University of the Western Cape

A Study to Investigate which Factors lead to a High Usage Day for Bike Rentals

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

GROUP 16

By

T Airdien, V Alweendo, R Hendricks

3870357, 3948288, 4030565

Supervisor:

Mr: Matthew Wayne Valentine

[SUBMISSION DATE]

Abstract

Bikes are a common form of transport in the city of Seoul. This form of transport is a faster way to get around the city without dealing with high traffic numbers in the city. However bike rental companies are struggling with the demand for their bikes when they are needed the most. These companies are faced with the problem of maximizing their efforts in making bikes available at the right times and if this problem is fixed, they can satisfy their existing and new customers while boosting their profits . The purpose of this study is to identify the factors that contribute to a high demand for rental bikes, in order to ensure that bikes are available and accessible to the public when they are needed, and to reduce waiting times. By finding the main factors, can indicate at which times the bike rentals should be making more bikes available to the public. We have explored the various factors to predict the target variable '*Rented bike count*'. The other problem is identifying if there is a high demand for bikes during warmer seasons such as summer and spring compared to the colder seasons. This problem arises from the fact the city of Seou has dry winters and humid summers, which would indicate that seasons should not be a factor since both winter and summer seasons have their disadvantages(Humidity and it being cold) for cycling. The multiple logistic regression model that was built demonstrated that the factors which led to a high usage day for bike rentals are seasons, temperature, hour of day(new hour) and humidity. The simple logistics regression built in this study will then show how warmer seasons have a higher demand for bikes compared to the colder seasons. On this basis, it is recommended that rental bike companies should focus on having a greater supply when the season is optimal, during peak hours and when the temperature is favorable to make sure they have a steady supply of rental bikes for the public.The recommendations would be that more advanced forms of machine learning should be used for predicting a high usage day count and due to the limitations in the knowledge of building regression models advanced selection techniques for the best possible model for this problem could not be used .It would be beneficial if the response variable was continuous instead of categorical.

UNIVERSITY of the WESTERN CAPE

Contents

Research Proposal 4

Research Questions 4

Timeline 4

Literature Review and Introduction 5

Bibliography 7

Methodology 9

Initial Analysis 10

Appendix 14

UNIVERSITY of the WESTERN CAPE

Research Proposal

Research Questions

1. Which factors are good predictors of a high usage day for bike rentals?
2. Is there a high demand for bikes during Warmer seasons like summer and spring compared to the cold seasons?

Timeline refer to *Appendix Table 1*

Literature Review and Introduction

Introduction

Bike sharing is a form of transportation introduced to urban cities in which bike rentals are made available to inhabitants for short or long distances. You can pick up and drop off these bikes at self-serving docking stations. (Davis, 2014).

Bike rentals are becoming increasingly popular as a means of transportation in cities around the world. As the demand for bike rentals grows, so does the need for a stable supply of rental bikes. The prediction of bike count required at each hour is crucial for ensuring a stable supply of rental bikes. The purpose of this study is to identify which factors are good predictors of a high usage day for bike rentals. The dataset used for this study contains weather information (temperature, humidity, wind speed, visibility, dew point, solar radiation, snowfall, rainfall), the number of bikes rented per hour, and date information. By analyzing this data, it is hoped that some insight can be gained into which factors are most likely to result in a high usage day for bike rentals.

Another purpose of this literature review is to investigate the potential relationship between Warm seasons and high usage of bike rentals. Type of season is a significant factor that can affect bike rental demand. The physical activity of people is greatly affected by their environment, including the weather. People tend to be more active during warmer seasons than colder seasons (Tucker and Gilland 2007). A possible reason people tend to be more active in warmer seasons is that the weather is more conducive to being outdoors. It is clear that the type of season could have a major influence on the amount of bike rented hence why it will be further discussed.

Throughout this review, predictor variables will be explored that would help anticipate which days would have a high rate of bike rentals. This will assist in providing a stable supply of bike rentals to cities, ensuring reliable transport and user satisfaction and also identifying if a logistic regression is best fit to predict the response variables.

Literature Review:

It can be difficult to identify which variables can be used to identify the best predictor variables since there are a lot of social economic characteristics, bicycle user habits and human behaviors to account for. (fear of losing bike) (Setyowati & Handayani, 2018). However, it makes more sense to believe that there is a high usage during the warmer seasons of the year compared to the cold seasons .Ogbe's findings indicate that the best predictors for high usage in bike rentals are weather, season and topography (Ogbe, 2022). Since people would rather do outdoor activities during these times, Ogbes's statement seem to have some validity behind it.Bobby Chandra's findings support part of this idea that seasons is the primary predictor since the investigation made in Washington D.C. the investigation revealed that the variables that affect bike rental count are temperature and season, summer and spring,to be precise(Chandra,2021).

Recent studies have also been consistent with the results from Bobby's paper. A paper from Zhifeng Wang has shown how to identify the best predictor variables by using machine learning. A supervised learning model was implemented with weather conditions as the predictor variable(Wang,2019). Although the main focus is on seasons, weather tends to indicate the same information as seasons .The findings produced a multiple linear regression model that indicated season, weather, feel temperature, humidity and wind speed to be the best predictor variables (Wang, 2019). The results from Wang are in line with the findings from bobby Chandra's findings which indicates that the temperature and seasons predictors are significant in identifying which is a high usage day for bike rentals.

Recent studies on building regression models for predicting high bike usage have been producing similar results of having predictor variables that have to do with the weather and temperature, this includes the study that came out Queensland University of Technology supports these ideas which shows that there is recurring theme around the topic of Modeling Bike Counts that weather is a good predictor for a high usage day of bike rentals (Ashqar, et al., 2019). The paper suggests that temperature, humidity and Time-of-the day are good predictors to be included in the regression model for predicting

high usage of rental bikes in the city (Ashqar, et al., 2019). From all the studies mentioned above a key takeaway is that they have different methods for creating their models, each set of authors wrote about a different model. Hence it is clear that a regression model alone cannot be used to build an accurate model for predicting bike count.

The Random Forest technique was used to rank the predictors in this study, which were then used to develop a regression model using a guided forward stepwise regression approach.(Ashqar, et al., 2019). Another method that was used was to create a linear regression model accompanied by ELM (Extreme learning machine) and Neural Network (Wang, 2019). The study continues to show how the ELM regression model has achieved the smallest errors in the test data and the best performance (Wang, 2019).

However, a report from Feng & Wang argues that even though the predictor variables are significant in finding a high usage day, the regression model produced will not be accurate and not suitable to be used in real life scenarios. (Feng & Wang, 2017) . One reason for this is that the model does not account for the fact that there are more high usage days during the summer than during the winter. This means that the model will tend to overestimate the number of high usage days during the winter and underestimate the number of high usage days during the summer. The paper continues to show this by creating a regression model which includes most of the variables used in Setyowati & Handayani's papers. The linear regression model used produced a R -squared value of 32.7% which strengthens their argument. R-Squared is a statistical measure that determines the proportion of variance in the dependent variable that can be explained by the independent variable. Feng and Wang suggest that a random forest model can help improve the multiple regression model and its accuracy (Feng & Wang, 2017). The findings from the study show that the accuracy of the model improved with 82% (Feng & Wang, 2017).

The type of season can have a significant impact on the number of bikes rented for a company. For example, warmer weather generally leads to increased bike rental demand, as people are more likely to want to enjoy the outdoors. In contrast, colder weather can lead to decreased demand, as people are less likely to want to ride a bike in the cold. In a recent study done in 2020 it was proven that this is in fact true, the author found through his analysis that bike rentals were the least in winter and gradually increased during spring to the peak in summer (Kim 2020). Additionally, other factors such as holidays and special events can also impact demand. For example, demand for bike rentals is likely to increase around summer holidays such as Memorial Day or Fourth of July, as people are more likely to travel and participate in outdoor activities, the analysis will reveal whether or not that is true..

The aim of this report is to identify the factors which are good predictors that lead to a high number of bike rentals as well as if the warmer seasons have a higher demand for bike rentals than the colder seasons. The literature above evaluates various predictors that contribute to the use of bike sharing. We can conclude that weather, season and temperature are the most prevalent factors that influence bike rentals from the studies discussed above. Despite the informative studies published, in order to further understand which factors are good predictors leading to a high rate of bike rentals, studies should be conducted in more cities throughout the world.

Methodology

This report looks at bike rentals, but more specifically which factors are good predictors that yield a high usage day for bike rentals. There are many factors each day which affect the amount of bike rentals however not all of them should be considered when predicting a high usage day. The question at hand is which factors are good predictors for high usage days so that bike rental companies can be prepared and have enough rentals at hand. A vast array of statistical methods will be looked at to determine those factors which include: include descriptive analysis, correlation analysis, simple logistic and multiple logistic regression models (full and reduced) with selection techniques, analysis of variance, hypothesis testing, r-squared along with coefficient of variance and receiver operating characteristic charts.

Descriptive statistics will be used as the procedures within them allow for evaluation of quantitative data in the dataset, defines means by groups, makes it easier to spot outliers within the data, produces a simple statistical summary report of the data, observe whether the data are approximately normally distributed in the form of a histogram and observe whether data distributions for variables differ by group. One of the procedures gives the ability to look at the skewness which measures symmetry as well as kurtosis which measures the shape of the distribution. Descriptive statistics can answer this question by providing information on the factors that are most associated with high usage days for bike rentals. This method will be used to see which factors are good predictors to high usage days by providing the average for the various predictors when a high usage day has occurred. It will also provide insight as to which season(s) have a high usage day either the warmer or colder seasons.

Correlation analysis will be conducted as it is useful for measuring the relationship between t quantitative variables measured on the same object. The correlation coefficient is a measure of the linear relationship between two quantitative variables measured on the same subject. The correlation *p* is a unitless quantity that ranges from a perfect negative and positive linear relationship, where p = -1 would indicate a strong negative correlation and p = +1 a strong positive correlation and when p = 0 no linear relationship The correlation coefficient *p* is estimated from the data using Pearson's correlation coefficient (Denoted by *r*). In order to understand the nature of the relationship between two variables, its good practice to examine a scatterplot of the

variables. We can take a look at our predictor variables vs our predicted variable (Pearson) is *rented bike count*. Lastly by looking at multicollinearity we can determine which factors are similar and do not add value to our model in which case we can just remove them from the model therefore improving the model which will allow us to better predict which of the factors are good predictors for a high usage day of bike rentals.

Multiple logistic regression will be used to create an equation to predict the probability of a high usage day for bike rentals, it will assess the relative importance of our factors such asWarm seasons or cold seasons, hour of the day, temperature , humidity . Simple logistic regression only has one predictor variable(warmcold) and our case a factor which will be used to answer the proposed research question Multiple logistic regression has more than one predictor variable (factor), from this we will be able to select the best set of predictors to create an effective prediction equation. The selection of the predictors for the multiple logistics regression model will be selected using various methods so that the most appropriate equation is created, these methods include a mix between manually selecting and automation that is *forward, backward and stepwise selection technique.* This process will help us determine which model is best fit for predicting a high usage day for bike rentals. A simple logistic regression can answer if warmer seasons see increase in a high usage day for bike rentals by. A multiple logistic regression can answer the main research question by looking at the factors that are best for predicting a high usage day Analysis of variance will be used to compare and test three or more variables. Allows for comparing of different groups and includes spreading out the variance. This process is essential to gain insight into how the groups vary. The method of analysis of variance can answer which factors are good predictors to a high usage bike rental day by testing for the significance of each factor in predicting a high usage day for bike rentals. Factors that are found to be significant predictors of a high usage day for bike rentals can be said to be good predictors of a high usage day for bike rentals.

Hypothesis testing within logistic regression will be used to see if there is sufficient statistical evidence to support a particular hypothesis regarding a parameter. This will prove to be useful to determine which factors play a role in a high usage day. The null hypothesis for factors affecting bike rentals would be that none of the factors are good predictors of a high usage day for bike rentals. The alternative hypothesis would be that at least one of the factors is a good predictor of a high usage day for bike rentals. To test this, the dataset provided will be used regarding bike rental usage and the potential predictor variables. Once the data is imported into SAS, analysis will take place to see if there is a relationship between the predictor variables and high bike rental usage. If there is a relationship, the next step is to determine if that relationship is statistically significant. If the relationship is statistically significant, the conclusion that the predictor variable is a good predictor of high bike rental usage can be made.

R-squared helps to understand the relationship between the movements of dependent variable and the movements of the independent variables as it calculates an estimate. The coefficient of variation depicts the degree of variability in data in a sample in comparison to the population mean. The R-Squared method can answer which factors are good predictors for a high bike rental day by looking at the correlation between the different factors and the number of bike rentals. A high R-Squared value indicates a strong correlation and means that the factor is a good predictor of high usage days for bike rentals. The coefficient of determination can identify which factors lead to a high usage of bike rentals by looking at how much variation in the bike rental usage can be explained by the variation in the factors. The higher the coefficient of determination, the better the factors are at predicting high usage days for bike rentals.

The ROC chart is a graphical representation of the performance of a binary classification system as its discrimination threshold is varied. The curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true-positive rate is also known as sensitivity, recall or probability of detection in machine learning. The false-positive rate is also known as the fall-out or probability of false alarm and can be calculated as (1 − specificity). Specificity is also known as the true negative rate. The method receiver operating characteristic can answer "Which factors are good predictors of a high usage day for bike rentals?" by looking at the factors that are most predictive of a high usage day and then ranking them in order of importance.

To conclude the methods being used each have their respective use to answer our main research question as well as our secondary research question. Combining the methods will construct a well-built argument to state which factors lead are good predictors of a high usage day for bike rentals.

Initial Analysis

Statistics gathered from a bike rental company for 2014 will be analyzed in this section. The main research question is: Which factors are good predictors of a high usage day for bike rentals? On the other hand the secondary research question is as follows:Is there a high demand for bikes during Warmer seasons like summer and spring compared to the cold seasons?? The intention is to point out which specific predictors play a role in rented bike count and to figure out if there is a high demand for bikes during Warmer seasons like summer and spring compared to the cold seasons. Both of these questions will be answered during the analysis. From the results of our analysis, knowledge will be gained which will improve the research field of study as more people will become aware of the predictors that lead to a high usage day of bike rentals and will make the companies in the industry more profitable. Taking the statistics gathered the following methods have been used to analyze the data: descriptive statistics, correlation analysis, simple and multiple logistic regression models (full and reduced) with selection

techniques , hypothesis testing analysis of variance, r-squared along with coefficient of variance and receiver operating characteristics.

The proposed research question:

Descriptive Statistics:

Firstly, the seasons are going to be separated into the warmer seasons and the colder seasons. Warmer seasons will include summer and spring while the colder seasons will include winter and autumn. However it's debatable in choosing if spring should be included in the warmer seasons and if autumn should be included in the colder seasons but itc common knowledge that spring is the transition from a colder temperature to warmer temperatures hence why it should be included in the warmer seasons category while autumn is a transition from the warmer temperatures to colder temperatures hence why it should be get to graze in the colder seasons

After categorizing the seasons into warm and cold seasons we are going to produce a frequency table which will indicate to us which of the two categories see a high usage of bike rentals. This part will be very important in answering the proposed question in this study.

The results from the frequency tables are as follows:

- The frequency distribution table and plot both indicate that 65% of the high usage days occur during warmer seasons while only 34.58% occurred during the colder seasons.*Appendix Tables 1.E and 1.F.*
- These findings agree with the findings from Bobby Chandra which support our findings in order to answer the proposed question.

The findings from the frequency tables indicated warmer seasons do have some sort of significance in determining the higher usage there; this can be further confirmed by the simple logistic regression.

Simple Logistic regression model

Simple regression was produced in order to study how it interacts with the response variable and to identify if the variable WarmCold can be used to predict a high usage of bike rentals.

The following simple regression function was formed*Appendix Tables 1.G* :

p(Estimate)=e^(-0.5254+1.1660X_1)/1+e^(-0.5254+1.1660X_1). ***Where X_1=WarmCold)***

In order to interpret this simple regression function we are going to make use of the odds ratio:

The odds ratio for WarmCold is 3.209. This indicates that the warmer seasons(summer and spring) are 3.209 times more likely to have high usage days than the colder seasons (winter and autumn).*Appendix Tables 1.G*

We can conclude that the Wald test statistic indicates a Chi_square value of 136.6830 with a P value of 0.0001 satisfies the significant level of 0.05 therefore we can reject the null hypothesis and conclude that the warm cold variable is predictive of the probability of a high usage day*Appendix Tables 1.G*

Since our findings are in line with our assumptions and the papers from Bobby's findings, this indicates that warmer seasons will see high usage days and this also indicates that the type of seasons are also good predictors of high usage days(Chandra,2021). Bobby's findings indicated that Summer and Spring are seasons with the highest demand for rental bikes(Chandra,2021)

The following will show the steps in building a multiple logistics regression model in order to identify the factors that are best for predicting a high usage day.

Descriptive Statistics:

The analysis done here will help give insight into the variables and average distribution of the data. From the literature above researchers have highlighted temperature, windspeed and rainfall to be influential in the number of bikes rented, however all the variables will be listed here. Since it is of interest which predictors will lead to a high usage day of bike rentals the analysis was run on the data where the Rented_Bike_Count had a value of 1 since it means a high usage day as had, by using SAS procedures, we obtain the following:

- The average hour of bike rentals is 14:00 with a standard deviation of 6.24
- The mean for temperature is 19.43 degrees celsius with a standard deviation of 8.89
- The mean for humidity is 54.67% with a standard deviation of 17.21
- The mean for windspeed is 1.84 miles/hour with a standard deviation of 0.95

- The mean for visibility in 10m range is 1541.82 with a standard deviation of 528.41
- The mean for dew point temperature is 9.28 degrees celsius with a standard deviation of 9.88
- The mean for solar radiation is 0.92 MJ/m^2 with a standard deviation of 1.02
- The average rainfall is 0.01 m with a standard deviation of 0.08
- The mean for snowfall is 0 mm with a standard deviation of 0.07
- The average day on which a high bike rental took place is 16th of the month with a standard deviation of 8.64
- The average month where high rental days took place was July (7) with a standard deviation of 2.58
- The average year with high rental days was 2018 with a standard deviation of 0.10
- The mean for seasons is 2 which implies that high rentals days took place in summer with a standard deviation of 0.83
- The average for functional day is 1 which means high rental days took place on a functional day with standard deviation of 0
- The mean for temperature is 13.16 degrees celsius with a standard deviation of 11.98

From the above it can be concluded that the data points tend to be close to their means as their standard deviations are not that large.It can be concluded that the above values are the days in which bike rental companies should expect the most rentals as they are the ideal conditions. This relates to our findings from before as the mean temperature is on the warmer side and our average for the season is summer which confirms that in the warmer season more bikes are rented.

Refer to Appendix 1.A

By using the Univariate procedure, we have the following results:

Normally distributed variables include hour, temperature, humidity, wind speed, visibility, dew point temperature, solar radiation, snowfall and rainfall as they each have a Shapiro-Wilk p value which is greater than 0.05.

Negatively skewed variables include hour, temperature, visibility and dew point temperature

Positively skewed variables include humidity, windspeed, solar radiation, rainfall and snowfall.

We can conclude that the following variables are good predictors for a high usage bike rental day hour, temperature, visibility and dew point temperature because

Refer to Appendix1.B

Correlation Analysis

Correlation analysis was done for the numerical variables from the data set. The numeric variables that have a normal distribution will undergo a correlation Pearson test while the variables that do not have a normal distribution will undergo a Spearman correlation test. This is done because Pearson correlation assumes normality while Spearman assumes non normality.
A scatterplot matrix will be produced to identify the relationships between all the independent variables. the correlation matrix has provided the following information:

Pearson Correlation Matrix

- It is evident that there is a strong linear relationship between dew point temperature and temperature Therefore dew point temperature variable should be removed as it explains the same information as temperature.
- There is a linear relationship between humidity, and do you point temperature which is another reason to remove Dew point temperature as available to be in the model
- All the other variables in the Pearson correlation matrix indicated there is no strong linear relationships among them so they can still stay in the model

Refer to Scatter plot Matrix 2 .A

Spearman Correlation Matrix

- This spearman correlation matrix shows that the variables visibility snowfall and rainfall don't have any relationship and therefore can be kept in the model

Variable selection

The months variable will be dropped from the data set because it would have to be grouped into quarters which will then give us the same information from the variable seasons. since seasons is a focus point for this study, the month variable will be dropped.

The functioning data variable will be dropped, and the holiday variable will be kept because when the holiday variable displays 0 it is also considered a functioning day, so they essentially indicated the same information.

A weekday variable is introduced, and this will indicate what day of the week is which would then lead to the day variable being dropped

Refer to Scatter plot Matrix 2 .B

The frequency tables indicate that season2 (summer) is the season where there is high usage of bikes with 38.22% of total count of bikes(High usage) followed by season3(Autumn) with 31.83%. These findings support the ideas from Bobby Chandra' findings when the paper indicated that season is one of the best indicators for predicting a Bike rental count whereby Chandra found that summer and spring have the highest demand for bikes(Chandra,2021) *.Refer to Appendix Table.A.*

The Hour variable was split into a Midnight(00:00-05:00),Morning(06:00-12:00), Afternoon(13:00-18:00)and Evening(19:00-23:00)session so it would be easier to tell which hours are most likely to have a high rented bike count.33.7% of the high rented bike count is seen in the afternoon session. The frequency table indicates that there is a high rented bike count in the Morning and afternoon sessions. 3 means midnight, 1 means afternoon,2 means evening and 0 means morning.*Refer to Appendix Table.B*

The Months of June,July and August are seen as the most popular months for bike renting as 13.22%,12.78% and 12.22% of the total high rented bike count respectively. These months are represented by 6,7 and 8 respectively. *Refer to Appendix Table.C*

Hypothesis Testing:

The hypothesis testing can aid in answering "Which factors are good predictors of a high usage day for bike rentals?" by helping to identify which factors are most likely to be associated with high usage days. This can be done by looking at the relationship between each factor and high usage days, and then testing to see if that relationship is statistically significant. If a factor is found to be a significant predictor of high usage days, then it is likely that it is a good predictor of high usage days. From our logistic analysis for rented bike count having a value of 1 which means high usage day we obtain the following results:

Predictors with a p-value < 0.05 therefore which should be included in the model are hour, humidity, visibility, dew point temperature, solar radiation, rainfall and snowfall. *Refer to appendix 2*

Therefore companies should consider the above factors to have enough bike for rentals to have a high usage day

It is important to note that there are certain variables that will not be added into the model due to different criterias. The variables that will not be added to the model Dew point temperature since there is multicollinearity between the variable temperature. The multicollinearity between these two variables are seen in the scatter plot matrix produced by sas. *Refer to Appendix Table D*

The other variable that will be removed from the the model is the Month variable due High pearson correlation coefficient 0f 77.54%.This result is seen in *Appendix Table E.*

Analysis of Variance:

Analysis of variance is used to determine whether there is a significant difference between the means of two or more groups. In a multiple logistic regression model, it can be used to determine whether the predictor variables are significantly associated with the outcome variable. From our analysis we obtain

SSR = 188.0321493

SSE = 249.3833758

SSTO = 437.4155251

The logistic regression model explains 188.0321493 of the variance in the dependent variable, leaving 249.3833758 unexplained. The model accounts

for 43.7% of the total variance. The results from the multiple logistic regression analysis can help to identify which factors are good predictors of a high usage day for bike rentals. In particular, the SSR indicates the amount of variation in the response variable that is explained by the predictors, while the SSE indicates the amount of variation in the response variable that is not explained by the predictors (Kim 2018). Based on the results, it appears that the predictor variables explain a significant amount of the variation in the response variable (bike rentals), and thus can be used to predict bike rental usage on high usage days.

ROC Chart:

The ROC chart for multiple logistic regression models is a graphical representation of the relative performance of each model. The models are represented by points on the chart, and the closer the points are to the top-left corner of the chart, the better the models are at predicting the outcomes. It is found that we have an area under the curve that is equal to 0.8961. The model does a good job of predicting the value of the response values, which is confirmed by the fact that the value is close to one.

Selecting the full multiple logistic regression equation

There are a total of three multiple logistic regression models however only one can be used to determine a high usage rented bike count. Sas has produced a full multiple logistic regression model which then will undergo 3 selection processes namely the forward backward and stepwise selection processes. After that, the best model will be selected based on the ROC curve, AIC and R-squared.

The first selection technique will be the forward selection technique. This technique enters variables that meet the significance level of 0.05. After the multiple logistics regression model has gone through the forward selection technique the following model was produced:

p(Estimates)=[1+2.8231+0.1081NewHour(0)+0.4903NewHour(1)+0.5402Season(1)+0.7813Season(2)+1.1670Season(3)+0.0957Temperature-0.0585Humdity-0.00032Visibility-0.1548Windspeed]^-1

Refer to Appendix Forward3.B

The second selection technique will be the backward selection technique. this technique enters all the variables in the model and then eliminates any model that does not meet the significance level of 0.01 the following model was produced by the backward selection technique

p(Estimate)= exp[1+1.681+0.1598NewHour(0)+0.4487NewHour(1)+0.8626(20+0.5547Season(1)+0.7589Season(2)+1.1543Season(3)+0.0925Temperature-0.0504Humdity]^-1

Refer to Appendix Backward3.C

The last selection technique is the stepwise selection technique. This model keeps variables that have P values which are greater than0.01 in variables that are less than 0.05 in the model. the following multiple logistic regression model was produced:

p(Estimate)=exp[1+1.681+0.1598NewHour(0)+0.4487NewHour(1)+0.8626(20+0.5547Season(1)+0.7589Season(2)+1.1543Season(3)+0.0925Temperature-0.0504Humdity]^-1

Refer to Appendix Backward.3c

The best model out of the three is the model produced by the backwards elimination technique. This particular model has a percent concordant of 91.3%. This indicates that this multiple logistic regression model has the predictive ability for the estimates of0.913 which is very good. The ROC curve for the selected model is 91.3 and it is on the top left corner of the graph which indicates that this is a good predictive accuracy of the multiple logistic regression model. The AIC is constant for all the models and R-squared is 0.4771.

This means that the predictor variables reduce the variability of the response variable by 47.71%.*Refer to Appendix Table 3A*

<u>The odds ratios</u>

The findings from the multiple logistic regression formula produced by the backward elimination selection method indicates that the variables Newhour season temperature and humidity I'm good predictive models for predicting a high usage day in bitcoins. The odds ratio helps to understand the effects these variables have on the response variable:

- The odds of having a high usage day are 1.097 times higher for every unit increase in the temperature
- The odds of having a high usage day are what is 0.951 times greater for every unit increase in humidity
- The odds of having a high usage of bike count during morning time is 5.109 times greater than if it were to be at midnight
- The odds of having a high usage of bike count during afternoon times is 6.820 times greater than if it were to be at midnight

The above points showed that there is significant evidence in concluding that time of day humidity and temperature are good predictors for a high usage day in bike rentals. The findings from the paper titled *Modeling bike counts in a bike-sharing system considering the effect of weather conditions also concludes that* Time of day, temperature,humidity are the best predictor variables (Ashqar, et al., 2019).

- Odds of having a high usage day during summer is 20.545 times greater than if it were to be in winter
- the odds of having a high usage day during spring is 25.199 times greater than if it were to be in winter

Conclusion

There is sufficient statistical evidence to conclude that seasons, temperature, humidity and Time of day are the best predictors for a high usage day in bike rentals in the city of Seoul although other academic papers would not agree there is also a good amount that have similar findings to this paper to build a multiple logistic regression to predict a high usage day in bike rentals. However, there are studies that indicate that multiple logistic regressions are not suitable for predicting a high usage bike count. Recommendations would be that more advanced forms of machine learning should be used for predicting a high usage day count and due to the limitations in the knowledge of advanced model selection techniques for the best possible model for this problem could not be used. It would also be beneficial if the response variable was continuous instead of categorical.

Appendix
Table 0: Timeline

ID	Task Description	Start Date	End Date	Duration	2022											
					Month											
					August				September				October			
					W1	W2	W3	W4	W1	W2	W3	W4	W1	W2	W3	W4
1	Research proposal	01/08/2022	15/08/2022	2 weeks	x	x	x									
2	Methodology	16/08/2022	02/09/2022	2 weeks			x	x	x							
3	Revised Research Proposal and Initial analysis	03/09/2022	23/09/2022	3 weeks					x	x	x					
4	Draft Report	26/09/2022	07/10/2022	2 weeks								x	x			
5	Final Report Submission	08/10/2022	17/10/2022	1 weeks										x	x	

1.A: Summary Statistics of all the variables in the data set

Summary statistics

The MEANS Procedure

Variable	Label	N	Mean	Std Dev
Hour	Hour	908	13.47	6.24
Temperature	Temperature	908	19.43	8.89
Humidity	Humidity	908	54.67	17.21
Wind_speed	Wind Speed	908	1.84	0.95
Visibility	Visibility	908	1541.82	528.41
Dew_point_temperature	Dew Point Temperature	908	9.28	9.88
Solar_Radiation	Solar Radiation	908	0.92	1.02
Rainfall	Rainfall	908	0.01	0.08
Snowfall	Snowfall	908	0.00	0.07
day_part	Day	908	15.92	8.64
month_part	Month	908	7.10	2.58
year_part	Year	908	2017.99	0.10
seasons1	Seasons	908	2.10	0.83
FunctioingnDay	Functioning Day	908	1.00	0.00
Holiday1	Holiday	0	.	.

1.B : Univariate Procedure for each quantitative variable

The UNIVARIATE Procedure
Variable: Hour (Hour)

Moments			
N	1752	Sum Weights	1752
Mean	11.5291096	Sum Observations	20199
Std Deviation	6.80202692	Variance	46.2675702
Skewness	-0.0265158	Kurtosis	-1.1468574
Uncorrected SS	313891	Corrected SS	81014.5154
Coeff Variation	58.9987186	Std Error Mean	0.16250671

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.956318	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.076694	Pr > D	<0.0100
Cramer-von Mises	W-Sq	2.451703	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	18.00123	Pr > A-Sq	<0.0050

The UNIVARIATE Procedure
Variable: Temperature (Temperature)

Moments			
N	1752	Sum Weights	1752
Mean	13.1574201	Sum Observations	23051.8
Std Deviation	11.9792616	Variance	143.502709
Skewness	-0.2058889	Kurtosis	-0.8192787
Uncorrected SS	554575.46	Corrected SS	251273.244
Coeff Variation	91.0456726	Std Error Mean	0.28619563

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.97984	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.071777	Pr > D	<0.0100
Cramer-von Mises	W-Sq	1.910774	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	10.95559	Pr > A-Sq	<0.0050

The UNIVARIATE Procedure
Variable: Humidity (Humidity)

Moments			
N	1752	Sum Weights	1752
Mean	58.3333333	Sum Observations	102200
Std Deviation	20.3466718	Variance	413.987055
Skewness	0.08800314	Kurtosis	-0.8369379
Uncorrected SS	6686558	Corrected SS	724891.333
Coeff Variation	34.8800089	Std Error Mean	0.48610079

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.980814	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.04716	Pr > D	<0.0100
Cramer-von Mises	W-Sq	1.098653	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	7.679424	Pr > A-Sq	<0.0050

The UNIVARIATE Procedure
Variable: Wind_speed (Wind_speed)

Moments			
N	1752	Sum Weights	1752
Mean	1.77751142	Sum Observations	3114.2
Std Deviation	1.04888343	Variance	1.10015645
Skewness	0.91777223	Kurtosis	1.02349791
Uncorrected SS	7461.9	Corrected SS	1926.37395
Coeff Variation	59.0085342	Std Error Mean	0.02505879

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.947156	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.096806	Pr > D	<0.0100
Cramer-von Mises	W-Sq	3.717882	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	22.49887	Pr > A-Sq	<0.0050

The UNIVARIATE Procedure
Variable: Visibility (Visibility)

Moments			
N	1752	Sum Weights	1752
Mean	1451.95548	Sum Observations	2543826
Std Deviation	603.724027	Variance	364482.7
Skewness	-0.7410511	Kurtosis	-0.897785
Uncorrected SS	4331731308	Corrected SS	638209209
Coeff Variation	41.5800646	Std Error Mean	14.4235249

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.829314	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.182936	Pr > D	<0.0100
Cramer-von Mises	W-Sq	18.80164	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	113.218	Pr > A-Sq	<0.0050

The UNIVARIATE Procedure
Variable: Dew_point_temperature (Dew_point_temperature)

Moments			
N	1752	Sum Weights	1752
Mean	4.32716895	Sum Observations	7581.2
Std Deviation	13.1453976	Variance	172.801477
Skewness	-0.4034347	Kurtosis	-0.7192327
Uncorrected SS	335380.52	Corrected SS	302575.387
Coeff Variation	303.787481	Std Error Mean	0.3140557

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.96428	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.053581	Pr > D	<0.0100
Cramer-von Mises	W-Sq	1.957848	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	14.94575	Pr > A-Sq	<0.0050

The UNIVARIATE Procedure
Variable: Solar_Radiation (Solar_Radiation)

Moments			
N	1752	Sum Weights	1752
Mean	0.59934361	Sum Observations	1050.05
Std Deviation	0.88158816	Variance	0.77719768
Skewness	1.40800076	Kurtosis	0.7867537
Uncorrected SS	1990.2139	Corrected SS	1360.87315
Coeff Variation	147.092277	Std Error Mean	0.02106196

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.722535	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.248819	Pr > D	<0.0100
Cramer-von Mises	W-Sq	37.82269	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	202.6751	Pr > A-Sq	<0.0050

The UNIVARIATE Procedure
Variable: Rainfall (Rainfall)

Moments			
N	1752	Sum Weights	1752
Mean	0.13287671	Sum Observations	232.8
Std Deviation	1.16459926	Variance	1.35629143
Skewness	19.7690463	Kurtosis	507.459366
Uncorrected SS	2405.8	Corrected SS	2374.8663
Coeff Variation	876.450989	Std Error Mean	0.02782335

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.08792	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.482063	Pr > D	<0.0100
Cramer-von Mises	W-Sq	127.9855	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	593.3371	Pr > A-Sq	<0.0050

The UNIVARIATE Procedure
Variable: Snowfall (Snowfall)

Moments			
N	1752	Sum Weights	1752
Mean	0.068379	Sum Observations	119.8
Std Deviation	0.42323183	Variance	0.17912518
Skewness	9.69520095	Kurtosis	134.131917
Uncorrected SS	321.84	Corrected SS	313.648196
Coeff Variation	618.950058	Std Error Mean	0.0101114

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.152954	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.517942	Pr > D	<0.0100
Cramer-von Mises	W-Sq	130.6715	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	604.2127	Pr > A-Sq	<0.0050

2: Logistic Procedure

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-7470.2	1166.3	41.0223	<.0001
Hour	1	0.1434	0.0129	123.4627	<.0001
Temperature	1	-0.0652	0.0783	0.6947	0.4046
Humidity	1	-0.0825	0.0219	14.1410	0.0002
Wind_speed	1	-0.1177	0.0866	1.8492	0.1739
Visibility	1	-0.00040	0.000163	6.0714	0.0137
Dew_point_temperatur	1	0.1698	0.0824	4.2466	0.0393
Solar_Radiation	1	0.8252	0.1586	27.0839	<.0001
Rainfall	1	-3.3917	0.6664	25.9049	<.0001
Snowfall	1	-1.3602	0.6182	4.8422	0.0278
day_part	1	-0.00069	0.00914	0.0056	0.9402
month_part	1	0.4056	0.0360	126.6473	<.0001
year_part	1	3.6933	0.4588	64.8023	<.0001
seasons1	1	-0.6543	0.0982	44.4034	<.0001
FunctioingnDay	1	20.3599	709.3	0.0008	0.9771

The FREQ Procedure

Rented_Bike_Count=1

Seasons1	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	247	27.20	247	27.20
2	347	38.22	594	65.42
3	289	31.83	883	97.25
4	25	2.75	908	100.00

Table.A

The FREQ Procedure

Rented_Bike_Count=1

NewHour	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	291	32.05	291	32.05
1	306	33.70	597	65.75
2	222	24.45	819	90.20
3	89	9.80	908	100.00

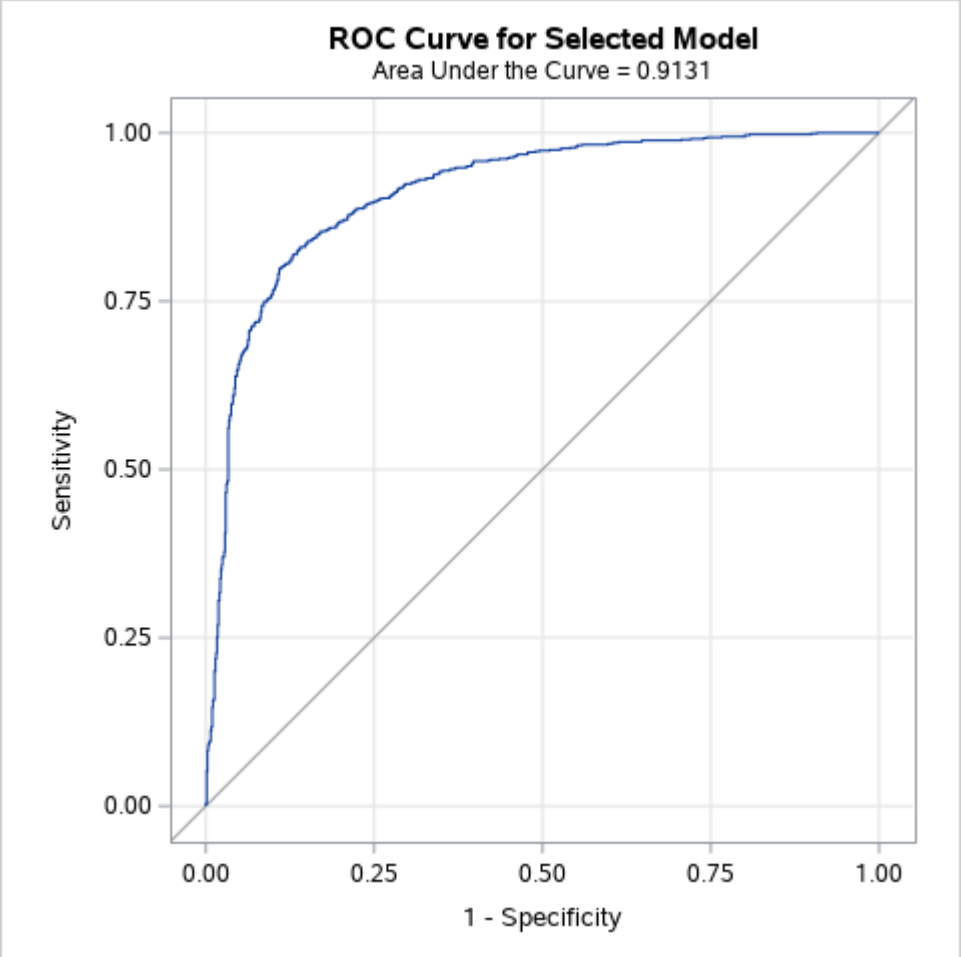
TableB

MONTH	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	6	0.66	6	0.66
2	10	1.10	16	1.76
3	73	8.04	89	9.80
4	88	9.69	177	19.49
5	86	9.47	263	28.96
6	120	13.22	383	42.18
7	116	12.78	499	54.96
8	111	12.22	610	67.18

9	93	10.24	703	77.42
10	100	11.01	803	88.44
11	96	10.57	899	99.01
12	9	0.99	908	100.00

Table C

TABLE E

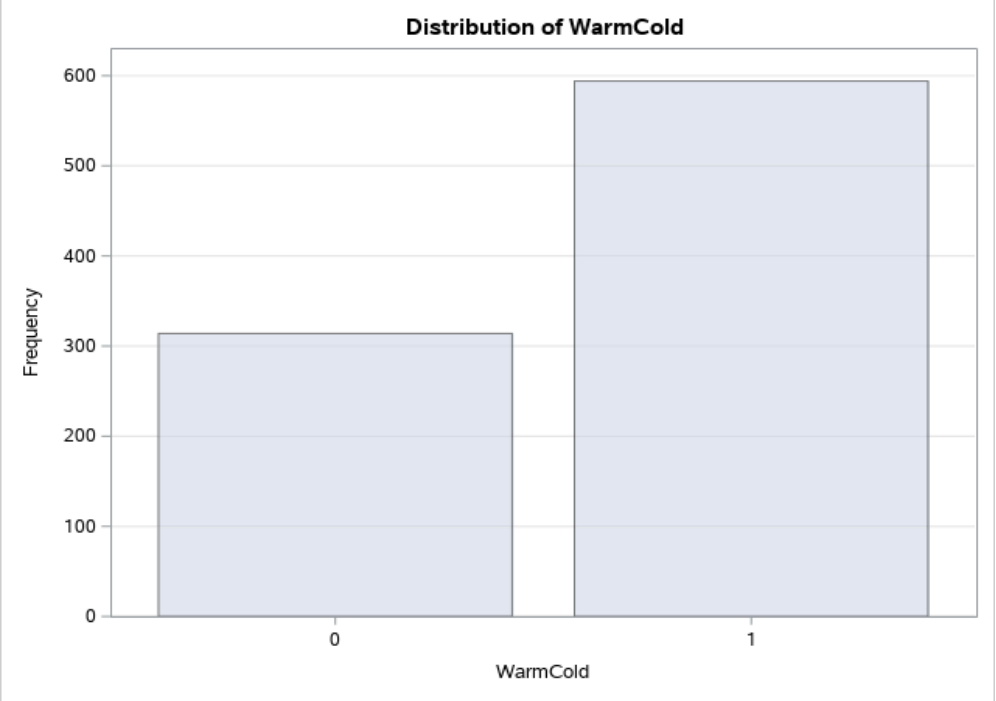


ROC for backwards selection

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	90.7	Somers' D	0.813
Percent Discordant	9.3	Gamma	0.813
Percent Tied	0.0	Tau-a	0.406
Pairs	766352	c	0.907

Warm Cold	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	314	34.58	314	34.58
1	594	65.42	908	100.00

1.E



1.F

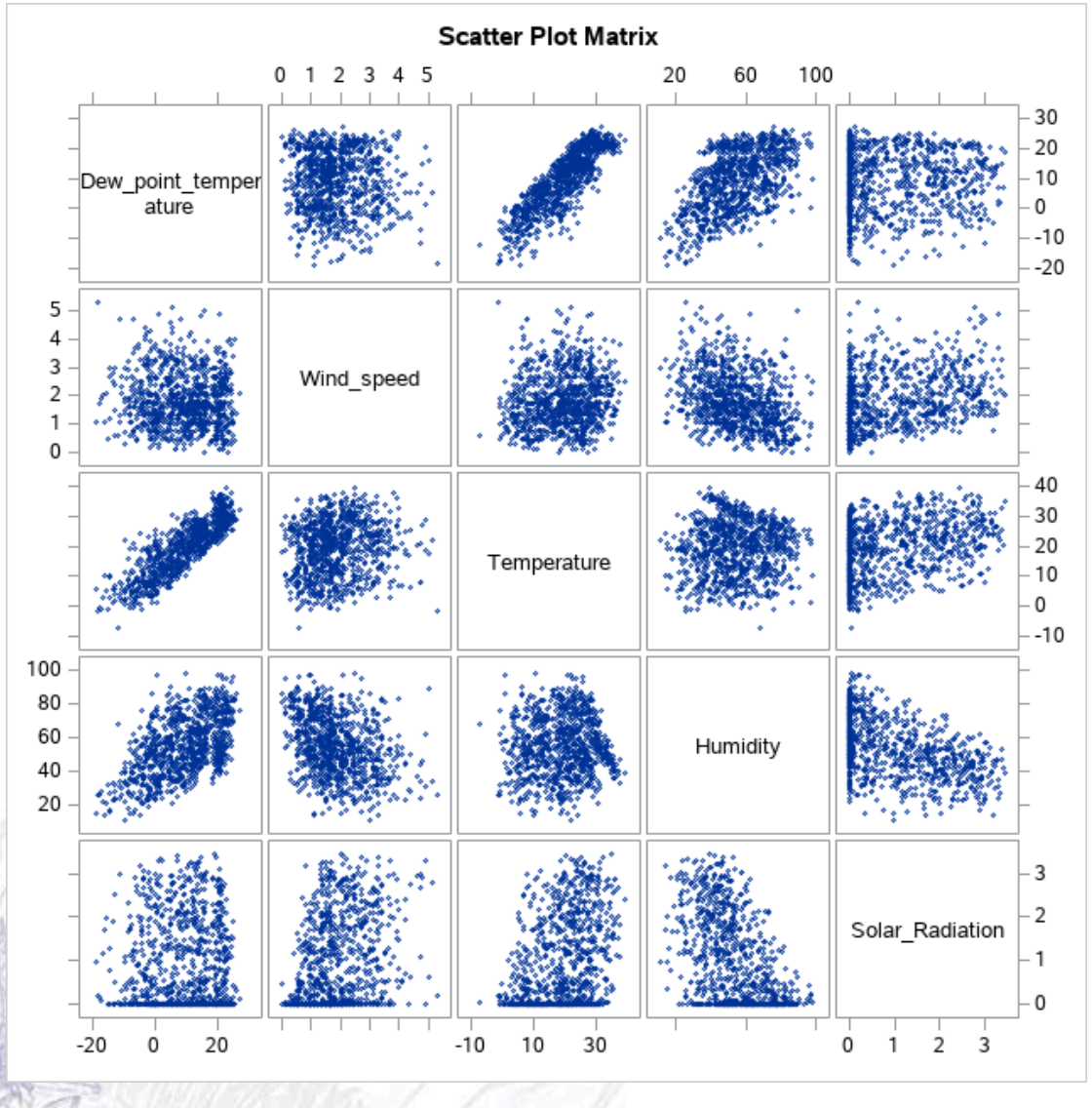
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	142.5239	1	<.0001
Score	140.6337	1	<.0001
Wald	136.6830	1	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.5254	0.0712	54.4594	<.0001
WarmCold	1	1.1660	0.0997	136.6830	<.0001

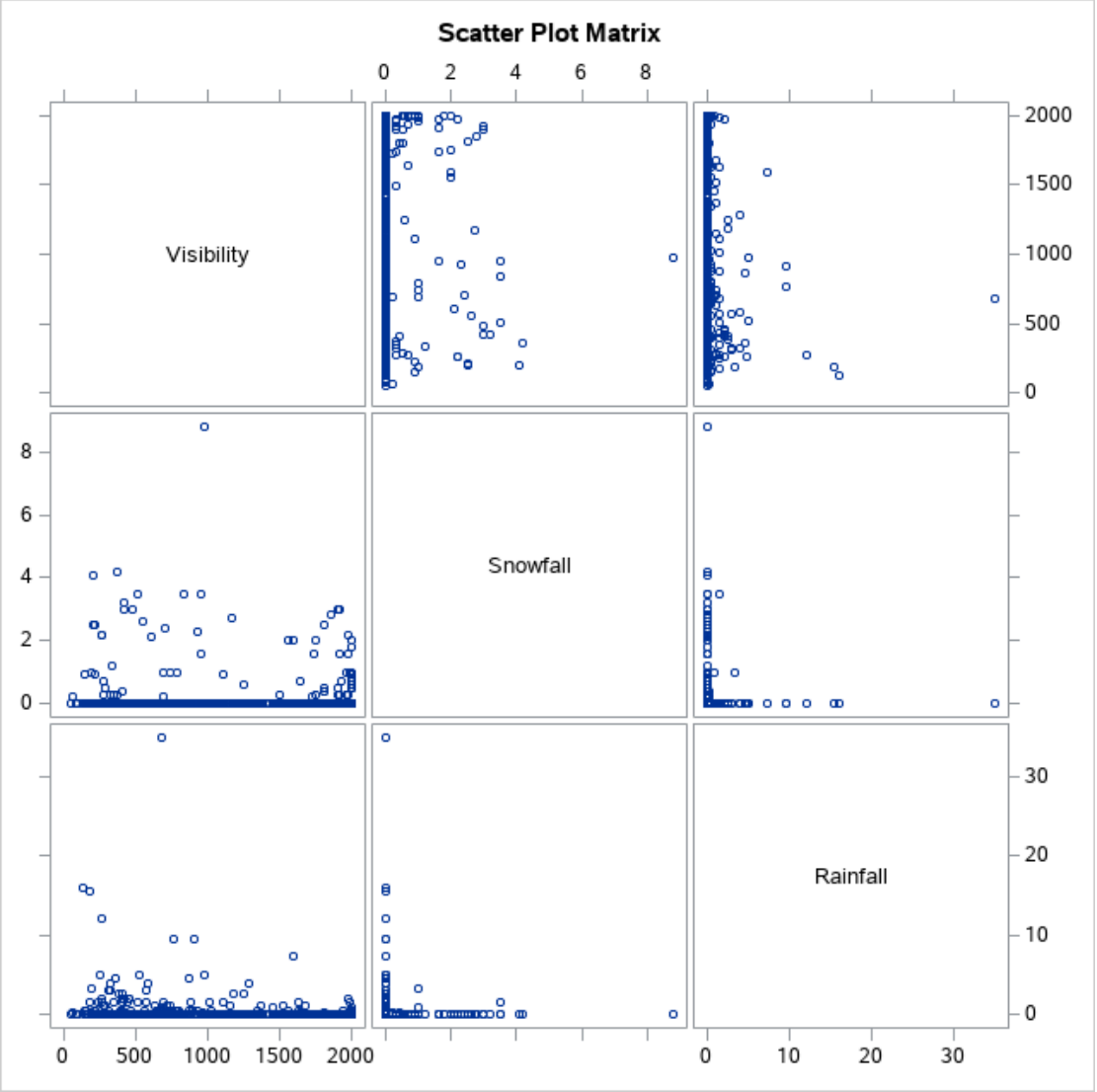
Association of Predicted Probabilities and Observed Responses			
Percent Concordant	41.2	Somers' D	0.283
Percent Discordant	12.8	Gamma	0.525
Percent Tied	46.0	Tau-a	0.142
Pairs	766352	c	0.642

Odds Ratio Estimates and Wald Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
WarmCold	1.0000	3.209	2.639	3.902

1.G



Scatter Plot Matrix 2.A



Scatter Plot Matrix 2.B

R-Square	Coeff Var	Root MSE	Rented_Bike_Count Mean
0.477125	69.89493	0.362241	0.518265

Table 3.A

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
Intercept		1	2.8231	0.5058	31.1496	<.0001	16.829
NewHour	0	1	0.1081	0.1143	0.8943	0.3443	1.114
NewHour	1	1	0.4903	0.1477	11.0237	0.0009	1.633
NewHour	2	1	0.9013	0.1385	42.3696	<.0001	2.463
Seasons1	1	1	0.5402	0.1234	19.1597	<.0001	1.716
Seasons1	2	1	0.7813	0.2102	13.8115	0.0002	2.184
Seasons1	3	1	1.1670	0.1272	84.1910	<.0001	3.212
Temperature		1	0.0957	0.0129	55.0413	<.0001	1.100
Humidity		1	-0.0585	0.00521	126.0809	<.0001	0.943
Visibility		1	-0.00032	0.000143	4.9956	0.0254	1.000
Wind_speed		1	-0.1548	0.0770	4.0412	0.0444	0.857

Association of Predicted Probabilities and Observed Responses

Percent Concordant	91.4	Somers' D	0.828
Percent Discordant	8.6	Gamma	0.828
Percent Tied	0.0	Tau-a	0.414
Pairs	766352	c	0.914

Odds Ratio Estimates and Wald Confidence Intervals

Effect	Unit	Estimate	95% Confidence Limits	
NewHour 0 vs 3	1.0000	4.991	3.479	7.161
NewHour 1 vs 3	1.0000	7.315	4.599	11.636
NewHour 2 vs 3	1.0000	11.033	7.127	17.079
Seasons1 1 vs 4	1.0000	20.669	10.963	38.970
Seasons1 2 vs 4	1.0000	26.306	10.995	62.934
Seasons1 3 vs 4	1.0000	38.685	20.026	74.732
Temperature	1.0000	1.100	1.073	1.129
Humidity	1.0000	0.943	0.934	0.953
Visibility	1.0000	1.000	0.999	1.000

Forward.3B

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	1.6381	0.2974	30.3327	<.0001
NewHour	0	1	0.1598	0.1122	2.0299	0.1542
NewHour	1	1	0.4487	0.1398	10.3015	0.0013
NewHour	2	1	0.8626	0.1373	39.4488	<.0001
Seasons1	1	1	0.5547	0.1196	21.4968	<.0001
Seasons1	2	1	0.7589	0.2106	12.9878	0.0003
Seasons1	3	1	1.1543	0.1230	88.0690	<.0001
Temperature		1	0.0925	0.0128	52.4344	<.0001
Humidity		1	-0.0504	0.00419	144.5535	<.0001

Association of Predicted Probabilities and Observed Responses

Percent Concordant	91.3	Somers' D	0.826
Percent Discordant	8.7	Gamma	0.826
Percent Tied	0.0	Tau-a	0.413
Pairs	766352	c	0.913

Odds Ratio Estimates and Wald Confidence Intervals

Effect	Unit	Estimate	95% Confidence Limits	
NewHour 0 vs 3	1.0000	5.109	3.576	7.299
NewHour 1 vs 3	1.0000	6.820	4.425	10.511
NewHour 2 vs 3	1.0000	10.317	6.735	15.803
Seasons1 1 vs 4	1.0000	20.545	10.867	38.842
Seasons1 2 vs 4	1.0000	25.199	10.490	60.534
Seasons1 3 vs 4	1.0000	37.422	19.435	72.057
Temperature	1.0000	1.097	1.070	1.125
Humidity	1.0000	0.951	0.943	0.959

Backward 3.C

SAS Code Used (With reference to appendix)

1.A

```
*Descriptive Statistics*;

PROC MEANS DATA = bike_1 MAXDEC=2 N MEAN STDDEV;

VAR Hour Temperature Humidity Wind_speed Visibility
Dew_point_temperature Solar_Radiation Rainfall Snowfall        day_part
month_part  year_part seasons1 FunctioingnDay Holiday1;

Label Wind_speed = 'Wind Speed'
```

```
                Dew_point_temperature = 'Dew Point Temperature'
                Solar_Radiation = 'Solar Radiation'
                day_part = 'Day'
                month_part = 'Month'
                year_part = 'Year'
                seasons1 = 'Seasons'
                FunctioingnDay = 'Functioning Day'
                Holiday1 = 'Holiday';
TITLE 'Summary statistics';
RUN;
```

1.B

```
ODS HTML;
PROC UNIVARIATE NORMAL PLOT DATA=bike2;
HISTOGRAM /NORMAL ;
TITLE 'PROC UNIVARIATE PROCEDURE';
RUN;
ODS HTML CLOSE;
```

2.

```
ODS HTML;
PROC LOGISTIC DATA=bike2 DESCENDING;
CLASS Holiday;
MODEL Rented_Bike_Count =  Hour Temperature Humidity Wind_speed
Visibility Dew_point_temperature Solar_Radiation Rainfall Snowfall
day_part month_part       year_part seasons1 FunctioingnDay;
RUN;
ODS HTML CLOSE;
```

```
PROC MEANS DATA =
"/home/u58258661/My_Folder/Project/group_16_train .sas7bdat"
MAXDEC=2 N MEAN STDDEV;
```

```
VAR Hour Temperature Humidity Wind_speed Visibility
Dew_point_temperature Solar_Radiation
        Rainfall Snowfall;
Label Wind_speed = "Wind Speed"
        Dew_point_temperature = "Dew Point Temperature"
        Solar_Radiation = "Solar Radiation";
TITLE 'Summary statistics';
ODS HTML;
PROC UNIVARIATE NORMAL PLOT
DATA="/home/u58258661/My_Folder/Project/group_16_train
(1).sas7bdat";
HISTOGRAM /NORMAL ;
TITLE 'PROC UNIVARIATE PROCEDURE';
RUN;
ODS HTML CLOSE;
ODS HTML;
PROC CORR
DATA="/home/u58258661/My_Folder/Project/group_16_train
(1).sas7bdat" PLOTS(MAXPOINTS=10000) = MATRIX SPEARMAN
PEARSON NOSIMPLE;
    VAR Temperature Humidity Wind_speed Visibility
Dew_point_temperature Solar_Radiation
        Rainfall Snowfall;
    WITH Rented_Bike_Count;
RUN;
ODS HTML CLOSE;
ODS HTML;
PROC GLM data = "/home/u58258661/My_Folder/Project/group_16_train
(1).sas7bdat";
  CLASS Date Seasons Holiday Functioning_Day;
   MODEL Rented_Bike_Count =  Hour Temperature Humidity
Wind_speed Visibility Dew_point_temperature Solar_Radiation
```

```
Rainfall Snowfall;
TITLE 'Analysis of Covariance Example';
RUN;
proc logistic data= "/home/u58258661/My_Folder/Project/group_16_train
(1).sas7bdat" descending plots(only)=roc;
 CLASS Date Seasons Holiday Functioning_Day;
 MODEL Rented_Bike_Count =  Hour Temperature Humidity Wind_speed
Visibility Dew_point_temperature Solar_Radiation
Rainfall Snowfall;
RUN;
LIBNAME Bike "/home/u58258106/Bike";
DATA Bike;
  set Bike.group_16_train;
RUN;
DATA BIKE1;
  SET BIKE;
  Seasons1=input(Seasons,7.);
DROP Seasons;
RUN;
PROC SORT DATA = BIKE1;
  BY Seasons1;
RUN;
PROC PRINT DATA= BIKE1;
RUN;
PROC CORR DATA = BIKE1 PEARSON NOSIMPLE;
  BY Seasons1 ;
  VAR Hour Temperature Rainfall Dew_point_temperature Wind_speed
Seasons1 Visibility Solar_Radiation Snowfall ;
  WITH Rented_Bike_Count;
RUN;
```

```
PROC CONTENTS DATA = BIKE1;
RUN;
PROC LOGISTIC DATA = BIKE1 DESCENDING ;
  CLASS Seasons1 ;
  MODEL Rented_Bike_Count(EVENT='1') = Hour Temperature Rainfall
Dew_point_temperature Wind_speed Seasons1 Visibility
Solar_Radiation Snowfall ;
RUN;
*ln(P/(1-P)= -2.9743+0.1240Hour 0.1788Temperature
-3.2964Rainfall-0.1124Dew_point_temperature-0.1288Wind_speed
+0.3841Season1(1)+0.8925Season1(2)+1.1287Seasons1(3)
-0.00020Visibility +0.5420Solar_Radiation -1.0809Snowfall*;
PROC LOGISTIC DATA = BIKE1 DESCENDING;
  MODEL Rented_Bike_Count(EVENT='1') = Hour Temperature Rainfall
Dew_point_temperature Wind_speed Seasons1 Visibility
Solar_Radiation Snowfall
  /EXPB SELECTION=FORWARD  SLENTRY=0.05 RISKLIMITS;
     TITLE 'LOGISTIC ON BIKE RENTALS';
RUN;
*ln(P/(1-P)=-2.2405+0.1174Hour +0.1796Temperature -3.5859Rainfall
-0.0604Dew_point_temperature -0.2543Wind_speed -0.3023Seasons1
+0.5375 Solar_Radiation -1.0332          Snowfall*;
PROC LOGISTIC DATA = BIKE1 DESCENDING;
  MODEL Rented_Bike_Count(EVENT='1') = Hour Temperature Rainfall
Dew_point_temperature Wind_speed Seasons1 Visibility
Solar_Radiation Snowfall
  / SELECTION=BACKWARD  SLSTAY=0.05 RISKLIMITS;
     TITLE 'LOGISTIC ON BIKE RENTALS';
RUN;
*ln(P/(1-P)= -2.2793+0.1166Hour +0.1889Temperature -3.5514Rainfall
-0.0668Dew_point_temperature +0.2586Wind_speed -0.3228Seasons1+
0.5189Solar_Radiation  *;
proc print data=bike1;
run;
```

Bibliography

Ashqar, H., Elhenawy, M. & Rakha, H., 2019. *Modeling bike counts in a bike-sharing system considering the effect of weather conditions,* Washington: Elsevier.

Bewick, V., Cheek, L. & Ball, J., 2004. Statistics review 13: Receiver operating characteristic curves. *Critical Care,* 4 November, Vol 8(No.6), pp. 508-512.

Christie, P., Georges, J., Thompson, J. & Wells, C., 2011. *Applied Analytics Using SAS Enterprise Miner.* North Carolina, USA: SAS Institute Inc.

Da Silva, I. et al., 2017. Multilayer Perceptron Networks. *Artificial Neural Networks,* Volume Vol 34, pp. 55-115.

Diez, M. D., Barr, D. C. & Cetinkaya-Rundel, M., 2015. *OpenIntro Statistics.* Chicago: CreateSpace.

Du, J., He, R. & Zhechev, Z., 2014. Forecasting Bike Rental Demand.

Fawcett, T., 2006. An introduction to ROC analysis. *Pattern Recognition Letters,* 19 December, Volume 27, pp. 861-874.

Feng, Y. & Wang, S., 2017. A Forcast for Bicycle Rental Demand Based on Random Forests and Multiple Linear Regression.

Garcia Neito, G., Alba, E., Jourdan, L. & Talbi, E., 2009. Sensitivity and specificity based multiobjective approach for feature selection: Application to cancer diagnosis. *Information Processing Letters,* 15 April, Volume 109, pp. 887 - 896.

Guyon, I. & Elisseeff, A., 2003. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research,* Volume 3, pp. 1157 - 1182.

Ogbe, M., 2022. *Predicting Bikeshare count in Washington DC,* Washington DC: RPubs.

Panchal, G., Ganatra, A., Kosta, Y. P. & Panchal, D., 2011. Behaviour Analysis of Multilayer Perceptrons with Multiple Hidden Neurons and Hidden Layers. *International Journal of Computer Theory and Engineering,* 2 April, Volume Vol3, pp. 332-337.

Setyowati, E. & Handayani, D., 2018. *Analysis of influencing factors on using rental bikes at shopping tourism sites in Surakarta,* Surakarta: MATEC Web of Conferences.

Team, C., 2022. *Corporate Finance Institute.* [Online]
Available at: https://corporatefinanceinstitute.com/resources/knowledge/other/r-squared/#:~:text=R%2DSquared%20(R%C2%B2%20or%20the,(the%20goodness%20of%20fit).

Wang, Z., 2019. Regression Model for Bike Sharing Service by Using Machine Learning. *Asian Journal of Social Science Studies,* 06 November.

Zhang, Q., Gupta, K. & Devabhaktuni, 2003. IEEE TRANSACTIONS ON MICROWAVE THEORY AND TECHNIQUES. *Artificial Neural Networks for RF and Microwave Design—From Theory to Practice,* pp. 1339 - 1350.

Herbert, K. (2021, August 4). Bike Share is a Boon for Public Health. Retrieved from: https://betterbikeshare.org/2021/08/04/bike-share-is-a-boon-for-public-health/

Davis, L. S. (2014). Rolling along the last mile: Bike-sharing programs blossom nationwide. Planning, 80(5), 10–16.

Elliot Fishman (2016) Bikeshare: A Review of Recent Literature, Transport Reviews, 36:1, 92-113, DOI: 10.1080/01441647.2015.1033036

Woodcock, J., Tainio, M., Cheshire, J., O'Brien, O., & Goodman, A. (2014). Health effects of the London bicycle sharing system: Health impact modelling study. BMJ, 348. doi:10.1136/bmj.g425

Fishman, E. (2014). Bikeshare: Barriers, facilitators and impacts on car use (PhD thesis by publication). Queensland University of Technology, Brisbane.

BobbyChandra(2021).https://bobby-js-chandra.medium.com/how-do-weather-and-season-affect-bike-rentals-d26164e3233b#:~:text=Higher%20temperatures%20and%20clear%2Fdry,the%20rest%20of%20the%20year. How do weather and season affect bike rentals?

Tucker P, Gilliland J. 2007. The effect of season and weather on physical activity: a systematic review. Public Health121(12):909–922, PMID: 17920646, 10.1016/j.puhe.2007.04.009. Crossref, Medline, Google Scholar

Kim, H. (2020). Seasonal Impacts of Particulate Matter Levels on Bike Sharing in Seoul, South Korea. *International Journal of Environmental Research and Public Health*, *17*(11). https://doi.org/10.3390/ijerph17113999