A variável fictícia é frequentemente denominada variável "dummy" ou ainda variável binária, pois assume apenas os valores 0 e 1 (HOFFMANN, 2016).

Suponha que A seja uma variável independente qualitativa com k categorias sendo elas $A_1, A_2, \dots, A_k$, segundo Charnet et al. (2008), a construção das variáveis binárias pode ser feita definindo-se $(k-1)$ variáveis do tipo x_i' assumindo valores 0 e 1, tal que para $i = 1, \dots, k-1$, tem-se:

$$x_i' = \begin{cases} 1, se\ pertence\ a\ categoria\ A_i \\ 0, se\ pertence\ a\ categoria\ A_j, com\ j \neq i \end{cases}$$

Pensando nas variáveis x_i' definidas como sequência, obtém-se:

$$\begin{gathered} x_1' \Longrightarrow (1, 0, \dots, 0) \Longrightarrow categoria\ A_1 \\ x_2' \Longrightarrow (0, 1, \dots, 0) \Longrightarrow categoria\ A_2 \\ \vdots \\ x_{k-1}' \Longrightarrow (0, 0, \dots, 1) \Longrightarrow categoria\ A_{k-1} \\ e \\ (0, 0, \dots, 0) \Longrightarrow categoria\ A_k \end{gathered}$$

2.2.7 Análise de Resíduos

Os resíduos (e_i) de um modelo de regressão são úteis na verificação das pressuposições do modelo e são obtidos pela subtração entre um valor observado (y_i) e um valor ajustado da variável dependente $(\hat{y}_i)$ (ANDERSON; SWEENEY; WILLIANS, 2007).

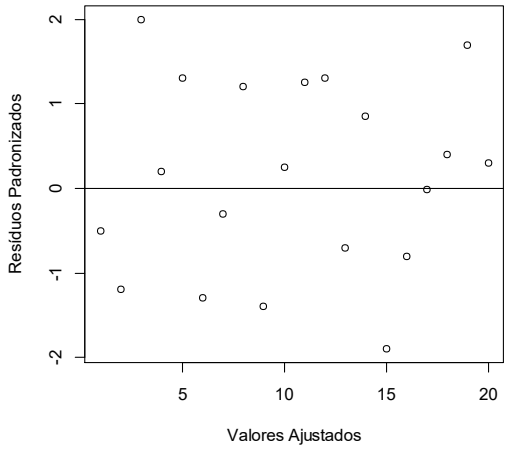
De acordo com Charnet et al. (2008) os resíduos possuem uma relação muito forte com a qualidade do ajuste e com a confiabilidade dos testes estatísticos sobre os parâmetros ajustados, e por isso, a análise desses resíduos é extremamente importante após o ajuste de qualquer modelo.

Um modelo de regressão com homoscedasticidade e sem autocorrelação assegura que o estimador de mínimos quadrados seja o melhor estimador linear não tendencioso e os erros-padrões calculados para este estimador sejam corretos, o que garante intervalos de confiança e testes de hipóteses não enganosos (HILL; GRIFFITHS; JUDGE, 2003). Além disso, os resíduos de um modelo de regressão funcional devem ser normalmente distribuídos, visto que os testes de hipóteses e as estimativas de intervalo se baseiam no pressuposto de que os erros possuem distribuição normal (HILL; GRIFFITHS; JUDGE, 2003).

Frequentemente utiliza-se os resíduos padronizados para realizar essa análise e para isso calcula-se $d_i = e_i/\sqrt{\hat{\sigma}^2}$, em que $\hat{\sigma}^2 = QMResíduo$ (MONTGOMERY; RUNGER, 2013). Os valores de d_i serão variáveis com média zero e variância um e ainda não terão unidade de medida, podendo ser comparados com os resíduos padronizados de outras regressões (GUJARATI; PORTER, 2011). De acordo com Montgomery e Runger (2013) se os erros seguirem uma distribuição normal, aproximadamente 95% dos resíduos padronizados devem estar no intervalo $(-2\,;+2)$.

Uma forma de analisar os resíduos é construir gráficos com os resíduos padronizados na ordenada e os valores ajustados da variável dependente na abcissa, sendo o gráfico ideal, indicando que nenhuma pressuposição foi violada, aquele que apresenta uma disposição de pontos bem aleatória, não apresentando nenhum tipo de tendência aparente, conforme figura 1 (CHARNET et al., 2008).

Figura 1 – Gráfico dos resíduos padronizados × valores ajustados



Fonte: A autora (2023)

Além da análise gráfica também pode-se verificar as pressuposições do modelo de regressão por meio de testes de hipóteses, como o de normalidade de Kolmogorov-Smirnov, o teste de independência de Durbin-Watson e o teste de heterocedasticidade de Breusch-Pagan descritos abaixo.

2.2.7.1 Teste de Kolmogorov-Smirnov

Segundo Fávero et al. (2009), o teste de Kolmogorov-Smirnov é um dos testes mais utilizados para se testar normalidade e consiste em comparar a distribuição de frequência acumulada de um conjunto de dados amostrais com a distribuição esperada.

Seja $F_{esp}(X)$ uma função de distribuição normal de frequências relativas acumuladas da variável X e $F_{obs}(X)$ uma função de distribuição observada de frequências relativas acumuladas da variável X. As hipóteses do teste são $F_{obs}(X) = F_{esp}(X)$, que representa a hipótese H_0 e afirma que a amostra provém de uma distribuição normal com média μ e desvio padrão σ ou $F_{obs}(X) \neq F_{esp}(X)$, que representa a hipótese H_1 e afirma que a amostra não provém de uma distribuição normal com média μ e desvio padrão σ. Sua estatística de teste é dada por:

$$D_{cal} = max\{|F_{esp}(X_i) - F_{obs}(X_i)|; |F_{esp}(X_i) - F_{obs}(X_{i-1})|\}, \quad i = 1, \dots, n \tag{11}$$

Rejeita-se a hipótese H_0 se $D_{cal} > D_c$, em que D_c representa o valor crítico obtido em uma tabela específica para o teste de Kolmogorov-Smirnov.

Quando este teste é utilizado apenas com as estimativas amostrais dos parâmetros μ e σ sugere-se efetuar uma correção ao teste, denominada correção de Lilliefors.

2.2.7.2 Teste de Durbin-Watson

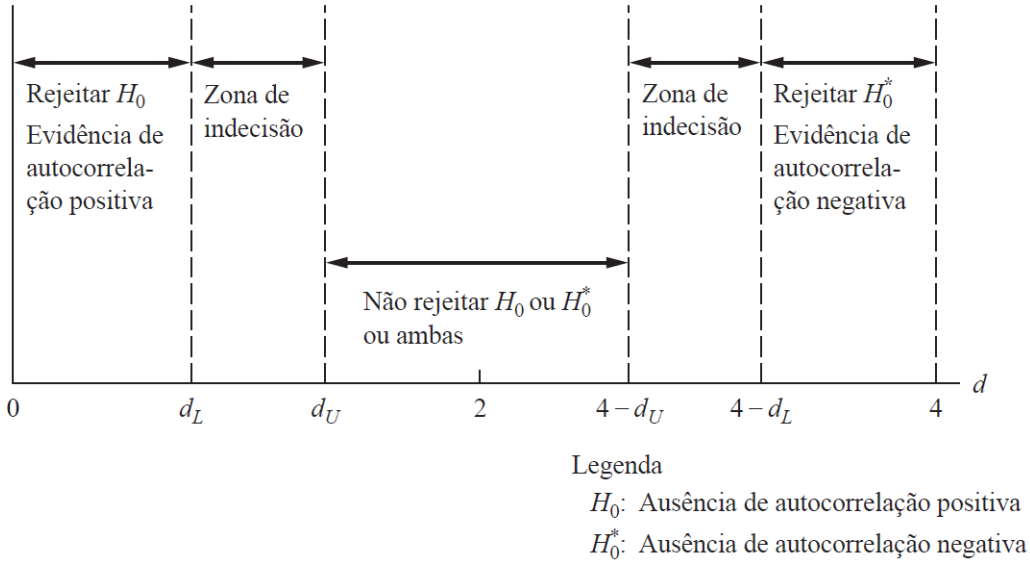
De acordo com Gujarati e Porter (2011), o teste de Durbin-Watson, conhecido como estatística *d* de Durbin-Watson, é utilizado para detectar a presença de correlação entre os resíduos e é definido por:

$$d = \frac{\sum_{t=2}^{t=n}(\hat{e}_t - \hat{e}_{t-1})^2}{\sum_{t=1}^{t=n}\hat{e}_t^2} \quad (12)$$

que é a razão entre a soma das diferenças, elevadas ao quadrado, entre resíduos sucessivos e a $SQResíduo$.

O valor estimado para d deve estar entre 0 e 4, porém os limites para tomada de decisão podem variar de acordo com o tamanho da amostra e para facilitar foram criados o limite inferior d_L e o limite superior d_U (GUJARATI; PORTER, 2011), cujas hipóteses são testadas de acordo com a figura 2, abaixo:

Figura 2 – Estatística *d* de Durbin Watson



Fonte: Gujarati e Porter (2011)

2.2.7.3 Teste de Breusch-Pagan-Godfrey

O teste de Breusch-Pagan-Godfrey descrito em Gujarati e Porter (2011) consiste em testar se a variância do erro é homocedástica e requer a normalidade dos resíduos. Supondo que a variância do erro, σ_i^2, seja:

$$\sigma_i^2 = f(\alpha_1 + \alpha_2 Z_{2i} + \cdots + \alpha_m Z_{mi}) \quad (13)$$

onde Z são variáveis não estocásticas. Suponha, especificamente, que:

$$\sigma_i^2 = \alpha_1 + \alpha_2 Z_{2i} + \cdots + \alpha_m Z_{mi} \quad (14)$$

sendo σ_i^2 uma função linear dos Z; Z pode ser algumas ou todas as variáveis independentes. Se $\alpha_2 = \cdots = \alpha_m = 0$ tem-se que $\sigma_i^2 = \alpha_1$, uma constante. Assim, para testar se σ_i^2 é homocedástico basta testar a hipótese de que $\alpha_2 = \cdots = \alpha_m = 0$.

Para a realização do teste obtém-se os resíduos $(\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_n)$ do ajuste da equação de regressão, calcula-se $\tilde{\sigma}^2 = \sum_{i=1}^{n} \hat{e}_i^2/n$, obtém-se as variáveis $p_i = \hat{e}_i^2/\tilde{\sigma}^2$ e posteriormente realiza-se uma análise de regressão de p_i construída sobre os Z's da seguinte forma:

$$p_i = \alpha_1 + \alpha_2 Z_{2i} + \cdots + \alpha_m Z_{mi} + v_i \quad (15)$$

em que v_i é o resíduo dessa regressão.

Considerando $SQRes$ como sendo a soma de quadrados de resíduos dessa análise de regressão, a estatística deste teste será dada por $\Theta = \frac{1}{2} SQRes$. A estatística Θ segue uma distribuição qui-quadrado com $m - 1$ graus de liberdade, quando os resíduos $(\hat{e}_i'\text{s})$ forem normalmente distribuídos, houver homoscedasticidade e o tamanho da amostra n aumentar indefinidamente.

A hipótese H_0 de homoscedasticidade será rejeitada se Θ for maior que o valor crítico de χ^2 ao nível de significância escolhido.

2.2.8 Observações Influentes

De acordo com Montgomery e Runger (2013), na regressão múltipla, eventualmente, encontra-se subconjuntos de observações influentes que estão distantes do resto do conjunto de dados. Assim, faz-se necessário analisar os pontos que são influentes para determinar se esses pontos influenciam de maneira errônea o modelo, e se for o caso, eliminá-los.

Há vários métodos para detectar pontos influentes, um deles é a Distância de Cook calculada por meio da fórmula:

$$D_i = \frac{r_i^2}{p}\frac{h_{ii}}{(1-h_{ii})} \tag{16}$$

com $i = 1, \dots, n$. Sendo h_{ii} o i-ésimo elemento da matriz $H = X(X'X)^{-1}X'$ correspondendo a variância do valor ajustado $\hat{Y}$, $h_{ii}/(1-h_{ii})$ é a distância do i-ésimo ponto a partir do centroide dos outros $(n-1)$ pontos e $r_i = e_i/\sqrt{\hat{\sigma}^2(1-h_{ii})}$.

Existem na literatura vários critérios para encontrar o valor limite de D_i a partir do qual considera-se os pontos como influentes. Dentre estes critérios considera um ponto como influente se $D_i > {}^{4}/_{n}$, sendo n o tamanho da amostra (ALTMAN; KRZYWINSKI, 2016).

2.2.9 Seleção de Variáveis

É comum na análise de regressão linear múltipla surgirem dúvidas em relação a inclusão de todas as variáveis independentes disponíveis ou a inclusão apenas de um subconjunto delas (CHARNET et al., 2008).

Para selecionar as variáveis independentes utiliza-se frequentemente o método denominado regressão por etapas (*stepwise*), no qual as variáveis são introduzidas uma por vez ou todas as variáveis possíveis são incluídas em um ajuste do modelo de regressão linear múltipla e, em seguida, rejeita-se uma a uma (GUJARATI; PORTER, 2011).

O método *stepwise* também denominado método "passo a passo", permite alternadamente, eliminações e inclusões de variáveis no modelo, e além dele existem os métodos (CHARNET et al., 2008):

i. Todas as regressões possíveis: a seleção é feita entre todos os modelos possíveis.

ii. Método "passo a frente" (*forward*): o método se inicia com o ajuste de um modelo contendo apenas uma das variáveis independentes e, segundo um critério específico, outras variáveis podem ser acrescentadas.

iii. Métodos “passo atrás” (*backward*): se inicia com o ajuste de um modelo com todas as variáveis independentes e, segundo um critério específico, as variáveis podem ser eliminadas.

Os passos para o procedimento do método stepwise são (CHARNET et al., 2008):

- **Passo 1:** ajustar o modelo reduzido composto por m variáveis e obter a $SQRegressão$ do modelo reduzido;

- **Passo 2:** para cada variável que não está no modelo do Passo 1, considerar o modelo completo, adicionando esta variável extra, e calcular a $SQRegressão$ do modelo completo e o estimador da variância do erro $\widehat{\sigma^2}$ para a obtenção do valor da estatística;

- **Passo 3:** achar o máximo dos valores das estatísticas obtidos no Passo 2, indicado por F_{max};

- **Passo 4:** seja F_{in} o quantil da distribuição F com 1 e $(n-m-2)$ graus de liberdade:

  Se $F_{max} > F_{in} \longrightarrow$ ir para o Passo 5 com o modelo completo formado por $(m+1)$ variáveis, sendo as m variáveis do passo 1 e a variável que teve o valor da estatística igual a F_{max};
  Se $F_{max} < F_{in} \longrightarrow$ ir para o Passo 5 com o modelo igual ao do passo 1 ou finalizar o processo se no passo 8 não houve nenhuma variável eliminada;

- **Passo 5:** ajustar o modelo completo composto por k variáveis, sendo k igual a m ou igual a $(m+1)$, e obter a $SQRegressão$ do modelo completo e o estimador de variância do erro $\widehat{\sigma^2}$;

- **Passo 6:** para cada uma das k variáveis do modelo do Passo 5, considerar o modelo reduzido, eliminando esta variável, e calcular a $SQRegressão$ do modelo reduzido para a obtenção do valor da estatística;

- **Passo 7:** achar o mínimo dos k valores da estatística obtidos no Passo 6, indicado por F_{min};

- **Passo 8:** seja F_{out} o quantil da distribuição F com 1 e $(n-k-1)$ graus de liberdade, tem-se duas suposições:

  Se $F_{min} > F_{out} \longrightarrow$ não fazer a eliminação de nenhuma variável e voltar ao Passo 1 com o modelo reduzido formado por k variáveis ou finalizar o processo se no Passo 4 não houve nenhuma variável anexada;
  Se $F_{min} < F_{out} \longrightarrow$ fazer a eliminação da variável que possuir o valor da estatística igual a F_{min} e voltar para o Passo 1 com o modelo reduzido com $(k-1)$ variáveis.

Uma sugestão para tornar a inclusão de variáveis mais fácil é escolher valores menores para F_{in} em relação a F_{out} (CHARNET et al., 2008).

3 METODOLOGIA

3.1 Conjunto de Dados

Para atender os objetivos desta pesquisa utilizou-se o conjunto de dados "Compartilhamento de Bicicletas", disponível na plataforma online Kaggle (KAGGLE, 2022). Os dados são referentes ao provedor de compartilhamento BikeIndia, dos Estados Unidos da América (EUA), que enfrentou diversas dificuldades para se manter no mercado durante a pandemia, e com o objetivo de alavancar seus negócios decidiu utilizar dados coletados em anos anteriores à pandemia para entender melhor a demanda por bicicletas compartilhadas (KAGGLE, 2022).

O conjunto de dados possui variáveis coletadas em 729 dias entre janeiro do ano de 2018 e dezembro do ano de 2019. As variáveis observadas foram:

i. estação:
    a. estação1: inverno
    b. estação2: primavera
    c. estação3: verão
    d. estação4: outono
ii. ano
iii. mês:
    a. mês1: janeiro
    b. mês2: fevereiro
    c. mês3: março
    d. mês4: abril
    e. mês5: maio
    f. mês6: junho
    g. mês7: julho
    h. mês8: agosto
    i. mês9: setembro
    j. mês10: outubro
    k. mês11: novembro
    l. mês12: dezembro
iv. feriado
v. dia da semana:
    a. dia1: domingo
    b. dia2: segunda-feira
    c. dia3: terça-feira
    d. dia4: quarta-feira
    e. dia5: quinta-feira

   f. dia6: sexta-feira
   g. dia7: sábado
vi. diau: dia útil
vii. situação climática:
   a. clima1 (situação climática 1): claro, poucas nuvens e/ou parcialmente nublado
   b. clima2 (situação climática 2): névoa + nublado, névoa + nuvens quebradas, névoa + poucas nuvens e/ou névoa
   c. clima3 (situação climática 3): neve fraca, chuva fraca + trovoada + nuvens dispersas e/ou chuva fraca + nuvens dispersas
viii. temp: temperatura
ix. senter: sensação térmica
x. umi: umidade
xi. vento: velocidade do vento
xii. cba: contagem de bicicletas alugadas

3.2 Análise Descritiva

Inicialmente realizou-se uma análise descritiva das variáveis. Para isso, procedeu-se o cálculo da média aritmética, desvio padrão, coeficiente de variação, mínimo e máximo para as variáveis quantitativas temperatura, sensação térmica, umidade, velocidade do vento e contagem de bicicletas alugadas. Depois, construiu-se gráficos de barras para as variáveis qualitativas ano, mês, dia da semana e estação e devido a grande diferença na quantidade de dias úteis e não úteis, feriados e não feriados e de cada situação climática optou-se por contruir gráficos boxplots para as variáveis qualitativas dia útil, feriado e situação climática.

3.3 Análise de Regressão Linear Múltipla

Devido ao fato dos dados terem sido coletados ao longo do tempo existia a possibilidade de ocorrer dependência nos resíduos da análise de regressão, não atendendo a uma das pressuposições do modelo. Dessa forma, antes de iniciar a análise de regressão aleatorizou-se as informações.

Para possibilitar a inclusão das variáveis qualitativas no ajuste do modelo de regressão realizou-se a transformação de suas categorias em variáveis binárias da seguinte forma:

- Para a variável estação:

$$estacao1' = \begin{cases} 1, se\ for\ inverno \\ 0, se\ não\ for\ inverno \end{cases} \qquad estacao3' = \begin{cases} 1, se\ for\ verão \\ 0, se\ não\ for\ verão \end{cases}$$

$$estacao2' = \begin{cases} 1, se\ for\ primavera \\ 0, se\ não\ for\ primavera \end{cases} \qquad estacao1' = estacao2' = estacao3' = 0, se\ for\ outono$$

- Para variável ano:

$$ano' = \begin{cases} 1, se\ for\ 2019 \\ 0, se\ for\ 2018 \end{cases}$$

- Para variável mês:

$$mes1' = \begin{cases} 1, se\ for\ janeiro \\ 0, se\ não\ for\ janeiro \end{cases} \qquad mes7' = \begin{cases} 1, se\ for\ julho \\ 0, se\ não\ for\ julho \end{cases}$$

$$mes2' = \begin{cases} 1, se\ for\ fevereiro \\ 0, se\ não\ for\ fevereiro \end{cases} \qquad mes8' = \begin{cases} 1, se\ for\ agosto \\ 0, se\ não\ for\ agosto \end{cases}$$

$$mes3' = \begin{cases} 1, se\ for\ março \\ 0, se\ não\ for\ março \end{cases} \qquad mes9' = \begin{cases} 1, se\ for\ setembro \\ 0, se\ não\ for\ setembro \end{cases}$$

$$mes4' = \begin{cases} 1, se\ for\ abril \\ 0, se\ não\ for\ abril \end{cases} \qquad mes10' = \begin{cases} 1, se\ for\ outubro \\ 0, se\ não\ for\ outubro \end{cases}$$

$$mes5' = \begin{cases} 1, se\ for\ maio \\ 0, se\ não\ for\ maio \end{cases} \qquad mes11' = \begin{cases} 1, se\ for\ novembro \\ 0, se\ não\ for\ novembro \end{cases}$$

$$mes6' = \begin{cases} 1, se\ for\ junho \\ 0, se\ não\ for\ junho \end{cases} \qquad mes1' = mes2' = \cdots = mes11' = 0, se\ for\ dezembro$$

- Para a variável feriado:

$$feriado' = \begin{cases} 1, se\ for\ feriado \\ 0, se\ não\ for\ feriado \end{cases}$$

- Para a variável dia da semana:

$$dia1' = \begin{cases} 1, \text{se for domingo} \\ 0, \text{se não for domingo} \end{cases} \qquad dia4' = \begin{cases} 1, \text{se for quarta} \\ 0, \text{se não for quarta} \end{cases}$$

$$dia2' = \begin{cases} 1, \text{se for segunda} \\ 0, \text{se não for segunda} \end{cases} \qquad dia5' = \begin{cases} 1, se\ for\ quinta \\ 0, se\ não\ for\ quinta \end{cases}$$

$$dia3' = \begin{cases} 1, \text{se for terça} \\ 0, \text{se não for terça} \end{cases} \qquad dia6' = \begin{cases} 1, se\ for\ sexta \\ 0, se\ não\ for\ sexta \end{cases}$$

$$dia1' = dia2' = \cdots = dia6' = 0, se\ for\ sábado$$

- Para variável dia útil:

$$diau' = \begin{cases} 1, se\ for\ dia\ útil \\ 0, se\ não\ for\ dia\ útil \end{cases}$$

- Para variável situação climática:

$$clima1' = \begin{cases} 1, se\ for\ situação\ climática\ 1 \\ 0, se\ não\ for\ situação\ climática \text{ tipo } 1 \end{cases}$$

$$clima2' = \begin{cases} 1, se\ for\ situação\ climática\ tipo\ 2 \\ 0, se\ não\ for\ situação\ climática\ tipo\ 2 \end{cases}$$

$$clima1' = clima2' = 0, se\ for\ situação\ climática\ tipo\ 3$$

Após as transformações calculou-se o coeficiente de correlação linear simples para todos os pares de variáveis, com os objetivos de identificar as variáveis mais correlacionadas com o número de bicicletas alugadas e também verificar a ocorrência de multicolinearidade substancial ($r > 0{,}7$ ou $r < -0{,}7$) entre as variáveis independentes, ou seja, verificar a ocorrência de multicolinearidade que pudesse gerar potenciais problemas para a análise.

Posteriormente realizou-se a análise de regressão linear múltipla considerando a contagem de bicicletas alugadas (cba) como variável dependente, e como independentes as variáveis ano (ano'), feriado (feriado'), dia útil (diau'), temperatura (temp), sensação térmica (senter), umidade (umi), velocidade do vento (vento), as categorias estacao1' (inverno), estacao2' (primavera), estacao3' (verão) da variável estação, as categorias mes1', mes2', mes3', mes4', mes5', mes6', mes7', mes8', mes9', mes10', mes11' da variável mês, as categorias dia1', dia2', dia3', dia4', dia5', dia6' da variável dia da semana e as categorias clima1', clima2' da variável situação climática.

Para selecionar as variáveis independentes que realmente fariam parte da modelagem utilizou-se o método de seleção *stepwise*. Após a seleção e o ajuste do modelo, verificou-se novamente a ocorrência de multicolinearidade substancial por meio do Fator de Inflação de Variância (FIV) e em seguida, verificou-se a presença de observações influentes por meio da distância de Cook.

Posteriormente, procedeu-se a análise de resíduos para testar se as pressuposições do modelo haviam sido atendidas. Esta análise foi feita tanto por meio de gráficos quanto por testes de hipóteses. Utilizou-se o teste de Kolmogorov-Smirnov com correção de Lilliefors, para testar a normalidade dos resíduos, o teste de Durbin-Watson para testar a independência dos resíduos e o teste de Breusch-Pagan-Godfrey para testar a homogeneidade de variância. A significância do ajuste foi verificada por meio do teste F (Análise de Variância) e de testes t. Para realização de todos os testes de hipóteses, utilizou-se um nível de significância de 5%.

Para testar as hipóteses de todos os testes citados acima, utilizou-se o valor *p*, que de acordo com Montgomery e Runger (2013) é o menor nível de significância que conduz à rejeição da hipótese nula, com os dados amostrais. Rejeitou-se as hipóteses nulas para valor $p < 0,05$. Por fim, calculou-se o coeficiente de determinação ajustado para verificar a qualidade do ajuste.

3.4 Software

Todas as análises estatísticas foram feitas no software R (R CORE TEAM, 2022). Para salvar os dados aleatorizados utilizou-se o pacote *clipr* (LINCOLN et al, 2022). Para fazer o gráfico de correlação utilizou-se o pacote *corrplot* (WEI; SIMKO, 2021). Realizou-se o teste de Kolmogorov-Smirnov com o pacote *nortest* (GROSS; LIGGES, 2015) e Durbin-Watson e Brescheu-Pagan com o pacote *lmtest* (HOTHORN et al., 2022). Para fazer os gráficos da distância de Cook e dos resíduos utilizou-se o pacote *olsrr* (HEBBALI, 2020)

4 RESULTADOS E DISCUSSÕES

4.1 Análise Descritiva

Uma análise inicial das variáveis quantitativas mostrou que a contagem de bicicletas alugadas (cba) foi a variável com maior variação em relação à média, apresentando coeficiente de variação (CV) próximo a 43% (Tabela 2). Enquanto em alguns dias alugou-se apenas 22 bicicletas, em outros essa quantidade foi superior a 8700, evidenciando o fato de determinados dias apresentarem características mais favoráveis a utilização desse meio de transporte.

A velocidade do vento (vento) e a temperatura (temp) também apresentaram grande variação em relação à média com CV de 40,73% e 36,95%, respectivamente, seguida pela sensação térmica (senter) com CV de 34,35%. Embora a temperatura média tenha ficado em torno de 20,33°C com sensação térmica média de 23,73ºC, em alguns dias a temperatura superou os 35°C e em outros esteve bem abaixo da média, com mínima registrada de 2,42°C. O mesmo ocorreu com a velocidade do vento que variou de 1,5 km/h a 34 km/h. A umidade do tempo (umi) apresentou o menor CV (22,36%) dentre as variáveis, indicando menor variação de seus valores em torno da umidade média de 62,85%.

Tabela 2 – Estatísticas descritivas das variáveis quantitativas

Estatísticas	temp (ºC)	senter (ºC)	umi (%)	vento (km/h)	cba
Média aritmética	20,33	23,73	62,85	12,76	4513
Desvio Padrão	7,51	8,15	14,06	5,20	1931,98
CV	36,95%	34,36%	22,36%	40,73%	42,81%
Mínimo	2,42	3,95	18,79	1,50	22
Máximo	35,33	42,05	97,25	34,00	8714

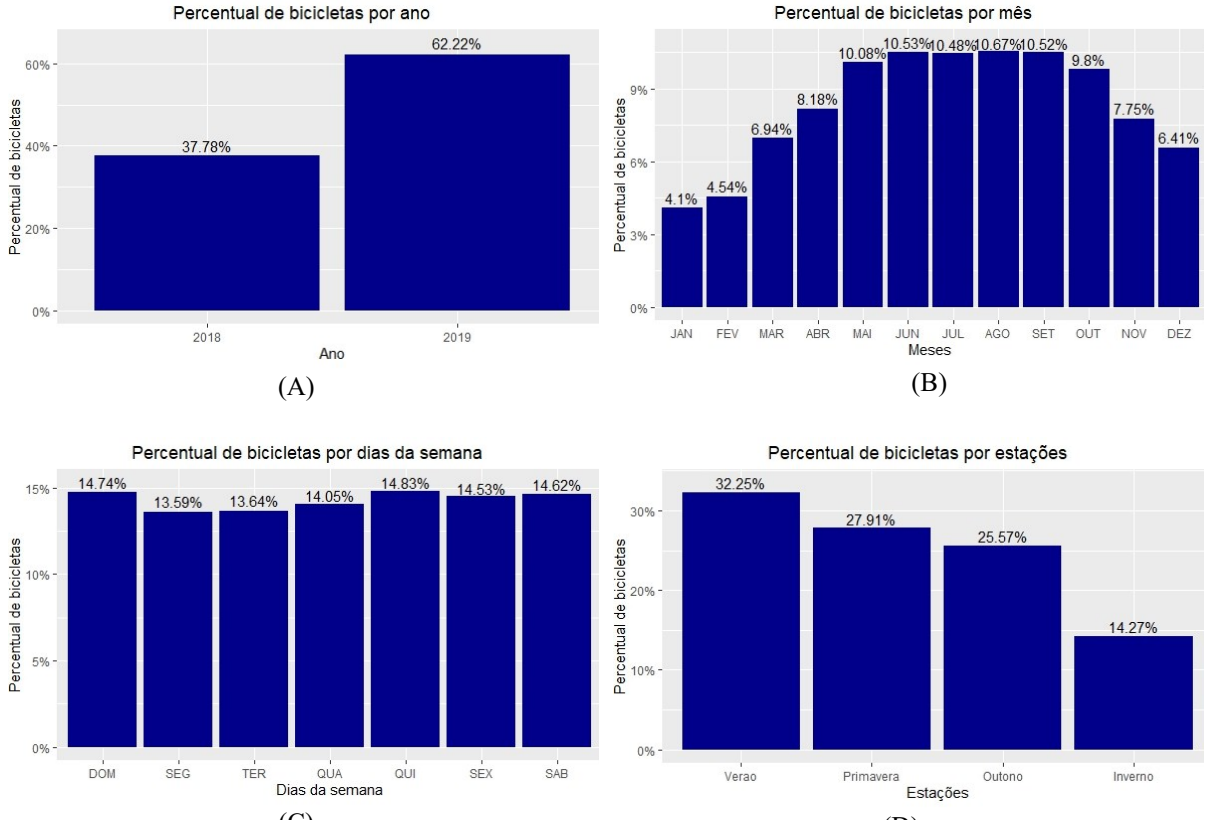
CV: Coeficiente de Variação; temp: temperatura; senter: sensação térmica, umi: umidade, vento: velocidade do vento e cba: contagem do total de bicicletas alugadas. Fonte: A autora (2023)

Analisando graficamente as variáveis qualitativas nota-se que os aluguéis de bicicletas tiveram um aumento de 24,44% no ano de 2019 em relação à 2018 (Figura 3A), sendo os meses de agosto, junho, setembro, julho, maio e outubro, responsáveis pelos maiores percentuais de aluguéis, com respectivamente 10,67%, 10,53%, 10,52%, 10,48%, 10,08% e 9,8% do total de bicicletas alugadas (Figura 3B).

Percebe-se que os dias da semana apresentaram percentuais parecidos de bicicletas alugadas, todos superiores a 13%, sendo quinta-feira o dia com mais aluguéis (Figura 3C), equivalente a 14,83%.

O verão é a estação com maior quantidade de bicicletas alugadas, equivalente a 32,25% dos aluguéis, seguida da primavera (27,91%), outono (25,57%) e inverno (14,27%) (Figura 3D). Analisando dados de cinco estações de Montreal no Canadá, Miranda-Moreno e Nosal (2011) também concluíram que no verão atingiu-se o pico de bicicletas alugadas. Além disso, Rudloff e Lackner (2014) verificaram que a demanda do sistema de compartilhamento Citybike é claramente menor no inverno.

Figura 3 – Percentuais de bicicletas alugadas, por ano, meses, dias da semana e estações do ano



Fonte: A autora (2023)

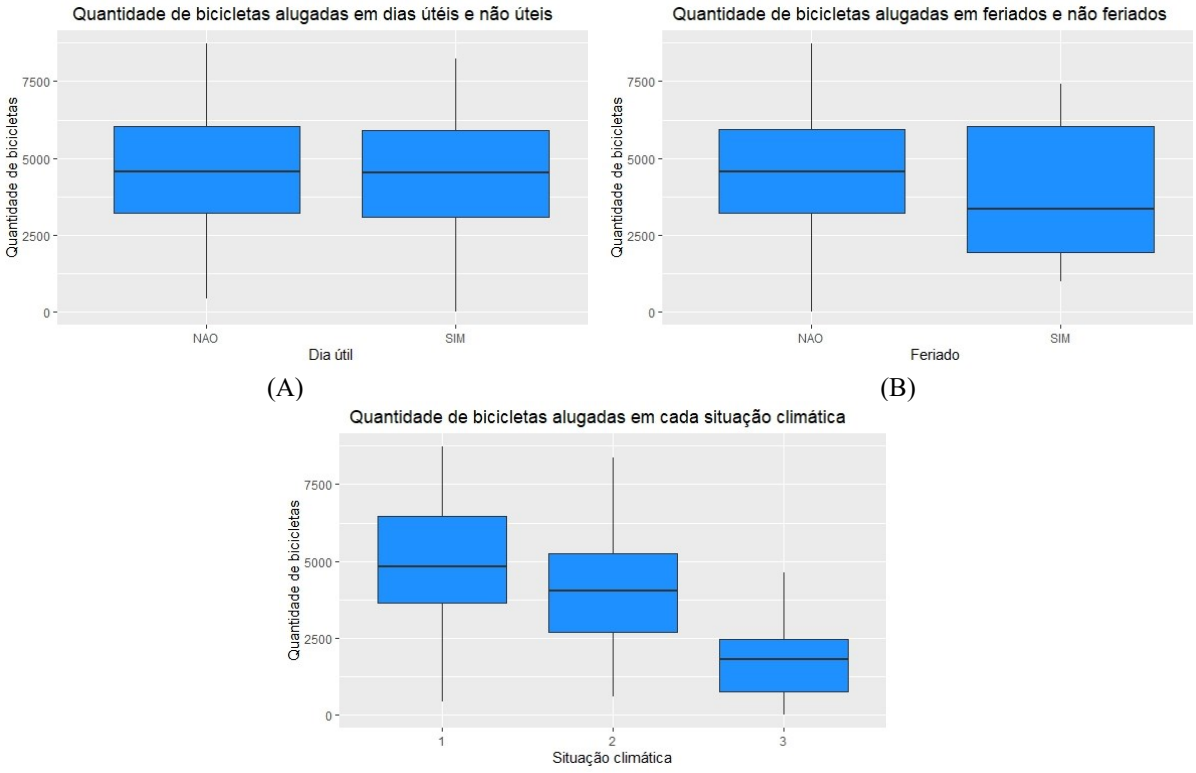
Comparando dias úteis com dias não úteis observou-se que tanto a contagem de bicicletas alugadas (cba) como a variabilidade média dessa contagem foi praticamente a mesma, pois apresentaram boxplots muito parecidos (Figura 4A). Além disso, verificou-se um comportamento aproximadamente simétrico da variável cba, indicando que o valor da mediana foi muito próximo ao valor da média aritmética, ou seja, aproximadamente 50% das contagens estavam acima da média e 50% abaixo da média (Figura 4A).

Nos feriados é possível verificar uma redução nos aluguéis em relação aos dias que não eram feriados, além de uma variabilidade média maior na quantidade de bicicletas alugadas. Verificou-se ainda comportamento assimétrico à direita, mostrando que em alguns dias de feriado alugou-se uma quantidade bem maior de bicicletas se comparado a maior parte dos dias de feriado, nos quais esses aluguéis apresentou redução (Figura 4B).

Observou-se na Figura 4C que os dias com o clima do tipo claro, poucas nuvens e/ou parcialmente nublado (situação climática 1) favoreceram o aluguel de bicicletas. Em contrapartida, verificou-se uma quantidade menor de aluguéis em dias com o clima do tipo névoa + nublado, névoa + nuvens quebradas, névoa + poucas nuvens e/ou névoa (situação climática 2) e grande redução nos aluguéis de bicicletas em dias com o clima do tipo neve fraca, chuva fraca + trovoada + nuvens dispersas e/ou chuva fraca + nuvens dispersas (situação climática 3) (Figura 4C).

Diferente das situaçãoes climáticas 1 e 2, na situação climática 3 a contagem de bicicletas alugadas apresentou comportamento assimétrico à esquerda, mostrando que em alguns desses dias essas contagens foram ainda mais baixas se comparadas a maioria dos dias da situação climática 3 (Figura 4C). Esses resultados já eram esperados uma vez que dias chuvosos e nublados dificultam a utilização de bicicletas como meio de transporte, influenciando negativamente sua demanda nos sistemas de compartilhamento. Kim et al. (2012) também verificaram redução no uso de bicicletas compartilhadas em dias com chuva.

Figura 4 - Boxplots da contagem bicicletas alugadas, em dias dias úteis e não úteis, feriados e não feriados e em cada situação climática.



Situação climática 1: clima do tipo claro, poucas nuvens e/ou parcialmente nublado; Situação climática 2: clima do tipo névoa + nublado, névoa + nuvens quebradas, névoa + poucas nuvens e/ou névoa; Situação climática 3: clima do tipo neve fraca, chuva fraca + trovoada + nuvens dispersas e/ou chuva fraca + nuvens dispersas.
Fonte: A autora (2023)

4.2 Análise de Regressão Linear Múltipla

Antes de ajustar um modelo de regressão é importante analisar as correlações entre as variáveis envolvidas. Ao realizar esta análise verificou-se que as variáveis mais correlacionadas com a quantidade de bicicletas alugadas foram temperatura (temp), sensação térmica (senter), ano (ano') e a categoria inverno (estacao1') da variável binária estação (Figura 4), sendo o inverno correlacionado negativamente, enquanto as demais (ano', temp e senter) apresentaram correlação positiva.

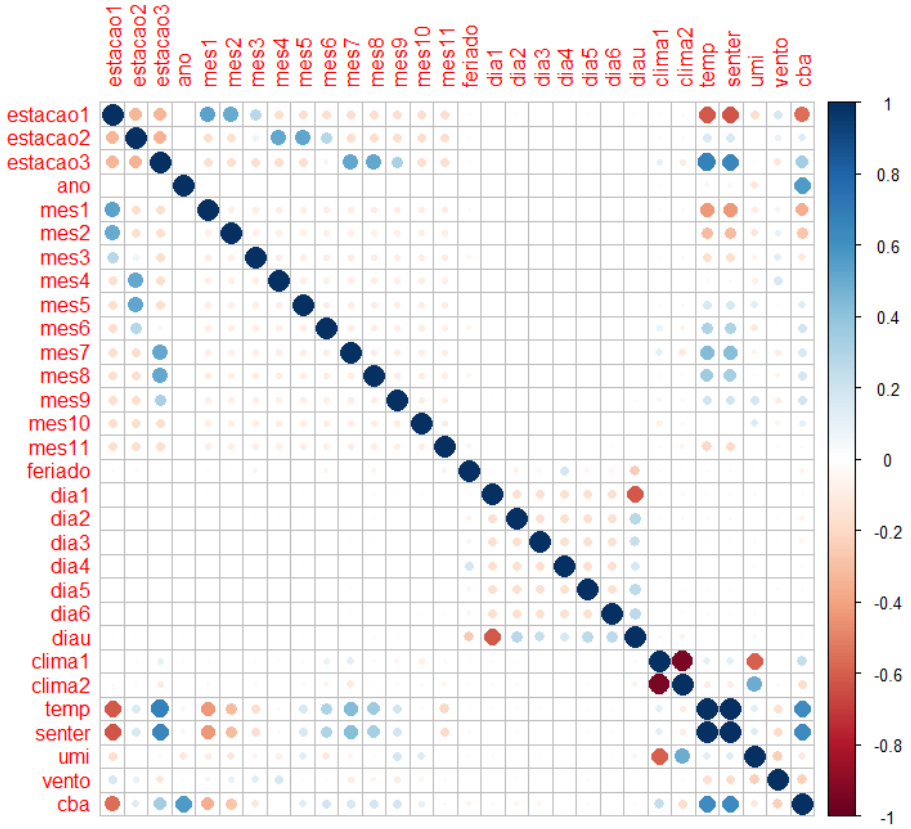
Estas correlações indicaram uma tendência de aumento da quantidade de bicicletas alugadas com o aumento da temperatura e da sensação térmica e ainda apontaram que no ano de 2019 a quantidade de bicicletas alugadas foi superior ao ano de 2018, como já havia sido observado no gráfico de barras (Figura 3A). Além disso, a correlação negativa da categoria

inverno reforçou a tendência de redução na quantidade de bicicletas alugadas durante esta estação.

Verificou-se também que as categorias inverno (estacao1') e verão (estacao3') da variável estação são altamente correlacionadas com as variáveis temp e senter, apresentando correlações inferiores a $-0{,}7$ e superiores a 0,7, respectivamente (Figura 4). Além disso, as categorias situação climática 1 (clima1') e 2 (clima2') da variável binária situação climática também estão altamente correlacionadas entre si ($r < -0{,}7$), assim como as variáveis temp e senter com r superior a 0,7. De acordo com Anderson, Sweeney e Willians (2007) estas correlações indicam um problema potencial para a análise e dessa forma optou-se por retirar do modelo as variáveis sensação térmica (senter), estação (estacao1', estacao2' e estacao3') e situação climática (clima1' e clima2').

A decisão de quais variáveis seriam retiradas foi baseada no fato de diversos estudos apresentarem a temperatura como um dos principais fatores de influência na quantidade de bicicletas alugadas como os estudos de Heinen, Van Wee e Maat (2010), Martinez (2017) e Kim (2018). Dessa forma, considerou-se importante manter a variável temp. Além disso, a necessidade de retirada de categorias das variáveis estação e situação climática impossibilitou a utilização das mesmas na análise.

Figura 5 - Gráfico de correlações



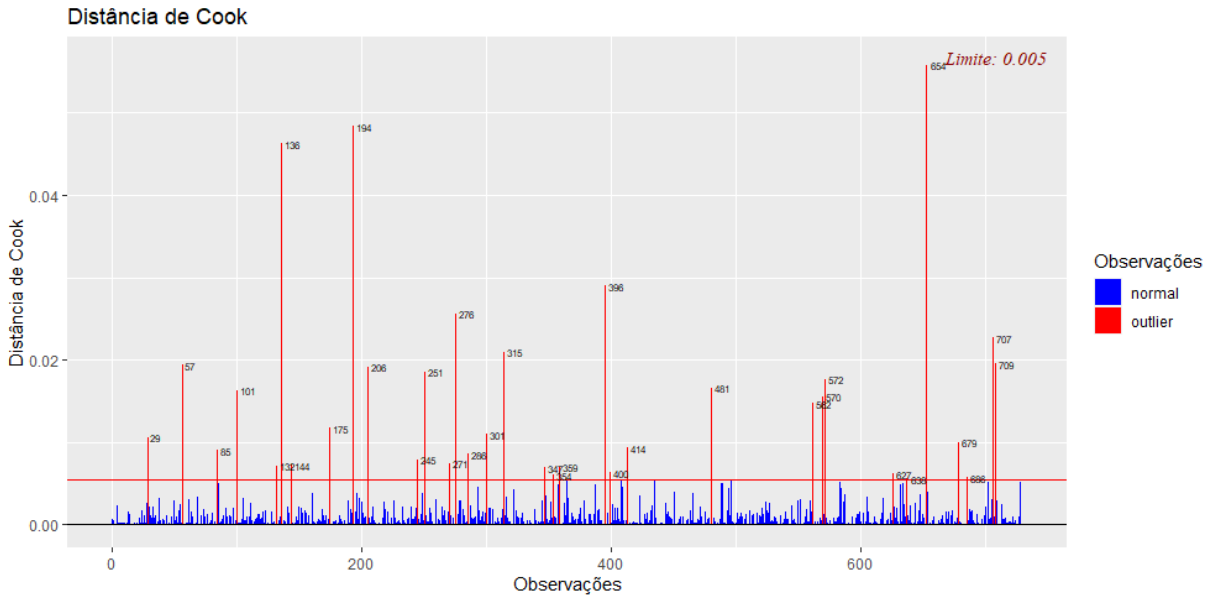
Fonte: A autora (2023)

Utilizando o método *stepwise* para seleção de variáveis no modelo contendo as variáveis independentes ano', mês (mes1', mes2', mes3', mes4', mes5', mes6', mes7', mes8', mes9', mes10', mes11'), feriado', dia da semana (dia1', dia2', weekday3', dia4', dia5', dia6'), diau', temp, um e vento, selecionou-se as variáveis independentes ano', feriado', temp, umi, vento, as categorias mes1', mes2', mes4', mes5', mes7', mes9', mes10', mes11' da variável binária mês e as categorias dia2', dia3' da variável binária dia da semana. Assim, com a exclusão de categorias das variáveis mês e dia da semana optou-se por retirar estas variáveis. Analisando o novo modelo, composto pelas variáveis ano', feriado', temp, um e vento, obteve-se todas as variáveis significativas, ao nível de significância de 5%.

Após o ajuste, verificou-se a presença de 34 observações influentes por meio da Distância de Cook (Figura 5), equivalente a menos de 5% do total de dados. Inicialmente estes pontos foram investigados para verificar se não se tratavam de possíveis erros de medida e posteriormente, analisou-se o comportamento do modelo com a retirada destas observações.

Não se indentificou indícios de erros de medida nestas observação e a retirada das mesmas praticamente não alterou o modelo, por isso decidiu-se manter estes pontos.

Figura 6 - Gráfico da Distância de Cook



Fonte: A autora (2023)

Posteriormente, verificou-se a ausência de multicolinearidade substancial entre as variáveis independentes com os valores de FIV inferiores a 10 (Tabela 3).

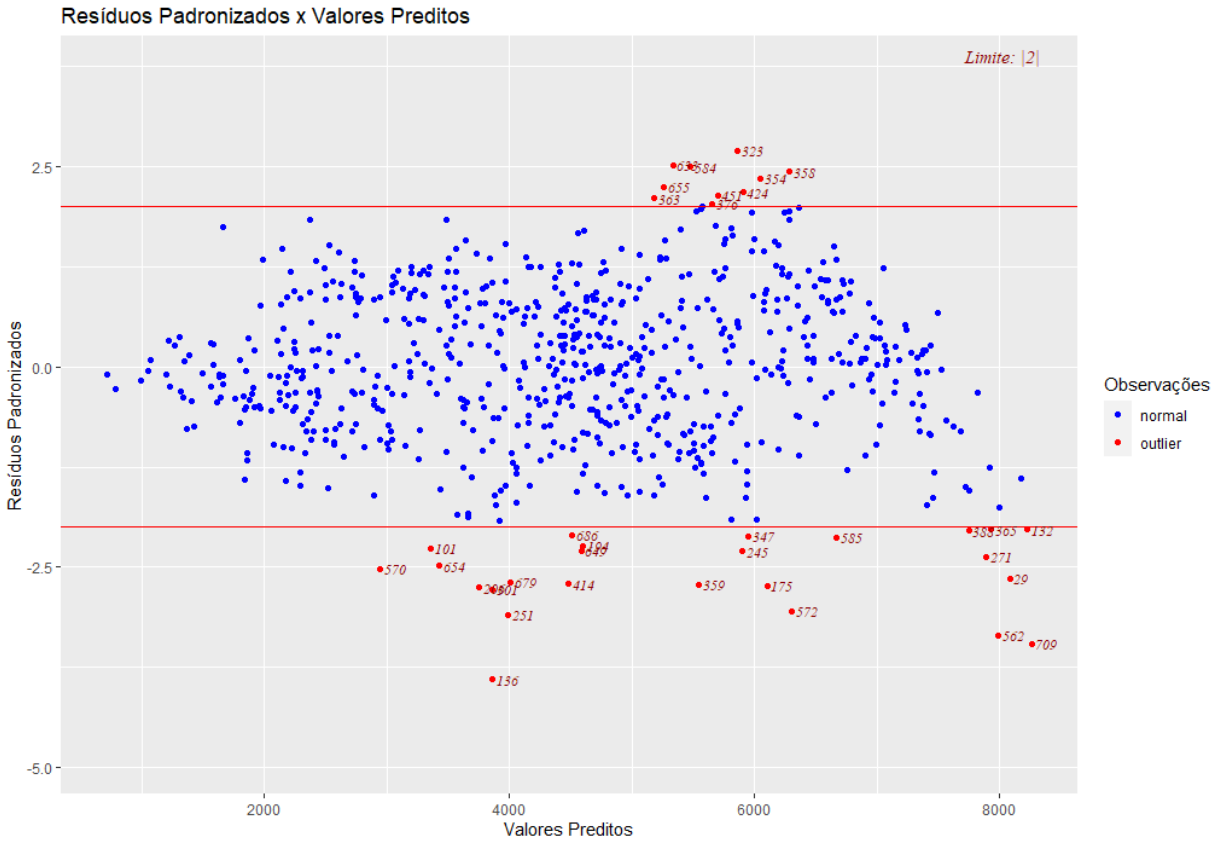
Tabela 3 - Valores do Fator de Inflação de Variância (FIV)

Variáveis	ano’	feriado’	temp	umi	vento
FIV	1,020097	1,001076	1,038630	1,092835	1,084558

ano’: ano; feriado’: feriado; temp: temperatura; umi: umidade; vento: velocidade do vento.
Fonte: A autora (2023)

Após o ajuste da equação analisou-se graficamente seus resíduos (Figura 6). Observou-se que os resíduos estão distribuídos aleatoriamente em torno do zero, não apresentando tendência. De acordo com Charnet et al. (2008) esse comportamento indica que todas as pressuposições do modelo foram satisfeitas indicando boa qualidade do ajuste e confiabilidade dos testes estatísticos sobre os parâmetros ajustados. Dessa forma, pode-se dizer pela análise gráfica que os resíduos são homocedásticos, não correlacionados entre si e seguem uma distribuição normal. Além disso, verifica-se que apenas 4,8% dos valores estão fora do intervalo entre -2 e 2 (Figura 6), reforçando a hipótese de normalidade dos resíduos.

Figura 7 - Gráfico dos resíduos



Fonte: A autora (2023)

Após a análise gráfica verificou-se as pressuposições por meio dos testes de hipóteses, e os resultados destes testes estão apresentados na tabela 4. A pressuposição de normalidade dos resíduos foi atendida, ao nível de significância de 5%, pelo teste de Kolmogorov-Smirnov e pelo teste de Durbin-Watson comprovou-se a independência dos resíduos, ao nível de significância de 5%. Entretanto, utilizando o teste de Breusch-Pagan-Godfrey rejeitou-se a hipótese de homogeneidade dos resíduos, ao nível de significância de 5%, contradizendo a análise gráfica.

Segundo Gujarati e Porter (2011), sob heterocedasticidade os estimadores de mínimos quadrados mantêm suas propriedades de consistência e não tendenciosidade, porém eles deixam de ser eficientes, tornando duvidosos os resultados dos testes de hipóteses. Os autores sugerem para estes casos a utilização de uma correção denominada correção de White, possibilitando a obtenção de erros padrão robustos de forma a obter inferências estatísticas válidas sobre os parâmetros do modelo.

Tabela 4 – Resultados dos testes de hipóteses da análise de resíduos

Testes	***Estatística***	***Valor p***
Kolmogorov-Smirnov	0,027289	0,2085
Durbin-Watson	2,014645	0,838
Breusch-Pagan- Godfrey	38,731	0,000000269

Fonte: A autora (2023)

Devido ao resultado do teste de homogeneidade utilizou-se a correção de heterocedasticidade de White sugerida por Gujarati e Porter (2011), de forma a obter erros padrão corrigidos (erros padrão robustos). Porém os resultados se mantiveram praticamente os mesmos, indicando que a heterocedasticidade não é um problema grave neste caso, o que confirma o resultado da análise gráfica. Dessa forma, optou-se pela utilização das estimativas de mínimos quadrados sem a correção de White (Tabela 6).

A análise de variância (Tabela 5) mostrou uma relação linear global significativa entre a contagem de bicicletas alugadas (cba) e as variáveis ano (ano'), feriado (feriado'), temperatura (temp), umidade (umi) e velocidade do vento (vento), ao nível de significância de 5%. Além disso, pode-se verificar na tabela 6 que todos os parâmetros do modelo foram estatisticamente significativos ao nível de significância de 5%.

Tabela 5 - Análise de Variância

FV	***GL***	***SQ***	***QM***	F_c	***Valor p***
Regressão	5	1989562230	397912446	395,33	<0,0000
Resíduo	723	727718648	1006526	-	-

FV: Fontes de Variação; *GL*: Graus de Liberdade; *SQ*: Soma de Quadrados; *QM*: Quadrado Médio; *Fc*: Valor F calculado. Fonte: A autora (2023)

Tabela 6 - Estimativas de mínimos quadrados

Coeficientes	***Estimativa***	***Erro padrão***	***Valor t***	***Valor p***
Intercepto	2721,785	246,747	11,031	< 0,0000
ano'	2002,445	75,059	26,678	< 0,0000
feriado'	-670,743	222,271	-3,018	0,00264
temp	152,950	5,046	30,312	< 0,0000
umi	-23,338	2,765	-8,439	< 0,0000
vento	-65,348	7,452	-8,769	< 0,0000

ano': ano; feriado': feriado; temp: temperatura; umi: umidade e vento: velocidade do vento. Fonte: A autora (2023)

Verificando-se a qualidade do ajuste por meio do coeficiente de determinação ajustado (R_a^2), obteve-se R_a^2 = 0,7303, ou seja, 73,03% da variabilidade na contagem de bicicletas alugadas pode ser explicada estatisticamente pela variabilidade nas variáveis ano (ano'), feriado (feriado'), temperatura (temp), umidade (umi) e velocidade do vento (vento).

O modelo de regressão ajustado é dado pela seguinte equação:

$$cba = 2721{,}785 + 2002{,}445 \times ano' - 670{,}743 \times feriado' + 152{,}950 \times temp - 23{,}338 \times umi - 65{,}348 \times vento$$

Interpretando o modelo estima-se que no ano de 2019 alugou-se em média 2002 bicicletas a mais do que em 2018. Nos feriados houve uma redução média no número de aluguéis de aproximadamente 671 bicicletas em relação aos outros dias. A redução de aluguéis de bicicleta nos feriados também foi constatada por Kim (2018) ao estudar o sistema de compartilhamento 'Tashu' de Daejon na Coréia do Sul. Acredita-se que essa redução esteja relacionada as finalidades de uso desse meio de transporte, Talavera-Garcia, Romanillos e Arias-Molinares (2021) verificaram que os utilizadores mais frequentes de um sistema de compartilhamentos na cidade de Madri, na Espanha, utilizavam a bicicleta para fins de estudo e/ou de trabalho.

Estima-se ainda que o aumento de 1°C na temperatura (temp) acarrete um aumento médio nos aluguéis de aproximadamente 153 bicicletas. Martinez (2017) e Kim (2018) também verificaram um aumento na quantidade de bicicletas alugadas devido ao aumento da temperatura. Porém Kim (2018) ressaltou que em dias com temperaturas acima de 30 ºC existe uma tendência de redução na quantidade de aluguéis.

Diferente da temperatura, a umidade e a velocidade do vento se relacionaram de forma negativa com os aluguéis de bicicleta. Com o acréscimo de 1% na umidade (umi) estimou-se uma redução média de 23,34 bicicletas alugadas. E com o aumento de 1 km/h na velocidade do vento (vento) estimou-se uma redução média de 65,35 bicicletas alugadas. Essa redução de bicicletas alugadas devido ao aumento da umidade e da velocidade do vento também foi verificada por Gebhart e Noland (2014) e por Kim (2018).

5 CONCLUSÃO

A demanda por aluguéis de bicicletas no sistema de compartilhamento BikeIndia, dos EUA, aumentou significativamente no ano de 2019 em relação a 2018, além de apresentar grande crescimento no número de bicicletas alugadas com o aumento da temperatura. Percebeu-se ainda que nos feriados houve grande redução nos aluguéis, o que também ocorreu com o aumento da umidade e da velocidade do vento.

A análise de regressão linear múltipla possibilitou a identificação dos principais fatores de influência na demanda por aluguéis de bicicletas, mostrando que 73,03% da variabilidade que ocorre no número de bicicletas alugadas pode ser explicada estatisticamente pela variabilidade nas variáveis ano, feriado, temperatura, umidade e velocidade do vento.

Diante disso, a análise de regressão pode ser vista com uma ferramenta útil para auxiliar os empresários do setor de compartilhamento de bicicletas a compreender melhor os fatores de influência para cada local, contribuindo para alavancar seus negócios além de favorecer o desenvolvimento sustentável.

REFERÊNCIAS

ALTMAN, N.; KRZYWINSKI, M. Analyzing outliers: influential or nuisance? **Nature Methods**, [*S. l.*], v. 13, n. 4, p. 281-283, 2016.

ANDERSON, D. R.; SWEENEY, D. J.; WILLIAMS, T. A. **Estatística aplicada à administração e economia.** 2. ed. São Paulo: Thomson Learning, 2007.

BITTENCOURT, R. C. **Análise dos fatores relevantes na escolha da localização de estações dos sistemas de compartilhamento de bicicletas.** 2020. Dissertação (Mestrado em Engenharia de Transportes). Programa de Pós-graduação em Engenharia de Transporte, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2020.

CARVALHO, M. L.; FREITAS, C. M. Pedalando em busca de alternativas saudáveis e sustentáveis. **Ciência & Saúde Coletiva**, [*S. l.*], v. 17, p. 1617 – 1628, 2012.

CASTRO A. E. et al. Algunas notas históricas sobre la correlación y regressión y su uso en el aula. **Números. Revista de Didáctica de las Matemáticas**, [*S. l.*], v. 81, p. 5 – 14, nov. 2012.

CHARNET, R. et al. **Análise de Modelos de Regressão Linear:** com aplicações. 2. ed. Campinas: Editora da Unicamp, 2008.

DE ARRUDA, H. R. et al. Consumo colaborativo e valores pessoais: o caso da bicicleta compartilhada. **Revista Brasileira de Marketing**, São Paulo, v. 15, n. 5, p. 683 - 698, out./dez. 2016.

FÁVERO, L. P. et al. **Análise de dados:** modelagem multivariada para tomada de decisões. Rio de Janeiro: Elsevier, 2009.

FELIPE, K. **Avaliação do desempenho do sistema de bicicletas compartilhadas de Brasília.** Monografia (Graduação em Engenharia Civil) - Faculdade de Tecnologia e Ciências Sociais Aplicadas, Centro Universitário de Brasília, Brasília, 2018.

FRADE, I.; RIBEIRO, A. Bicycle sharing systems demand. **Procedia - Social and Behavioral Sciences**, Portugal, v. 111, p. 518 – 527, 2014.

GEBHART, K.; NOLAND, R. B. The impact of weather conditions on bikeshare trips in Washington, DC. **Transportation**, [*S. l.*], v. 41, n. 6, p. 1205 – 1225, 06 ago. 2014. DOI: https://doi.org/10.1007/s11116-014-9540-7.

GROSS, J.; LIGGES, U. **nortest:** tests for normality. R package version 1.0-4, 2015. Disponível em: https://CRAN.R-project.org/package=nortest/. Acesso em: 29 nov. 2020.

GUJARATI, D. N.; PORTER, D. C. **Econometria básica**. 5. ed. Porto Alegre: Artimed, 2011.

HEBBALI, A. **olsrr:** Tools for Building OLS Regression Models. R package version 0.5.3, 2020. Disponível em: https://CRAN.R-project.org/package=olsrr/. Acesso em: 4 jan. 2023.

HEINEN, E.; VAN WEE, B.; MAAT, K. Commuting by bicycle: An overview of the literature. **Transport Reviews**, [*S. l.*], v. 30, n. 1, p. 59 – 96, 2010. DOI: https://doi.org/10.1080/01441640903187001.

HEYDARI, S.; KONSTANTINOUDIS, G.; BEHSOODI, A W. Efeito da pandemia de COVID-19 na demanda de compartilhamento de bicicletas e no tempo de aluguel: evidências do Santander Cycles em Londres. **PloS ONE**, [*S. l.*], v. 16, n.12, 2 dez. 2021. DOI: https://doi.org/10.1371/journal.pone.0260969.

HILL, R. C.; GRIFFITHS, W. E.; JUDGE, G. G. **Econometria**. 2. ed. São Paulo: Saraiva, 2003.

HOFFMANN, R. **Análise de Regressão:** Uma Introdução à Econometria. 5. ed. Piracicaba: HUCITEC, 2016.

HOTHORN, T. et al. **lmtest:** Testing Linear Regression Models. R package version 0.9-40, 2022. Disponível em: https://CRAN.R-project.org/package=lmtest/. Acesso em: 29 nov. 2022.

IBOLD, S. et al. **El brote de COVID-19 y las implicancias para la movilidad sostenible: algunas observaciones.** SUTP – Sustanable Urban Transport Project, [S. l.], 14 abr. 2020. Disponível em: https:// www.sutp.org/el-brote-de-covid-19-y-las-implicancias-para-la-movilidadsostenible-algunas-observaciones-2/. Acesso em: 14 dez. 2022.

ITDP - INSTITUTO DE POLÍTICAS DE TRANSPORTE E DESENVOLVIMENTO. **Guia de planejamento de sistemas de bicicletas compartilhadas.** Rio de Janeiro: ITDP, 2014.

ITDP - INSTITUTO DE POLÍTICAS DE TRANSPORTE E DESENVOLVIMENTO. **Guia de planejamento de sistemas de bicicletas compartilhadas.** lugar: ITDP, 2018.

KAGGLE. **Bike Sharing**. Hyderabad, Telangana, [2022]. Disponível em: https://www.kaggle.com/code/gauravduttakiit/bike-sharing-multiple-linear-regression. Acesso em: 24 jun. 2022.

KIM, D. et al. Factors influencing travel behaviors in bikesharing. **Transportation Research Board 91st Annual Meeting**, Washington DC, 2012.

KIM, K. Investigation on the effects of weather and calendar events on bikesharing according to the trip patterns of bike rentals of stations. **Journal of Transport Geography**, [*S. l.*], v. 66, p. 309 – 320, jan 2018. DOI: https://doi.org/10.1016/j.jtrangeo.2018.01.001.

LINCOLN, M. et al. **clipr:** Read and Write from the System Clipboard. R package version 0.8.0, 2022. Disponível em: https://CRAN.R-project.org/package=clipr/. Acesso em: 20 nov. 2022.

LOPES, T. F. **Comparando métodos de aprendizado de máquina para previsão da demanda de viagens de bicicletas da bixi montreal e análise do efeito da pandemia de covid-19 na demanda de 2020.** 2021. TCC (Pós-graduação Lato Sensu em Ciência de Dados e Machine Learning), Centro Universitário de Brasília, Instituto CEUB de Pesquisa e Desenvolvimento – ICPD, 2021.

MARTINEZ, M. The impact weather has on NYC citi bike share company activity. **Journal of Environmental and Resource Economics at Colby**, [*S. l.*], v. 4, n. 1, p. 12, 2017. Disponível em: https://digitalcommons.colby.edu/jerec/vol4/iss1/12. Acesso em: 06 jan. 2023.

MEMÓRIA, J. M. P. **Breve história da estatística.** 1. ed. Brasília: Embrapa Informações Tecnológicas, 2004

MIRANDA-MORENO, L.; NOSAL, T. Weather or not to cycle: Temporal trends and impact of weather on cycling in an urban environment. **Transportation Research Record: Journal of the Transportation Research Board**, [*S. l.*], v. 2247, n. 1 p. 42 – 52, 01 jan. 2011. DOI: https://doi.org/10.3141/2247-06.

MONTGOMERY, D. C.; RUNGER, G. C. **Estatística aplicada e probabilidade para engenheiros.** 2. ed. Rio de Janeiro: LTC, 2013.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. **Introduction to linear regression analysis**. Hoboken: John Wiley & Filhos, 2021.

R CORE TEAM. **R:** A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2022. Disponível em: https://www.R-project.org/. Acesso em: 10 nov. 2022.

REZENDE, A. A.; MARCELINO, J. A.; MIYAJI, M. A reinvenção das vendas: as estratégias das empresas brasileiras para gerar receitas na pandemia de covid-19. **Boletim de Conjuntura (BOCA)**, Boa Vista, v. 2, n. 6, 2020.

RUDLOFF, C.; LACKNER, B. Modeling demand for bikesharing systems: Neighboring stations as source for demand and reason for structural breaks. **Transportation Research Record: Journal of the Transportation Research Board**, [*S. l.*], v. 2430, n. 1, p. 1 – 11, 01 jan. 2014. DOI: https://doi.org/10.3141/2430-01.

SAMPAIO, N. Aplicações da correlação e regressão linear. **Associacao Educacional Dom Boco**, [*S. l.*], 2015.

SCHMIDT, A. F.; FINAN, C. Linear regression and the normality assumption. **Journal of Clinical Epidemiology**, [*S. l.*], 98 ed., p. 146 – 151, 2018. DOI: https://doi.org/10.1016/j.jclinepi.2017.12.006.

STANTON, J. M. Galton, Pearson, and the Peas: A Brief Historyof Linear Regression for Statistics Instructors. **Journal of Statistics Education**, [*S. l.*], v. 9, n. 3, 2001. DOI: https://doi.org/10.1080/10691898.2001.11910537.

TALAVERA-GARCIA, R.; ROMANILLOS, G.; ARIAS-MOLINARES, D.. Examining spatio-temporal mobility patterns of bike-sharing systems: the case of BiciMAD (Madrid). **Journal of Maps**, [*S. l.*], v. 17, n. 1, p. 7 - 13, 2021. DOI: https://doi.org/10.1080/17445647.2020.1866697.

WEI, T.; SIMKO, V. **corrplot:** Visualization of a Correlation Matrix. R package version 0.92, 2021. Disponível em: https://CRAN.R-project.org/package=corrplot/. Acesso em: 19 dez. 2022.

APÊNDICE A – SCRIPT

```
# Carregando o banco de dados

bike=read.table("dayof1.txt", header=T)
attach(bike)
head(bike)

# O arquivo acima encontra-se disponível através do link:
#https://www.kaggle.com/code/gauravduttakiit/bike-sharing-multiple-linear-regression

# Análise Descritiva

# Médias aritmética, mínimos e máximos.

summary(bike)



# Coeficiente de variação

cv = function(x)
{
  coef = sd(x)/mean(x)*100
  return(coef)
}

cv(temp)
cv(senter)
cv(umi)
cv(vento)
cv(cba)



# Desvio padrão
```

```
sd(temp)
sd(senter)
sd(umi)
sd(vento)
sd(cba)


# Gráficos

# Gráficos de Barras

# Instalar pacote ggplot2

library("ggplot2")

bikegraf=read.table("TABELA2.txt", header=T)
attach(bikegraf)

# O arquivo acima corresponde a uma modificação do arquivo original da
# seguinte maneira:
# 1) substituiu-se os valores das categorias das variáveis estação, ano, mês
# e dia da semana por seus respectivos nomes
# 2) nas variáveis feriado e dia útil trocou-se "1" por "sim" e "0" por "não"


cba1=prop.table(cba)
head(cba1)


# Ano

anos1<-factor(bikegraf$ano, levels = c("2018","2019"))
```

```
(a=ggplot(bikegraf, aes(y = cba1, x = anos1)) +
   geom_bar(fill = "darkblue",stat = "identity")+
   xlab("Ano")+
   ylab("Percentual de bicicletas")+
   ggtitle("                    Percentual de bicicletas por ano")+
   scale_y_continuous(labels = scales::percent))

anos= read.table("ano.txt", header=T)
(prop.table(anos))*100

(b=a + annotate(geom="text", x=1, y=0.4, label="37.78%",
            color="black"))
(c=b + annotate(geom="text", x=2, y=0.65, label="62.22%",
            color="black"))



# Mês

meses1<-factor(bikegraf$mes, levels = c("JAN","FEV","MAR","ABR","MAI",
                                        "JUN","JUL","AGO","SET","OUT",
                                        "NOV","DEZ"))

(a=ggplot(bikegraf, aes(y = cba1, x = meses1)) +
   geom_bar(fill = "darkblue" ,stat = "identity")+
   xlab("Meses")+
   ylab("Percentual de bicicletas")+
   ggtitle("                    Percentual de bicicletas por mês")+
   scale_y_continuous(labels = scales::percent))

meses= read.table("meses.txt", header=T)
(prop.table(meses))*100

(b=a + annotate(geom="text", x=1, y=0.045, label="4.1%",
            color="black"))
```

```
(c=b + annotate(geom="text", x=2, y=0.05, label="4.54%",
                color="black"))
(d=c + annotate(geom="text", x=3, y=0.074, label="6.94%",
                color="black"))
(e=d + annotate(geom="text", x=4, y=0.086, label="8.18%",
                color="black"))
(f=e + annotate(geom="text", x=5, y=0.105, label="10.08%",
                color="black"))
(g=f + annotate(geom="text", x=6, y=0.109, label="10.53%",
                color="black"))
(h=g + annotate(geom="text", x=7, y=0.1082, label="10.48%",
                color="black"))
(i=h + annotate(geom="text", x=8, y=0.1087, label="10.67%",
                color="black"))
(j=i + annotate(geom="text", x=9, y=0.1086, label="10.52%",
                color="black"))
(k=j + annotate(geom="text", x=10, y=0.102, label="9.8%",
                color="black"))
(l=k + annotate(geom="text", x=11, y=0.082, label="7.75%",
                color="black"))
(m=l + annotate(geom="text", x=12, y=0.07, label="6.41%",
                color="black"))


# Dia da semana

diasdasemana<-factor(bikegraf$dia, levels = c("DOM", "SEG",
                                    "TER","QUA","QUI","SEX","SAB"))

(a=ggplot(bikegraf, aes(y = cba1, x = diasdasemana)) +
   geom_bar(fill = "darkblue",stat = "identity")+
   xlab("Dias da semana")+
   ylab("Percentual de bicicletas")+
   ggtitle("              Percentual de bicicletas por dias da semana")+
```

```
   scale_y_continuous(labels = scales::percent))

dias= read.table("dias.txt", header=T)
(prop.table(dias))*100

(b=a + annotate(geom="text", x=1, y=0.153, label="14.74%",
            color="black"))
(c=b + annotate(geom="text", x=2, y=0.142, label="13.59%",
            color="black"))
(d=c + annotate(geom="text", x=3, y=0.143, label="13.64%",
            color="black"))
(e=d + annotate(geom="text", x=4, y=0.146, label="14.05%",
            color="black"))
(f=e + annotate(geom="text", x=5, y=0.153, label="14.83%",
            color="black"))
(g=f + annotate(geom="text", x=6, y=0.15, label="14.53%",
            color="black"))
(h=g + annotate(geom="text", x=7, y=0.152, label="14.62%",
            color="black"))

# Estação

estacoes1<-factor(bikegraf$estacao,levels=c("Verao","Primavera","Outono","Inverno"))

(a=ggplot(bikegraf, aes(y = cba1, x = estacoes1)) +
   geom_bar(fill = "darkblue", stat = "identity")+
   xlab("Estações")+
   ylab("Percentual de bicicletas")+
   ggtitle("                Percentual de bicicletas por estações")+
   scale_y_continuous(labels = scales::percent))

estacoes= read.table("estacao.txt", header=T)
(prop.table(estacoes))*100
```

```
(b=a + annotate(geom="text", x=1, y=0.335, label="32.25%",
                color="black"))
(c=b + annotate(geom="text", x=2, y=0.29, label="27.91%",
                color="black"))
(d=c + annotate(geom="text", x=3, y=0.268, label="25.57%",
                color="black"))
(e=d + annotate(geom="text", x=4, y=0.155, label="14.27%",
                color="black"))


# Boxplots

# Dia útil

ggplot(bikegraf, aes(x = diau, y =cba)) +
  geom_boxplot(fill = "dodgerblue") +
  labs(y = "Quantidade de bicicletas",
       x = "Dia útil",
       title = "   Quantidade de bicicletas alugadas em dias útéis e não úteis")

# Feriado

ggplot(bikegraf, aes(x = feriado, y =cba)) +
  geom_boxplot(fill = "dodgerblue") +
  labs(y = "Quantidade de bicicletas",
       x = "Feriado",
       title = "  Quantidade de bicicletas alugadas em feriados e não feriados")


# Situação climática

sitcli<-factor(bikegraf$clima, levels = c("1","2","3"))

ggplot(bikegraf, aes(x = sitcli, y =cba)) +
```

```
  geom_boxplot(fill = "dodgerblue") +
  labs(y = "Quantidade de bicicletas",
       x = "Situação climática",
       title = "  Quantidade de bicicletas alugadas em cada situação climática")


# Aleatorizando os dados

linhas <- sample(1:length(bike$cba),length(bike$cba)*1)

bikeal = bike[linhas,]

# Salvando os dados aleatorizados

# Instalar o pacote clipr

library(clipr)

write_clip(bikeal)

# Após rodar essa função os dados podem ser copiados, utilizando a função
# Ctrl+V para uma planilha externa ao R.


# Carregando os dados aleatorizados

bikeal=read.table("dadosaleatorizados2012SO.txt", header=T)
attach(bikeal)


# Regressão Múltipla

# Ajuste do modelo com todas as variáveis
```

```
lm(formula = cba ~ estacao1 + estacao2 + estacao3 + ano + mes1 + mes2 +
    mes3 + mes4 + mes5 + mes6 + mes7 + mes8 + mes9 + mes10 + mes11 +
    feriado + dia1 + dia2 + dia3 + dia4 + dia5 + dia6 + diau + clima1 +
    clima2 + temp + senter + umi + vento,
  data = bikeal)

modc <- lm(formula = cba ~ estacao1 + estacao2 + estacao3 + ano + mes1 + mes2 +
           mes3 + mes4 + mes5 + mes6 + mes7 + mes8 + mes9 + mes10 + mes11 +
           feriado + dia1 + dia2 + dia3 + dia4 + dia5 + dia6 + diau + clima1 +
           clima2 + temp + senter + umi + vento,
         data = bikeal)

summary(modc)



# Gráfico de correlação

# Instalar pacote corrplot

library(corrplot)

M <- cor(bikeal)
corrplot(M, method = "circle")



# Ajuste do modelo sem variáveis altamente correlacionadas
# (estacao1, estacao2, estacao3, cliam1, clima2 e senter)

modsem <- lm(formula = cba ~ ano + mes1 + mes2 + mes3 + mes4 + mes5 + mes6 +
             mes7 + mes8 + mes9 + mes10 + mes11 + feriado + dia1 + dia2 +
             dia3 + dia4 + dia5 + dia6 + diau + temp + umi + vento,
           data = bikeal)
```

```
# Seleção de variáveis pelo método stepwise

modstep <- step(modsem)
summary(modstep)

lm(formula = cba ~ ano + mes1 + mes2 + mes4 + mes5 + mes7 + mes9 + mes10 +
    mes11 + feriado + dia2 + dia3 + temp + umi + vento,
  data = bikeal)


# Verificação do FIV

car::vif(modstep)


# Novo Modelo (com a exclusão das variáveis meses e dias da semana)

mod1 <- lm(formula = cba ~ ano + feriado + temp + umi + vento,
          data = bikeal)

summary(mod1)


# Análise de resíduos

# Gráfico da Distância de Cook

# Instalar pacote olsrr

library(olsrr)

ols_plot_cooksd_bar(mod1)
```

```
# Verificação do modelo sem pontos influentes

# Carregando o banco de dados

bike2=read.table("dadosS7PI.txt", header=T)
attach(bike2)


# Ajuste do modelo sem pontos influentes

mod2 <- lm(formula = cba ~ ano + feriado + temp + umi + vento,
           data = bike2)

summary(mod2)

# Como retirada dos pontos praticamente não alterou o modelo, decidiu-se
# mantê-los e prosseguir com a análise de resíduos


# Verificação do FIV

car::vif(mod1)


# Gráfico dos resíduos

ols_plot_resid_stud_fit(mod1)


# Testes de hipóteses

# Normalidade (Kolmogorov-Smirnov)

# Instalar pacote nortest
```

```
library("nortest")

mod1res=scale(mod1$residuals)
lillie.test(mod1res)


# Independencia dos residuos (Durbin-Watson)

# Instalar pacote car

library(car)

durbinWatsonTest(mod1)


# Homocedasticidade (Breusch-Pagan)

# Instalar pacote pacman

library(pacman)
pacman::p_load(dplyr, car, rstatix, lmtest, ggpubr,
               QuantPsyc, psych, scatterplot3d)

bptest(mod1)


# Devido ao resultado do teste de homogeneidade utilizou-se a correção de
# heterocedasticidade de White

# Correção de White

vcov.white0 <- hccm(mod1, type = c("hc1"))
```

```
coeftest(mod1, vcov.white0)
summary(mod1)

# Como os resultados se mantiveram praticamente os mesmos, optou-se pela
# utilização das estimativas de mínimos quadrados sem a correção de White


# Qualidade do ajuste

# Análise de Variância (anova)

simpleAnova <- function(object, ...) {

  # Compute anova table
  tab <- anova(object, ...)

  # Obtain number of predictors
  p <- nrow(tab) - 1

  # Add predictors row
  predictorsRow <- colSums(tab[1:p, 1:2])
  predictorsRow <- c(predictorsRow, predictorsRow[2] / predictorsRow[1])

  # F-quantities
  Fval <- predictorsRow[3] / tab[p + 1, 3]
  pval <- pf(Fval, df1 = p, df2 = tab$Df[p + 1], lower.tail = FALSE)
  predictorsRow <- c(predictorsRow, Fval, pval)

  # Simplified table
  tab <- rbind(predictorsRow, tab[p + 1, ])
  row.names(tab)[1] <- "Predictors"
  return(tab)
}
```

```
simpleAnova(mod1)


# Estimativas de mínimos quadrados
summary(mod1)


# R ao quadrado ajustado
summary(mod1)$adj.r.squared
```