BIKE RENTAL DEMAND PREDICTION IN WASHINGTON D.C

TABLE OF CONTENT

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 models or per-use charges. By encouraging cycling, bike sharing promotes a convenient, healthy, and eco-friendly mode of transportation in numerous cities globally.

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, increasing 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 indicates that the time of day plays a crucial role in estimating bike rental demand. Considering the busy periods when individuals are commuting can provide valuable insights 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 spring and summer often encourage outdoor activities, including biking.

02 | Data Understanding

DATA UNDERSTANDING

- There are three 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 17379 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

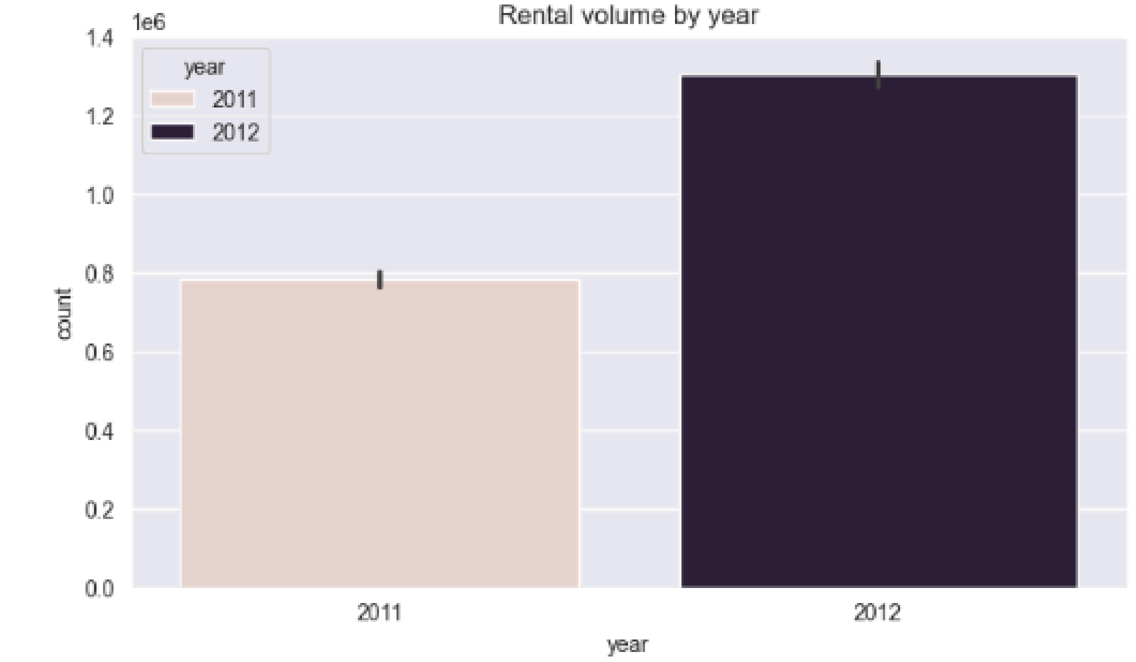
04 | EDA & Insight

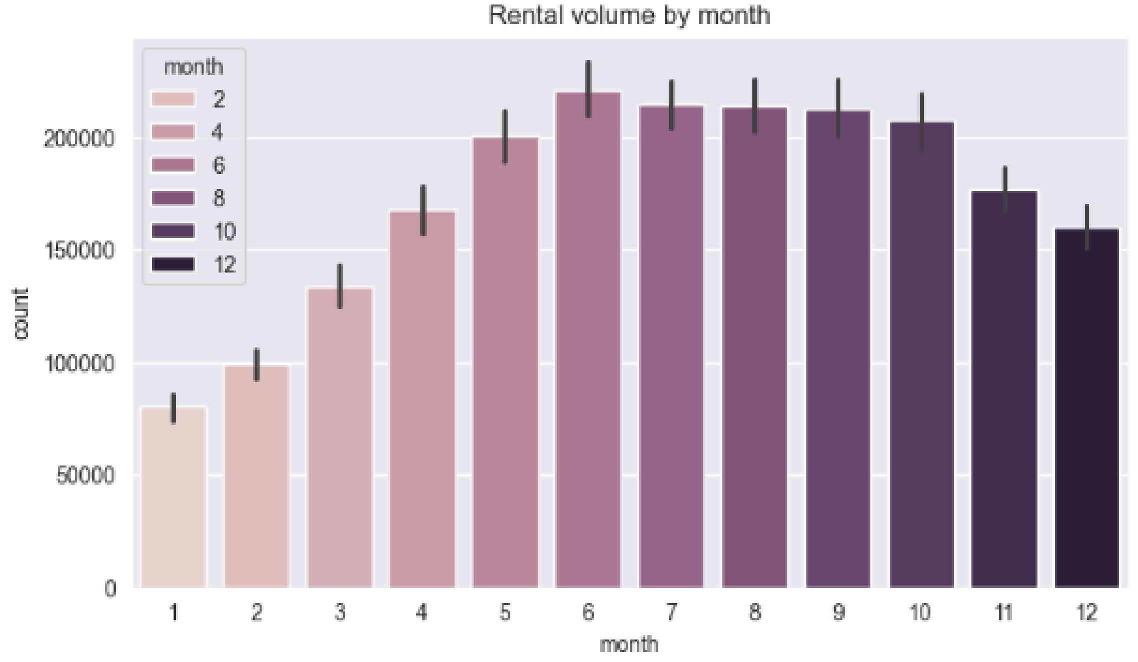
RENTAL BIKE VOLUME BY THE TIME

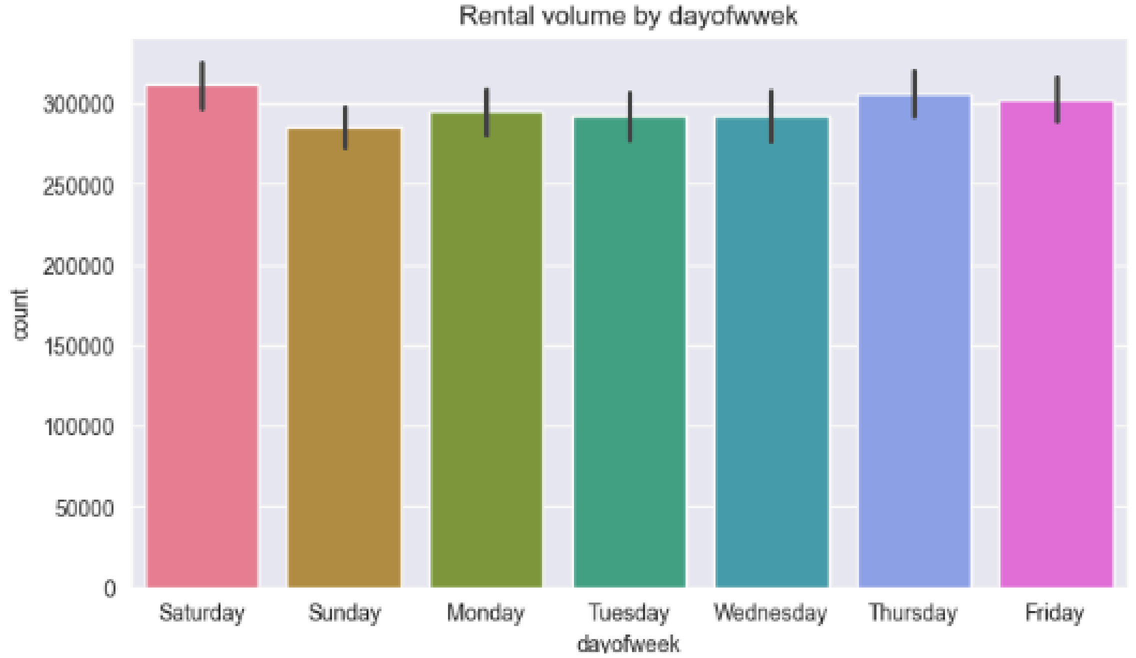
- **Seasonal Trends:** Rental volume was higher in 2022 compared to the previous year, with a steady increase from January to June, stabilizing until October, and then significantly decreasing. This suggests higher bike usage during warmer months.

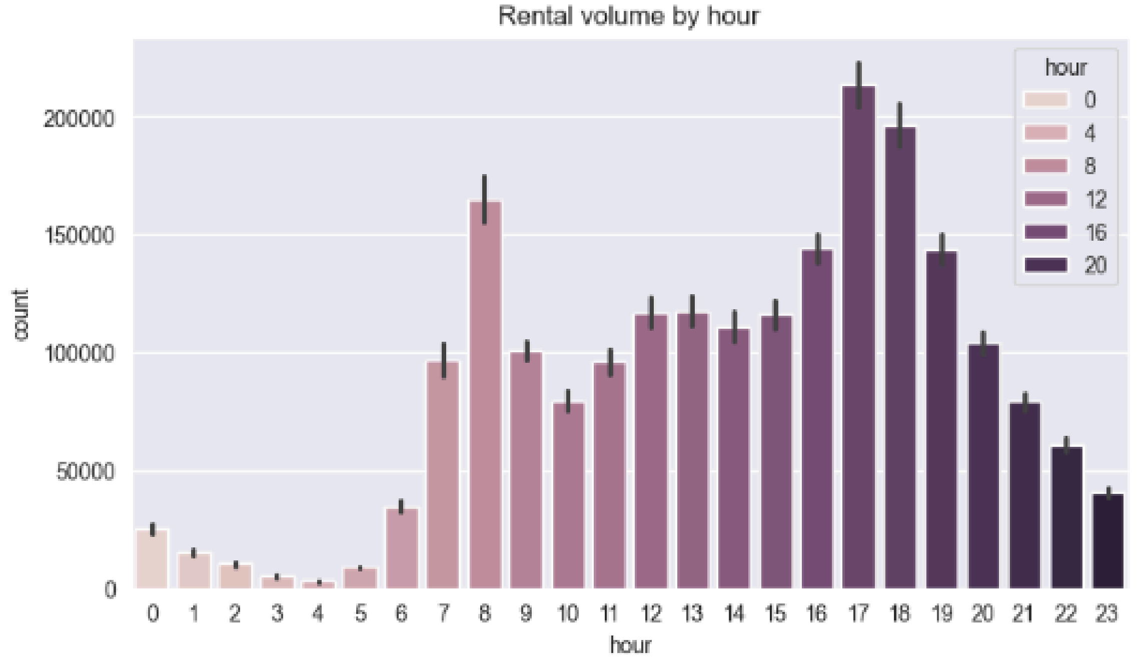
- **Weekly Patterns:** Rental volume steadily increases during weekdays, peaks on Saturday, and drops to its lowest on Sunday. This suggests Sunday is a rest day, resulting in common low demand.

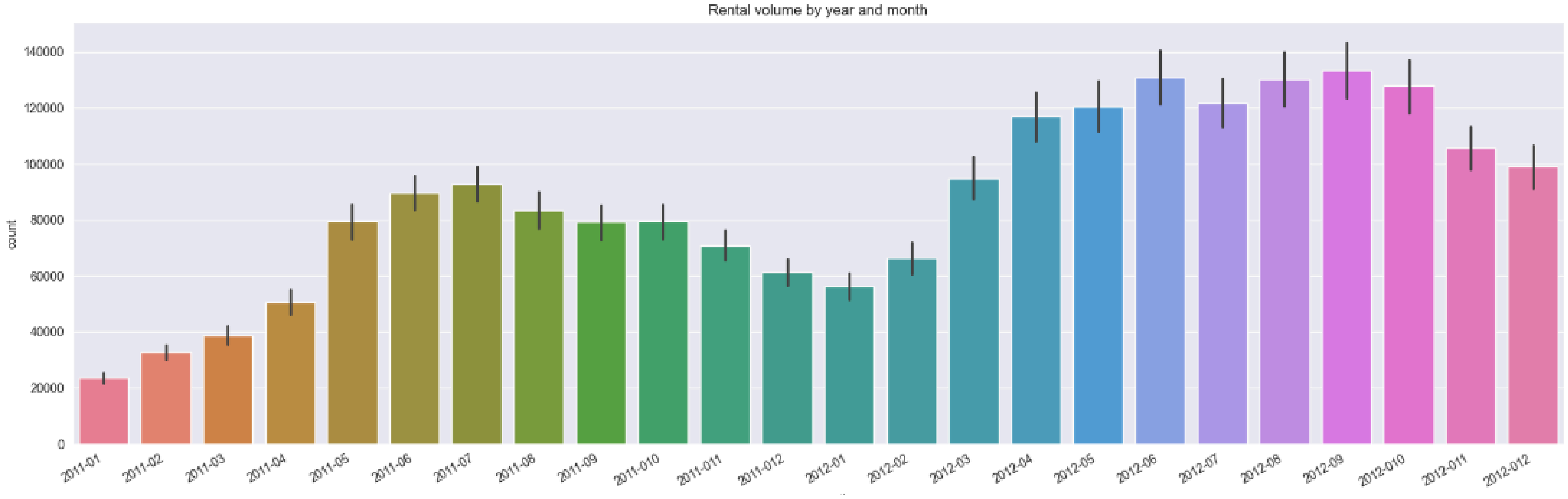
- **Hourly Demand:**
  - **High demand:** from 7-9 AM and 4-7 PM
  - **Medium demand:** from 10 AM to 3 PM
  - **Low demand:** from 12-6 AM and 9 PM to midnight.



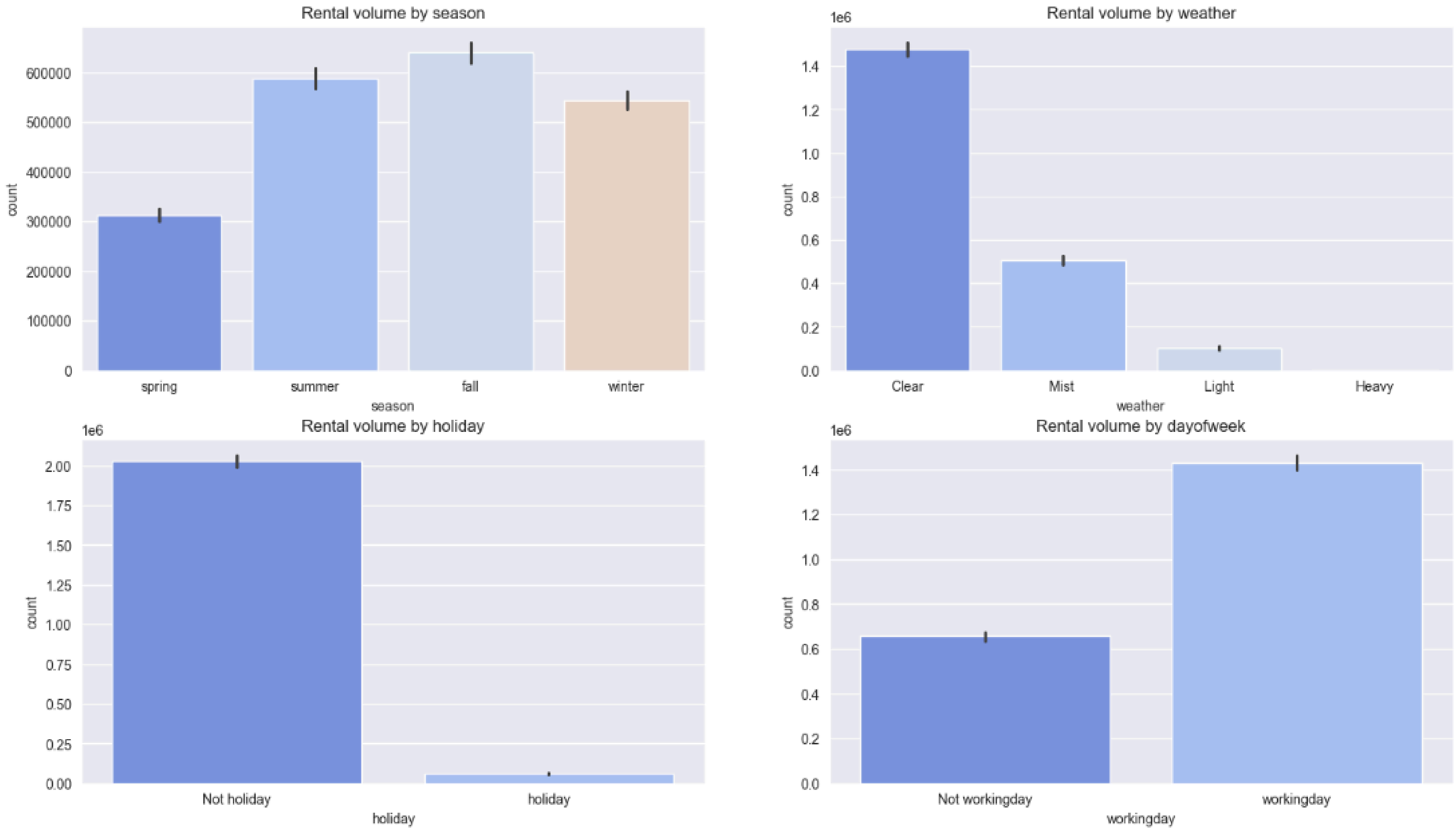








RENTAL BIKE VOLUME BY CONDITION AND USAGE PATTERN



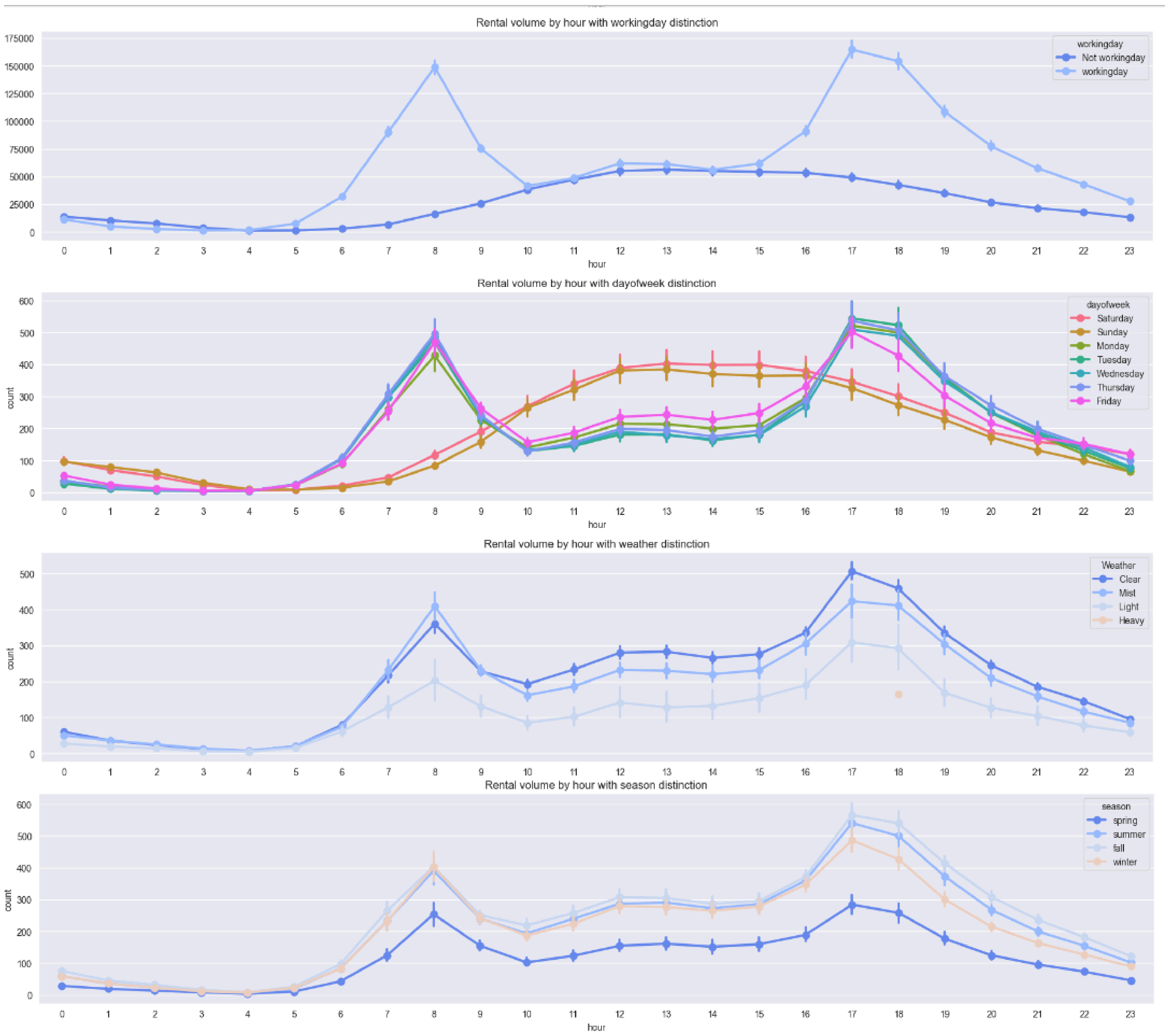
- **Seasonal Trends:** Fall, followed by summer, emerges as the peak season for bike usage, likely due to more favorable weather conditions. Conversely, winter sees a decline in usage due to colder weather. Spring exhibits a lower volume compared to others.
- **Weather Conditions:** Clear weather is ideal for bike usage, but there's still a subset of individuals who prefer biking in misty or overcast conditions. However, during seasons with heavier precipitation, such as heavy rain or snow, the preference for biking decreases significantly.
- **Usage Patterns:** Bike usage is more prevalent on regular workdays compared to holidays, indicating that many individuals use bikes as their primary mode of transportation for daily commutes.

RENTAL BIKE VOLUME BY HOUR WITH OTHER CONDITIONS

- **Weekend Usage Patterns:** On non-working days, particularly Saturdays and Sundays, high demand for bikes is consistently observed from 9 AM to 8 PM, indicating leisure-hour usage patterns among users.

- **Weekday Usage Patterns**: Conversely, on weekdays demand follows a different pattern, with high demand from 7 AM to 9 AM and from 5 PM to 7 PM, indicating peak commuting hours. There's an average demand from 10 AM to 4 PM and low demand from 12 AM to 6 AM and from 8 PM to 12 PM.

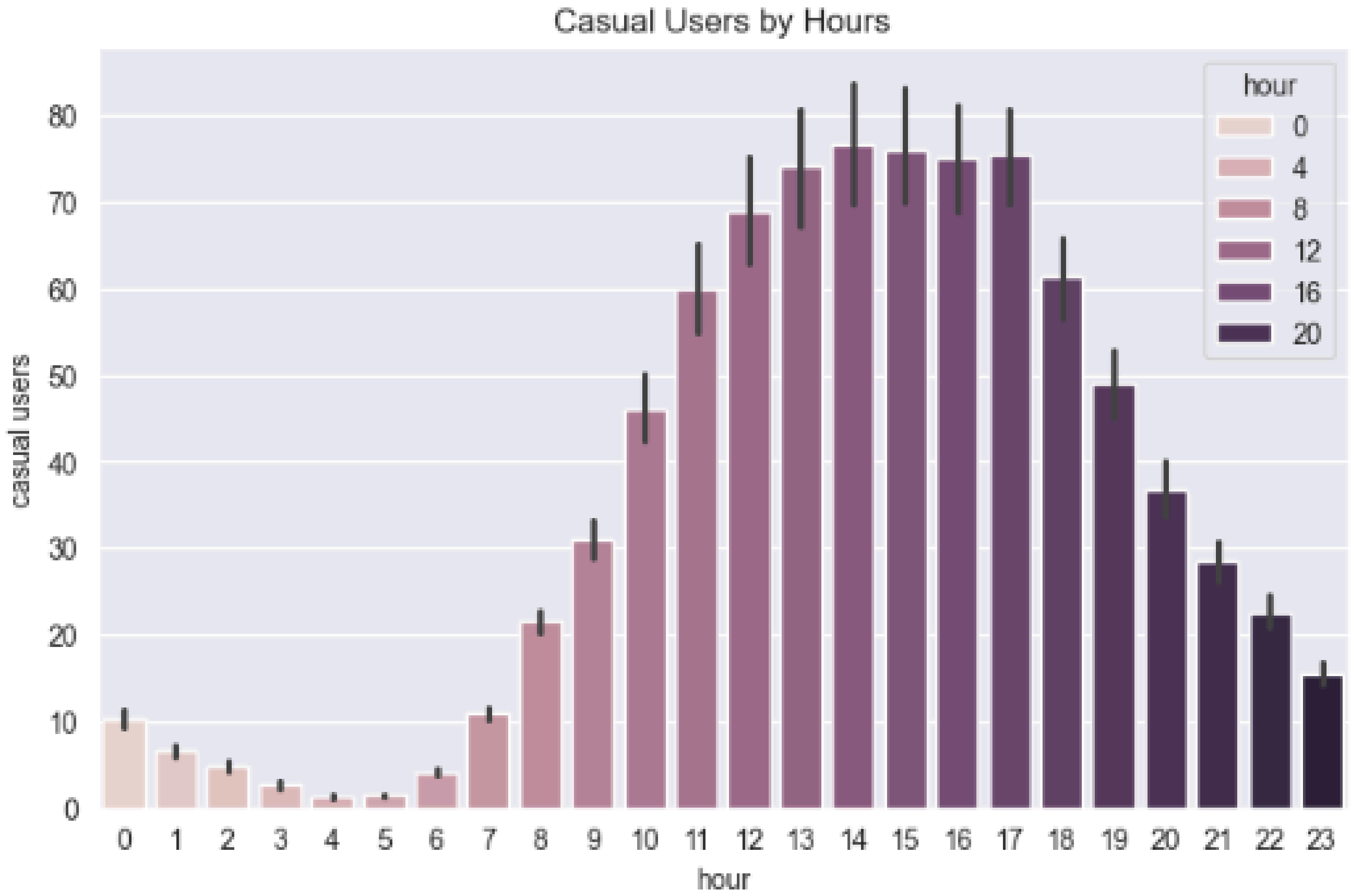
- **Weather and Seasonal Influence:** Weather and Seasonal Influence: Weather and seasonal patterns coincide with rental volume, following the weekday and working day patterns.

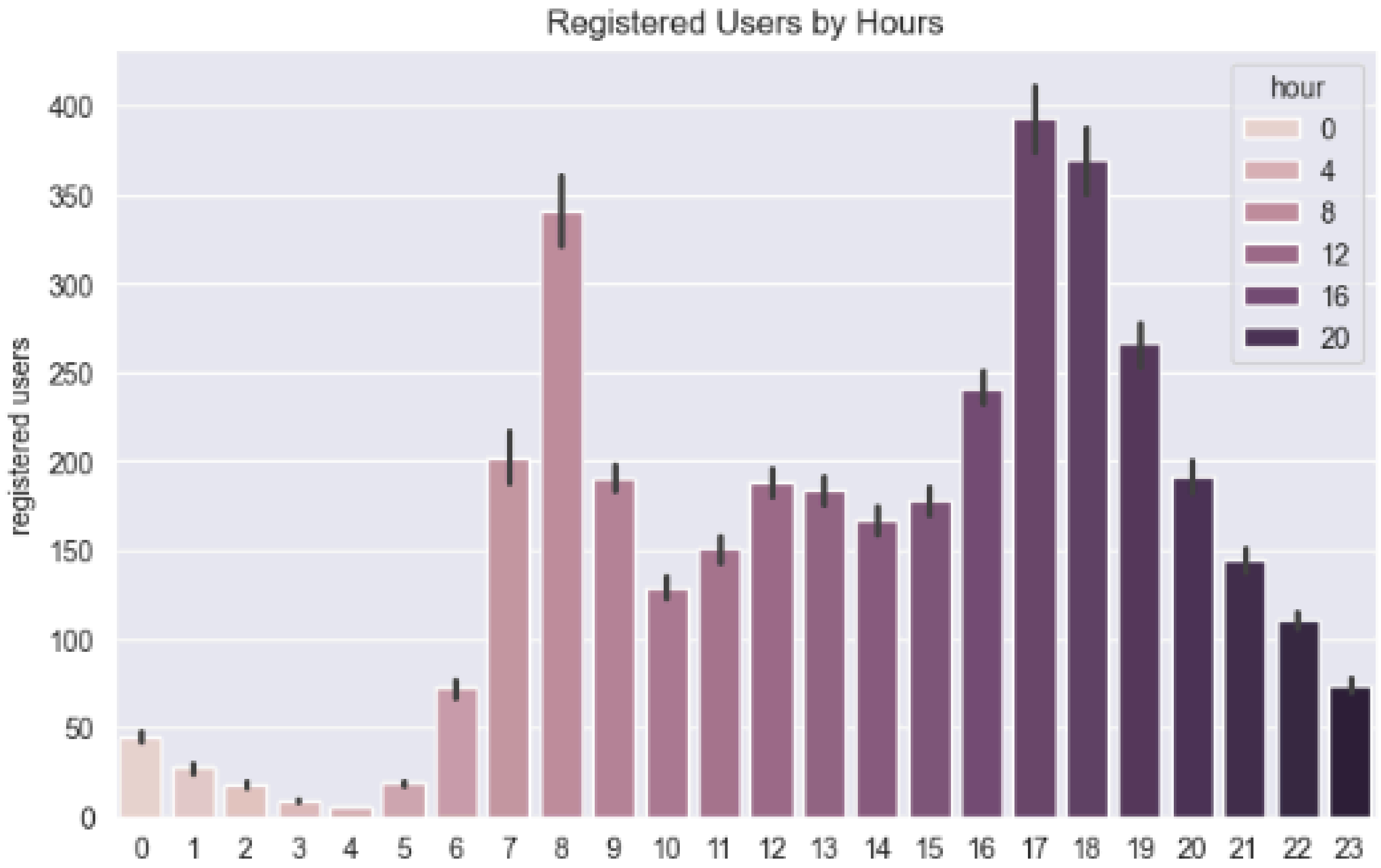


RENTAL BIKE VOLUME BY HOUR WITH TYPE OF USERS

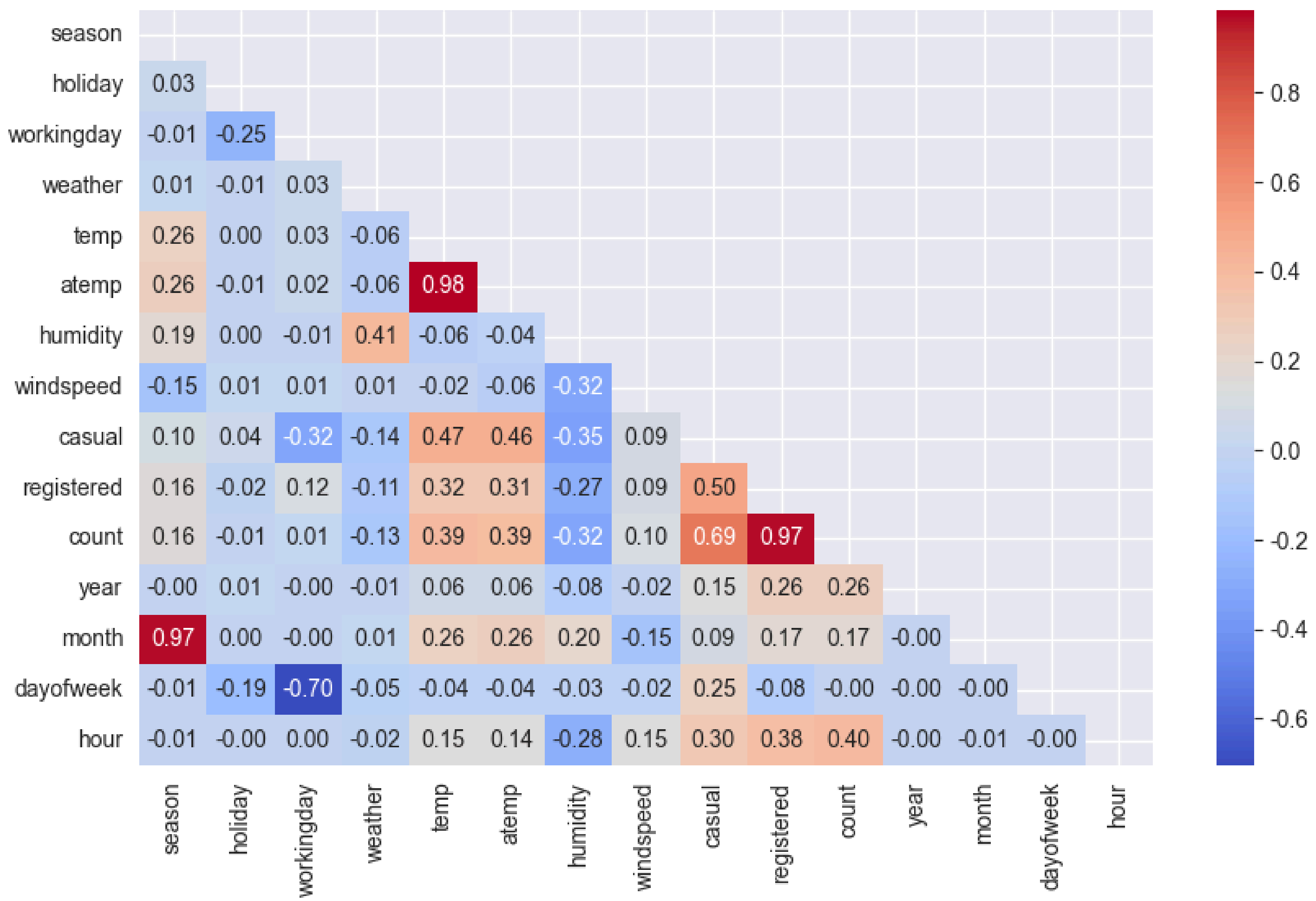
- **User Behavior Patterns:** The dataset contains two user types: registered and casual users. Casual users exhibit a usage pattern resembling weekends and non-working days, while registered users display a pattern more akin to weekdays and typical commuting hours.

- **Usage Preferences:** Casual users tend to use the bike-sharing service more during weekends and non-working days, while registered users prefer weekday commuting hours for utilizing bikes as a means of transportation to work or other commitments.





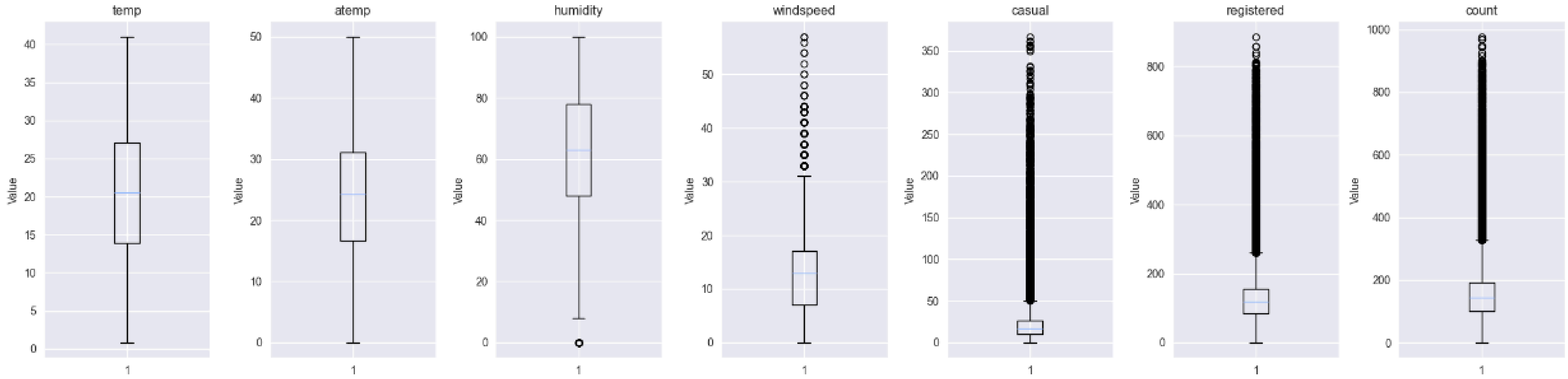
CORRELATION EACH FEATURES ON HEATMAP



- "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 natural occurrences and not errors, the presence of many outliers in the counts of registered and casual users might indicate variability in user behavior or exceptional circumstances that led to unusually high or low rental counts.**

06 | MACHINE LEARNING MODEL

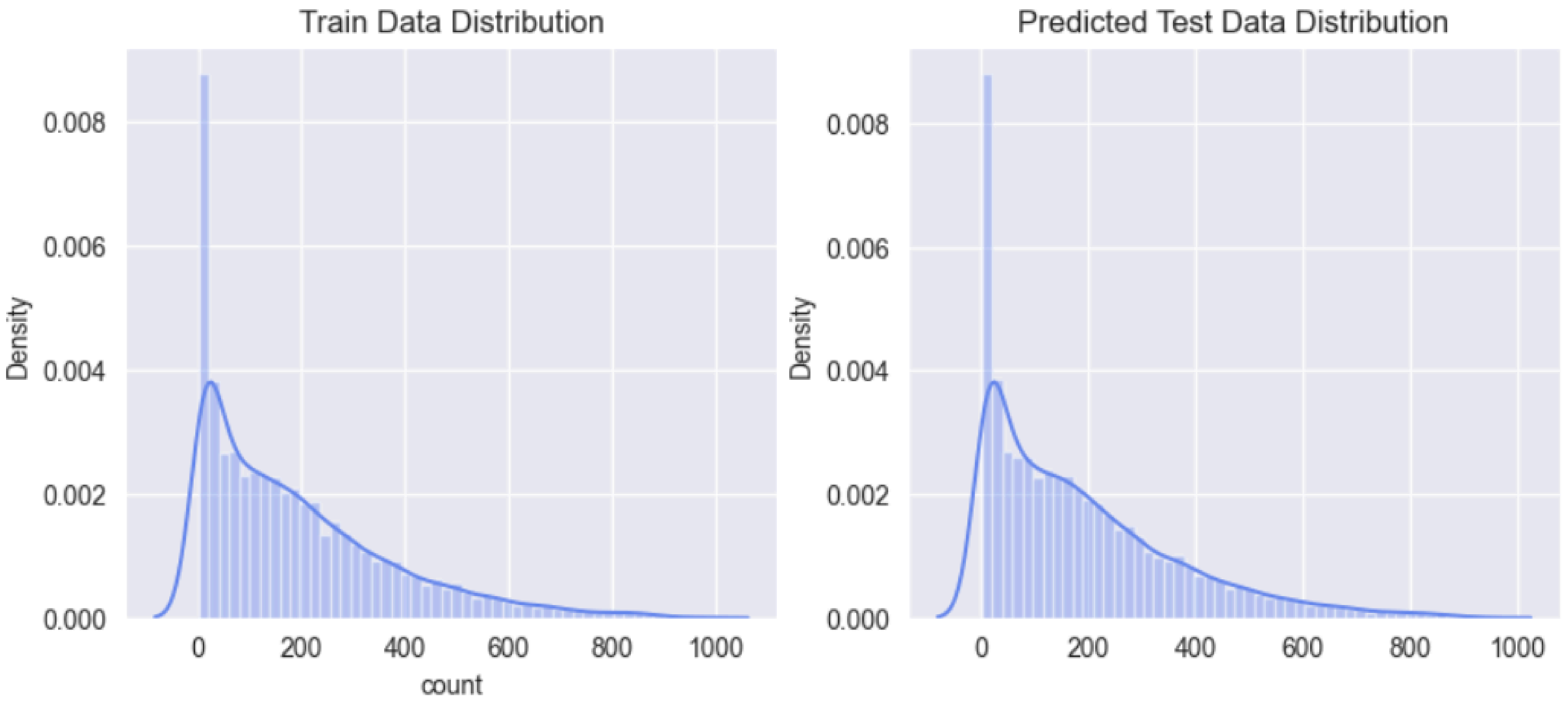
BANCKMARK MODEL

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

These models are evaluated using the Root Mean Squared Logarithmic Error (RMSLE) metric to determine the best-performing model.

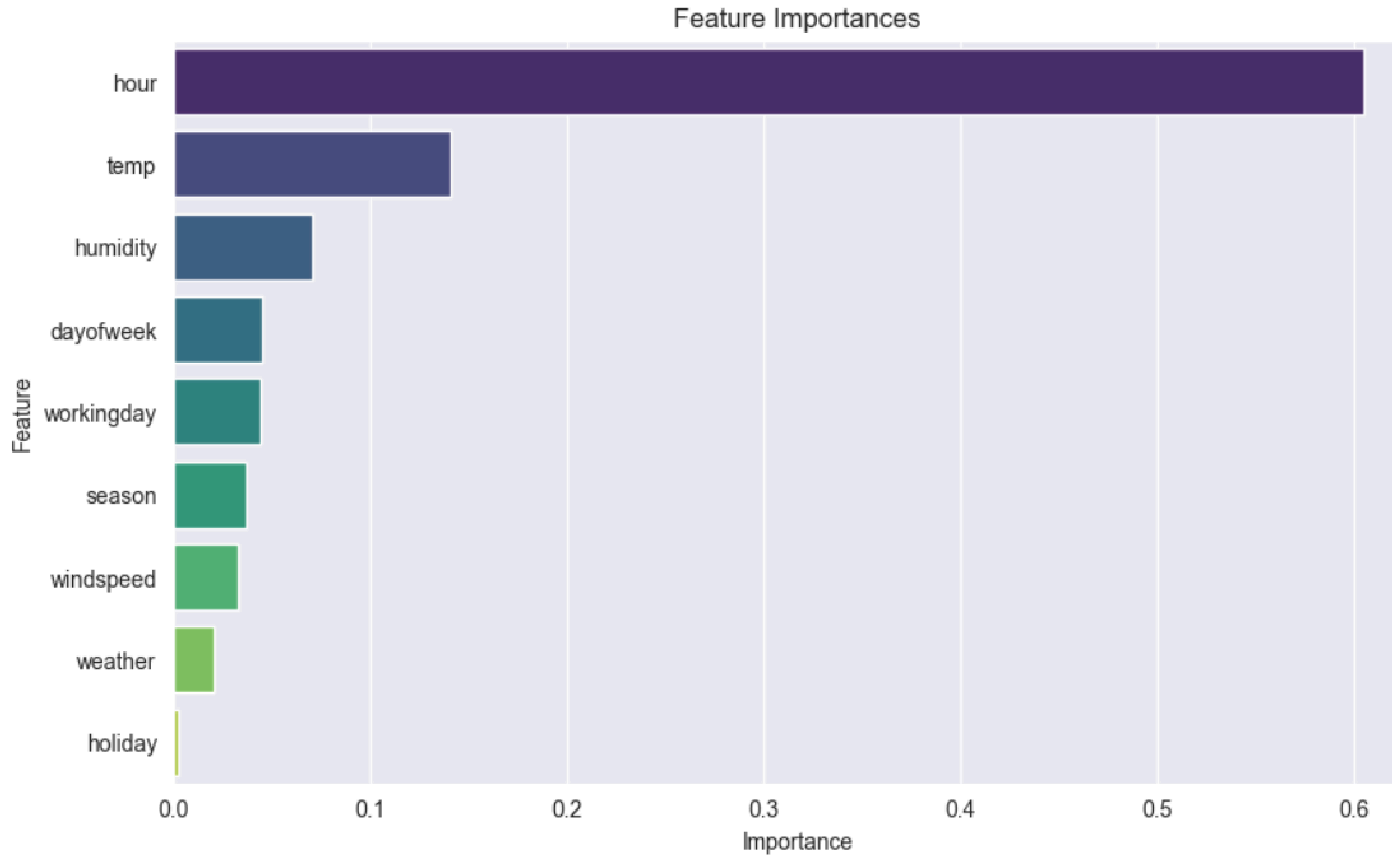
The Random Forest model proved to be the best. Before tuning, it showed good and consistent performance, with both RMSLE and cross-validation scores being low and close to each other. This indicates that the model generalizes well to unseen data. Although it showed some overfitting after tuning, this was less severe compared to the other models. The KNN and Gradient Boosting models overfitted significantly after tuning, while the Decision Tree model's performance deteriorated. Hence, **the Random Forest model 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) in the DataFrame submission 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, workingday, and others.**

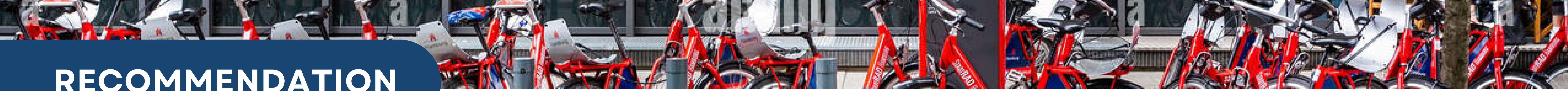
Adding the prediction results to the submission aims **to present an estimate of the expected demand quantity for each specific time period, this is for aiding in inventory planning, resource allocation, and other decision-making processes in business management.**

	datetime	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	count
0	2011-01-20 00:00:00	1	0	1	1	10.66	11.365	56	26.0027	14
1	2011-01-20 01:00:00	1	0	1	1	10.66	13.635	56	0.0000	5
2	2011-01-20 02:00:00	1	0	1	1	10.66	13.635	56	0.0000	4
3	2011-01-20 03:00:00	1	0	1	1	10.66	12.880	56	11.0014	3
4	2011-01-20 04:00:00	1	0	1	1	10.66	12.880	56	11.0014	3
5	2011-01-20 05:00:00	1	0	1	1	9.84	11.365	60	15.0013	16
6	2011-01-20 06:00:00	1	0	1	1	9.02	10.605	60	15.0013	57
7	2011-01-20 07:00:00	1	0	1	1	9.02	10.605	55	15.0013	153
8	2011-01-20 08:00:00	1	0	1	1	9.02	10.605	55	19.0012	295
9	2011-01-20 09:00:00	1	0	1	2	9.84	11.365	52	15.0013	176
10	2011-01-20 10:00:00	1	0	1	1	10.66	11.365	48	19.9995	113
11	2011-01-20 11:00:00	1	0	1	2	11.48	13.635	45	11.0014	105
12	2011-01-20 12:00:00	1	0	1	2	12.30	16.665	42	0.0000	123
13	2011-01-20 13:00:00	1	0	1	2	11.48	14.395	45	7.0015	120
14	2011-01-20 14:00:00	1	0	1	2	12.30	15.150	45	8.9981	133
15	2011-01-20 15:00:00	1	0	1	2	13.12	15.910	45	12.9980	119
16	2011-01-20 16:00:00	1	0	1	2	12.30	15.150	49	8.9981	115
17	2011-01-20 17:00:00	1	0	1	2	12.30	15.910	49	7.0015	225
18	2011-01-20 18:00:00	1	0	1	2	10.66	12.880	56	12.9980	209
19	2011-01-20 19:00:00	1	0	1	1	10.66	11.365	56	22.0028	160
20	2011-01-20 20:00:00	1	0	1	2	10.66	12.120	60	19.0012	107
21	2011-01-20 21:00:00	1	0	1	2	9.84	11.365	60	16.9979	65
22	2011-01-20 22:00:00	1	0	1	2	9.84	10.605	65	19.0012	57
23	2011-01-20 23:00:00	1	0	1	2	9.84	10.605	65	22.0028	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

Operational Planning

To effectively manage fluctuating demand, prioritize increasing bike availability during peak hours and warmer seasons. By employing dynamic fleet adjustments based on real-time weather forecasts, resources can be optimally allocated, ensuring consistent service reliability.

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. Offering targeted discounts during off-peak hours, weekends, holidays, and adverse weather conditions can incentivize usage during traditionally quieter periods, maximizing revenue potential while maintaining user satisfaction.

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!