Washington, D.C. Bikeshare Demand Analysis and Prediction

—Utilizing Machine Learning to Forecast Daily Bike Rentals

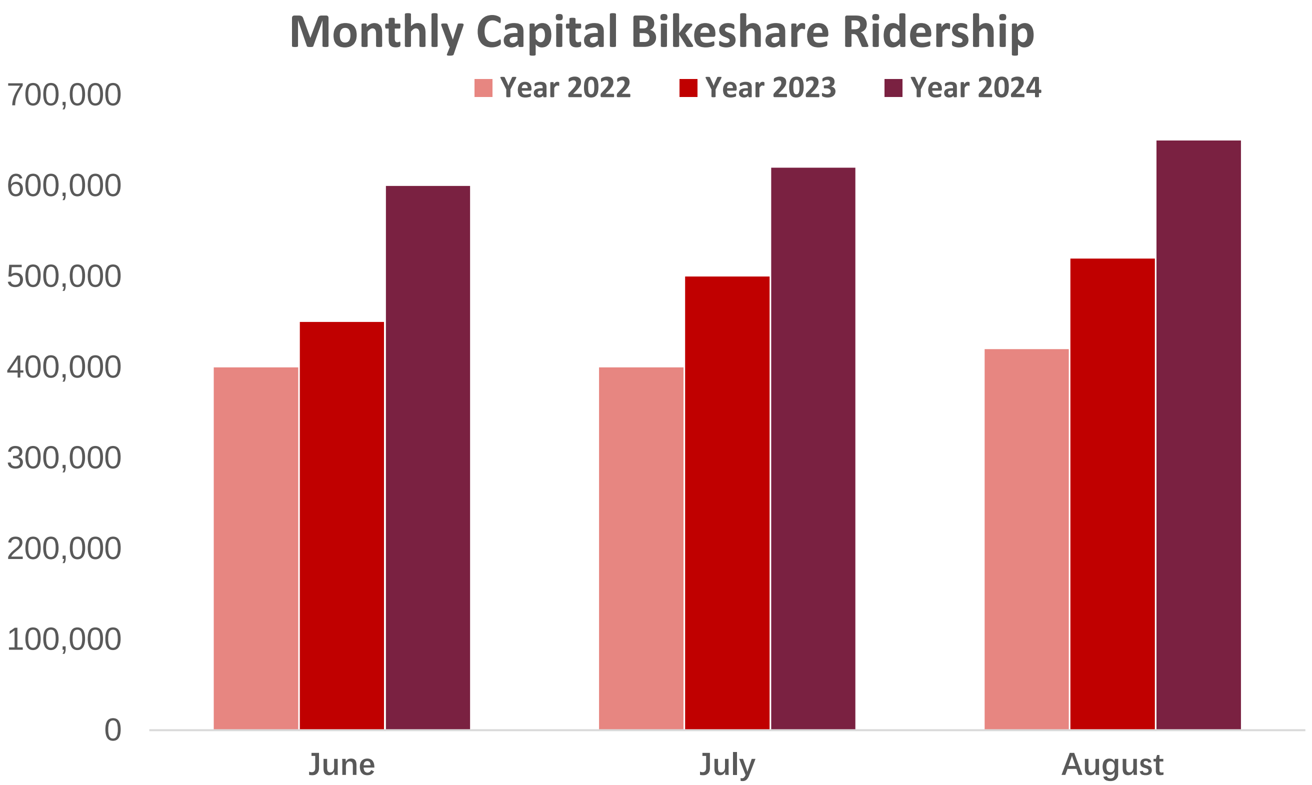
Group6- Hui Gao, Jerry Lin, ManYi Hong, Vivian Huang

Group 6

Contents

Background

Bike-sharing has become an essential part of modern urban transportation.



Month-to-Month

In August, CaBi recorded 614,639 rides. This is a **31.1%** increase from August 2023.

For the Record

Through August 2024, CaBi's yearly ridership has increased **31.1%** from 2023.

	Rides through August
2024	**3,788,634**
2023	2,890,520
2022	2,322,438
2021	1,735,789

However, with the growing demand for shared bikes and the diversification of usage patterns, accurately predicting bike demand and efficiently allocating resources has become a significant challenge for operators.

Problem Statement

In Washington D.C, the **lack of accurate demand forecasting and resource optimization** in bike-sharing systems leads to **operational inefficiencies** and **user dissatisfaction**.

Problems

01 Unpredictable Demand
Growing and varied usage patterns challenge accurate demand forecasting.

02 Inefficient Resource Allocation
Over- or under-stocking results in user dissatisfaction and operational inefficiencies.

03 Complex Influencing Factors
Weather, seasons, and user behaviors are difficult to model.

GOALS

01 Predicting Daily Bike Demands

02 Evaluating the Impact of Weather on Demand

03 Optimizing Bike Availability at Peak Locations and Times

Contents

Data Source

The Capital Bikeshare System Data from Capital Bikeshare provides extensive information on bike-sharing activity in the Washington, D.C., metro area and its neighboring regions.

Station list (2)
station_id
station_name

Daily rent data (13)	
ride_id	***started_at (time)***
rideable_type	*start_lng*
started_lat	*end_lat*
ended_at	*end_lng*
start_station_id	*end_station_id*
member_casual	*end_station_name*
start_station_name	

Usage frequency (3)
date
station_name
pickup_counts
dropoff_counts

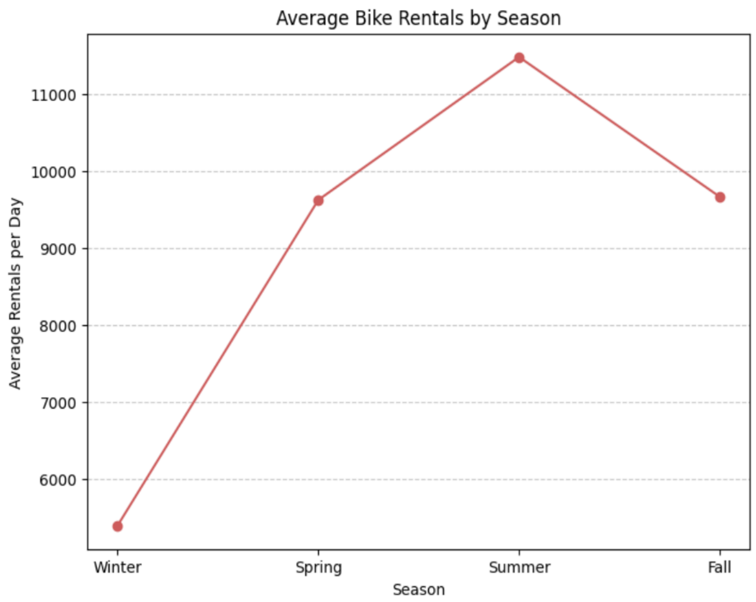
Weather (33)	
cloudcover	*feelslike*
datetime	*dew*
tempmax	*humidity*
feelslikemax	*precipprob*
feelslikemin	*snow*
……	

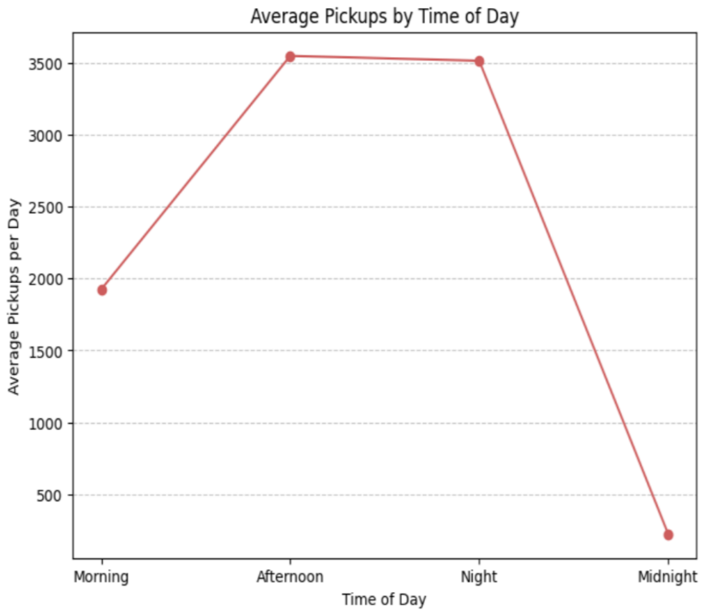
Source: https://capitalbikeshare.com/system-data , https://www.visualcrossing.com/

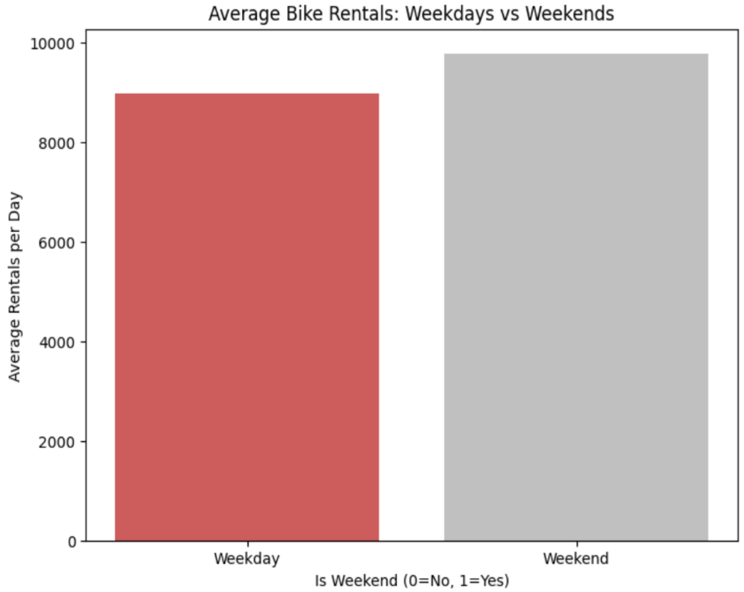
Contents

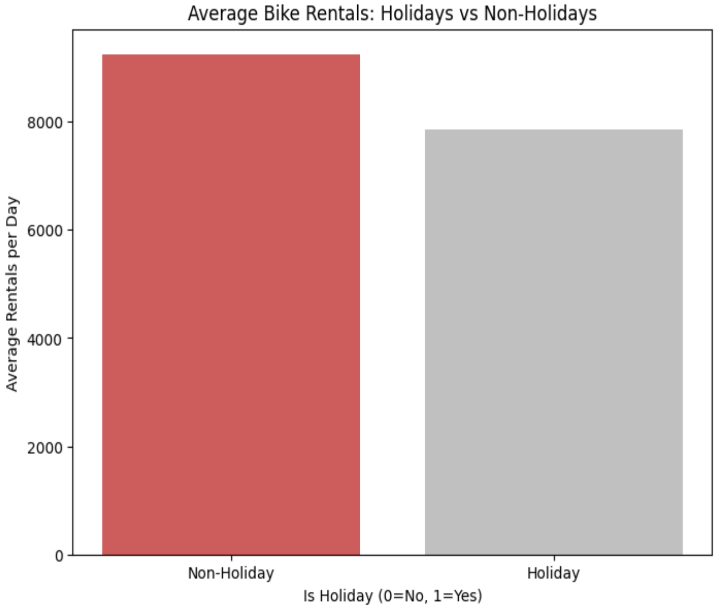
EDA

Time Factors



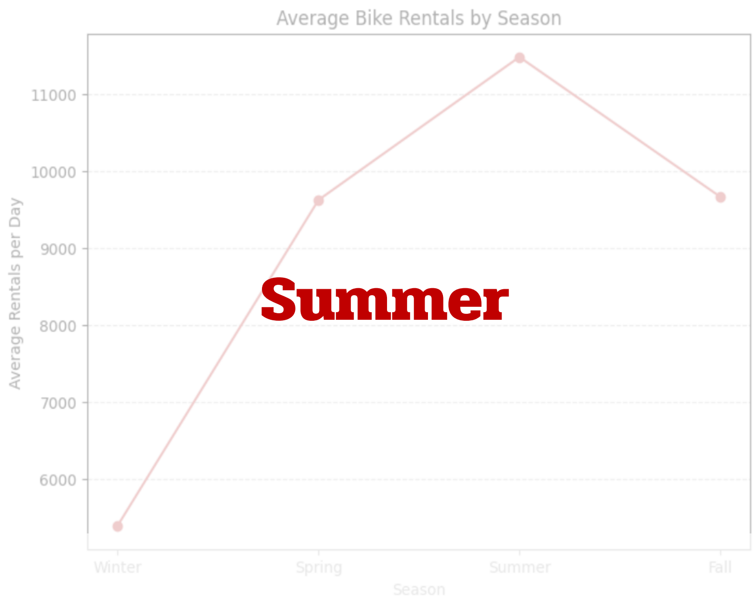


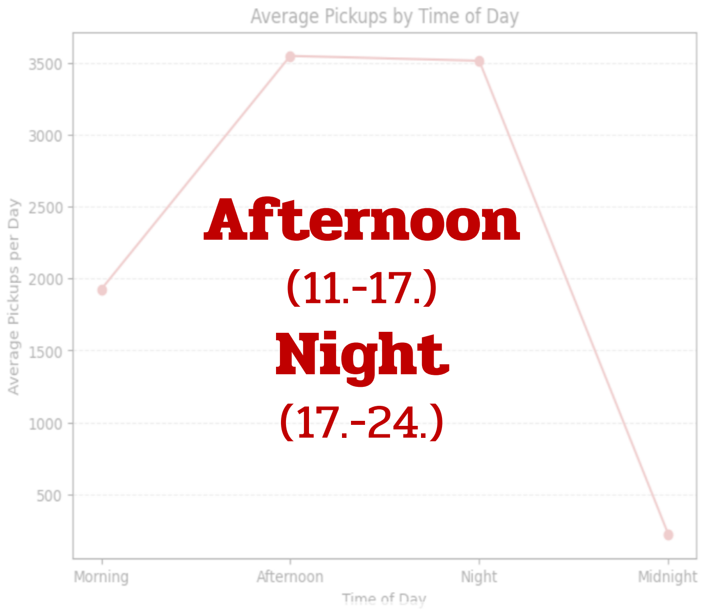


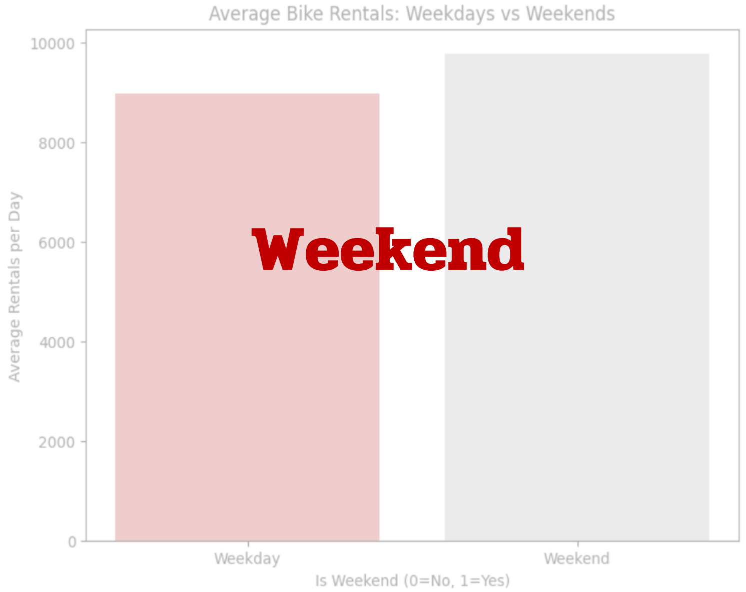


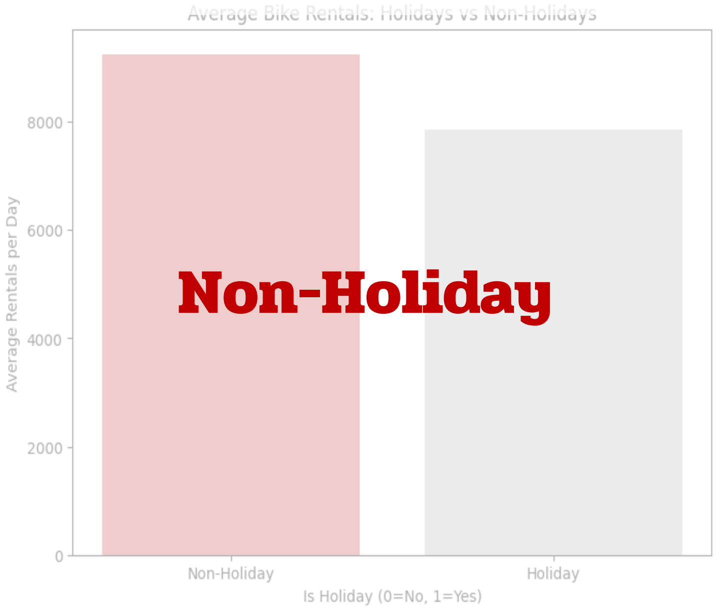
EDA

Time Factors

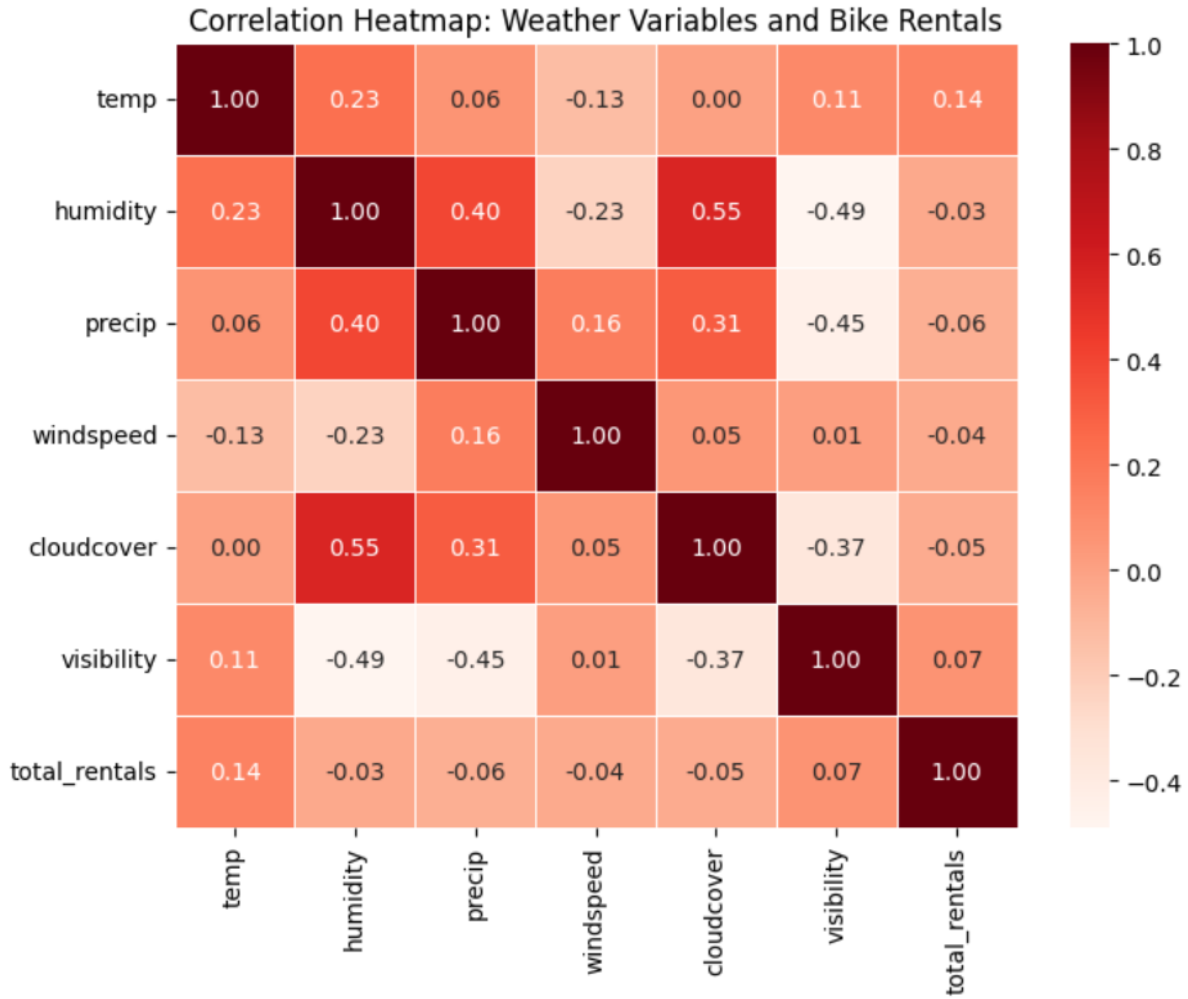


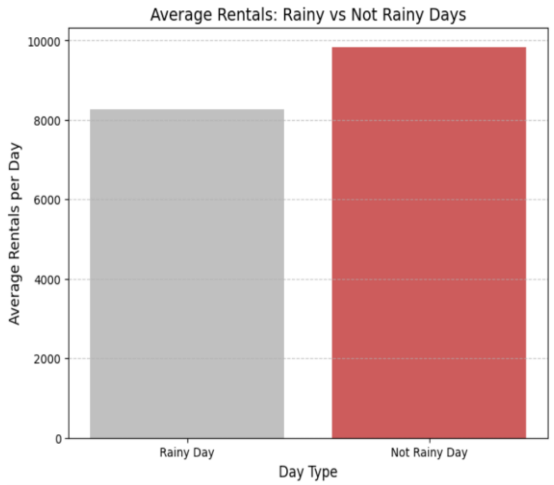


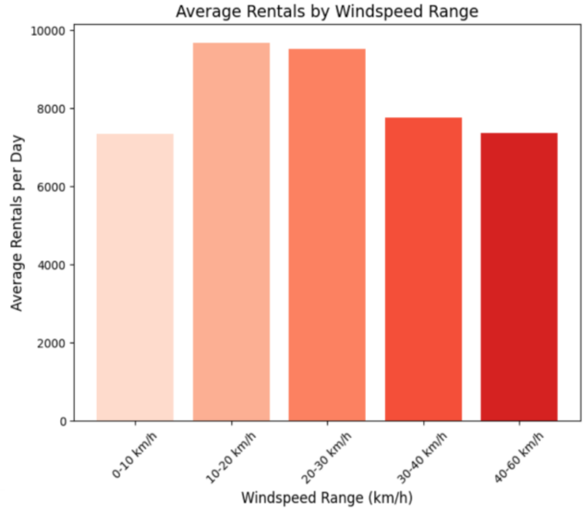


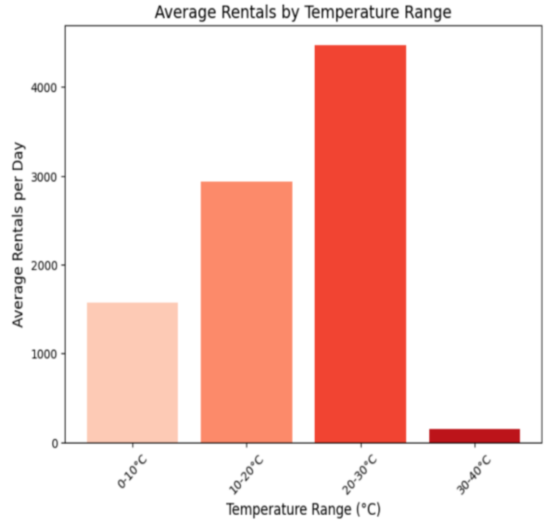


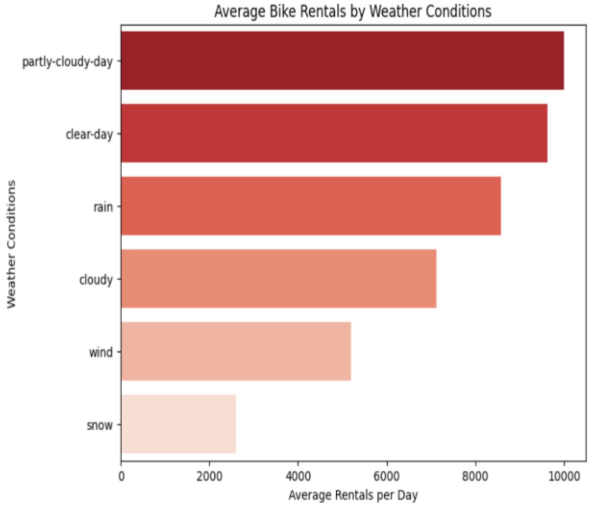
EDA- Weather Factors



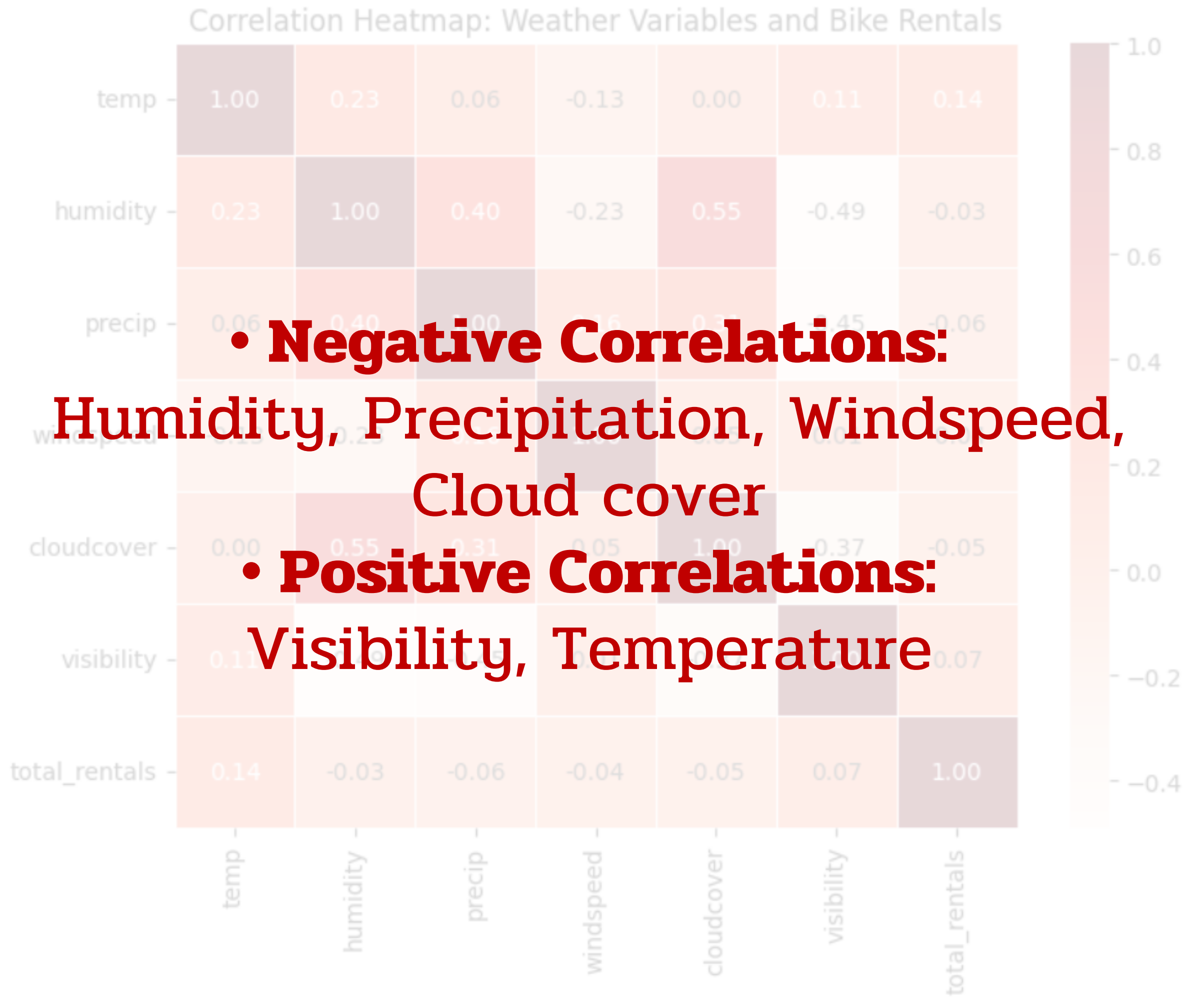






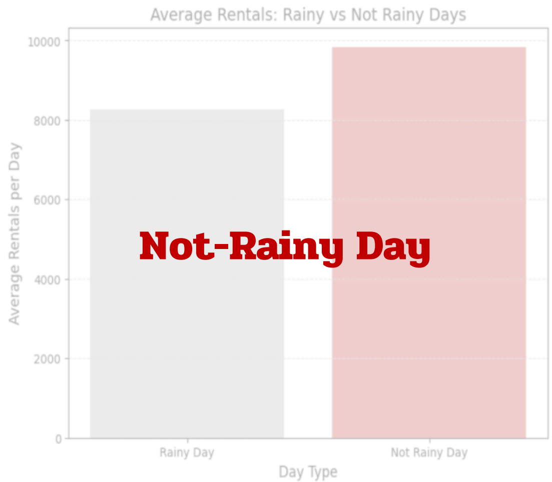


EDA- Weather Factors

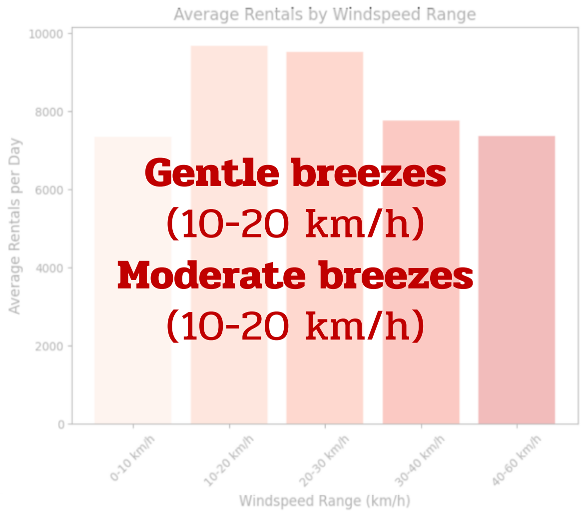


• **Negative Correlations:**
Humidity, Precipitation, Windspeed, Cloud cover

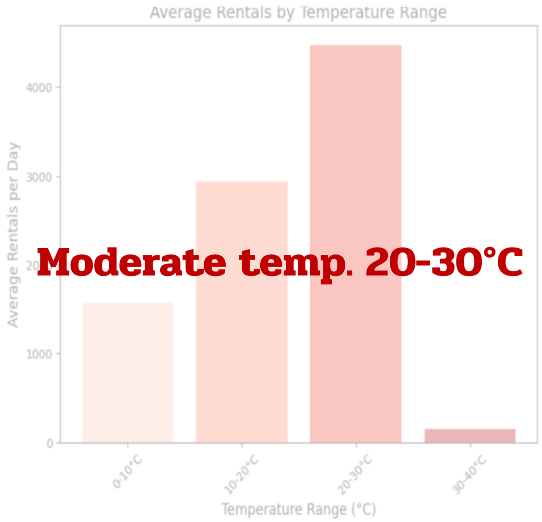
• **Positive Correlations:**
Visibility, Temperature



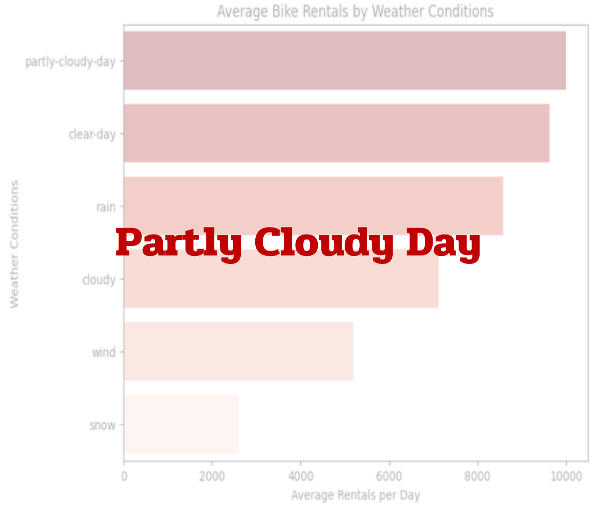
Not-Rainy Day



Gentle breezes
(10-20 km/h)
Moderate breezes
(10-20 km/h)



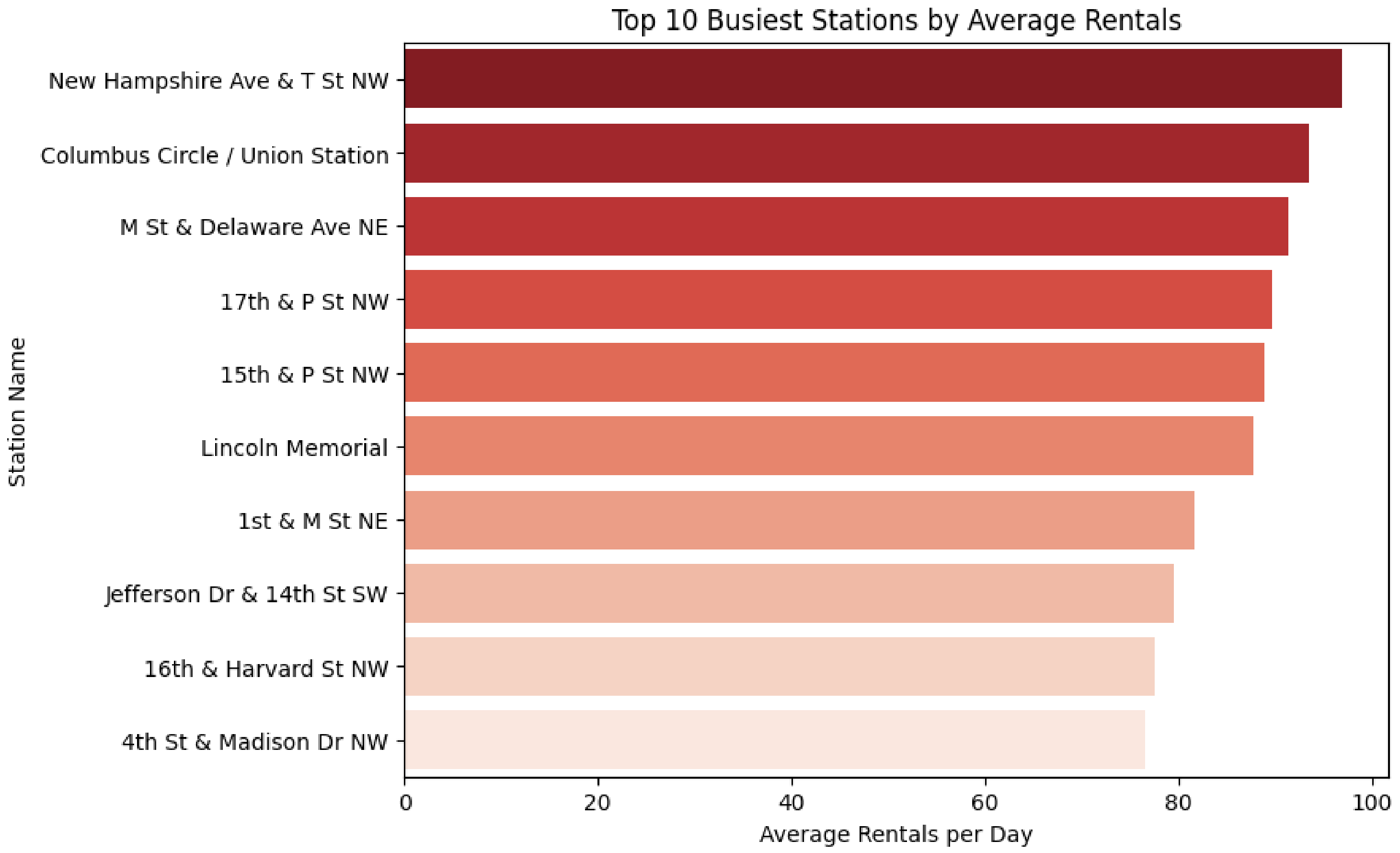
Moderate temp. 20-30°C



Partly Cloudy Day

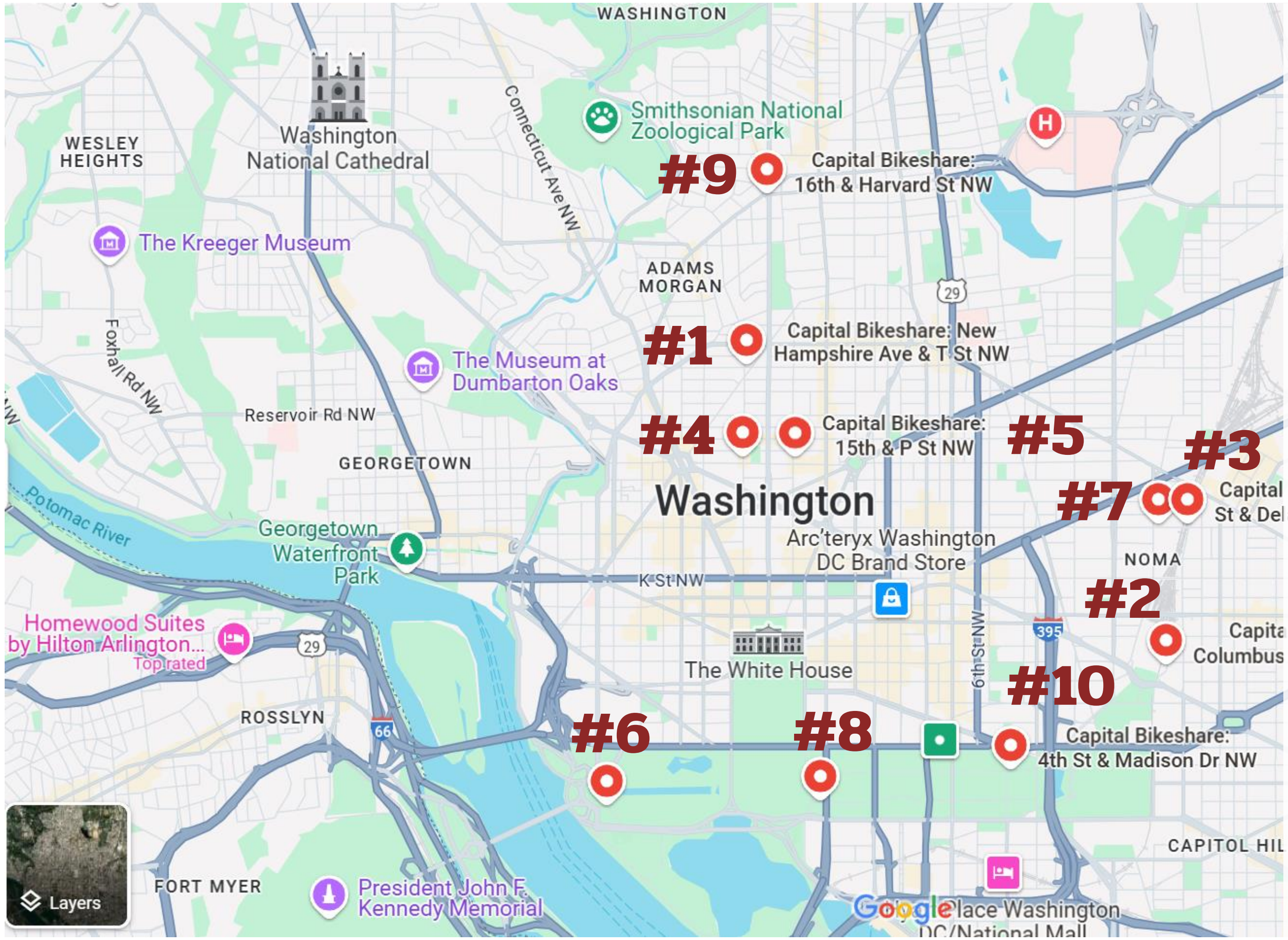
EDA- Spatial Factors (Station Activity)

Top 10 busiest stations



EDA- Spatial Factors (Station Activity)

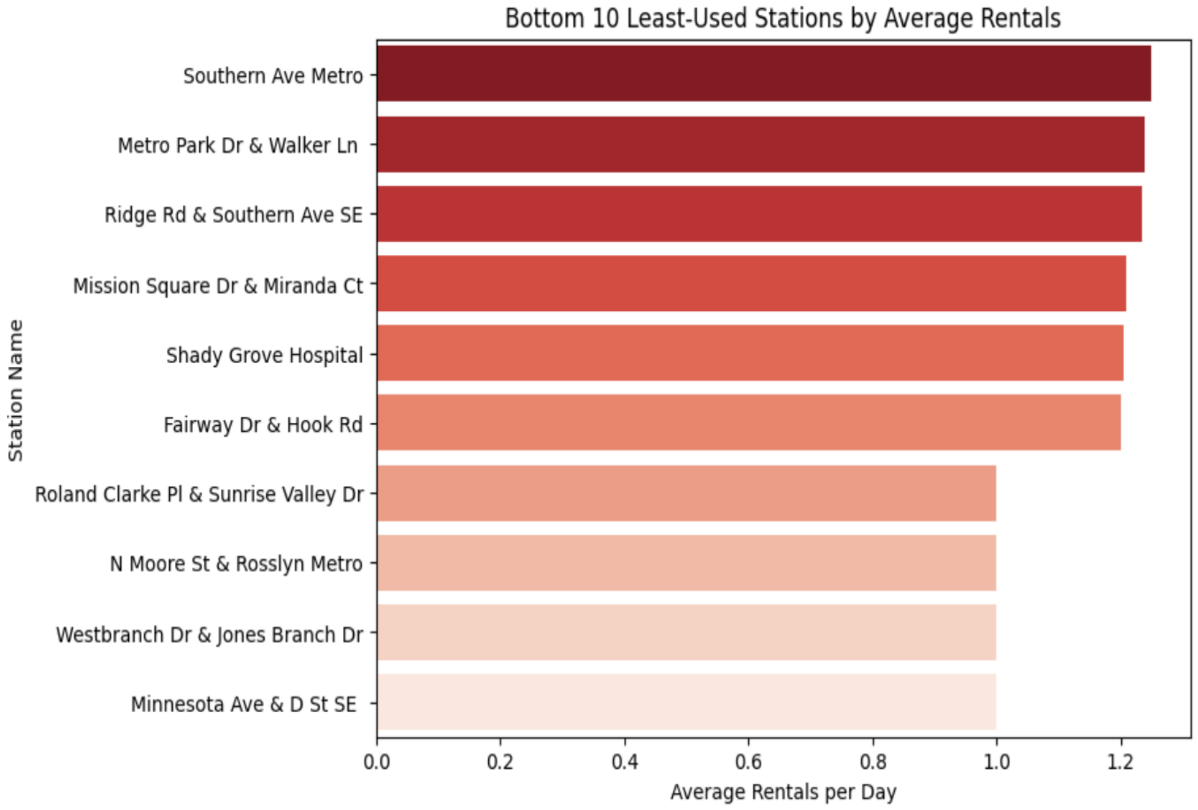
Top 10 busiest stations



The majority of the top 10 busiest stations are concentrated in **central Washington, D.C.**, near high-traffic areas such as **tourist landmarks, government buildings, and popular commuter hubs**. Ex, #1, #2, #6, #10

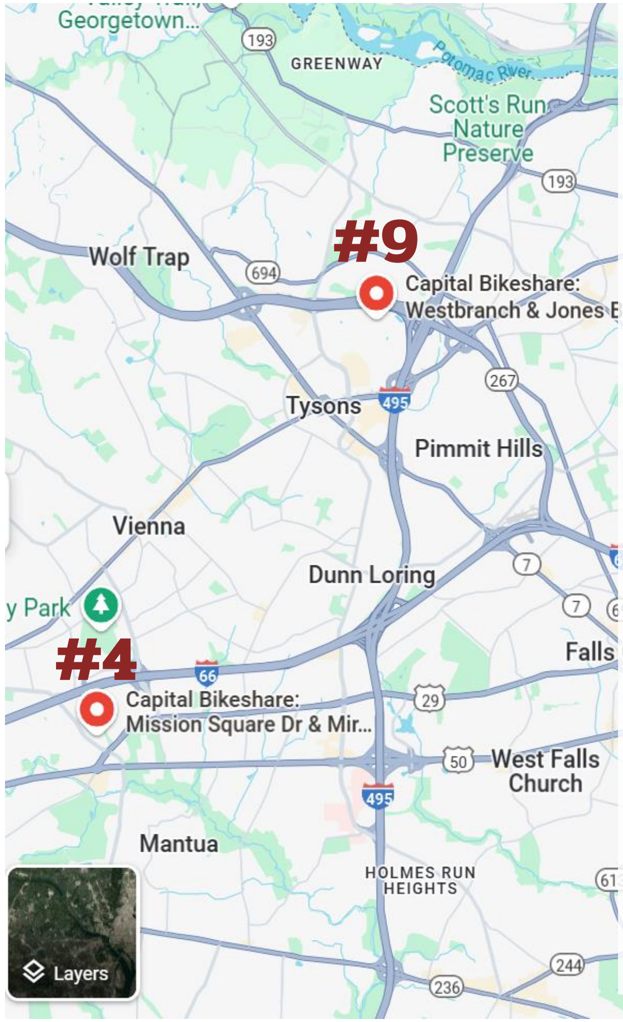
EDA- Spatial Factors (Station Activity)

Bottom 10 least-used stations

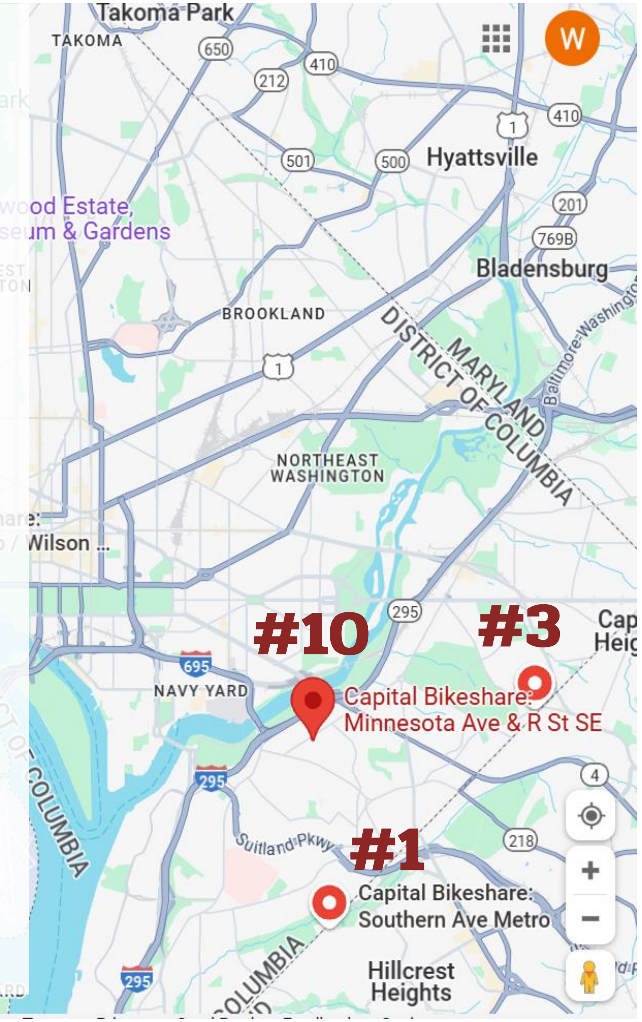


EDA- Spatial Factors (Station Activity)

Bottom 10 least-used stations



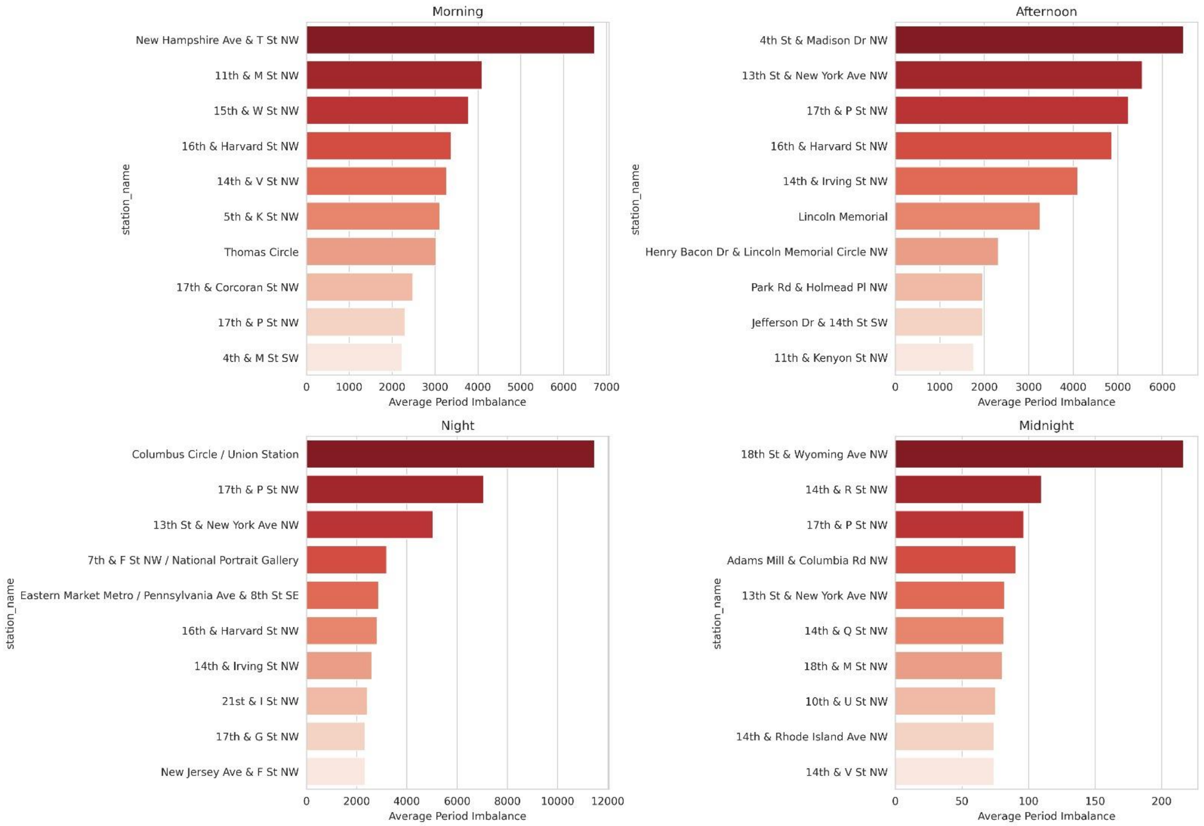
Many of the least-used stations are located on the outskirts of Washington, D.C., in **suburban or less densely populated areas**. Some also have low usage due to **weaker connectivity** to other high-demand stations or destinations. ex, #1, #4, #8, #9, #10



EDA- Station Imbalance

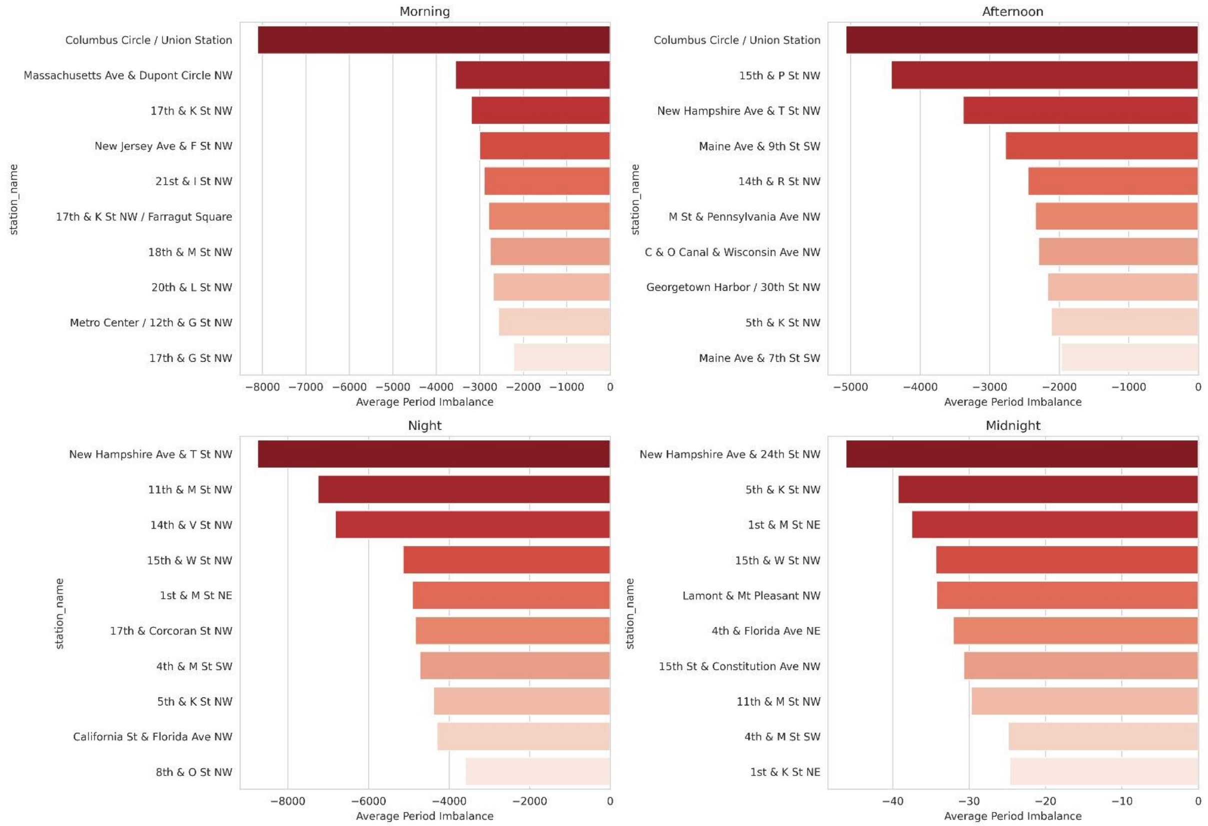
Demand > Supply

Top 10 Stations by Average Period Imbalance for Each Period



Supply > Demand

Low 10 Stations by Average Period Imbalance for Each Period



EDA- Station Imbalance

• Imbalance (Pick up-Drop off) happened:

Night (17.-24.) > Morning > Afternoon > Midnight

• Many of the **busiest stations** experience **high imbalances**.

EX, New Hampshire Ave & T St NW (#1 busiest) Columbus Circle
EX, Union Station (#2 busiest)

• Some Stations show **reverse trends**

(Ex, New Hampshire Ave & T St NW: morning pickups and midnight drop-offs)

• **Transit hubs** have sharp demand differences during commuting hours

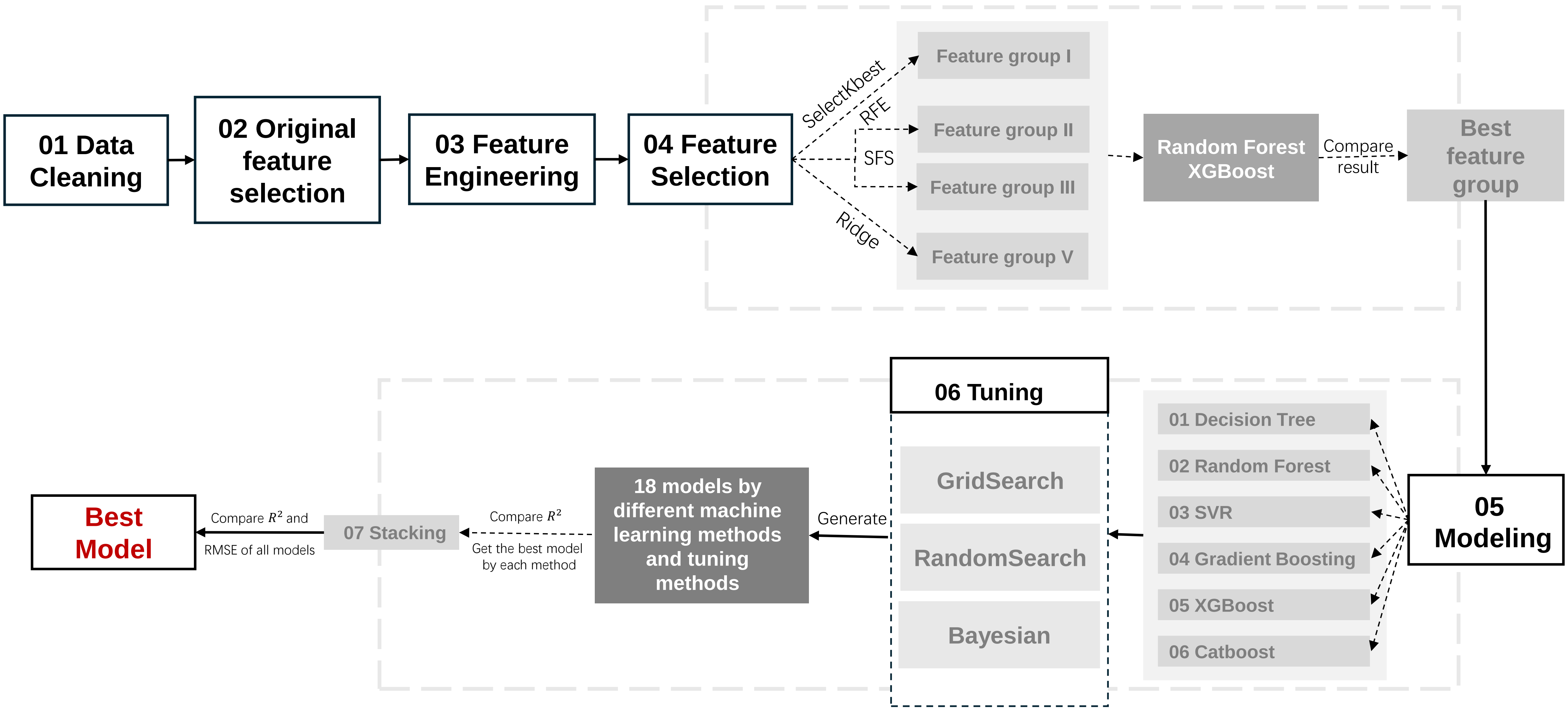
(Ex, Union Station)

• **Stations near tourist attractions** have high afternoon usage

(Ex, 4th St & Madison Dr NW)

Contents

Processing Map of Machine Learning



01 Data Cleaning → 02 Original Feature Selection → 03 Feature Engineering → 04 Feature Selection → 05 Modeling → 06 Tuning → Best Model

01 Data Cleaning

- Handling Missing Values
- Changing Data type
- Examine potential outliers
- Inspect for duplicate rows
- Check the cleaned data

02 Original Feature Selection

- **Removed similar variables**
  kept overall temperature instead of max, min, average, or "feels like" temperature
- **Excluded irrelevant variables**
  sunrise and sunset times
- **Dropped non-distinguishable variables**
  like "snow" with all values as 0

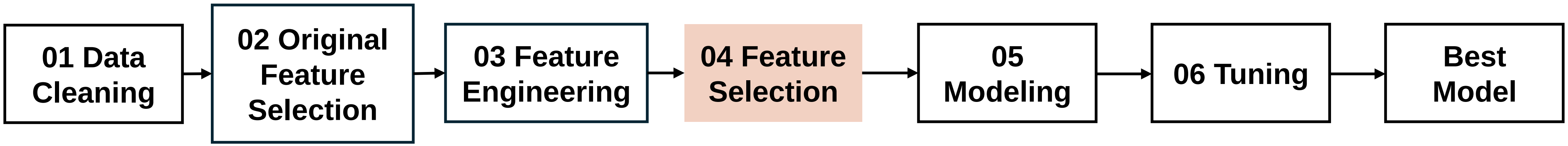
03 Feature Engineering

Create features to understand and model patterns in bikeshare usage.

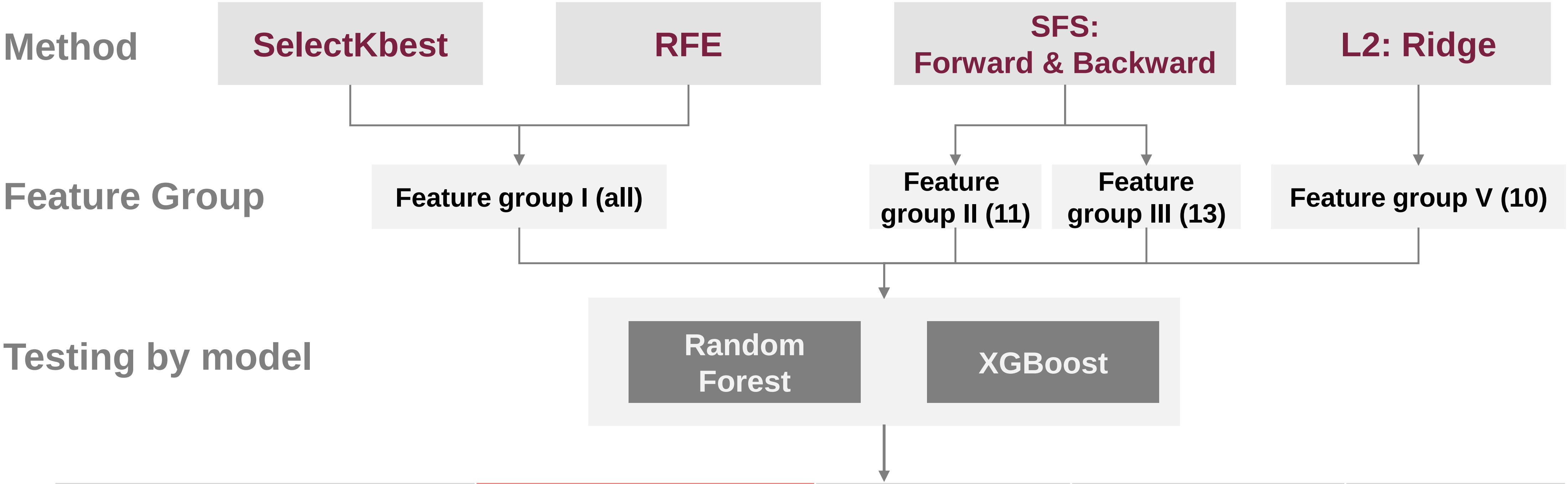
is_holiday	Indicates if the date is a public holiday (based on the US Federal Holiday Calendar)
is_weekend	Captures whether the date falls on a weekend (Saturday or Sunday).
season	Categorizes the date into seasons (1=Spring, 2=Summer, 3=Autumn, 4=Winter)

01 Data Cleaning → 02 Original Feature Selection → 03 Feature Engineering → 04 Feature Selection → 05 Modeling → 06 Tuning → Best Model

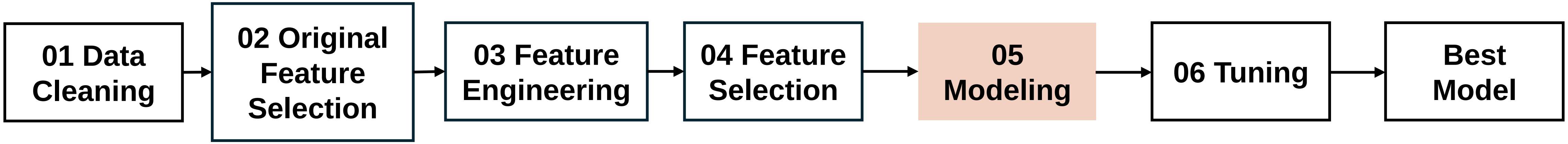
Final Dataset	
date	*total _pickup*
Is_holiday	*Is_weekend*
season	*temp*
humidity	*precip*
precipprob	*snow*
windgust	*windspeed*
cloudcover	visibility
severerisk	icon



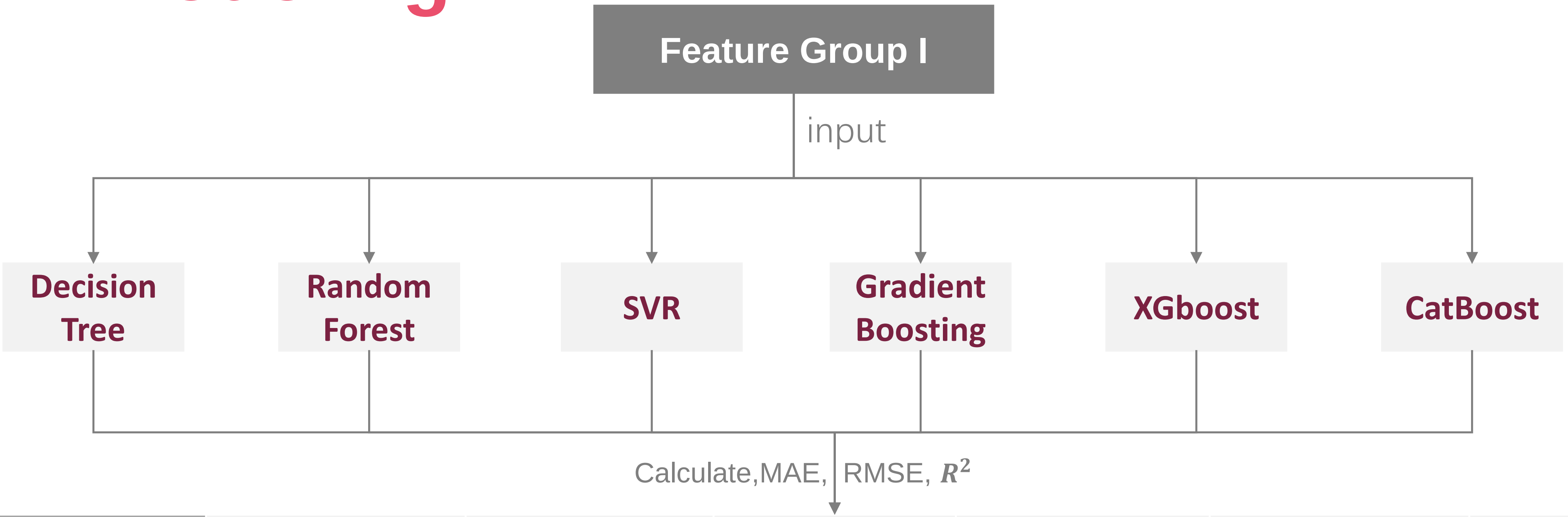
04 Feature Selection



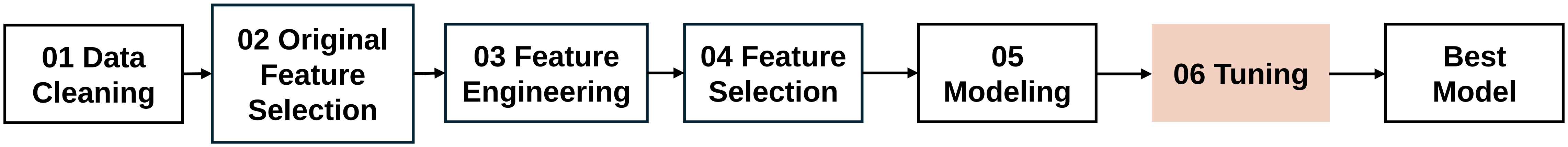
Feature Group	I (Best Feature group)	II	III	V
R^2 of Random Forest	0.8110	0.788	0.796	0.774
R^2 of XGBoost	0.781	0.759	0.771	0.7908



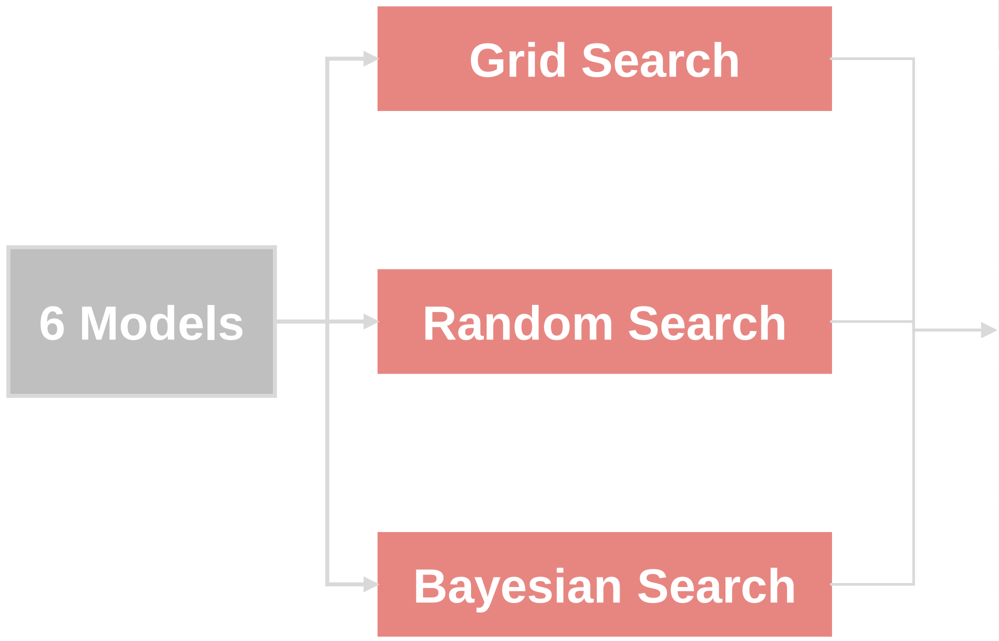
05-1 Modeling



MAE	1931.92	1363.14	3353.90	1344.15	1470.82	1316.72
RMSE	2500.46	1757.28	4024.65	1745.49	1890.69	1720.18
R^2	0.617	0.811	0.009	0.814	0.781	0.819



06 Tuning



	Original	Grid	Random	Bayesian
Decision Tree	0.617	0.718	0.761	0.753
Random Forest	0.811	0.813	0.815	0.815
SVR	0.009	0.560	0.560	0.558
Gradient Boosting	0.814	0.820	0.826	0.824
XGBoost	0.781	0.826	0.823	0.822
CatBoost	0.819	0.833	0.815	0.834

01 Data Cleaning → 02 Original Feature Selection → 03 Feature Engineering → 04 Feature Selection → **05 Modeling** → 06 Tuning → Best Model

05-2 Stacking

	Original	Grid	Random	Bayesian
Decision Tree	0.617	0.718	**0.761**	0.753
Random Forest	0.811	0.813	**0.81504**	0.81500
SVR	0.009	**0.56047**	0.56004	0.558
Gradient Boosting	0.814	0.820	**0.826**	0.824
XGBoost	0.781	**0.826**	0.823	0.822
CatBoost	0.819	0.833	0.815	**0.834**

Stacking →
Get the best model by each method

Model 7

MAE	1275.94
RMSE	1664.76
R^2	0.8304

01 Data Cleaning → 02 Original Feature Selection → 03 Feature Engineering → 04 Feature Selection → 05 Modeling → 06 Tuning → Best Model

07- Model Selection

	Original	Grid	Random	Bayesian
Decision Tree	0.617	0.718	0.761	0.753
Random Forest	0.811	0.813	0.815	0.815
SVR	0.009	0.560	0.560	0.558
Gradient Boosting	0.814	0.820	0.826	0.824
XGBoost	0.781	0.826	0.823	0.822
CatBoost	0.819	0.833	0.815	0.834
Stacking	0.8304	/	/	/

Best Model

CONTENT

Conclusion

1. **Relationship Between Weather and Bikeshare Usage**

   Moderate weather conditions increase bikeshare demand, while extreme weather reduces usage. Including weather data improves forecasting accuracy.

2. **Station Daily Imbalance**

   Imbalances in pick-ups and drop-offs, influenced by location and time, can cause shortages or overcrowding. Dynamic rebalancing enhances system efficiency.

3. **Predictive Model of Daily Demand**

   Machine learning models effectively predict daily ridership, aiding resource allocation and improving user satisfaction.

Insight

Washington D.C. Bikeshare Demand Analysis and Prediction

Challenges & Future Steps

Challenges

01 Large and Messy Dataset

4 datasets, including daily_rent_detail, which contains over 14 million rows, and weather data with more than 30 columns.

02 Excessive Repetition or Similarity in Features

a large number of repeated or highly similar columns, feature duplication, negatively impacting model accuracy

03 Necessity of Feature Engineering

The columns in the raw dataset lack sufficient meaningful features to capture patterns in bike rental behavior.

Future Step

01 Collecting More Individual Station Information

- **Incorporate station-specific variables**, such as nearby attractions, schools, shopping malls, and alternative transportation options.

02 Predicting the Imbalance in Bike Rentals at Individual Stations

- Extend the analysis to include drop-off volumes in addition to pick-up volumes.

Thank You

QnA

Colab Link

https://colab.research.google.com/drive/10uTzHQwv6FxMMWWbc_w48idcSWnHnW4Q?usp=sharing