

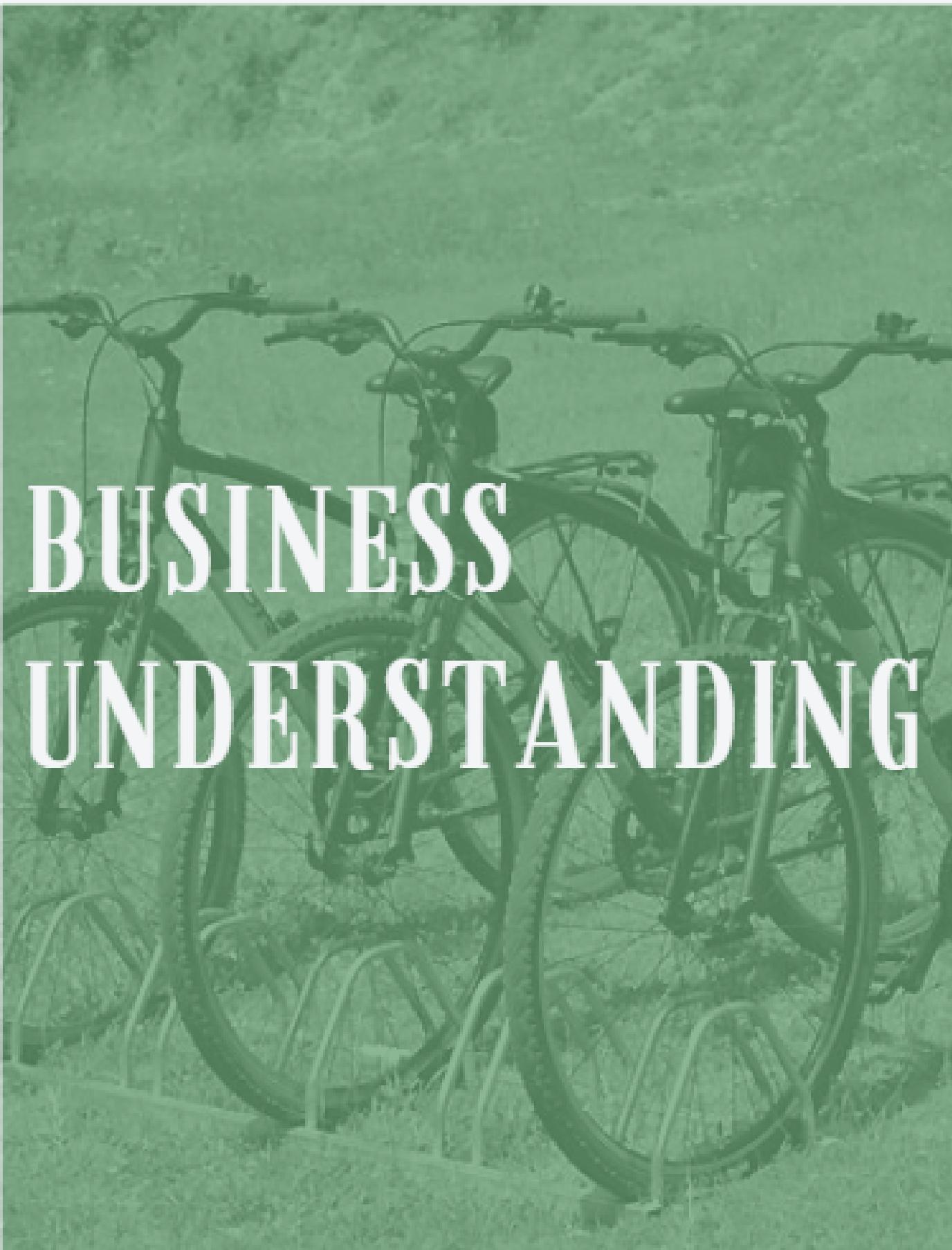


PROJETO APLICADO EM CIÊNCIA DE DADOS I

PREVISÃO NO DATASET

SEOUL BIKES

Business Understanding



BUSINESS UNDERSTANDING

OBJETIVO

O objetivo deste projeto consiste em prever o número de bicicletas necessárias a disponibilizar a cada hora de forma a tornar o serviço o mais eficiente possível.

CRITÉRIOS DE SUCESSO

Ser estabelecida uma margem mínima de bicicletas disponíveis.

Obter um modelo com capacidade de generalização e um erro de previsão mínimo.

FERRAMENTAS

5 alunos de 2º ano da licenciatura de Ciência de Dados, um docente supervisor, dataset proveniente do UCI Machine Learning Repository, computadores pessoais , Jupyter Notebook, linguagem de programação Python, Prezi e Django.

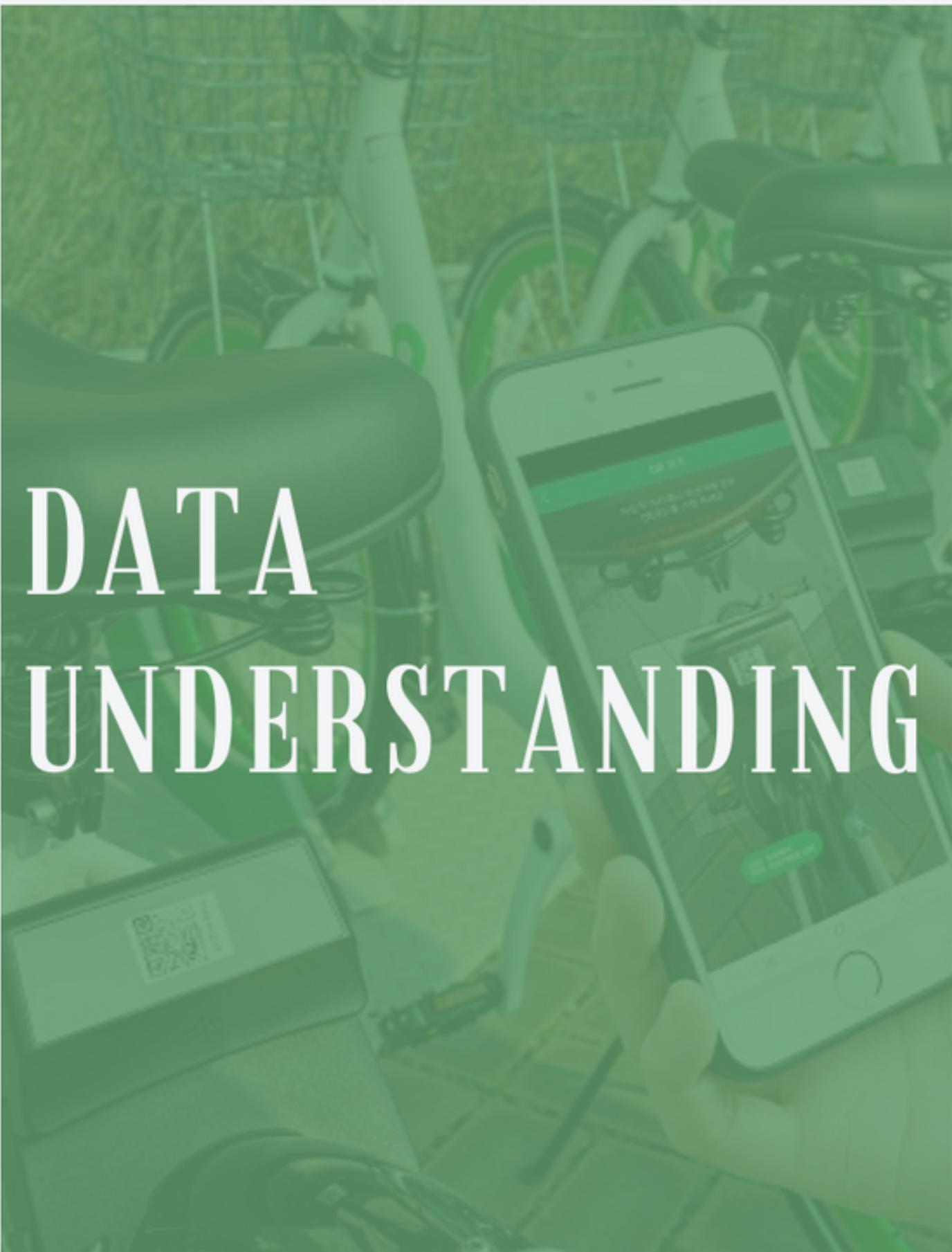


PROJETO APLICADO EM CIÊNCIA DE DADOS I

PREVISÃO NO DATASET

SEOUL BIKES

Análise Exploratória de Dados



DATA UNDERSTANDING

APRESENTAÇÃO DO DATASET

- Dados de alugueres de bicicletas entre Dez. 2017 e Nov. 2018
- 8760 observações e 14 variáveis

Coluna	Tipo	Descrição
Date	String	O dia do ano, durante 365 dias
Rented Bike Count	Inteiro	Número de bicicletas alugadas por hora
Hour	Inteiro	Hora do dia
Temperature(°C)	Float	Temperatura por hora
Humidity(%)	Inteiro	Humidade no ar
Wind Speed (m/s)	Float	Velocidade do vento em metros por segundo
Visibility(10m)	Inteiro	Visibilidade por 10 metros
Dew Point Temperature(°C)	Float	Temperatura no inicio do dia
Solar Radiation (MJ/m2)	Float	Radiação solar
Rainfall(mm)	Float	Chuva por milímetro
Snowfall(cm)	Float	Neve por centímetro
Seasons	String	Estação do ano
Holiday	String	Se é ou não feriado nacional
Functioning Day	String	Se o serviço de aluguer estava ou não em funcionamento

ALTERAÇÃO DO TIPO DE "DATE" DE STRING PARA DATETIME

```
seoul_bikes['Date']=pd.to_datetime(seoul_bikes['Date'],format="%d/%m/%Y")
```

ESTATÍSTICAS DESCRIPTIVAS BÁSICAS

	count	mean	std	min	25%	50%	75%	max
Rented Bike Count	8760.0	704.602055	644.997468	0.0	191.00	504.50	1065.25	3556.00
Hour	8760.0	11.500000	6.922582	0.0	5.75	11.50	17.25	23.00
Temperature(°C)	8760.0	12.882922	11.944825	-17.8	3.50	13.70	22.50	39.40
Humidity(%)	8760.0	58.226256	20.362413	0.0	42.00	57.00	74.00	98.00
Wind speed (m/s)	8760.0	1.724909	1.036300	0.0	0.90	1.50	2.30	7.40
Visibility (10m)	8760.0	1436.825799	608.298712	27.0	940.00	1698.00	2000.00	2000.00
Dew point temperature(°C)	8760.0	4.073813	13.060369	-30.6	-4.70	5.10	14.80	27.20
Solar Radiation (MJ/m2)	8760.0	0.569111	0.868746	0.0	0.00	0.01	0.93	3.52
Rainfall(mm)	8760.0	0.148687	1.128193	0.0	0.00	0.00	0.00	35.00
Snowfall (cm)	8760.0	0.075068	0.436746	0.0	0.00	0.00	0.00	8.80

DIAS DE "NO" FUNCTIONING DAY

No Functioning Day	Rented Bike Count
02/10/2018	0
03/11/2018	0
04/10/2018	0
06/10/2018	0
06/11/2018	0
09/10/2018	0
09/11/2018	0
10/05/2018	0
11/04/2018	0
18/09/2018	0
19/09/2018	0
28/09/2018	0
30/09/2018	0

Valores Omissos

Nenhum.

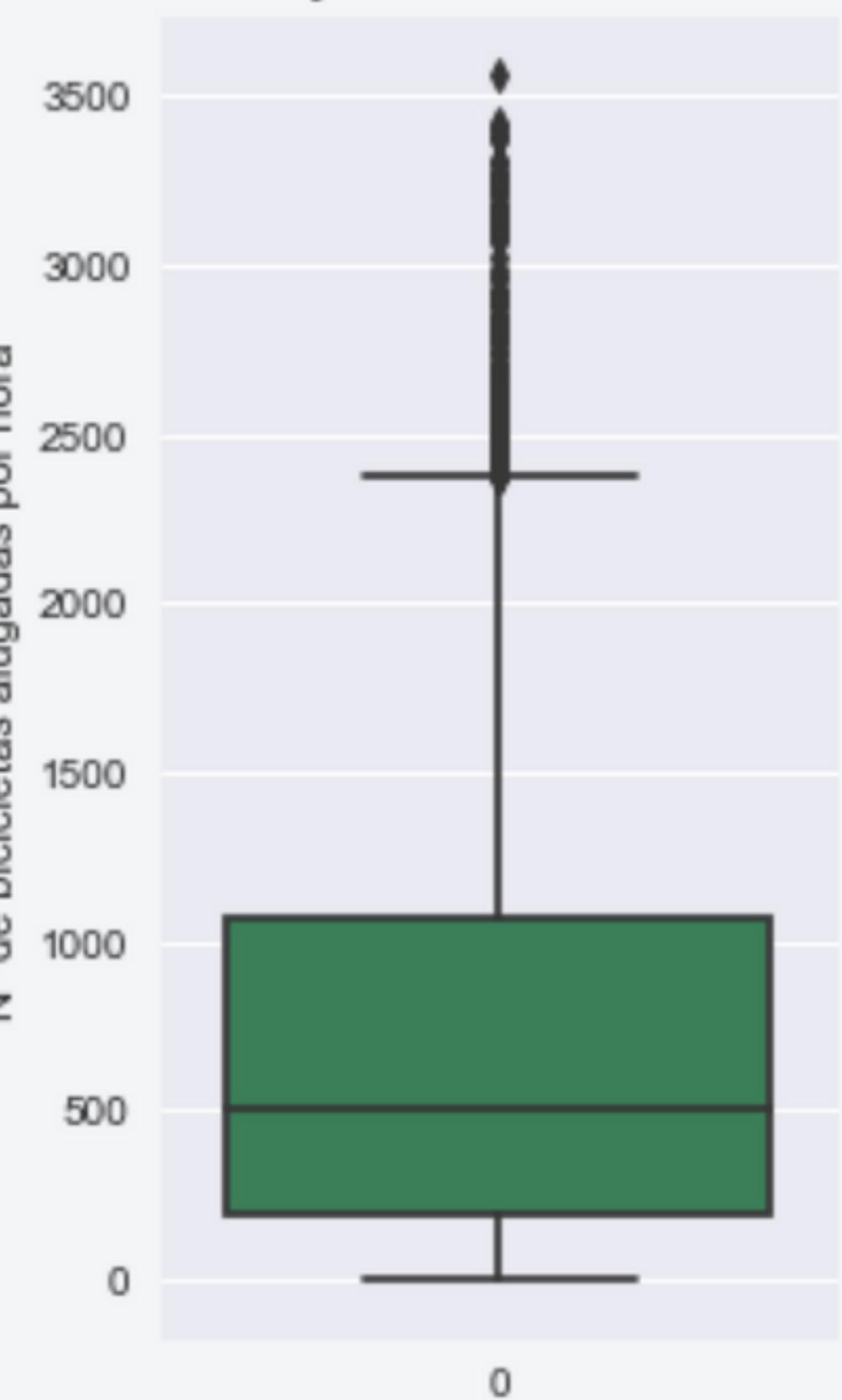
Duplicados

Nenhum.

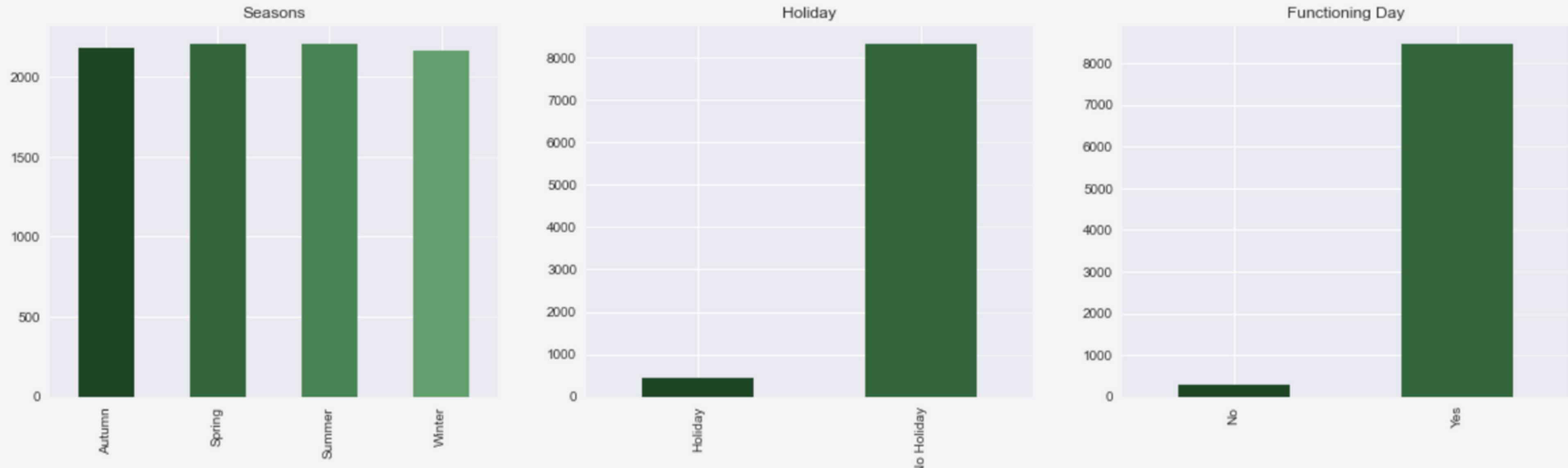
Rented Bike Count

Distribuição da variável target ao longo do dataset.

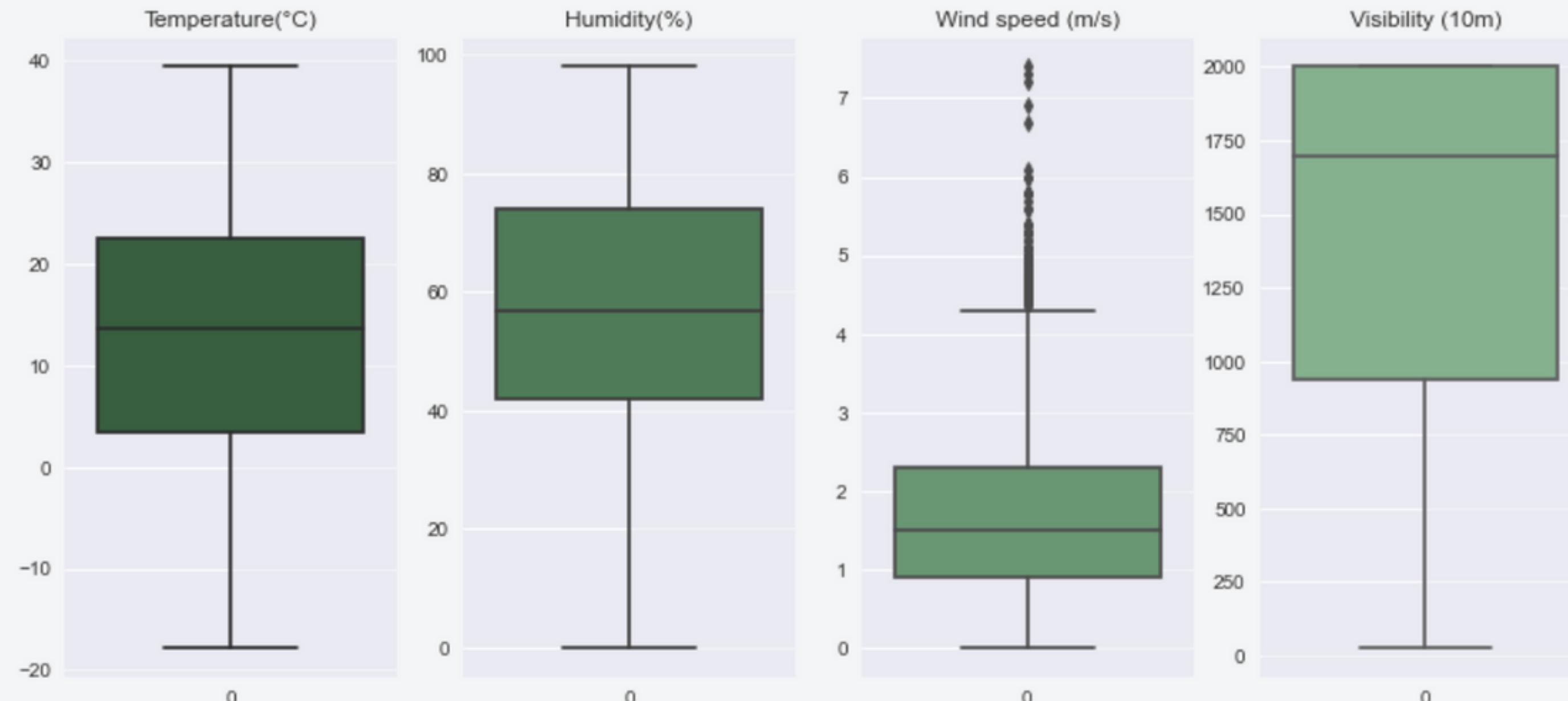
Distribuição de Rented Bike Count



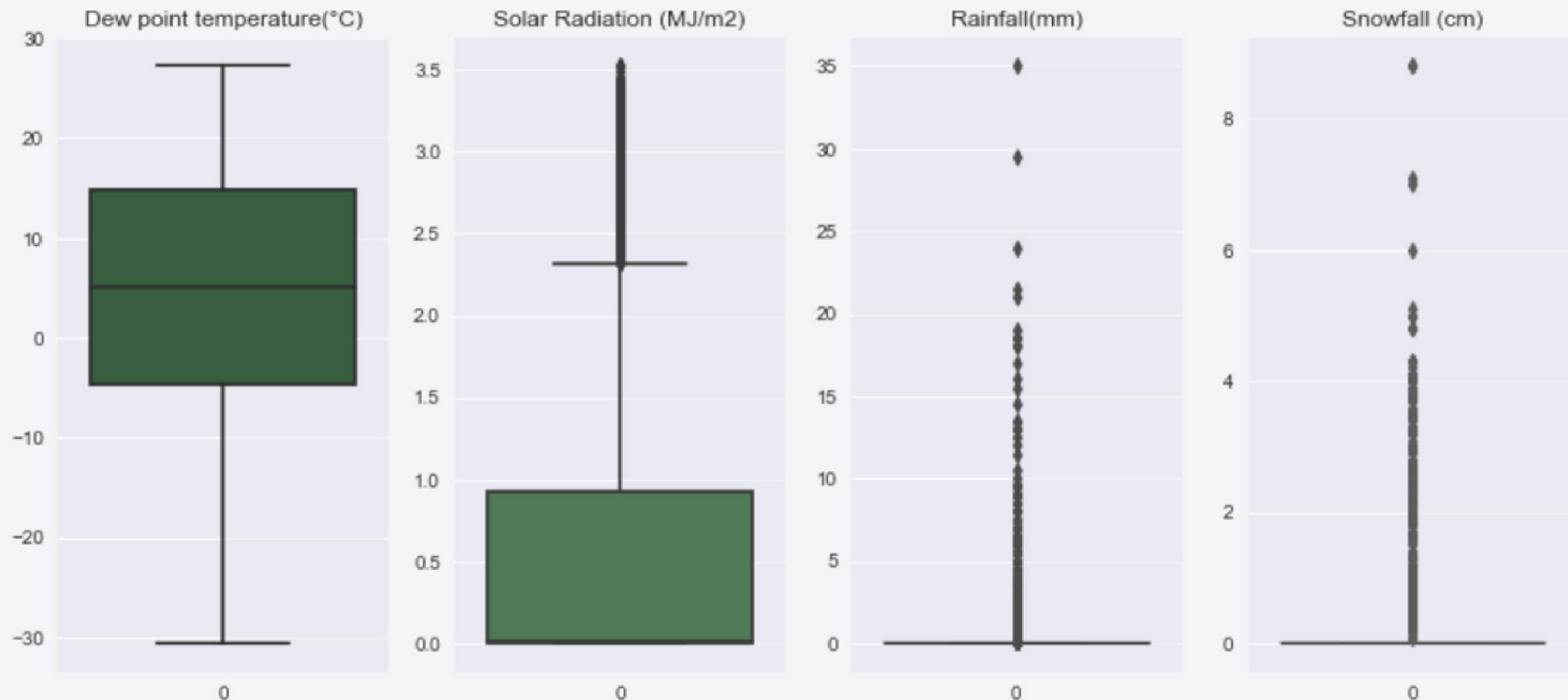
Distribuição das variáveis categóricas



Distribuição das variáveis numéricas



Distribuição das variáveis numéricas



Outliers

Rainfall(mm)

Total de 528.

Todas as observações com Rainfall(mm)>0 são consideradas outliers severos.

Snowfall(cm)

Total de 443.

Todas as observações com Snowfall(cm)>0 são consideradas outliers severos.

Wind speed (m/s)

Total de 161.

Rented Bike Count

Total de 158.

Não se removeram os outliers.

Condições atmosféricas em Seoul

Seasons	Rainfall(mm)	Snowfall (cm)
Autumn	0.122756	0.056319
Spring	0.182880	0.000000
Summer	0.253487	0.000000
Winter	0.032824	0.247500

Ao contrário do que seria de esperar obtiveram-se os maiores níveis de precipitação durante o Verão. As condições atmosféricas influenciam o nº de alugueres de bicicletas portanto os "outliers" que representavam estas condições teriam que ser mantidos.



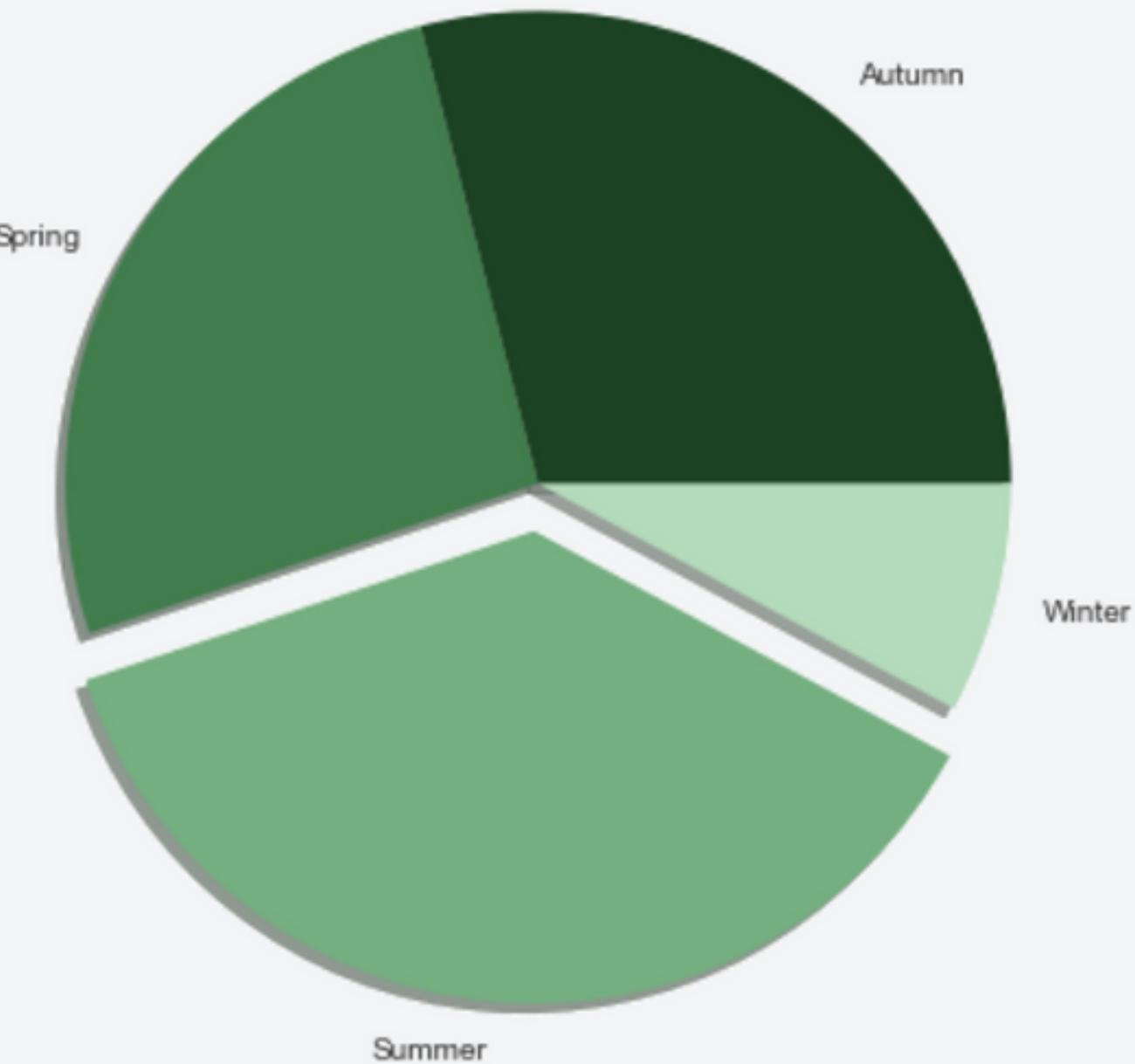
PROJETO APLICADO EM CIÊNCIA DE DADOS I

PREVISÃO NO DATASET SEOUL BIKES

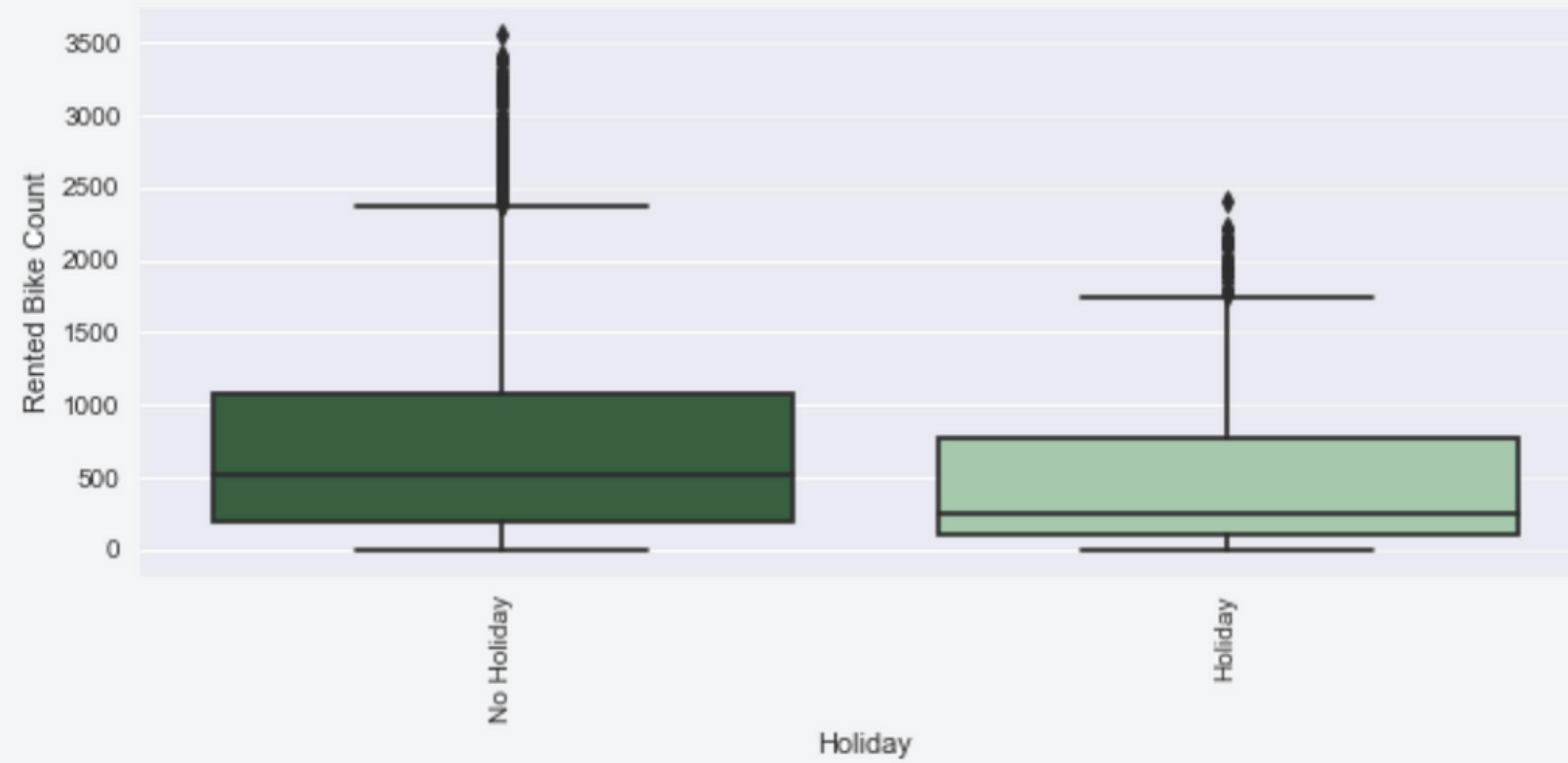
Data Visualization



"Rented Bike Count" por estação do ano



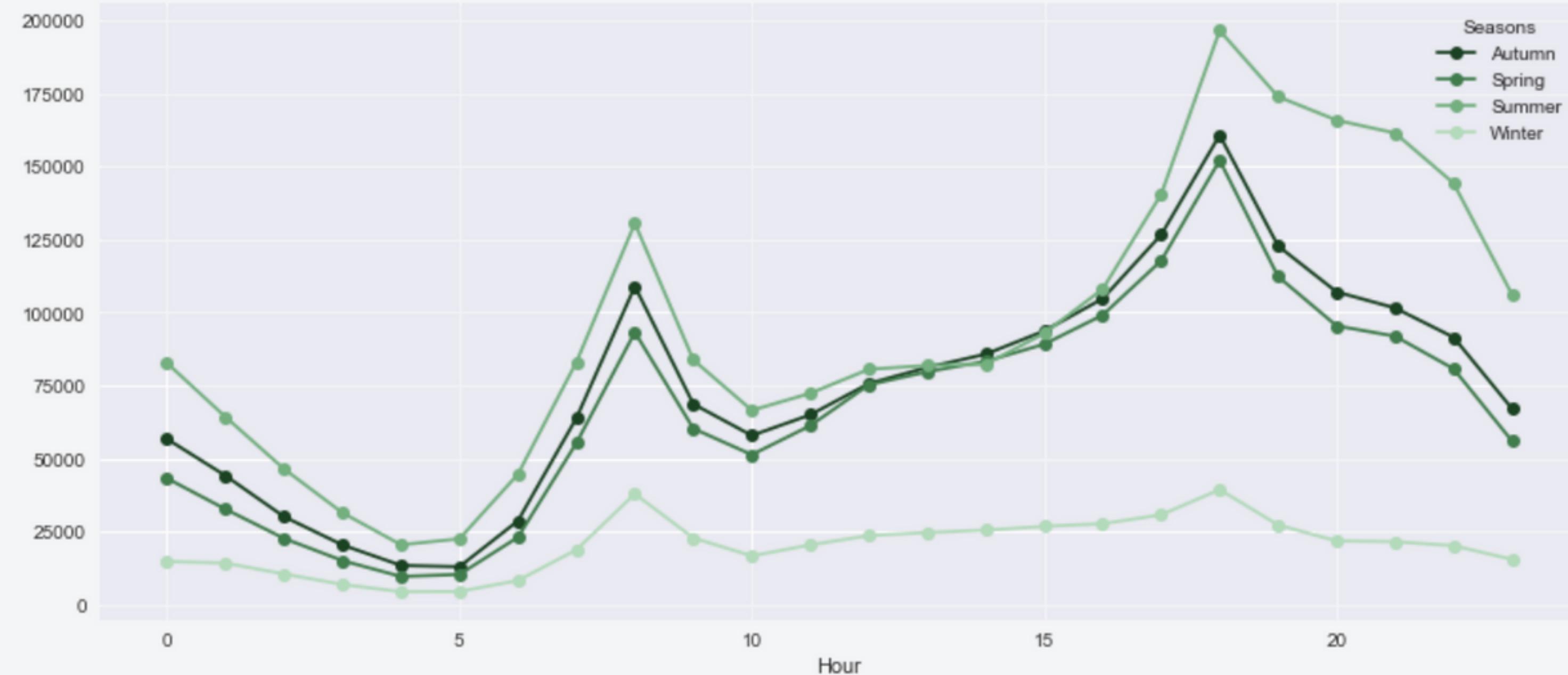
Alugueres de Bicicletas sendo ou não feriado



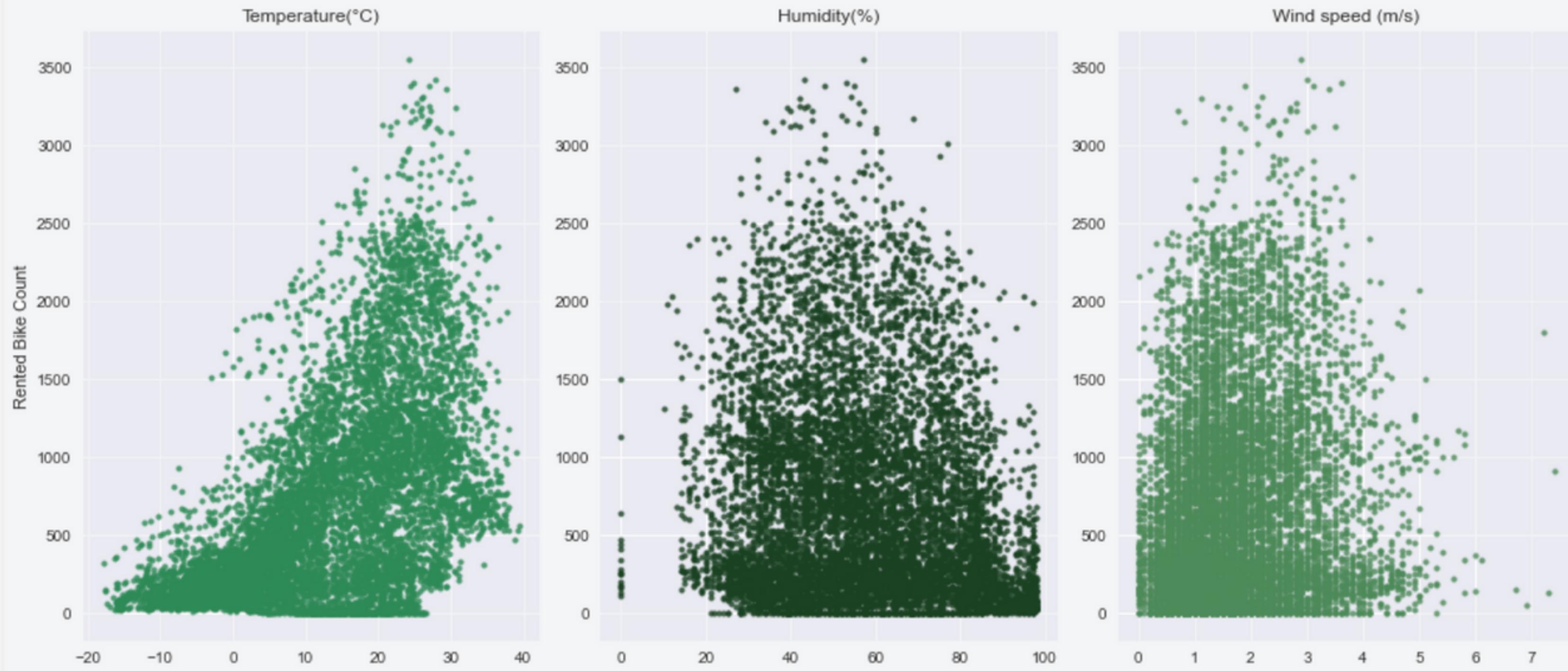
Nº médio de bicicletas alugadas por dia



Nº médio de bicicletas alugadas por hora e por estação do ano



Como varia o nº de alugueres consoante as condições atmosféricas?





PROJETO APLICADO EM CIÊNCIA DE DADOS I

PREVISÃO NO DATASET SEOUL BIKES

Data Modelling



Correlações entre as variáveis



Holiday

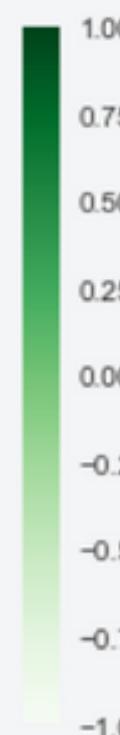
$-1.4e-15$

Seasons

$3.3e-14$

Functioning Day

$1.114e-13$



Rainfall_scale

Categorias	Intervalo (em mm)
No Rain	0mm
Light	>0 e <=2.4
Moderate	>2.4 e <= 10
Heavy	>10

Escalas

Foram criadas escalas para as variáveis, de modo a facilitar a sua interpretação.

Snowfall_scale

Categorias	Intervalo (em cm)
No	0
Yes	>0

Dummies

Conversão das variáveis categóricas a dummy, para posterior uso das mesmas na modelação.

Seasons_Spring	Seasons_Summer	Seasons_Winter	Holiday_No Holiday	Functioning Day_Yes	Rainfall_scale_Light	Rainfall_scale_Moderate	Rainfall_scale_Heavy	Snowfall_scale_Yes
8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000
0.252055	0.252055	0.246575	0.950685	0.966324	0.045776	0.011986	0.002511	0.050571
0.434217	0.434217	0.431042	0.216537	0.180404	0.209011	0.108830	0.050054	0.219132
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000
1.000000	1.000000	0.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000
1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Remoção de variáveis

Removeram-se as variáveis Dew Point Temperature e Functioning Day.

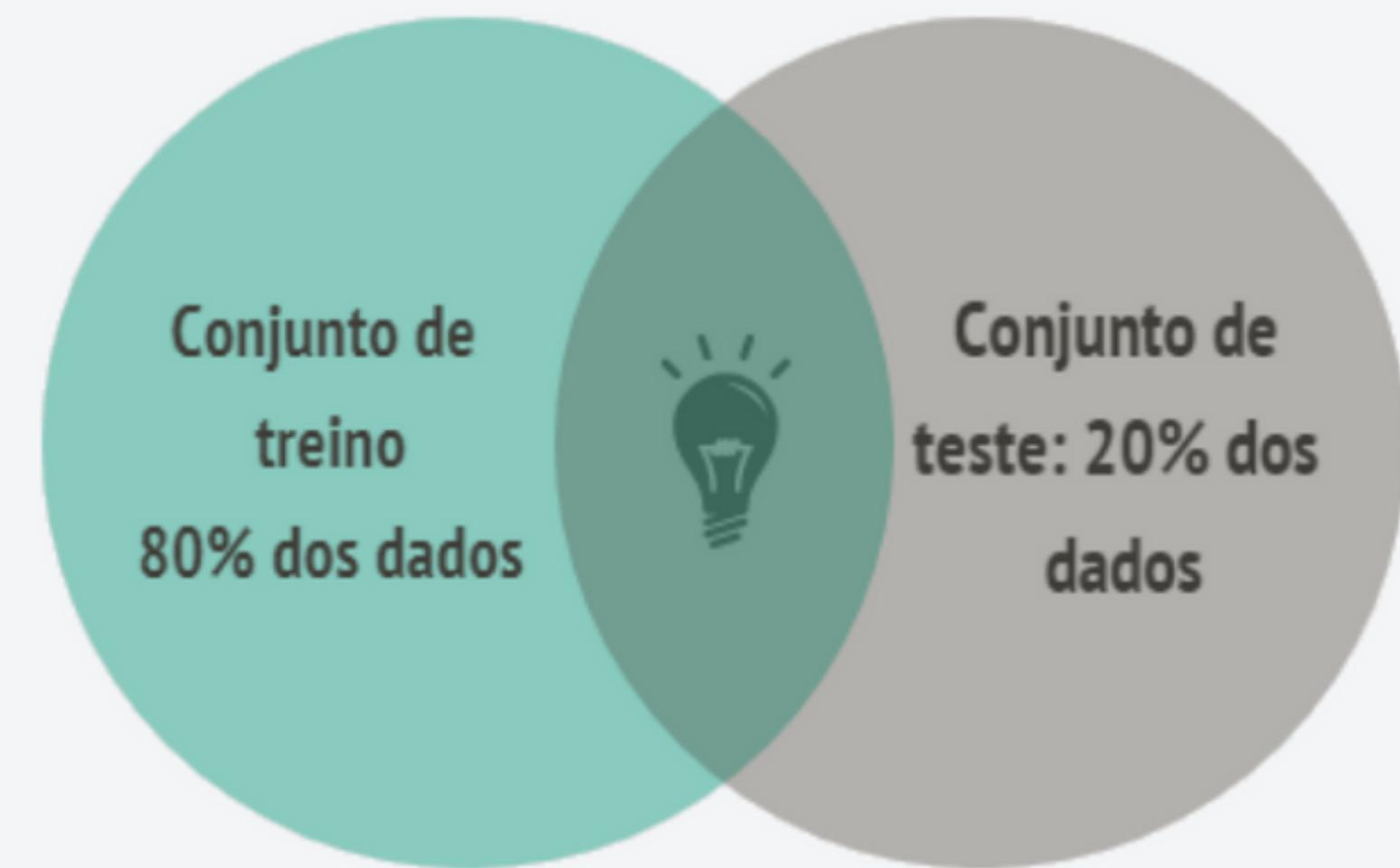
Estandardização

Estandardizaram-se as variáveis numéricas, uma vez que se encontravam em escalas distintas.

Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)
0	254	-1.662748	-1.484762	-1.032395	0.458429	0.929577	-0.654079	-0.132495 -0.174951
1	204	-1.518249	-1.509548	-0.983575	-0.895248	0.929577	-0.654079	-0.132495 -0.174951
2	173	-1.373751	-1.550858	-0.934756	-0.701865	0.929577	-0.654079	-0.132495 -0.174951
3	107	-1.229252	-1.567382	-0.885937	-0.798556	0.929577	-0.654079	-0.132495 -0.174951
4	78	-1.084754	-1.550858	-1.081214	0.555121	0.929577	-0.654079	-0.132495 -0.174951

Divisão em conjuntos de treino e teste

O dataset foi dividido nas seguintes proporções:



1^a fase de modelação

Primeiro modelaram-se os dados estandardizados tendo as variáveis Rainfall e Snowfall com e sem escala.

	Modelos	R ²
1	Regressão Linear standardized com escala	0.557071
2	LASSO standardized com escala	0.557022
3	Regressão Linear standardized com var. originais	0.537481
4	LASSO standardized com var. originais	0.537430

De seguida, comparou-se com a modelação em dados não estandardizados usando Rainfall_scale e Snowfall_scale):

Modelos	R ²	MAPE	RMSE
Regressão Linear	0.557071	153.791141	437.617583
LASSO	0.557038	153.591243	437.633552

Modelação

Aplicaram-se diferentes modelos no intuito de melhorar os resultados e os erros de previsão:

	Modelos	R^2	MAPE	RMSE
0	Regressão Linear	0.557071	153.791141	437.617583
1	LASSO	0.557038	153.591243	437.633552
2	LASSO Afinado	0.557011	153.558042	437.646986
3	Ridge	0.557068	153.628636	437.618644
4	Decision Tree Regressor	0.766208	63.115568	317.937701
5	Extra Tree Regressor	0.882172	47.588345	225.710424
6	Random Forest Regressor	0.873874	57.862877	233.523399
7	Regressão Linear com Target Log	0.541650	73.751154	445.170030
8	Extra Tree com Target Log	0.869221	31.852998	237.792017
9	Random Forest Tree com Target Log	0.860088	35.400087	245.954629

Interpretação do melhor modelo

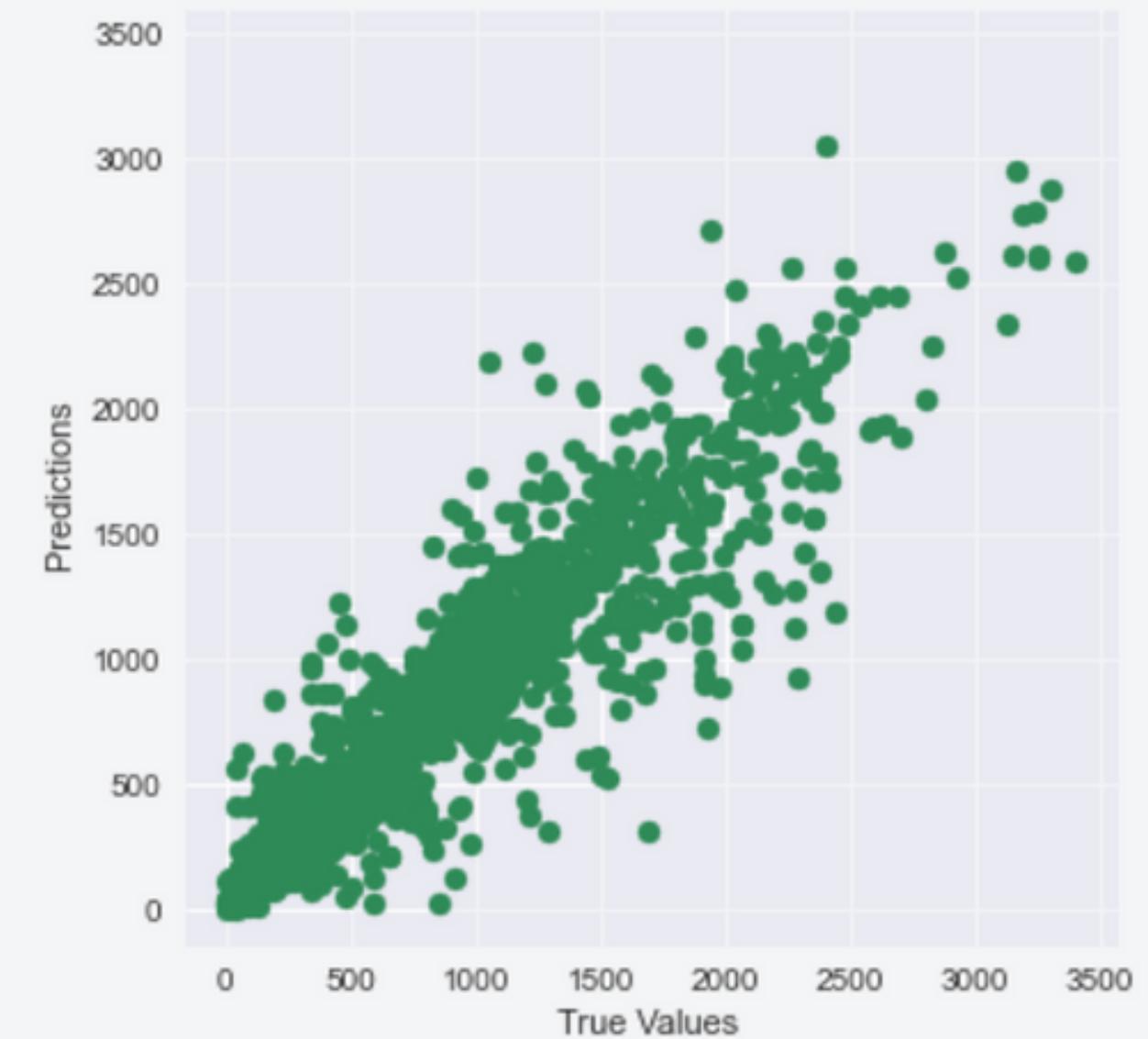
Extra Tree com Target Log

- R²: 86.9%
- MAPE: 31.85%
- RMSE: 237.79

Esta técnica consiste na definição inicial de um parâmetro `max_features=15`.

De seguida foram geradas diversas árvores de decisão nas subamostras aleatoriamente criadas, que geraram cada uma previsões. Ao fazer a média de todas as previsões obtidas, foi obtido um valor mais representativo das valores preditivos.

Forma que ajustou melhor os dados



Variável Binária Eventos

0 = dia sem evento

1= dia de evento



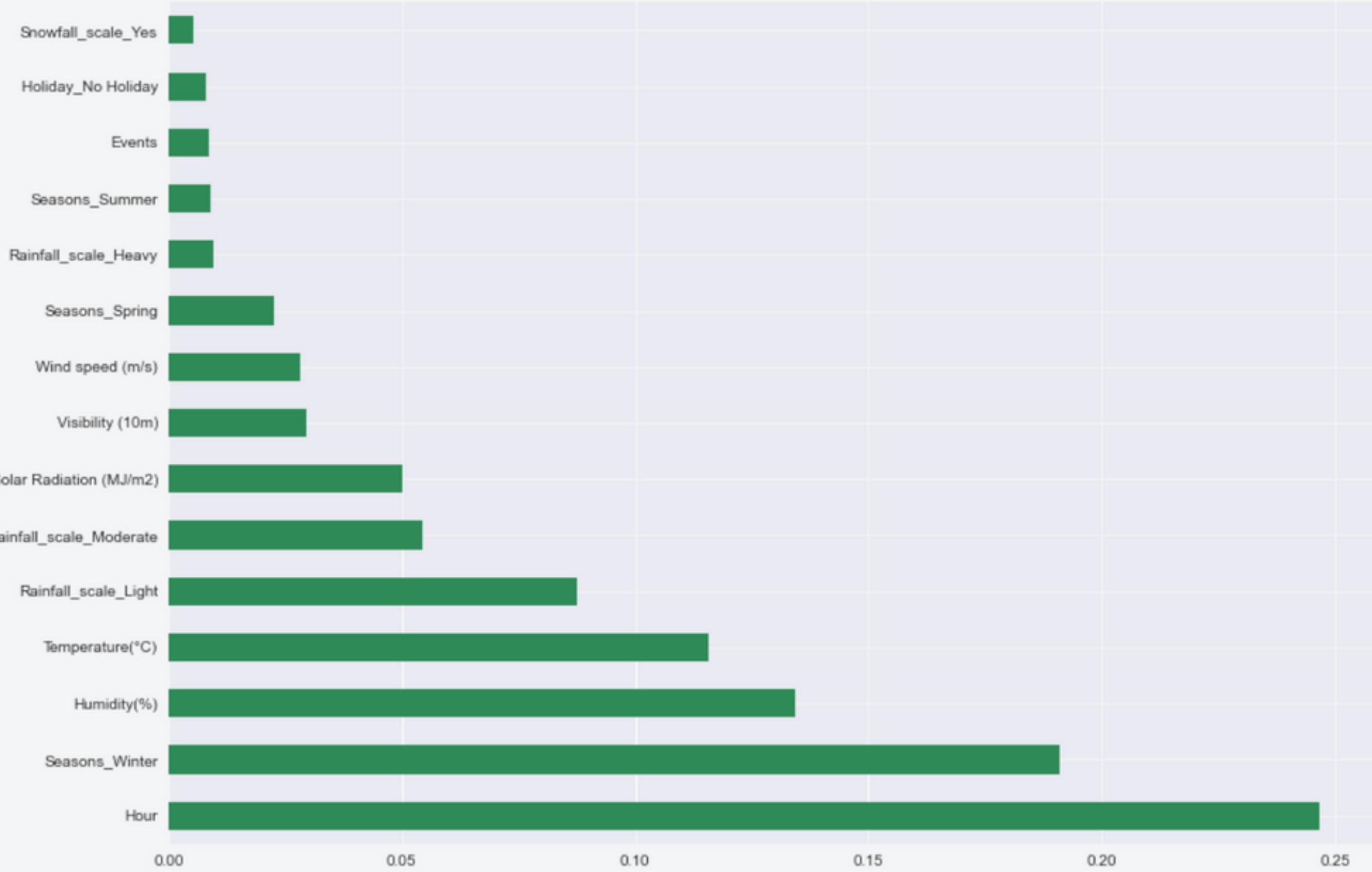
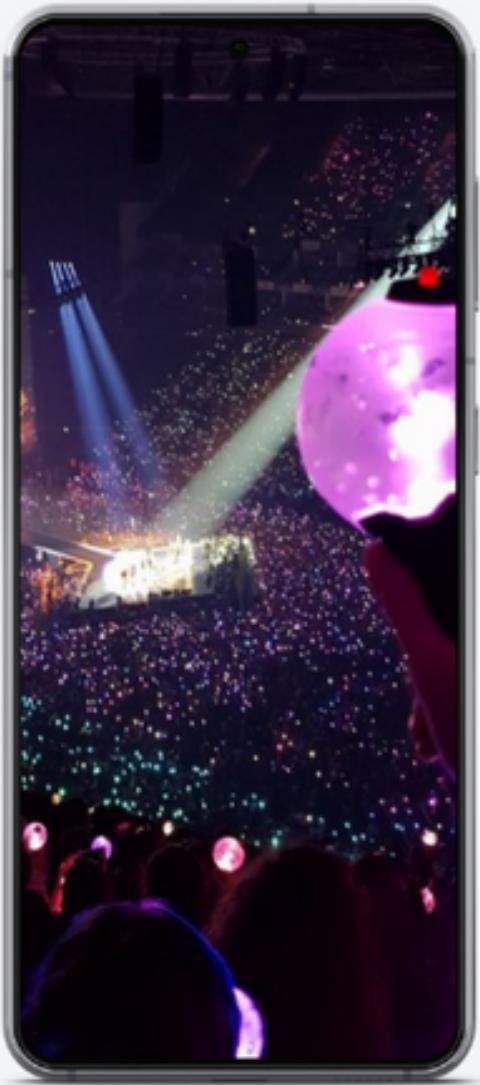
Datas de eventos:

- Dia de eleições em Seoul: 13 de Junho
- Protesto em Seoul: 4 de Maio
- Seoul International Fireworks Festival 2018: 6 de Outubro de 2018
- Seoul Lantern Festival 2018: 2 a 18 de Novembro de 2018
- BTS Concert: 25 e 26 de Agosto de 2018
- Seoul Comic Word: 3 e 4 de fevereiro de 2018
- Lunar New Year Fun: 17 de Fevereiro de 2018
- Santa Run 2018: 8 de Dezembro de 2018
- Monsta X concerto: 10 de Outubro 2018
- Concerto Twice: 19 e 20 de Maio 2018
- EXO Concerto: 13,14,15 de julho de 2018
- UltraKorea festival: 8, 9 e 10 de Junho de 2018
- Korean Liberation Day: 15 de Agosto de 2018
- ICON Concerto: 18 de agosto de 2018

Extra Tree Log com a variável Events

Resultados:

- Melhoria do R²: 87,34%
- Melhoria do MAPE: 30%



Previsão em dias de No Functioning Day

Aplicou-se o melhor modelo Extra Tree Log no dataset cujos dias de Functioning Day = No, logo o nº de alugueres de bicicletas seria nulo.

Rented Bike Count	Hour	Temperature(°C)	Humidity(%)
3144	0	14.4	82
3145	0	13.6	81
3146	0	12.7	80
3147	0	11.6	81
3148	0	10.2	83

Resultados:

