Applied Data Science with R Capstone Project

Laïd ATTIA

24 November 2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

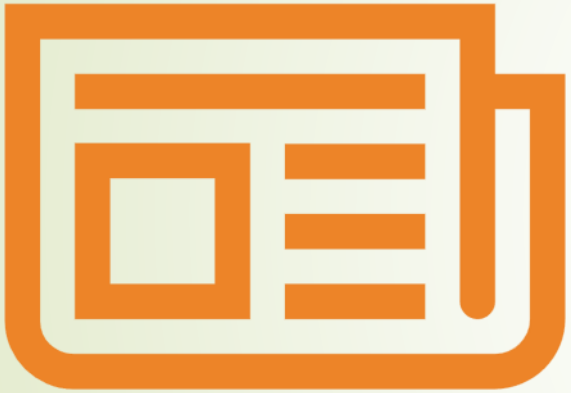
- Bicycle demand can be influenced by a variety of factors:
  - City/Location
  - Availability of Bikes
  - Seasonality
  - Outside Temperature
  - Time of Day
  - Events/Holidays
- A **Linear Regression Model** can be used to predict bicycle rental demand

Introduction

- The models replicated in this project intend to show how the weather can predict demand for ride-sharing bicycles in urban areas.

- This project incorporated the following data strategies:
  - Data Collection and Sources
  - Data Exploration and Analysis
  - Data Wrangling
  - Data Modeling
  - Interactive Dashboards

Methodology



- Perform data collection
- Perform data wrangling
- Perform exploratory data analysis (EDA) using SQL and visualization
- Perform predictive analysis using regression models
  - How to build the baseline model
  - How to improve the baseline model
- Build a R Shiny dashboard app

Methodology

Data collection

- Data was extracted from the HTML Table “Bicycle Sharing Systems” found on Wikipedia.
- For this project we also focused on using the Seoul Bike Sharing Demand Data Set which was designed for this purpose.

Data wrangling

- The data was processed by importing the table as a dataframe and then exporting the dataframe as a csv table for further analysis.

```
In [18]: df
```

Out[18]:

	country	city	name	system	operator	launched	discontinued	stations	bicycles	daily_ridership
0	Albania	Tirana	Ecovolis	NaN	NaN	March	NaN	8	200	NaN
1	Argentina	Mendoza	Metrobici	NaN	NaN		NaN	2	40	NaN
2	Argentina	San Lorenzo Santa Fe	Biciudad	Biciudad	NaN	November	NaN	8	80	NaN
3	Argentina	Buenos Aires	Ecobici	Serttel Brasil	Bike In Baires Consortium		NaN	400	4000	21917
4	Argentina	Rosario	Mi Bici Tu Bici	NaN	NaN	December	NaN	47	480	NaN
...	...	...	...	...	...	...	...	...	...	...
479	United States	Santa Monica California	Breeze Bike Share	Gen CycleHop and Social Bicycles	NaN	August	NaN	80	500	NaN
480	United States	Savannah Georgia	CAT Bike	Gen BCycle	NaN	January	NaN	2	16	NaN
481	United States	Seattle Washington	Pronto Cycle Share	D	Motivate	October	March	50	500	NaN
482	United States	Spartanburg South Carolina	Spartanburg BCycle	Gen BCycle	NaN		NaN	5	40	NaN
483	United States	St Paul	Yellow Bike Project	Gen w BikeCard	volunteers and city council		NaN	NaN	NaN	NaN

Then the table needed to be cleaned to replace empty values and to create uniformity.

EDA with SQL

- **Exploratory Data Analysis** tasks were performed using SQL.

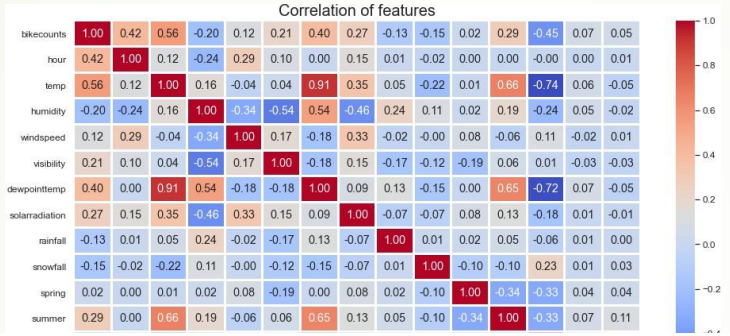
- Examples:
  - Determine how many records are in the **seoul_bike_sharing** dataset.
  - Determine how many hours had non-zero rented bike count.
  - Query the weather forecast for Seoul over the next 3 hours.
  - Find which seasons are included in the Seoul bike sharing dataset.
  - Find the first and last dates in the Seoul Bike Sharing dataset.
  - Determine which date and hour had the most bike rentals.
  - Determine the average hourly temperature and the average number of bike rentals per hour over each season.
  - Find the average hourly bike count during each season.
  - Consider the weather over each season.
  - Use an implicit join across the WORLD_CITIES and the BIKE_SHARING_SYSTEMS tables to determine the total number of bikes available in Seoul.
  - Find all cities with total bike counts between 15000 and 20000.

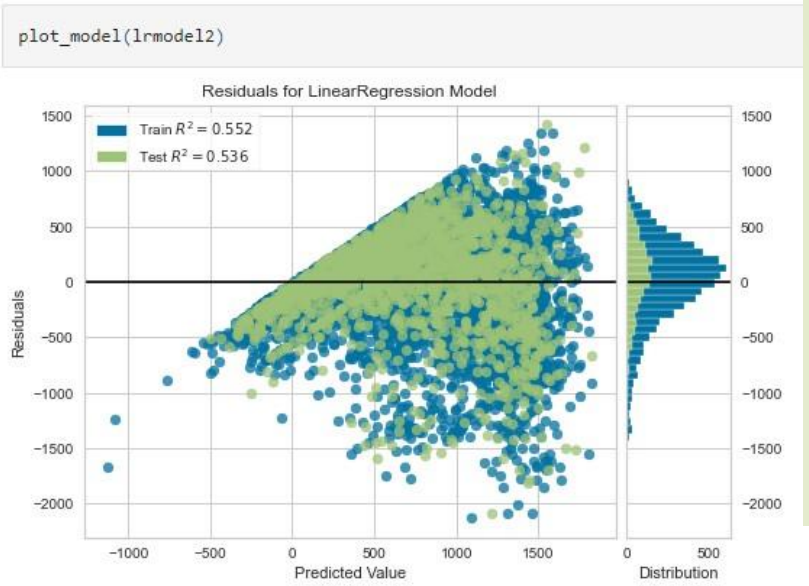
Attribute Information:

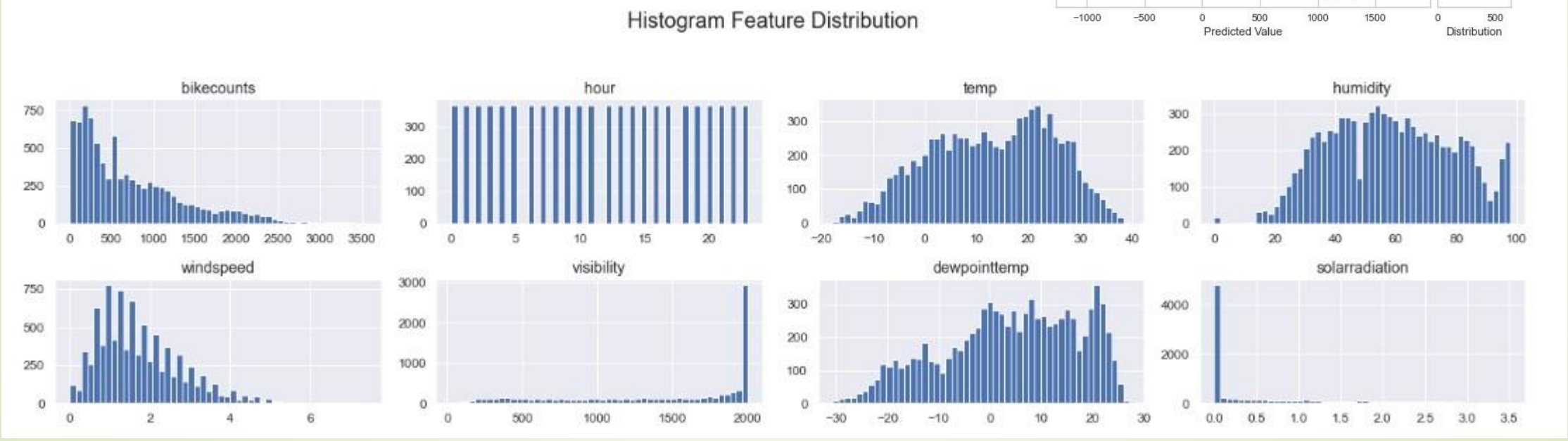
- Date : year-month-day
- Rented Bike count - Count of bikes rented at each hour
- Hour - Hour of he day
- Temperature-Temperature in Celsius
- Humidity - %
- Windspeed - m/s
- Visibility - 10m
- Dew point temperature - Celsius
- Solar radiation - MJ/m2
- Rainfall - mm
- Snowfall - cm
- Seasons - Winter, Spring, Summer, Autumn
- Holiday - Holiday/No holiday
- Functional Day - NoFunc(Non Functional Hours), Fun(Functional hours)

EDA with data visualization

- Charts were plotted using:
  - Histograms
  - Box Plots
  - Heatmaps
  - Plot Models

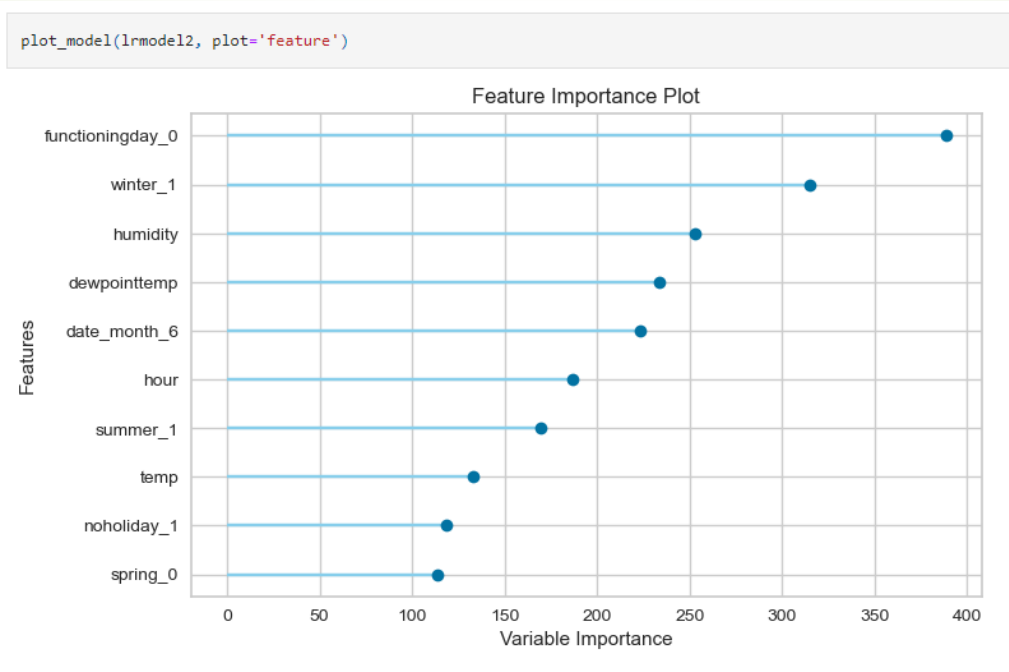






Predictive analysis

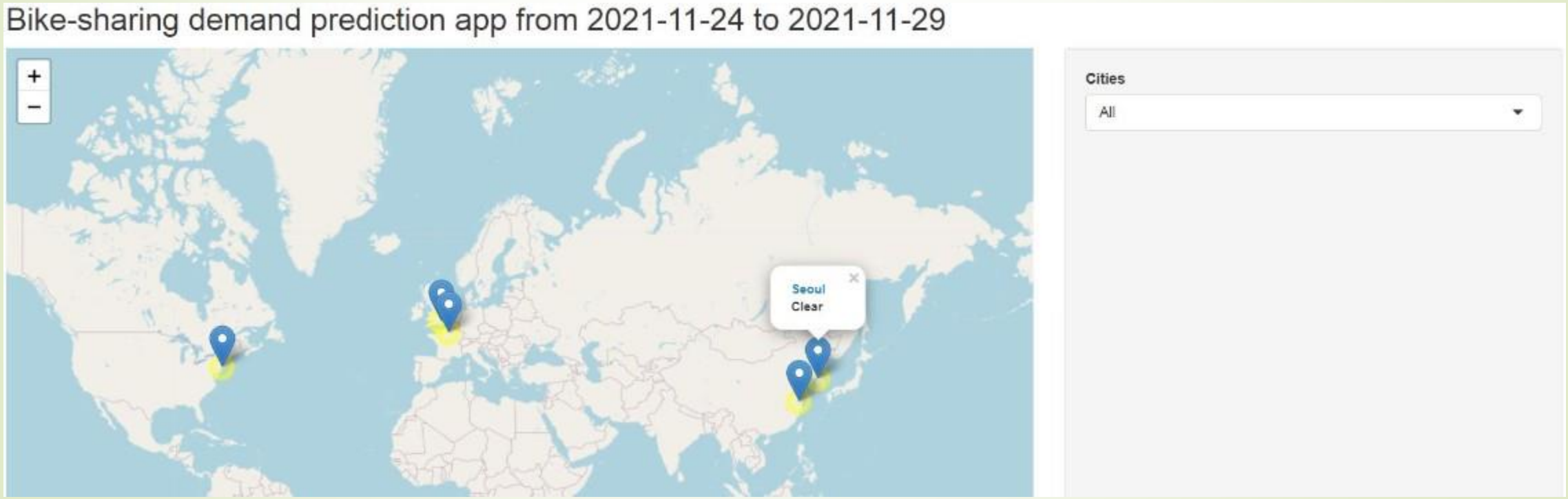
- I created various charts and models to find the best fit to represent the data. An example of this is by creating the Feature Importance Plot that helped guide which predictors might hold key insights.



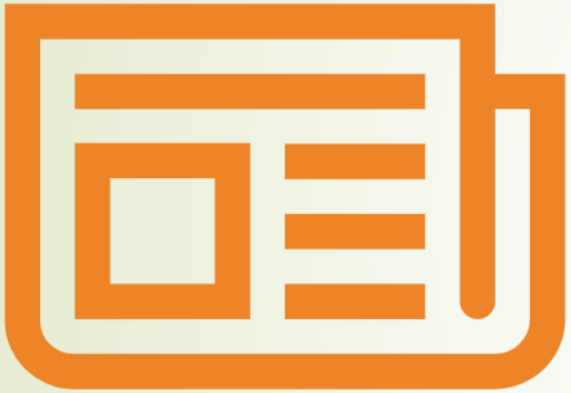
- The most important predictor of bicycle rental is **Functioning Day**.

Build a R Shiny dashboard

- The dashboard is made up of a World Map that allows you to select (input) to select a specific city, that will then visualize weather forecast data and predicted hourly bike-sharing demand for the following cities.



Results



- Exploratory data analysis results
- Predictive analysis results
- A dashboard demo in screenshots

EDA with SQL

Busiest bike rental times

- The busiest bike rental day was a **Tuesday late-afternoon at 6PM**. The conditions were: Summer, 2.9 wind speed, 57% humidity, 24 degrees Celsius.

```
seoul_bike_sharing["RENTED_BIKE_COUNT"].max()
```

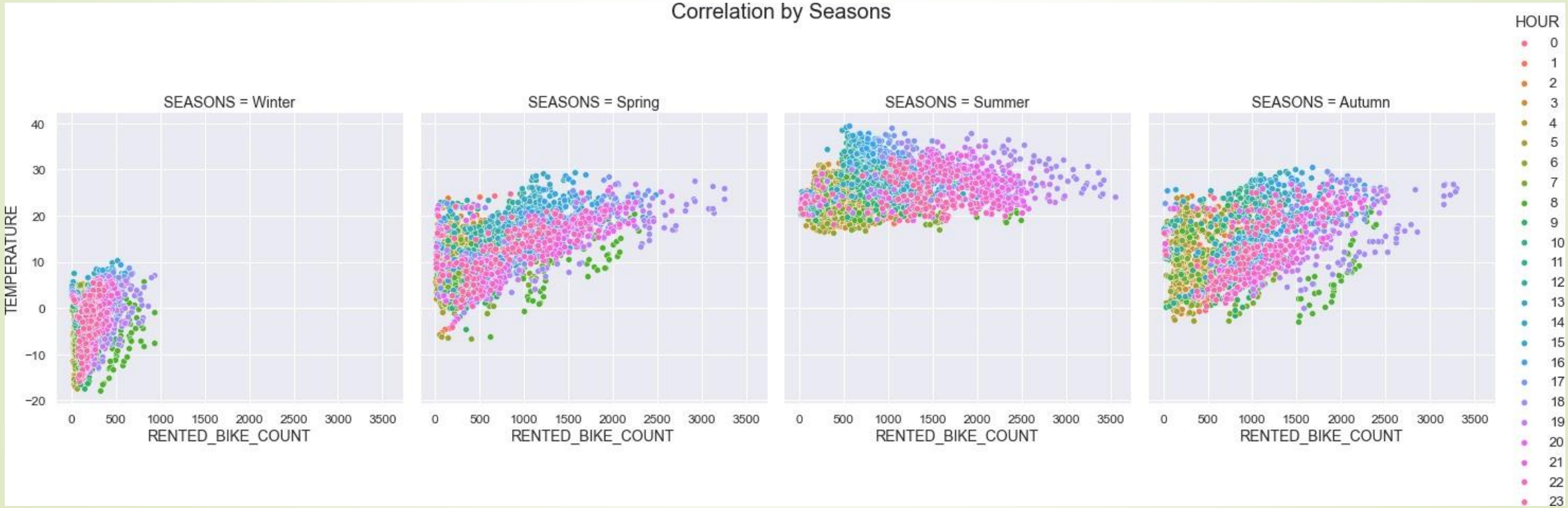
```
3556
```

```
seoul_bike_sharing[seoul_bike_sharing["RENTED_BIKE_COUNT"] == 3556]
```

	DATE	RENTED_BIKE_COUNT	HOUR	TEMPERATURE	HUMIDITY	WIND_SPEED	VISIBILITY	DEW_POINT_TEMPERATURE	SOLAR_RADIATION	RAINFALL	SNOWFALL	SEASONS
4770	2018-06-19	3556	18	24.1	57	2.9	1301	15.0	0.56	0.0	0.0	Summer

Hourly popularity and temperature by seasons

- There was a clear preference for bike rentals to occur throughout the **Afternoon** (11-20 HOUR) and during the **Summer** season.



Renter Seasonality

- The **Summer** season showed the clearest preference based on renter count by season.

	RENTED_BIKE_COUNT							
	count	mean	std	min	25%	50%	75%	max
SEASONS								
Autumn	1937.0	924.110480	617.547879	2.0	427.00	856.0	1271.0	3298.0
Spring	2160.0	746.254167	618.667962	2.0	225.00	599.0	1118.0	3251.0
Summer	2208.0	1034.073370	690.244759	9.0	526.75	905.5	1442.5	3556.0
Winter	2160.0	225.541204	150.372236	3.0	110.00	203.0	305.0	937.0

Weather Seasonality

- The weather was determined by overall temperature, humidity, wind speed, visibility, dew point, solar radiation, rainfall, and snowfall.
- The data showed a clear uptick in rentals when the weather was warm and temperate conditions outdoors, with a clear decline in high snow or rainfall.

SEASONS	RENTED_BIKE_COUNT	HOUR	TEMPERATURE	HUMIDITY	WIND_SPEED	VISIBILITY	DEW_POINT_TEMPERATURE	SOLAR_RADIATION	RAINFALL	SNOWFALL
Autumn	924.110480	11.530718	13.821580	59.044915	1.492101	1558.174497	5.150594	0.522783	0.117656	0.0635
Spring	746.254167	11.500000	13.021685	58.758333	1.857778	1240.911574	4.091389	0.680301	0.186944	0.0000
Summer	1034.073370	11.500000	26.587711	64.981431	1.609420	1501.745471	18.750136	0.761255	0.253487	0.0000
Winter	225.541204	11.500000	-2.540463	49.744907	1.922685	1445.987037	-12.416667	0.298181	0.032824	0.2475

Bike-sharing info in Seoul

- To do this we first had to Use an implicit join across the WORLD_CITIES and the BIKE_SHARING_SYSTEMS tables to determine the total number of bikes available in Seoul, plus the following city information about Seoul: CITY, COUNTRY, LAT, LON, POPULATION, in a single view.

```
world_cities[world_cities["CITY"] == "Seoul"]
```

	CITY	CITY_ASCII	LAT	LNG	COUNTRY	ISO2	ISO3	ADMIN_NAME	CAPITAL	POPULATION	ID
7	Seoul	Seoul	37.5833	127.0	Korea, South	KR	KOR	Seoul	primary	21794000.0	1.410836e+09

Cities similar to Seoul

- Here are the cities with comparable bike scale to Seoul's bike sharing system:

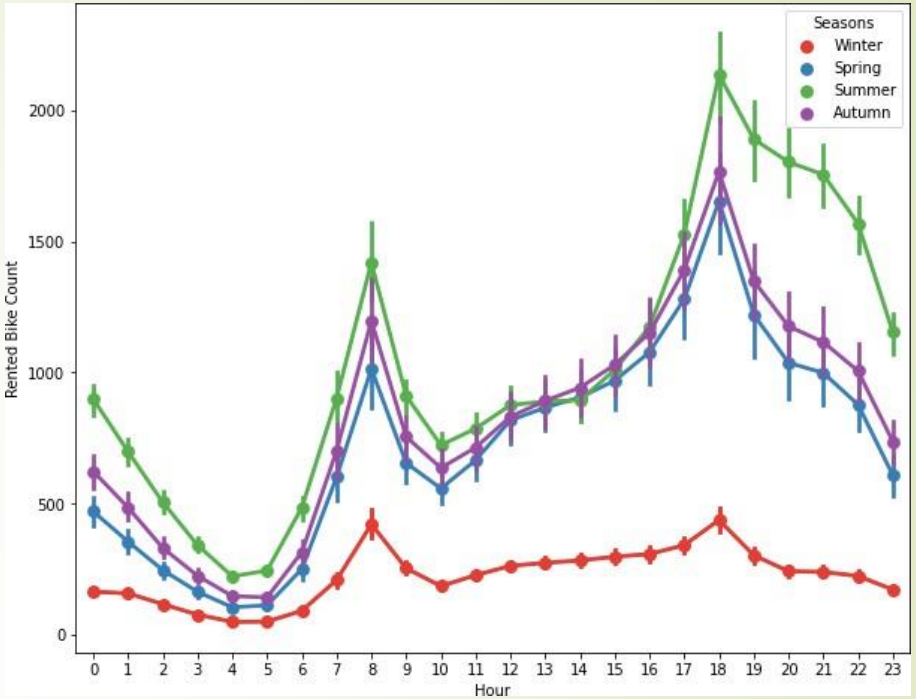
```
df_joined[(df_joined["BICYCLES"] >= 15000) & (df_joined["BICYCLES"] <= 20000)]
```

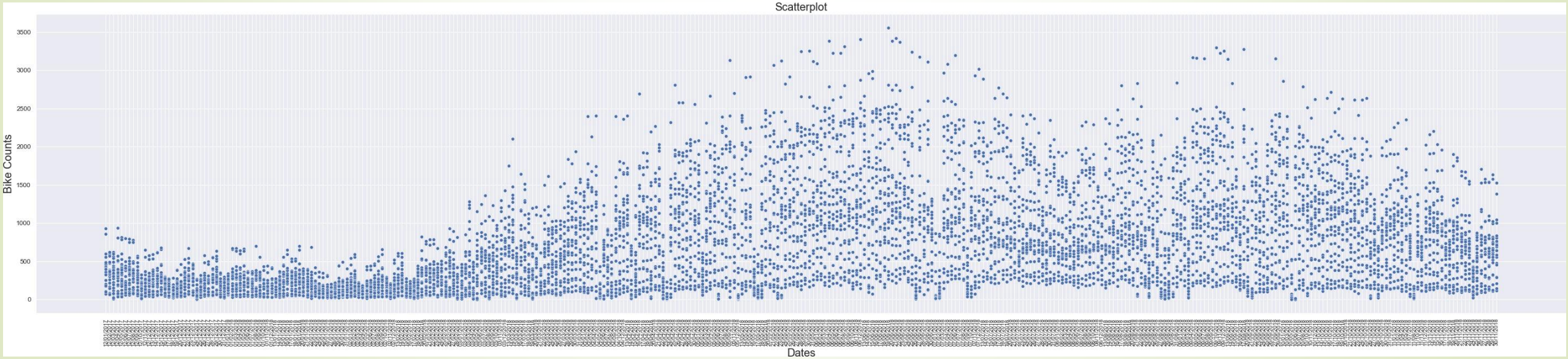
	CITY	CITY_ASCII	LAT	LNG	COUNTRY_x	ISO2	ISO3	ADMIN_NAME	CAPITAL	POPULATION	ID	COUNTRY_y	SYSTEM	BICYCLES
1	Shanghai	Shanghai	31.1667	121.4667	China	CN	CHN	Shanghai	admin	22120000.0	1.156074e+09	China	Forever Bicycle	19165.0
3	Seoul	Seoul	37.5833	127.0000	Korea, South	KR	KOR	Seoul	primary	21794000.0	1.410836e+09	South Korea	NaN	20000.0
6	Beijing	Beijing	39.9050	116.3914	China	CN	CHN	Beijing	primary	19433000.0	1.156229e+09	China	NaN	16000.0
25	Weifang	Weifang	36.7167	119.1000	China	CN	CHN	Shandong	NaN	9373000.0	1.156913e+09	China	NaN	20000.0
27	Ningbo	Ningbo	29.8750	121.5492	China	CN	CHN	Zhejiang	minor	7639000.0	1.156171e+09	China	NaN	15000.0
68	Zhuzhou	Zhuzhou	27.8407	113.1469	China	CN	CHN	Hunan	minor	3855609.0	1.156042e+09	China	Foshan Tianzhou	20000.0

EDA with Visualization

Bike rental vs. Date

- A scatterplot of RENTED_BIKE_COUNT vs. DATE may be too busy, but does show us the start of a pattern.

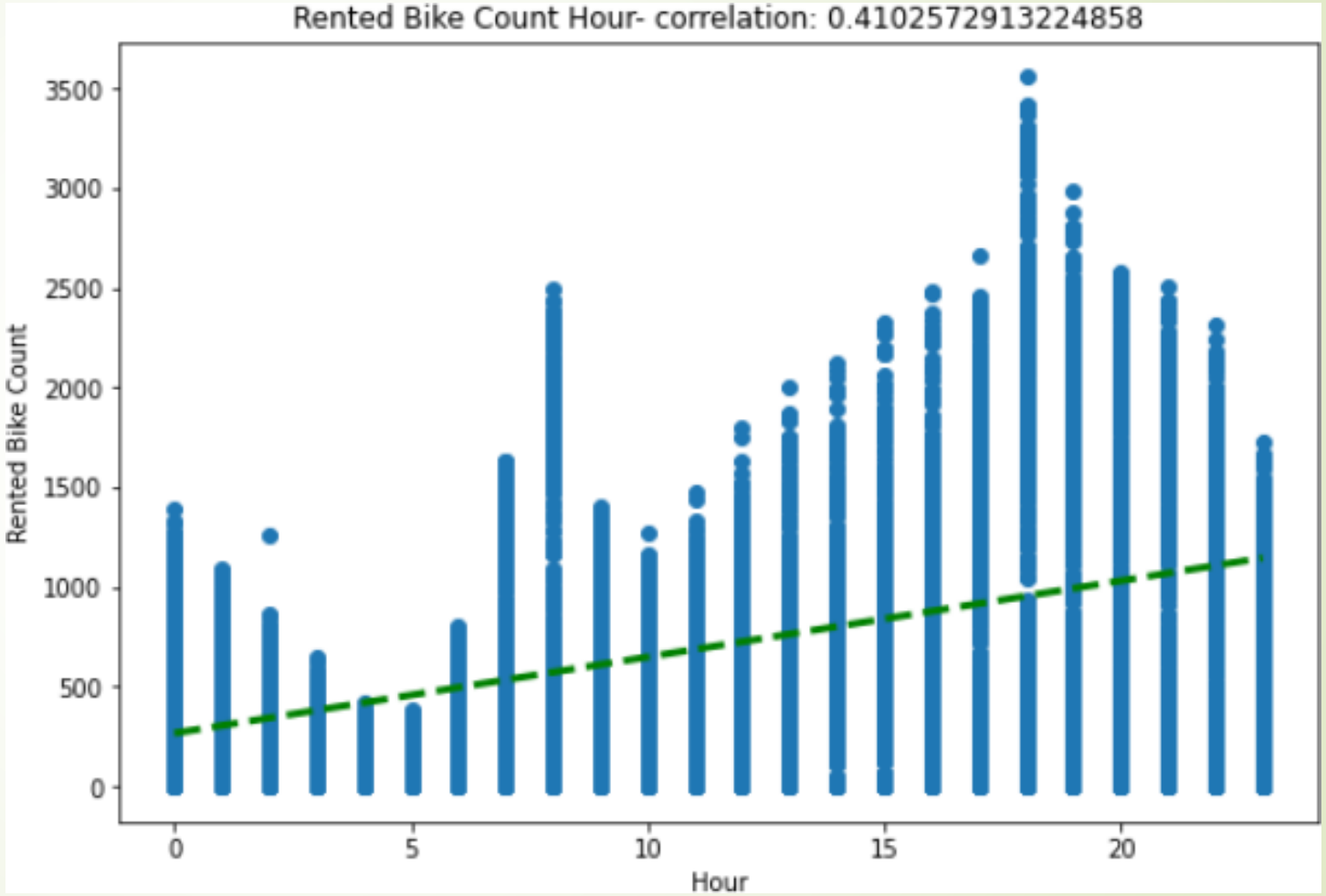




Bike rental vs. Datetime

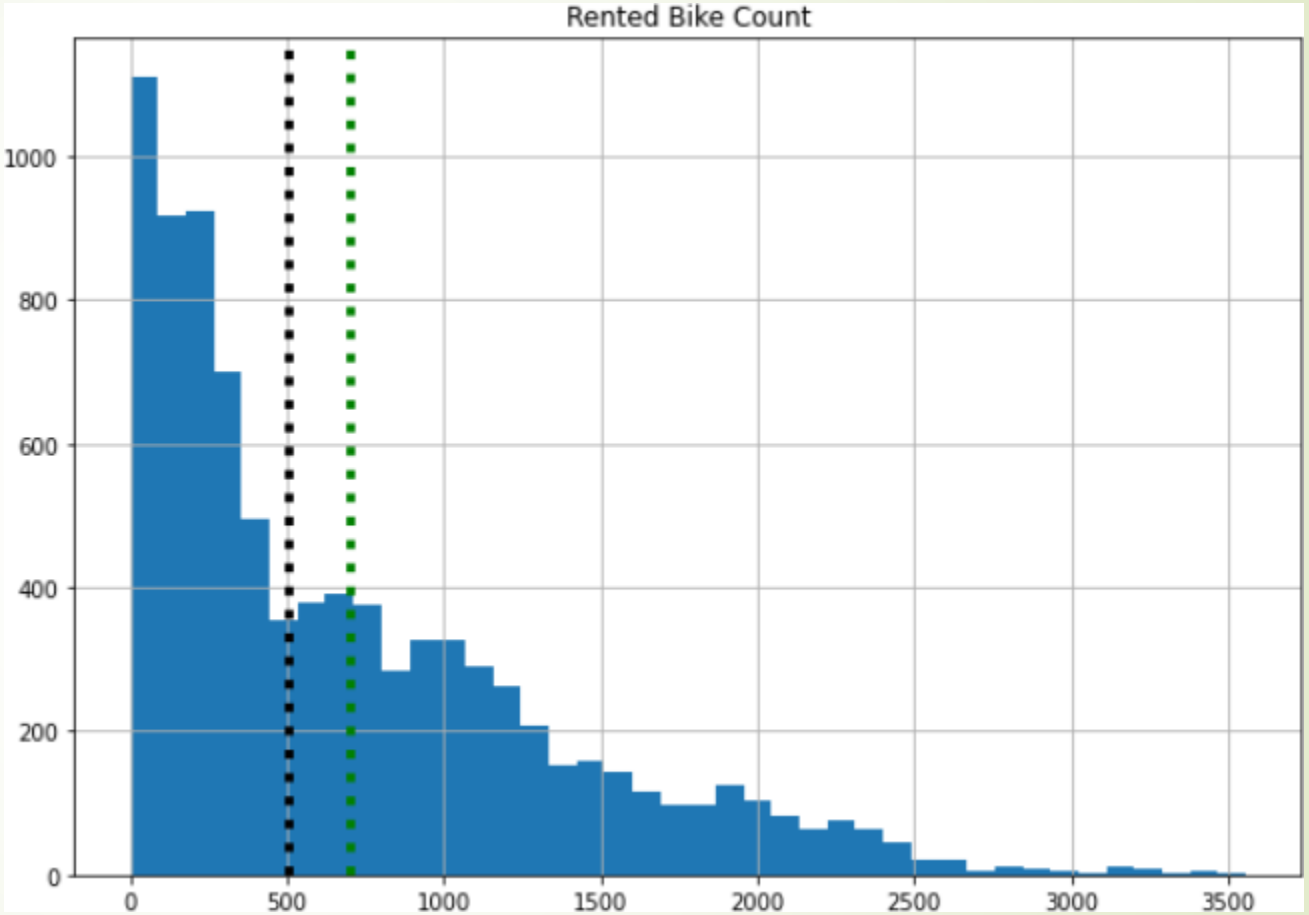
Show the same plot of
the RENTED_BIKE_COUNT time series,
but now add HOURS as the colour

Show the screenshot of the scatter plot with explanations



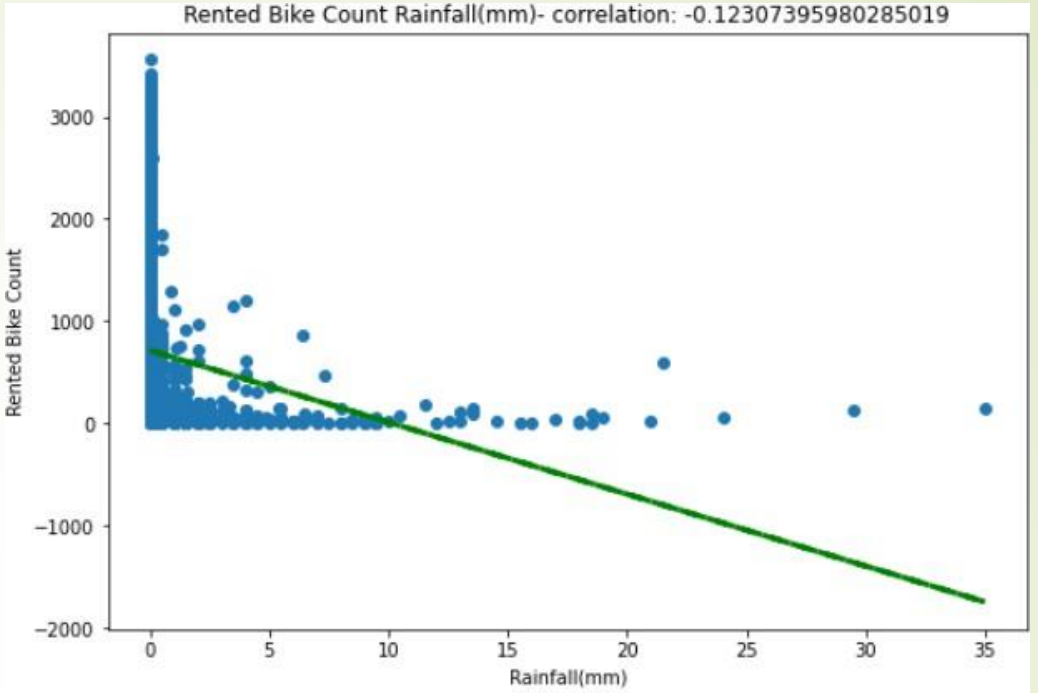
Bike rental histogram

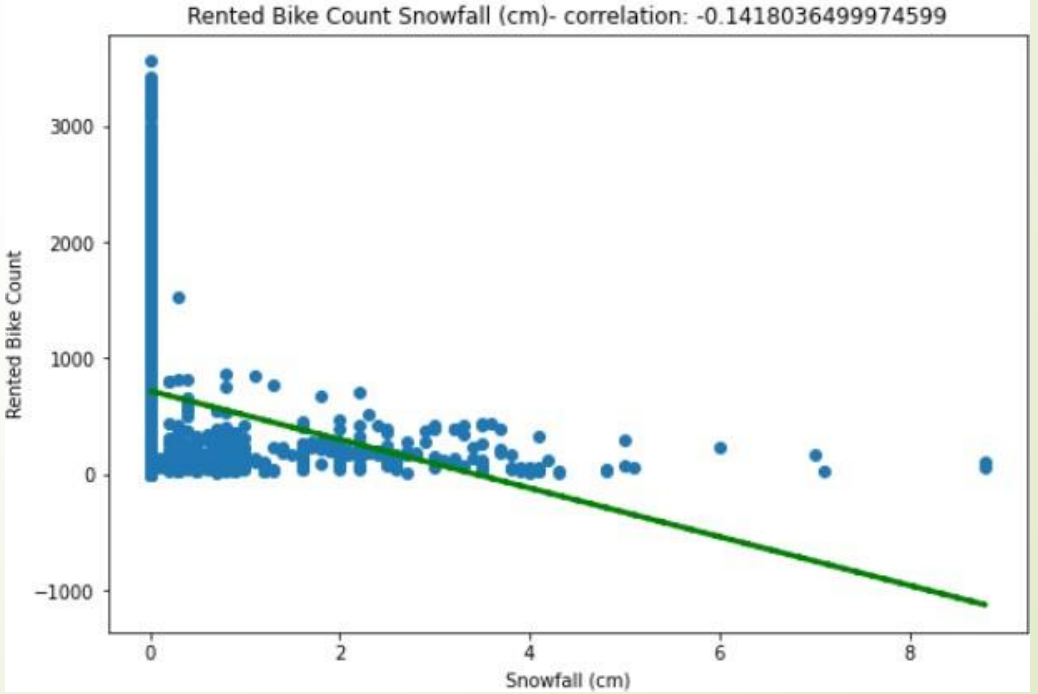
Show a histogram overlaid with a kernel density curve



Daily total rainfall and snowfall

There is a clear connection between when Rainfall/Snowfall occur, there will be less daily bicycle rentals.



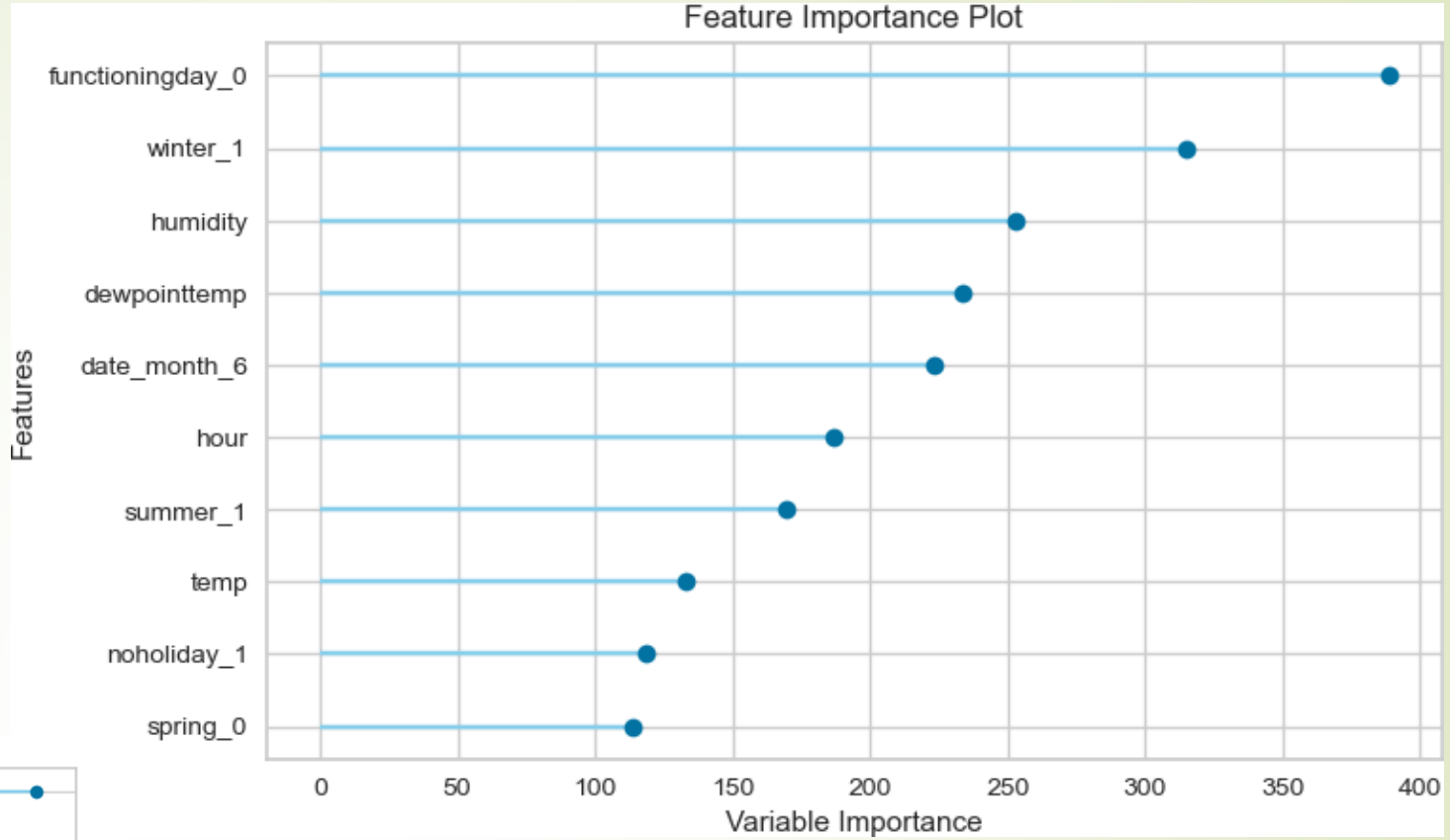


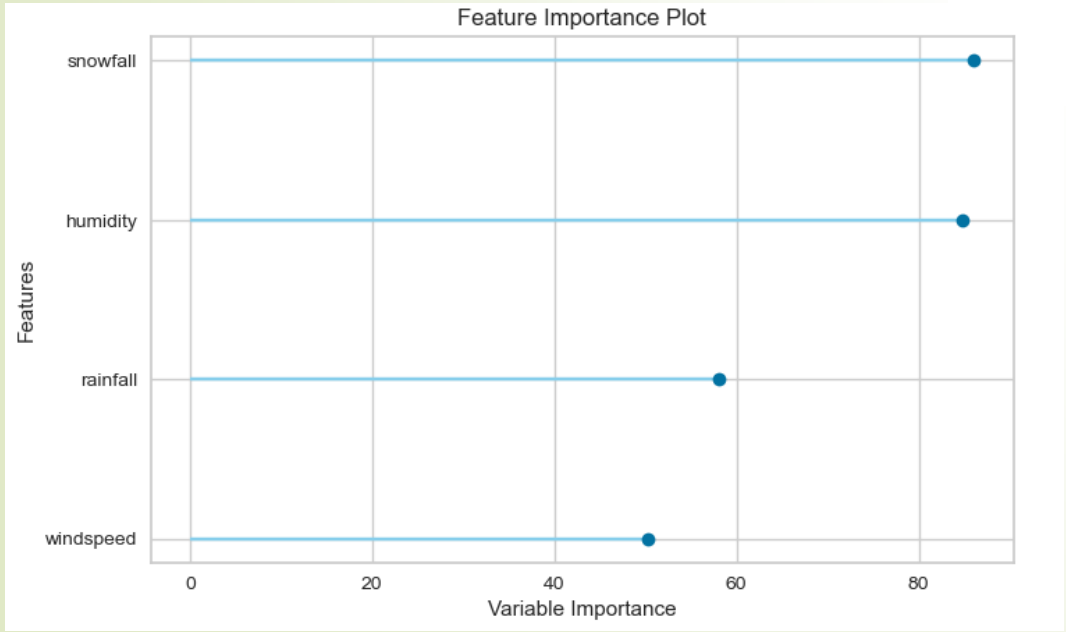
Predictive analysis

Ranked coefficients

A ranked coefficients bar chart helps show the different between specific variable, or all variables.

Some variables are show a higher chance of correlation than others, and sometimes researchers will want to separate specific variables for analysis.

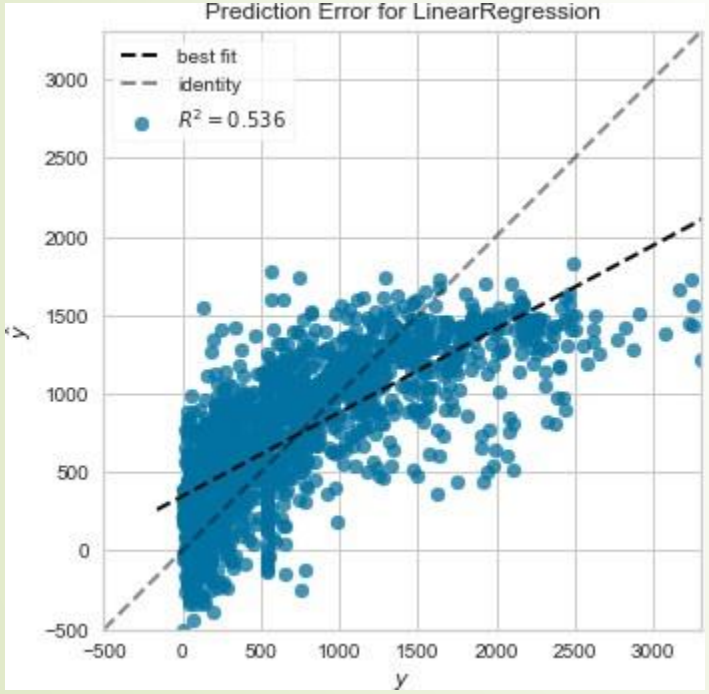




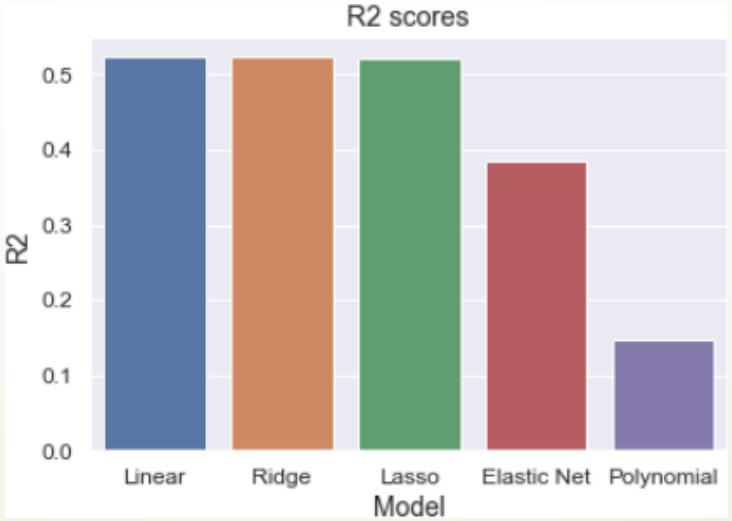
In this case, we wanted to study the baseline effects of weather over the other variables.

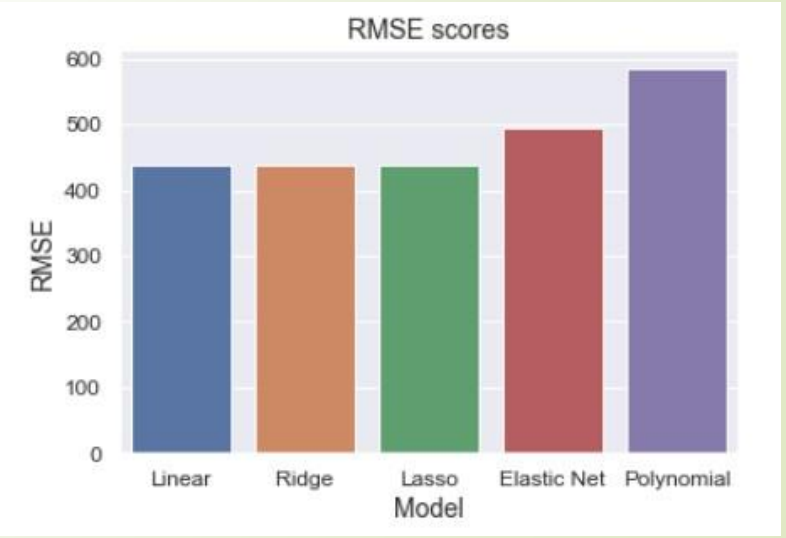
Model evaluation

Built at least 5 different models using polynomial terms, interaction terms, and regularizations.



	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	317.1563	180032.8750	424.3028	0.5386	0.9043	1.3594
1	307.2601	163261.6094	404.0565	0.5518	0.8835	1.3457
2	318.2002	180363.7344	424.6925	0.5567	0.9860	1.7550
3	318.9995	189241.0000	435.0184	0.5626	0.9107	1.5847
4	311.6680	167021.3125	408.6824	0.5276	0.8937	1.5680
5	305.2763	171933.0625	414.6481	0.5392	0.9245	1.6530
6	328.0816	197317.3594	444.2042	0.5637	0.9522	1.7299
7	330.1945	193396.0469	439.7682	0.5044	0.9320	1.8311
8	320.6869	186026.3594	431.3077	0.5396	0.9060	1.5617
9	323.8840	182132.9375	426.7704	0.5683	0.9208	1.6885
Mean	318.1407	181072.6297	425.3451	0.5453	0.9214	1.6077
SD	7.8171	10525.8306	12.4163	0.0185	0.0285	0.1515





Find the best performing model

- The best performing model was the **Ridge Regression Model**.

- Model formula:

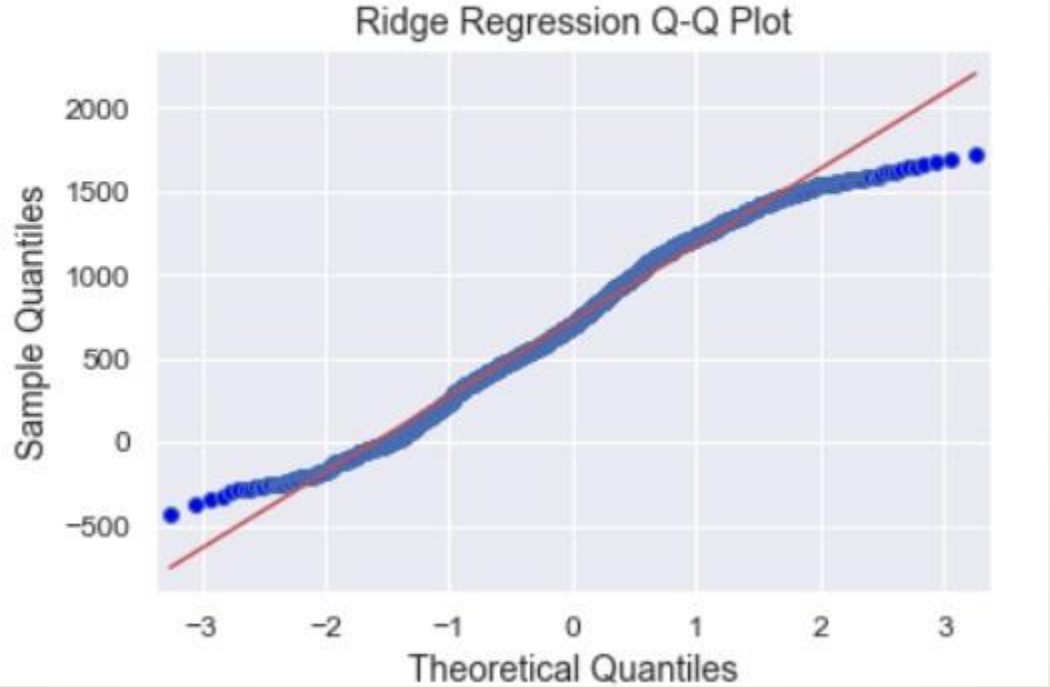
```
Ridge(alpha=1.0, copy_X=True, fit_intercept=True, max_iter=None,
      normalize=False, random_state=0, solver='auto', tol=0.001)
```

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
lr	Linear Regression	283.0143	142660.9000	377.6294	0.6422	0.8924	1.5912	1.1580
ridge	Ridge Regression	282.9418	142645.2344	377.6091	0.6422	0.8941	1.5886	0.0140
lar	Least Angle Regression	283.0429	142767.5492	377.7662	0.6420	0.8922	1.5900	0.0160
lasso	Lasso Regression	283.3254	143644.9844	378.9253	0.6398	0.8928	1.5563	0.0400
llar	Lasso Least Angle Regression	354.6801	227398.6736	476.6986	0.4301	0.9387	1.9962	0.0160
en	Elastic Net	359.0337	235343.4406	484.8899	0.4107	0.9293	1.7635	0.0160

Display Q-Q plot of best model

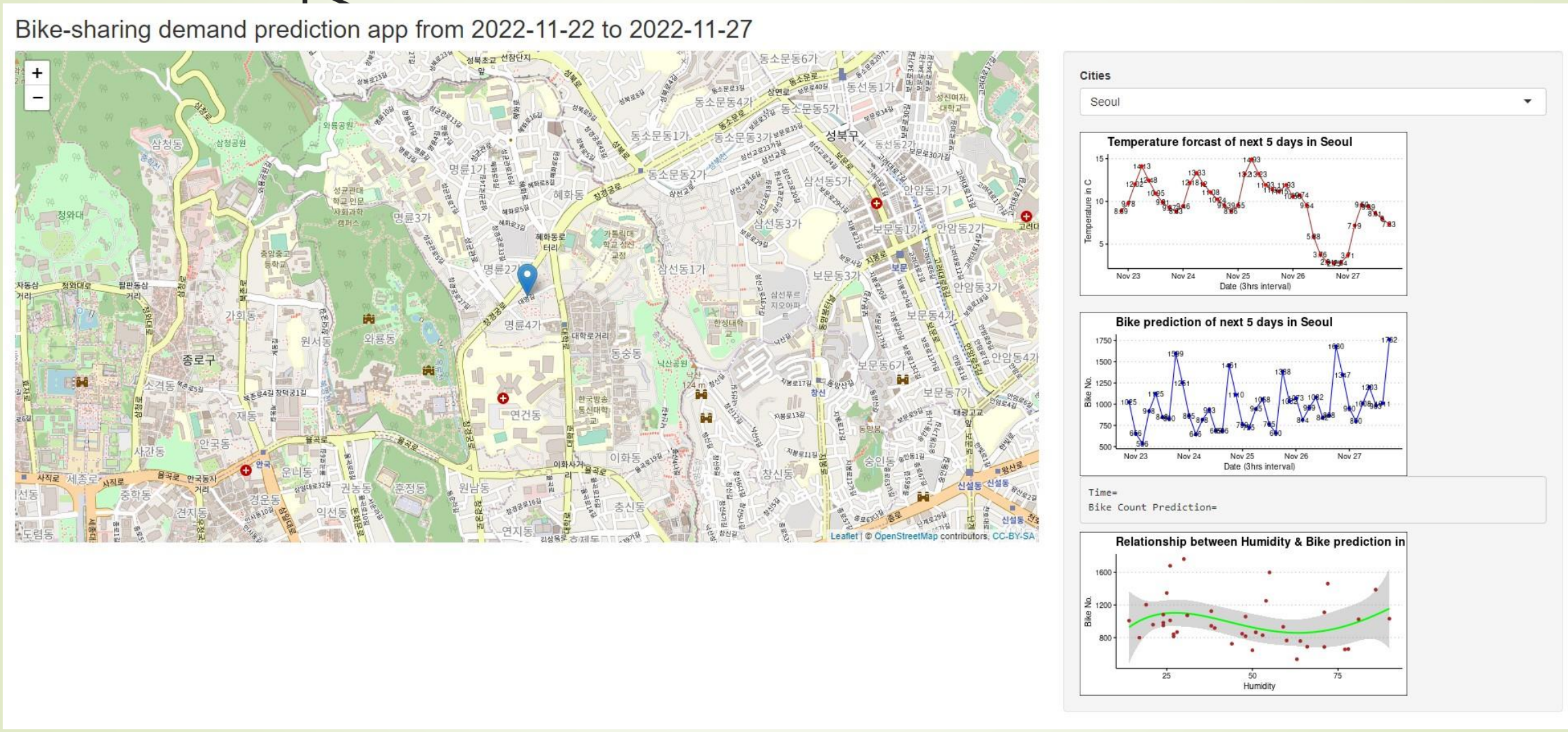
- The best performing model was the **Ridge Regression Model**.
- Model formula:

```
sm.qqplot(ridge_pred, line='s')
plt.title("Ridge Regression Q-Q Plot")
plt.show()
```



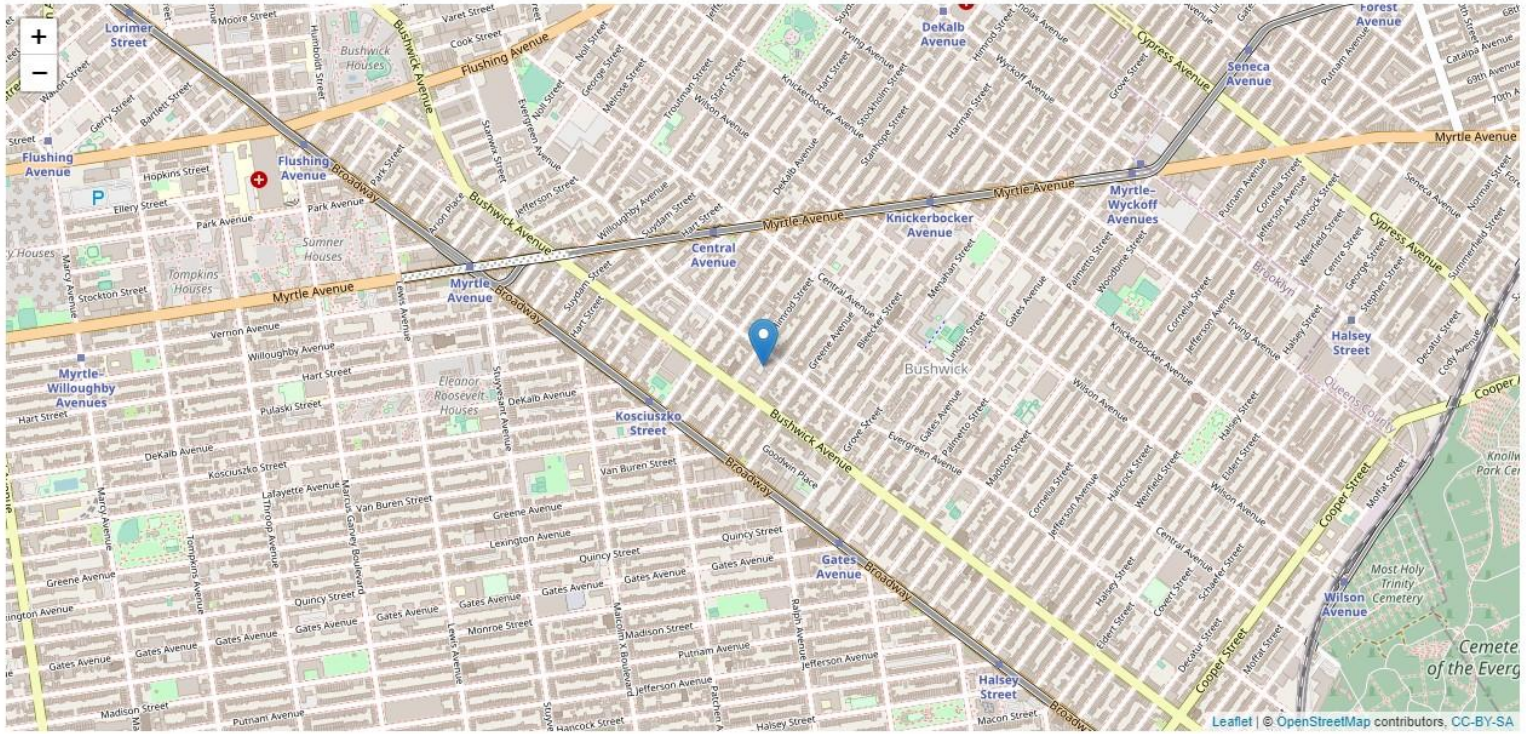
Dashboard

<Dashboard screenshot 1>



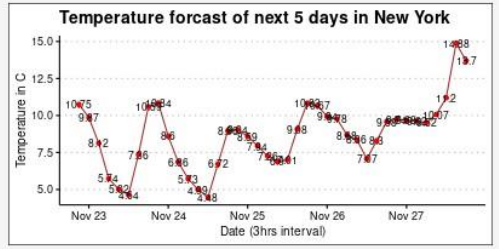
<Dashboard screenshot>

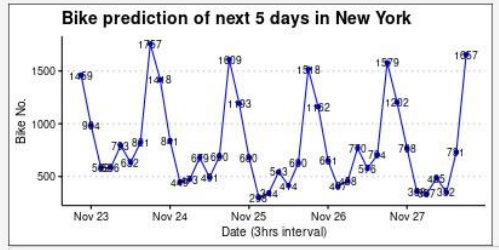
Bike-sharing demand prediction app from 2022-11-22 to 2022-11-27



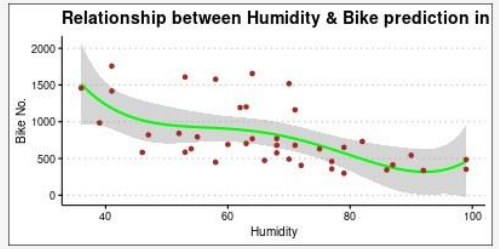
Cities

New York



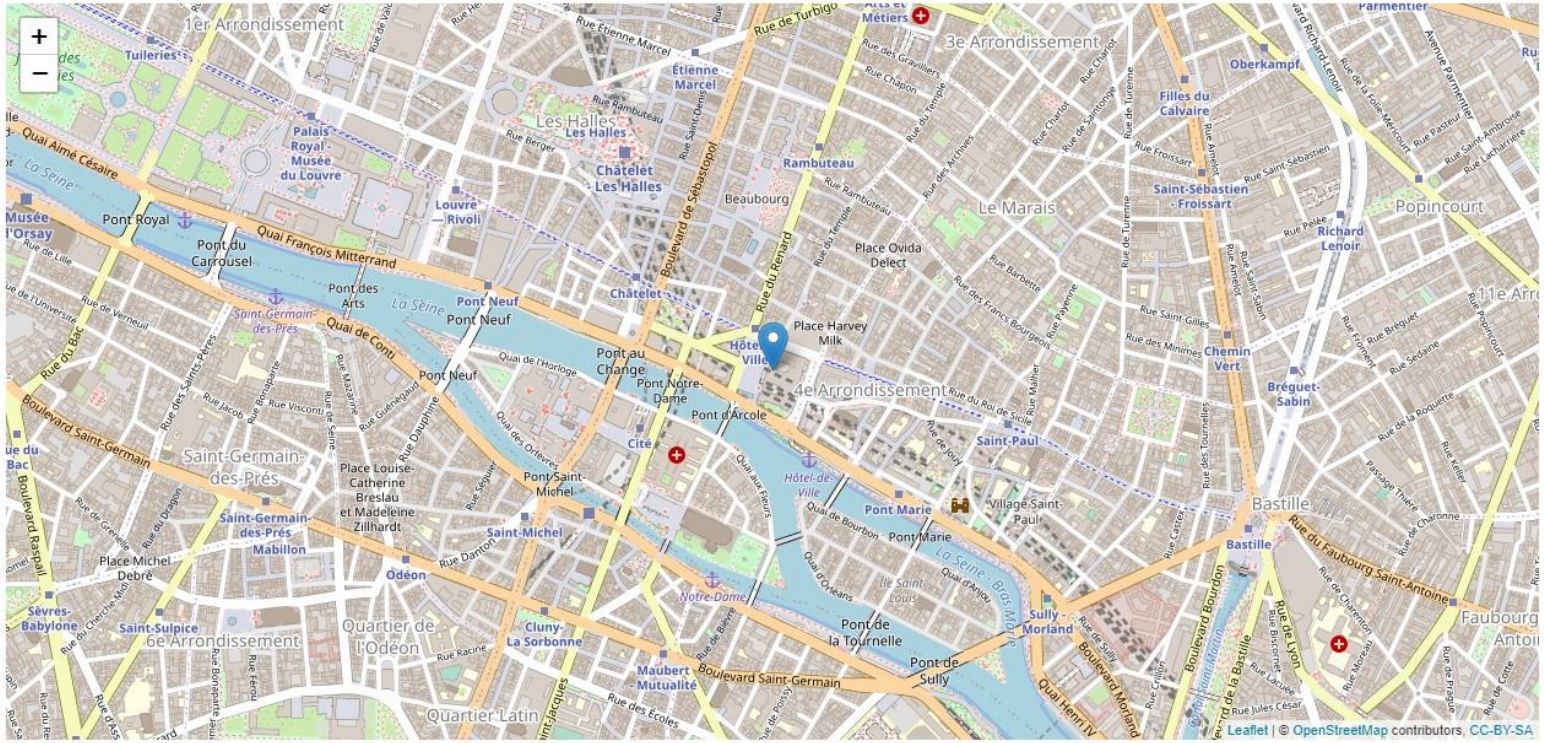


Time=
Bike Count Prediction=



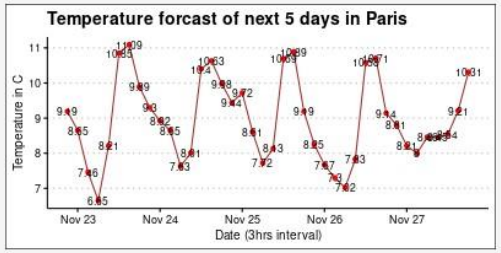
<Dashboard screenshot 2>

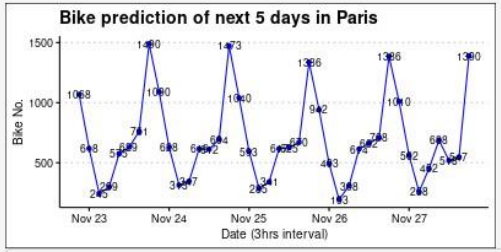
Bike-sharing demand prediction app from 2022-11-22 to 2022-11-27



Cities

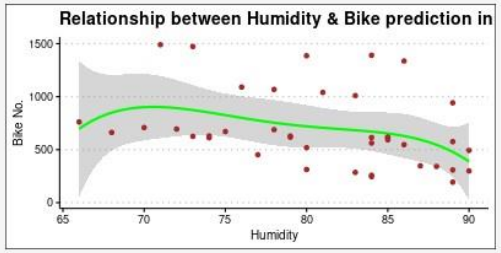
Paris





Time=

Bike Count Prediction=



CONCLUSION

- There are clear connections that can be made using common data analysis techniques such as data wrangling, regression models, and interactive visualizations.
- The insights used here showed us insights into the data including which days, times, and weather conditions lead to a clear increase in bike rentals between major cities across the world.

APPENDIX

- Key Webscrapping Snippets
- OpenWeather API Snippets
- Key Data Wrangling Code Using Regular Expressions
- Key Data Wranglign Codes Using Diplyr
- All Other Required SQL Queries

```
url = "https://en.wikipedia.org/wiki/List_of_bicycle-sharing_systems"
# Get the root HTML node by calling the `read_html()` method with URL

df = pd.read_html(url)
```

```
# URL for Current Weather API
current_weather_url = 'https://api.openweathermap.org/data/2.5/weather?'

df_current_weather = pd.DataFrame()
```

```
df = pd.read_csv("raw_bike_sharing_systems.csv")
```

```
df['city'] = df['city'].str.replace(r"[^a-zA-Z ]","")
df['city']
```

```
df.to_csv("bikesharing.csv", index=False)
```

```
df = pd.read_csv("raw_seoul_bike_sharing.csv",parse_dates=['Date'])  df.describe(include='all')
```

```
df.hist(bins=50, figsize=(20,10))
plt.suptitle('Histogram Feature Distribution', x=0.5, y=1.02, ha='center', fontsize=20)
plt.tight_layout()
plt.show()
```

```
df.boxplot(figsize=(20,10))
plt.suptitle('BoxPlots Feature Distribution', x=0.5, y=1.02, ha='center', fontsize=20)
plt.tight_layout()
plt.show()
```

```
df2_cat = pd.get_dummies(data=df_cat,drop_first=True)  df3 = pd.concat([df_num,df2_cat],axis=1)
```