

PROJETO FINAL - APRENDIZAGEM SUPERVISIONADA

Bike Sharing Dataset(Regressão Linear Múltipla)

Pedro Henrique Mello Pereira - 230353025

Bernardo Miguel Esperança Sousa - 230353006

Table of contents

Introdução:	3
Definição dos objetivos:	3
Apresentação do conjunto de dados e identificação das variáveis dependente e independentes:	3
Visão das 5 primeiras entradas do Data set bike sharing:	5
Metadados do data set objeto do presente estudo:	5
Limpeza dos dados:	6
Análise Exploratória dos Dados	8
Explorando a distribuição das variáveis independentes (numéricas):	9
Explorando a relação entre o número de bikes alugadas e a estação do ano: . .	11
Explorando a relação entre o número de bikes alugadas e o ano:	12
E em relação aos meses?	13
Explorando a relação entre o número de bikes alugadas e o dia da semana: . . .	14
E como comportam-se os números de alugueis de bicicletas em relação à condição climática?	15
Desenvolvimento do Modelo De Regressão Linear Múltipla	16
Pressupostos para a validação de um modelo de Regressão Linear Múltipla:	17
Desenvolvendo e avaliando o Modelo 1 de Regressão Linear Múltipla:	18
Avaliação geral do desempenho do Modelo 1:	19
Problemas identificados no modelo	24
Modelo 2 - Novo modelo com menos variáveis	24
Modelo 3: Excluindo “mnth” do conjunto de variáveis independentes:	26

Extra - Avaliação da performance de uma árvore de decisão para a mesma tarefa de	
Regressão.	29
Aplicação da validação cruzada e avaliação dos resultados:	32
Conclusão	34

Introdução:

Este trabalho visa desenvolver um modelo de regressão linear múltipla para prever o número de alugueis de bicicleta por dia com base em condições ambientais e sazonais. A análise será realizada utilizando o dataset “bikesharing” e tendo como consideração sua sub-divisão “day”, não tendo sido feito portanto com base no dataset “hour”. Neste estudo, pretendemos explorar como diferentes variáveis, como estação do ano, ano, mês, feriado, dia da semana e condições meteorológicas, influenciam o número de alugueis de bicicleta.

Além disso, é importante destacar que este trabalho será fundamentado na análise estatística do modelo de regressão linear múltipla. Através desta abordagem, pretendemos identificar quais variáveis independentes têm uma influência significativa no número de alugueis de bicicleta, bem como avaliar a força e a direção dessas relações. Utilizaremos técnicas estatísticas para ajustar o modelo aos dados, testar sua adequação e interpretar os resultados. Ao compreendermos melhor como as variáveis ambientais e sazonais impactam a demanda por alugueis de bicicleta, poderemos fornecer insights valiosos para empresas e organizações envolvidas no compartilhamento de bicicletas.

Definição dos objetivos:

O principal objetivo deste trabalho é desenvolver um modelo de regressão linear múltipla que seja capaz de prever o número de alugueis de bicicleta por dia com base nas variáveis ambientais e sazonais fornecidas no conjunto de dados. Pretende-se entender a relação de fatores como estação do ano, condições meteorológicas, feriados e dia da semana com a demanda por alugueis de bicicleta.

Assim sendo, a variável que queremos prever é a “cnt” (contagem total de bicicletas alugadas, incluindo tanto as casuais quanto as registradas), que neste estudo será chamada de variável dependente e por vezes também pode assumir as nomenclaturas de variável resposta ou variável-alvo (target) e que pode ser compreendida como “a contagem total de bicicletas alugadas por dia, incluindo utilizadores casuais e registados”;

Apresentação do conjunto de dados e identificação das variáveis dependente e independentes:

O dataset utilizado no presente estudo contém a contagem diária de bicicletas alugadas entre os anos de 2011 e 2012 no sistema de compartilhamento de bicicletas Capital Bikeshare, juntamente com as informações meteorológicas e sazonais correspondentes.

A seguir apresentamos o dicionário das variáveis presentes no conjunto de dados, as quais utilizaremos como variáveis independentes/preditores para prever o valor de “cnt”, a saber:

- **instant:** Índice do registo.

- **dteday**: Data.
- **season**: Estação (1: inverno, 2: primavera, 3: verão, 4: outono).
- **yr**: Ano (0: 2011, 1: 2012).
- **mnth**: Mês (1 a 12).
- **hr**: Hora (0 a 23).
- **holiday**: Dia de tempo é feriado ou não (extraído de <http://dchr.dc.gov/page/holiday-schedule>).
- **weekday**: Dia da semana.
- **workingday**: Se o dia não é fim de semana nem feriado, é 1, caso contrário é 0.
- **weathersit**:
 - 1: Limpo, Poucas nuvens, Parcialmente nublado, Parcialmente nublado.
 - 2: Nevoeiro + Nublado, Nevoeiro + Nuvens quebradas, Nevoeiro + Poucas nuvens, Nevoeiro.
 - 3: Neve fraca, Chuva fraca + Trovoada + Nuvens dispersas, Chuva fraca + Nuvens dispersas.
 - 4: Chuva forte + Granizo + Trovoada + Nevoeiro, Neve + Nevoeiro.
- **temp**: Temperatura normalizada em Celsius. Os valores são derivados via $(t - t_{\min}) / (t_{\max} - t_{\min})$, $t_{\min} = -8$, $t_{\max} = +39$ (apenas em escala horária).
- **atemp**: Sensação térmica normalizada em Celsius. Os valores são derivados via $(t - t_{\min}) / (t_{\max} - t_{\min})$, $t_{\min} = -16$, $t_{\max} = +50$ (apenas em escala horária).
- **hum**: Humidade normalizada. Os valores são divididos por 100 (máximo).
- **windspeed**: Velocidade do vento normalizada. Os valores são divididos por 67 (máximo).
- **casual**: Contagem de utilizadores casuais.
- **registered**: Contagem de utilizadores registados.
- **cnt**: Contagem total de bicicletas alugadas, incluindo utilizadores casuais e registados.

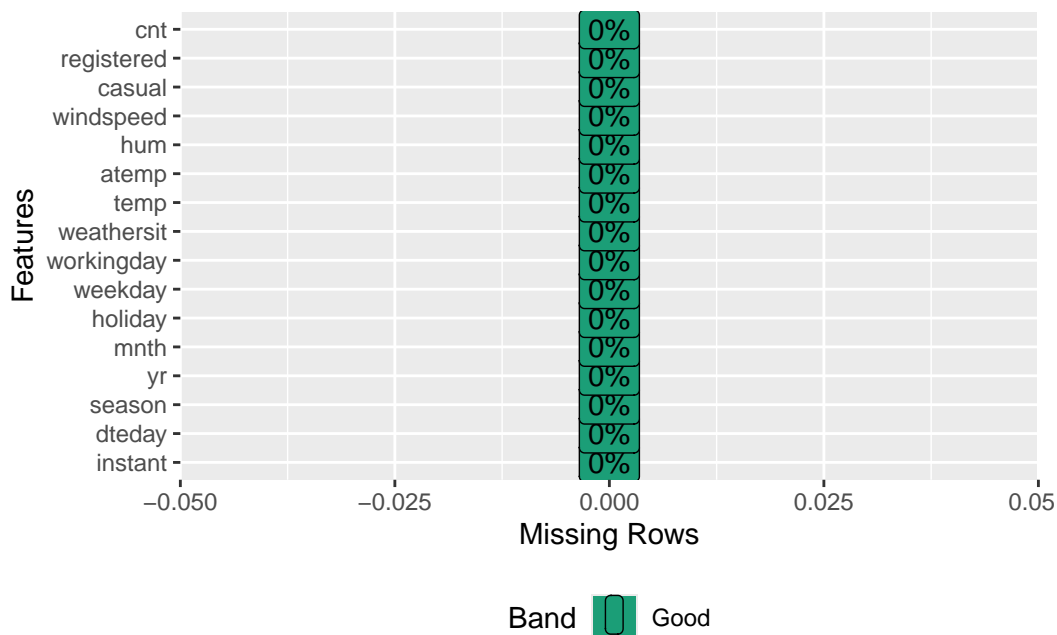
Visão das 5 primeiras entradas do Data set bike sharing:

instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
1	2011-01-01	1	0	1	0	6	0	2	0.344167	0.363625	0.805833	0.1604460	331	654	985
2	2011-01-02	1	0	1	0	0	0	2	0.363478	0.353739	0.696087	0.2485390	131	670	801
3	2011-01-03	1	0	1	0	1	1	1	0.196364	0.189405	0.437275	0.2485090	120	1229	1349
4	2011-01-04	1	0	1	0	2	1	1	0.200000	0.212122	0.596435	0.1602960	108	1454	1562
5	2011-01-05	1	0	1	0	3	1	1	0.226957	0.229270	0.436957	0.1869000	82	1518	1600

Metadados do data set objeto do presente estudo:

Usando a função “introduce” do pacote DataExplorer, teremos um inimportante panorama geral sobre o conjunto de dados bike sharing:

rows	731
columns	16
discrete_columns	1
continuous_columns	15
all_missing_columns	0
total_missing_values	0
complete_rows	731
total_observations	11696
memory_usage	112208



Como podemos perceber, trata-se de um data set de dimensões relativamente pequenas, a saber 731 linhas (observações) e 16 colunas.

Um dado importante trazido no gráfico de missing values é que não existe nenhuma entrada com valores ausentes, o que constitui um importante indicador de qualidade em relação ao preenchimento dos dados.

Na próxima seção nos ocuparemos em limpar os dados, sobretudo para remover colunas que não fazem sentido e tornar tudo mais claro para a etapa de análise exploratória de dados e posterior desenvolvimento dos modelos de regressão linear.

Limpeza dos dados:

Em primeiro lugar, removemos as colunas `casual` e `registered`, vez que tal condição fora mencionada como obrigatória no guião do projeto. Desta feita, excluimos as colunas ora mencionadas pois não serão objeto deste estudo.

Em seguida, é digno de atenção que cada linha do dataframe em questão diz respeito a um dia do ano, indo desde 01-01-2011 a 31-12-2012. Ademais, a variável `'instant'` servia como índice.

Logo, com vistas a melhor organizar o data set e diminuir o tempo de processamento para tarefas de regressão, decidimos por excluir a coluna `'instant'` e utilizar a coluna `'dteday'` como índice.

Assim está a nova configuração dos dados:

rows	731
columns	13
discrete_columns	1
continuous_columns	12
all_missing_columns	0
total_missing_values	0
complete_rows	731
total_observations	9503
memory_usage	102928

Busca por outliers:

Agora que já excluimos da análise as colunas desnecessárias e checamos a não existência de valores nulos, é importante atentarmos para a existência de outliers, que são valores que se afastam significativamente da maioria dos outros valores num conjunto de dados. Eles podem distorcer análises estatísticas e modelos, influenciando os resultados.

Além disso, uma das maiores desvantagens da utilização de modelos de regressão linear é a sua alta sensibilidade a outliers. Isto posto, identificar e lidar com outliers é de suma importância para garantir a precisão do modelo a ser desenvolvido.

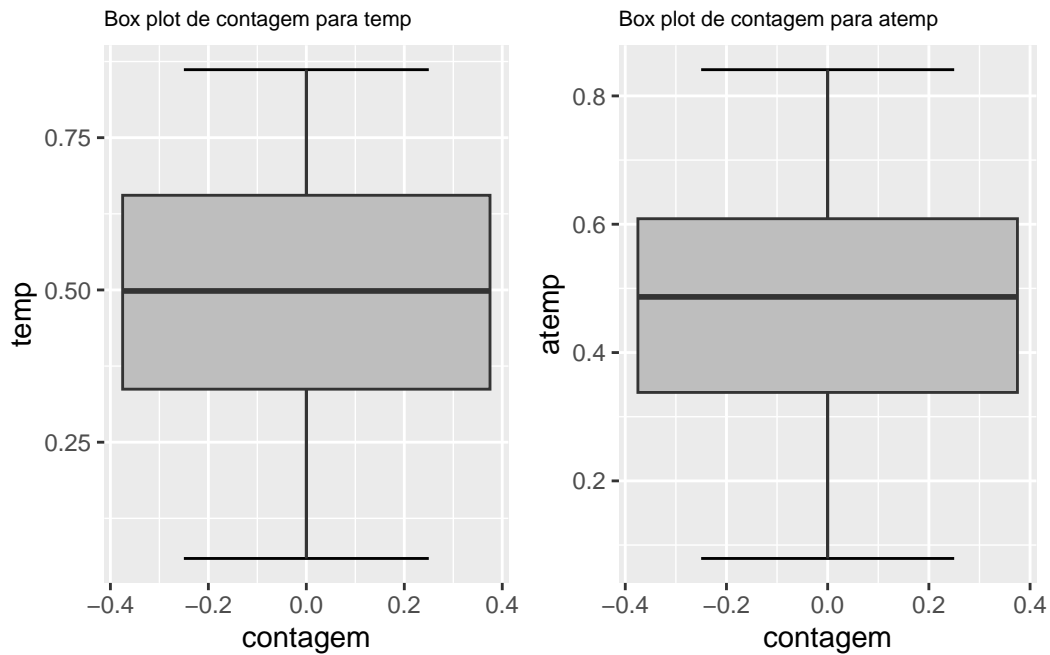


Figure 1: Distribuições com Outliers

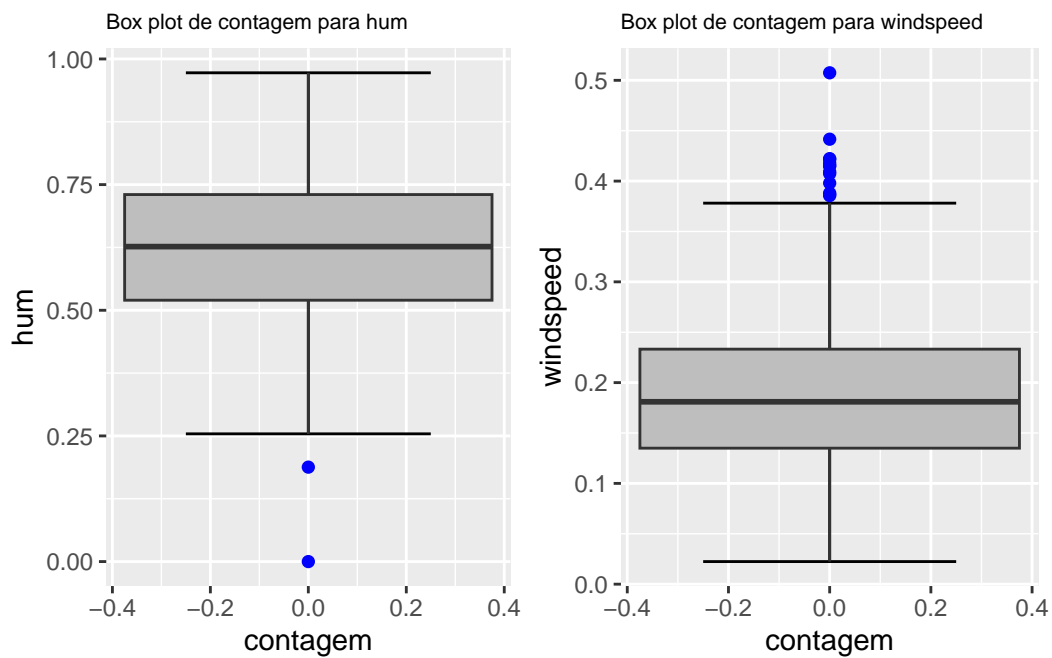


Figure 2: Distribuições com Outliers

Conforme depreende-se da análise dos boxplots, as variáveis ‘hum’ e ‘windspeed’ apresentaram outliers em sua composição. Para lidar com este problema, haja vista serem poucos os valores classificados como outliers, excluiríamos tais valores discrepantes do data set `df_bike_new`.

Exclusão dos outliers por meio da definição dos limites superiores e inferiores pelo IQR SCORE:

A exclusão dos outliers será feita por meio do IQR Score. O IQR (Intervalo Interquartil) é uma medida de dispersão que indica a amplitude dos dados em torno da mediana. O IQR score é uma medida estatística usada para identificar outliers, calculada como a diferença entre o terceiro quartil (Q3) e o primeiro quartil (Q1). Outliers geralmente são definidos como valores que caem abaixo de $Q1 - 1,5 * IQR$ (Limite inferior) ou acima de $Q3 + 1,5 * IQR$ (Limite superior).

Após realizarmos a exclusão dos outliers, apenas 18 observações foram apagadas do conjunto de dados. Apesar de ser interessante contar com o máximo de dados possíveis, é ainda mais importante para o nosso modelo a não existência de outliers.

Utilizamos também o parâmetro `distinct` do R para eliminar duplicatas do nosso conjunto de dados. Entretanto, nenhuma duplicata foi identificada.

Após a finalização da limpeza dos dados, vamos salvá-los em um novo ficheiro o qual chamaremos de “`df_clean`” e que será utilizado de agora em diante na análise exploratória e modelagem.

Análise Exploratória dos Dados

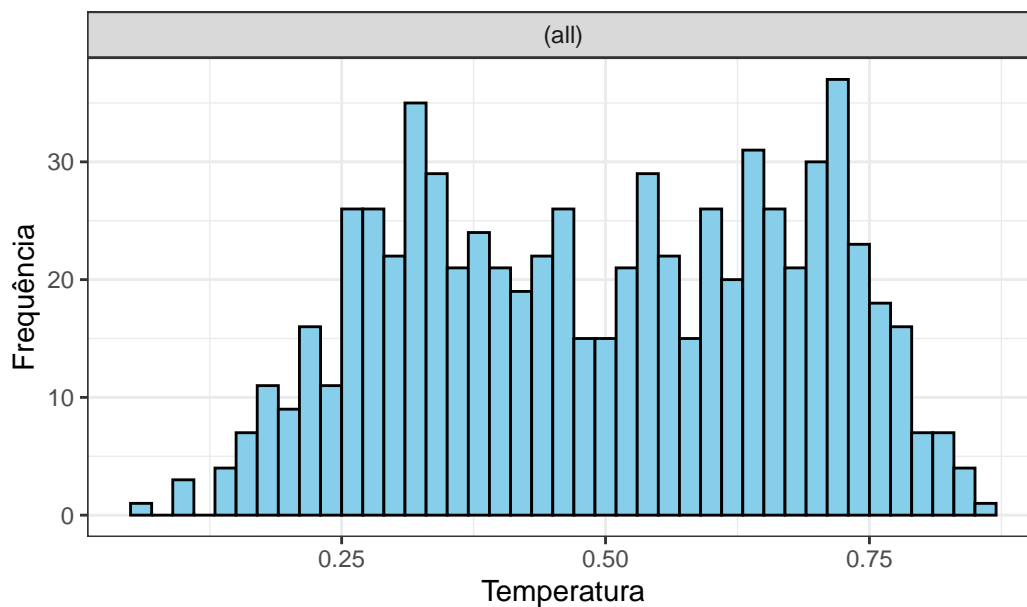
Ao implementar as etapas de limpeza anteriores, criamos um conjunto de dados mais refinado e confiável, estabelecendo a base para nossa subsequente análise exploratória de dados (AED), bem como para o desenvolvimento de modelos preditivos. O tratamento cuidadoso dos problemas de qualidade dos dados é crucial para garantir a precisão e confiabilidade dos insights que serão derivados do conjunto de dados.

Prosseguindo, nosso foco irá mudar para a Análise Exploratória de Dados (AED) para obter insights mais profundos sobre os dados de compartilhamento de bicicletas e identificar padrões acionáveis. Esta fase analítica tem como objetivo descobrir tendências significativas, correlações e padrões dentro do conjunto de dados.

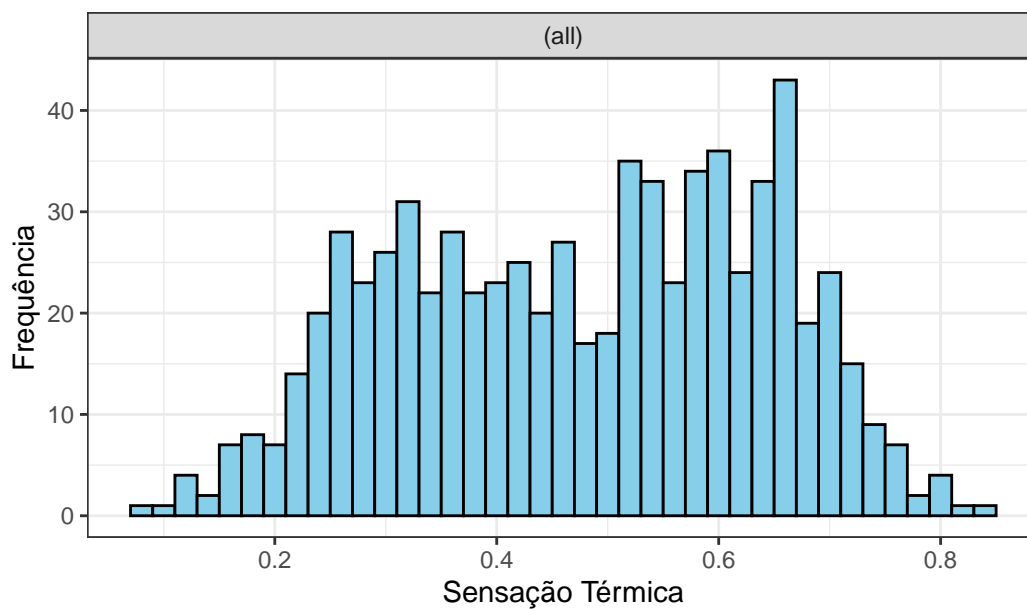
A primeira medida será transformar em fator as variáveis categóricas, utilizando para tanto a função `factor` do R.

Explorando a distribuição das variáveis independentes (numéricas):

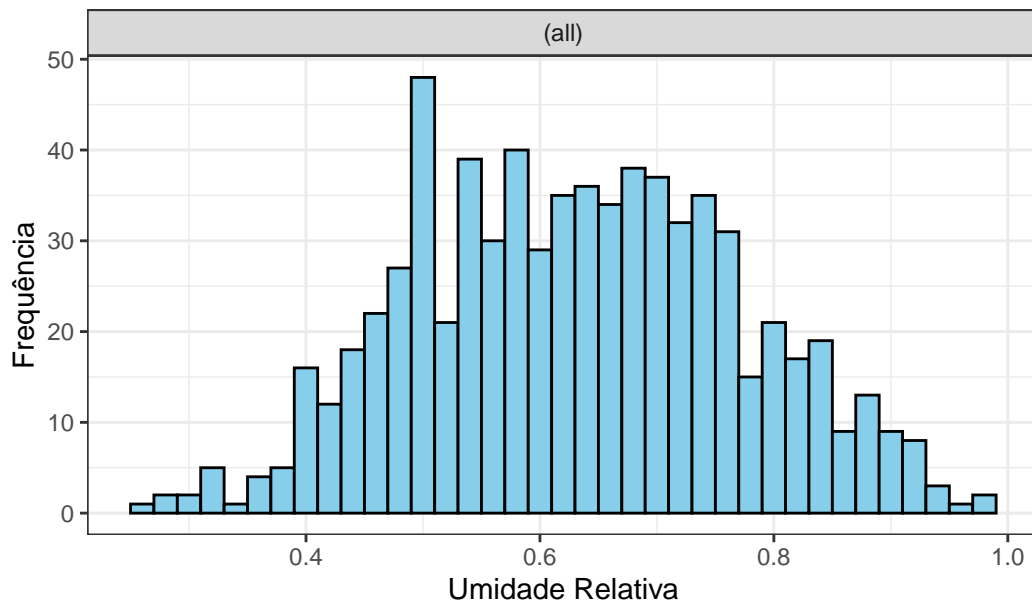
Distribuição de Temperatura



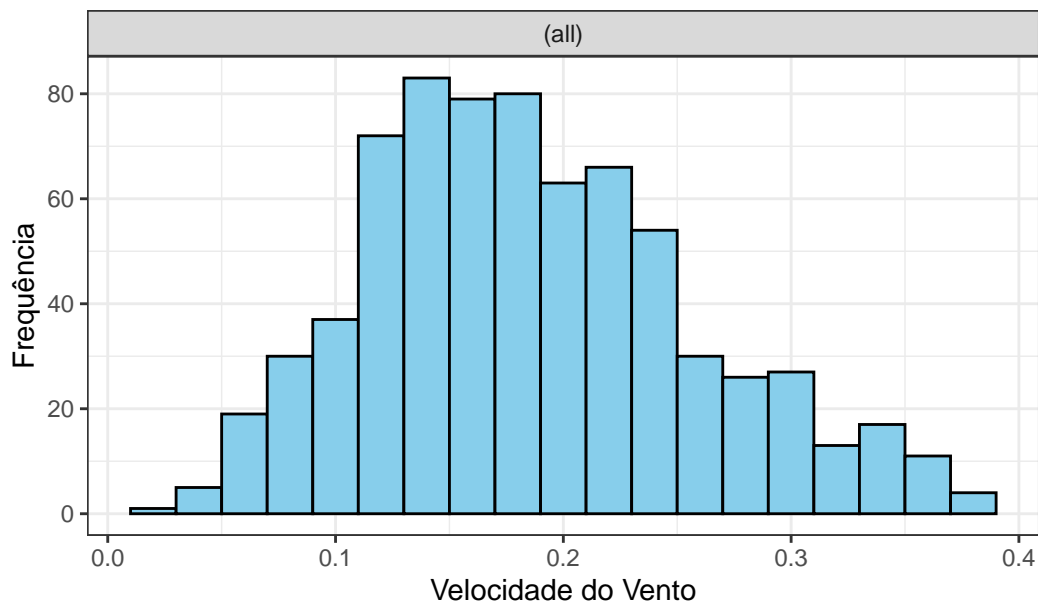
Distribuição de Sensação Térmica



Distribuição de Umidade Relativa



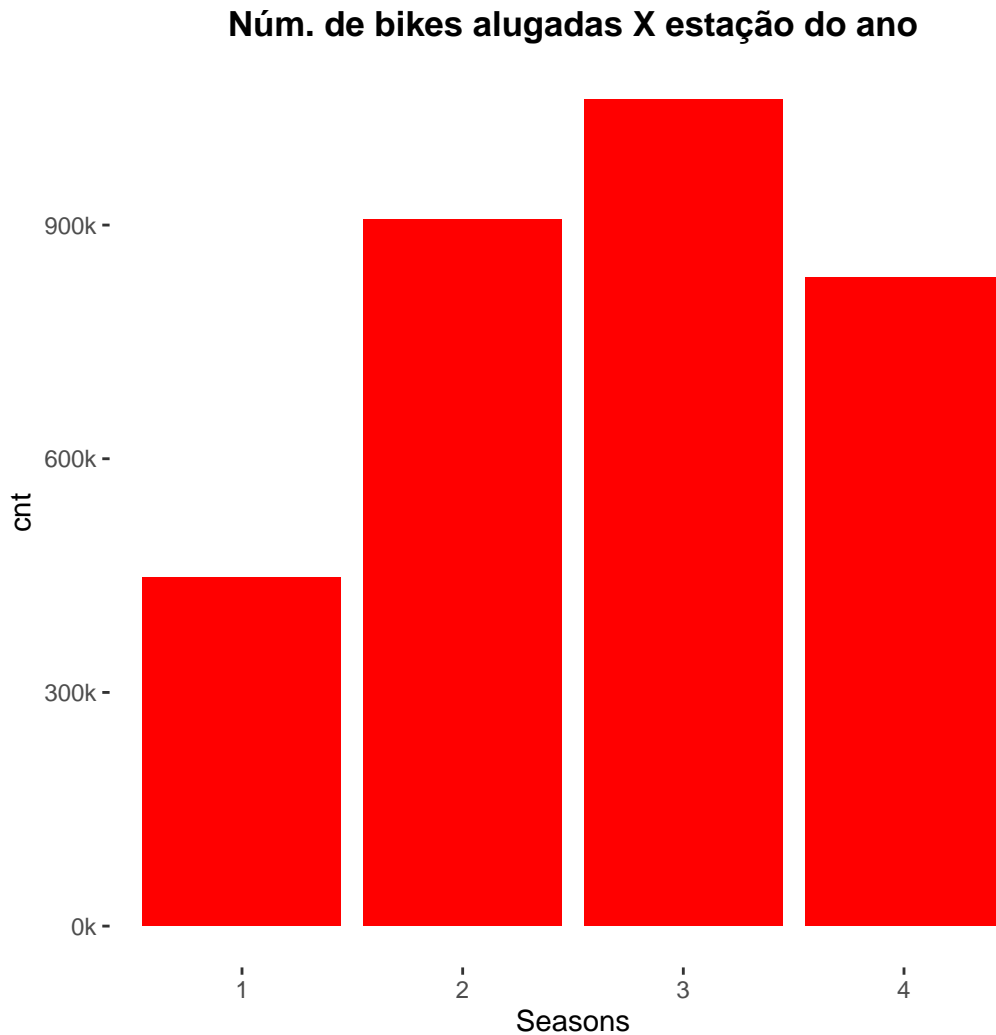
Distribuição de Velocidade do Vento



- Quanto à variável que diz respeito à umidade relativa, a distribuição aproxima-se da normalidade.

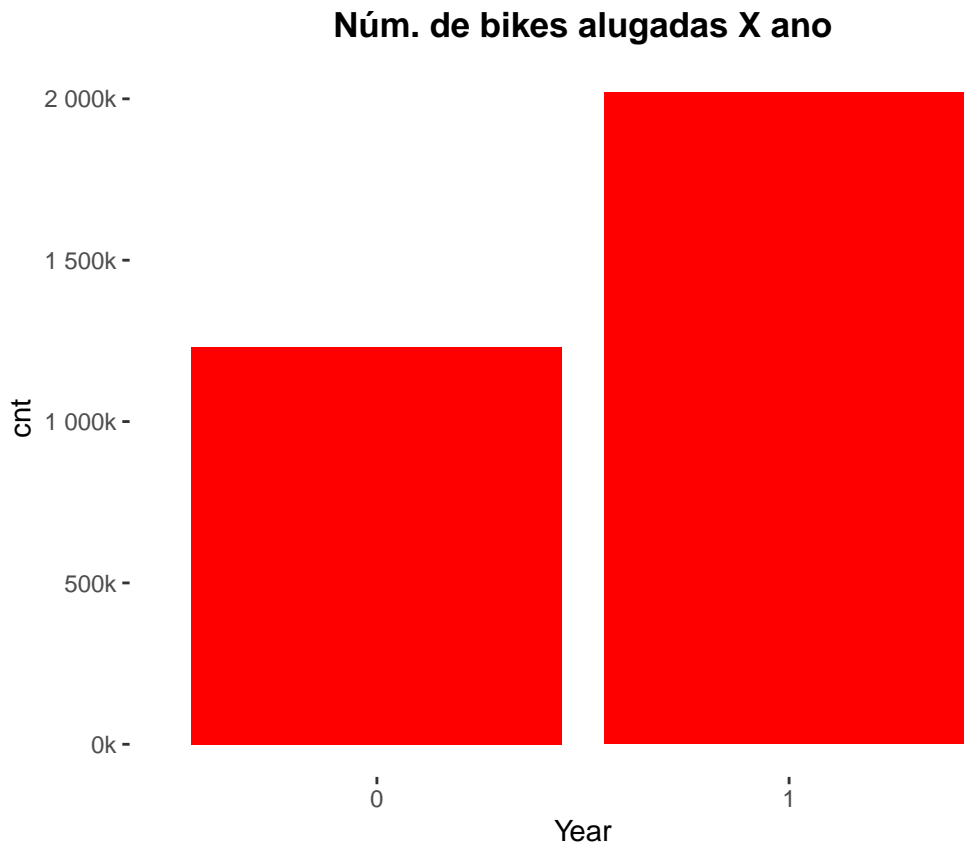
- Já em relação à distribuição das variáveis “sensação térmica” e “temperatura”, podemos dizer que a análise dos gráficos de distribuição nos leva a concluir que pode existir uma distribuição bimodal em ambos os casos.
- A variável relativa à velocidade do vento, por sua vez, traz uma assimetria positiva em sua distribuição.

Explorando a relação entre o número de bikes alugadas e a estação do ano:



Aqui podemos perceber que há uma clara tendência pela maior procura de aluguel de bicicletas na primavera e verão, tendo como contraponto o inverno, estação em que os alugueis reduzem-se a um terço do observado no verão.

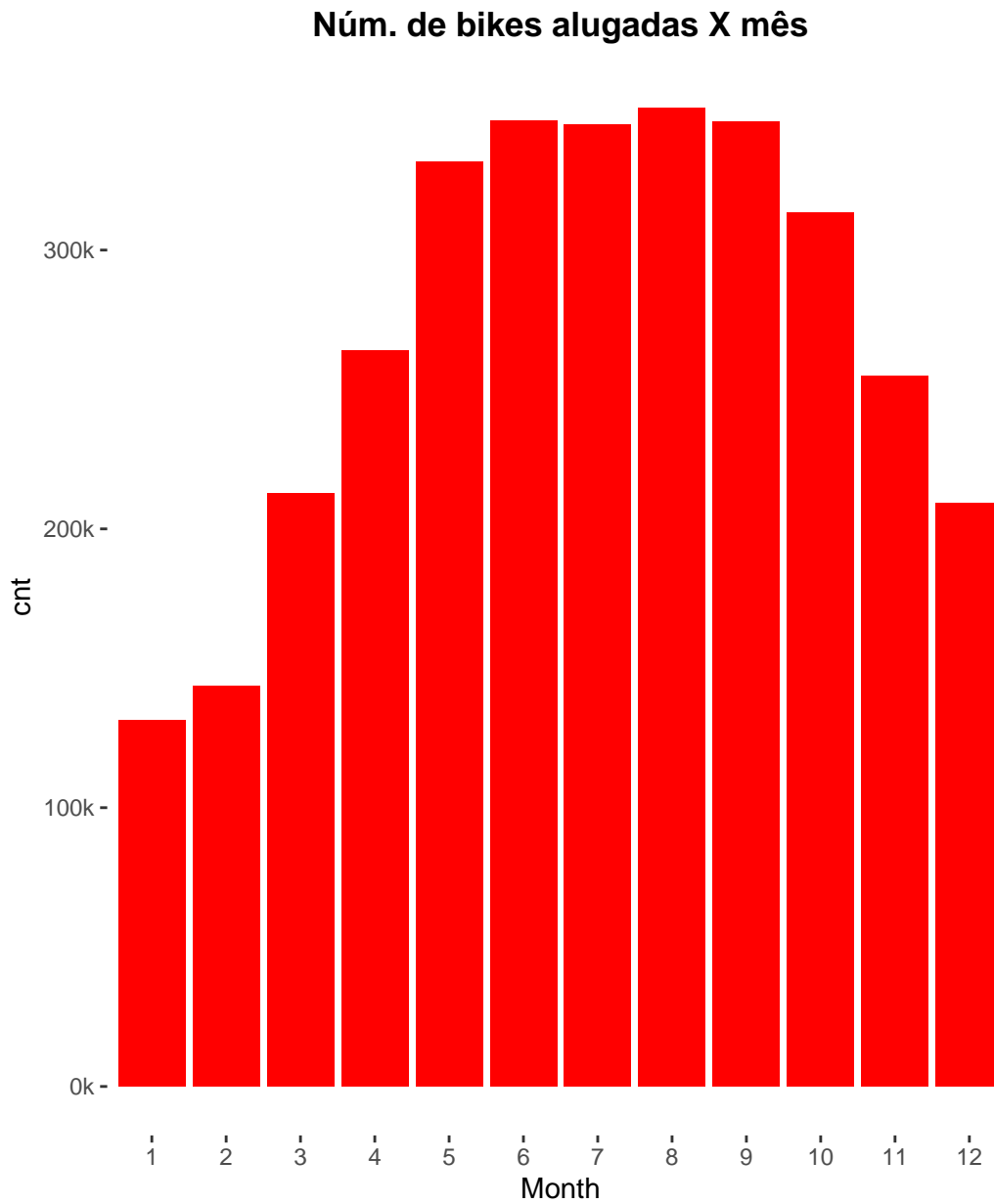
Explorando a relação entre o número de bikes alugadas e o ano:



Conforme depreende-se da análise do gráfico, houve um aumento substancial no número de bicicletas no ano de 2012, se comparado ao ano anterior.

Isto pode significar uma tendência de crescimento para este mercado. Para confirmar tal asunção seria necessário ter os dados dos anos posteriores e proceder à análise de séries temporais para saber se a tendência se confirma.

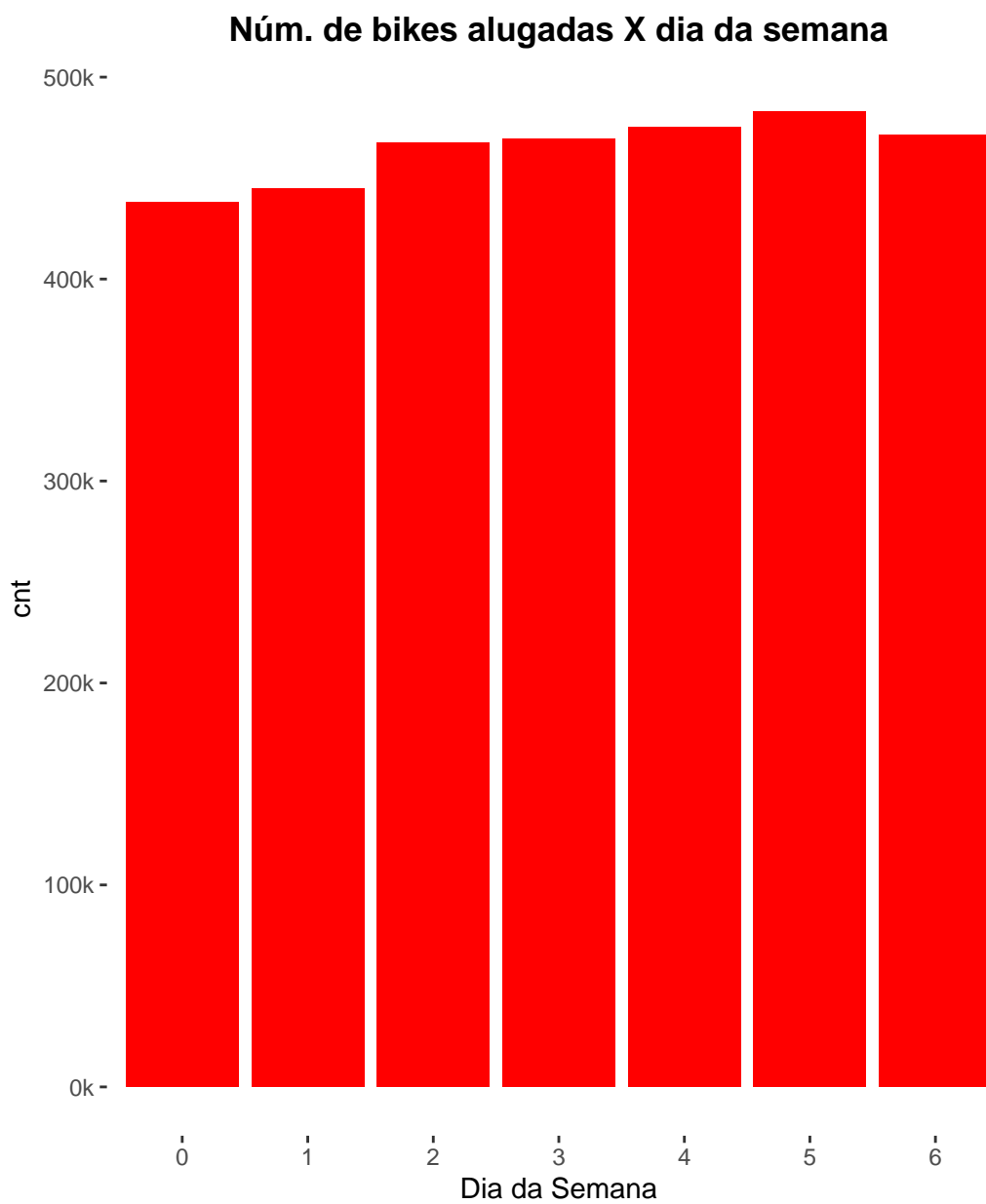
E em relação aos meses?



Tendo em consideração a distribuição das estações climáticas para os países do hemisfério norte, podemos depreender da análise do gráfico acima uma confirmação da constatação relativa ao primeiro ponto desta análise exploratória, onde observamos uma tendência muito maior ao aluguel de bicicletas durante a primavera e o verão.

Aqui a tendência se confirma, demonstrando que nos meses mais quentes há um aumento sensível no número de bicicletas alugadas.

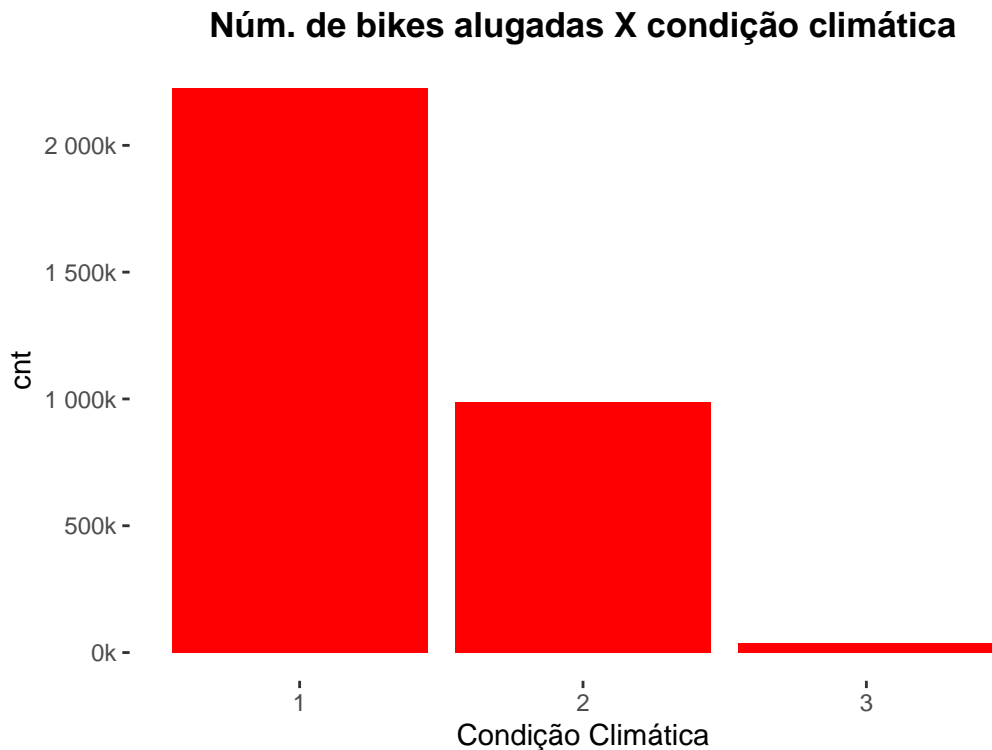
Explorando a relação entre o número de bikes alugadas e o dia da semana:



Em relação ao dia da semana, a análise do gráfico nos mostra que há quase um equilíbrio na distribuição ao longo dos dias, com uma tendência um pouco menor de alugueis aos domingos e segundas-feiras, bem como uma leve tendência a um maior número de alugueis aos sábados.

Desta feita, caso a empresa queira aumentar o número de alugueis no domingo/segunda-feira com vistas a igualar os demais dias da semana, pode ser interessante endereçar campanhas de marketing com descontos ou programas de fidelidade destinados a estes dias em específico.

E como comportam-se os números de alugueis de bicicletas em relação à condição climática?



Conforme era de se esperar, a maioria dos alugueis acontece quando as condições climáticas são: Limpo, Poucas nuvens ou Parcialmente nublado.

Em seguida sob as condições “Nevoeiro + Nublado, Nevoeiro + Nuvens quebradas, Nevoeiro + Poucas nuvens, Nevoeiro”, os alugueis caem para menos da metade em relação à condição número 1.

É importante termos atenção à variável independente ‘weathersit’ pois pode ser que ela contribua sobremaneira para explicar as variações em ‘cnt’.

Desenvolvimento do Modelo De Regressão Linear Múltipla

A regressão linear múltipla é uma técnica estatística que busca modelar a relação entre uma variável dependente, que é aquela que queremos prever, e duas ou mais variáveis independentes. Ela é uma extensão da regressão linear simples, que envolve apenas uma variável independente. Na regressão linear múltipla, o objetivo é estimar os coeficientes das variáveis independentes para prever ou explicar a variabilidade na variável dependente.

A equação da regressão linear múltipla é representada a seguir:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon$$

Após a fase inicial de carregamento e limpeza dos dados, assim como uma análise exploratória para compreender melhor as características do conjunto de dados, avançaremos para a etapa de desenvolvimento de modelos de regressão linear múltipla. Esta etapa é crucial para o projeto, pois visa prever o número de bicicletas alugadas (`cnt`) com base em diversas variáveis explicativas disponíveis. Neste contexto, será adotada a métrica de avaliação do coeficiente de determinação ajustado (R^2 ajustado), dada sua relevância para a precisão de modelos de regressão linear múltipla.

O coeficiente de determinação ajustado (R^2 ajustado) é uma medida estatística que avalia o quão bem o modelo de regressão linear múltipla se ajusta aos dados, levando em consideração o número de variáveis independentes incluídas no modelo. Ele é uma versão ajustada do coeficiente de determinação (R^2), que quantifica a proporção da variabilidade na variável dependente que é explicada pelo modelo.

A fórmula do coeficiente de determinação ajustado é representada a seguir:

$$R^2_{ajustado} = 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)}$$

onde:

- (RSS) é a soma dos quadrados dos resíduos (erro quadrático médio residual).
- (TSS) é a soma total dos quadrados.
- (n) é o número total de observações.
- (p) é o número de variáveis independentes no modelo.

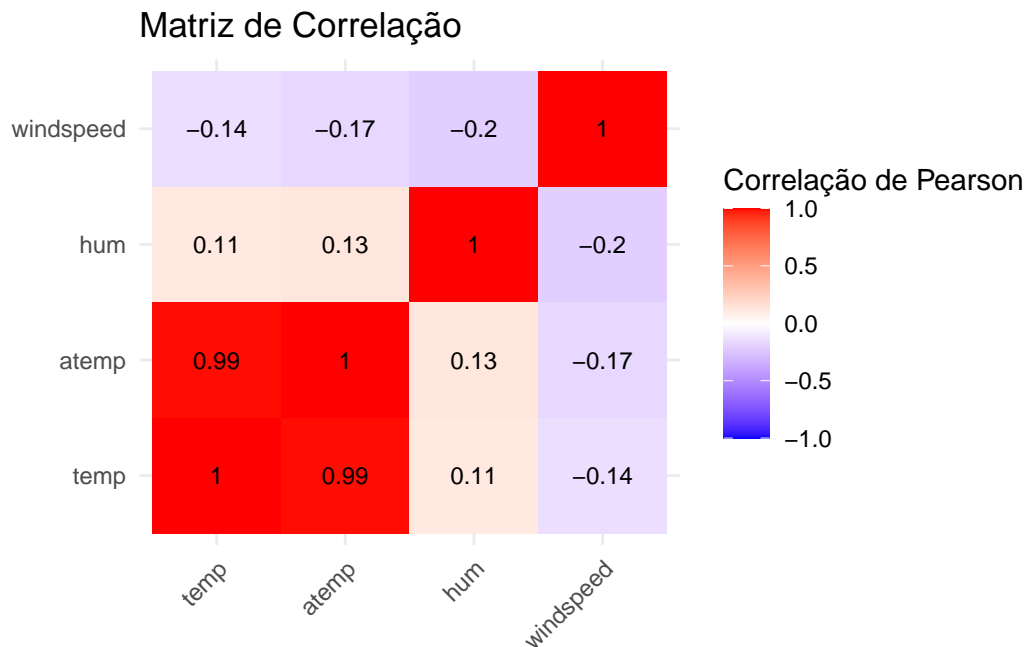
Durante o processo de modelagem, serão exploradas algumas combinações de variáveis, incluindo algumas que foram identificadas durante a análise exploratória como potencialmente influentes no número de alugueis de bicicletas. Esta abordagem permitirá testar a influência relativa de cada variável sobre a variável de interesse (`cnt`). Além disso, serão consideradas iterações do modelo, incluindo ou excluindo variáveis, para determinar qual combinação oferece o melhor desempenho com base no R^2 ajustado. Este procedimento visa encontrar um equilíbrio entre a simplicidade do modelo e sua capacidade de explicar a variação nos dados observados.

Pressupostos para a validação de um modelo de Regressão Linear Múltipla:

Para que seja considerado como estatisticamente significativo e apresente resultados com qualidade e confiabilidade, um modelo de regressão linear deve reunir alguns pressupostos básicos. Não abordaremos todos em seus mínimos detalhes neste estudo, portanto partiremos do pressuposto que existe relação linear entre a variável dependente “cnt” e ao menos uma das variáveis independentes.

- 1) Não existência de Multicolinearidade entre as variáveis preditoras.

Antes de avançar, é importante plotar a matriz de correlação para saber se não há multicolinearidade entre as nossas variáveis preditoras. Desta feita, no primeiro momento analisaremos apenas a matriz de correlação de Pearson para avaliar se há correlação linear entre as nossas variáveis numéricas e o quão forte ela é.



Salta aos olhos a forte correlação linear entre as variáveis temp e atemp. Ora, sendo temp a temperatura real medida naquela data e atemp a sensação térmica, fica fácil perceber a razão desta correlação linear positiva tão forte.

Assim sendo, com vistas a evitar a violação do pressuposto da não existência de multicolinearidade entre as variáveis independentes no modelo de regressão linear, excluiremos da nossa análise a variável “atemp”, mantendo apenas “temp” no modelo.

Quanto às variáveis categóricas, que constituem uma importante parte das variáveis independentes do nosso modelo, analisaremos se existe multicolinearidade entre elas mais adiante, após a construção do modelo, por meio do VIF (Variance Inflation Factor). Por ora, o que sabemos é que a variável preditora “atemp” de partida já não estará presente nos dados.

- 2) Os resíduos devem ter média = 0 e distribuição normal.
- 3) Homocedasticidade das variâncias dos erros. Partiremos do princípio que este pressuposto está satisfeito para os dados em questão.

Desenvolvendo e avaliando o Modelo 1 de Regressão Linear Múltipla:

Desenvolveremos a seguir um modelo de regressão Linear Múltipla saturado com todas as variáveis independentes que temos à disposição, à exceção de “atemp” que foi eliminada na etapa anterior.

O objetivo aqui é avaliar a acurácia do modelo, sobretudo no que diz respeito à significância e grau de contribuição de cada variável independente para a assertividade das previsões.

Ressaltamos que todas as análises realizadas a partir deste momento terão em consideração um nível de significância $\alpha = 0.05$.

Call:

```
lm(formula = cnt ~ ., data = train_data)
```

Residuals:

Min	1Q	Median	3Q	Max
-3320.2	-339.6	66.3	439.7	2522.2

Coefficients: (1 not defined because of singularities)

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	83356.785	52998.883	1.573	0.116343	
dteday	-5.451	3.537	-1.541	0.123799	
season2	889.497	208.418	4.268	2.33e-05	***
season3	800.525	243.107	3.293	0.001056	**
season4	1545.817	204.540	7.558	1.74e-13	***
yr1	3957.887	1294.695	3.057	0.002345	**
mnth2	320.500	184.566	1.737	0.083038	.
mnth3	797.331	278.345	2.865	0.004336	**
mnth4	967.271	419.089	2.308	0.021369	*
mnth5	1358.431	514.285	2.641	0.008493	**
mnth6	1276.007	613.166	2.081	0.037897	*

mnth7	982.290	719.175	1.366	0.172545	
mnth8	1619.025	813.581	1.990	0.047088	*
mnth9	2295.850	902.711	2.543	0.011256	*
mnth10	2072.195	1002.698	2.067	0.039240	*
mnth11	1646.744	1107.966	1.486	0.137782	
mnth12	1773.898	1200.361	1.478	0.140036	
holiday1	-683.278	196.741	-3.473	0.000555	***
weekday1	254.329	122.514	2.076	0.038369	*
weekday2	380.692	116.303	3.273	0.001130	**
weekday3	472.973	115.479	4.096	4.84e-05	***
weekday4	376.212	117.433	3.204	0.001436	**
weekday5	515.156	117.820	4.372	1.47e-05	***
weekday6	468.829	117.845	3.978	7.87e-05	***
workingday1	NA	NA	NA	NA	
weathersit2	-499.501	85.406	-5.849	8.54e-09	***
weathersit3	-2040.012	232.645	-8.769	< 2e-16	***
temp	4466.890	462.041	9.668	< 2e-16	***
hum	-1673.201	337.446	-4.958	9.50e-07	***
windspeed	-2876.164	473.284	-6.077	2.30e-09	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 748.8 on 547 degrees of freedom

Multiple R-squared: 0.8566, Adjusted R-squared: 0.8493

F-statistic: 116.7 on 28 and 547 DF, p-value: < 2.2e-16

Avaliação geral do desempenho do Modelo 1:

Em primeiro medida, analisemos as hipóteses associadas ao F-statistic, que nos traz a informação a respeito do quão significativo é o modelo, informando-nos ainda a respeito da existência de ao menos uma variável independente que seja suficientemente significativa para explicar a variável-alvo cnt:

- H0: Não há relação significativa entre as variáveis independentes e a variável dependente no modelo.
- H1: Pelo menos uma das variáveis independentes tem uma relação significativa com a variável dependente no modelo.

Logo, haja vista que o valor de F-statistic encontra-se distante de 1 (F-Statistic = 116.7), tendo ainda o p-value associado a este teste um valor $< \alpha$ (p-value: < 2.2e-16), possuímos evidências estatísticas suficientes para rejeitar a hipótese nula do Teste F e afirmar que **O**

modelo é globalmente significativo e pelo menos uma das variáveis independentes tem uma relação significativa com a variável dependente no modelo.

Acurácia do modelo e Regressão Linear Múltipla:

Vejamos os números relacionados às métricas de avaliação da acurácia e precisão do modelo preditivo:

```
[1] "RMSE (Training): 729.70166960945"
```

```
[1] "MAE (Training): 536.39163608996"
```

```
[1] "R-squared (Training): 0.856648892418335"
```

```
[1] "R-squared (Adjusted, Training): 0.849035005751909"
```

Na presente análise nos concentraremos apenas nas medidas do Coeficiente de Determinação Ajustado, haja vista que tal métrica possui um padrão de análise que varia de 0 a 1, sendo que quanto mais próximo de 0, pior é a acurácia do modelo e quanto mais próximo de 1 melhor será o seu desempenho preditivo.

Além disso, o motivo pelo qual consideraremos o Coeficiente de Determinação Ajustado ao invés do Coeficiente de Determinação “simples” é que esta métrica é a mais recomendável para regressões lineares múltiplas, sobretudo aquelas que possuem muitas variáveis independentes, como é o caso nesta análise. O coeficiente de Determinação Ajustado penaliza mais os modelos maiores e que possuem mais variáveis.

Isto posto, **é possível afirmar que o nosso modelo performou bem ao realizar predições para os dados de treino, ao apresentar um R-squared (Adjusted) = 0.849035005751909.**

A seguir, apresentaremos as conclusões relativas à performance do modelo quando aplicado à predição dos dados de teste:

```
[1] "RMSE: 855.10298986998"
```

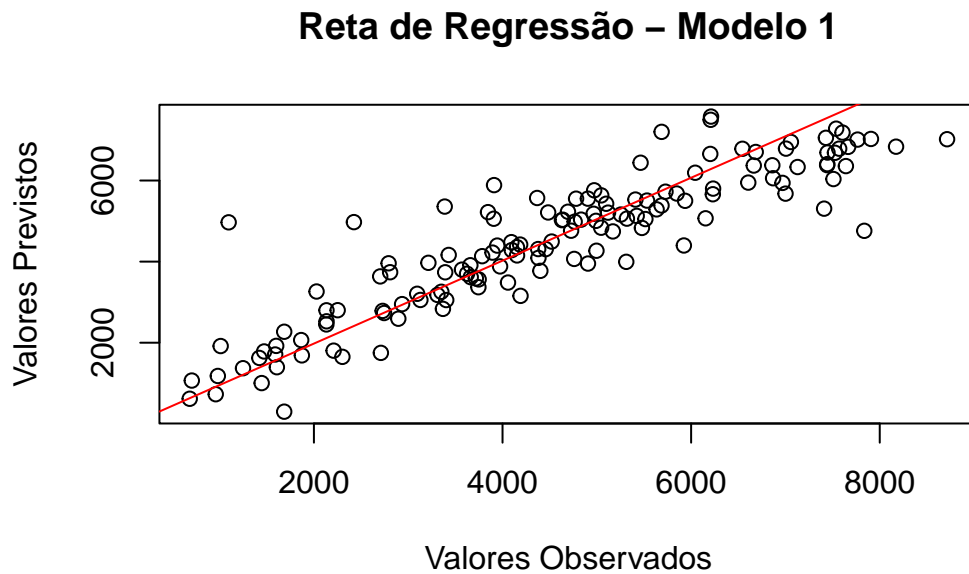
```
[1] "MAE: 608.153511116337"
```

```
[1] "R-squared: 0.808121594695134"
```

```
[1] "R-squared (Adjusted): 0.757991200516386"
```

Aqui há um alerta: Ao aplicarmos o modelo aos dados “novos” do conjunto de Teste, o valor do Coeficiente de Determinação Ajustado cai consideravelmente, vez que neste caso $R\text{-squared (Adjusted)} = 0.757991200516386$.

Isto não quer dizer que seja um modelo ruim, haja vista que 0.75 ainda é um valor relativamente alto, mas indica que pode haver espaço para melhorias, sobretudo ao retirar do modelo aquelas variáveis que não contribuem para explicar a variável-alvo, o que será assunto do próximo tópico.



Significância das variáveis preditoras para o modelo:

Ao analisar os resultados do treino do nosso modelo, sob a ótica da contribuição das variáveis independentes na explicação da variável-alvo, devemos considerar o teste T associado a elas e o p-value associado, sendo estas as hipóteses:

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

tendo em consideração o nível de significância definido neste trabalho ($\alpha = 0.05$), consideraremos como significativas para o modelo (ou seja, rejeita-se a hipótese nula apenas para) as seguintes variáveis independentes:

Variável	Estimate	Std. Error	t value	Pr(>)	Significância
weathersit3	-2040.012	232.645	-8.769	< 2e-16	Muito significativo
temp	4466.890	462.041	9.668	< 2e-16	Muito significativo
season4	1545.817	204.540	7.558	1.74e-13	Muito significativo
yr1	3957.887	1294.695	3.057	0.002345	Muito significativo
windspeed	-2876.164	473.284	-6.077	2.30e-09	Muito significativo
weathersit2	-499.501	85.406	-5.849	8.54e-09	Muito significativo
holiday1	-683.278	196.741	-3.473	0.000555	Muito significativo
weekday5	515.156	117.820	4.372	1.47e-05	Muito significativo
weekday1	254.329	122.514	2.076	0.038369	significativo
season2	889.497	208.418	4.268	2.33e-05	Muito significativo
season3	800.525	243.107	3.293	0.001056	Muito significativo
weekday2	380.692	116.303	3.273	0.001130	Muito significativo
weekday3	472.973	115.479	4.096	4.84e-05	Muito significativo
weekday4	376.212	117.433	3.204	0.001436	Muito significativo
weekday6	468.829	117.845	3.978	7.87e-05	Muito significativo
mnth3	797.331	278.345	2.865	0.004336	Muito significativo
mnth4	967.271	419.089	2.308	0.021369	Significativo
mnth5	1358.431	514.285	2.641	0.008493	Muito significativo
mnth6	1276.007	613.166	2.081	0.037897	Significativo
mnth8	1619.025	813.581	1.990	0.047088	Significativo
mnth9	2295.850	902.711	2.543	0.011256	Muito significativo
mnth10	2072.195	1002.698	2.067	0.039240	Significativo
hum	-1673.201	337.446	-4.958	9.50e-07	Muito significativo

Aquelas que consideramos como pouco ou não significativas de acordo com o resultado do teste T associado são (incluindo-se o intercepto):

Variável	Estimate	Std. Error	t value	Pr(>)	Significância
(Intercept)	83356.785	52998.883	1.573	0.116343	Não significativo
dteday	-5.451	3.537	-1.541	0.123799	Não significativo
workingday1	NA	NA	NA	NA	NA
mnth2	320.500	184.566	1.737	0.083038	Não significativo
mnth7	982.290	719.175	1.366	0.172545	Não significativo
mnth11	1646.744	1107.966	1.486	0.137782	Não significativo
mnth12	1773.898	1200.361	1.478	0.140036	Não significativo

E quanto ao pressuposto da distribuição dos resíduos?

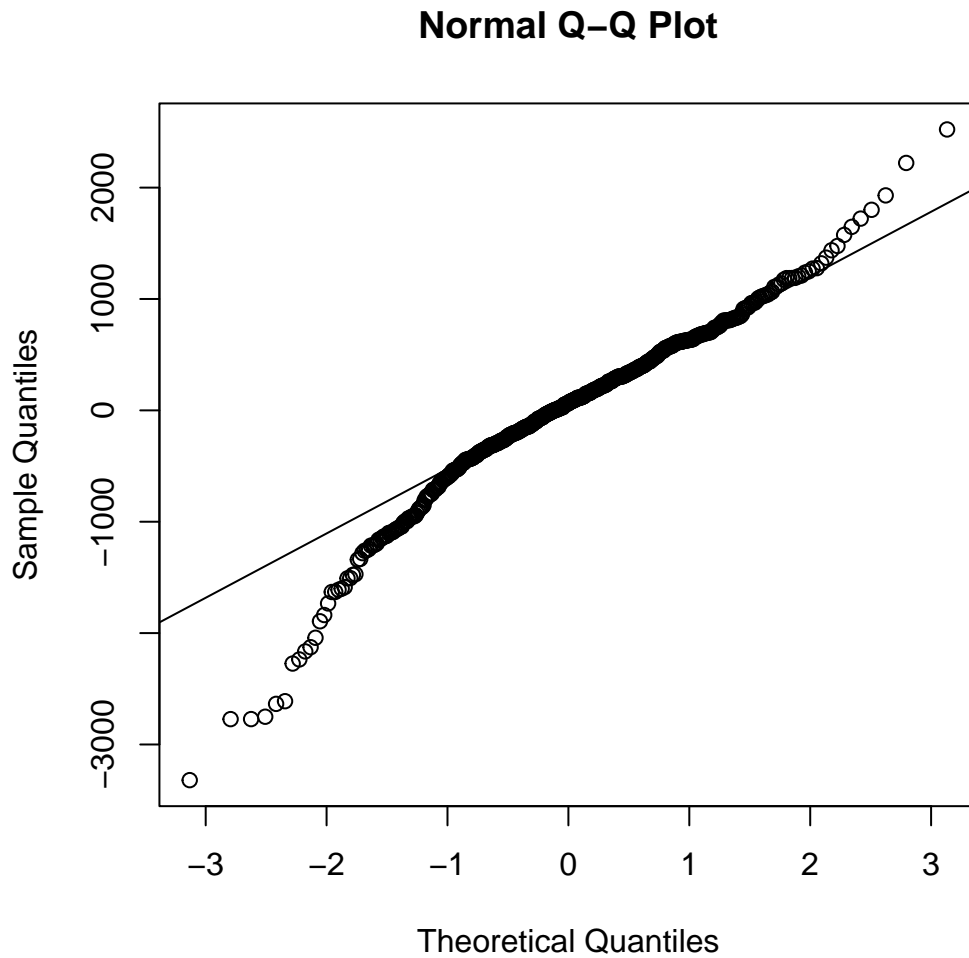


Figure 3: Distribuições dos Resíduos

Podemos observar que mesmo com um certo “peso” nas caudas, sobretudo do lado negativo, há uma concentração muito considerável de observações dos resíduos sobre a curva, o que nos indica que a distribuição destes resíduos está muito próxima da normal e a média está próxima de 0, vindo a confirmar portanto a satisfação do pressuposto.

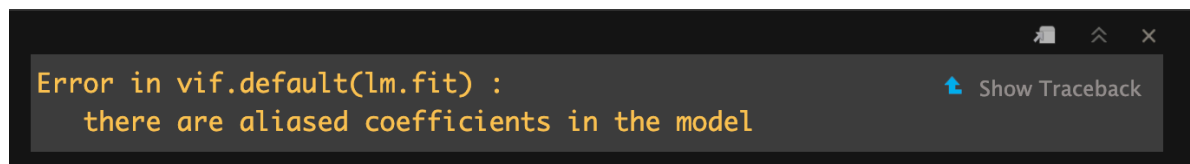
Importante ressaltar que pode ser que os próximos modelos, com menos variáveis, sejam ainda mais aderentes ao pressuposto e venham a corrigir um pouco este “peso” a maior nas caudas da distribuição dos resíduos.

Problemas identificados no modelo

Apesar de apresentar indicadores bons, o primeiro modelo possui alguns problemas. Um deles é o excesso de variáveis, que costuma ser computacionalmente custoso a modelos de regressão linear, sobretudo pela quantidade de variáveis categóricas existentes e que acabam por gerar um alto número de variáveis dummy.

Assim sendo, pode ser benéfico ao modelo - em questões computacionais e tendo-se em consideração os possíveis custos para uma organização relacionados ao deploy deste, a redução da dimensionalidade do conjunto de variáveis preditoras.

Entretanto, um outro problema ainda mais grave foi diagnosticado: Ao tentar utilizar o VIF para calcular a existência de multicolinearidade entre as variáveis independentes do modelo, o R traz o seguinte erro como output:



```
Error in vif.default(lm.fit) :  
  there are aliased coefficients in the model
```

Este erro indica que pode existir multicolinearidade entre as variáveis preditoras do modelo. Ora, sendo a não existência de multicolinearidade entre as variáveis preditoras um dos principais pressupostos para a assunção de que um modelo de regressão linear é bom e confiável, resta-nos a opção de investigar se está a ocorrer multicolinearidade e excluir as variáveis problemáticas.

Modelo 2 - Novo modelo com menos variáveis

Neste momento nos dedicaremos à melhora do modelo anterior, retirando inicialmente as variáveis “dteday” e “workingday”, baseados em sua não contribuição e significância e analisaremos o VIF para este caso.

Call:

```
lm(formula = cnt ~ . - workingday - dteday, data = train_data)
```

Residuals:

Min	1Q	Median	3Q	Max
-3382.3	-355.7	61.1	444.5	2568.6

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	1665.61	267.73	6.221	9.80e-10	***
season2	885.19	208.66	4.242	2.60e-05	***
season3	796.77	243.40	3.273	0.001129	**
season4	1555.49	204.70	7.599	1.30e-13	***
yr1	1964.74	64.74	30.348	< 2e-16	***
mnth2	164.25	154.43	1.064	0.287985	
mnth3	479.48	187.19	2.561	0.010690	*
mnth4	487.98	281.33	1.735	0.083382	.
mnth5	714.68	300.48	2.378	0.017728	*
mnth6	468.57	319.10	1.468	0.142566	
mnth7	14.21	350.81	0.041	0.967700	
mnth8	477.29	336.96	1.416	0.157208	
mnth9	983.05	299.55	3.282	0.001097	**
mnth10	585.06	273.42	2.140	0.032815	*
mnth11	-14.11	258.37	-0.055	0.956476	
mnth12	-49.11	205.42	-0.239	0.811154	
holiday1	-676.65	196.94	-3.436	0.000636	***
weekday1	252.80	122.66	2.061	0.039786	*
weekday2	376.90	116.42	3.237	0.001280	**
weekday3	472.34	115.62	4.085	5.06e-05	***
weekday4	375.12	117.58	3.190	0.001502	**
weekday5	514.62	117.97	4.362	1.54e-05	***
weekday6	467.94	117.99	3.966	8.28e-05	***
weathersit2	-489.14	85.25	-5.738	1.59e-08	***
weathersit3	-2041.73	232.93	-8.765	< 2e-16	***
temp	4447.61	462.45	9.617	< 2e-16	***
hum	-1700.92	337.39	-5.041	6.29e-07	***
windspeed	-2897.36	473.68	-6.117	1.82e-09	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 749.7 on 548 degrees of freedom

Multiple R-squared: 0.856, Adjusted R-squared: 0.8489

F-statistic: 120.7 on 27 and 548 DF, p-value: < 2.2e-16

Como podemos perceber, o modelo 2 aparenta ter bons indicadores, tendo um Adjusted R-squared = 0.8489, ou seja, pouquíssimo inferior ao seu antecessor e com um valor de F-statistic = 120.7, corroborado pelo p-value < α .

Aqui o intercepto já passa a ser significativo para o modelo, ao contrário do que ocorrera anteriormente e a variável “mnth”(representada pelas variáveis dummy) parece seu pouco significativa para explicar a variável resposta “cnt”.

Entretanto, vejamos o VIF associado ao novo modelo:

	GVIF	Df	GVIF ^{1/(2*Df)}
season	211.907282	3	2.441693
yr	1.073706	1	1.036198
mnth	522.363646	11	1.329059
holiday	1.138400	1	1.066958
weekday	1.219839	6	1.016698
weathersit	2.001208	2	1.189387
temp	7.458795	1	2.731080
hum	2.209591	1	1.486469
windspeed	1.208740	1	1.099427

Tendo em consideração que um valor de $VIF > 10$ denota a existência de multicolinearidade, é certo dizer que as variáveis season ($VIF = 211.907282$) e mnth ($VIF = 522.363646$) possuem multicolinearidade e fazem com que o modelo viole o pressuposto relacionado à não existência desta condição.

Chama atenção ainda que a variável “temp” apresente um VIF muito próximo do limite aceitável. Entretanto, na presente análise não a descartaremos.

Desta feita, desenvolveremos um novo modelo, desta vez sem a variável “mnth”, haja vista ter apresentado níveis de significância para o modelo mais modestos que a variável “season”, bem como o maior VIF se comparada às demais.

Modelo 3: Excluindo “mnth” do conjunto de variáveis independentes:

Call:

```
lm(formula = cnt ~ . - workingday - dteday - mnth, data = train_data)
```

Residuals:

Min	1Q	Median	3Q	Max
-3122.09	-375.11	62.75	505.89	2300.57

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1389.57	261.71	5.310	1.59e-07 ***
season2	1160.66	126.79	9.154	< 2e-16 ***
season3	857.26	166.30	5.155	3.53e-07 ***
season4	1581.21	105.74	14.953	< 2e-16 ***
yr1	1978.79	67.16	29.466	< 2e-16 ***

holiday1	-729.85	203.73	-3.582	0.000370	***
weekday1	236.55	128.54	1.840	0.066250	.
weekday2	375.94	121.89	3.084	0.002141	**
weekday3	477.36	120.92	3.948	8.89e-05	***
weekday4	371.77	122.85	3.026	0.002590	**
weekday5	538.35	123.84	4.347	1.64e-05	***
weekday6	490.39	123.76	3.963	8.38e-05	***
weathersit2	-506.90	88.64	-5.719	1.75e-08	***
weathersit3	-2092.16	241.54	-8.662	< 2e-16	***
temp	4978.92	339.96	14.646	< 2e-16	***
hum	-1304.42	333.20	-3.915	0.000102	***
windspeed	-2771.34	492.62	-5.626	2.92e-08	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 788.6 on 559 degrees of freedom

Multiple R-squared: 0.8375, Adjusted R-squared: 0.8328

F-statistic: 180.1 on 16 and 559 DF, p-value: < 2.2e-16

	GVIF	Df	GVIF^(1/(2*Df))
season	3.738169	3	1.245785
yr	1.044119	1	1.021821
holiday	1.101030	1	1.049300
weekday	1.162640	6	1.012637
weathersit	1.909766	2	1.175561
temp	3.642921	1	1.908644
hum	1.947716	1	1.395606
windspeed	1.181545	1	1.086989

Ao avaliarmos a performance do modelo, agora sem a presença das variáveis independentes “workingday”, “dteday” e “mnth”, bem como o valor do VIF associado às variáveis independentes restantes, podemos concluir que:

- 1) **Trata-se de um modelo globalmente significativo** e ao menos uma variável independente é significativa para explicar a variável dependente “cnt”. Tal assunção é corroborada pelo valor de F-statistic, o qual encontra-se distante de 1 (F-Statistic = 180.1), tendo ainda o p-value associado a este teste um valor $< \alpha$ (p-value: < 2.2e-16)
- 2) O pressuposto da não existência de multicolinearidade entre as variáveis independentes não foi violado. **Portanto, não há multicolinearidade no modelo em questão.**
- 3) Os resíduos possuem distribuição muito próxima da normal e sua média aproxima-se sobremaneira de 0, conforme demonstrado no histograma a seguir:

Histograma dos Resíduos – Modelo 3

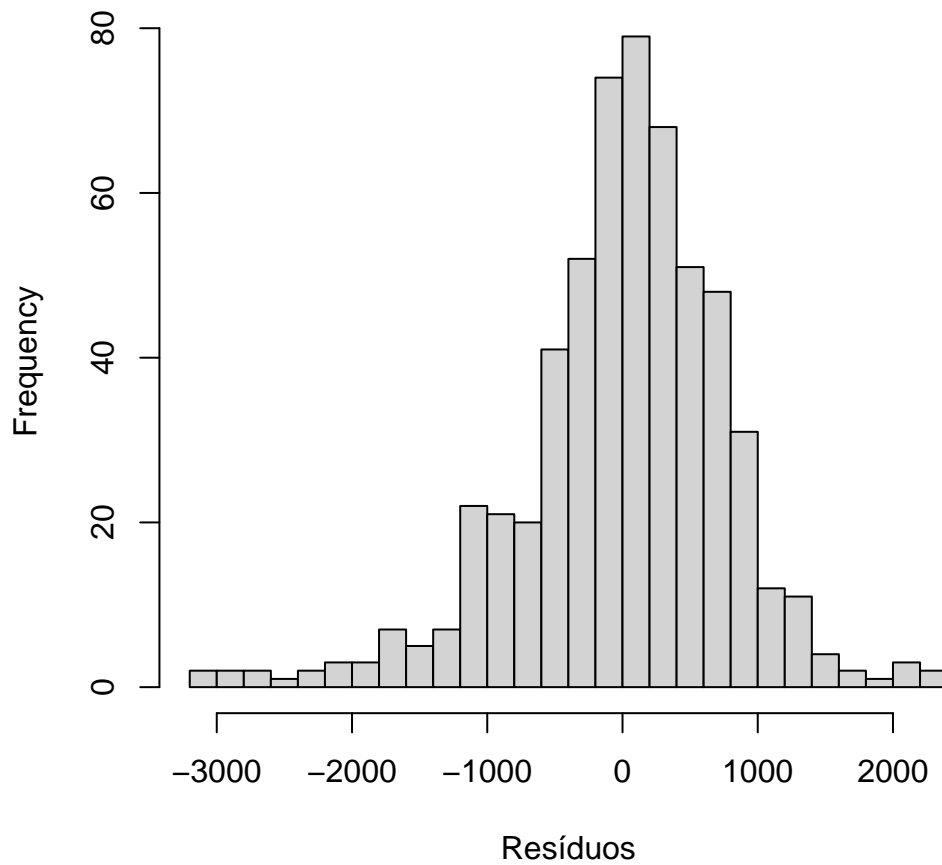


Figure 4: Distribuições dos Resíduos

- 4) O modelo possui uma boa acurácia quando aplicado aos dados de treino, haja vista o coeficiente de determinação ajustado = 0.8328.

Resta-nos entretanto aplicá-lo aos dados de teste para saber qual será a acurácia para dados não vistos anteriormente e saber se há um possível overfitting, ou até mesmo underfitting.

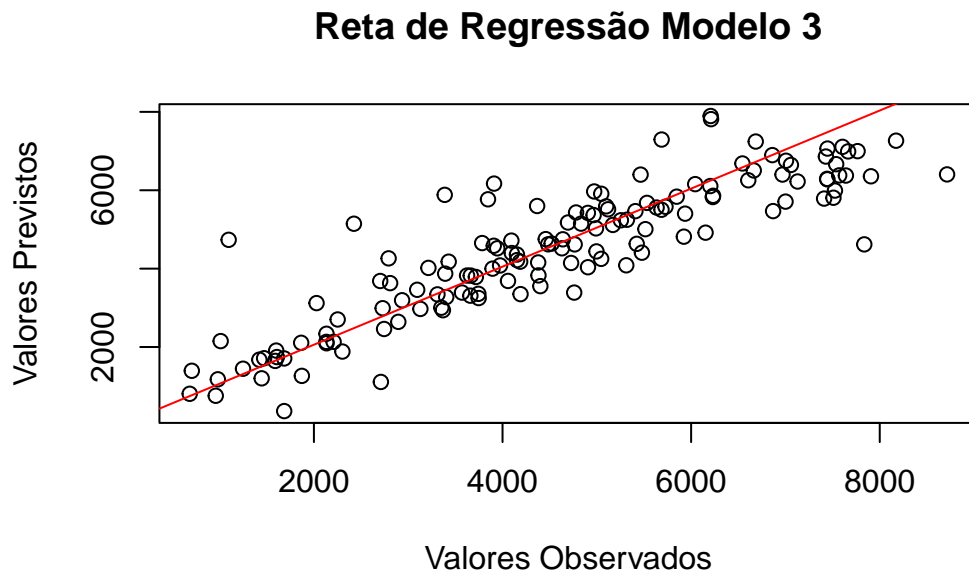
Aplicação do último modelo aos dados de teste:

```
[1] "R-squared: 0.776348088959387"
```

[1] "R-squared (Adjusted): 0.747489777857372"

Apesar de o modelo apresentar uma ligeira queda no valor do coeficiente de determinação ajustado em relação ao primeiro modelo que fora definido neste estudo, é possível ainda assim dizer que trata-se de um bom coeficiente ($R\text{-squared (Adjusted)} = 0.7474$). Logo, pode-se dizer que o modelo está apto a fazer boas previsões para novos dados, e que este é melhor por não possuir os problemas de multicolinearidade e excesso de variáveis que os dois anteriores possuíam.

Analisemos a seguir o quão bem ajustada está a reta de regressão do modelo em questão:



O gráfico em questão confirma a assunção de que o modelo 3, o qual não conta com as variáveis preditoras “workingday”, “dteday” e “mnth”, é adequado para prever o número de bicicletas que serão alugadas, utilizand-se para tanto do conjunto de variáveis independentes restantes no conjunto de dados bike sharing, as quais são significativas para o contexto do nosso modelo.

Extra - Avaliação da performance de uma árvore de decisão para a mesma tarefa de Regressão.

Após desenvolver e analisar a performance de 3 modelos de Regressão Linear Múltipla para a tarefa de previsão do número total de bicicletas alugadas/dia em função de uma série de

variáveis independentes, testaremos como será a performance de uma árvore de decisão para o mesmo conjunto de dados.

[1] 216 13

Conforme observamos acima, 216 observações do nosso dataframe foram separadas para Teste do modelo de árvore de decisão. As observações restantes constituem o conjunto de Treino do modelo.

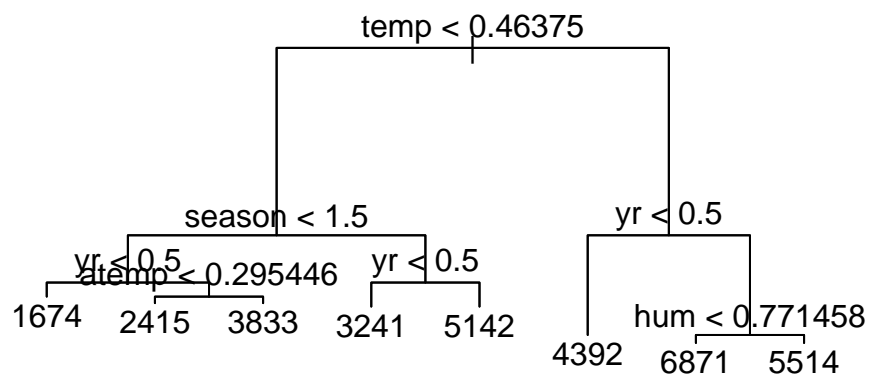


Figure 5: Árvore com 8 Nós-Folha

```

Regression tree:
tree(formula = cnt ~ ., data = df_limpo, subset = traint.df)
Variables actually used in tree construction:
[1] "temp" "season" "yr" "atemp" "hum"
Number of terminal nodes: 8
Residual mean deviance: 737800 = 363700000 / 493
Distribution of residuals:
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-5120.0 -438.2   117.3     0.0   524.5   2140.0

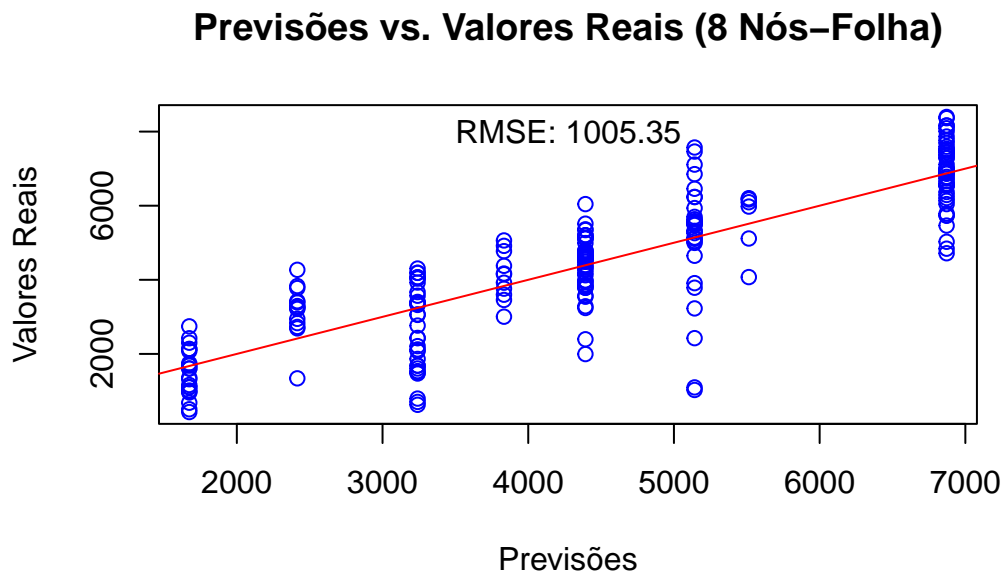
```

Ao treinarmos o modelo aplicando-o ao data set de treino, obtivemos como output automático do R uma árvore de decisão com 8 nós-folha.

Isto está diretamente ligado à complexidade do modelo. Sendo assim, quanto maior o número de nós folha, mais complexo se torna o modelo, o que pode levá-lo ao sobreajuste.

É salutar ressaltar que o R utilizou como variáveis independentes apenas o conjunto “temp”, “season”, “yr”, “atemp” e “hum”.

Vejamos a capacidade de realizar previsões do nosso modelo e a distância entre o observado e aquilo que foi predito para os dados de teste:



O RMSE (Root Mean Square Error) é uma medida que quantifica a média das diferenças entre os valores previstos por um modelo e os valores reais de uma variável de interesse. É comumente usado para avaliar a precisão de modelos de previsão, como modelos de regressão, **onde valores menores indicam uma melhor capacidade de previsão do modelo.**

Com vistas a melhorar a capacidade preditiva do modelo, utilizaremos o mecanismo da validação cruzada para tentar encontrar o número ideal de nós-folha para o nosso modelo de árvore de decisão para regressão.

Para a primeira árvore com 8 nós-folha, alcançamos um RMSE de 1005.35.

Aplicação da validação cruzada e avaliação dos resultados:

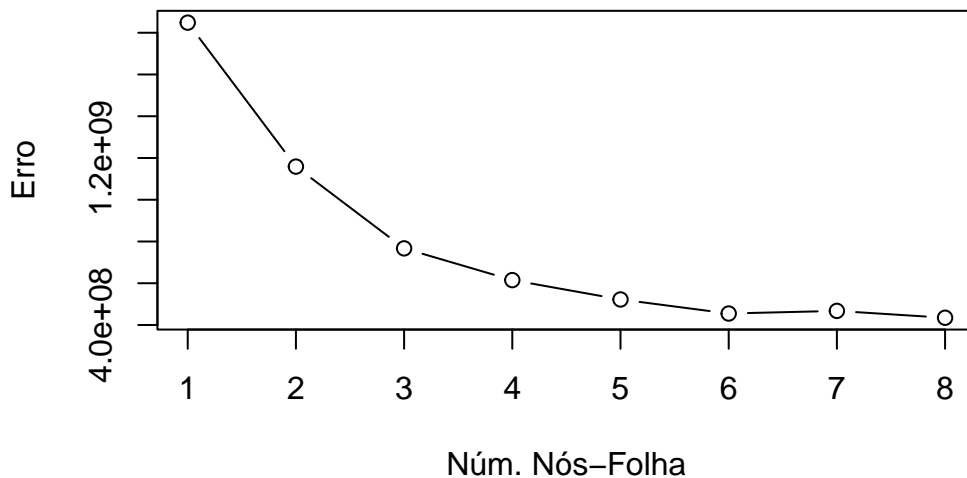


Figure 6: Erro de Validação Cruzada vs. Tamanho da Árvore

Mesmo que a validação cruzada não aponte uma grande diferença de RMSE da árvore com 6 ou 8 nós, é importante que façamos o treinamento do modelo com 6 nós-folha com o objetivo de diminuir a complexidade do modelo e saber qual será o tamanho da perda, em termos de precisão, uma vez que a depender deste resultado pode ser mais vantajoso utilizar o modelo com menos nós.

Um nó folha (nó terminal) é o ponto final onde uma decisão ou classificação é feita. Ele não possui ramificações adicionais, representando uma conclusão ou resultado final baseado nos atributos e critérios analisados ao longo do caminho na árvore de decisão.

Desta feita, após a avaliação dos resultados, treinaremos uma nova árvore de decisão, desta vez com 6 nós-folha e avaliaremos a performance das suas previsões por meio da comparação do valor da raiz quadrada do erro médio(RMSE):

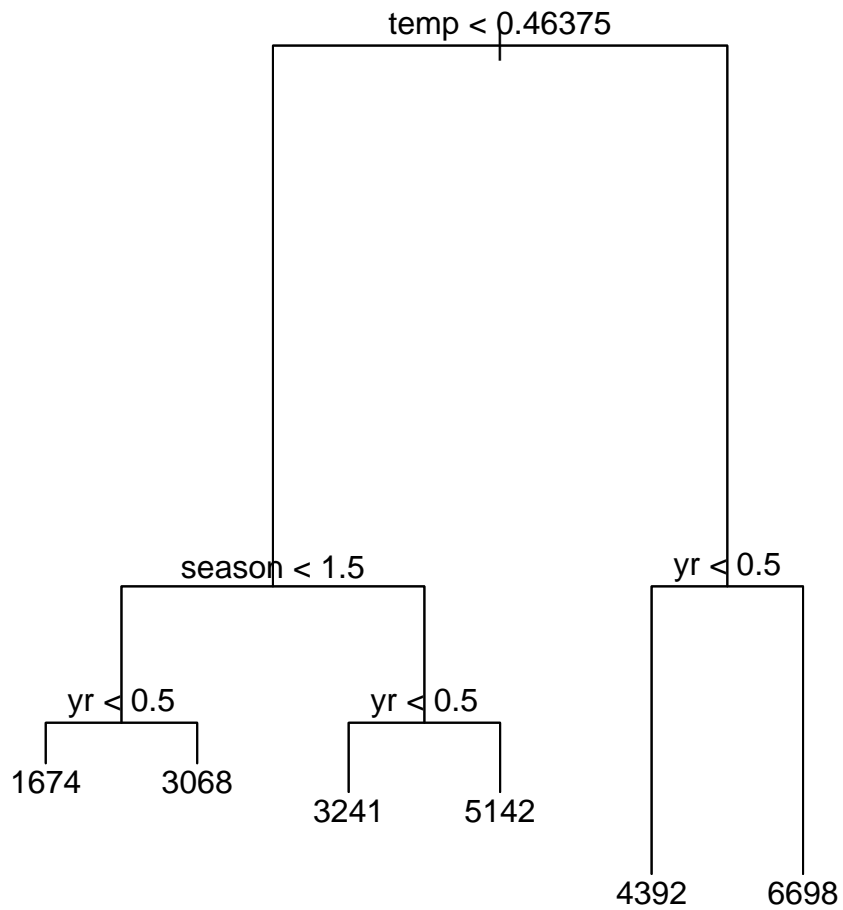


Figure 7: Nova árvore com apenas 6 Nós-Folha

[1] "RMSE da árvore com 6 Nós-Folha: 1033.20818180359"

Como é possível perceber, a diminuição da complexidade da árvore por meio da poda, levou o modelo a uma perda significativa de RMSE, tendo este resultado em 1033.20, frente aos 1005.35 da árvore com 8 nós.

Para diminuir a complexidade da árvore, o R eliminou duas variáveis independentes deste último modelo, sendo elas “atemp” e “hum”, o que pode ter levado ainda ao aumento do RMSE.

Haja vista que o desenvolvimento da árvore de decisão apresenta-se neste estudo como um “extra”, sendo o seu principal objeto as regressões lineares múltiplas, utilizaríamos a primeira árvore de decisão se quiséssemos realizar previsões por meio deste tipo de modelo (decision trees) e não a última, com apenas 6 nós-folha.

Conclusão

Concluimos o presente projeto com a assunção de que o último modelo de regressão linear múltipla desenvolvido (modelo 3) foi o que mais se aproximou daquilo que consideramos ideal para prever novos valores para o número de alugueis de bicicletas, dadas as mesmas informações constantes ao data set bike sharing.

Desenvolvemos e construímos o modelo tendo por base as etapas iniciais de limpeza e exploração dos dados, passando posteriormente pela verificação dos pressupostos básicos para o desenvolvimento do modelo de regressão linear ,múltipla, divisão dos dados entre treino e teste e, por fim, a comparação de modelos até chegar àquele que consideramos o que melhor generaliza sobre os dados e os padrões relacionados, tendo como benchmark para comparação de modelos o coeficiente de determinação ajustado.

Em seguida, como uma etapa extra, desenvolvemos duas árvores de decisão com vistas a efetuar a mesma tarefa de regressão para a variável dependente CNT do nosso conjunto de dados, sendo uma com 8 nós-folha (a qual seria escolhida caso fossemos optar neste estudo por modelos de árvore de decisão) e outra com apenas 6 nós-folha.

Desta feita, como próximas etapas do presente projeto poderíamos citar a aplicação e desenvolvimento de novos modelos em árvore para o mesmo conjunto de dados, como por exemplo Random Forest ou XGboost, com vistas a compará-los a tudo o que fora desenvolvido neste estudo e buscar aquele que melhor se adeque aos dados, sem contudo perder a capacidade de generalizar e prever sobre novos dados de teste e/ou validação..

Por fim, após a escolha do melhor modelo baseado em critérios objetivos seria possível construir a etapa de deploy para colocá-lo em produção em benefício da organização, ao utilizar para tanto ferramentas cloud e esteiras de CI/CD.

O código fonte em R do presente trabalho consta ao link do repositório a seguir: [Repositório GitHub](#).