# An Analysis of Seoul Bike Sharing Demand: The Data Science Perspective

By Ho Yin, Tam

**Abstract:** With the concern of environmental issues, the popularity of the sharing bike scheme rose significantly due to the reduced emission of air pollutants. Also, the program provides convenience to users for commuting to the workplace. In this paper, the analysis from the data science perspective commences with a broad approach depicting the trend of the number of shared bike rentals during the period of December 2017 to November 2018. Then, by the visual analytics approach, the analysis is narrowed down to the average rented bikes over hours in different seasons, holidays, functioning days, and weekdays. Furthermore, a multiple linear regression model is developed, utilizing the hour and the weather conditions as predictors, to predict the number of shared bike rentals. The performance of the model is evaluated by the mean absolute error, the mean squared error and the coefficient of determination.

## Introduction:

First launched in 1965 in Amsterdam, the Netherlands, the bike-sharing scheme has now been adopted worldwide. According to the Meddin Bike-Sharing World Map report published in 2021, the number of open bike-sharing systems dramatically rose from 5 systems to 1999 systems from the year 2000 to 2021. In 2021, China has the largest number of active bike-sharing systems with 673 systems while the United States has the second most open bike-sharing systems with 174 systems [1]. The scheme allows users to locate the available bikes nearby via websites or apps and then rent a bike in an unmanned rental station. The fees are charged through the websites or apps by online payment method. After the ride, users can return the bike to any station. There is no doubt that the bike-sharing scheme has become more common in various countries and provides an alternative convenient transportation mode for commuting and tourism.

The bike-sharing scheme not only contributes to the reduction of air pollutants but also provides convenience to commuters for travelling to their workplace. Individuals opt to choose cycling as their mode of transportation. Therefore, having an accurate prediction of the demand for shared bikes is of paramount importance and also plays a vital role in the success of the bike-sharing scheme. This paper will focus on the analysis of the bike-sharing program in Seoul, Korea from the data science perspective.

## Analytical questions:

The scope of this research mainly focuses on the three questions below.

1. What is the trend of shared bike rental from 1st December 2017 to 30th November 2018?
2. Are there variations in the demand for shared bikes across different seasons, holidays, functioning days, and weekdays?
3. What are the correlations between weather conditions (temperature, humidity, wind speed, visibility, solar radiation, snowfall, and rainfall) and time of the day and the demand for shared bikes (response)?

## Data (materials):

### A. Key Characteristics

The dataset extracted from the UCI Machine Learning Repository comprises 8760 observations and 14 columns [2]. The features consist of weather conditions, such as temperature, humidity, wind speed, visibility, dew point temperature, solar radiation, rainfall, and snowfall which are numerical data. Also, the features include the types of seasons, functioning days and holidays which are categorical data. The dataset records the number of shared bike rentals in Seoul from 1st December 2017 to 30th November 2018.

### B. Suitability to answer the research questions

The day, month, and year are clearly recorded. The line chart can be plotted to visualize the trend of the demand for shared bikes in each month from 2017 to 2018.

The type of seasons, holidays, and functioning day data can be used to identify the patterns that exist with the demand for shared bikes.

The hour, the weather conditions and the demand

for the shared bikes are predictors and responses respectively. As they are both numerical data, a multiple linear regression model can be developed.

### C. Key Assumptions

1. For the 'hour' column, a total of 24 unique values should appear, with '0' referring to the first hour and '23' representing the last hour. The column 'rented bike count' should not contain any negative numbers.
2. As Seoul, Korea is in the northern hemisphere, it is the fact that the summer starts from 1st June to 31st August and winter starts from 1st December to 28th February.
3. The observations are independent of each other, and the residuals are normally distributed for the multiple linear regression model [3].

## Analysis:

### A. Data Preparation and Data Derivation

Data understanding and cleaning are imperative as they have a significant impact on the upcoming process. In the dataset, no missing value is found. The 'Date' column is transformed to 'Year-Month', 'Year', 'Month', 'Day' and 'Weekday' in which the five columns are newly appended to the table. The number of shared bike rentals can now be aggregated by year and month, allowing for the creation of a line chart to visualize the overall trend from 1st December 2017 to 30th November 2018. Also, in the 'Weekday' column, 'Monday' to 'Friday' are encoded to 'weekdays' while 'Saturday' and 'Sunday' are encoded to 'weekends'. The purpose is to compare the number of shared bike rentals between weekdays and weekends.

Table 1 shows the details information of the weather conditions and the shared bike rentals with the mean, median, and standard deviation. It is reasonable for the negative value in temperature as the temperature in Seoul, Korea will drop to below 0 degrees Celsius in winter. Also, it is noticeable that more than 75% of rainfall (mm) and snowfall (cm) data are 0 mm and 0 cm respectively.

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| Rented Bike Count | 8760.0 | 704.602055 | 644.997468 | 0.0 | 191.00 | 504.50 | 1065.25 | 3556.00 |
| Hour | 8760.0 | 11.500000 | 6.922582 | 0.0 | 5.75 | 11.50 | 17.25 | 23.00 |
| Temperature(°C) | 8760.0 | 12.882922 | 11.944825 | -17.8 | 3.50 | 13.70 | 22.50 | 39.40 |
| Humidity(%) | 8760.0 | 58.226256 | 20.362413 | 0.0 | 42.00 | 57.00 | 74.00 | 98.00 |
| Wind speed (m/s) | 8760.0 | 1.724909 | 1.036300 | 0.0 | 0.90 | 1.50 | 2.30 | 7.40 |
| Visibility (10m) | 8760.0 | 1436.825799 | 608.298712 | 27.0 | 940.00 | 1698.00 | 2000.00 | 2000.00 |
| Dew point temperature(°C) | 8760.0 | 4.073813 | 13.060369 | -30.6 | -4.70 | 5.10 | 14.80 | 27.20 |
| Solar Radiation (MJ/m2) | 8760.0 | 0.569111 | 0.868746 | 0.0 | 0.00 | 0.01 | 0.93 | 3.52 |
| Rainfall(mm) | 8760.0 | 0.148687 | 1.128193 | 0.0 | 0.00 | 0.00 | 0.00 | 35.00 |
| Snowfall (cm) | 8760.0 | 0.075068 | 0.436746 | 0.0 | 0.00 | 0.00 | 0.00 | 8.80 |

Table 1

On the other hand, the histogram (Figure 1) visualises the distribution of the numerical data, including the number of shared bike rentals and the weather conditions. The temperature and humidity are close to having a normal distribution while the rainfall and snowfall have a heavily skewed right distribution.

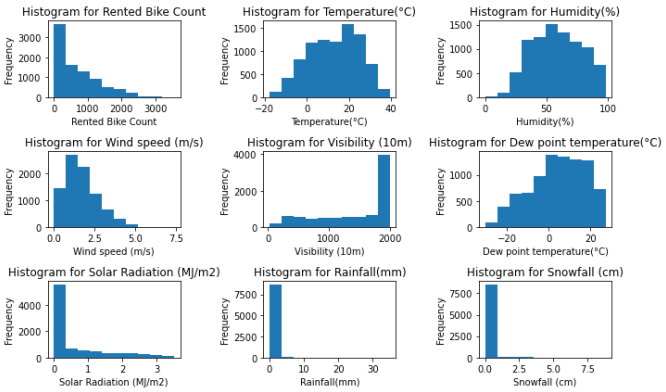


Figure 1

### B. Trend analysis

The section of analysis commences with the broad idea of the overall trend of shared bike rentals from 1st December 2017 to 30th November 2018 (Question 1). Then the analysis will narrow down to the comparison of the number of shared bike rentals in various seasons, holidays, functioning days and weekdays (Question 2) by the visual analytics approach.

In Figure 2, the number of shared bike rentals declined gradually from December 2017 to February 2018. The circumstances drastically changed as it rose sharply with a peak of approximately 900,000 bikes until in June 2018. Since then, the number of shared bike rentals decreased by over 40% and dropped to 500,000 bikes in November 2018, despite a minor growth observed from August 2018 to September 2018.

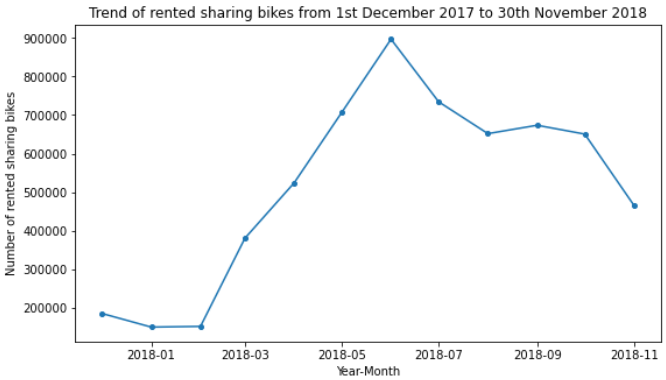


Figure 2

Figure 3 provides a summary of the cumulative count of rented bike count throughout the hours spanning from 1st December 2017 to 30th November 2018. The peak in total number of shared bike rentals occurs between 18:00 and 19:00. It is believed that the surge in demand is attributed to individuals departing from their workplaces and opting for cycling as a means of commuting to their residences.

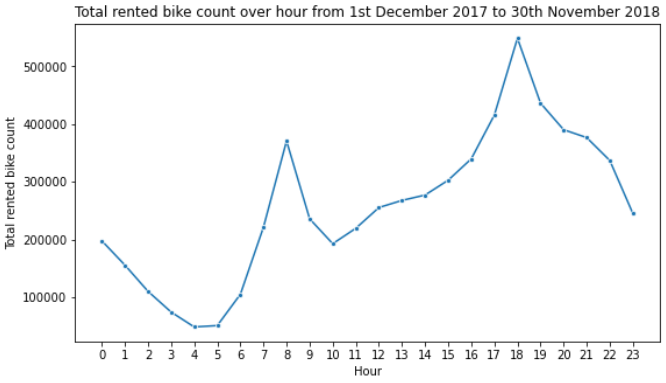


Figure 3

Figure 4 shows the average rented bike count in summer is the highest while the average rented bike count in winter is the lowest. In autumn, the average number of shared bike rentals is slightly higher than in spring. For all four seasons, the highest average number of rented bikes is between 18:00 to 19:00 while the lowest is between 4:00 to 5:00.

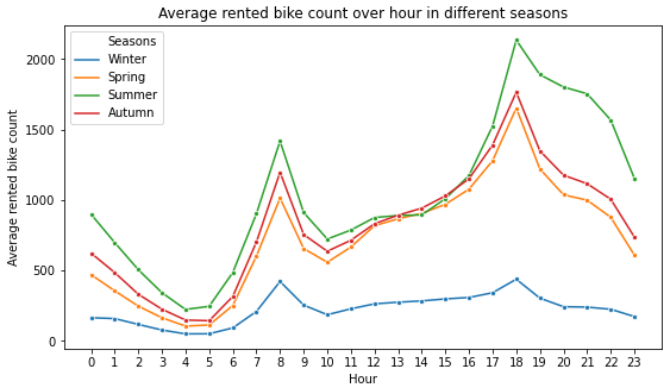


Figure 4

Figure 5 depicts the average rented bike count on non-holiday is greater than on holiday. The peak average rented bike count on non-holiday days, reaching approximately 1,600 bikes, occurs between 18:00 to 19:00 which is nearly twice as high as the maximum number during holidays within the same time frame.

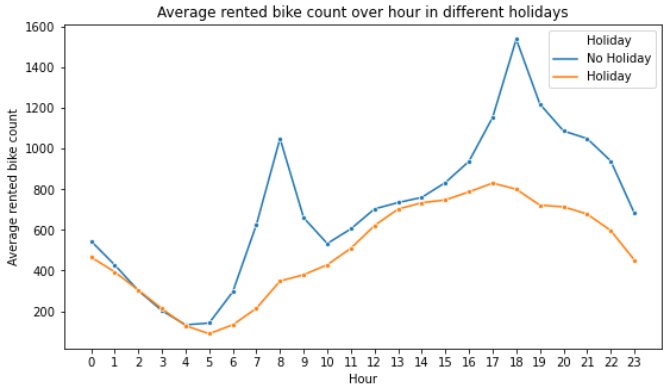


Figure 5

Figure 6 illustrates the average rented bike count on functioning days reaches the peak number with approximately 1,600 bikes between 18:00 and 19:00 while the average rented bike count on non-functioning days remains at zero.

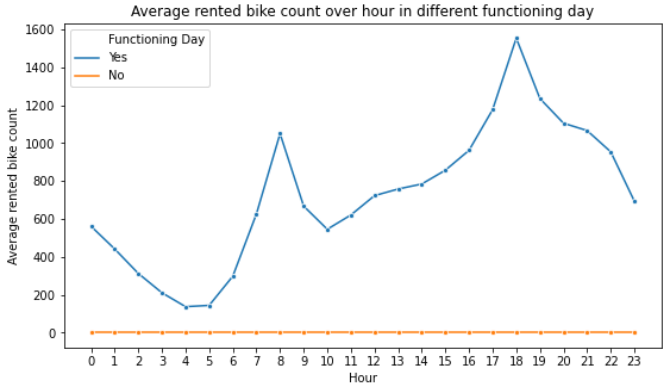


Figure 6

Figure 7 describes the peak average number of rented bikes on weekdays reaching over 1,600 bikes between 18:00 and 19:00, while the peak average

number of rented bikes on weekends reaches around 1,200 bikes between 17:00 and 18:00. The lowest average rented bikes count occurs between 4:00 and 5:00 on weekdays, and between 5:00 and 6:00 on weekends. In addition, between 10:00 and 17:00, the number of rented bikes on weekends surpasses that of the weekdays.

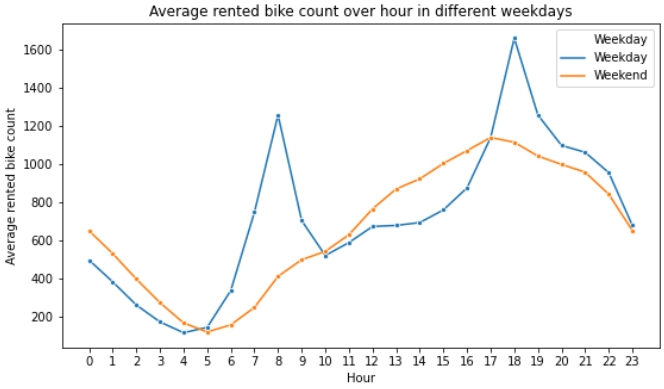


Figure 7

## C. Construction of models

This section of the analysis is dedicated to building the multiple linear regression model to analyse the relationship between the weather conditions (predictors), the hour (predictor) and the number of shared bike rentals (response).

The correlation matrix heatmap (Figure 8) reveals that the temperature has the strongest correlation with the rented bike count with a correlation coefficient of 0.54 among other features. On the other hand, within the features, the dew point temperature has a remarkably strong correlation to the temperature with a correlation coefficient of 0.91. To avoid multicollinearity, the dew point temperature feature is dropped as it has a weaker correlation to the rented bike count with a correlation coefficient of 0.38.

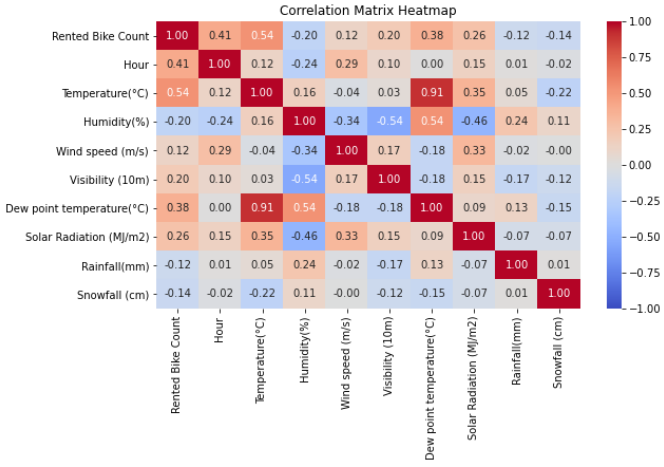


Figure 8

The dataset is split into an 80% training set and a 20% test set in which the training set is utilized for model generalization while the test set is employed to assess the model performance. After the multiple linear regression model is trained, the slope is as follows: 28.19 for the hour, 31.44 for temperature, -7.20 for humidity, 5.91 for wind speed, 0.02 for visibility, -81.70 for solar radiation, -61.42 for rainfall, and 18.87 for snowfall. These indicate that, for instance, a one-degree Celsius rise in temperature is associated with the 31 bikes increase, assuming all other predictors remain constant. Also, the intercept is 401.12, which is the fixed portion of bike rentals that are not affected by the predictors.

Figure 9 visualizes the histogram of residuals which exhibits a normal distribution. This provides evidence to support the assumption that the residuals follow a normal distribution.

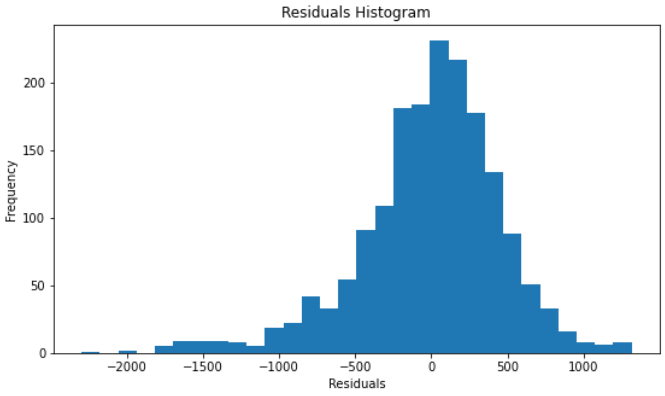


Figure 9

## D. Validation of results

To evaluate the performance of the multiple linear regression model, the mean absolute error, the mean squared error, and the coefficient of determination (R-squared) are assessed (Table 2).

First, the mean absolute error is the measure of the average difference between the predicted values by the model and the actual values of the observations. The mean absolute error is equal to 349 bikes which indicates the model predictions have an error of approximately 349 bikes on average.

Secondly, the mean squared error is a measure of the average squared difference between the predicted values by the model and the actual values of the observations. The mean-squared error is 222,944 bikes-squared which indicates the average variability of the model is 222,944 bikes squared.

Thirdly, the coefficient of determination (R-

squared) is the percentage of the variation in the dependent variable that can be explained by the independent variables. The R-squared is 0.46, indicating that 46% of the variation in the rented bike count can be explained by the hour and weather conditions.

| | |
|---|---|
| **Mean Absolute Error** | 349.49 |
| **Mean Squared Error** | 222944.06 |
| **Coefficient of Determination** | 0.4649 |

Table 2

**Findings, reflections, and future work:**

The trend analysis provides a comprehensive overview of the trend, by visual analytics approach, depicting the number of shared bike rentals from 1st December 2017 to 30th November 2018. In this period, the rented bike count starts with approximately 200,000 bike rentals and ends with nearly 500,00 bike rentals. The percentage growth is nearly 150%. Also, the line chart of the rented bike count over hours illustrates that the peak occurs in the period of 18:00 to 19:00 while the minimum occurs in the period of 4:00 to 5:00. The phenomenon is attributed to the fact that individuals leaving their workplace and choose cycling as the transportation mode for commuting to their residences. This remains true for the different seasons, holidays and function days, except the weekends.

The multiple linear regression model is constructed to investigate the relationship between the hour (predictor), the weather conditions (predictor) and the rented bike count (response). It is found that the temperature (degree Celsius) has the strongest correlation to the number of shared bike rentals with a correlation coefficient of 0.54. At the same time, to avoid multicollinearity, the feature of dew point temperature is removed. The model is developed under the approach of an 80% training set and a 20% test set splitting. The model is evaluated to have a moderate performance as the coefficient of determination is 0.46, revealing that only 46% of the variation in the dependent variables can be explained by the independent variables.

There is no doubt that the demand for shared bikes fluctuates in different seasons, holidays, functioning days and weekdays. The hour and weather conditions also influence the demand for shared bikes. There is a need to study other factors, such as the socio-demographic features, the location of the unmanned rental station, the coverage of the public transportation system and the price of bike rental, impacting the number of shared bike rentals.

**References:**

[1] Meddin Bike-sharing World Map, 2021. The Meddin bike-sharing world Map mid-2021 report. Available at: https://bikesharingworldmap.com/reports/.

[2] Seoul Bike Sharing Demand. (2020). UCI Machine Learning Repository. https://doi.org/10.24432/C5F62R.

[3] Osborne, Jason W. and Waters, Elaine (2019) "Four assumptions of multiple regression that researchers should always test," Practical Assessment, Research, and Evaluation: Vol. 8, Article 2.

**Words count:**

| Report | Word Counts |
|---|---|
| Abstract | 141 |
| Introduction | 232 |
| Analytical Questions | 77 |
| Data (Materials) | 265 |
| Analysis | 1130 |
| Finding, reflections, and further work | 320 |
| Total | 2165 |