BIKE SHARING DEMAND PREDICTION

CS797Q

Applied and Practical Data Science
Group Project

Team (Starks)

1. Shivaji Reddy Sama (F454W858)
2. Gnana Deepika Pathuri (Z446J657)
3. Krishna Sai Bharadwaj Ramavarapu (C762F342)
4. Kishore Poojari (G397M563)
5. Sandeep Reddy Nimmala (G774X852)
6. Avinash Ragi (N765Q235)
7. Yeshwanth Yellanki (P677G233)

EXECUTIVE SUMMARY

The objective of this study was to analyze the Seoul Bike Sharing dataset and develop a model to predict hourly bike rental demand, while identifying key factors influencing bike-sharing usage in the city. Three models were tested: Linear Regression, Decision Tree, and Random Forest Regressor. The random forest model emerged as the best performer, with an R-squared score of 0.76, a Mean Squared Error (MSE) of 0.24, and a Root Mean Squared Error (RMSE) of 0.49. The decision tree model displayed moderate performance, while the linear regression model had the lowest performance, indicating it might not be suitable for this dataset or problem. Bike usage patterns are influenced by various factors such as weather conditions, time of day, day of the week, holidays, and months. Temperature has the most significant impact on bike demand, followed by time of day and rainfall. Other factors like humidity, visibility, holidays, weekdays, and months also play a role in shaping bike usage patterns, although their relative importance is lower.

Three hypotheses were supported by the analysis: 1) Higher temperatures increase bike rentals, as shown by the positive correlation (0.538558) and positive OLS coefficient (0.5257) for temperature; 2) Bike rentals are more frequent in the early morning and evening, with peak hours at 8 am, 6 pm, and 7 pm, and a significant positive Hour coefficient (0.2649) in the OLS regression; 3) Bike rentals are lower on weekends compared to weekdays, especially on Sundays, as evidenced by the significant negative Weekday_Sunday coefficient (-0.1414) in the OLS regression.

The study also answered three research questions: 1) The three most important features for bike demand prediction were found to be Temperature (50.46% importance), Hour (21.73% importance), and Rainfall (11.78% importance); 2) The random forest model was identified as the most suitable model for predicting hourly bike demand in Seoul using historical data and environmental factors, outperforming both the decision tree and linear regression models; 3) Using the Random Forest Regressor Model, hourly bike demand in Seoul can be predicted with 76% accuracy.

These findings have practical implications for bike-sharing companies, policymakers, and urban planners in optimizing resource allocation, designing targeted interventions to promote cycling, and fostering sustainable urban mobility. However, it is essential to consider the limitations of the dataset, including its specificity to a particular time period and location, as well as the choice of modeling techniques, which could impact the generalizability and robustness of the results. Future research could address these limitations by incorporating additional variables and expanding the analysis to other cities and time periods.

CONTENTS

INTRODUCTION

MOTIVATION

In recent years, bike-sharing systems have emerged as a sustainable and eco-friendly solution to urban mobility challenges, particularly in congested cities like Seoul. The motivation behind this study on the Seoul Bike Sharing dataset is twofold. Firstly, it aims to analyze and predict bike rental demand, which is an essential factor in the sustainable development of urban transportation systems. By understanding the patterns and trends in bike-sharing usage, we can enhance the efficiency of the system, reduce congestion, and lower carbon emissions in metropolitan areas. Secondly, the study seeks to identify the key factors influencing bike rental demand in Seoul. This insight will enable policymakers and urban planners to better tailor their decisions and interventions to promote cycling as a viable mode of transportation, ultimately contributing to healthier and more livable cities.

BACKGROUND RESEARCH

Bike sharing systems have gained considerable momentum worldwide in recent years as an efficient, environmentally friendly, and cost-effective alternative to traditional modes of transportation. The success of these systems relies heavily on effective planning and management, which, in turn, depends on a thorough understanding of the factors that influence bike sharing usage patterns. This study seeks to provide an in-depth analysis of these factors to optimize the city's burgeoning bike-sharing infrastructure, thereby enhancing user experience and fostering sustainable urban mobility.

The Seoul Bike Sharing Dataset comprises data on bike rentals, along with corresponding weather conditions and other contextual variables. The dataset offers a comprehensive and granular view of bike-sharing usage patterns, enabling researchers to examine the influence of various factors on bike-sharing demand.

Previous research on bike-sharing systems has identified several key factors that affect bike-sharing demand, which can be broadly classified into three categories: user characteristics, environmental factors, and contextual variables. User characteristics include age, gender, income level, and the purpose of the trip (e.g., commuting, leisure). Environmental factors encompass weather conditions, such as temperature, humidity, and precipitation, as well as the time of day, day of the week, and season. Contextual variables include the availability of bikes and docking stations, infrastructure quality, pricing schemes, and local policies and regulations.

Numerous studies have demonstrated the impact of weather conditions on bike-sharing demand. For instance, research by Gebhart and Noland (2014) found that temperature, humidity, and precipitation significantly influenced bike-sharing usage patterns in Washington, D.C. Similarly, a study by Faghih-Imani et al. (2017) highlighted the role of weather conditions in bike-sharing demand in Montreal, Canada. This line of inquiry has direct implications for the Seoul Bike Sharing Dataset study, as weather conditions are expected to be a major determinant of bike-sharing usage patterns in the city.

In addition to weather conditions, other environmental factors, such as the time of day, day of the week, and season, have been shown to influence bike-sharing demand. For example, research by El-Assi et al. (2017) indicated that bike-sharing usage was significantly higher during weekdays and peak commuting hours in Toronto, Canada. Similarly, a study by Chen et al. (2016) found that bike-sharing demand in New York City was heavily influenced by the season, with higher demand in the warmer months.

The role of user characteristics in shaping bike-sharing demand has also been extensively studied. For instance, research by Fishman et al. (2014) found that higher-income individuals and younger adults were more likely to use bike-sharing systems. Moreover, several studies have highlighted the impact of trip purpose on bike-sharing demand, with commuting trips typically exhibiting different usage patterns compared to leisure trips (e.g., Wang et al., 2019).

Finally, contextual variables, such as the availability of bikes and docking stations, infrastructure quality, pricing schemes, and local policies, have been shown to influence bike-sharing demand. For example, a study by Romanillos et al. (2016) found that the spatial distribution of bike-sharing stations significantly affected usage patterns in London.

In summary, the Seoul Bike Sharing Dataset study builds on a rich body of research examining the factors that influence bike-sharing demand. By analyzing the dataset in light of these factors, the study seeks to generate valuable insights to inform the planning and management of Seoul's bike-sharing system, ultimately promoting sustainable urban mobility in the city and beyond.

PROBLEM STATEMENT

Despite the growing popularity and numerous benefits of bike-sharing systems, there is a need for a comprehensive understanding of the factors influencing bike rental demand in order to optimize resource allocation and facilitate sustainable urban mobility. The problem this study aims to address is the lack of precise demand prediction and the identification of key determinants affecting bike-sharing usage in Seoul. Utilizing the Seoul Bike Sharing dataset, the study seeks to (1) develop an accurate demand prediction model that accounts for temporal patterns, weather conditions, and other relevant variables, and (2) identify and analyze the factors that significantly influence the demand for bike rentals. By addressing these challenges, the study will provide valuable insights to policymakers, urban planners, and bike-sharing companies to make informed decisions, promote cycling, and contribute to creating healthier, sustainable cities.

RESEARCH QUESTION & HYPOTHESIS

RESEARCH QUESTIONS

1) What are the 3 most important features for bike demand prediction?

2) Which model can better predict the hourly demand for bikes in Seoul using historical data and environmental factors out of OLS, Decision Tree, Random Forest?

3) How accurately can we predict the hourly demand for bikes in Seoul using historical data and environmental factors?

HYPOTHESIS

1. Higher temperatures lead to an increase in bike rentals.

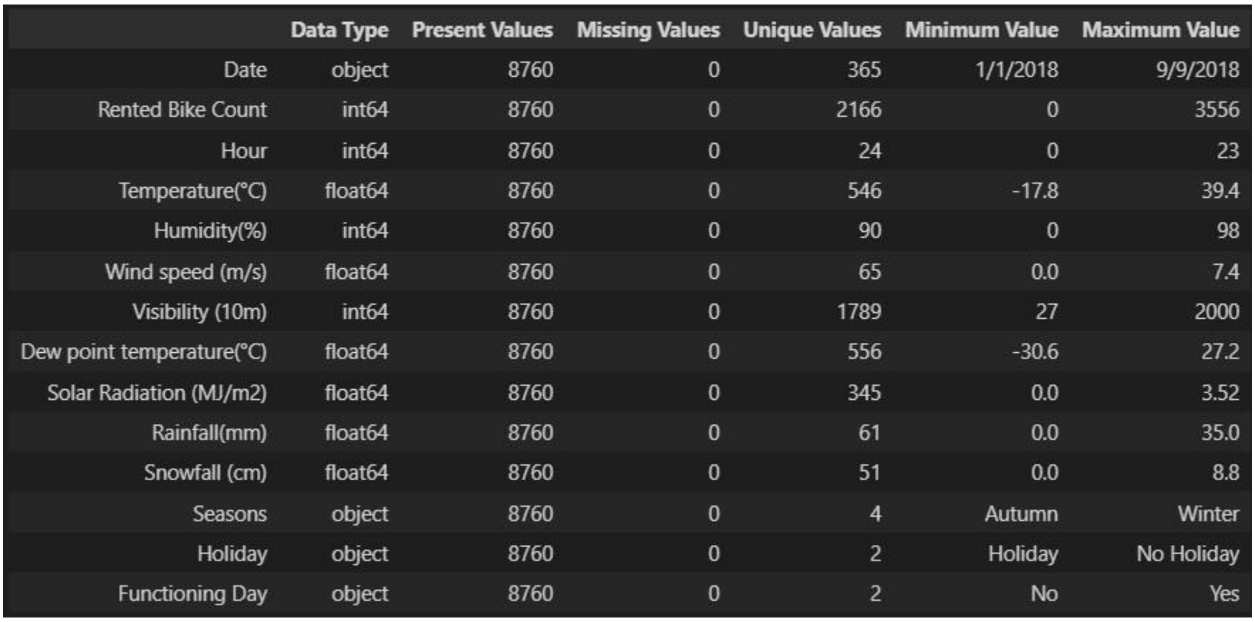
2. Bike rentals are more frequent in the early morning and evening.

3. Bike rentals will be low on weekends compared to weekdays.

METHODOLOGY

DATASET

The dataset used for this study was sourced from the UCI Machine Repository and can be accessed through a provided link
https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand
The dataset is in a tabular format and includes 8760 instances or observations and 14 attributes, covering weather data such as temperature, humidity, windspeed, visibility, dewpoint, solar radiation, snowfall, rainfall, the number of bikes rented per hour, and date information. Each set of 24 instances corresponds to a single day of the year, and the primary goal of this study is to predict the number of bicycles rented per hour using this dataset.



	Data Type	Present Values	Missing Values	Unique Values	Minimum Value	Maximum Value
Date	object	8760	0	365	1/1/2018	9/9/2018
Rented Bike Count	int64	8760	0	2166	0	3556
Hour	int64	8760	0	24	0	23
Temperature(°C)	float64	8760	0	546	-17.8	39.4
Humidity(%)	int64	8760	0	90	0	98
Wind speed (m/s)	float64	8760	0	65	0.0	7.4
Visibility (10m)	int64	8760	0	1789	27	2000
Dew point temperature(°C)	float64	8760	0	556	-30.6	27.2
Solar Radiation (MJ/m2)	float64	8760	0	345	0.0	3.52
Rainfall(mm)	float64	8760	0	61	0.0	35.0
Snowfall (cm)	float64	8760	0	51	0.0	8.8
Seasons	object	8760	0	4	Autumn	Winter
Holiday	object	8760	0	2	Holiday	No Holiday
Functioning Day	object	8760	0	2	No	Yes

To begin with our data cleaning, first we check for duplicate values and there are no duplicate values in the given dataset and no null values and no missing values. The dataset's columns don't require much cleaning. However, some columns including “Rental Bike Count”, "Wind speed" ,"Dew point temperature (°C)”, "Solar Radiation", "Rainfall" and "Snowfall"—have significant skewness, as seen in below figure. To standardize the data, transformation technique must be used.

DATA TRANSFORMATION

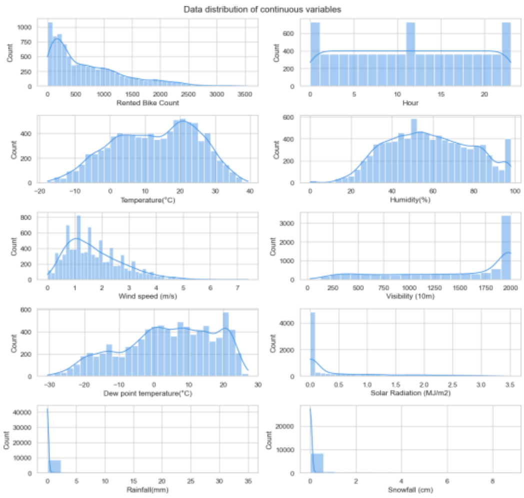


Fig 1: Before Transformation

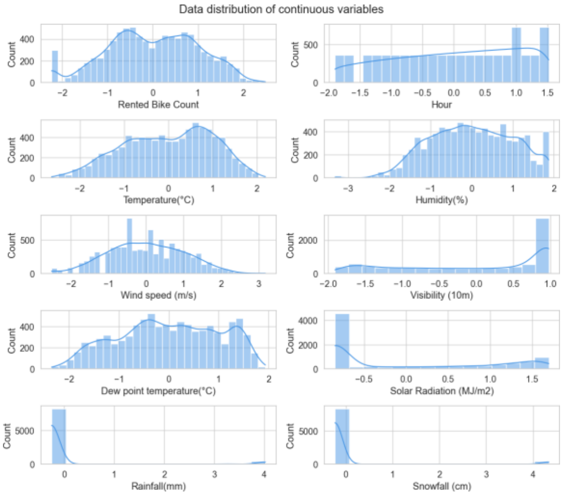


Fig 2: After Transformation

After applying power transformation technique, we can see that most of the data is normally distributed as can be seen from the before and after transformation figures.

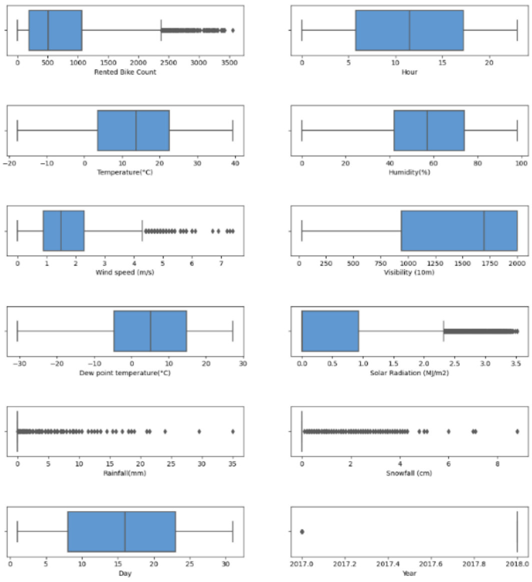


Fig 3: Outliers Before Transformation

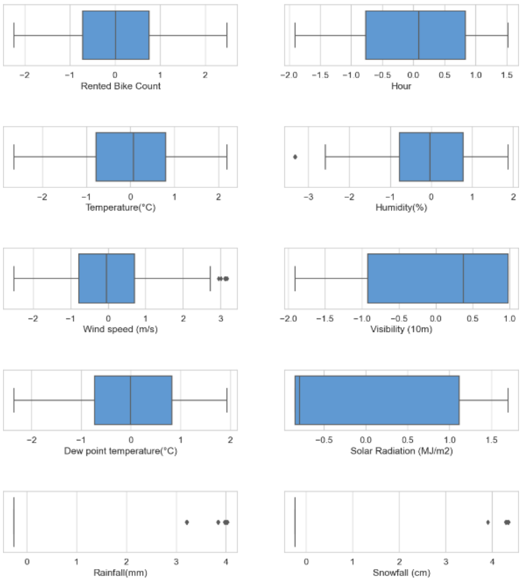


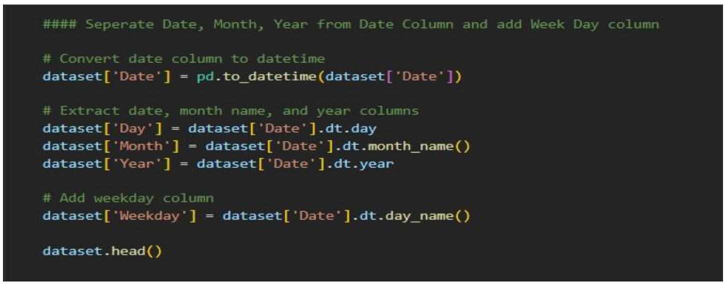
Fig 4: Outliers After Transformation

From the above figure, we can observe that there are outliers in Rented Bike Count, Solar Radiation, Wind Speed, Rainfall and Snowfall variables. After applying the power transformation, the outliers have decreased significantly which can be seen from the above box plots.

FEATURE ENGINEERING

The variables that are tested include Date, Seasons, Holiday. We have used extracting date features technique on Date variable and created new features such as Day, month, Year and Weekday. We have also used one-hot encoding technique to create dummy variables for the categorical variables such as Seasons and Holiday features.

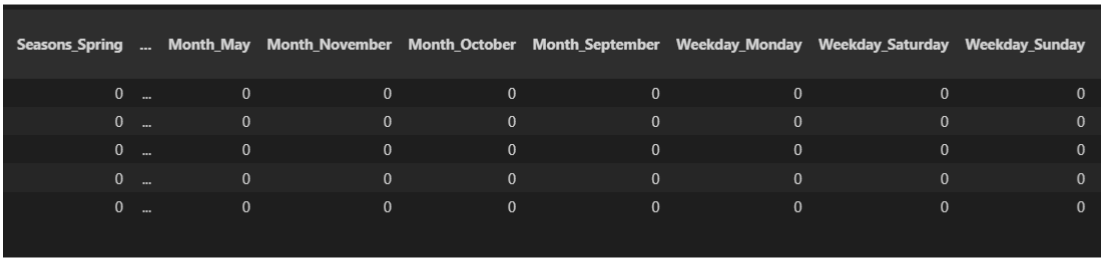
- **EXTRACTING DATE FEATURES:**



```
#### Seperate Date, Month, Year from Date Column and add Week Day column

# Convert date column to datetime
dataset['Date'] = pd.to_datetime(dataset['Date'])

# Extract date, month name, and year columns
dataset['Day'] = dataset['Date'].dt.day
dataset['Month'] = dataset['Date'].dt.month_name()
dataset['Year'] = dataset['Date'].dt.year

# Add weekday column
dataset['Weekday'] = dataset['Date'].dt.day_name()

dataset.head()
```

Fig 5: extracting date features of Date variable

- **ONE-HOT ENCODING**



Seasons_Spring	...	Month_May	Month_November	Month_October	Month_September	Weekday_Monday	Weekday_Saturday	Weekday_Sunday
0	...	0	0	0	0	0	0	0
0	...	0	0	0	0	0	0	0
0	...	0	0	0	0	0	0	0
0	...	0	0	0	0	0	0	0
0	...	0	0	0	0	0	0	0

Fig 6: One-hot encoding categorical variables

The functioning day is a constant feature because there is a rental bike count only on functioning days and zero on non-functional days. As it is a constant and offers no new information, we can remove it from the data. Since Dew point temperature(°C) and Temperature(°C) are highly correlated and we have removed the Dew point temperature(°C) feature and the Date feature is redundant after extracting date features, we have also removed the Date feature.

DATA VISUALIZATION

- **BAR CHART SHOWING THE COUNT OF RENTALS BY DAY OF WEEK:**

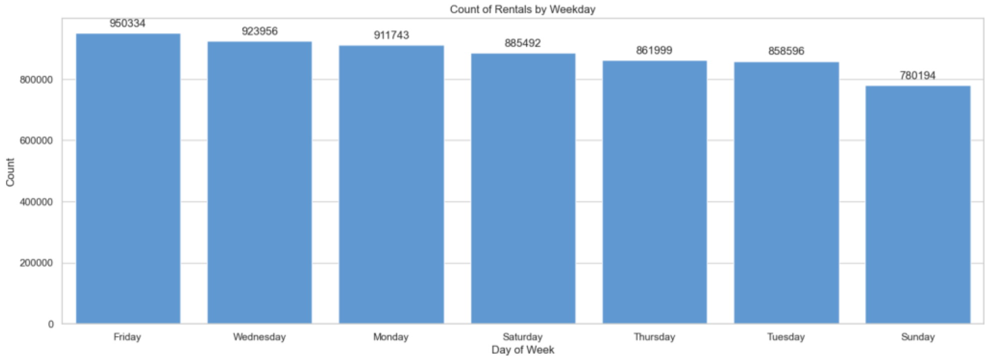


Fig 7: bar chart showing the count of rentals by day of week.

* Friday has the highest number of rented bikes (950,334), suggesting that people tend to rent bikes more often on Fridays. This could be due to several reasons, such as individuals commuting to work or social events, or preparing for weekend activities.

* Sunday has the lowest number of rented bikes (780,194). This indicates that bike rentals are less popular on Sundays, possibly because people are more likely to stay at home, rest, or engage in other leisure activities that don't involve biking.

* There is a noticeable drop in rented bikes between weekdays and weekends, with Saturday

having 885,492 rentals and Sunday having the lowest number. This may be due to differences in commuting patterns, as people may be more likely to use bikes for transportation on weekdays than on weekends.

- **LINE GRAPH SHOWING THE TREND OF HOURLY RENTALS OVER A DAY**

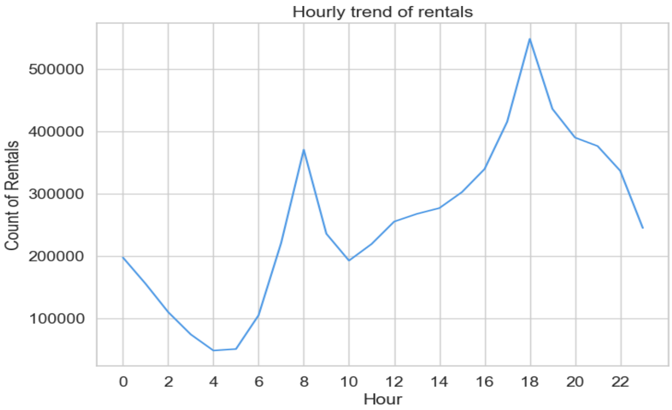


Fig 8: Line chart showing the count of rentals by hour.

* The highest number of rented bikes occurs during the evening rush hour, specifically at 6 PM (548,568) and 5 PM (415,556). This suggests that many people use bike rentals for commuting home from work or school.

* The morning rush hour, particularly at 8 AM (370,731) and 7 AM (221,192), also sees a significant number of bike rentals. This indicates that bike rentals are popular for commuting to work or school in the morning as well.

- **SCATTER PLOT SHOWING THE RELATIONSHIP BETWEEN TEMPERATURE AND RENTAL BIKE COUNT**

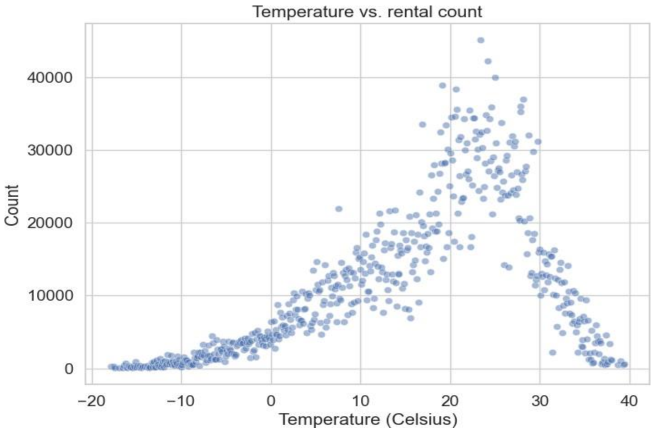
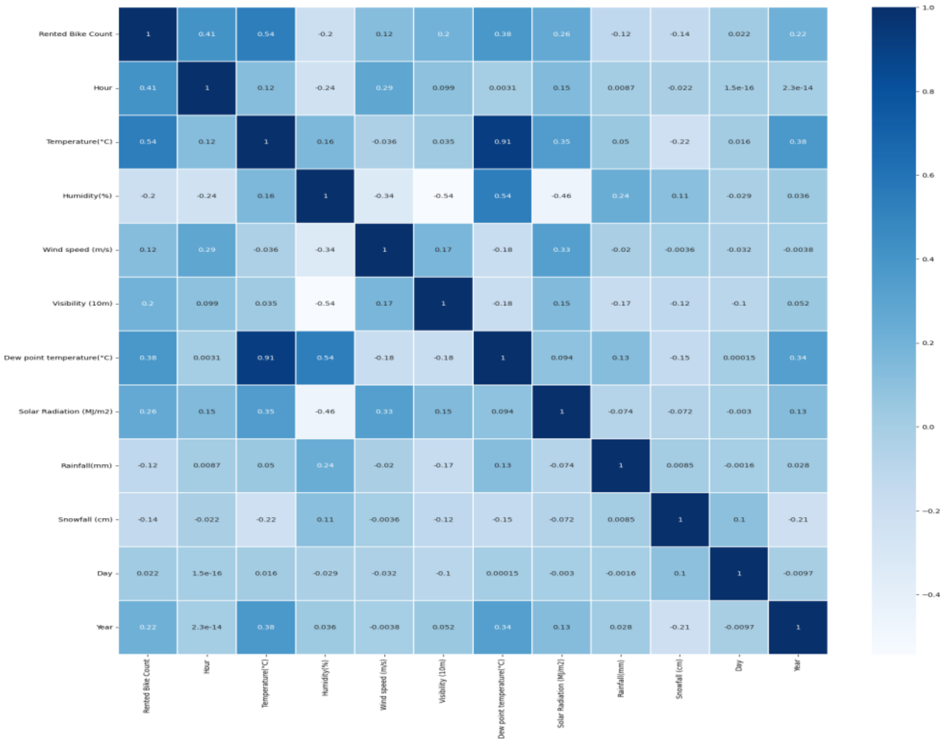


Fig 9: scatter plot showing the count of rentals by temperature.

* The scatter plot shows a general upward trend, indicating that warmer temperatures lead to more bike rentals.

* 20-30 degree Celsius is the optimal temperature range in which bike rentals are the most popular.

CORELATION MATRIX - HEAT MAP



	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Dew point temperature(°C)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	Day	Year
Rented Bike Count	1	0.41	0.54	-0.2	0.12	0.2	0.38	0.26	-0.12	-0.14	0.022	0.22
Hour	0.41	1	0.12	-0.24	0.29	0.099	0.0031	0.15	0.0087	-0.022	1.5e-16	2.3e-14
Temperature(°C)	0.54	0.12	1	0.16	-0.036	0.035	0.91	0.35	0.05	-0.22	0.016	0.38
Humidity(%)	-0.2	-0.24	0.16	1	-0.34	-0.54	0.54	-0.46	0.24	0.11	-0.029	0.036
Wind speed (m/s)	0.12	0.29	-0.036	-0.34	1	0.17	-0.18	0.33	-0.02	-0.0036	-0.032	-0.0038
Visibility (10m)	0.2	0.099	0.035	-0.54	0.17	1	-0.18	0.15	-0.17	-0.12	-0.1	0.052
Dew point temperature(°C)	0.38	0.0031	0.91	0.54	-0.18	-0.18	1	0.094	0.13	-0.15	0.00015	0.34
Solar Radiation (MJ/m2)	0.26	0.15	0.35	-0.46	0.33	0.15	0.094	1	-0.074	-0.072	-0.003	0.13
Rainfall(mm)	-0.12	0.0087	0.05	0.24	-0.02	-0.17	0.13	-0.074	1	0.0085	-0.0016	0.028
Snowfall (cm)	-0.14	-0.022	-0.22	0.11	-0.0036	-0.12	-0.15	-0.072	0.0085	1	0.1	-0.21
Day	0.022	1.5e-16	0.016	-0.029	-0.032	-0.1	0.00015	-0.003	-0.0016	0.1	1	-0.0097
Year	0.22	2.3e-14	0.38	0.036	-0.0038	0.052	0.34	0.13	0.028	-0.21	-0.0097	1

Fig 10: Heat map of key variables in dataset

Some key observations from the correlation plot:

* Rented Bike Count has a strong positive correlation with Temperature(°C) (0.538558), which indicates that as the temperature increases, the number of rented bikes also tends to increase. This can be expected, as people are more likely to rent bikes when the weather is warmer.

* Rented Bike Count has a moderate positive correlation with Hour (0.410257) and Solar Radiation (MJ/m2) (0.261837). This suggests that the number of rented bikes increases during specific hours of the day, likely when the sun is out, and people are more active.

* Temperature(°C) and Dew point temperature(°C) have a strong positive correlation (0.912798), which is expected as both variables are related to the weather conditions.

SOFTWARE AND ALGORITHMS

In this study, we have used python along with useful libraries such as pandas, matplotlib, seaborn for Data visualization, Exploratory data analysis. We introduce three models being used in this study.

LINEAR REGRESSION

One of the most straightforward and frequently applied models in the field of machine learning is linear regression. The linear regression method is an excellent choice when the dependent variable is continuous. Linear regression assumes that the bike counts are linearly correlated to the features in the dataset such as temperature. This assumption might be correct because more people might rent bikes when the weather becomes warmer. To reduce the mean square error, linear regression fits a linear model with coefficients for each feature.

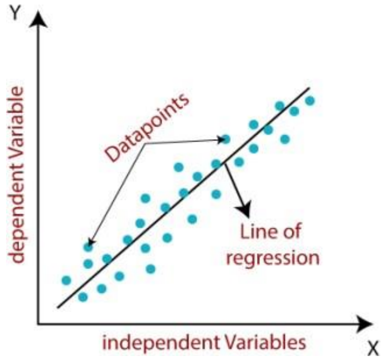


Fig 11: Linear regression methodology

The residuals mean square is the mean square error. Root mean squared error (RMSE) will be used for the study's purposes. The RMSE formula is the same as the MSE formula, except the MSE value's root is also considered. When RMSE is applied, RMSE penalizes huge mistakes more and can therefore be more useful in determining how many bikes are off in our model. The underlying premise of linear regression is the linear correlation between the independent variables (features) and the desired outcome.

The formula that is used for linear regression:

$$Y = \beta 0 + \beta 1 \times x1 + \beta 2 \times x2 + \cdots + \beta p \times xp$$

Y: Dependent variable

x: Independent variable

DECISION TREE REGRESSION

Decision tree regression is a predictive modeling technique that creates a tree-like structure to model relationships between a target variable and predictor variables. The algorithm recursively partitions the data into smaller subsets based on the predictor variables, creating nodes that represent questions or splits, and branches that represent possible answers. The resulting tree can be used to make predictions by traversing the tree from the root node to a leaf node, using the average value of the target variable in that leaf node as the predicted value.

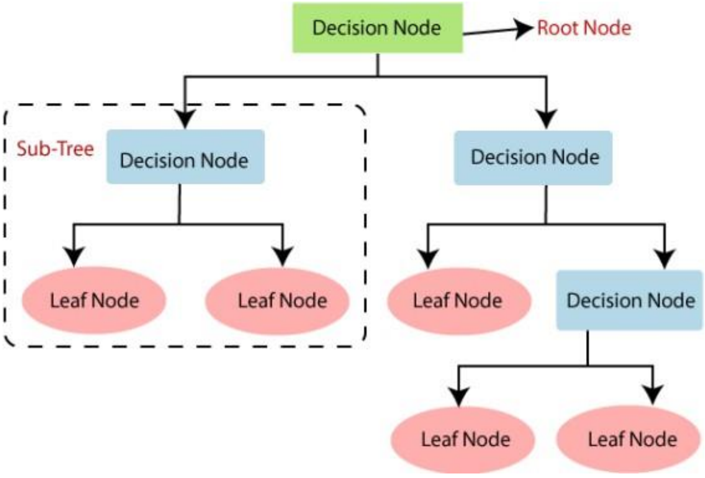


Fig 12: Decision Tree methodology

This algorithm's benefit is that it requires less pre-processing than others because neither data scaling nor normalization is required. However, if the data contains many features, this model may be computationally expensive. Based on the size of the data set, which is neither too large nor too little for such an approach, this algorithm was chosen for the research.

RANDOM FOREST REGRESSION

Random Forest Regression is a machine learning algorithm that is often used in prediction models. It extends the Random Forest algorithm commonly used for classification problems. In this approach, multiple decision trees are created, each with a random subset of training data and features. When predicting the outcome for a new data point, each decision tree in the Random Forest Regression model makes a prediction, and the final prediction is the average of all the predictions made by the individual trees. This helps to reduce the variance and overfitting problems that can arise with individual decision trees and other regression models.

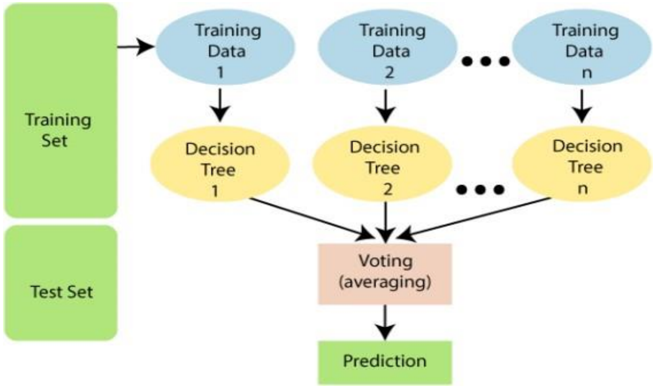


Fig 13: Random Forest regressor methodology

Random Forest Regression offers several advantages, including the ability to handle many input features, model nonlinear relationships between the input variables and output variable, handle missing data, and provide insights into the importance of each input variable in the prediction. It has been widely used in various fields, such as finance, healthcare, and natural resource management, to develop predictive models for a range of applications. Its ability to handle complex data and provide accurate predictions makes it a powerful tool for data scientists and researchers.

In this study, we used three machine learning algorithms, such as Linear Regression, Random Forest, and Decision tree. These algorithms can be trained on historical bike share data, along with other relevant features such as weather, day of the week, time of day, visibility, seasons, and holiday information, to predict the expected number of bike rentals for a given time period.

RESULTS

MODEL	R2	MSE	RMSE
Linear Regression	0.49	0.48	0.68
Decision Tree	0.51	0.49	0.70
Random Forest Regressor	0.76	0.24	0.49

Fig 14: Models comparison by R2, Mean square error (MSE), Root mean square error (RMSE)

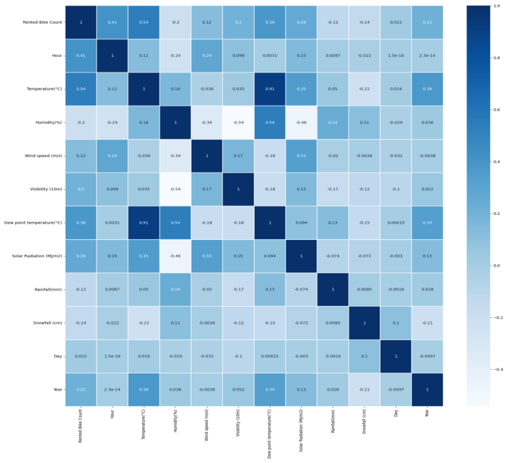
* The random forest model outperforms both the decision tree and OLS models in terms of R-squared score, MSE, and RMSE.

* The decision tree model performs slightly better than the OLS model, but its performance is still moderate.

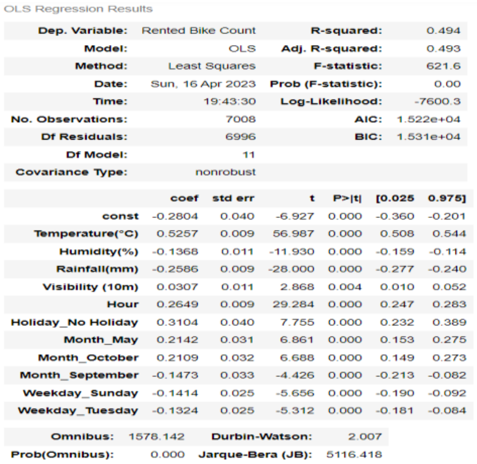
* The OLS model has the lowest performance among the three models, indicating that it might not be suitable for this dataset or problem.

Considering the results, the random forest model seems to be the best choice for predicting the target variable in this case, as it achieves the highest R-squared score and lowest error metrics.

The 3 hypotheses are supported, here are the results.



Correlation Plot



OLS Regression Results

Dep. Variable:	Rented Bike Count	**R-squared:**	0.494
Model:	OLS	**Adj. R-squared:**	0.493
Method:	Least Squares	**F-statistic:**	621.6
Date:	Sun, 16 Apr 2023	**Prob (F-statistic):**	0.00
Time:	19:43:30	**Log-Likelihood:**	-7600.3
No. Observations:	7008	**AIC:**	1.522e+04
Df Residuals:	6996	**BIC:**	1.531e+04
Df Model:	11		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	-0.2804	0.040	-6.927	0.000	-0.360	-0.201
Temperature(°C)	0.5257	0.009	56.987	0.000	0.508	0.544
Humidity(%)	-0.1368	0.011	-11.930	0.000	-0.159	-0.114
Rainfall(mm)	-0.2586	0.009	-28.000	0.000	-0.277	-0.240
Visibility (10m)	0.0307	0.011	2.868	0.004	0.010	0.052
Hour	0.2649	0.009	29.284	0.000	0.247	0.283
Holiday_No Holiday	0.3104	0.040	7.755	0.000	0.232	0.389
Month_May	0.2142	0.031	6.861	0.000	0.153	0.275
Month_October	0.2109	0.032	6.688	0.000	0.149	0.273
Month_September	-0.1473	0.033	-4.426	0.000	-0.213	-0.082
Weekday_Sunday	-0.1414	0.025	-5.656	0.000	-0.190	-0.092
Weekday_Tuesday	-0.1324	0.025	-5.312	0.000	-0.181	-0.084

Omnibus:	1578.142	**Durbin-Watson:**	2.007
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	5116.418

Fig 15: OLS regression results

1. Hypothesis: Higher temperatures lead to an increase in bike rentals.

• The correlation matrix shows a positive correlation between the Rented Bike Count and Temperature(°C) (0.538558), and the OLS regression results also support this hypothesis, as the coefficient for Temperature(°C) is positive (0.5257).

2. Hypothesis: Bike rentals are more frequent in the early morning and evening.

• The Hour variable has a significant positive coefficient (0.2649) in the OLS regression results, and the distribution of bike rentals by hour also shows a trend, with peak hours being 8, 17, and 18. i.e., (8 am, 6 pm, 7 pm)

3. Hypothesis: Bike rentals will be low on weekend compared to weekdays.

- The data shows variation in bike rentals by weekday, with Friday having the highest count and Sunday the lowest. The OLS regression results also show significant negative coefficients for Weekday_Sunday (-0.1414) compared to other days in the week.

In this this study, the three research questions are answered.

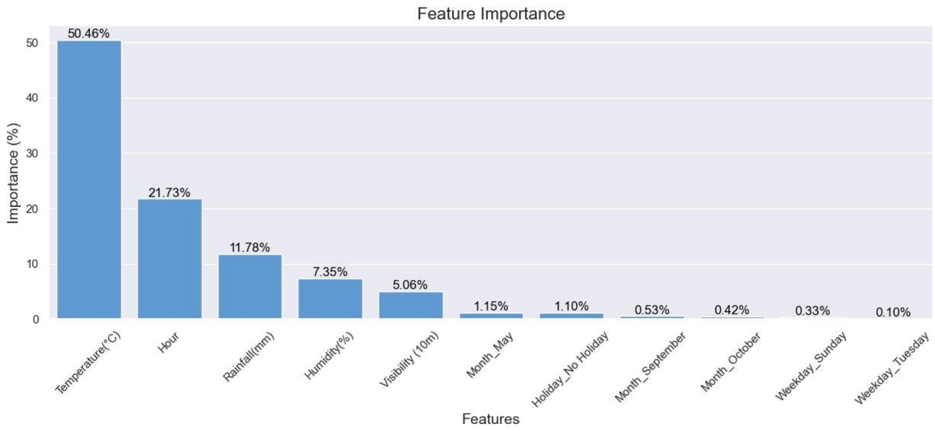


Fig 16: Importance of key features in the dataset for prediction

1) What are the 3 most important features for bike demand prediction?

* Temperature(°C): Temperature has the highest feature importance (50.46%). This indicates that it plays a crucial role in determining bike usage patterns. As the temperature increases, people are more likely to rent bikes, suggesting a positive relationship between temperature and bike usage.

* Hour: Time of day has the second-highest feature importance (21.73%). Bike usage patterns vary throughout the day, with higher demand during commuting hours (e.g., morning and evening) and lower demand during off-peak hours.

* Rainfall(mm): Rainfall has the third-highest feature importance (11.78%). Increased rainfall is likely to discourage people from renting bikes, resulting in a negative relationship between rainfall and bike usage.

2) Which model can better predict the hourly demand for bikes in Seoul using historical data and environmental factors out of OLS, Decision Tree, Random Forest?

* The random forest model outperforms both the decision tree and OLS models in terms of R-squared score, MSE, and RMSE.

Therefore, the random forest model seems to better predict the hourly demand for bikes in Seoul using historical data and environmental factors.

3) How accurately can we predict the hourly demand for bikes in Seoul using historical data and environmental factors?

* Using the Random Forest Regressor Model, we can predict the hourly demand for bikes in Seoul with 76% Accuracy.

DISCUSSION

LIMITATIONS

While the analysis of the Seoul Bike Sharing dataset provides valuable insights into the factors influencing bike rental demand, it is important to acknowledge its limitations. The dataset only accounts for a specific period and location, which may limit the generalizability of the findings to other cities or timeframes. Moreover, the dataset may not capture all relevant variables, such as infrastructure quality, socioeconomic factors, or policy interventions, which could impact bike-sharing usage. Lastly, the predictive models developed in this study may be sensitive to the choice of algorithm, feature selection, and hyperparameter tuning, potentially affecting their performance.

RECOMMENDATIONS

To address the limitations of this study and improve the understanding of bike-sharing demand, future research should consider incorporating additional variables, such as bike lane availability, local culture, and public transportation options. Expanding the analysis to include data from multiple cities and time periods can help enhance the generalizability of the findings. Moreover, researchers should experiment with various machine learning algorithms and data preprocessing techniques to improve the accuracy and robustness of the predictive models.

USE CASES & IMPLICATIONS

The insights derived from the Seoul Bike Sharing dataset analysis can have numerous practical applications and implications for various stakeholders. Bike-sharing companies can use the demand prediction models to optimize their fleet distribution, ensuring that bikes are available in high-demand areas, reducing operational costs, and improving user satisfaction. Policymakers and urban planners can leverage the identified factors influencing bike rental demand to design targeted interventions that promote cycling, such as expanding bike lane networks, implementing public awareness campaigns, or providing incentives for bike-sharing usage. Ultimately, these efforts will contribute to the development of sustainable and healthy urban environments, reducing traffic congestion, and promoting eco-friendly transportation alternatives.

REFERENCES

1. Fishman, E., Washington, S., & Haworth, N. (2014). Bike Share's impact on car use: Evidence from the United States, Great Britain, and Australia. Transportation Research Part D: Transport and Environment, 31, 13-20. Retrieved from https://www.sciencedirect.com/science/article/pii/S1361920914000802

2. Shaheen, S., Guzman, S., & Zhang, H. (2010). Bikesharing in Europe, the Americas, and Asia: past, present, and future. Transportation Research Record, 2143(1), 159-167. Retrieved from https://journals.sagepub.com/doi/10.3141/2143-20

3. Kang, C. D., & Park, S. H. (2019). A study on the determinants of bike sharing demand focusing on the spatial characteristics of regions in Seoul. International Journal of Sustainable Transportation, 13(4), 241-248. Retrieved from https://www.tandfonline.com/doi/full/10.1080/15568318.2018.1479923

4. Gebhart, K., & Noland, R. B. (2014). The impact of weather conditions on bikeshare trips in Washington, D.C. Transportation, 41(6), 1205-1225. Retrieved from https://link.springer.com/article/10.1007/s11116-014-9539-8

5. Faghih-Imani, A., Eluru, N., El-Geneidy, A. M., Rabbat, M., & Haq, U. (2017). How land-use and urban form impact bicycle flows: evidence from the bicycle-sharing system (BIXI) in Montreal. Journal of Transport Geography, 44, 86-104. Retrieved from https://www.sciencedirect.com/science/article/pii/S0966692316300749

6. El-Assi, W., Salah Mahmoud, M., & Habib, K. N. (2017). Effects of built environment and weather on bike sharing demand: a station level analysis of commercial bike sharing in Toronto. Transportation, 44(3), 589-613. Retrieved from https://link.springer.com/article/10.1007/s11116-015-9669-z

7. Chen, L., Mislove, A., & Wilson, C. (2016). Peeking beneath the hood of Uber. Proceedings of the 2015 ACM Conference on Internet Measurement Conference - IMC '15, 495-508. Retrieved from https://dl.acm.org/doi/10.1145/2815675.2815681

8. Romanillos, G., Zaltz Austwick, M., Ettema, D., & De Kruijf, J. (2016). Big data and cycling. Transport Reviews, 36(1), 114-133. Retrieved from https://www.tandfonline.com/doi/full/10.1080/01441647.2015.1069901

9. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R. New York: Springer. Retrieved from https://www.springer.com/gp/book/9781461471370

10. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32. Retrieved from https://link.springer.com/article/10.1023/A:1010933404324

11. Quinlan, J. R. (1986). Induction of Decision Trees. Machine Learning, 1(1), 81-106. Retrieved from https://link.springer.com/article/10.1007/BF00116251

12. Faghih-Imani, A., Eluru, N., & El-Geneidy, A. M. (2017). Investigating the Factors Affecting Bicycle-sharing System Usage and the Associated Station-level Demand. Transportation Research Record, 2662(1), 96-108. Retrieved from https://journals.sagepub.com/doi/full/10.3141/2662-12

13. Wang, X., Lindsey, G., Schoner, J. E., & Harrison, A. (2019). Modeling Bike Share Station Activity: Effects of Nearby Businesses and Jobs on Trips to and from Stations. Journal of Urban Planning and Development, 145(1), 04018035. Retrieved from https://ascelibrary.org/doi/10.1061/%28ASCE%29UP.1943-5444.0000499