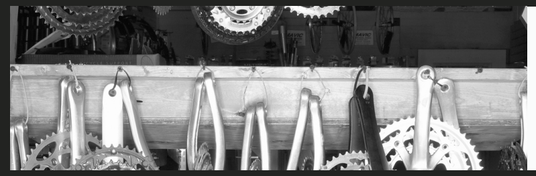
Seoul Bike Sharing

PREPARED BY MOHAMED-HOUSSEM REZGUI, YAHYA EL OUDOUNI & SASCHA CAUCHON

Outline



PROJECT DESCRIPTION

THE DATA

CLEANING AND MANIPULATION

SUMMARY OF DATA ANALYSIS

FEATURE ENGINEERING

MODELS

Project Description

BIKE RENTAL PROGRAMS

What exactly are they?

Rent out bicycles for short periods of time, usually for a few hours.

BIKE RENTAL PROGRAMS

Multiple benefits

- Healthier people & cities
- Green & sustainable way of travel
- Reduce urban traffic

BIKE RENTAL PROGRAMS

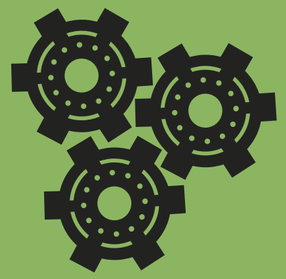
What problem are we solving?

- Providing the city with a stable supply of rental bikes
- Predict the number if bikes required each hour

The Data

The dataset contains weather information , the number of bikes rented per hour and date information between the years 2017-2018.

1 Target and 13 features.



8760 rows with 14 columns.

THE TARGET

This is the value that we want to predict based on the different features available in the dataset:

- Rented Bike Count

THE FEATURES

TIME

- Date
- Hour
- Seasons
- Holiday
- Functional Day

WEATHER

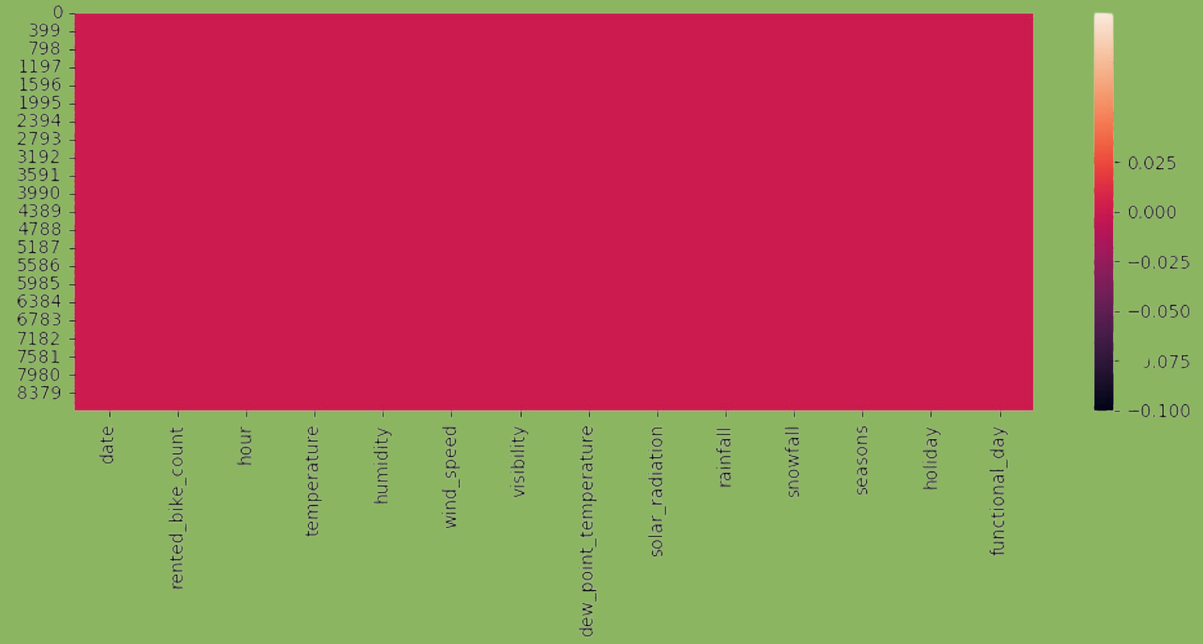
- Temperature
- Humidity
- Windspeed
- Visibility
- Dew Point Temperature
- Solar Radiation
- Rainfall
- Snowfall

Cleaning & manipulation

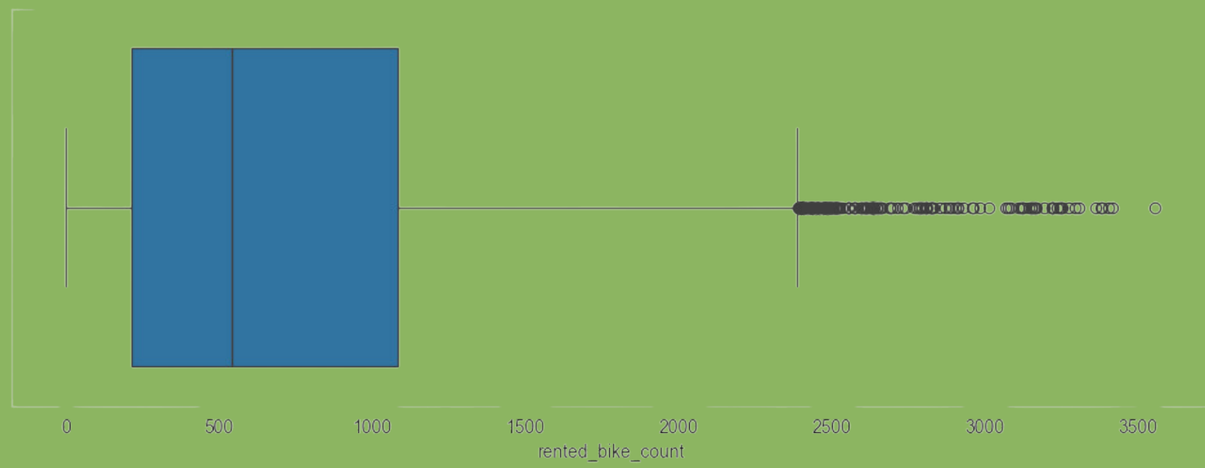
- Downloaded dataset & stored it with appropriate filename
- Imported the necessary Python Libraries and loaded the dataset
- Understanding the data by looking at the rows, columns, shape, types,...
- Preprocessed the data: changing column names, null values, convertion, outliers,...
- Encoding: transformed categorical values of relevant features into numerical ones. (eg: Holiday/No holiday, Seasons)
- Feature extraction: added new columns (eg: date column to year, month, day)

Cleaning & manipulation

NULL VALUES

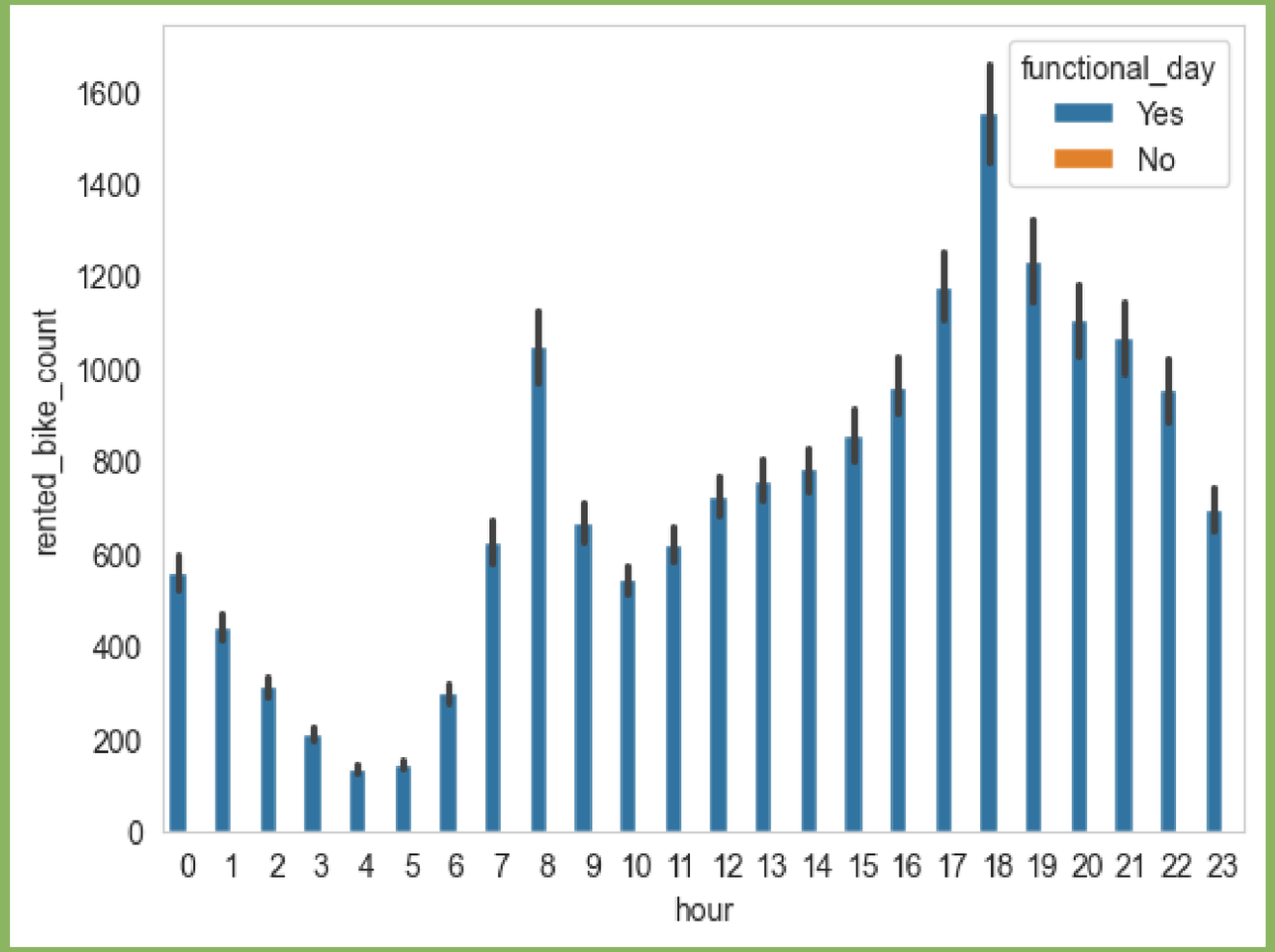


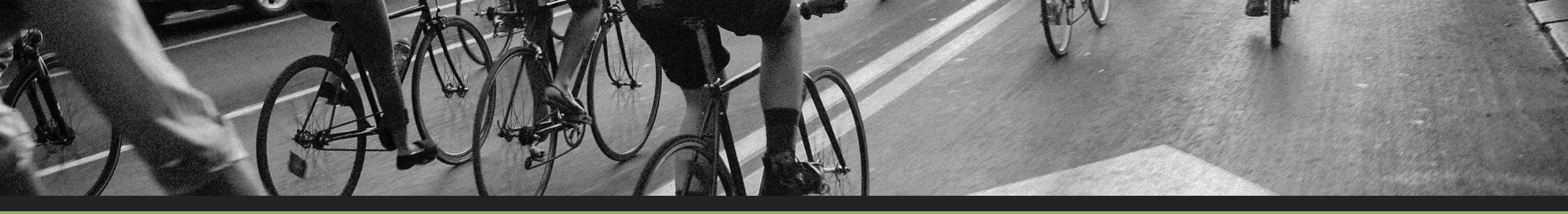
OUTLIERS



Cleaning & manipulation

FEATURE EXTRACTION



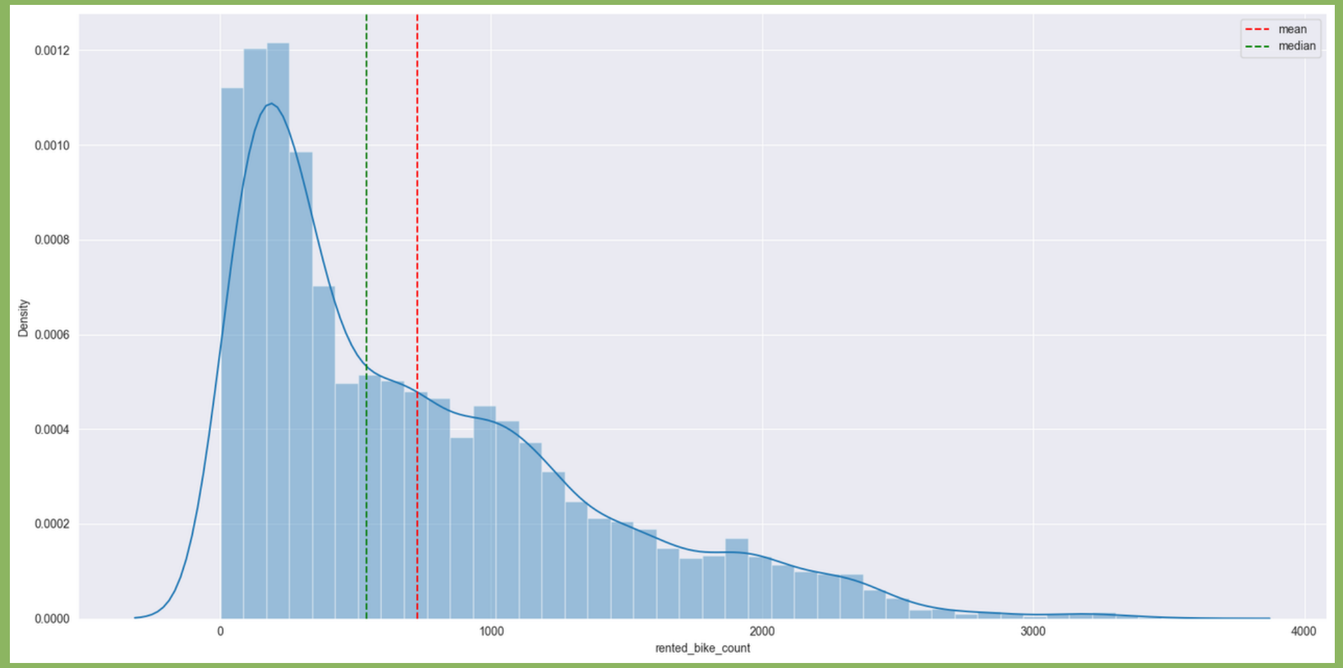


Summary of Data Analysis

Distribution of numerical features

RIGHT SKEWED DISTRIBUTION

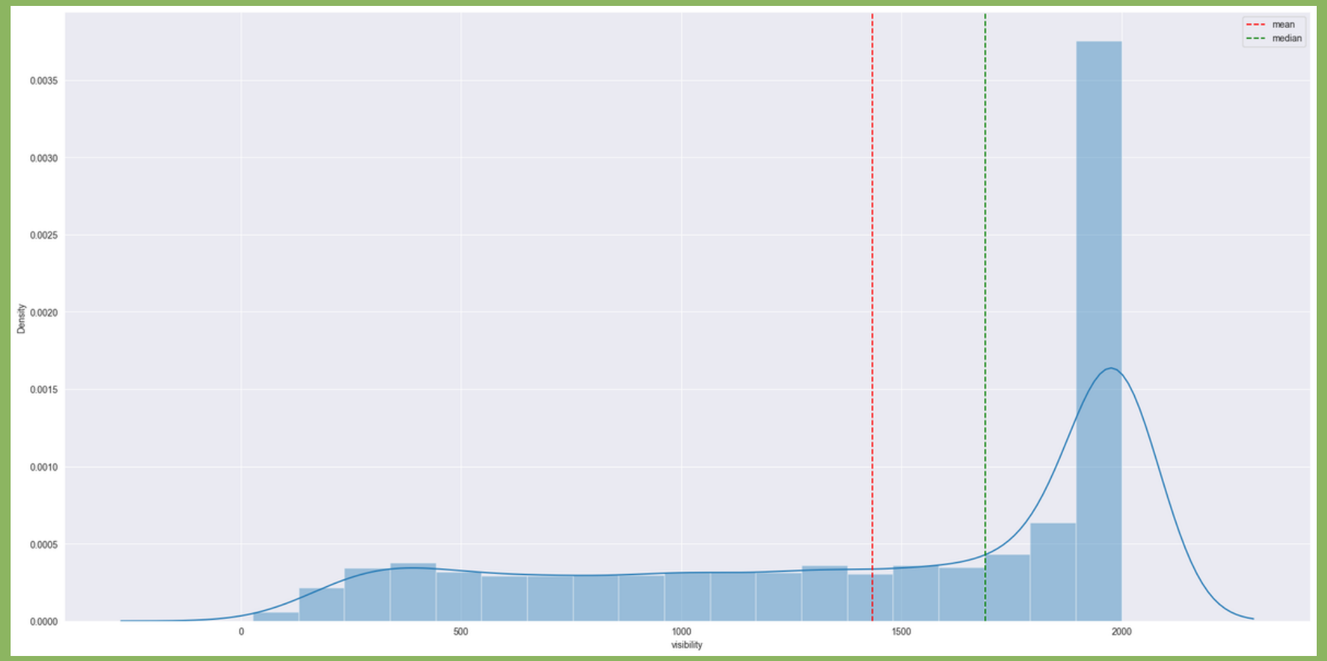
- Rented Bike Count, Wind Speed, Solar Radiation, Rainfall,...



Distribution of numerical features

LEFT SKEWED DISTRIBUTION

- Visibility, year



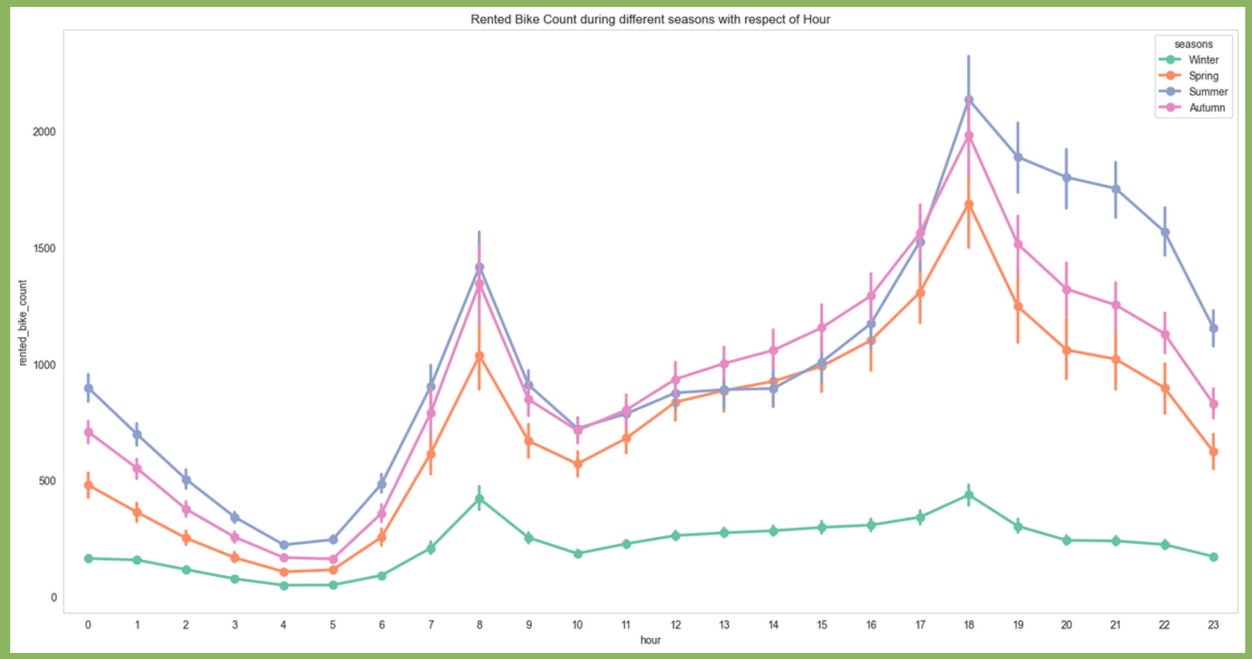
When do people rent the most bikes?

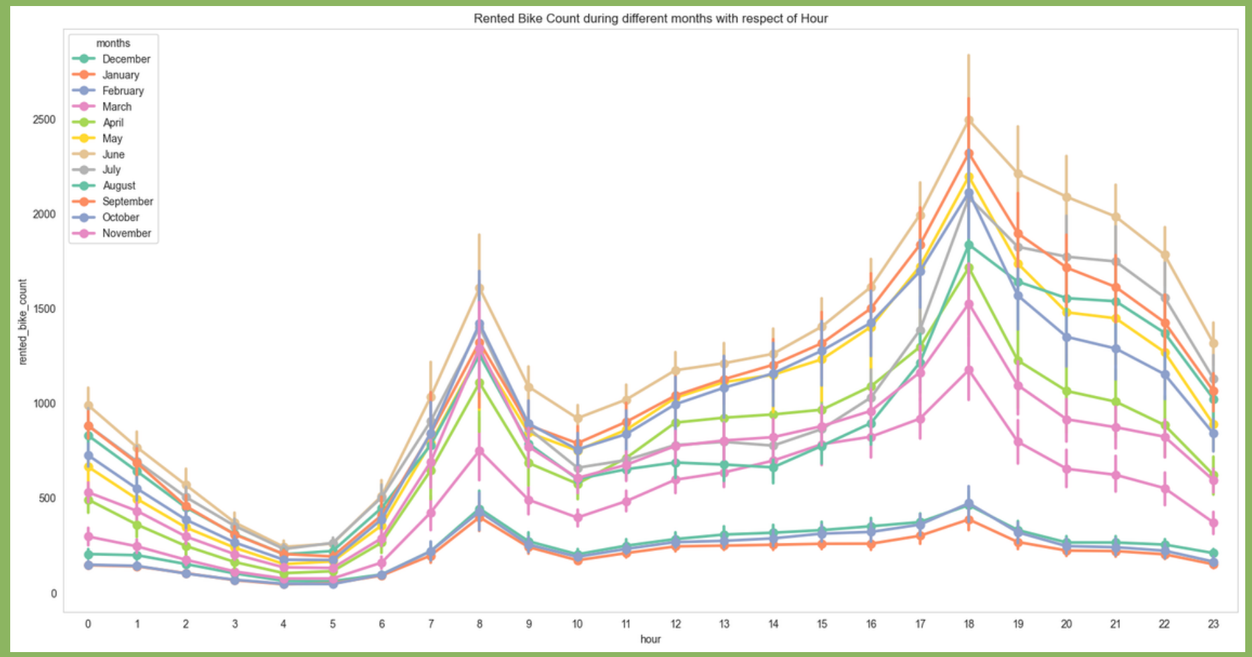
SEASONS

- Demand is low during the winter
- Demand is high during the Autumn/Summer seasons

MONTHS

- Demand is low during the months of December, January, and February





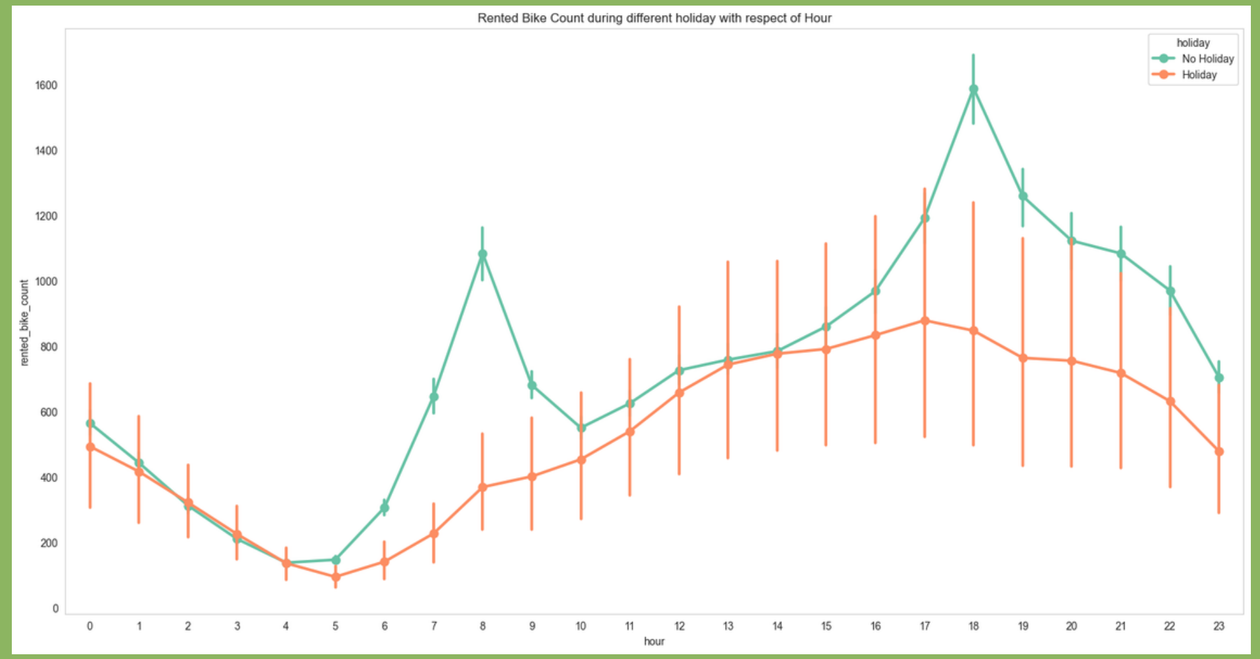
When do people rent the most bikes?

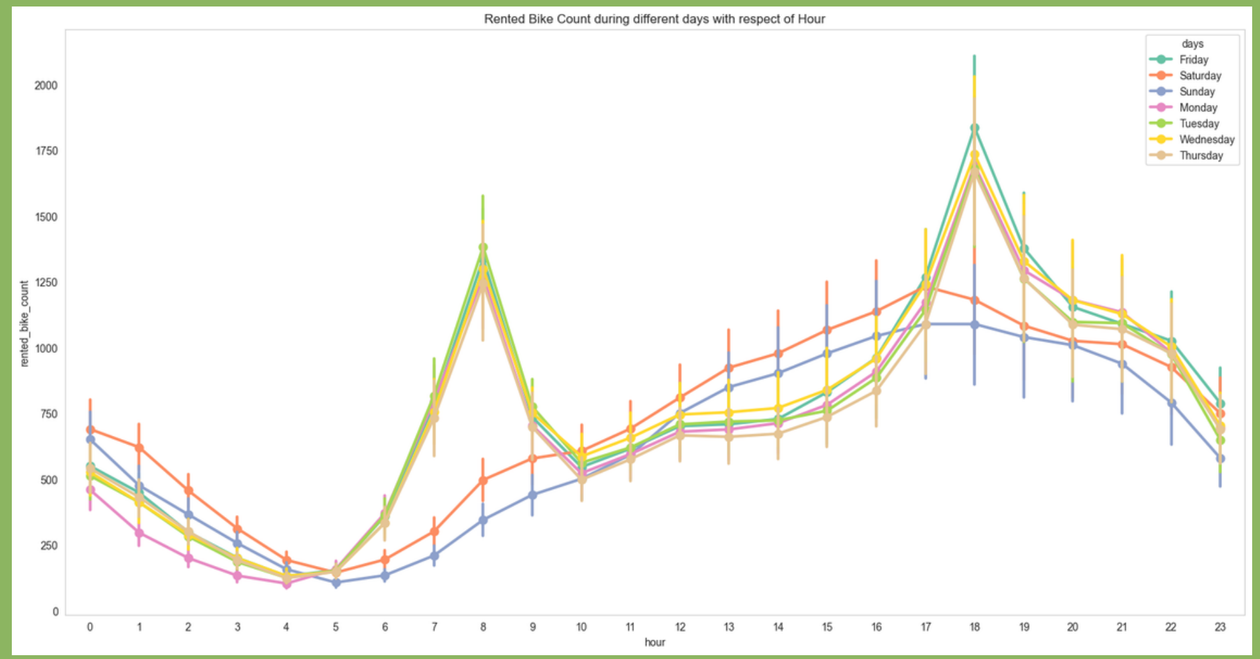
HOLIDAYS

- Demand is low during holidays
- Demand is high during the no holidays (work)

DAYS OF THE WEEK

- Pattern for weekdays and weekends is different

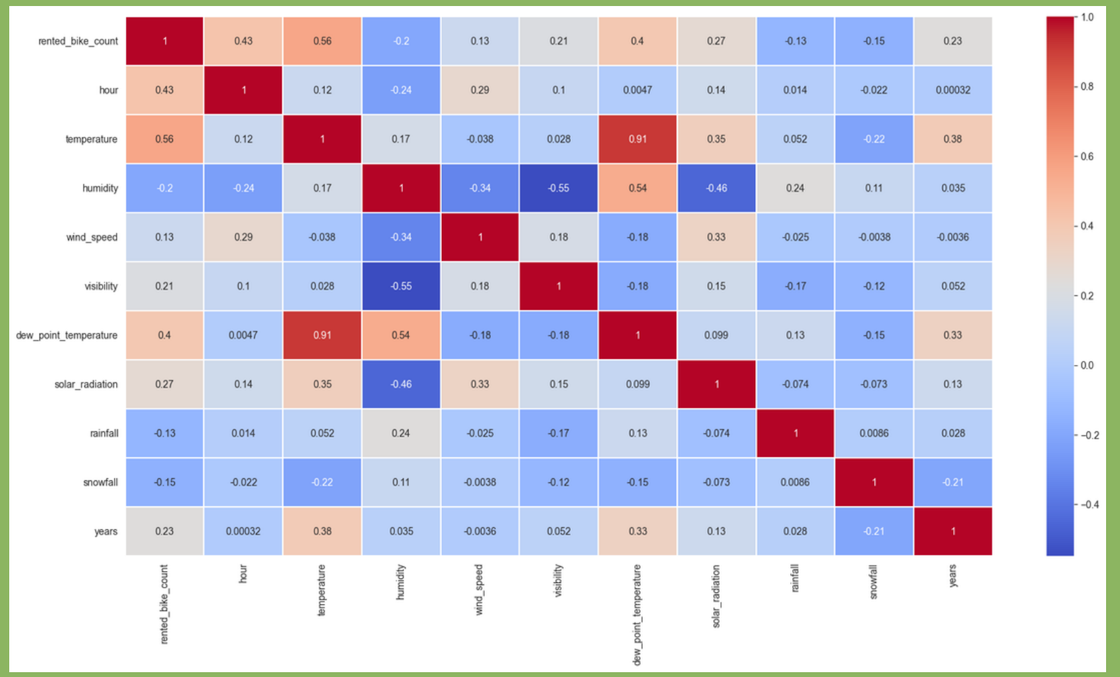


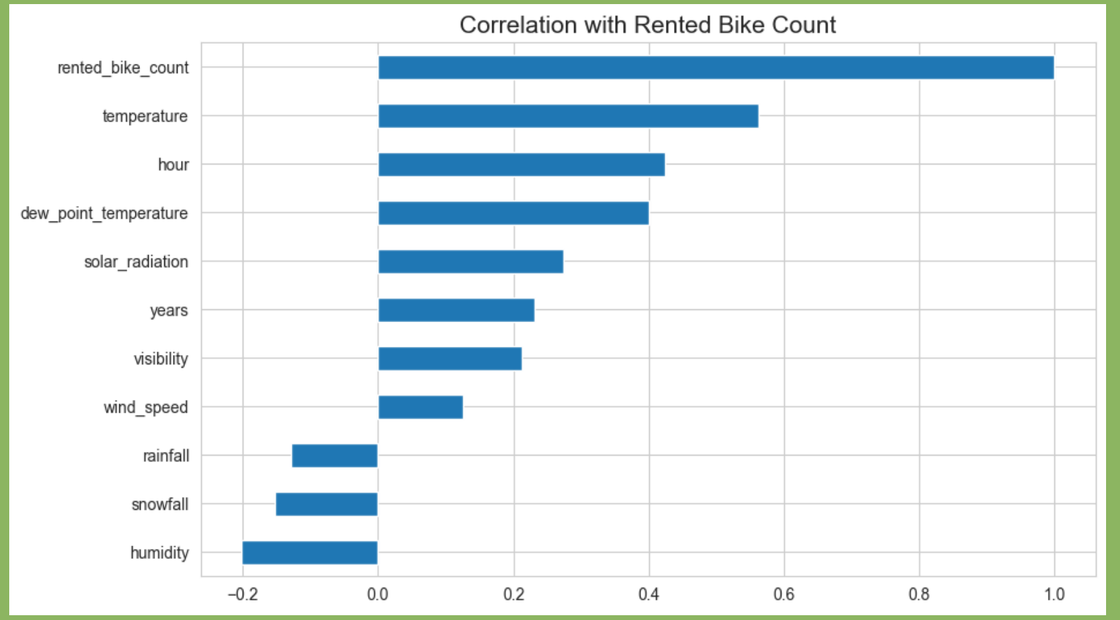


Correlation Analysis

HEATMAP

- Dew point temperature and temperature are highly correlated
- Multicollinearity
- Drop the feature that is the least correlated with the target variable (Dew Point Temperature)





Key findings

HIGH TEMPERATURE

People tend to rent more bikes when the temperatures are warmer (Autumn and Summer).

SPECIFIC HOURS

The number of bikes rented is higher during “peak hours” (around 8am and 6pm).

WORK, WORK, WORK

People are commuting to work and back home. On the weekends they demand is higher during the afternoon.

HOLIDAYS

During the holidays, there is not su much demand, users rent bicycles mostly for work.

Encoding

ONE-HOT ENCODING

- Holiday, seasons, hours, days and months will all be transformed

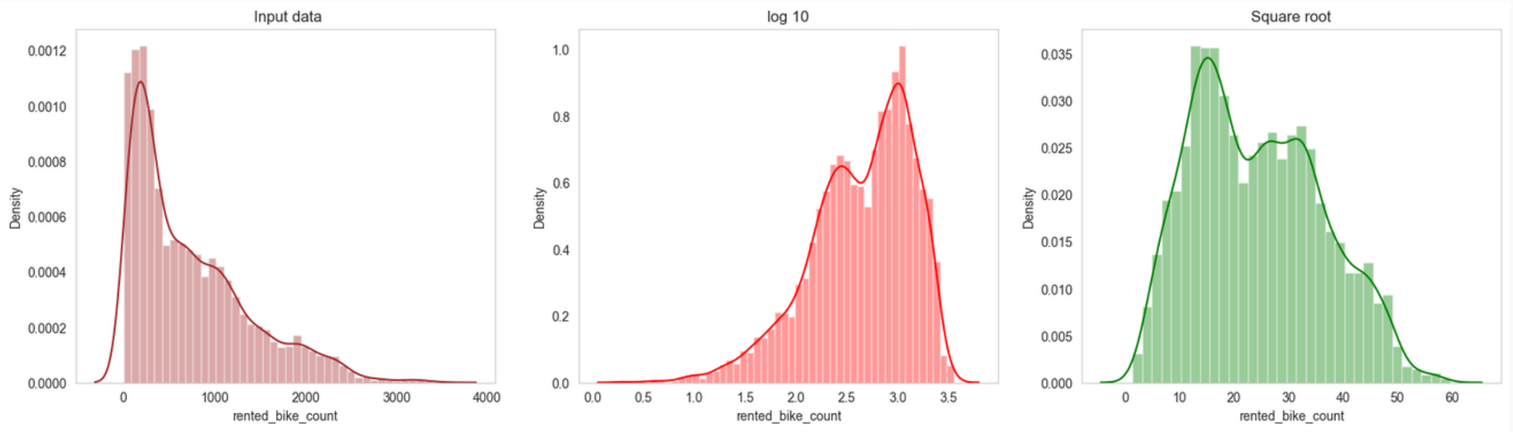
id	color
1	red
2	blue
3	green
4	blue

One Hot Encoding

id	color_red	color_blue	color_green
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0

Transform the target variable .....

- Before transformation: right skewed
- After Transformation using square root transformation: in the green plot our target variable is normalized to some extent



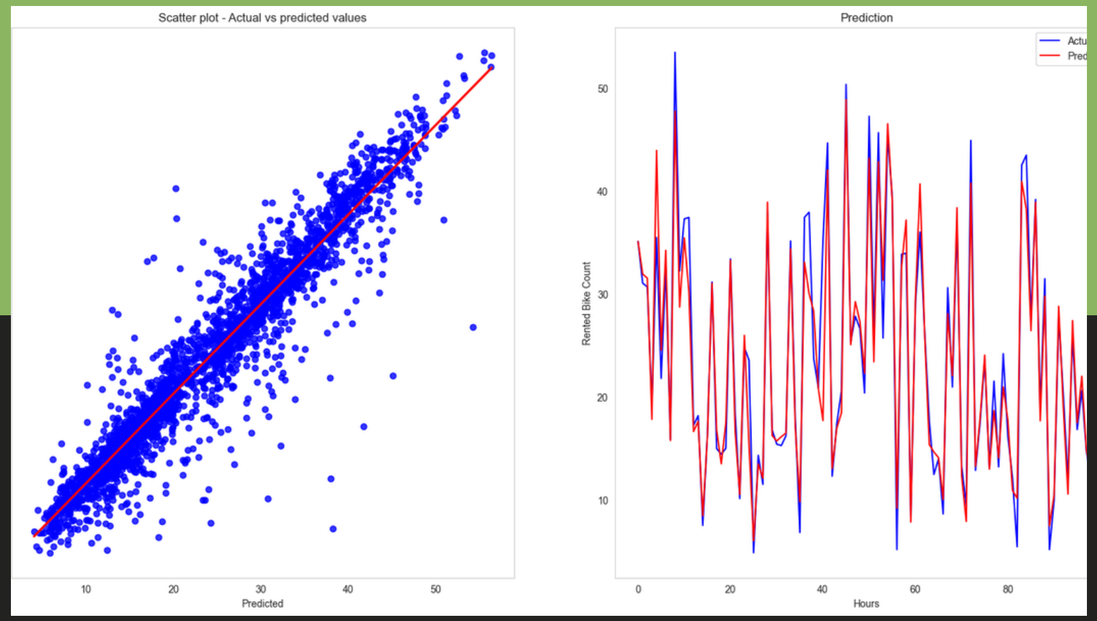
Feature Scaling

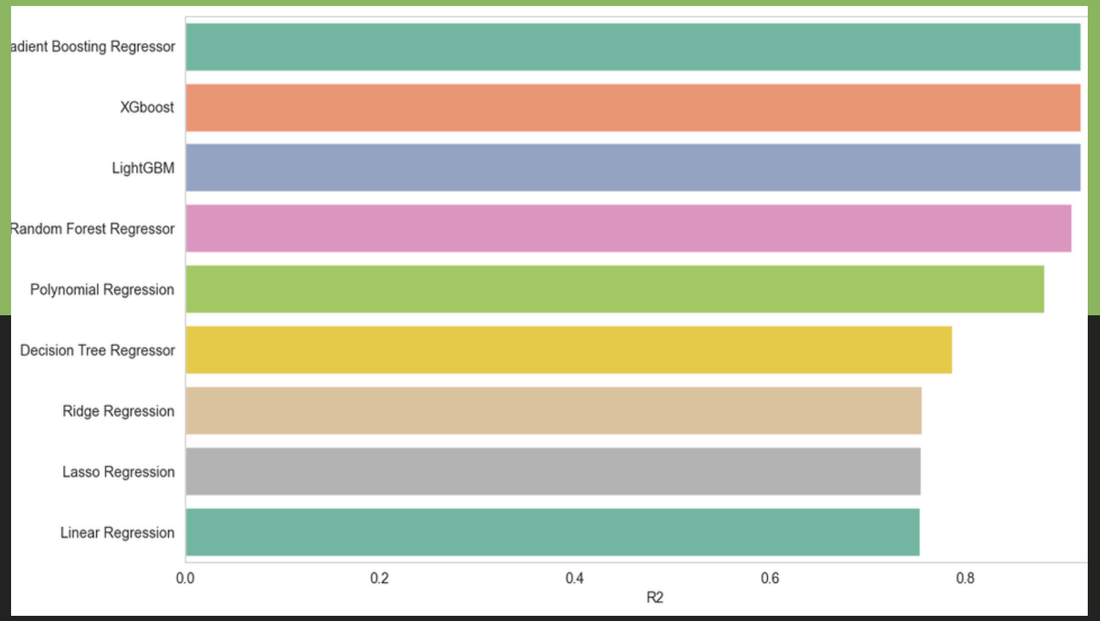
- For scaling the data we decided to go with Standardization
- More efficient to compare measurments with different measurments & with outliers

Models

REGRESSION MODELS

- Linear regression
- Polynomial regression
- Decision Tree regression
- Random forest regression
- Ridge regression
- Lasso regression





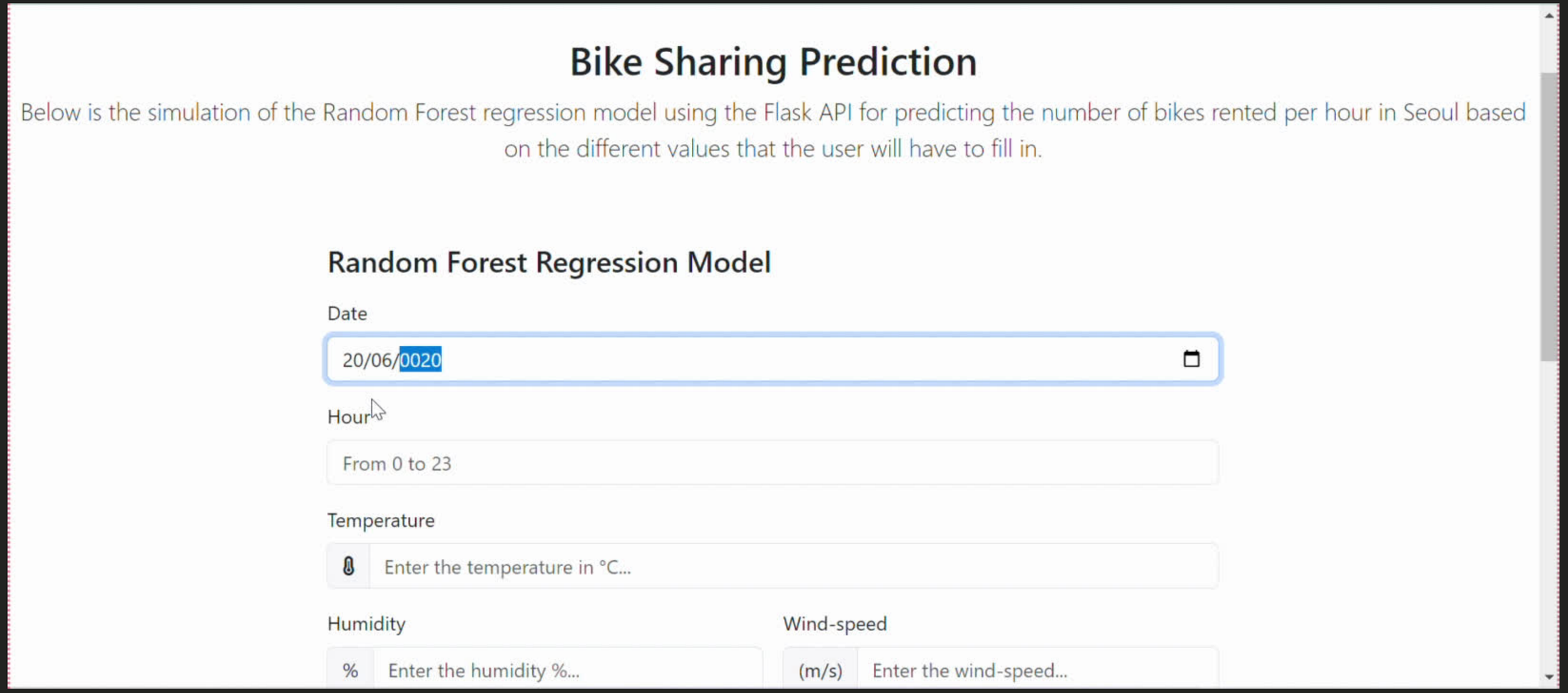
BEST MODEL

- Random Forest Regression
- R2 = 0.909010
- Training Score = 0.987905

EVALUATION OF THE DIFFERENT MODELS

- Polynomial Regression comes in 2nd
- The model with the best performance will be used in the model deployment

API Model Deployment



OUR PARTING WORDS

THANK YOU FOR LISTENING

REZGUI MOHAMED-HOUSSEM
YAHYA EL OUDOUNI
SASCHA CAUCHON