BIKE RENTAL DEMAND PREDICTION IN WASHINGTON D.C

By **Bintang Ary Pradana**

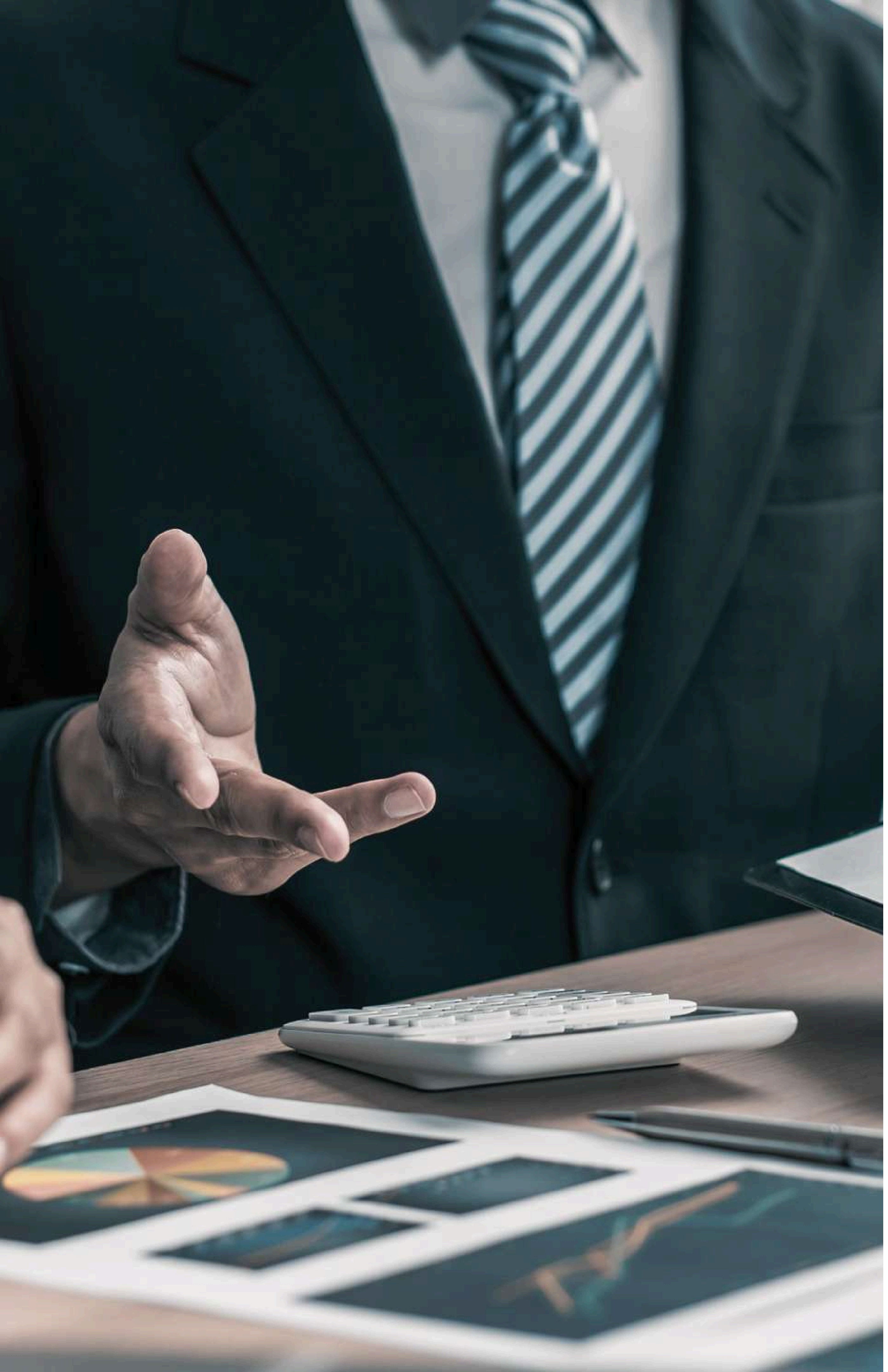


TABLE OF CONTENT

01 Business Understanding

02 Data Understanding

03 Data Preparation

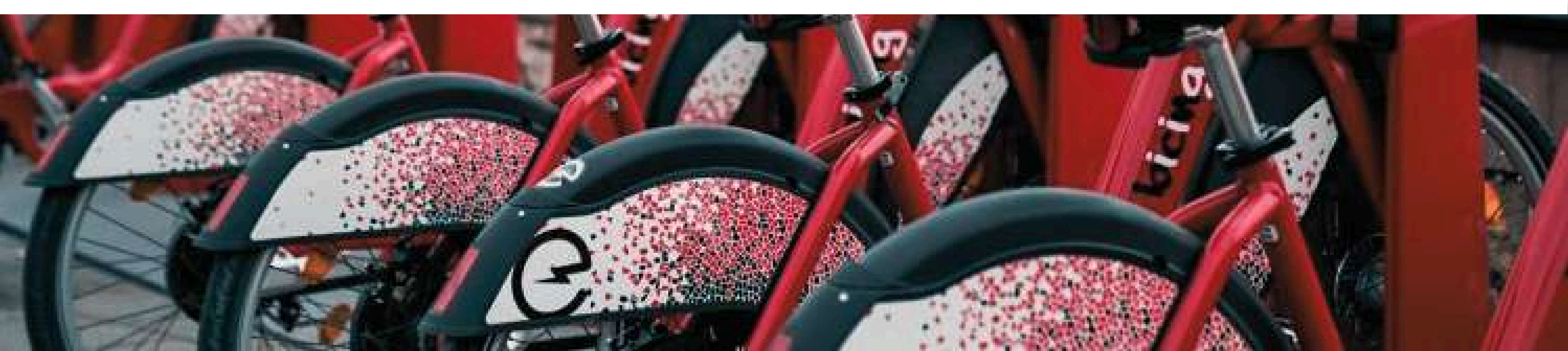
04 Exploratory Data Analysis & Insight

05 Data Preprocessing

06 Machine Learning Model

07 Conclusion and Recommendation

01 | Business Understanding



BACKGROUND

Bicycle sharing Washington D.C, are offer short-term transportation service by providing bicycles for communal use. These systems enable users to rent a bike from one location and return it to another within the network.

Accessible through automated kiosks or mobile applications, users can pay for rentals via subscription or per-use charges. By encouraging cycling, bike sharing promotes a convenient, healthy, and eco-friendly.

CURRENT PROBLEM

Inconsistent Bike Availability

Bike-sharing companies often struggle to accurately estimate the number of bikes needed at specific times. This fluctuating demand can result in either too many bikes, leading to increased operational costs, or too few bikes, causing frustration among users and undermining their trust in the service.

Limited Customer Behavior Insights

Bike sharing companies face inefficient resource allocation due to a lack of insight into customer behavior. This results in bike oversupply or undersupply in different areas, deterring potential users and reducing operational efficiency.

Following the current problem, our aim is to develop accurate demand predictions to ensure optimal bike availability by analyzing the number of customers in the bike-sharing system.

RESARCH QUESTION

As a **Data Scientist** in Bike-sharing company, we have responsibilities to give business insights and recommendations over a current problem to enhance company perfomance. **So, we need to answer this following deep dive questions**:

- **Which machine learning algorithm performs better** and has the most accurate result in bike rental prediction? and why?
- **What are the primary factors** that influence bike rental demand and contribute to fluctuations in availability throughout the day?
- How do these factors interact with each other to shape customer behavior and usage preferences regarding bike rental services, and **what actions should the company take** based on these insights?

HYPOTESIS

The demand for bike sharing services in Washington, D.C. can be influenced by various factors such as weather conditions, season, day of the week and time of day.

Time Of Day

Renting a bike is more common during peak commuting hours, typically when people are traveling to work or school. This suggests that the time of day plays a crucial role in estimating bike rental demand. By considering the busy periods when individuals are commuting, valuable insights can be gained into predicting the fluctuations in bike usage throughout the day.

Day of The Week

People are likely to rent bikes more frequently on weekdays due to the necessity for commuting and practical use, in contrast to weekends. During weekdays, individuals often have work or school commitments, necessitating transportation options like bike rentals for their daily commute. However, it's essential to verify this hypothesis with data since the day of the week significantly influences rental counts.

Weather Conditions

Generally, people are more likely to rent bikes when the weather is nice, such as sunny and mild days, rather than when it's raining or snowing heavily. It's reasonable to expect that bike rental demand will increase when the weather is favorable.

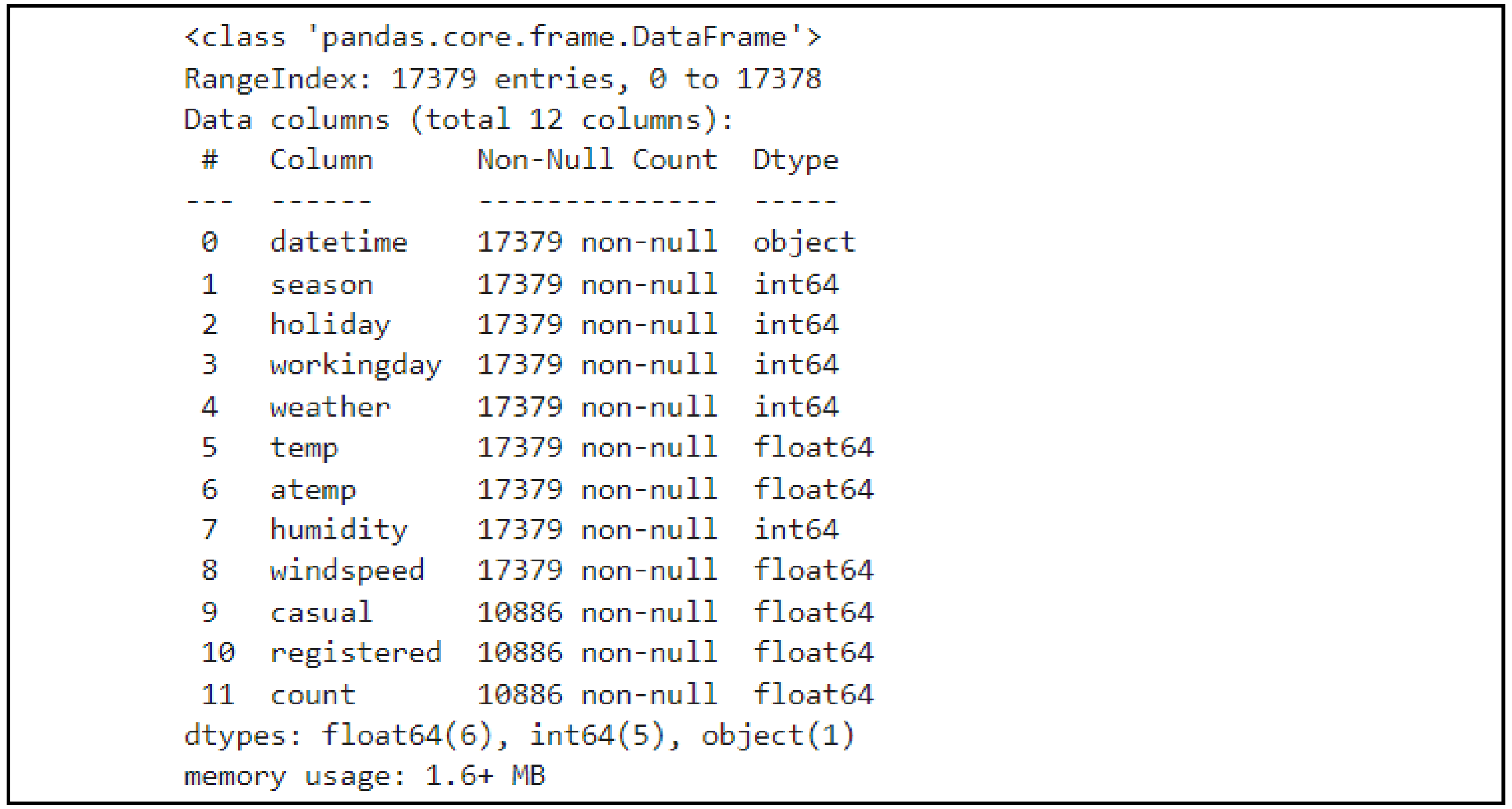
Season

People may lean towards renting a bike during the best seasons. They are more inclined to prefer biking in warmer temperatures, as it provides a more comfortable riding experience. Additionally, favorable seasons like fall and summer often encourage outdoor activities, including biking.

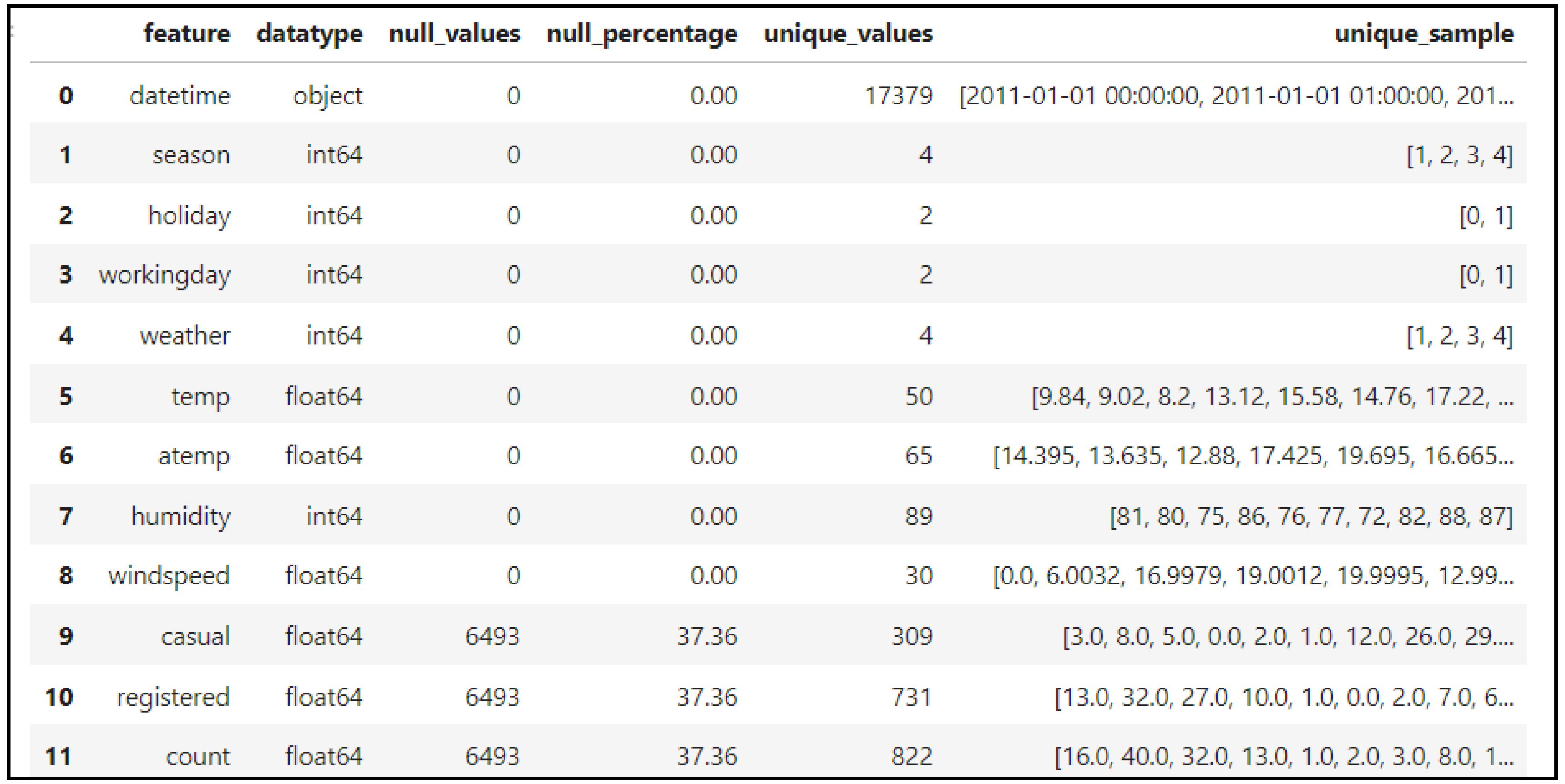
02 Data Understanding

DATA UNDERSTANDING

- There are **3 datasets used**, namely: train, test, and submission.
- The training dataset (train) consists of 12 columns and 10,886 rows.
- The test dataset (test) consists of 9 columns and 6,493 rows.
- The simpleSubmission dataset (submission) consist 2 columns and 6,493 rows.
- The total dataset (train + test) consist 12 columns and 17,379 rows.
- The dataset comprises **1 datetime column, 4 categorical columns, and 7 numerical columns**. The categorical data has been converted into integer format, although it was originally categorical.
- There are no duplicated data.
- There are missing values in the dataset due to the concatenation of the training and test data.
- The target or label is represented by the 'count' column.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 17379 entries, 0 to 17378
Data columns (total 12 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   datetime    17379 non-null  object 
 1   season      17379 non-null  int64  
 2   holiday     17379 non-null  int64  
 3   workingday  17379 non-null  int64  
 4   weather     17379 non-null  int64  
 5   temp        17379 non-null  float64
 6   atemp       17379 non-null  float64
 7   humidity    17379 non-null  int64  
 8   windspeed   17379 non-null  float64
 9   casual      10886 non-null  float64
 10  registered  10886 non-null  float64
 11  count       10886 non-null  float64
dtypes: float64(6), int64(5), object(1)
memory usage: 1.6+ MB
```



	feature	datatype	null_values	null_percentage	unique_values	unique_sample
0	datetime	object	0	0.00	17379	[2011-01-01 00:00:00, 2011-01-01 01:00:00, 201...
1	season	int64	0	0.00	4	[1, 2, 3, 4]
2	holiday	int64	0	0.00	2	[0, 1]
3	workingday	int64	0	0.00	2	[0, 1]
4	weather	int64	0	0.00	4	[1, 2, 3, 4]
5	temp	float64	0	0.00	50	[9.84, 9.02, 8.2, 13.12, 15.58, 14.76, 17.22, ...
6	atemp	float64	0	0.00	65	[14.395, 13.635, 12.88, 17.425, 19.695, 16.665...
7	humidity	int64	0	0.00	89	[81, 80, 75, 86, 76, 77, 72, 82, 88, 87]
8	windspeed	float64	0	0.00	30	[0.0, 6.0032, 16.9979, 19.0012, 19.9995, 12.99...
9	casual	float64	6493	37.36	309	[3.0, 8.0, 5.0, 0.0, 2.0, 1.0, 12.0, 26.0, 29....
10	registered	float64	6493	37.36	731	[13.0, 32.0, 27.0, 10.0, 1.0, 0.0, 2.0, 7.0, 6...
11	count	float64	6493	37.36	822	[16.0, 40.0, 32.0, 13.0, 1.0, 2.0, 3.0, 8.0, 1...

SCOPE OF DATA

FEATURE

Categorical Description

datetime : hourly date + timestamp

season: 1 = spring, 2 = summer, 3 = fall, 4 = winter

holiday: whether the day is a holiday or not (1/0)

workingday: whether the day is neither a weekend nor holiday (1/0)

weather:

- 1: Clear, Few clouds, Partly cloudy, Partly cloudy (clear).
- 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist (Mist)
- 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds (Light)
- 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog temp - temperature in Celsius (heavy)

LABEL

Count

number of total rentals.

count = casual + registered.

FEATURE

Numerical Description

temp: hourly temperature in Celsius

atemp: "feels like" temperature in Celsius

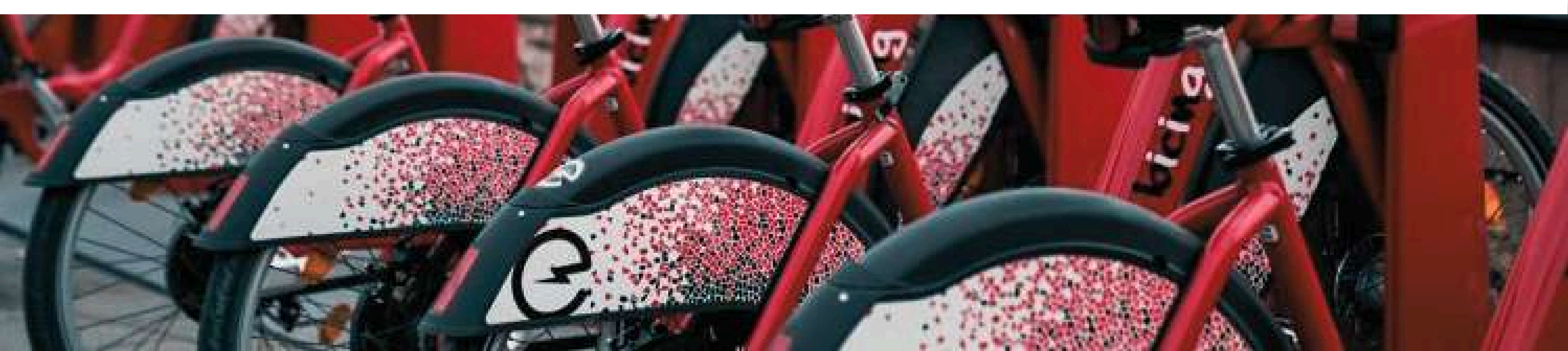
humidity: relative humidity

windspeed : wind speed

casual : number of non-registered user rentals initiated

registered : number of registered user rentals initiated

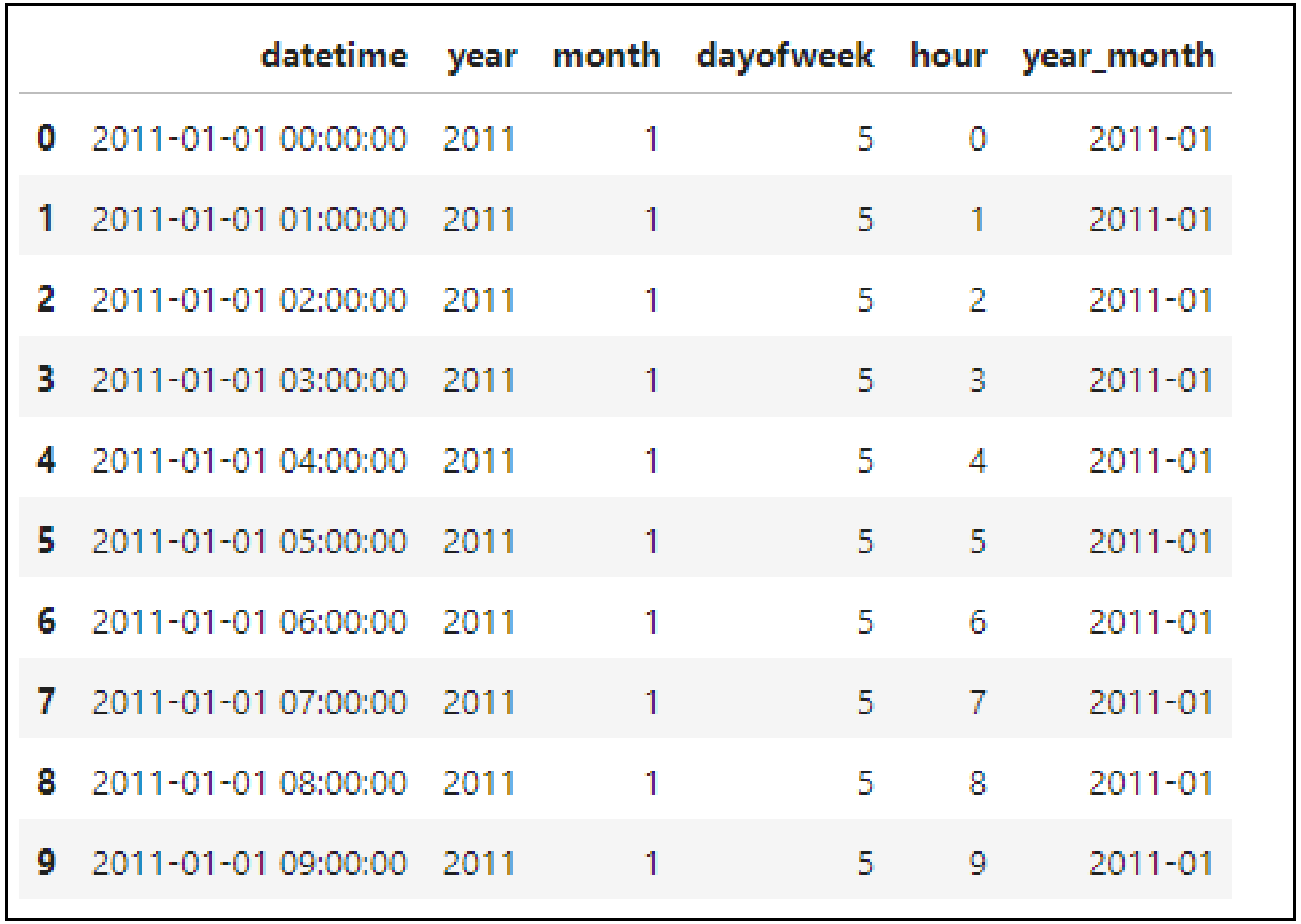
03 | Data Preparation



DATA EXTRACTION

We are current focusing on identifying factors and customer behavior to analyze and predict demand based on the number of customers.

To enhance our analysis, **we will extract and transform the datetime data into new features** such as year, month, day of the week, hour, and year-month. Adding these features will enrich our dataset, providing more detailed information and improving the accuracy of our demand predictions.



	datetime	year	month	dayofweek	hour	year_month
0	2011-01-01 00:00:00	2011	1	5	0	2011-01
1	2011-01-01 01:00:00	2011	1	5	1	2011-01
2	2011-01-01 02:00:00	2011	1	5	2	2011-01
3	2011-01-01 03:00:00	2011	1	5	3	2011-01
4	2011-01-01 04:00:00	2011	1	5	4	2011-01
5	2011-01-01 05:00:00	2011	1	5	5	2011-01
6	2011-01-01 06:00:00	2011	1	5	6	2011-01
7	2011-01-01 07:00:00	2011	1	5	7	2011-01
8	2011-01-01 08:00:00	2011	1	5	8	2011-01
9	2011-01-01 09:00:00	2011	1	5	9	2011-01

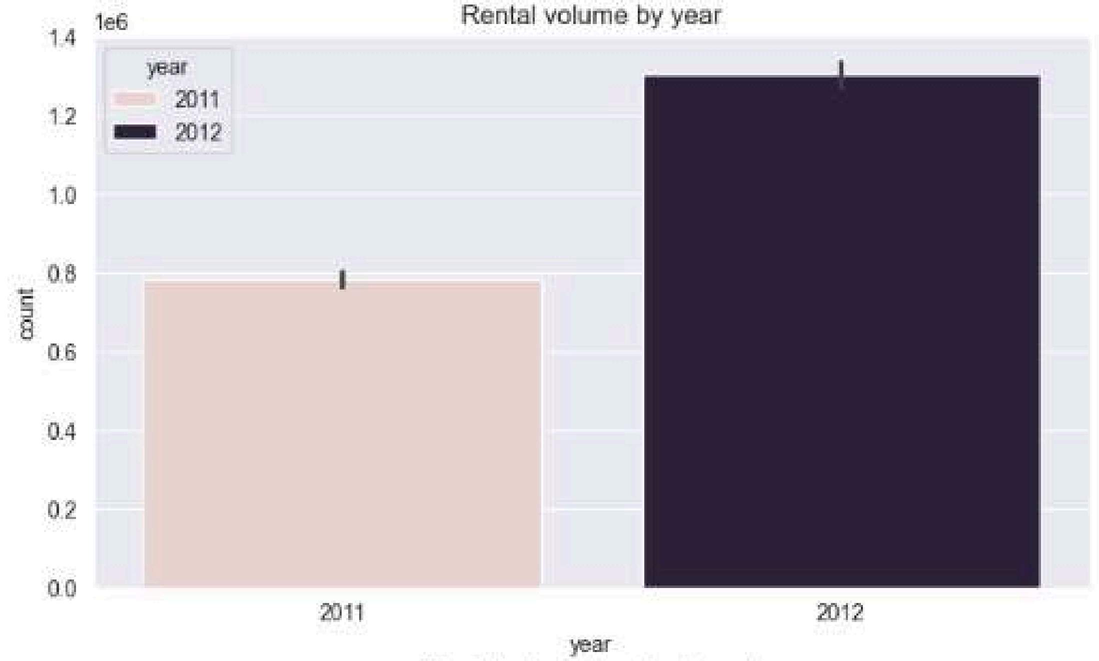
04 | EDA & Insight

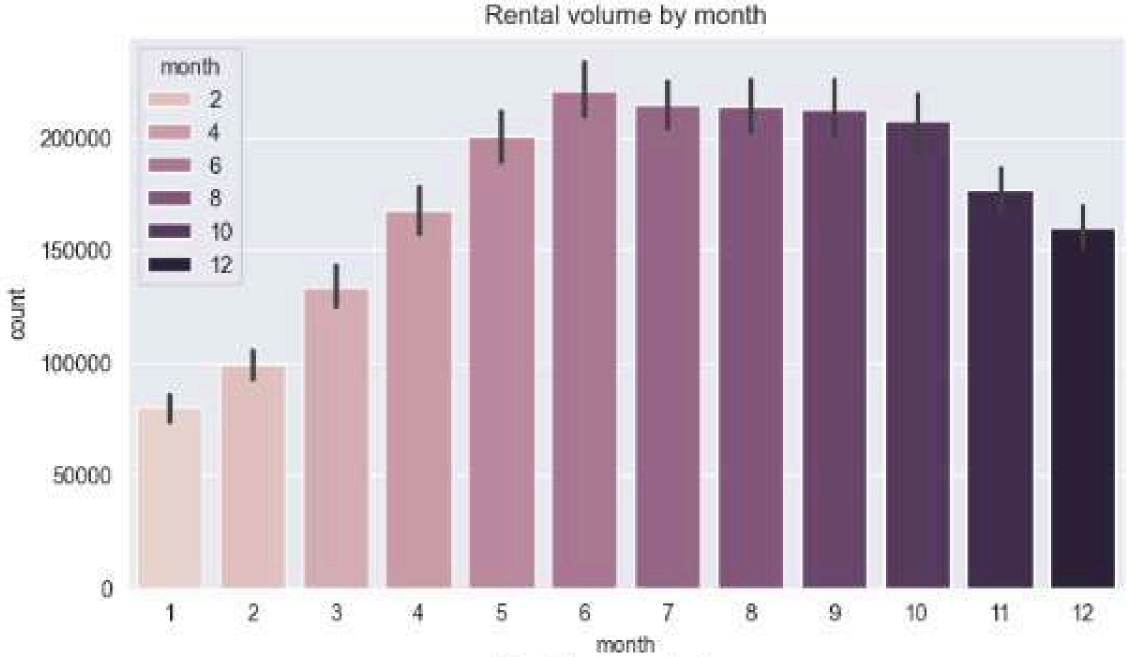
RENTAL BIKE VOLUME BY THE TIME

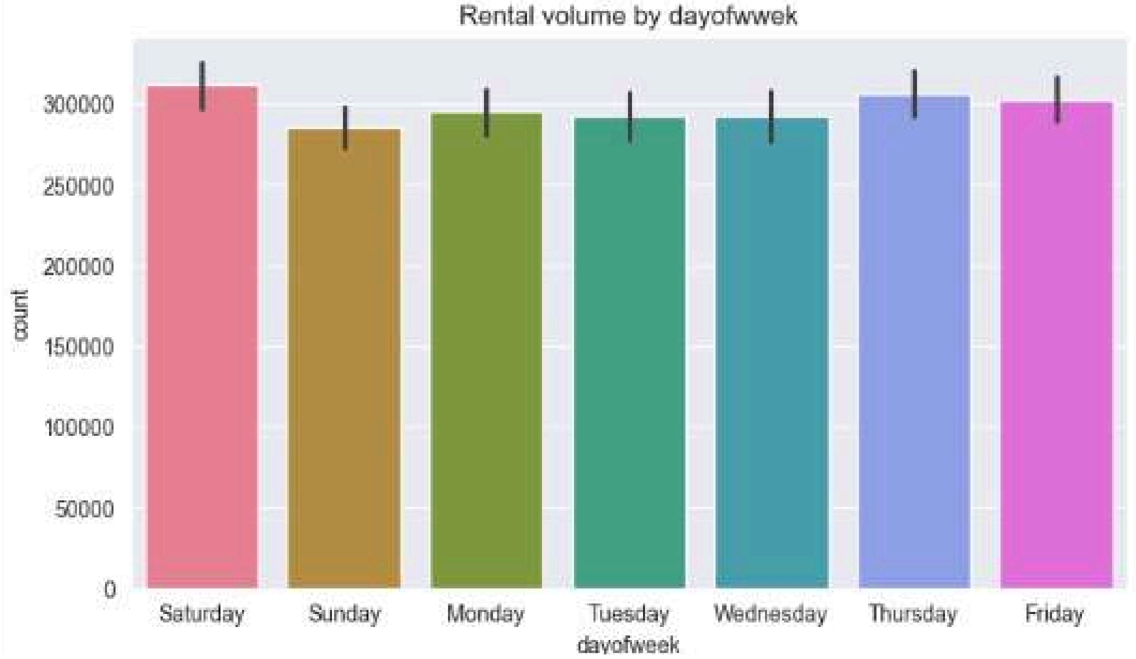
- **Seasonal Trends:** Rental volume was higher in 2012 compared to the previous year, with a steady increase from January to June, stabilizing until October, and then significantly decreasing. This suggests higher bike usage during warmer months.

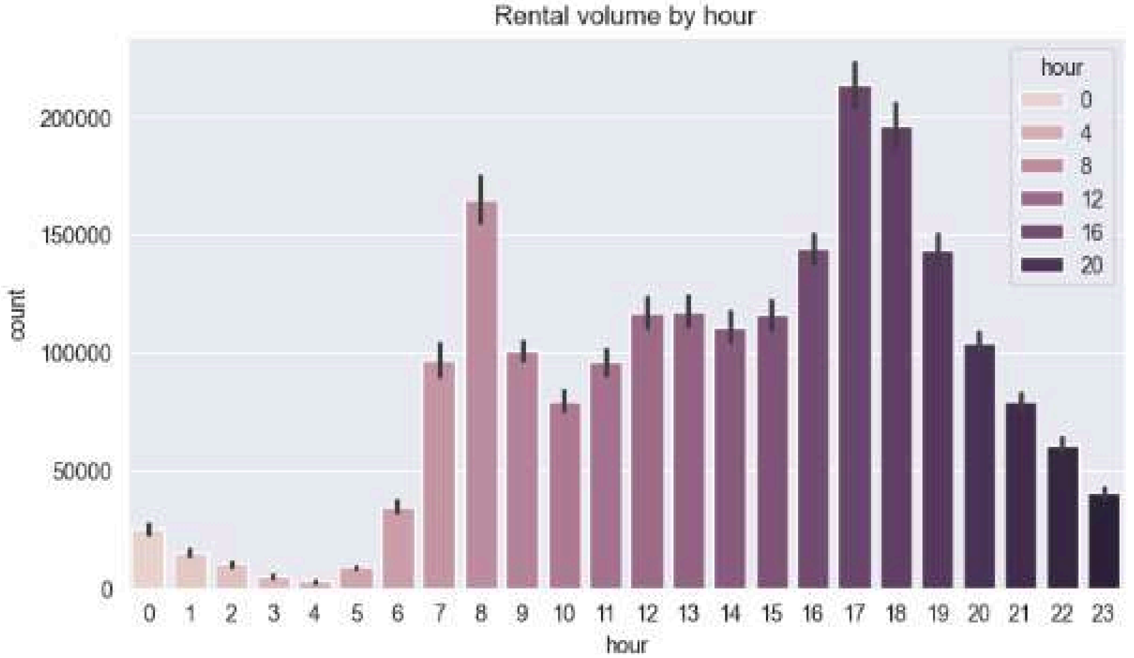
- **Weekly Patterns:** Rental volume steadily increases during weekdays, peaks on Saturday, and drops to its lowest on Sunday. This suggests Sunday is a rest day, resulting in common low demand.

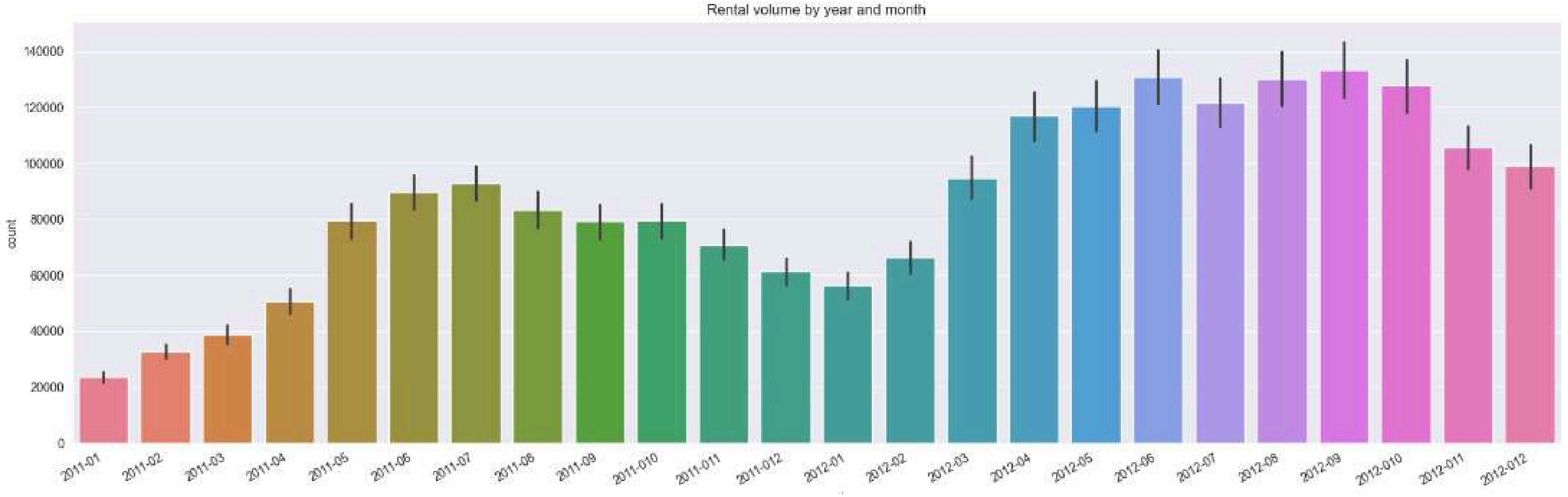
- **Hourly Demand:**
  - **High demand:** from 7-9 AM and 4-7 PM
  - **Medium demand:** from 10 AM to 3 PM
  - **Low demand:** from 12-6 AM and 9 PM to midnight.


Rental volume by year

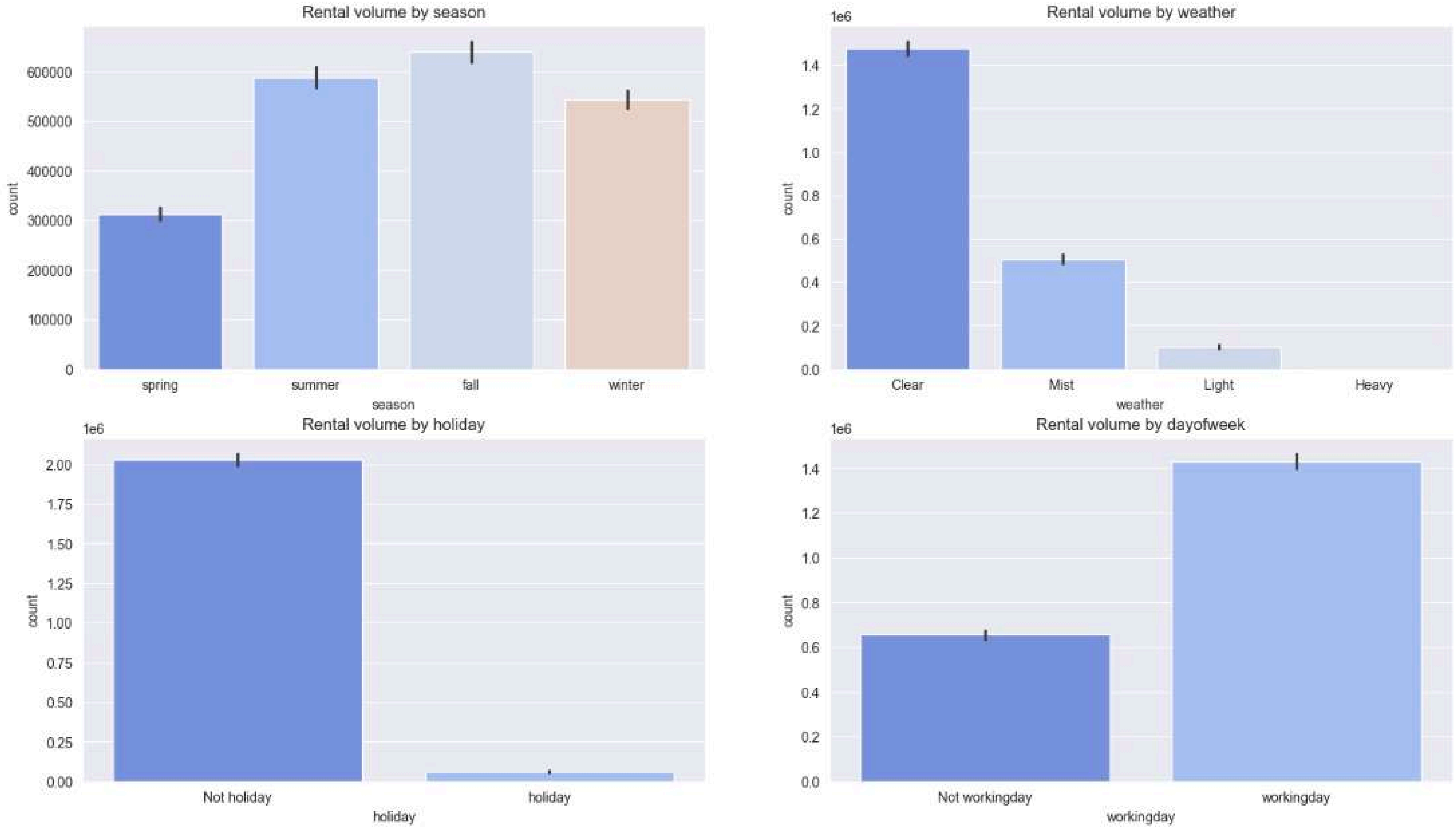

Rental volume by month


Rental volume by dayofwweek


Rental volume by hour


Rental volume by year and month

RENTAL BIKE VOLUME BY CONDITION AND USAGE PATTERN



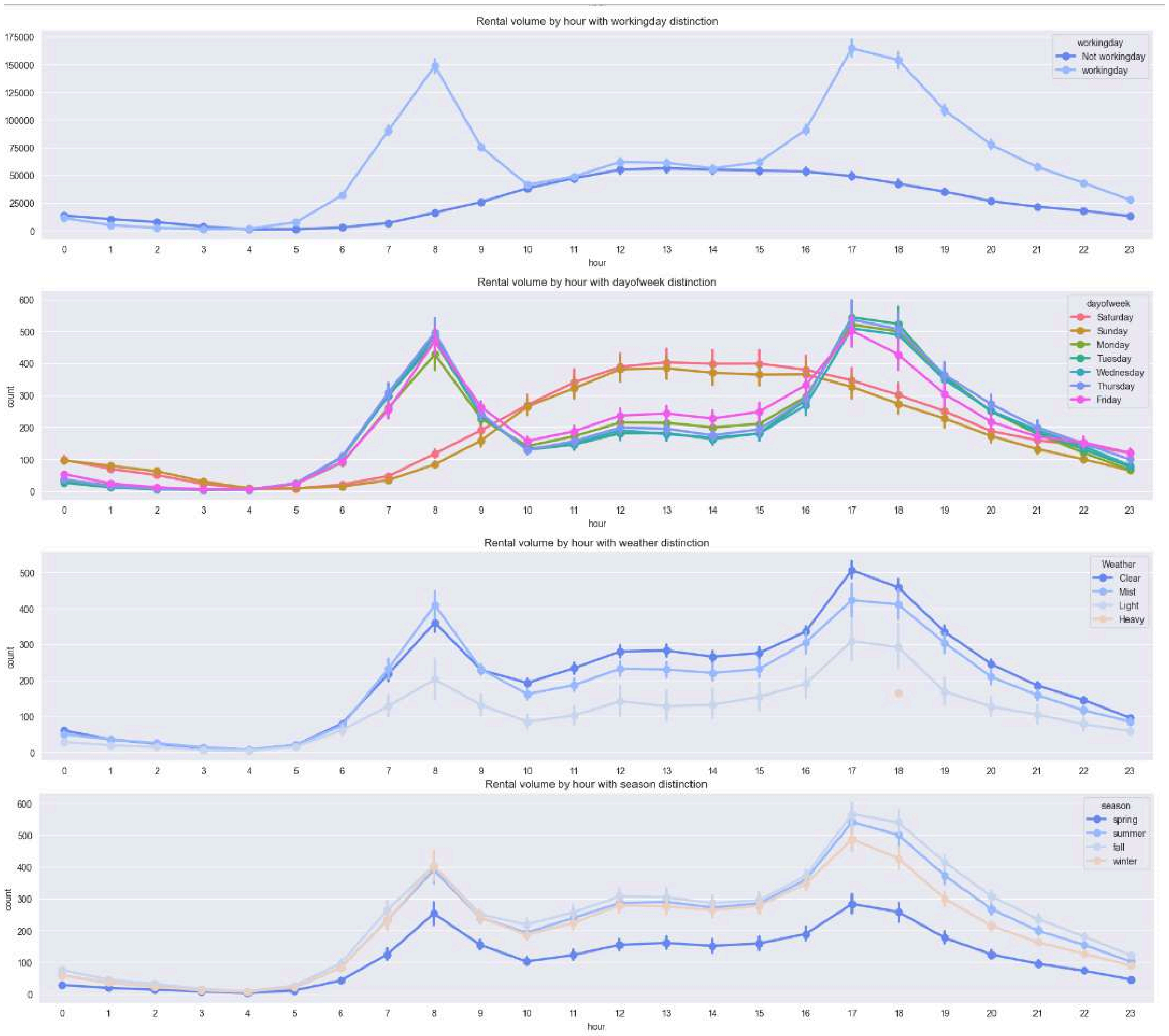
- **Seasonal Trends:** Peak bike usage in Fall and Summer due to favorable weather.
- **Weather Conditions:** Ideal conditions are clear weather, some prefer biking in misty or overcast conditions, while heavy rain or snow decreases usage.
- **Usage Patterns:** More bike usage on workdays than on holidays, indicating bikes are used mainly for commuting.

RENTAL BIKE VOLUME BY HOUR WITH OTHER CONDITIONS

- **Weekend Usage Patterns:** High bike demand from 9 AM to 8 PM, indicating leisure use.

- **Weekday Usage Patterns:** Peak demand from 7 AM to 9 AM and 5 PM to 7 PM for commuting. Average demand from 10 AM to 4 PM, low demand from 12 AM to 6 AM and 8 PM to 12 AM.

- **Weather and Seasonal Influence:** Rental volume follows similar patterns to weekday and seasonal trends.



Rental volume by hour with workingday distinction

Rental volume by hour with dayofweek distinction

Rental volume by hour with weather distinction

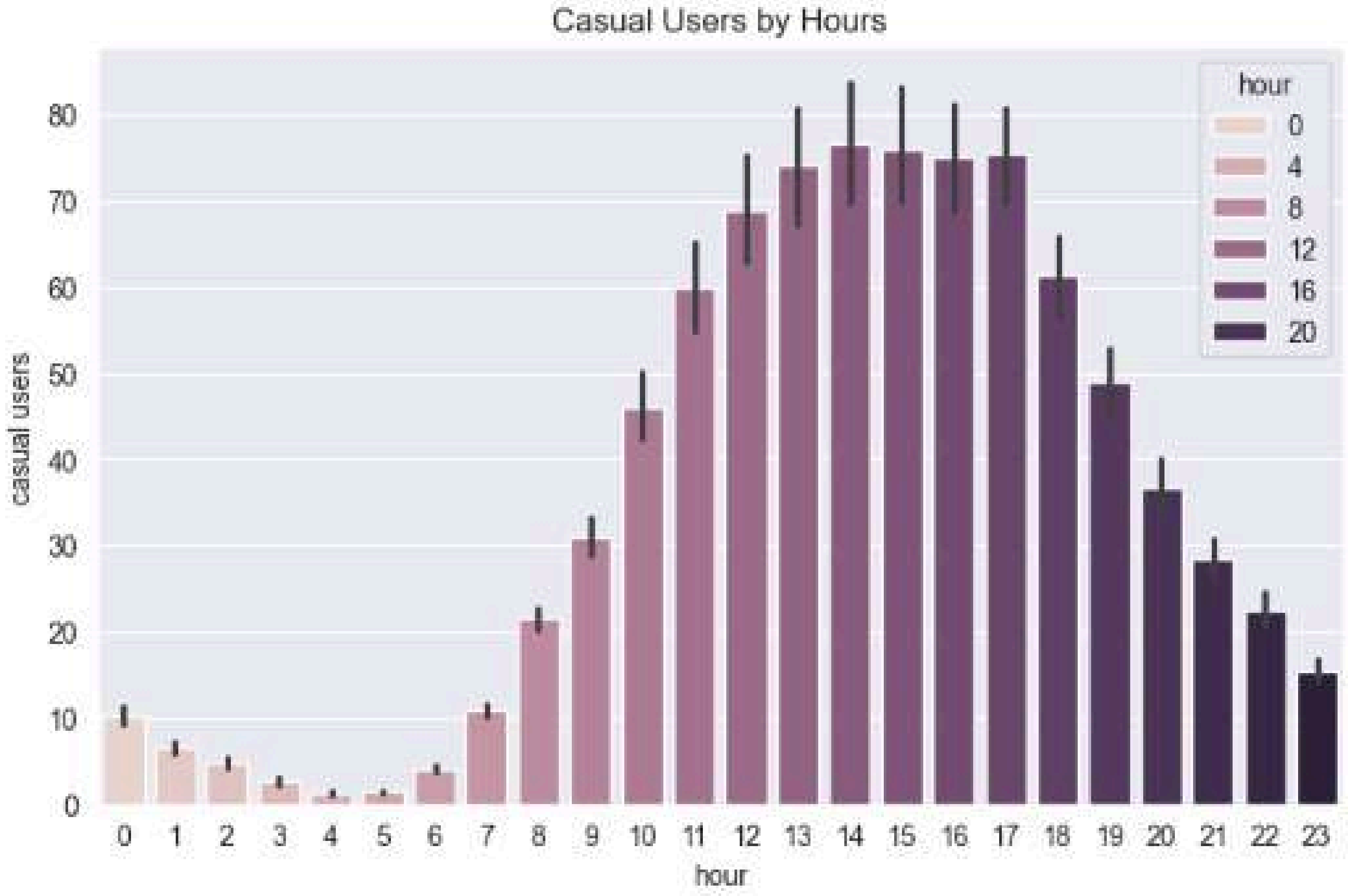
Rental volume by hour with season distinction

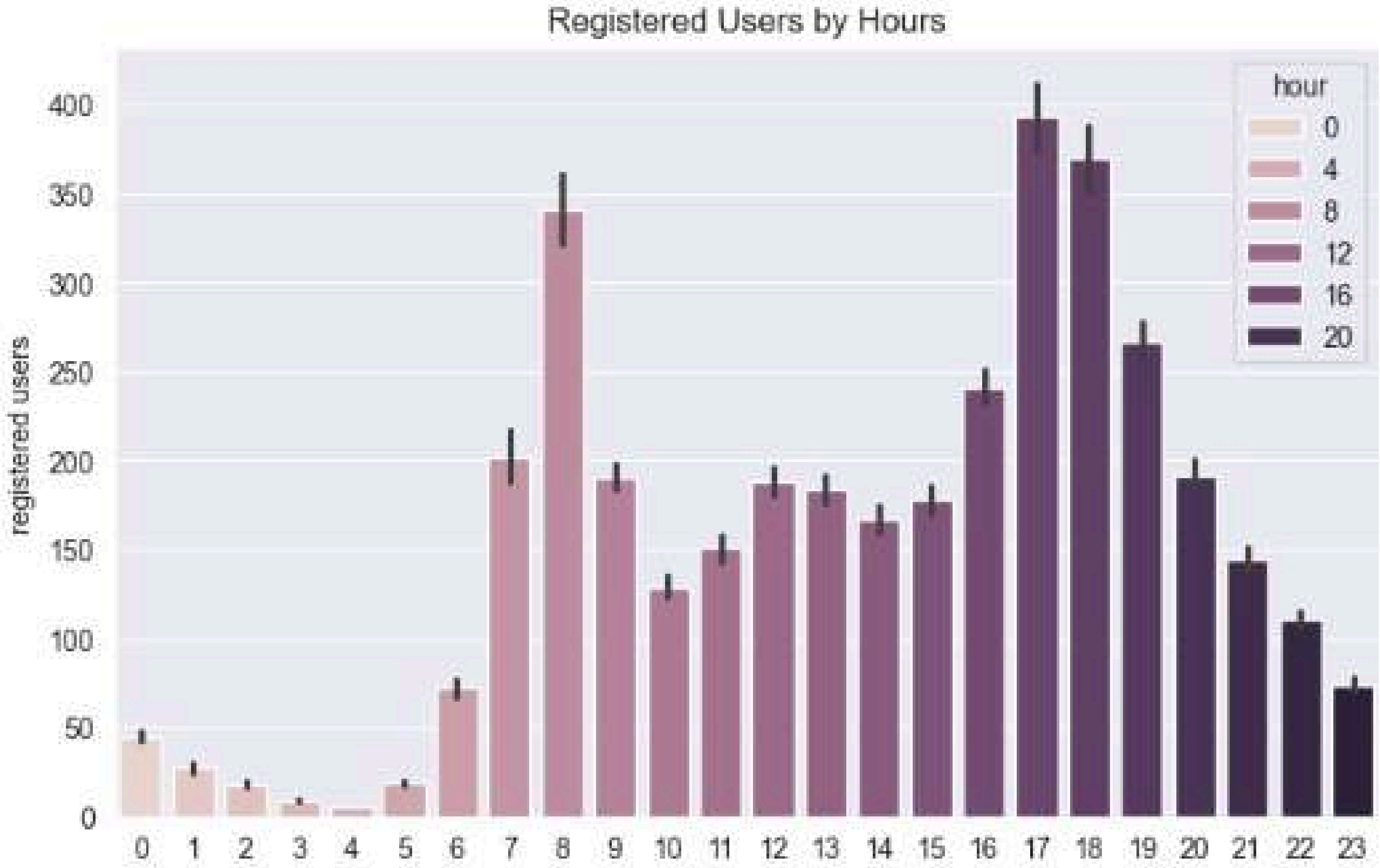
RENTAL BIKE VOLUME BY HOUR WITH TYPE OF USERS

User Behavior Patterns

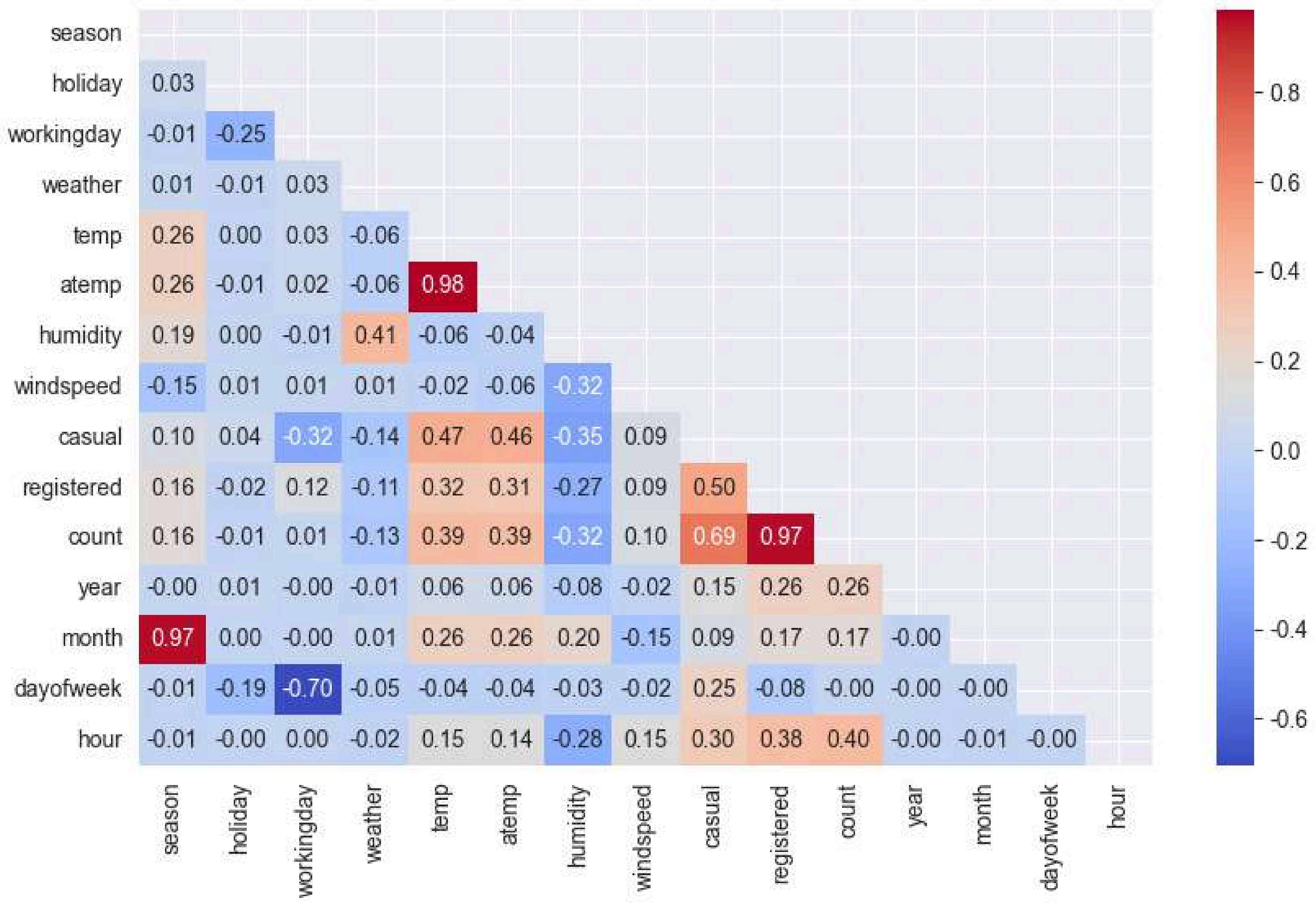
The dataset includes two types of users: registered and casual.

- **Casual Users:** These users usually ride bikes more frequently on weekends and non-working days. Their usage patterns suggest that they use bikes primarily for leisure and recreational activities rather than for daily commuting.
- **Registered Users:** In contrast, registered users tend to ride bikes more often on weekdays, especially during peak commuting hours. This pattern indicates that registered users likely use bikes as part of their regular commute to work or school.





CORRELATION EACH FEATURES ON HEATMAP



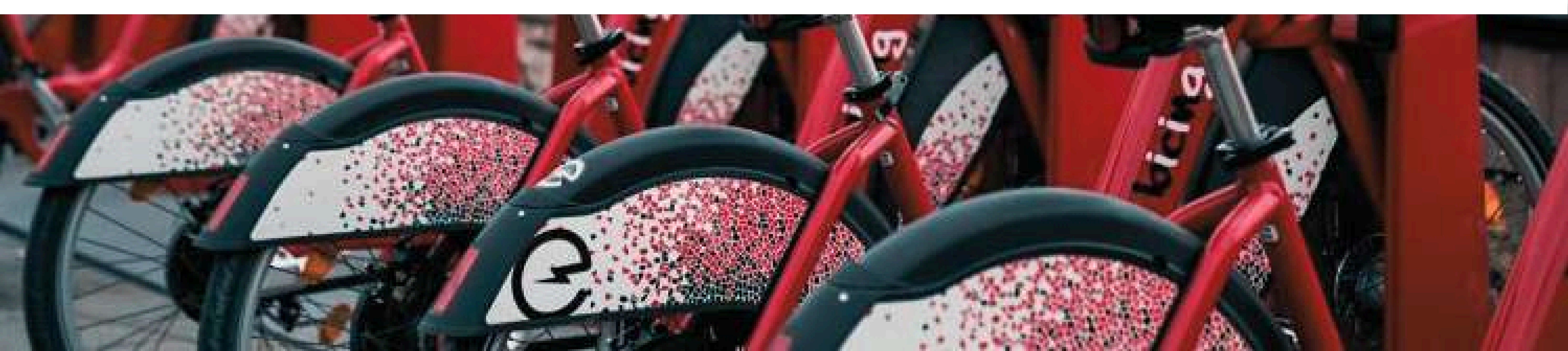
	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	casual	registered	count	year	month	dayofweek	hour
season															
holiday	0.03														
workingday	-0.01	-0.25													
weather	0.01	-0.01	0.03												
temp	0.26	0.00	0.03	-0.06											
atemp	0.26	-0.01	0.02	-0.06	0.98										
humidity	0.19	0.00	-0.01	0.41	-0.06	-0.04									
windspeed	-0.15	0.01	0.01	0.01	-0.02	-0.06	-0.32								
casual	0.10	0.04	-0.32	-0.14	0.47	0.46	-0.35	0.09							
registered	0.16	-0.02	0.12	-0.11	0.32	0.31	-0.27	0.09	0.50						
count	0.16	-0.01	0.01	-0.13	0.39	0.39	-0.32	0.10	0.69	0.97					
year	-0.00	0.01	-0.00	-0.01	0.06	0.06	-0.08	-0.02	0.15	0.26	0.26				
month	0.97	0.00	-0.00	0.01	0.26	0.26	0.20	-0.15	0.09	0.17	0.17	-0.00			
dayofweek	-0.01	-0.19	-0.70	-0.05	-0.04	-0.04	-0.03	-0.02	0.25	-0.08	-0.00	-0.00	-0.00		
hour	-0.01	-0.00	0.00	-0.02	0.15	0.14	-0.28	0.15	0.30	0.38	0.40	-0.00	-0.01	-0.00	

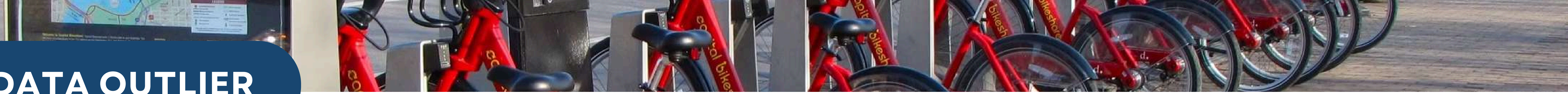
- "Feels-like Temperature" (atemp) and "Temperature" (temp) exhibit a strong correlation (> 0.75) with bike rental demand. Therefore, **atemp may be excluded from modeling to avoid redundancy.**

- **"Month" can be omitted to prevent redundancy with other temporal** features like "Season" and "Day of the Week," enhancing model efficiency.

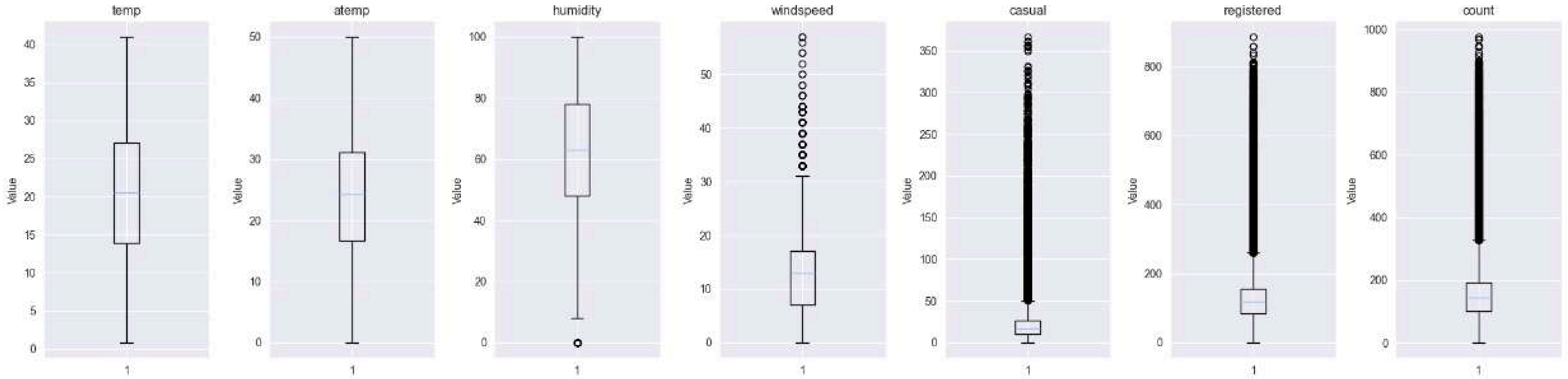
- **"Casual" and "Registered" users may be removed due to redundancy with the total count of rentals**, simplifying model complexity.

05 | Data Preprocessing



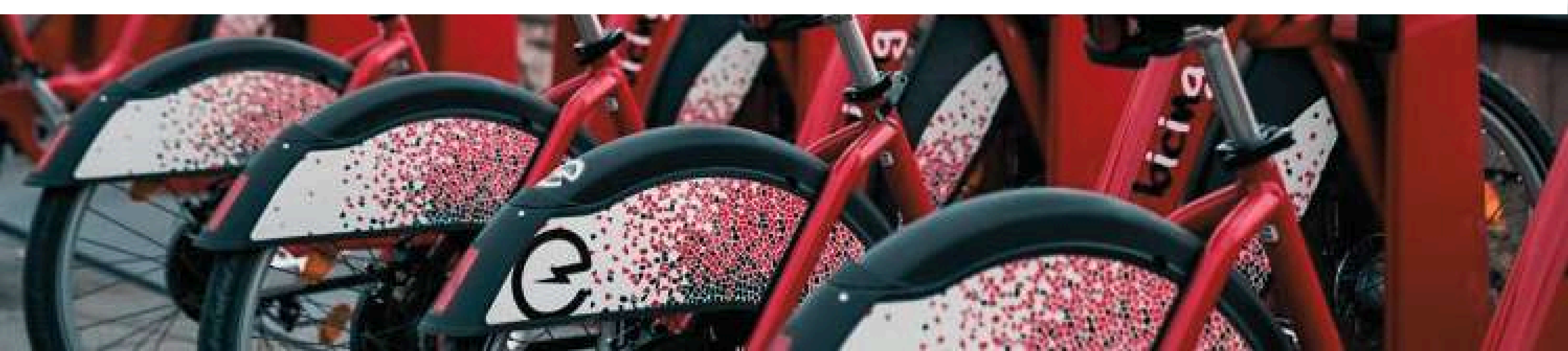


DATA OUTLIER



We also examined outliers in each numeric feature, particularly noting outliers in windspeed, humidity. Additionally, outliers were observed in the counts of casual and registered users. **While these outliers are not mistakes but normal variations, having many outliers in the number of registered and casual users may suggest that user behavior is quite different at certain times or that special situations caused unusually high or low bike rental numbers.**

06 | MACHINE LEARNING MODEL

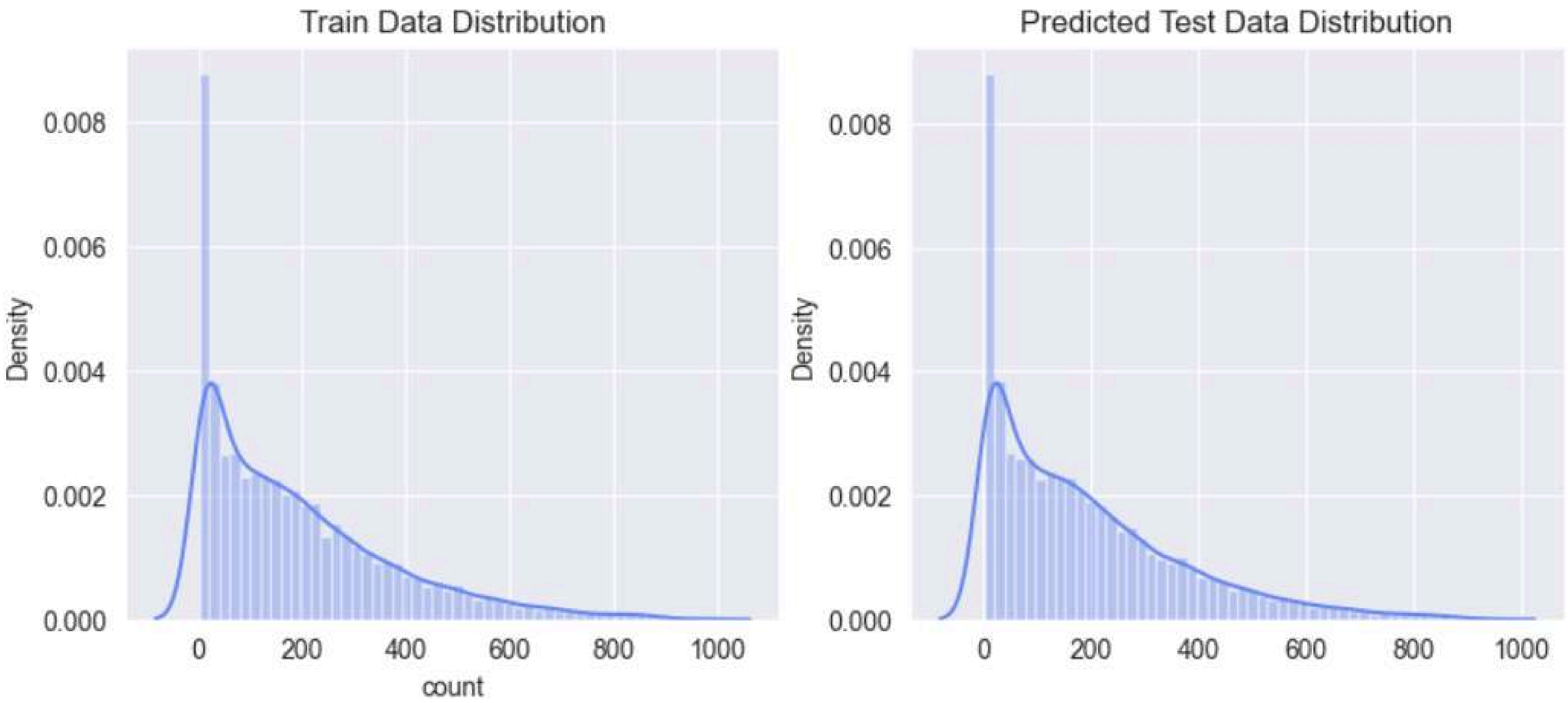


BENCKMARK MODEL

Algorithm	RMSLE Score	RMSLE Cross Val Score	RMSLE Score (Tuned)	RMSLE Cross Val Score (Tuned)
KNN Regressor	0.1332	0.19214	0.01804	0.79392
Decision Tree Regressor	0.00531	0.14725	0.34117	0.5205
Random Forest Regressor	0.03645	0.12158	0.44429	0.60559
Gradient Boosting Regressor	0.09976	0.12394	0.60705	0.70698

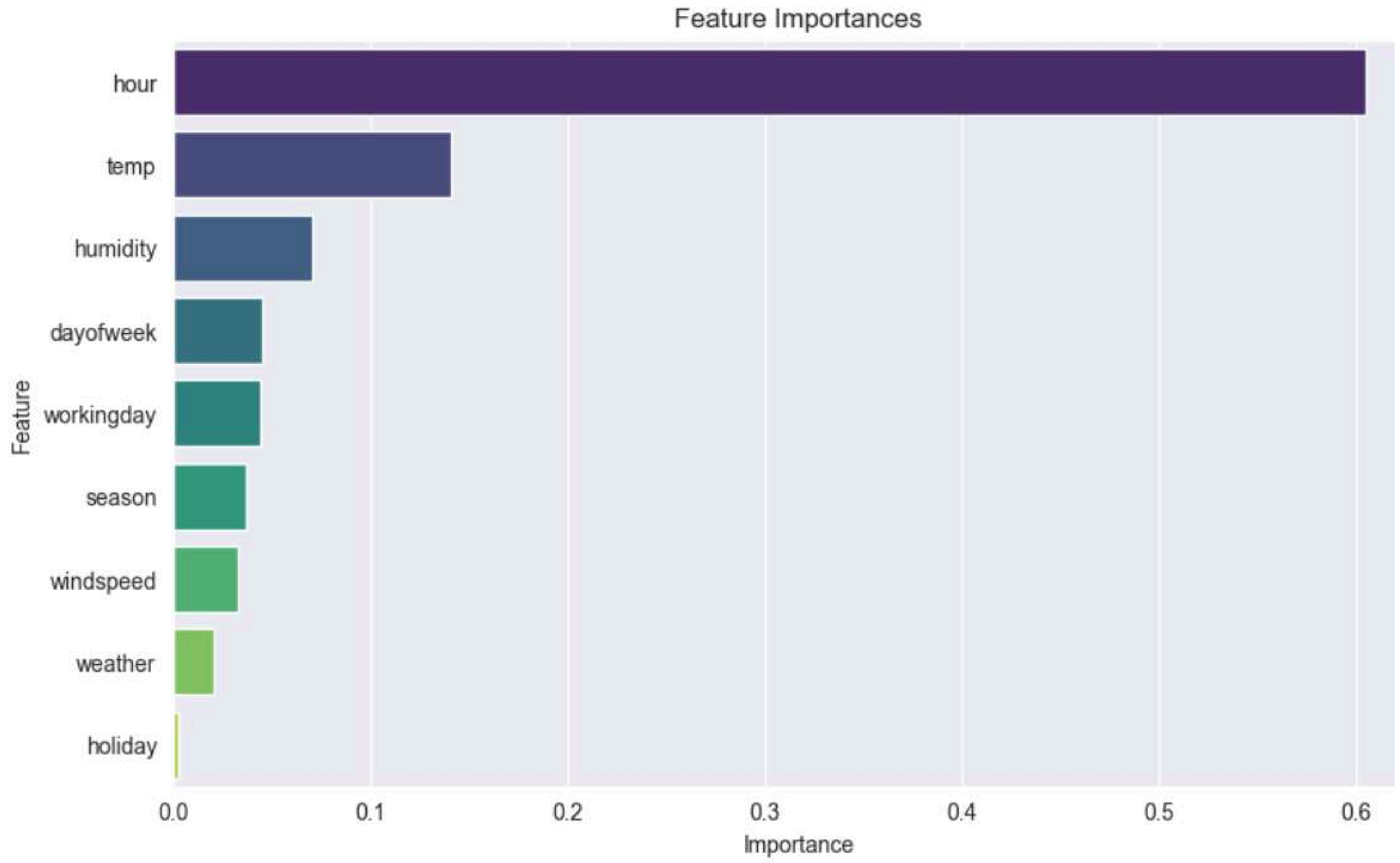
- **Evaluation Metric:** Models are evaluated using RMSLE to identify the best performer.
- **Best Model:** Random Forest showed the best performance, with low RMSLE and cross-validation scores, indicating good generalization.
- **Model Comparison:**
  - **Random Forest:** Slight overfitting after tuning, but less severe than others.
  - **KNN and Gradient Boosting:** Significant overfitting after tuning.
  - **Decision Tree:** The performance got worse after tuning.
- **Conclusion:** Random Forest stands out for its balance of accuracy and stability.

DISTRIBUTION OF BEST MODEL



Here's a visualization of the distribution of the train and prediction data

FEATURE IMPORTANT

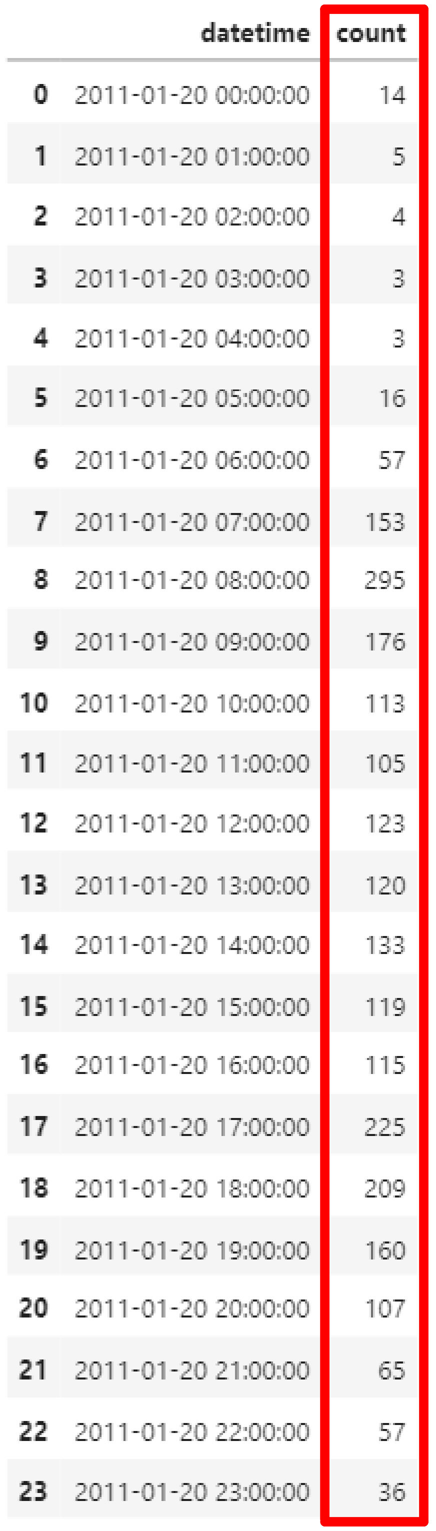


- **Hour:** The hour of the day is the most critical factor affecting rental demand. Demand peaks during morning and evening rush hours when people commute to work or school and drops during late hours.
- **Temp**: Moderate temperatures lead to higher demand as they are more comfortable for cycling, whereas extreme temperatures (very hot or very cold) reduce the number of rentals.
- **Humidity:** High humidity levels can make cycling uncomfortable, thereby decreasing the number of rentals.
- **Day of Week, Working Day, and Season:** These features collectively influence rental demand patterns. Rental demand is generally higher on weekdays compared to weekends, with Saturday being an exception where demand peaks and Sunday sees a significant drop in rentals. Working days see higher rentals during commute times. Seasonality also affects demand, with higher rentals during favorable weather conditions.

ADD PREDICTION TO SUBMISSION DATA

The 'count' column (feature) represents the predicted demand for each time period (hour) within a day. **These predictions are obtained through a machine learning model trained using relevant features such as hour, weather, season, working day, and others.**

Adding the prediction results to the submission aims to present an estimate of the expected demand quantity for each specific time period. **This aids in inventory planning, resource allocation, and other decision-making processes in business management.**



	datetime	count
0	2011-01-20 00:00:00	14
1	2011-01-20 01:00:00	5
2	2011-01-20 02:00:00	4
3	2011-01-20 03:00:00	3
4	2011-01-20 04:00:00	3
5	2011-01-20 05:00:00	16
6	2011-01-20 06:00:00	57
7	2011-01-20 07:00:00	153
8	2011-01-20 08:00:00	295
9	2011-01-20 09:00:00	176
10	2011-01-20 10:00:00	113
11	2011-01-20 11:00:00	105
12	2011-01-20 12:00:00	123
13	2011-01-20 13:00:00	120
14	2011-01-20 14:00:00	133
15	2011-01-20 15:00:00	119
16	2011-01-20 16:00:00	115
17	2011-01-20 17:00:00	225
18	2011-01-20 18:00:00	209
19	2011-01-20 19:00:00	160
20	2011-01-20 20:00:00	107
21	2011-01-20 21:00:00	65
22	2011-01-20 22:00:00	57
23	2011-01-20 23:00:00	36

07 | CONCLUSION & RECOMMENDATION

CONCLUSION

1. **Seasonal Trends:**
   - Bike rental volume was higher in 2022 compared to the previous year, with an increase from March to June, stabilizing until October, and then decreasing significantly. Higher usage is observed during warmer months, particularly in summer and fall, with a decline in winter.
2. **Weekly and Hourly Patterns:**
   - Rentals increase during weekdays, peak on Saturdays, and drop on Sundays.
   - High demand occurs from 7-9 AM and 4-7 PM (commuting hours), medium demand from 10 AM to 3 PM, and low demand from 12-6 AM and 9 PM to midnight.
3. **Weather Conditions:**
   - Clear weather promotes higher bike usage. Moderate biking occurs in misty or overcast conditions, but demand significantly decreases during heavy rain or snow.
4. **Usage Patterns:**
   - Higher bike usage is observed on regular workdays compared to holidays, indicating bikes are a primary mode of transport for daily commutes.
   - Weekends show high demand from 9 AM to 8 PM, indicating leisure use.
   - Weekdays show peak demand during commuting hours (7-9 AM and 5-7 PM), with average demand mid-day and low demand late night.
5. **User Behavior:**
   - Casual users tend to use bikes more on weekends and non-working days.
   - Registered users primarily use bikes for weekday commuting

RECOMMENDATION

Optimizing Bike Availability by Hour and Weather

To effectively manage fluctuating demand, prioritize increasing bike availability during peak hours such as 7-9 AM and 4-7 PM (commuting hours) and warmer seasons. By adjusting the fleet based on real-time weather forecasts, resources can be used more effectively, helping to keep the service reliable.

Promotional Campaigns

Launch targeted promotional campaigns tailored to specific user segments to drive demand and foster consistent usage habits. Highlighting the financial benefits of riding during off-peak times or the convenience of biking in any weather can effectively encourage adoption and retention among diverse user groups.

Pricing Strategies

Implement dynamic pricing system to considering multiple factors such as time of day, season, day of the week, and weather conditions. Providing special discounts during off-peak hours, weekends, holidays, or bad weather can encourage more users during slower times, helping to increase revenue while keeping users satisfied.

Urban Planning

Invest in enhancing bike infrastructure and weather-proofing measures to elevate the biking experience. Prioritize integrating green spaces with biking routes to enhance environmental quality and community wellness.

THANK YOU!