University of the Western Cape

**The factors that form part of good predictors of a high usage day for bike rentals.**

A Report submitted in fulfillment of the requirements for the STA332 module

Group 6

By

Edwin Mahlangu

Xabsa Mohumed

Justine Linda Chiloane

4020163

4048604

3931585

Supervisor:

Mr. Matthew Wayne Valentine

[26 September 2022]

**Authors:**

Edwin Mahlangu

Xabsa Mohumed

Justine Linda Chiloane

Contents

[Abstract 2](#_heading=h.30j0zll)

[Research Question 2](#_heading=h.1fob9te)

[Literature Review 2](#_heading=h.3znysh7)

[Methodology 6](#_heading=h.tyjcwt)

[Analysis and Results 8](#_heading=h.3dy6vkm)

[Summary and Conclusion 15](#_heading=h.4d34og8)

[References 15](#_heading=h.2s8eyo1)

[Codes 16](#_heading=h.17dp8vu)

[Appendix 16](#_heading=h.3rdcrjn)

[*Table 1: Descriptive statistics* 16](#_heading=h.26in1rg)

[*Table 2: Summary of the forward selection(variables entered)* 16](#_heading=h.lnxbz9)

[*Table 3: Odds ratio of forward selection method* 16](#_heading=h.35nkun2)

[*Table 4: Summary of the Backward selection method (variables Removed)* 16](#_heading=h.1ksv4uv)

[*Table 5: Odds ratio of backward selection method* 16](#_heading=h.44sinio)

[*Table 6: Summary of the Stepwise selection(variables entered)* 16](#_heading=h.2jxsxqh)

[*Table 7: Odds ratios of Stepwise selection* 16](#_heading=h.z337ya)

[*Table 8: Analysis of Maximum Likelihood Estimates of backward selection method* 16](#_heading=h.3j2qqm3)

[*Figure 1: The Roc curve of Full logistic regression model* 16](#_heading=h.1y810tw)

[*Figure 2: The Roc curve of Forwarding Selection method* 16](#_heading=h.4i7ojhp)

[*Figure 3: The Roc cure of Backward selection method* 16](#_heading=h.2xcytpi)

[*Figure 4: The Roc curve for the Stepwise selection method* 16](#_heading=h.1ci93xb)

# **Abstract**

Bike rental is an enterprise that rents out bicycles as a return for profit, and it enhances comfortable mobility. The main objective of this study is to determine factors that influence the high usage of bike rentals. This study contributes to researching how the change in weather conditions (like an increase in rainfall, humidity, or temperature) affects bike rental usage by analyzing the given data and commenting on whether the recent studies have produced good analysis results. This paper will provide a predictive multiple logistic regression model that will help predict high usage bike rentals. The multiple logistic regression model is produced from a succession of statistical methods and selection methods that use analytical data in different ways to come up with the best final multiple logistic regression model to predict high usage bike rentals. The methods used include descriptive statistics, simple logistic regressions, hypothesis testing, odds ratio interpretation, multiple logistic regression including the selection techniques forward, backward, and stepwise, and the roc curve. A final model is created by using different multiple regression selection methods in SAS software, which eliminates all unnecessary variables.

The final multiple logistic regression model was selected using the forward selection technique as it is statistically proven to be the best with a roc curve area of 0.9429. Not all variables are included in the model; all the insignificant ones were dropped. The variables kept in the model are “Hour, Humidity, Visibility, Dew\_point\_temperatur, Solar\_Radiation, Rainfall, Snowfall, Spring, Summer, Autumn, Function\_Day". Therefore, any business or company that is involved in bike rental infrastructure should consider looking at factors like temperature, humidity, and wind speed when supplying the bikes. This answers our created research question and the main research question.

# **Research Question**

The main research question is stated below:

Which factors are good predictors of a high usage day for bike rentals?

Formulated research questions:

Does a change in weather conditions (like an increase in rainfall, humidity, or temperature) affects the usage of bike rentals?

**Literature Review**

**Introduction**

Many cities around the world have become more environmentally conscious, and as a result, they have implemented bike rental systems to reduce pollution caused by automobiles while earning profit. Seoul is one of these cities. However, despite people's efforts to reduce their carbon footprint, some factors impact bike rentals of them is weather conditions. Unfavourable weather conditions including heavy rainfalls, strong winds, snowfall, and low temperatures pose significant cycling hurdles. Cycling in these sorts of weather is difficult for those who rely only on bicycles for transportation because, firstly, it increases the chance of accidents, and, secondly, it is uncomfortable. Contrarily, pleasant weather conditions include warm temperatures, or no wind, which are ideal for commuters who ride bicycles. Overall good weather is associated with increased bike usage, while bad weather has been reported to cause discomfort, making bicycles impossible to ride, thus resulting in lowered utilization. (Amiri & Sadeghpour, 2015; Flynn, Dana, Sears, & Aultman-Hall, 2012). This literature review examines the factors that lead to high bike rentals. Additionally, it illustrates how weather conditions like an increase in rainfall, humidity or temperature affect the usage of bike rentals.

**Weather as a factor:**

Poor weather affects high bike usage, shifting demand to alternatives as it reduces vision and makes cycling unsafe. (Corcoran et al., 2014; Dell’Olio et al., 2014). Swiers et al. believe that the greatest impediment to increased bike utilization is the weather. (Swiers et al., 2017). Bad weather is distinguished by heavy rain, high winds, cold, and snow (Flynn et al., 2012; Parkin et al., 2008). Extreme weather, such as heavy rain, snow, and storms, will halt the usage of bike rental services. (El-Assi et al., 2017; Heaney et al., 2019). Higher temperatures have been reported to increase bike utilization in cold climate countries (Medard de Chardon et al., 2017; Zhao et al., 2019), whilst the opposite is true in warm climate countries. Faghih-Imani and Eluru discovered that low temperatures have a deleterious influence on Montreal. Sweating and discomfort are caused by high humidity, whereas low humidity can cause skin peeling and eye irritation. (Faghih-Imani and Eluru, 2016) As a result, bottlenecks might arise on both the temperature and humidity spectrums. However, encountering both extremes in the same region would be unusual. In cold climate areas, a greater temperature is required for high-rental bike services. Higher humidity is a required factor for the demand for rental bike services in cold climate areas. High bike rentals have a negative connection with wind velocity (Heinen et al., 2011; Lin et al., 2020). Slow winds, according to Mattson and Godavarthy, attract more riders. As the wind speed increases, so does the resistance and difficulty in balancing the bike (Mattson and Godavarthy, 2017). Furthermore, it reduces visibility on the road. Lower wind speed, on the other hand, is a requirement for high bike rentals. (Autio et al., 2013). This literature tends to focus on how an increase in rainfall, temperature and humidity impacts high bike rental utilization.

*Increase in rainfall:*

Increased Rain is generally considered to be the most adverse weather condition and there is a negative relationship between rainy days and high bike usage. (Corcoran, Li, Rohde, Charles-Edwards, & Mateo-Babiano, 2014; Gebhart & Noland, 2014; Hyland et al., 2018; Kim, 2018; Sun, Chen, & Jiao, 2018). Additionally, heavy rain significantly reduces high bike usage, and 3 hours later, the demand for a journey after heavy rain returns to normal (Reiss & Bogenberger, 2016).

Humidity is a weather variable that, along with temperature and precipitation, is a key indicator of whether people will rent bikes. High bike rentals are negatively impacted by both high and low humidity. (Gallop, Tse, & Zhao, 2012). Kim investigated the temperature-humidity index, relative humidity, and wind velocity meteorological variables at all hours of the day to assess the combined effect of temperature and humidity dependency and discovered that all the relevant factors had a negative influence on bike sharing. (Kim, 2018) Furthermore, Miranda-Moreno and Nosal found that a 10% rise from 14.7 °C results in an average increase of 4% to 5% in hourly volume. (Miranda-Moreno and Nosal, 2014) However, they discovered that if the humidity hit 60% at temperatures over 28 °C, it would decrease. Using data from the BSP in Washington, Gebhart and Noland discovered that humidity harmed bike sharing. (Gebhart and Noland, 2014). Furthermore, many researchers believe that increasing humidity reduces trip generation (Corcoran et al., 2014; Croci & Rossi, 2014; Sun et al., 2018; Wang et al., 2018).

*Temperature:*

One of the most researched factors is the impact of temperature on bike-sharing. (Spencer et al., 2013). It is discovered that there is a positive relationship between rising temperatures and increased demand for bike sharing. Trip production is favourably connected with weather temperatures ranging from 0 to 20 degrees Celsius; bike-sharing demand peaks between 20 and 30 degrees Celsius (Heinen et al., 2010; Hyland, Hong, de Farias Pinto, & Chen, 2018; Kim, 2018; Wang, Akar, & Chen, 2018).

The conclusions of studies on the impact of temperatures over 30 degrees Celsius on bike sharing are contradicting since some researchers regard temperatures over 30 degrees Celsius to be the optimal temperature for high bike-rental demand, while others consider it to be "scorching hot". (Castells-Graells et al., 2020). El-Assi, Mahmoud, and Habib found in their study that the demand for bike-sharing in Toronto, where temperatures may exceed 42 °C, increases with temperatures of +30 °C. (El-Assi, Mahmoud, and Habib, 2018). In contrast to this report, Kim claimed that South Korea only saw +30 °C on 49 days in 2015 and that the Tashu Bike sharing program referred to +30 °C as "scorching heat." The intense heat, therefore, had a detrimental impact on travel demand. (Kim,2018) However, Jing and Zhao discovered that the optimum weather temperature for increased bike-sharing demand is between 30-35 °C. (Jing and Zhao, 2015). Thus, finding an ideal temperature for bike ridership that works for everyone may be challenging because various places may experience different effects from the same temperature.

**Conclusion:**

Bike-sharing system program (BSP) development is now believed to be important for reducing carbon footprint. As a result, the number of researchers interested in studies that advocate the use of BSP has expanded in recent years. This study focused on weather conditions that influenced high bike rentals in the literature. Increased rain, high temperatures, and humidity all decrease demand for bike-sharing. If there is no precipitation at 20-30 °C weather temperatures, the number of bike sharing is more likely to grow in comparison to other weather circumstances. Temperatures that are higher than the maximum limit value for the local climate have a detrimental impact on high bike rentals. The necessary condition hypotheses can be retested in future research at tourist attractions for casual/leisurely bike rides and close to workplaces for regular use. The researcher looking into the use of public bikes can also look into the future effects of any new obstacles brought forth by COVID-19.

# **Methodology**

**Introduction**

This chapter explains the methodologies that were used to analyse the data as applicable to the research. It gives a full description of each research method that was employed, through sections with subheadings. The main theme of the research is to identify factors that will predict a high-usage day for bike rentals. To determine the factors the following methodologies are used; firstly, descriptive statistics are analysed, then secondly the correlation analysis, thirdly the hypothesis testing, and lastly the predictive model with selection techniques.

**Descriptive statistics**

The descriptive statistics were calculated to get a clear understanding of how the predictor variables of bike rental relate to each other. The method was implemented to help gain insight into variables and the spread of the data. Under this method, histograms were implemented for every variable, to understand the distribution of the data, furthermore, the skewness and kurtosis were computed to confirm the symmetry of the data from the histogram statistically.

**Correlation analysis.**

This type of analysis is undertaken to understand if there is an association or relationship between the continuous variables. This analysis will shed light on which predictor variables are highly correlated and might contain the same information. Thus, these variables might be affected by multicollinearity.

**Simple logistic regressions**

Simple logistic regressions are used to check Odds ratios which measure the importance of predictor variables related to High usage of bike rentals. Moreover, to determine if there’s an association between the response variable (High usage) and each of the predictor variables, using the Wald-test hypothesis.

**Hypothesis testing**

The Likelihood ratio test hypothesis is run to test the significance of the variables on the model through the ANOVA approach; this is an overall test on all Bi’s if they are equal to zero. The test statistic and the p-value are produced by the SAS software. The constraint is that if the p-value is less than 0.05 which is the significance level, then the conclusion is that some of the Bi’s are not equal to zero, hence the model is predictive.

**Multiple Logistic regression model:**

The response variable is binary hence, a multiple binary logistic regression model is utilized to predict the probability of bikes rented per hour. The Multiple logistic regression selection techniques were utilized using the criterions SLSTAY at 0.05 and SLENTRY at 0.05 as the cutoff point. We use these selection methods to identify the best model that answers the main research question by determining the good predictors of a high usage day for bike rentals and respond to our created research question if an increase in rainfall, humidity and temperature do cause an increase in the number of bike rentals supplied. The logistic regression estimates for the response variable were calculated using the multiple logistic regression and the selection techniques which are the forward, backward, and stepwise selection methods. The odds ratios were calculated to measure the importance of the independent variables relative to the response.

**Roc Curve**

The ROC curves were constructed to measure the predictive ability of the model, this method will help us by finding the best model that to respond to the main research question which is to identify good factors that are good predictors of a high usage day for bike rentals and respond to the created research question.

**Conclusion**

In conclusion, the above-mentioned methods will be employed to aid in answering the main question which is finding the good predictors or factors of high usage day for bike rentals. We shall therefore look in detail at the factors in the next section and create a model that best answers the research questions. In the coming section, the results will be discussed.

# **Analysis and results**

**Descriptive statistics**

The univariate analyses for continuous variables help analyse the variables on how they are distributed and how the data in between the variables vary from each other e.g., on the variable hour, we will be able to see at what hour most bikes consider to be used at high usage. The distribution discussed are namely the normal distribution, platykurtic and leptokurtic. Platykurtic distribution is defined as a distribution that has a negative kurtosis( Gen,2022, no pagination). and leptokurtic distribution is defined as distribution that has excess positive kurtosis, where the kurtosis is greater than 3( Gen,2022, no pagination). *Refer to table 1 in the appendix for descriptive statistics results.*

The average of the variable hour is 11.52 with a skewness of -0.0265 indicating that the variable is slightly skewed to the left and kurtosis of -1.1469 meaning that the variable’s distribution is platykurtic with a bit shorter and thinner tail to the left and it’s low peaked, almost normally distributed. The mean of the variable temperature is 13.1574 degrees Celsius, with skewness of -0.20589 telling us the variable temperature is skewed to the left and kurtosis of -0.8192 meaning the variable’s distribution has a thin tail and a central peak that is lower as compared to the normal distribution. The average humidity in column data is 58.33 degrees Celsius and the skewness of this variable is 0.0888 meaning that the variable is slightly skewed to the right. The variable humidity has a kurtosis of -0.8369 indicating that the variable’s distribution is platykurtic, with just a bit broader peak than the peak of a normal distribution. The average wind speed is 1.7775 km/h, and the skewness of this variable is 0.91777 telling us that the variable is skewed to the right. The variable Wind\_speed has a kurtosis of 1.02 indicating that the variable’s distribution is slightly leptokurtic with a longer and fatter tail, and a bit higher and sharper central peak compared to normal distribution. The average visibility in the column data is 1451.96, and the skewness of this variable is -0.74105 indicating that the variable is skewed to the left. The variable visibility has a kurtosis of -0.8978 indicating that the distribution has a bit shorter and thinner tail. The mean of the variable dew\_point\_temperature is 4.327 degrees Celsius, and the skewness of this variable is -0.4034 indicating that the variable is skewed to the left with a shorter tail and shallow peak, this is indicated by a kurtosis of -0.7192. The average solar radiation in the column data is 0.5993 and the skewness of this is 1.4080 indicating that the variable is skewed to the right. The variable has a kurtosis of 0.7868 which indicates that is approximately highly symmetrical. The rainfall variable has an average of 0.1329 mm with skewness of 19.7690 which indicates that the variable is skewed to the right. The variable has a kurtosis of 507.4593 which is greater than the expected value 3, this indicates that the variable’s distribution has a longer tail to the right and a higher central peak relative to the normal distribution. Lastly, we have the variable snowfall with a mean of 0.0684 mm and skewness of 9.6952 indicating that the variable is skewed to the right. The variable has a kurtosis of 134.1319 which is greater than the expected value 3, this indicates that the variable’s distribution has a longer tail to the left and a higher central peak.

**Correlation analysis**

The Pearson correlation analysis is undertaken to understand the association or relationship between the continuous variables. This analysis will give light on which predictor variables are highly correlated and might contain the same information. Thus, these variables might be affected by multicollinearity. *Refer to table 2 in the appendix for the correlation matrix results.*

Results found by one of the researchers were “Temperature and Hour of the day is the most influential variable for rental bike sharing demand prediction” (Sathishkumar and Cho, 2020), now back to our research results of correlation to find out if temperature and hour will or will not affect the high usage day for bike rentals, we find that the correlation between temperature and hour is a very weak positive correlation 0.09026. From the above findings, temperature and hour variables are unlikely to suffer from multicollinearity. From the correlation analysis, we found that Dew\_point\_temperature and temperature have a strong positive correlation of 0.91327 indicating that these variables contain the same information, and this might lead to our final chosen multiple logistic regression that has both variables to suffer from multicollinearity.

We are now eager to know if the three predictor variables temperature, hour and dew\_point\_tempeture will be included in the multiple logistic regression model.

**Simple logistic regressions:**

As mentioned, the odds ratios are calculated to measure the importance of predictor variables related to the count of bikes rented at each hour, and the hypothesis using the Wald test is conducted to determine if there is an association between the binary variable and the predictor variables, all, this is done by implementing simple logistic regressions using the SAS software.

*Hypothesis testing:*

Ho: There is no association between the two variables.

Ha: There is an association between the two variables.

*Odds ratios and conclusions:*

For Hour, odds ratio estimates are 1.097. Meaning the odds of high usage bike rentals are 1.097 times greater for each hour increase. Since the Wald Test indicates that 56.6328 with a p-value of 0.0001 satisfies the significance level of 0.05, therefore the null hypothesis is rejected, there is an association between the response variable (high usage) and hour.

For Humidity, the odds of high usage bike rentals are 0.981 times greater for each unit (%) increase in Humidity. Since the Wald Test indicates that 59.6328 with a p-value of 0.0001 satisfies the significance level of 0.05, therefore the null hypothesis is rejected, there is an association between the response variable (high usage) and humidity.

For Temperature, the odds of high usage bike rentals are 1.125 times greater for each unit increase in Temperature. This means that the odds ratio Temperature indicates that the higher the temperature the more the bikes are rented. Since the Wald Test indicates that 401.3296 with a p-value of 0.0001 satisfies the significance level of 0.05, therefore the null hypothesis is rejected, there is an association between the response variable (high usage) and temperature.

In this model, the odds of high usage bike rentals are 1.121 times greater for each Miles/hour increase In Wind speed. Since the Wald Test indicates that 6.1518 with a p-value of 0.0131 satisfies the significance level of 0.05, therefore the null hypothesis is rejected, there is an association between the response variable (high usage) and wind speed.

For Visibility, the odds of high usage bike rentals are 1.001 times greater for each 10m increase in visibility range. Since the Wald Test indicates that 41.0818 with a p-value of 0.0001 satisfies the significance level of 0.05, therefore the null hypothesis is rejected, there is an association between the response variable (high usage) and visibility.

For Dew\_point\_temperature, the odds of high usage bike rentals are 1.070 times greater for each unit increase in dew point temperature. Since the Wald Test indicates that 235.3791 with a p-value of 0.0001 satisfies the significance level of 0.05, therefore the null hypothesis is rejected, there is an association between the response variable (high usage) and dew point temperature.

In this model, the odds of high usage bike rentals are 3.058 times greater for each unit increase in Solar radiation. Since the Wald Test indicates that 200.6367 with a p-value of 0.0001 satisfies the significance level of 0.05, therefore the null hypothesis is rejected, there is an association between the response variable (high usage) and solar radiation.

For Rainfall, the odds ratio estimates for Rainfall is 0.058, this means that the odds of high usage bike rentals are 0.058 times greater for each 1mm increase in Rainfall fall. Since the Wald Test indicates that 26.005 with a p-value of 0.0001 satisfies the significance level of 0.05, therefore the null hypothesis is rejected, there is an association between the response variable (high usage) and rainfall.

For Snowfall, the odds ratio estimates for Snowfall is 0.013, this means that the odds of high usage of bike rentals are 0.013 times greater for each 1mm increase in Snowfall. Since the Wald Test indicates that 17.7682 with a p-value of 0.0001 satisfies the significance level of 0.05, therefore the null hypothesis is rejected, there is an association between the response variable (high usage) and snowfall.

For Season, odds ratio (1 vs 4) estimates are 19.597. This means that cases with a 1 value for the season are 19.597 times more likely to result in high usage of bike rentals than bikes with a 4 value for season. Since the Wald Test indicates that 21.2802 with a p-value of 0.0001 satisfies the significance level of 0.05, therefore the null hypothesis is rejected, there is an association between the response variable (high usage) and spring.

For Season, odds ratio (2 vs 4) estimates are 43.685. This means that cases with a 2 value for the season are 43.685 times more likely to result in high usage bike rentals than bikes with a 4 value for season. Since the Wald Test indicates that 152.5768 with a p-value of 0.0001 satisfies the significance level of 0.05, therefore the null hypothesis is rejected, there is an association between the response variable (high usage) and summer.

For Season, odds ratio (3 vs 4) estimates are 29.349. This means that cases with a 3 value for the season are 74.3488 times more likely to result in high usage bike rentals than bikes with a 4 value for the season. Since the Wald Test indicates that 56.6328 with a p-value of 0.0001 satisfies the significance level of 0.05, therefore the null hypothesis is rejected, there is an association between the response variable (high usage) and autumn.

For Holiday, odds ratio (0 vs 1) estimates are 2.126. This means that cases with a 0 value for holiday are 2.126 times more likely to result in high usage of bike rentals than days with a 1 value for holiday. Since the Wald Test indicates that 11.0225 with a p-value of 0.0009 satisfies the significance level of 0.05, therefore the null hypothesis is rejected, there is an association between the response variable (high usage) and No Holiday.

For Functioning Day, odd ratio (0 vs 1) estimates are <0.001. This means that cases with 0 value for a Functioning day are 0.001 more like to result in high usage of bike rentals than days with a 1 value for a functioning day. Since the Wald Test indicates that 0.0026 with a p-value of 0.0001 does not satisfies the significance level of 0.05, therefore the null hypothesis is not rejected, there is no association between the response variable (high usage) and No function day.

**Multiple logistic regression analysis**

**Full Multiple Logistic regression model:**

Computing a model that will produce the best-predicted outcomes. This model will help us get the variables that are highly related to the response variable high usage. We also used the subset selection methods to help us eliminate variables that are not important and keep the significant ones. We used this model to report whether the “null hypothesis that all the ”( Elliott, Woodward,2010, p.261). We also analyzed the roc curve under this model, which measures the predictiveness of the model. *Refer to table 3 in the appendix for the following results.*

*hypothesis*

Ho: Bi = 0: The ith independent variable is not predictive of the probability of occurrence

Ha: Bi# 0: The ith independent variable is predictive of the occurrence.

*Conclusion:*

We can conclude that the Likelihood ratio test indicates that 1369.1238 with p-value of <0.001 satisfies the significance level of 0.05. Therefore, it can be assumed that at least some of the variables are useful in predicting the response variable.

**Forward Selection Model**:

The backward selection method is defined as the “method [the] considers all predictor variables and eliminates the ones that do not meet the minimal SLSTAY criterion until only those meeting the criterion remain[s]” (Elliott, Woodward,2010, p.231). The SLSTAY criterion, in this case, is 0.05. *Table 4* shows the summary of the forward selection method, the variables entered the model are namely “Temperature, Humidity, Seasons,Function\_Day,Hour,Solar\_Radiation,Rainfall,Visibility,Dew\_point\_temperature, Snowfall”. All these variables satisfy the significance level of 0.05 and they play an important role in predicting the high usage of bike rentals.

As mentioned, the odds ratios are calculated to measure the importance of predictor variables related to count of bikes rented at each hour. *Table 5* shows the odds ratios of this model which are interpreted as follows:

Odds conclusion:

For Hour, odds ratio estimates are 1.146. Meaning the odds of high usage bike rentals are 1.097 times greater for each hour increase.

For Temperature, the odds of high usage bike rentals are 0.954 times greater for each unit increase in Temperature. This means that the odds ratio Temperature indicates that the higher the temperature the more the bikes are rented

For Humidity, the odds of high usage bike rentals are 0.919 times greater for each unit (%) increase in Humidity.

For Visibility, the odds of high usage bike rentals are 1.00 times greater for each 10m increase in visibility range.

For Dew\_point\_temperature, the odds of high usage bike rentals are 1.189 times greater for each unit increase in dew point temperature.

In this model, the odds of high usage bike rentals are 2.109 times greater for each unit increase in Solar radiation.

For Rainfall, the odds ratio estimate for Rainfall is 0.035, this means that the odds of high usage bike rentals are 0.033 times greater for each 1mm increase in Rainfall fall.

For Snowfall, the odds ratio estimates for Snowfall is 0.308, this means that the odds of high usage of bike rentals are 0.310 times greater for each 1mm increase in Snowfall.

For Season, odds ratio (1 vs 4) estimates are 14.601. This means that cases with a 1 value for the season are 14.686 times more likely to result in high usage of bike rentals than bikes with a 4 value for the season.

For Season, odds ratio (2 vs 4) estimates are 15.535. This means that cases with a 2 value for the season are 15.606 times more likely to result in high usage bike rentals than bikes with a 4 value for the season.

For Season, odds ratio (3 vs 4) estimates are 55.588. This means that cases with a 3 value for the season are 56.163 times more likely to result in high usage bike rentals than bikes with a 4 value for the season.

For Functioning Day, odd ratio (0 vs 1) estimates are <0.001. This means that cases with a 0 value for a Functioning day are 0.001 more like to result in high usage of bike rentals than days with a 1 value for a functioning day.

**Backward selection Model:**

The forward selection method is defined as the “method brings in the most significant variable that meets the SLENTRY criterion and continues entering variables until none meets the criterion” (Elliott, Woodward,2010, p.231). The SLSTAY criterion in this case is 0.05. *Table 6* shows the summary of the Backward selection method, the variables removed from the model are namely “Functioning\_Day, Temperature, day\_part, Holiday, month\_part, Wind\_speed, Snowfall”. All these variables do not satisfy the significance level of 0.05 and they do not play a significant role in predicting the high usage of bike rentals, so they were dropped, and variables kept due to significant importance are namely “Hour, Humidity, Visibility, Dew\_point\_temperature, Solar\_Radiation and Seasons (Spring, Summer, Autumn)”

Odd conclusions: *Refer to table 7 for the odds estimates*

For Hour, odds ratio estimates are 1.123. Meaning the odds of high usage bike rentals are 1.097 times greater for each hour increase.

For Humidity, the odds of high usage bike rentals are 0.945 times greater for each unit (%) increase in Humidity.

For Visibility, the odds of high usage bike rentals are 1.00 times greater for each 10m increase in visibility range.

For Dew\_point\_temperature, the odds of high usage bike rentals are 1.093 times greater for each unit increase in dew point temperature.

In this model, the odds of high usage bike rentals are 1.606 times greater for each unit increase in Solar radiation.

For Rainfall, the odds ratio estimate for Rainfall is 0.048, this means that the odds of high usage bike rentals are 0.048 times greater for each 1mm increase in Rainfall fall.

For Season, odds ratio (1 vs 4) estimates are 15.859. This means that cases with a 1 value for the season are 15.859 times more likely to result in high usage of bike rentals than bikes with a 4 value for the season.

For Season, odds ratio (2 vs 4) estimates are 23.708. This means that cases with a 2 value for the season are 23.708 times more likely to result in high usage bike rentals than bikes with a 4 value for the season.

For Season, odds ratio (3 vs 4) estimates are 33.386. This means that cases with a 3 value for the season are 33.386 times more likely to result in high usage bike rentals than bikes with a 4 value for the season.

The roc curve of this model is shown in Figure 3 in the appendix, the roc curve was constructed to measure the predictiveness of the model. The “Area Under the Curve” (AUC) is 0.9176, which is greater than 0.7 and closer to 1.0. This indicates that the model is strong in predicting the response variable Rented\_bike\_count of high usage.

**Stepwise Selection Model:**

The stepwise selection method is defined as “a mixture of the two [backward and forward selection methods] ; it begins like the FORWARD method but reevaluates variables at each step and may eliminate a variable if it does not meet the SLSTAY criterion” (Elliott, Woodward,2010, p.231).The SLSTAY criterion in this case is 0.05. *Table 8* in the appendix shows the summary of the stepwise method. The variables left in the model according to the summary table “Temperature, Seasons, Humidity”.

As previously mentioned, the odds ratios are calculated to measure the importance of predictor variables related to count of bikes rented at each hour. *Table 9 in the appendix* shows the odds ratios of this model which are interpreted as follows:

Odds conclusion:

For Temperature, the odds of high usage bike rentals are 1.108 times greater for each unit increase in Temperature. This means that the odds ratio Temperature indicates that the higher the temperature the more the bikes are rented

For Humidity, the odds of high usage bike rentals are 0.949 times greater for each unit (%) increase in Humidity.

For Season, odds ratio (1 vs 4) estimates are 83.608. This means that cases with a 1 value for the season are 83.608 times more likely to result in high usage of bike rentals than bikes with a 4 value for the season.

For Season, odds ratio (2 vs 4) estimates are 321.995. This means that cases with a 2 value for the season are 321.995 times more likely to result in high usage bike rentals than bikes with a 4 value for the season.

For Season, odds ratio (3 vs 4) estimates are 158.356. This means that cases with a 3 value for the season are 158.356 times more likely to result in high usage bike rentals than bikes with a 4 value for the season.

**Roc curve and Conclusion:**

The ROC chart is used to measure how well the model predicts the outcomes (Elliott, Woodward,2010, p.237). The roc curves of backward, forward, and stepwise selection methods of the multiple logistic regression models are compared to choose the best predictive model. The selection methods roc curves are also compared with the full multiple logistic regression roc curve.

The full multiple regression model is constructed without considering any of the selection methods, so it can be used to compare its importance against the selection methods. So, the comparison of these methods will assist in seeing if the selection methods are effective or not. The aim of these selection methods is to create a model that will contain what are necessary in predicting the high usage count of rental bikes per day. The selection methods are namely forward, stepwise, and backward, which are procedures that can be run through SAS.

*Roc curve of full multiple logistic regression: Refer to figure 1 in the appendix.* The “Area Under the Curve” (AUC) is 0.9432 which is greater than 0.7, and closer to 1.0. This indicates that the model is strong in predicting the response variable Rented\_bike\_count of high usage

*Roc curve of forward selection model: Refer to figure 2 in the appendix.* The “Area Under the Curve” (AUC) is 0.9429, which is greater than 0.7 and closer to 1.0. This indicates that the model is strong in predicting the response variable Rented\_bike\_count of high usage

*Roc curve of backward selection model: Refer to figure 3 in the appendix.* The “Area Under the Curve” (AUC) is 0.9176, which is greater than 0.7 and closer to 1.0. This indicates that the model is strong in predicting the response variable Rented\_bike\_count of high usage.

*Roc curve of stepwise selection model: Refer to figure 4 in the appendix.* The “Area Under the Curve” (AUC) is 0.8878, which is greater than 0.7 and closer to 1.0. This indicates that the model is strong in predicting the response variable Rented\_bike\_count of high usage.

**Conclusion:**

To choose the best model from the selection techniques. We focused on the Roc curve of each selection technique discussed since it measures the predictiveness of the model. As mentioned, the Roc curve for the forward selection is 0.9429, for backward is 0.9176, and for stepwise is 0.8878 then from these three techniques we chose the forward model as it proves to be more predictive, and its Roc curve closer to 1, indicating that the regression fits the data very well. The final model will be used to predict the high usage of bike rentals in a day.

Based on the forward selection method, it is observed from the results in Table 10 (Analysis of Maximum Likelihood Estimates) in the appendix that the only predictor variables that are important in predicting high usage of bike rentals in a day are namely as “Hour, Humidity, Visibility, Dew\_point\_temperatur, Solar\_Radiation, Rainfall, Snowfall, Spring, Summer, Autumn, Function\_Day”. Table 8 also provides the respective parameter estimates for the model.

Estimated Multiple logistic regression model:

Let Q = -6.6142 +0.1366\*Hour – 0.04668\*Temperature -0.0843\*Humidity-0.00043\*Visibility + 0.1734\*Dew\_point\_temperatur+0.7462\*Solar\_Radiation-3.3624\*Rainfall -1.1776\*Snowfall+0.3206\*Spring + 0.3826\*Summer+1.6574\*Autumn-10.3879\*Function\_Day

Therefore, the estimated logistic regression model is

P(High usage) =

# **Summary and Conclusion:**

The purpose of this study was to run an analysis on the Seoul Bike Sharing Demand and provide companies and businesses with the factors that are related to high usage bike rentals. And lastly, build a multiple logistic regression which will help the business to predict the number of bikes to supply the city with. The final multiple logistic regression model included the variables “Hour, Humidity, Visibility, Dew\_point\_temperatur, Solar\_Radiation, Rainfall, Snowfall, Spring, Summer, Autumn, Function\_Day”. Some variables have proved not be normally distributed. The logistic regression is produced using different selection models namely, forward, backward, and stepwise. The Roc curves which are produced in all the selection models helped us choose the best model that answers the main research question and the formulated research question. High bike rentals are reduced by increased rain, high temperatures, and humidity. If there is no precipitation and the temperature is between 20 and 30 degrees Celsius, the number of bike share users is more likely to grow than in other weather situations. Temperatures that exceed the maximum limit for the local climate have a negative impact on bike-sharing

# **References**

Alan C. Elliott, & Wayne A. Woodward. (2010). SAS essentials: a guide to mastering SAS for research.

Giot, R., Cherrier, R., (2014). Predicting bikeshare system usage up to one day ahead. In: 2014 IEEE Symposium on Computational Intelligence in Vehicles and Transportation Systems (CIVTS). Presented at the 2014 IEEE Symposium on Computational Intelligence in Vehicles and Transportation Systems (CIVTS), pp. 22–29. <https://doi.org/10.1109/CIVTS.2014.7009473>.

Lin, L., He, Z., & Peeta, S. (2018). Predicting station-level hourly demand in a large-scale bike sharing network: A graph convolutional neural network approach. *Transportation Research Part C*. <https://doi.org/www.elsevier.com/locate/trc>

Eren, E., & Uz, V. E. (2020). A review on bike-sharing: The factors affecting bike-sharing demand. *Sustainable Cities and Society*. <https://doi.org/https://www.sciencedirect.com/journal/sustainable-cities-and-society>

Sathishkumar V E & Yongyun Cho (2020) A rule-based model for Seoul Bike sharing demand prediction using weather data, European Journal of Remote Sensing, 53:sup1, 166-183. <https://doi.org/10.1080/22797254.2020.1725789>

Kumar, D. (2021). Meteorological barriers to bike rental demands: A case of Washington D.C. using NCA approach. *Case Studies on Transport Policy*. <https://doi.org/https://www.journals.elsevier.com/case-studies-on-transport-policy>

[Stephanie Glen](https://www.statisticshowto.com/contact/). "Kurtosis: Definition, Leptokurtic, Platykurtic" From [StatisticsHowTo.com](https://www.statisticshowto.com/): Elementary Statistics for the rest of us! <https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/kurtosis-leptokurtic-platykurtic/>

# **Codes**

/\* Converting date format to numeric \*/

**data** new project.group\_6\_train\_new;

set project.group\_6\_train;

date\_new = input(Date,DDMMYY10.);

**run**;

**data** new PROJECT.group\_6\_train\_new1;

set PROJECT.group\_6\_train\_new;

day\_part = day(date\_new);

month\_part = month(date\_new);

year\_part = year(date\_new);

drop date\_new;

drop \_dataobs\_;

drop Date;

**run**;

/\* Calculating the descriptive statistics \*/

**proc** **means** data=project.group\_6\_train\_new1 mean skewness kurtosis;

var Hour temperature humidity wind\_speed visibility Dew\_point\_temperature Solar\_Radiation Rainfall Snowfall day\_part month\_part year\_part;

**run**;

/\* Running correlation analysis between continuous variables \*/

**proc** **corr** data = project.group\_6\_train\_new1;

var Hour temperature humidity wind\_speed visibility Dew\_point\_temperature Solar\_Radiation Rainfall Snowfall day\_part month\_part year\_part;

**run**;

/\* Running simple logistic regression between the binary response variable and predictor variables \*/

**proc** **logistic** data = project.group\_6\_train\_new1 descending;

model rented\_bike\_count = hour/expb ;

**run**;

**proc** **logistic** data = project.group\_6\_train\_new1 descending;

model rented\_bike\_count = humidity/expb ;

**run**;

**proc** **logistic** data = project.group\_6\_train\_new1 descending;

model rented\_bike\_count = temperature/expb;

**run**;

**proc** **logistic** data =project.group\_6\_train\_new1 descending;

model rented\_bike\_count = wind\_speed/expb;

**run**;

**proc** **logistic** data = project.group\_6\_train\_new1 descending;

model rented\_bike\_count = visibility/expb;

**run**;

**proc** **logistic** data = project.group\_6\_train\_new1 descending;

model rented\_bike\_count = Dew\_point\_temperature/expb;

**run**;

**proc** **logistic** data = project.group\_6\_train\_new1 descending;

model rented\_bike\_count = Solar\_radiation/expb;

**run**;

**proc** **logistic** data =project.group\_6\_train\_new1 descending;

model rented\_bike\_count = Rainfall/expb ;

**run**;

**proc** **logistic** data = project.group\_6\_train\_new1 descending;

model rented\_bike\_count = Snowfall/expb ;

**run**;

**proc** **logistic** data =project.group\_6\_train\_new1 descending;

class seasons;

model rented\_bike\_count = seasons /expb ;

**run**;

**proc** **logistic** data =project.group\_6\_train\_new1 descending;

class holiday;

model rented\_bike\_count = holiday /expb ;

**run**;

**proc** **logistic** data =project.group\_6\_train\_new1 descending;

class functioning\_day;

model rented\_bike\_count = functioning\_day /expb ;

**run**;

**proc** **logistic** data =project.group\_6\_train\_new1 descending;

model rented\_bike\_count = day\_part /expb ;

**run**;

**proc** **logistic** data =project.group\_6\_train\_new1 descending;

model rented\_bike\_count = month\_part /expb ;

**run**;

**proc** **logistic** data =project.group\_6\_train\_new1 descending;

model rented\_bike\_count = year\_part /expb ;

**run**;

/\* Computing the multiple logistic regression model \*/

**proc** **logistic** data= project.group\_6\_train\_new1 descending;

class seasons holiday functioning\_day;

model Rented\_Bike\_Count = Hour temperature humidity wind\_speed visibility Dew\_point\_temperature Solar\_Radiation Rainfall Snowfall seasons holiday functioning\_day day\_part month\_part year\_part/ outroc= Roc1 ;

**run**;

/\* Computing the foward selection method for multiple logistic regression model \*/

**proc** **logistic** data= project.group\_6\_train\_new1 descending ;

class seasons holiday functioning\_day;

/\* Computing the Backward selection method for multiple logistic regression model \*/

**proc** **logistic** data= project.group\_6\_train\_new1 descending ;

class seasons holiday functioning\_day;

model Rented\_Bike\_Count = Hour temperature humidity wind\_speed visibility Dew\_point\_temperature Solar\_Radiation Rainfall Snowfall seasons holiday functioning\_day day\_part month\_part year\_part / EXPB SELECTION=BACKWARD SLSTAY=**0.05** RISKLIMITS outroc= Roc1 ;

**run**;

/\* Computing the stepwise selection method for multiple logistic regression model \*/

**proc** **logistic** data= project.group\_6\_train\_new1 descending ;

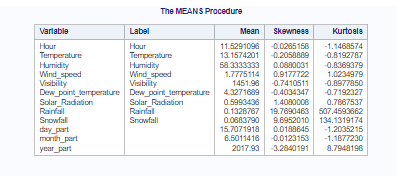
class seasons holiday functioning\_day;

model Rented\_Bike\_Count = Hour temperature humidity wind\_speed visibility Dew\_point\_temperature Solar\_Radiation Rainfall Snowfall seasons holiday functioning\_day day\_part month\_part year\_part / EXPB SELECTION=stepwise SLENTRY=**0.05** SLSTAY=**0.05** RISKLIMITS outroc= Roc1 ;

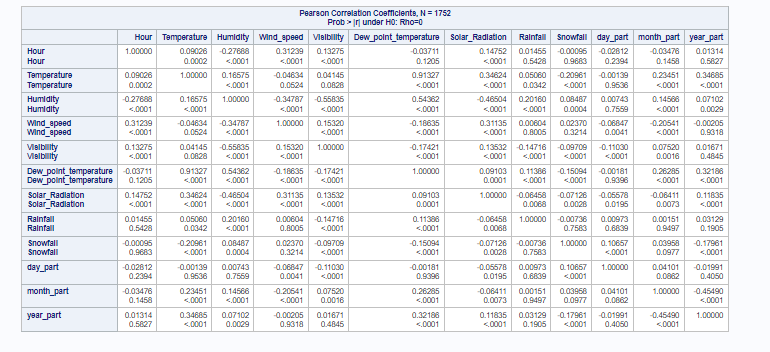
**run**;

# **Appendix**

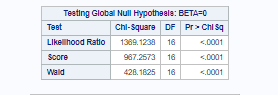
### Table 1 : The Descriptive statistics



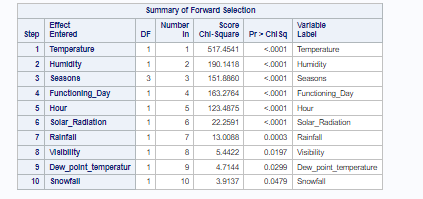
### Table 2: Matrix Correlation of continuous variables



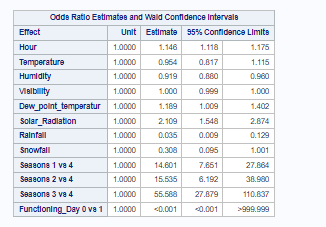
### Table 3: Likelihood ratio Test hypothesis.



### Table 4: Summary of the Forward selection method (Variables entered)



### Table 5: Odds ratio estimates of the forward selection method



### Table 6: Summary of the Backward selection method (Variables removed)

Table, calendar

Description automatically generated with medium confidence

### Table 7: Odds ratio estimates of the Backward selection method

Table

Description automatically generated

### Table 8: Summary of the Stepwise selection method (Variables entered)

Calendar

Description automatically generated

### Table 9: Odds ratio estimates of the stepwise selection method

Table

Description automatically generated

### Table 10: Analysis of Maximum Likelihood Estimates of Backward selection method

Table

Description automatically generated

### Figure 1 : The Roc curve of the Full Multiple logistic regression

Chart, line chart

Description automatically generated

### Figure 2: The Roc curve of the Forward Selection method

Chart, line chart

Description automatically generated

### Figure 3: The Roc curve of the Backward selection method

Chart, line chart

Description automatically generated

### Figure 4: Roc curve of the Stepwise selection method

Chart, line chart

Description automatically generated