University of the Western Cape

Factors that influence a high usage day for bike rentals

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

Group 7

By
Vanick J. Semegni
Sibablo A. Joja
Mbali Kalako

4165406
4028996
4027600

Supervisor:
Mr: Matthew Wayne Valentine

6 October 2023

Table of Contents

Abstract

This research essay examines the core challenge of predicting whether there is a high or low usage of rented bikes to ensure a stable and reliable supply in urban bike-sharing systems. Our main goal is to determine which factors are good predictors of a high usage day for bike rentals. As well as determining whether an increase in temperature increases the overall high usage bike rentals. According to earlier research published in the literature, biking is generally impacted by the days weather. The interaction and patterns of bike usage showed that as temperature changes during a day, the usage of bikes changes. Descriptive analysis is performed on the dataset to understand its characteristics and structure. A simple logistic regression model is used to analyse the relationship between the rented bike counts and temperature. A multiple logistic regression model is used to analyse the relationship between rented bike count and all the various predictors. Various selection techniques will be used to determine which predictors do not contribute significantly and be ignored from the model.

Main research question:

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

Additional research question:

- Does an increase in temperature increase the overall high usage bike rentals in Seoul.

Key findings:

The simple logistic regression model between the predictor variable Temperature and the response Rented Bike Count, is given by the following estimated logistic regression equation: $\log\left(\frac{p}{1-p}\right) = -1.4801 + 0.1181X_1$. It shows temperature emerges as a significant determinant of a high usage day for Rented Bike Count. However, when considered alongside other predictor variables in multiple logistic regression models, its significance decreases. Nonetheless, the Simple Logistic Regression model, with an AUC of 0.8108, effectively predicts high and low usage days, indicating that the temperature indeed influences high usage bike rentals.

Across all the multiple logistics regression models created, the Seasons variable consistently stands out. The Autum Season in particular, results in the most significant influence, making it a good robust predictor for high usage day. Its impact is greater than the other seasons, underlining its importance in predicting a high usage day for rented bike count.

The full multiple logistic regression model, with the highest AUC of 0.9430 and a R-squared value of 0.551237 is given by the following estimated logistic regression model:

$$\log\left(\frac{p}{1-p}\right) = -6.7556 + 0.1400X_1 - 0.0430X_2 - 0.0839X_3 - 0.1098X_4 - 0.0004X_5 + 0.01689X_6 + 0.7804X_7 - 3.2556X_8 - 1.1980X_9 + 0.3325X_{10} + 0.3628X_{11} + 1.6225X_{12} + 0.2114X_{13} - 10.3888X_{14}$$

It reveals which factors are good predictors of a high usage day for rented bike count. Specifically, an increase in Hour (as the day goes by), Dew Point Temperature and Solar Radiation, coupled with reduced Humidity, Rainfall and Snowfall during the Autumn Season are seen as good factors in predicting a high usage day for bike rentals.

By understanding these various influences on bike usage, businesses can strategically manage their resources ensuring that there are available bikes for high usage days and reducing waiting times.

Literature review

Introduction

In recent years, urban mobility patterns have seen a significant transformation, with the rise of sustainable transportation modes becoming a focal point. Among these sustainable transportation modes, urban Bike Sharing Systems have gained prominence in cities worldwide, serving various purposes from reducing traffic congestion to promoting environmental sustainability (Kim, Shin, Im & Park, 2011). This study delves into identifying which factors are good predictors of high usage day for bike rentals. This literature review discusses the aspects of bike-sharing usage patterns focusing on the impact of temperature, seasons variation, weather-related risk, and user perceptions on bike rentals. The rise in urban bike sharing systems underlines the importance of a thorough understanding of factors that influence its usage dynamics. findings are crucial for optimizing bike availability, minimizing waiting times, and enhancing user experiences. While existing research has explored aspects of weathers influence on outdoor activities, a detailed analysis of its impact on bike-sharing usage, especially in varying seasons and weather conditions, remain limited (Kim D. , 2011). The results obtained in this result will provide practical implications for urban panners and bike-sharing businesses looking to improve their user experience and encourage sustainable transportation choices. As cities around the world continue to embrace brace bike sharing systems as an important aspect of their transportation network, understanding the understanding the influence of weather, seasons, risks exposure and user satisfaction becomes important.

Impact of Temperature on Bike Rentals

Temperature significantly influences bike rental patterns. Research indicates a direct correlation between temperature and outdoor activities, with rentals increasing as temperatures rise, up to 30 degrees Celsius (Kim K. , 2018) .However, extreme weather conditions such such as temperature over 30 degrees Celsius deter users (Kim K. , 2018), emphasizing the importance of mild, comfortable temperatures for optimal bike rental usage. Temperature nuances attract diverse user groups, indicating the necessity of accommodating various preferences (Langley, 2022). Economic repercussions arise from reduced rentals during adverse weather, underlining the need for strategic management to maximize profitability.

Impact of Seasons on Bike Rentals

Seasonal variations profoundly affect bike rentals. Spring and autumn witness heightened bike usage due to pleasant weather, contrasting with reduced rentals in winter due to cold weather challenges (Kim H. , 2020). Tourists amplify summer rentals, while locals dominate autumnal biking activities. Peak seasons bolster the local economy, aligning with sustainable mobility goals by curbing emissions and enhancing air quality. Seasonal fluctuations necessitate adaptive strategies, incorporating technology-driven solutions and infrastructural adjustments to sustain user engagement and address storage requirements.

Interplay of Weather and cyclist risk exposure

In road safety models, traffic volume is the primary independent variable of risk exposure. The most susceptible road users in terms of weather are cyclists, who are more exposed than motorists. As a result, the weather has a significant impact on their choice to cycle. It suggested that any chance in the weather may have a significant effect on bicycle use. Results indicated that air temperature, humidity, cloud cover and wind strength have a substantial effect on bicycle use (Quach, 2020).

Bike users perception and satisfaction

Cycling has proven to be an important strategy in decreasing the risk of non-communicable diseases. (Sharma, Nam, Yan, & Kim, 2019) studied to discover barriers and enabling factors influencing satisfaction and safety perception towards the use of bicycle roads in Korea. The results showed that having enough bicycle parking spaces, a moderate slope, and enough bicycle signs were all important enabling variables for people to be happy using bicycle roads. Based on thee findings, healthy cities should promote cycling behaviour encouraging enabling factors and initiating attempts to improve the factors that act as barriers through urban planning.

Conclusion

Understanding the interplay between temperature, seasons, and user preferences is pivotal for refining bike rental systems. This research underscores the significance of weather-related dynamics in program enhancement. Providers can optimize their services by aligning bike supply with demand fluctuations induced by temperature and seasonal changes. Technological solutions, real-time weather updates, and indoor storage facilities offer strategies to sustain user engagement throughout the year. The study illuminates the intricate relationship between user behaviour, weather patterns, economic impacts, and seasonal dynamics within bike rental systems. Adapting strategies based on these insights enhances user experiences, boosts peak-season revenues, and advances sustainable urban mobility initiatives.

Methodology

In this section, the research methodology used in this study is outlined in great details, presenting a systematic approach to investigating the influence of various factors on the usage patterns of rented bikes. The methodology comprises of various analysis, ranging from descriptive analysis to simple and multiple logistic regression modelling strategies. As well as model evaluation techniques such as Area Under the ROC Curve to evaluate the accuracy of the logistic regression models. Each methodology used is aimed at unravelling the factors which are good predictors of a high usage day for bike rentals.

1. Descriptive Analysis. Analysing the mean, median and standard deviation of each quantitative variable provides an overview of the central tendencies and variability in the given dataset. As well as using the counts method for qualitative variables to view frequency tables, providing an idea of which results appears the most in the data. Descriptive analysis gives an initial understanding of the given dataset's basic properties.
2. Correlation Analysis. Used to measure the relationship between the continuous predictor variables. Used Cramers V value to measure the relationship between Seasons and Rented Bike Count. Lastly, used the Point Biserial correlation analysis to evaluate the correlation between Temperature and the Rented Bike Count. Results of the correlation analysis can be used to identify the possible predictor variables that have a significant effect on the response variable, (rented bike count) where these variables can later be selected in the regression analysis.
3. Simple Logistic Regression Models. Use to build models to gain an initial understanding of the individual predictor variables influence in determining the desired response - a high usage day for bike rentals. The main predictor variable of interest used here would be the temperature variable. Building a simple logistic regression model between temperature (predictor) and rented bike count (response) can be used to determine whether an increase in the temperature, increases the rented bike count.
4. Multiple Logistic Regression model (Full model): Used to develop a compete logistic model that includes all predictor variables. Building a model that includes all the predictor variables helps to explore how each of those predictors influence the likelihood of the response variable when in the presence of other predictors variables.
5. Analysis of Variance. Used to test whether there is indeed a regression relation between the response variable and the set of predictor variables. In the ANOVA table, looking at whether the SSE is smaller than the SSR as this is a way of determining whether the model is a good fit.

6. Hypothesis test: Is used to test whether a particular predictor variable has any significance in influencing the response variable. Can be used to validate whether the temperature significantly contributes to a high usage in bike rented count.
7. R-squared – Coefficient of (Multiple) determination. Determining whether a regression relation exists does not imply that the model is accurate at making predictions. To measure how accurate a model predicts the response, the R-squared is used. The closer the R-squared value is to 1, the more accurate the model is at predicting a high usage day for rented bike count.
8. Multiple logistic regression (reduced model): Having the entire set of predictors in the model may be redundant. A reduced model can be built using the result of the correlation analysis, or from making use of the various selection techniques. The use of the selection techniques helps to refine the and optimize the model by selecting only the most relevant predictor variables which results in a model with an improved accuracy at making prediction of the response variable.
9. ROC Chart – Area under the curve. Used to assess the model's ability to differentiate between high and low usage days. This measure has a range of [0,1]. A ROC result greater than or equal to 0.5 indicated the model does not discriminate between high or low usage days.

By using these methods mention, this study aims to construct a comprehensive and accurate predictive model for rental bike count, and identifying which factors are good predictors of a high usage day for bike rentals.

Initial Analysis

It is essential to comprehend the factors that influence the bike-sharing systems' usage patterns to improve user experiences and enhance system efficiency. The weather stands out as a significant factor among the many variables that can affect how often people rent bikes. It has long been assumed that seasonality, humidity, and temperature have an impact on how often people use bike-sharing programs. We explore the complex relationship between these weather-related factors and the total number of leased bikes in this analysis. With descriptive statistics as the beginning point and logistic regression and ROC analysis as the conclusion, our objective is to present a thorough overview of the results acquired using various analytical techniques. This investigation aims to shed light on the interactions between temperature, humidity, and season and provides information that can be highly beneficial to bike-sharing system operators and urban planners.

Temperature Analysis (Figure 1.1)

First, descriptive analysis is performed on the various variables to understand their underlying structure. It can be seen in figure 1.1 the Temperature variable has a mean of 13.157 degrees Celsius, represents the average temperature in the dataset and a standard deviation of 11.979 which indicates a significant spread in the temperatures from the mean. A skewness of -0.207 suggests a slight negative skew, which means there might be fewer cooler days (left tail) that are significantly cooler than most of the days. A kurtosis of -0.819 indicates a platykurtic distribution which mean the distribution of the temperature variables has a lighter tail and is less peaked than a normal distribution, indicating a more uniform spread without significant outliers as can be seen in the histogram in figure 1.1.

Humidity Analysis (Figure 1.2.1 and 1.2.2)

Descriptive analysis on the humidity shown in figure 1.2.1, shows a mean of 53.333% humidity in the dataset. With a large standard deviation of 20.347 suggesting that the humidity levels in the dataset are spread out from the average humidity. A skewness of 0.088 suggest as slight positive skew, indicating that the right tail of the humidity distribution may be slightly longer/stretched than the left tail as can be seen in the histogram in figure 1.2.2. A negative kurtosis of -0.837 suggests that the distribution of humidity variable is less peaked than the normal distribution. The humidity is less concentrated around the mean and there are fewer extreme values than in a normal distribution.

Season and Rented Bike Counts Analysis (Figure 1.3)

Analysing the frequencies between Seasons and Rented bike counts (figure 1.3), shows that the Season with the many high usage days for Bike Rentals occurred during Season 2, which is in Summer. This reinforces Kim (2018) suggestion that an increase in temperature is a good indicator in determining a high usage day for Rented Bike Count due to the favourable weather conditions. Notice that Autum and Spring have relatively large percentage of high usage day for Rented Bike Count which also backs research from Kim (2020) indicating that Autum and Spring have high usage of bike rentals due to the pleasant weather conditions.

Correlation Analysis

Correlations analysis performed on the variables reveals the relationship between all the continuous variables. Results of the correlation analysis shows that all the predictor variables have somewhat of a relationship between. With some having a strong correlation with another. (figure 2.1)

A noticeable result is the correlation coefficient of 0.91327 between the two predictor variables, Temperature and Dew Point Temperature. This indicates a strong positive correlation between these two predictor variables. The same can be seen with the Humidity and Visibility predictor variables as well, which have a correlation coefficient of 0.55835, indication a strong positive correlation between these two predictor variables. However, using two highly correlated predictor variables in regression analysis is not encourage due to the issue of multicollinearity. Thus, only one of the variables with high correlation coefficients may be selected to be included in the regression analysis.

Another standout result from correlation analysis is the relationship between Seasons and Rented Bike Count which has a Cramer’s V value of 0.5182. Which suggests a significant positive relationship between the two variables. Implying that the Seasons may be a valuable predictor in determining a high usage day for Bike Rentals. (See Figure 2.2.1). As is also mentioned in their paper (Chardon, Caruso, & Thomas, 2016) that bike sharing usage can change according to the season, weather, and day of the week. Which concurs with (Langley, 2022) which said the number of bike rentals increases in the Summer, but during the winter the number decreases.

Simple Logistic Regression between Temperature and Rented Bike Count

Taking a closer look at the Temperature variable (predictor) and the Rented Bike Count (response), simple logistic regression is performed to model the relationship between these two variables to help better understand their relationship, see figure 3. The following estimated logistic regeneration model was obtained $\log\left(\frac{p}{1-p}\right) = -1.4801 + 0.1181X_1$, where $X_1 = Temperature$. The Point Biserial correlation coefficient is 0.1181 which indicates a weak positive relationship between the Temperature and the Rented Bike Count. Although it has a weak relationship, the p-value is less than 0.0001 (hence also less than α=0.05) which suggests that the correlation between these variables is statistically significant implying an increase in temperature may increase the likelihood of a high usage day for bike rentals.

This is reinforced by the Odds Ratio of 1.125 for Temperature which indicates that for each degree Celsius increase in Temperature, the odds of a high usage day for Rented Bike Count is 1.125 times higher than the odds of when Temperature decreases. This reinforces the claim made by (Kim,2018) who stated that there is an increase in bike rentals as the temperature increases, however, his model may have included other variables, it did not include variables such as Dew Point temperature, Solar radiation, and Humidity. Further analysis is performed to see Temperatures effect in the presence of other significant variables.

Multiple Logistic regression (Full Model)

Multiple Logistic regression analysis is used to model the relationship between all the predictor variables and the response variable. The estimated logistic regression function is given by:

$$\log\left(\frac{p}{1-p}\right) = -6.7556 + 0.1400X_1 - 0.0430X_2 - 0.0839X_3 - 0.1098X_4 - 0.0004X_5 + 0.01689X_6 + 0.7804X_7 - 3.2556X_8 - 1.1980X_9 + 0.3325X_{10} + 0.3628X_{11} + 1.6225X_{12} + 0.2114X_{13} - 10.3888X_{14}$$

Where $X_1\ to\ X_9$ represent the continues predictor variables (see figure 4.1 for full variable names with their estimates) and for categorical variable Seasons:

$$X_{10=}\begin{cases}1\ if\ Season = 1(Spring)\\ \quad 0\ otherwise\end{cases} X_{11=}\begin{cases}1\ if\ Season = 2(Summer)\\ \quad 0\ otherwise\end{cases} X_{12=}\begin{cases}1\ if\ Season = 3(Autum)\\ \quad 0\ otherwise\end{cases}$$

$$X_{13=}\begin{cases}1\ if\ Holiday = 1\\ \quad 0\ otherwise\end{cases} X_{14=}\begin{cases}1\ if\ Functioning\_Day = 1\\ \quad 0\ otherwise\end{cases}$$

An interesting result from this model shows that contrary to the simple logistic regression model with Temperature as the only predictor variable, when adding in all other predictor variables, the Temperature now has a p-value=0.5957 which is greater than α=0.05 and a Odds Ratio=0.958, hence implying that with all other predictor variables added to the model, Temperature does not have a significant influence in predicting a high usage day for Rented Bike Count. Predictors such as the Temperature, Wind speed, Season (summer), and

Functional day may be omitted from the model as they too have a p-value greater than α=0.05. (See figure 4.1).

The predictor with the most significant influence in the model is Rainfall, with a p-value <0.0001, and a coefficient= -3.2556. It indicates that for each mm increase in rainfall, the log odds of a high usage day for Rented Bike Count decrease by 3.2556. With the odds=0.039, if inverted (1/0.039) it results in a odds of 25.64. Indicating that as rainfall decreases, the odds of a high usage day for Rented Bike Count is 25.64 time more than that when it is raining. This is backed by the study by (Quach, 2020) which states that the higher the precipitation the lower the bike usage as well as stating that the converse also applies. A lower precipitation results in higher bike usage count.

The next predictor with the most significant influence in the model is the Autum Season with a p-value<0.001 and a coefficient = 1.6225. It shows that during Autum, the log odds of a high usage day for Rented Bike Count Increases by 1.6225. With the odds ratio = 51.345, indicates that during Autum the odds of a high usage day for Rented Bike count is a staggering 51.345 time higher than compared to other Seasons. This concurs with the earlier correlation analysis performed between Seasons and Rented Bike count which indicated that the is a strong correlation between the two variables. As mentioned in a paper written by (Chardon, Caruso, & Thomas, 2016) they mention that bike sharing usage can vary during Seasons.

From the result of the full multiple logistic regression, it is evident that although all the variables are included in the model and have some influence in the Rented Bike Count, these influences are statistically insignificant as their p-values of greater than the significant level of 0.05. On the other hand, there are predictors such a Rainfall and Autum Season which has a significant influence in determining a high or low usage day for Rented Bike Count. (figure 4.1.)

ANOVA for Full Multiple Logistic Regression model

Performing ANOVA (see figure 5.1) on the full logistic regression model reveals the sum of squares error, SSE=196.2959 and the Regression sum of squares, SSR=241.1196. Since the SSE is less than the SSR, it indicates that the model explains some proportion of the variation in the response variable. However, the SSE is still quite large hence the model can be improved. Obtaining a coefficient of multiple determination, R-squared=0.551237, means that variance in the Rented Bike Count can be explained by all the predictors included in the above model. The R-square value also suggests a moderate level of “goodness fit” of the model. The model is moderately effective at prediction the daily usage in Rented Bike Count but can still be improved as a large percent of variation of response variable is not explained by the model.

Multiple Logistic regression (Forward Selection)

To try and improve the full logistic regression model, Forward selection technique is implemented, and the results are discussed below.

The variables Temperature, Humidity, Seasons, Functioning Day, Hour, Solar Radiation, Rainfall, Visibility, Dew Point Temperature and Snowfall were sequentially added to the model based on their Chi-Square score and were all found to be statistically significant. Forward Selection resulted in the following estimated logistic regression model:

$$\log\left(\frac{p}{1-p}\right) = -6.6142 + 0.1366X_1 - 0.0468X_2 - 0.0843X_3 - 0.0004X_4 + 0.1734X_5 + 0.7462X_6 - 3.3624X_7 - 1.1776X_8 + 0.3208X_9 + 0.3826X_{10} + 1.6574X_{11} - 10.3879X_{12}$$

Where X_1 to X_9 and X_{12} represent the continues predictor variables (see figure 6.1 for full variable names with their estimates) and for Seasons:

$$X_{9=}\begin{cases}1\ if\ Season = 1(Spring)\\ 0\ otherwise\end{cases} \quad X_{10=}\begin{cases}1\ if\ Season = 2(Summer)\\ 0\ otherwise\end{cases} \quad X_{11=}\begin{cases}1\ if\ Season = 3(Autum)\\ 0\ otherwise\end{cases}$$

$$X_{12=}\begin{cases}1\ if\ Functioning\_day = 1\\ 0\ otherwise\end{cases}$$

As with the full logistic regression model, the predictors with the most significant influence on the Rented Bike Count is the Rainfall and the Autum season.

For a mm increase in Rainfall, the log odds of a high usage day for Rented Bike Count decreases by 3.3624. Which is slightly greater than that observed in the full logistic regression model. It has a p-value<0.0001 solidifying the significance of the Rainfall variable. The odds ratio of Rainfall is 0.035, if inverted (1/0.035 = 28.57), indicates that for a mm decrease in Rainfall, the odds of a high usage day for Bike Rental Count is 28.57 times than when the rain increases.

During Autum, the log odds for high usage of Rented Bike Count increases by 1.6574 and it is statistically significant with a p-value < 0.0001. The odds ratio = 55.88, indicating that the odds of a high usage day for Rented Bike Count is 55.88 times higher than when compared to other Seasons.

The Hour predictor variable has an Odds Ratio = 1.146, indicating that for each increase in Hour during the fay, the odds of a high usage day for rented bike count increases by 1.146 times.

Temperature, with a p-value of 0.5555, is not statistically significant at the α=0.05 level, indicating it may not have a significant effect on the Rented Bike Count when including other predictor variables. The odds ratio of 0.954 suggests that for each one-degree Celsius increase in temperature, the odds of a high usage day for Rented Bike Count occurring decrease by approximately 0.954 (4.6%). However, this decrease is not statistically significant.

Humidity has an odds ratio of 0.919, indicating that a one percent increase in humidity corresponds to a 1.081 (8.1%) decrease in the odds of high usage day for rented bike count. This effect is statistically significant.

Visibility has an odds ratio of 1.000, indicating no change in the odds for a 10m change in visibility. It is statistically significant.

Dewpoint temperature has a odds ratio of 1.189, indicating that a one-degree Celsius increase in dew point temperature corresponds to an 18.9% increase in the odds of a high usage day for rented bike count.

The Hour, Humidity, Visibility, Dewpoint temperature, Solar radiation, Rainfall and Snowfall all have P-values less than significant level α=0.05, (p-values < 0.05), hence their effect on influencing the rented bike count is statistically significant.

Backwards Selection

The variables Functioning Day, Temperature, Holiday, Wind speed and Snowfall were removed from the model as they were deemed insignificant in influencing a high usage day for Rented Bike Count, with all of them having a p-value greater than significant level α=0.05. Using Backwards selection, the model obtained is given by:

$$\log\left(\frac{p}{1-p}\right) = 1.8814 + \ 0.1156X_1 - 0.0570X_2 - 0.000032X_3 + 0.0890X_4$$
$$+ 0.4736X_5 - 3.0314X_6 + 0.4043X_7 + 0.8064X_8 + 1.1487X_9$$

(see figure 6.2 for summary tables on Backwards selection)

Where $X_1\ to\ X_6$ represent the continues predictor variables (see figure 6.2 for full variable names with their estimates) and for the categorical variable Seasons:

$$X_{7=}\begin{cases}1\ if\ Season = 1(Spring)\\ \quad 0\ otherwise\end{cases} \quad X_{8=}\begin{cases}1\ if\ Season = 2(Summer)\\ \quad 0\ otherwise\end{cases} \quad X_{9=}\begin{cases}1\ if\ Season = 3(Autum)\\ \quad 0\ otherwise\end{cases}$$

As with the full logistic regression model and the Forward selection model, in this Backwards Selection model the predictors with the most significant influence on the Rented Bike Count is the again the Rainfall and the Autum season.

For a mm increase in Rainfall, the log odds of a high usage day for Rented Bike Count decreases by 3.0314. It has a p-value<0.0001 solidifying the significance of the Rainfall variable. The odds ratio of Rainfall is 0.048, if inverted (1/0.048 = 20.833), indicates that for a mm decrease in Rainfall, the odds of a high usage day for Bike Rental Count is 20.833 times higher than when the rain increases.

During Autum, the log odds for high usage of Rented Bike Count increases by 1.1487 and it is statistically significant with a p-value < 0.0001. The odds ratio = 55.88, indicating that the odds of a high usage day for Rented Bike Count is 33.386 time higher than when compared to other Seasons.

The results show that for each increase in Hour, the log odds of a high usage day increase by 0.1156 and the odds of high usage day for bike rental increases by 1.123. Suggesting that people prefer to ride bikes as the day goes on backing the statement made by (Kim H. , 2020) that there is a high usage of bike rentals during the afternoon-evening time of the day.

For each one percent increase in Humidity, the log odds of the event decrease by 0.0570, indicating that a higher Humidity is associated with a decrease in the probability of a high usage day for bike rentals. The odds of a high usage day for Rented Bike Count decreases by 0.945 (5.5%) for a percentage increase in Humidity.

An interesting observation is that for a 10m increase in Visibility, the log odds of a high usage day for Rented Bike Count decreases by 0.00032, however this effect is very little and may not be practically significant. This is reinforced by the odds ratio of 1, indicating that the odds of a high usage day for Rented Bike Count remains the same regardless of the Visibility.

For each degree Celsius in Dew Point Temperature, the log odds of a high usage day for Rented Bike Count increases by 0.0890. The odds ratio indicates that the odds of a high usage day for Rented Bike Count increases by 1.093 (9.93%).

For a one Watts per square meter increase in Solar radiation, the log odds of a high usage day for bike rentals increases by 0.4736 with the odds ratio of 1.606 suggesting that the odds of a high usage day for bike rental increases by 1.606 (60.6%).

For a one-millimetre increase in rainfall, the log odds of a high usage day for bike rentals decreases by 3.0314, indicating a significant negative impact on the Rented Bike Count when it is raining. This is reinforced by the odd ratio of 0.048 suggesting the odds of a high usage day for Rented Bike Count reduces by 0.048 (4.8%) for a one-millimetre increase in rainfall.

Amongst all the Seasons, Autum is seen to have the greatest influence on the Rented Bike count. In Autum, the log odds of a high usage day for Rented Bike Count increase by 1,1487 compared to only 0.4043 in Spring and 0.8064 in Summer. The odds ratio for Autum is 33.386 suggesting that the odds of a high usage day for Rented Bike Count increases by 33.386 times in Autum, compared to only 15.859 in Spring and 23.708 in Summer.

All the variable included in this model created using the Backward selection technique have p-values less than the significance level of α=0.05, indicating that these variables have a significant influence on determining a high usage day for Rented Bike Count.

Stepwise Selection

The variables Temperature, Humidity and Seasons were the only variables included in the final model after stepwise selection technique was used. Using Stepwise selection, the model obtained is given by:

$$\log\left(\frac{p}{1-p}\right) = 1.8472 + 0.1029X_1 - 0.0564X_2 + 0.5785X_3 + 0.4621X_4 + 1.0558X_5$$

$$where\ X_1 = Temperature\ ,\ X_2 = Humidity\ and\ Season\ is\ given\ by:$$

$$X_{3=}\begin{cases}1\ if\ Season = 1(Spring)\\ 0\ otherwise\end{cases}\quad X_{4=}\begin{cases}1\ if\ Season = 2(Summer)\\ 0\ otherwise\end{cases}\quad X_{5=}\begin{cases}1\ if\ Season = 3(Autum)\\ 0\ otherwise\end{cases}$$

(see figure 6.3 for summary tables on Stepwise selection)

Notice that Temperature was not included in the Backwards selection model, it has however been included in the Stepwise Selection and deemed to have a significant influence on the Rented Bike count with a p-value<0.0001. For a one-degree Celsius increase in Temperature, the log odds of a high usage day for bike rentals increase by 0.1019 and for a one-degree Celsius increase in temperature, the odds of a high usage day for Rented Bike count increases by 1.108 (10.8%).

As with the all the above-mentioned models, the Seasons variable has also been included in this model and again the Season with the most influence in determining a high usage day for Rented Bike Count is Autum. In Autum, the log odds of a high usage day increase by 1.0558 compared to only 0.5785 in Spring and 0.4621 in Summer. In Autum, the odds of a high usage day for bike rentals increases by 23.388 compared to only 14.510 in Spring and 12.916 in Summer. Supporting the work done previously by (Kim H. , 2020)indicating that Spring and Autum results in a high usage of bikes because of the pleasant weather conditions.

For a one percent increase in Humidity, the log odds of a high usage day for Bike rentals decrease by 0.0564 which indicates a weak influence on the Rented Bike Count. And the odds ratio is 0.952. If inverted (1/0.952=1.058) it indicates that for each percent decrease in Humidity, the odds of a high usage day for bike rentals are 1.058 times higher than if there was an increase in the Humidity. This relationship is statistically significant with p-value<0.0001. This same result has been achieved by (Quach, 2020) who stated what humidity and cloud cover has a weak impact on the Rented Bike Count.

ROC Charts

ROC curves are used to evaluate and compare the performance of each of the models mentioned above. The Area Under the Curve (AUC) value is what is used to quantify the model's ability to distinguish between a high usage day for Rented Bike Count and a Low usage day for Rented Bike Count. An AUC value closer to 1 indicates a better performing model.

- The Simple Logistic Regression Model: AUC = 0.8108 (see Figure 7.1)
- The Multiple Logistic Regression – Full Model z: AUC = 0.9430 (see Figure 7.2)
- The Multiple Logistic Regression – Forward Selection: AUC = 0.9429 (see Figure 7.3)
- The Multiple Logistic Regression – Backward Selection: AUC = 9176 (see Figure 7.4)
- The Multiple Logistic Regression – Stepwise Selection: AUC = 0.8878 (see Figure 7.5)

The results shows that every model is good at distinguishing between a high usage day and a low usage day for Rented Bike Count. However, the Full Multiple Logistic Regression Model performs the best. The Multiple Logistic Regression model with Forward selection shows a slightly lower ability to determine high or low usage day than the Full Model but is still a very good model. The Multiple logistic Regression with Backwards and Stepwise selection led to models with reduced performance, they may have removed variables that were relevant.

Summary and Conclusion

The results of the simple logistic regression model shows that Temperature alone is statistically significant in determining a high usage day for Rented Bike Count, but when included in the multiple logistic regression models, it shows that in the presence of other variables, Temperature is not as significant in determining a high usage day for Rented Bike Count. However, based on the ROC curve and AUC=0.8108 of the Simple logistic regression and as stated by (Quach, 2020) in her study, the model is good at determining a high or low usage day for Renter Bike Count. Hence by the simply logistic regression model, an increase in Temperature does increase the overall high usage day in Rented Bike count in Seoul.

From all the multiple logistic regression models obtained, the only variable included in all the models and had a significant influence on the Rented Bike Count were the Seasons variable. In each of the models, the Seasons variables has a significant influence in determining a high usage day for Rented Bike Count. With the Autum Season always having the greatest influence between all the seasons. This indicates that the Autum Seasons is a good predictor for a high usage day for Rented Bike Count. The study by (Kim H. , 2020) also results in a similar conclusion indicating that Autum is indeed a good predictor of a high usage day for Bike Rentals due to the pleasant weather conditions.

Another noticeable result was the significant influence that Rainfall had in determining a high usage day for Rented Bike Count. In every model that included Rainfall, it had the most significant influence. Hence it can be said that a decrease in Rainfall, can be regarded as a good predictor of a high usage day for Rented Bike Count. The increase in rainfall causes a low usage day for Rented bike count meanwhile a decrease in rainfall causes an increase in the Rented Bike Count (Quach, 2020).

Based on the results of the full multiple logistic regression model with the highest AUC value of 0.9430, an increase in the Hour, Dew Point Temperature and Solar Radiation with a decrease in the Humidity, Rainfall and Snowfall during the Autum Season are good predictors of a high usage day for the Rented Bike Count.

References

Chardon, C. M., Caruso, G., & Thomas, I. (2016). Bike-share rebalancing strategies, patterns, and purpose. *Journal of Transport Geography*, 22-39.

Kim, D. (2011, November 14). *Nocto.* Retrieved from Factors inluencing Travel Behaviours in Bikesharing: https://nacto.org/wp-content/uploads/2012/02/Factors-Influencing-Travel-Behaviors-in-Bikesharing-Kim-et-al-12-1310.pdf

Kim, H. (2020, June 4). *Seasonal Impact of Particulate Matter Levels on Bike Sharing in Seoul, South Korea.* Retrieved from International Journal of Environmental Research and Public Health: https://doi.org/10.3390/ijerph17113999

Kim, K. (2018). Investigation on the effects of weather and calender events on bike-sharing according to trip patterns of bike-rental stations. *Journal of Transport Geography*, 309-320.

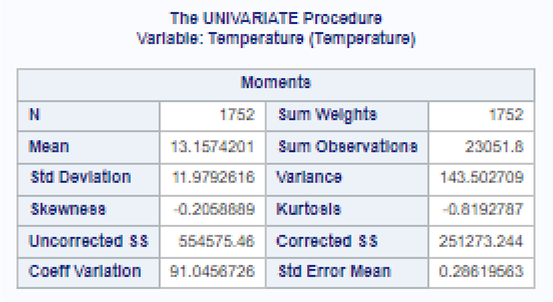
Langley, J. J. (2022, November 2). *Fifteen.* Retrieved from Seasonal Planning: How to design a bike-sharing scheme that works year-round: https://fifteen.eu/en/resources/blog/seasonal-planning-how-to-manage-bike-sharing-schemes-throughout-the-year

Quach, J. (2020, 06 1). *MALMO University.* Retrieved from Exploring the weather impact on bike sharing usage: https://www.diva-portal.org/smash/get/diva2:1480419/FULLTEXT01.pdf

Sharma, B., Nam, H. K., Yan, W., & Kim, H. Y. (2019). Barriers and Enabling Factors Affecting Satisfaction and Safety Perception with Use of Bicycle Roads in Seoul, South Korea. *Int. J. Environ. Res. Public Health*, 773.

Appendix A – Result tables

Figure 1.1. Descriptive statistics for Temperature



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

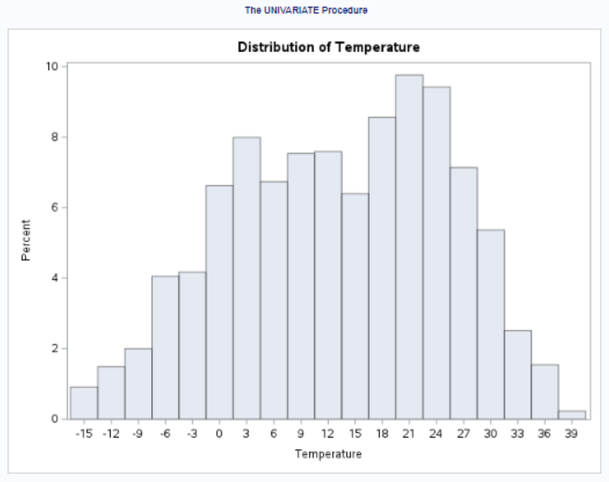
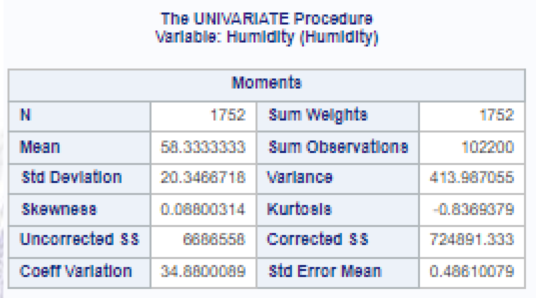


Figure 1.2. Descriptive statistics for Humidity



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

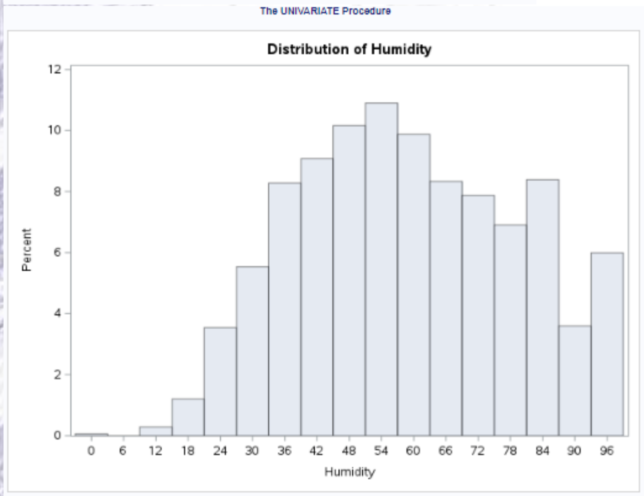
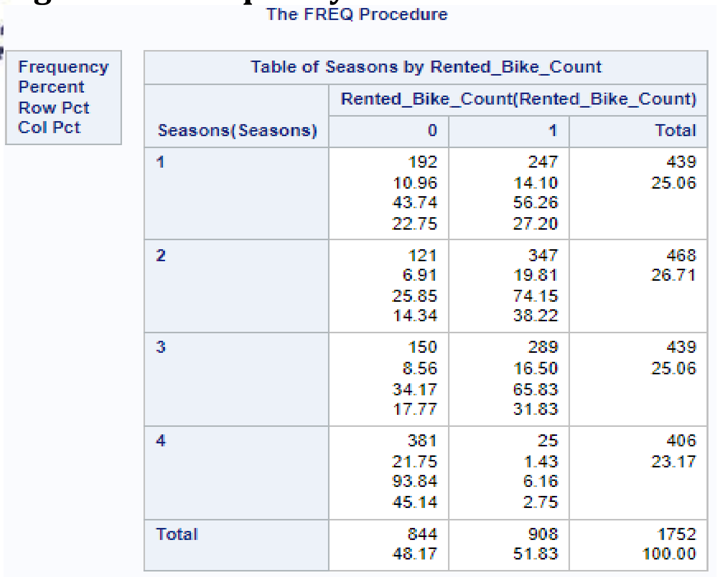


Figure 1.3. Frequency statistics for Seasons and Rented Bike Count



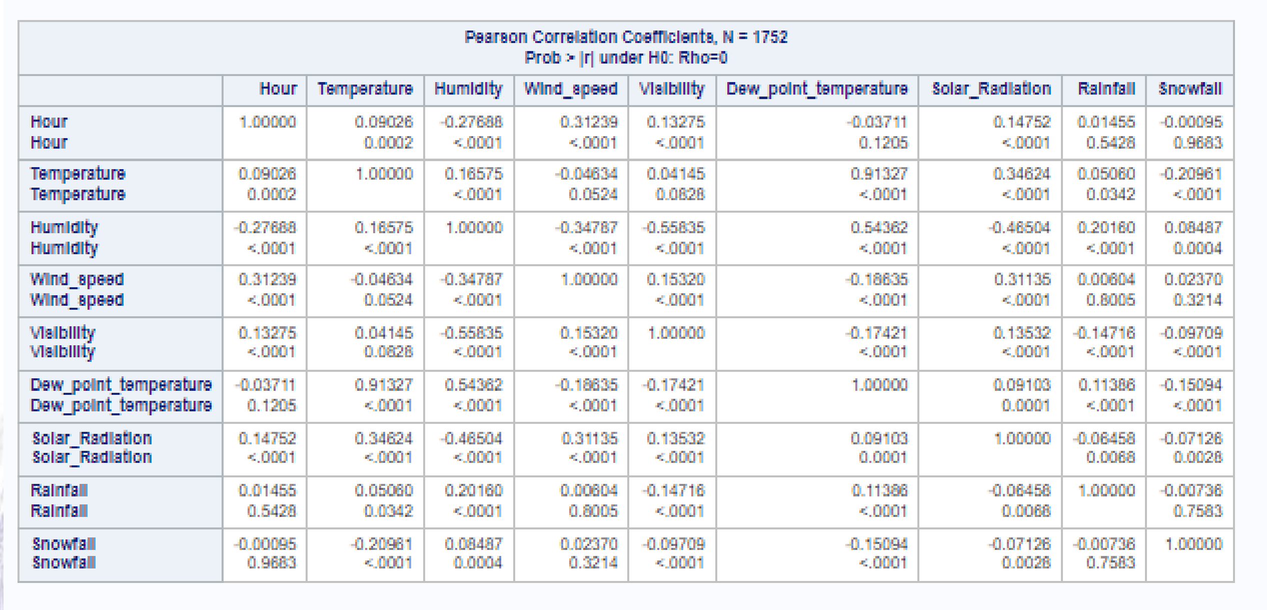
The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of Seasons by Rented_Bike_Count

	Rented_Bike_Count(Rented_Bike_Count)		
Seasons(Seasons)	0	1	Total
1	192 10.96 43.74 22.75	247 14.10 56.26 27.20	439 25.06
2	121 6.91 25.85 14.34	347 19.81 74.15 38.22	468 26.71
3	150 8.56 34.17 17.77	289 16.50 65.83 31.83	439 25.06
4	381 21.75 93.84 45.14	25 1.43 6.16 2.75	406 23.17
Total	844 48.17	908 51.83	1752 100.00

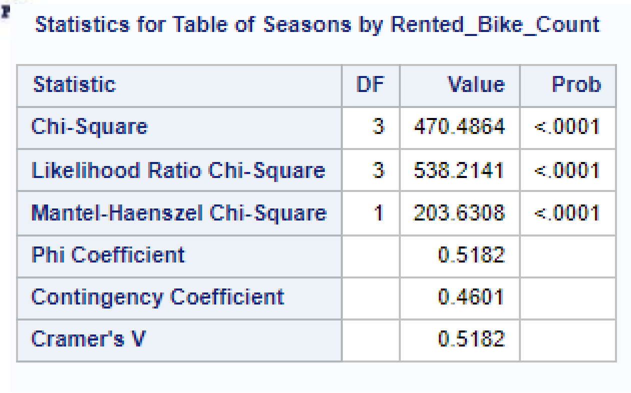
Figure 2.1 Correlation analysis between the continuous variables



Pearson Correlation Coefficients, N = 1752
Prob > \|r\| under H0: Rho=0

	Hour	Temperature	Humidity	Wind_speed	Visibility	Dew_point_temperature	Solar_Radiation	Rainfall	Snowfall
Hour Hour	1.00000	0.09026 0.0002	-0.27688 <.0001	0.31239 <.0001	0.13275 <.0001	-0.03711 0.1205	0.14752 <.0001	0.01455 0.5428	-0.00095 0.9683
Temperature Temperature	0.09026 0.0002	1.00000	0.16575 <.0001	-0.04634 0.0524	0.04145 0.0828	0.91327 <.0001	0.34624 <.0001	0.05060 0.0342	-0.20961 <.0001
Humidity Humidity	-0.27688 <.0001	0.16575 <.0001	1.00000	-0.34787 <.0001	-0.55835 <.0001	0.54362 <.0001	-0.46504 <.0001	0.20160 <.0001	0.08487 0.0004
Wind_speed Wind_speed	0.31239 <.0001	-0.04634 0.0524	-0.34787 <.0001	1.00000	0.15320 <.0001	-0.18635 <.0001	0.31135 <.0001	0.00604 0.8005	0.02370 0.3214
Visibility Visibility	0.13275 <.0001	0.04145 0.0828	-0.55835 <.0001	0.15320 <.0001	1.00000	-0.17421 <.0001	0.13532 <.0001	-0.14716 <.0001	-0.09709 <.0001
Dew_point_temperature Dew_point_temperature	-0.03711 0.1205	0.91327 <.0001	0.54362 <.0001	-0.18635 <.0001	-0.17421 <.0001	1.00000	0.09103 0.0001	0.11386 <.0001	-0.15094 <.0001
Solar_Radiation Solar_Radiation	0.14752 <.0001	0.34624 <.0001	-0.46504 <.0001	0.31135 <.0001	0.13532 <.0001	0.09103 0.0001	1.00000	-0.06458 0.0068	-0.07126 0.0028
Rainfall Rainfall	0.01455 0.5428	0.05060 0.0342	0.20160 <.0001	0.00604 0.8005	-0.14716 <.0001	0.11386 <.0001	-0.06458 0.0068	1.00000	-0.00736 0.7583
Snowfall Snowfall	-0.00095 0.9683	-0.20961 <.0001	0.08487 0.0004	0.02370 0.3214	-0.09709 <.0001	-0.15094 <.0001	-0.07126 0.0028	-0.00736 0.7583	1.00000

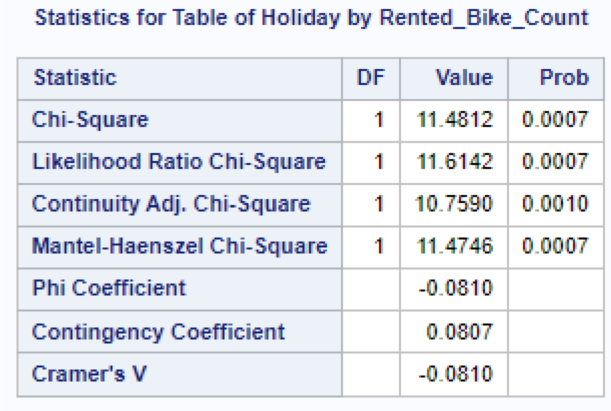
figure 2.2.1. Correlation analysis between Seasons and Rented Bike Count



Statistics for Table of Seasons by Rented_Bike_Count

Statistic	DF	Value	Prob
Chi-Square	3	470.4864	<.0001
Likelihood Ratio Chi-Square	3	538.2141	<.0001
Mantel-Haenszel Chi-Square	1	203.6308	<.0001
Phi Coefficient		0.5182	
Contingency Coefficient		0.4601	
Cramer's V		0.5182	

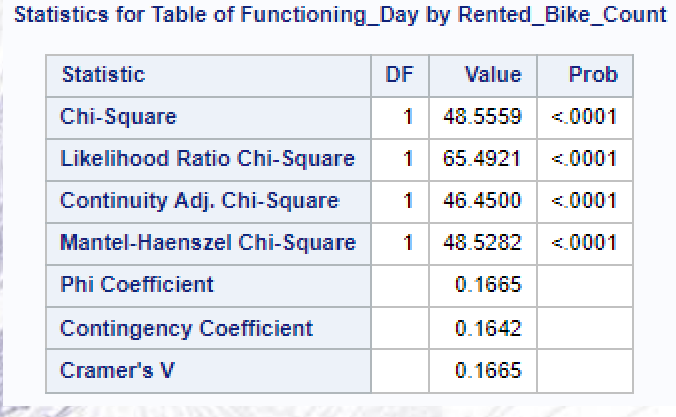
Figure 2.2.2. Correlation analysis between Holiday and Rented Bike Count



Statistics for Table of Holiday by Rented_Bike_Count

Statistic	DF	Value	Prob
Chi-Square	1	11.4812	0.0007
Likelihood Ratio Chi-Square	1	11.6142	0.0007
Continuity Adj. Chi-Square	1	10.7590	0.0010
Mantel-Haenszel Chi-Square	1	11.4746	0.0007
Phi Coefficient		-0.0810	
Contingency Coefficient		0.0807	
Cramer's V		-0.0810	

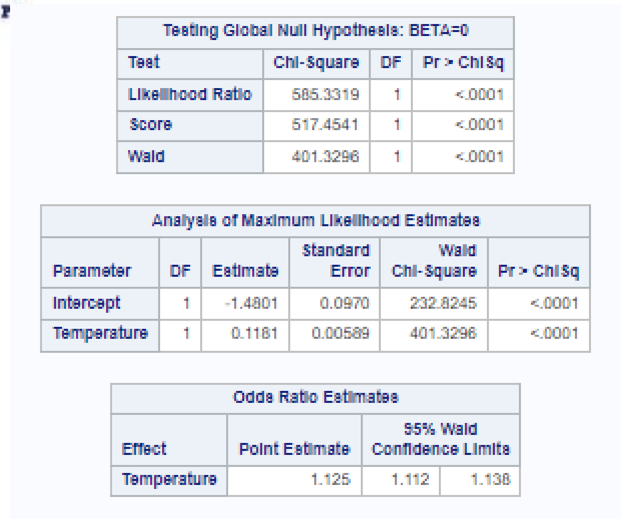
Figure 2.2.3. Correlation analysis between Functioning Day and Rented Bike Count



Statistics for Table of Functioning_Day by Rented_Bike_Count

Statistic	DF	Value	Prob
Chi-Square	1	48.5559	<.0001
Likelihood Ratio Chi-Square	1	65.4921	<.0001
Continuity Adj. Chi-Square	1	46.4500	<.0001
Mantel-Haenszel Chi-Square	1	48.5282	<.0001
Phi Coefficient		0.1665	
Contingency Coefficient		0.1642	
Cramer's V		0.1665	

Figure 3 Simple logistic regression between Temperature and Rented Bike Count

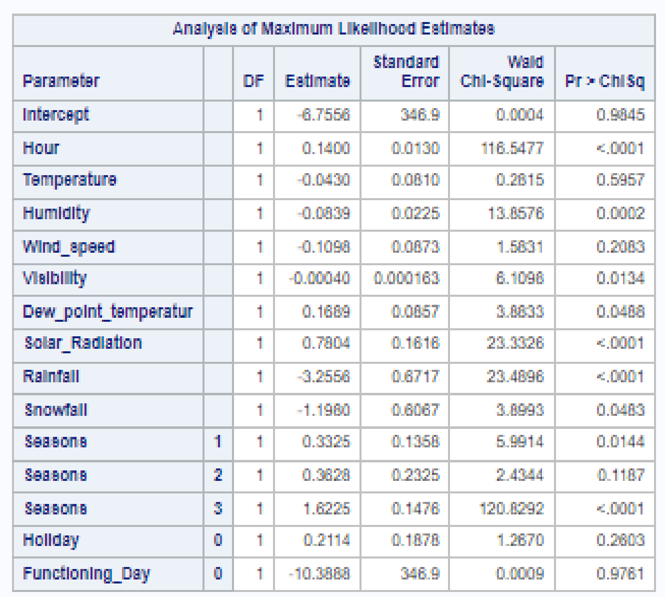


Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	585.3319	1	<.0001
Score	517.4541	1	<.0001
Wald	401.3296	1	<.0001

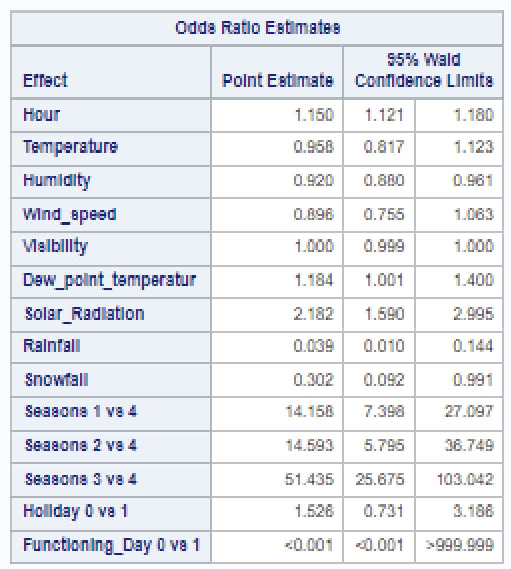
Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.4801	0.0970	232.8245	<.0001
Temperature	1	0.1181	0.00589	401.3296	<.0001

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Temperature	1.125	1.112	1.138

Figure 4.1. Multiple logistic regression (full model) Estimates and Odds Ratios

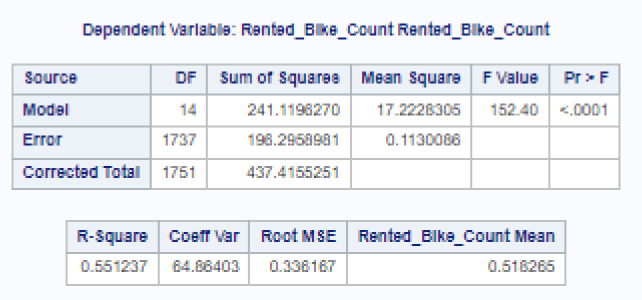


Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-6.7556	346.9	0.0004	0.9845
Hour		1	0.1400	0.0130	116.5477	<.0001
Temperature		1	-0.0430	0.0810	0.2815	0.5957
Humidity		1	-0.0839	0.0225	13.8576	0.0002
Wind_speed		1	-0.1098	0.0873	1.5831	0.2083
Visibility		1	-0.00040	0.000163	6.1098	0.0134
Dew_point_temperatur		1	0.1689	0.0857	3.8833	0.0488
Solar_Radiation		1	0.7804	0.1616	23.3326	<.0001
Rainfall		1	-3.2556	0.6717	23.4896	<.0001
Snowfall		1	-1.1980	0.6067	3.8993	0.0483
Seasons	1	1	0.3325	0.1358	5.9914	0.0144
Seasons	2	1	0.3628	0.2325	2.4344	0.1187
Seasons	3	1	1.6225	0.1476	120.8292	<.0001
Holiday	0	1	0.2114	0.1878	1.2670	0.2603
Functioning_Day	0	1	-10.3888	346.9	0.0009	0.9761



Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Hour	1.150	1.121	1.180
Temperature	0.958	0.817	1.123
Humidity	0.920	0.880	0.961
Wind_speed	0.896	0.755	1.063
Visibility	1.000	0.999	1.000
Dew_point_temperatur	1.184	1.001	1.400
Solar_Radiation	2.182	1.590	2.995
Rainfall	0.039	0.010	0.144
Snowfall	0.302	0.092	0.991
Seasons 1 vs 4	14.158	7.398	27.097
Seasons 2 vs 4	14.593	5.795	36.749
Seasons 3 vs 4	51.435	25.675	103.042
Holiday 0 vs 1	1.526	0.731	3.186
Functioning_Day 0 vs 1	<0.001	<0.001	>999.999

Figure 5. ANOVA table for Multiple logistic regression

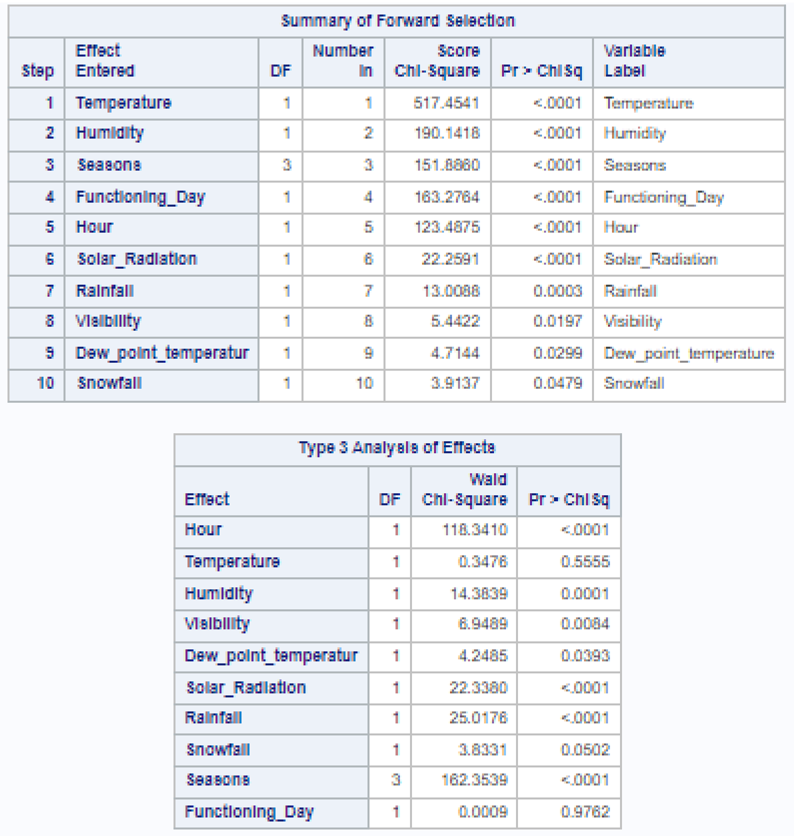


Dependent Variable: Rented_Bike_Count Rented_Bike_Count

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	241.1196270	17.2228305	152.40	<.0001
Error	1737	196.2958981	0.1130086		
Corrected Total	1751	437.4155251			

R-Square	Coeff Var	Root MSE	Rented_Bike_Count Mean
0.551237	64.86403	0.336167	0.518265

Figure 6.1.1 Results of forward selection



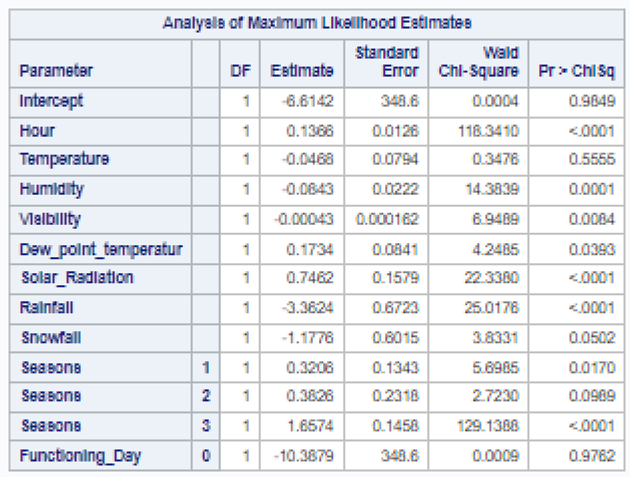
Summary of Forward Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq	Variable Label
1	Temperature	1	1	517.4541	<.0001	Temperature
2	Humidity	1	2	190.1418	<.0001	Humidity
3	Seasons	3	3	151.8860	<.0001	Seasons
4	Functioning_Day	1	4	163.2764	<.0001	Functioning_Day
5	Hour	1	5	123.4875	<.0001	Hour
6	Solar_Radiation	1	6	22.2591	<.0001	Solar_Radiation
7	Rainfall	1	7	13.0088	0.0003	Rainfall
8	Visibility	1	8	5.4422	0.0197	Visibility
9	Dew_point_temperatur	1	9	4.7144	0.0299	Dew_point_temperature
10	Snowfall	1	10	3.9137	0.0479	Snowfall

Type 3 Analysis of Effects

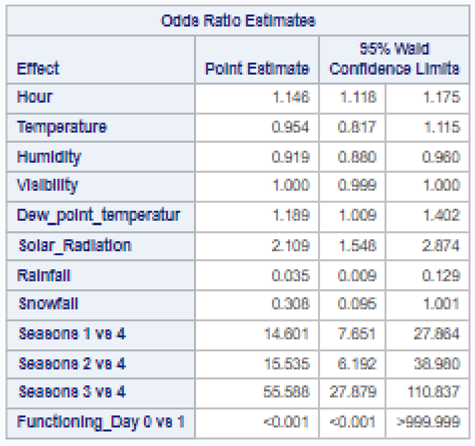
Effect	DF	Wald Chi-Square	Pr > ChiSq
Hour	1	118.3410	<.0001
Temperature	1	0.3476	0.5555
Humidity	1	14.3839	0.0001
Visibility	1	6.9489	0.0084
Dew_point_temperatur	1	4.2485	0.0393
Solar_Radiation	1	22.3380	<.0001
Rainfall	1	25.0176	<.0001
Snowfall	1	3.8331	0.0502
Seasons	3	162.3539	<.0001
Functioning_Day	1	0.0009	0.9762

Figure 6.1.2. Multiple logistic regression (Forward Selection) Estimates and Odd Ratio table



Analysis of Maximum Likelihood Estimates

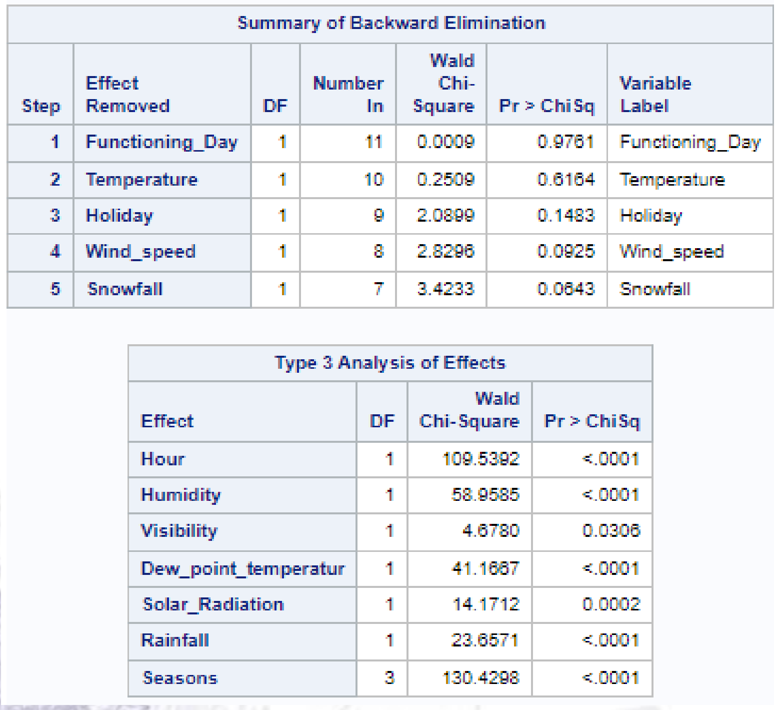
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-6.6142	348.6	0.0004	0.9849
Hour		1	0.1366	0.0126	118.3410	<.0001
Temperature		1	-0.0468	0.0794	0.3476	0.5555
Humidity		1	-0.0843	0.0222	14.3839	0.0001
Visibility		1	-0.00043	0.000162	6.9489	0.0084
Dew_point_temperatur		1	0.1734	0.0841	4.2485	0.0393
Solar_Radiation		1	0.7462	0.1579	22.3380	<.0001
Rainfall		1	-3.3624	0.6723	25.0176	<.0001
Snowfall		1	-1.1776	0.6015	3.8331	0.0502
Seasons	1	1	0.3206	0.1343	5.6985	0.0170
Seasons	2	1	0.3826	0.2318	2.7230	0.0989
Seasons	3	1	1.6574	0.1458	129.1388	<.0001
Functioning_Day	0	1	-10.3879	348.6	0.0009	0.9762



Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Hour	1.146	1.118	1.175
Temperature	0.954	0.817	1.115
Humidity	0.919	0.880	0.960
Visibility	1.000	0.999	1.000
Dew_point_temperatur	1.189	1.009	1.402
Solar_Radiation	2.109	1.548	2.874
Rainfall	0.035	0.009	0.129
Snowfall	0.308	0.095	1.001
Seasons 1 vs 4	14.601	7.651	27.864
Seasons 2 vs 4	15.535	6.192	38.980
Seasons 3 vs 4	55.588	27.879	110.837
Functioning_Day 0 vs 1	<0.001	<0.001	>999.999

Figure 6.2.1. Results of backward selection



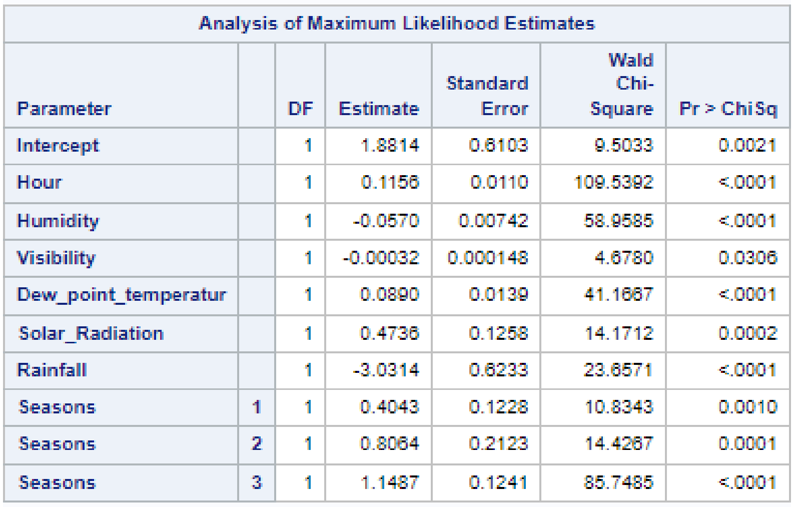
Summary of Backward Elimination

Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq	Variable Label
1	Functioning_Day	1	11	0.0009	0.9761	Functioning_Day
2	Temperature	1	10	0.2509	0.6164	Temperature
3	Holiday	1	9	2.0899	0.1483	Holiday
4	Wind_speed	1	8	2.8296	0.0925	Wind_speed
5	Snowfall	1	7	3.4233	0.0643	Snowfall

Type 3 Analysis of Effects

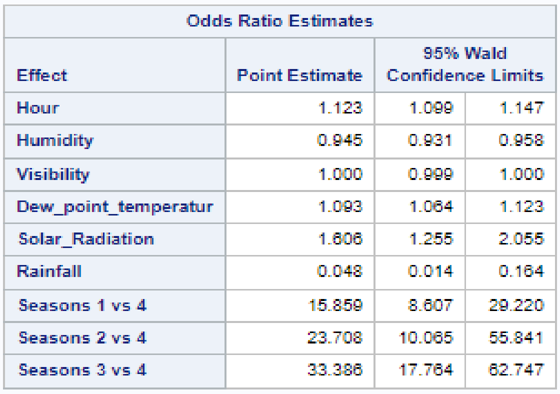
Effect	DF	Wald Chi-Square	Pr > ChiSq
Hour	1	109.5392	<.0001
Humidity	1	58.9585	<.0001
Visibility	1	4.6780	0.0306
Dew_point_temperatur	1	41.1667	<.0001
Solar_Radiation	1	14.1712	0.0002
Rainfall	1	23.6571	<.0001
Seasons	3	130.4298	<.0001

Figure 6.2.2. Multiple logistic regression (Backwards Selection) Estimate and Odd Ratio table



Analysis of Maximum Likelihood Estimates

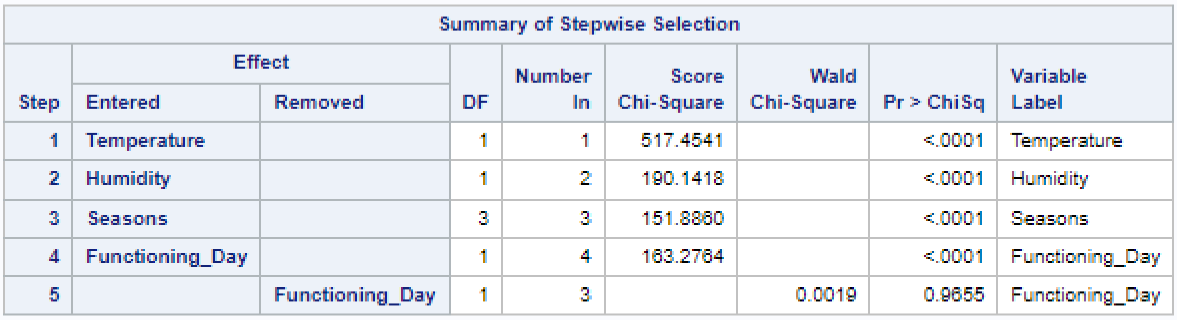
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	1.8814	0.6103	9.5033	0.0021
Hour		1	0.1156	0.0110	109.5392	<.0001
Humidity		1	-0.0570	0.00742	58.9585	<.0001
Visibility		1	-0.00032	0.000148	4.6780	0.0306
Dew_point_temperatur		1	0.0890	0.0139	41.1667	<.0001
Solar_Radiation		1	0.4736	0.1258	14.1712	0.0002
Rainfall		1	-3.0314	0.6233	23.6571	<.0001
Seasons	1	1	0.4043	0.1228	10.8343	0.0010
Seasons	2	1	0.8064	0.2123	14.4267	0.0001
Seasons	3	1	1.1487	0.1241	85.7485	<.0001



Odds Ratio Estimates

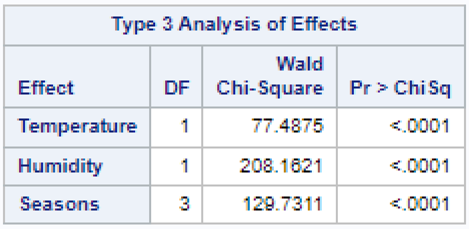
Effect	Point Estimate	95% Wald Confidence Limits	
Hour	1.123	1.099	1.147
Humidity	0.945	0.931	0.958
Visibility	1.000	0.999	1.000
Dew_point_temperatur	1.093	1.064	1.123
Solar_Radiation	1.606	1.255	2.055
Rainfall	0.048	0.014	0.164
Seasons 1 vs 4	15.859	8.607	29.220
Seasons 2 vs 4	23.708	10.065	55.841
Seasons 3 vs 4	33.386	17.764	62.747

Figure 6.3.1 Stepwise selection results



Summary of Stepwise Selection

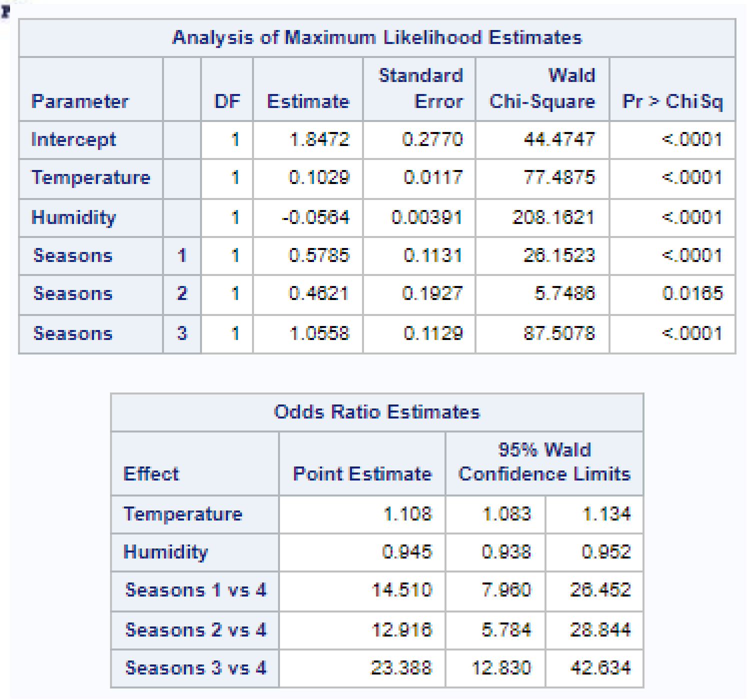
	Effect							
Step	Entered	Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Variable Label
1	Temperature		1	1	517.4541		<.0001	Temperature
2	Humidity		1	2	190.1418		<.0001	Humidity
3	Seasons		3	3	151.8860		<.0001	Seasons
4	Functioning_Day		1	4	163.2764		<.0001	Functioning_Day
5		Functioning_Day	1	3		0.0019	0.9655	Functioning_Day



Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Temperature	1	77.4875	<.0001
Humidity	1	208.1621	<.0001
Seasons	3	129.7311	<.0001

Figure 6.3.2. Multiple logistic regression (Stepwise Selection) Estimates and Odd Ratio table



Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	1.8472	0.2770	44.4747	<.0001
Temperature		1	0.1029	0.0117	77.4875	<.0001
Humidity		1	-0.0564	0.00391	208.1621	<.0001
Seasons	1	1	0.5785	0.1131	26.1523	<.0001
Seasons	2	1	0.4621	0.1927	5.7486	0.0165
Seasons	3	1	1.0558	0.1129	87.5078	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Temperature	1.108	1.083	1.134
Humidity	0.945	0.938	0.952
Seasons 1 vs 4	14.510	7.960	26.452
Seasons 2 vs 4	12.916	5.784	28.844
Seasons 3 vs 4	23.388	12.830	42.634

Figure 7.1. Roc Curve for Simple Logistic Regression Model

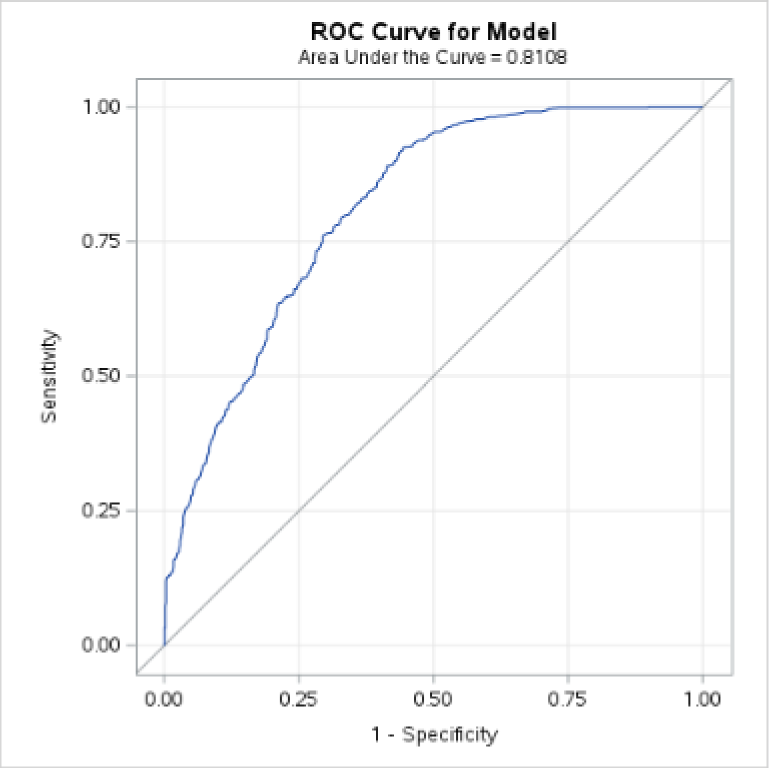


Figure 7.2 ROC curve for Multiple Logistic regression (Full model)

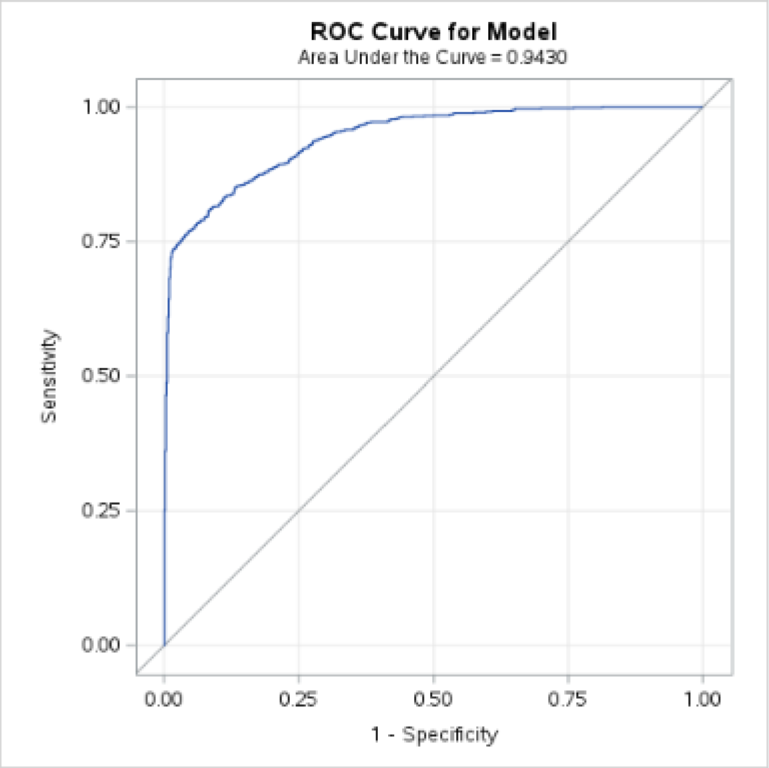


Figure 7.2. ROC curve for a Multiple Logistic Regression model (with Forward Selection)

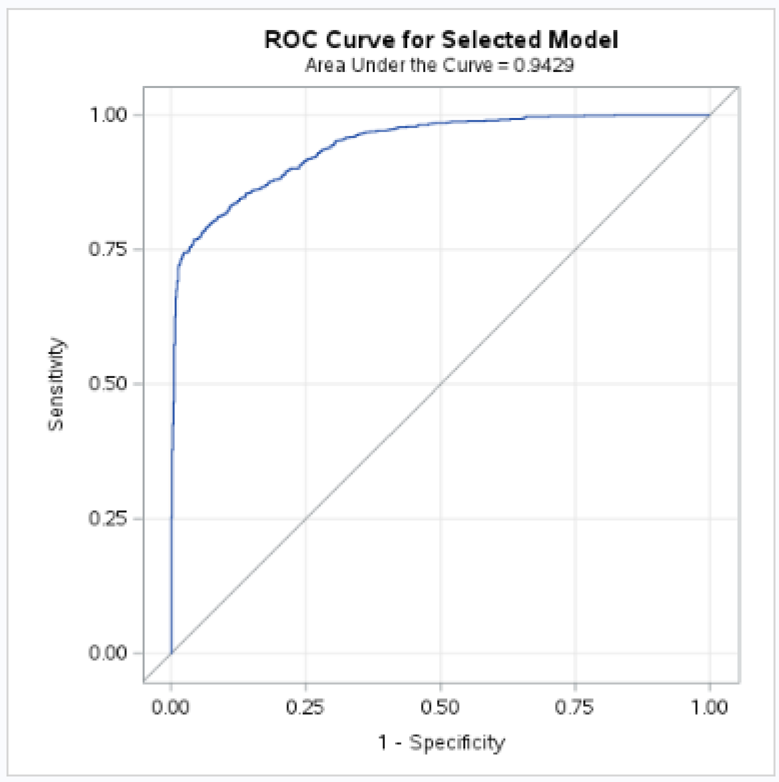


Figure 7.3. ROC curve for a Multiple Logistic Regression model (with Backwards Selection)

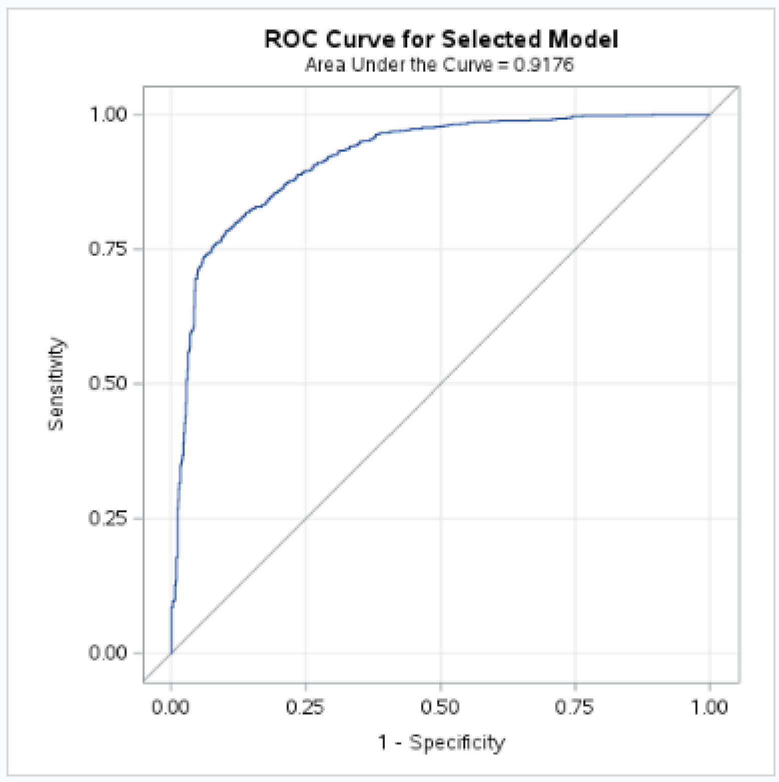
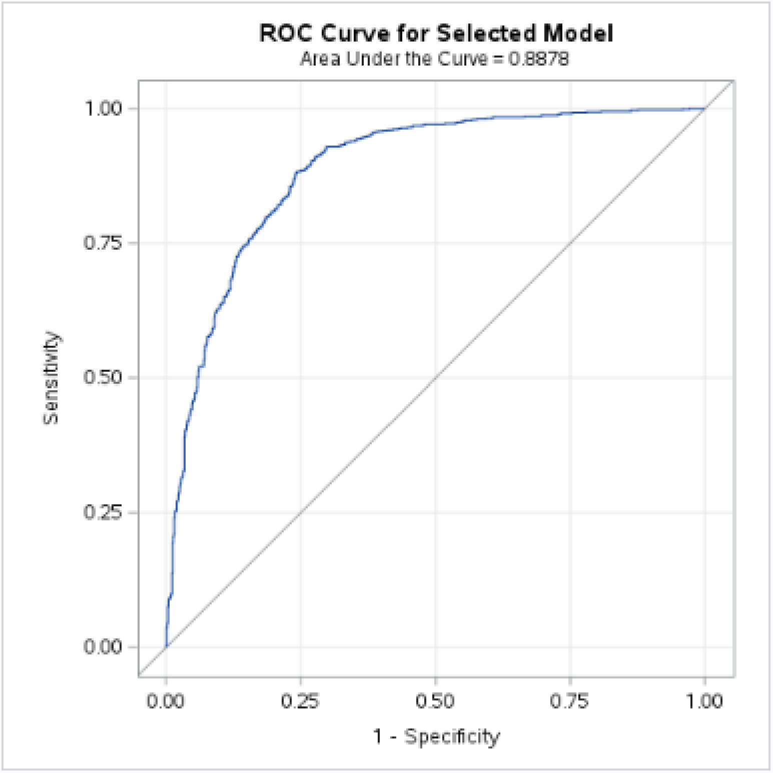
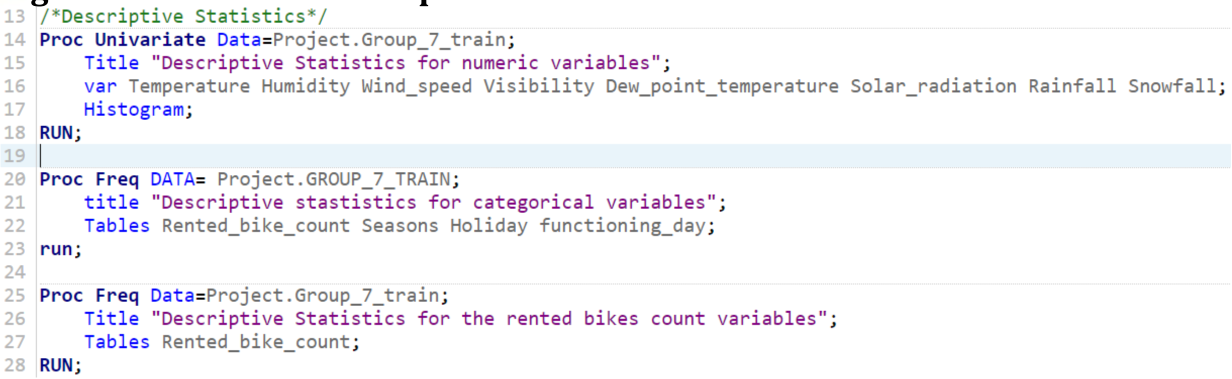


Figure 7.3. ROC curve for a Multiple Logistic Regression model (with Stepwise Selection)



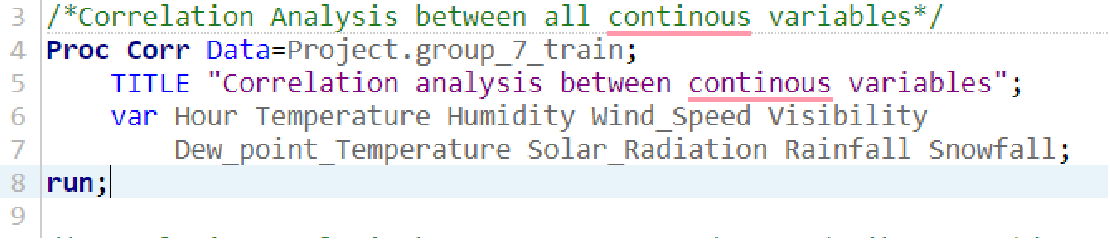
Appendix B: SAS Code

Figure 8.1. Code for Descriptive Statistics



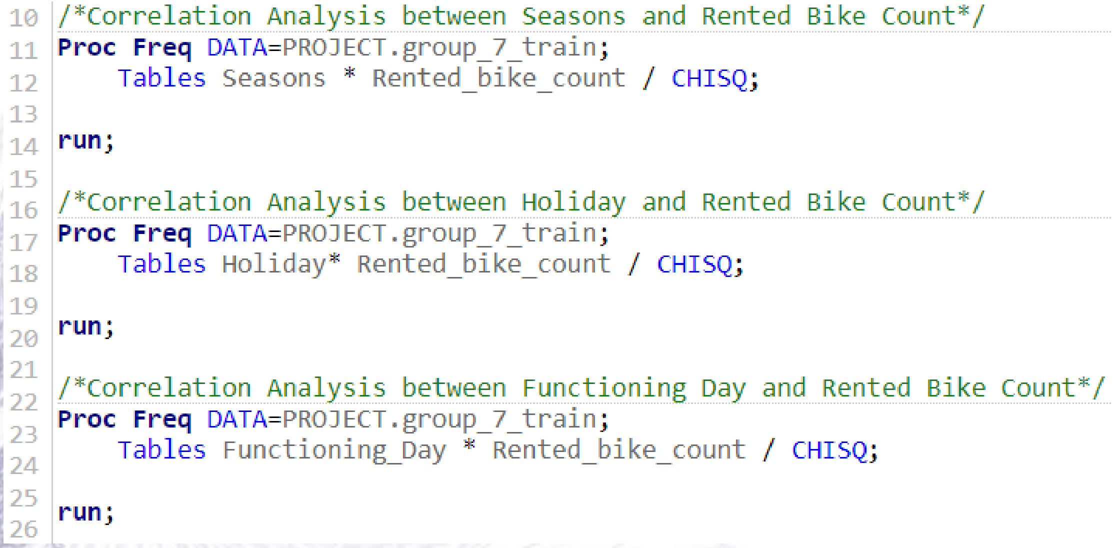
```
/*Descriptive Statistics*/
Proc Univariate Data=Project.Group_7_train;
    Title "Descriptive Statistics for numeric variables";
    var Temperature Humidity Wind_speed Visibility Dew_point_temperature Solar_radiation Rainfall Snowfall;
    Histogram;
RUN;

Proc Freq DATA= Project.GROUP_7_TRAIN;
    title "Descriptive stastistics for categorical variables";
    Tables Rented_bike_count Seasons Holiday functioning_day;
run;

Proc Freq Data=Project.Group_7_train;
    Title "Descriptive Statistics for the rented bikes count variables";
    Tables Rented_bike_count;
RUN;
```

Figure 8.2.1. Code for correlation Analysis between all continuous variables



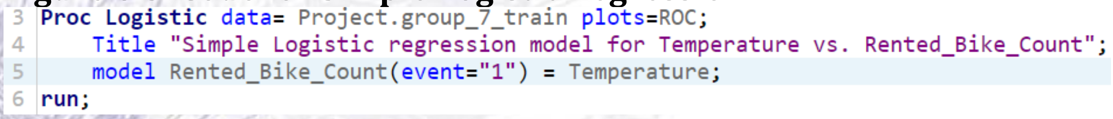
```
/*Correlation Analysis between all continous variables*/
Proc Corr Data=Project.group_7_train;
    TITLE "Correlation analysis between continous variables";
    var Hour Temperature Humidity Wind_Speed Visibility
        Dew_point_Temperature Solar_Radiation Rainfall Snowfall;
run;

```

Figure 8.2.2 Code for Correlation Analysis between categorical variables



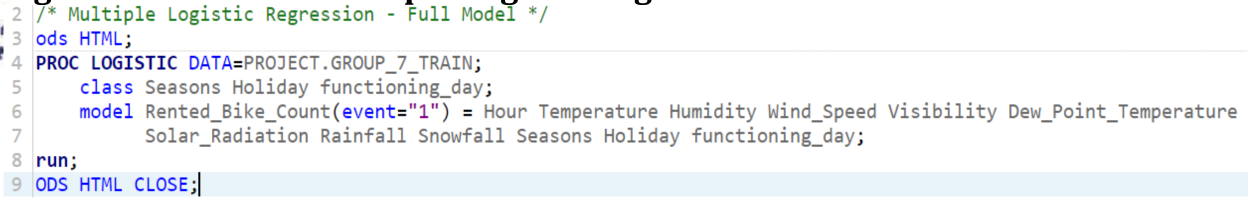
```
/*Correlation Analysis between Seasons and Rented Bike Count*/
Proc Freq DATA=PROJECT.group_7_train;
    Tables Seasons * Rented_bike_count / CHISQ;

run;

/*Correlation Analysis between Holiday and Rented Bike Count*/
Proc Freq DATA=PROJECT.group_7_train;
    Tables Holiday* Rented_bike_count / CHISQ;

run;

/*Correlation Analysis between Functioning Day and Rented Bike Count*/
Proc Freq DATA=PROJECT.group_7_train;
    Tables Functioning_Day * Rented_bike_count / CHISQ;

run;

```

Figure 8.3. Code for Simple Logistic Regression



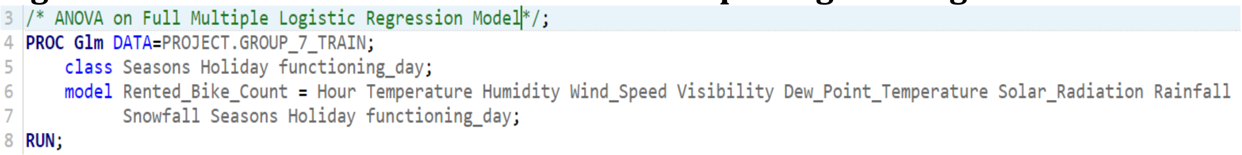
```
Proc Logistic data= Project.group_7_train plots=ROC;
    Title "Simple Logistic regression model for Temperature vs. Rented_Bike_Count";
    model Rented_Bike_Count(event="1") = Temperature;
run;
```

Figure 8.4. Code for Multiple Logistic Regression: Full Model



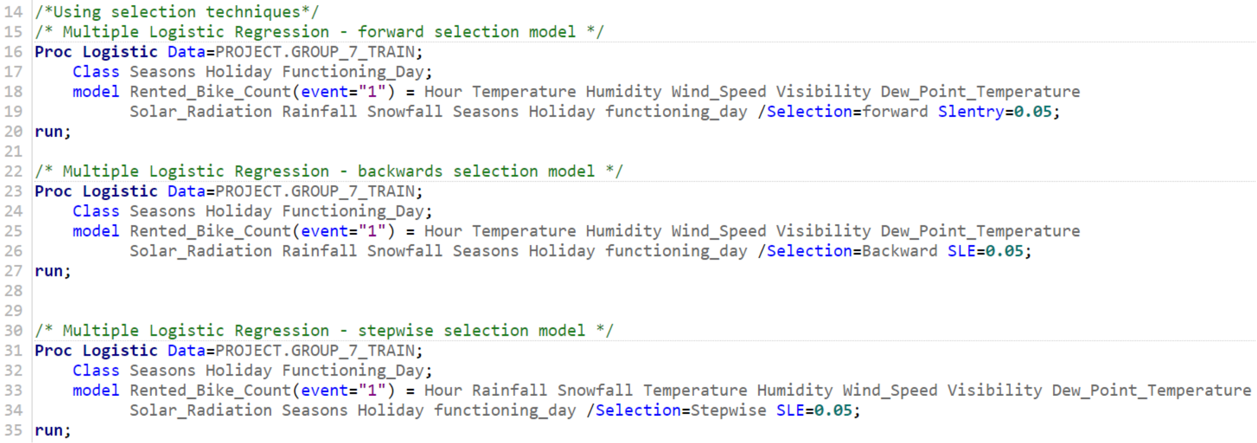
```
/* Multiple Logistic Regression - Full Model */
ods HTML;
PROC LOGISTIC DATA=PROJECT.GROUP_7_TRAIN;
    class Seasons Holiday functioning_day;
    model Rented_Bike_Count(event="1") = Hour Temperature Humidity Wind_Speed Visibility Dew_Point_Temperature
          Solar_Radiation Rainfall Snowfall Seasons Holiday functioning_day;
run;
ODS HTML CLOSE;
```

Figure 8.5. Code for ANOVA table on the Multiple Logistic Regression



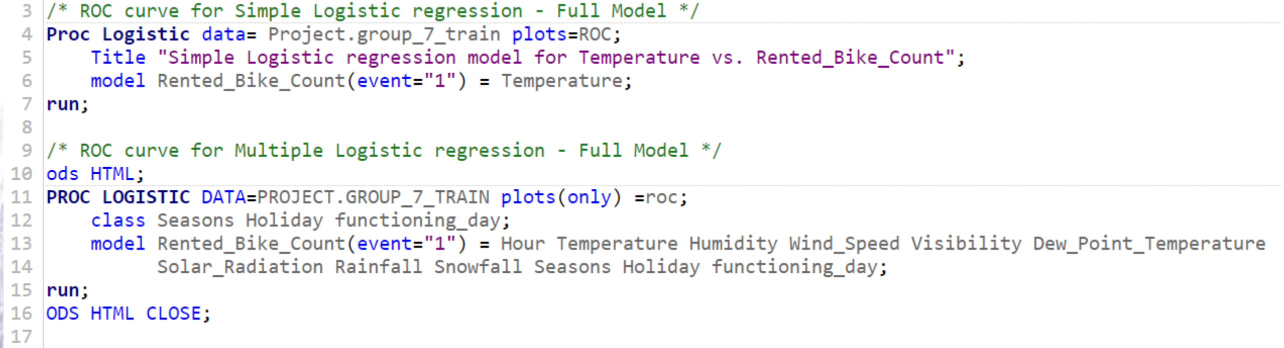
```
/* ANOVA on Full Multiple Logistic Regression Model*/;
PROC Glm DATA=PROJECT.GROUP_7_TRAIN;
    class Seasons Holiday functioning_day;
    model Rented_Bike_Count = Hour Temperature Humidity Wind_Speed Visibility Dew_Point_Temperature Solar_Radiation Rainfall
          Snowfall Seasons Holiday functioning_day;
RUN;
```

Figure 8.6. Code for Multiple Logistic Regression – Reduced Models

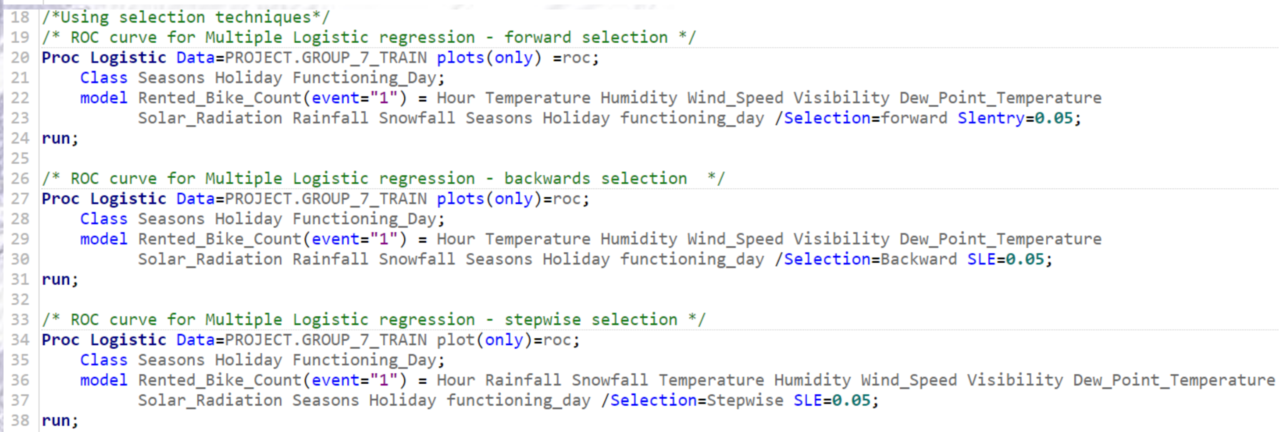


```
/*Using selection techniques*/
/* Multiple Logistic Regression - forward selection model */
Proc Logistic Data=PROJECT.GROUP_7_TRAIN;
    Class Seasons Holiday Functioning_Day;
    model Rented_Bike_Count(event="1") = Hour Temperature Humidity Wind_Speed Visibility Dew_Point_Temperature
          Solar_Radiation Rainfall Snowfall Seasons Holiday functioning_day /Selection=forward Slentry=0.05;
run;

/* Multiple Logistic Regression - backwards selection model */
Proc Logistic Data=PROJECT.GROUP_7_TRAIN;
    Class Seasons Holiday Functioning_Day;
    model Rented_Bike_Count(event="1") = Hour Temperature Humidity Wind_Speed Visibility Dew_Point_Temperature
          Solar_Radiation Rainfall Snowfall Seasons Holiday functioning_day /Selection=Backward SLE=0.05;
run;


/* Multiple Logistic Regression - stepwise selection model */
Proc Logistic Data=PROJECT.GROUP_7_TRAIN;
    Class Seasons Holiday Functioning_Day;
    model Rented_Bike_Count(event="1") = Hour Rainfall Snowfall Temperature Humidity Wind_Speed Visibility Dew_Point_Temperature
          Solar_Radiation Seasons Holiday functioning_day /Selection=Stepwise SLE=0.05;
run;
```

Figure 8.7. Code for ROC charts for all the Logistic Reression Models



```
/* ROC curve for Simple Logistic regression - Full Model */
Proc Logistic data= Project.group_7_train plots=ROC;
    Title "Simple Logistic regression model for Temperature vs. Rented_Bike_Count";
    model Rented_Bike_Count(event="1") = Temperature;
run;

/* ROC curve for Multiple Logistic regression - Full Model */
ods HTML;
PROC LOGISTIC DATA=PROJECT.GROUP_7_TRAIN plots(only) =roc;
    class Seasons Holiday functioning_day;
    model Rented_Bike_Count(event="1") = Hour Temperature Humidity Wind_Speed Visibility Dew_Point_Temperature
          Solar_Radiation Rainfall Snowfall Seasons Holiday functioning_day;
run;
ODS HTML CLOSE;
```



```
/*Using selection techniques*/
/* ROC curve for Multiple Logistic regression - forward selection */
Proc Logistic Data=PROJECT.GROUP_7_TRAIN plots(only) =roc;
    Class Seasons Holiday Functioning_Day;
    model Rented_Bike_Count(event="1") = Hour Temperature Humidity Wind_Speed Visibility Dew_Point_Temperature
          Solar_Radiation Rainfall Snowfall Seasons Holiday functioning_day /Selection=forward Slentry=0.05;
run;

/* ROC curve for Multiple Logistic regression - backwards selection  */
Proc Logistic Data=PROJECT.GROUP_7_TRAIN plots(only)=roc;
    Class Seasons Holiday Functioning_Day;
    model Rented_Bike_Count(event="1") = Hour Temperature Humidity Wind_Speed Visibility Dew_Point_Temperature
          Solar_Radiation Rainfall Snowfall Seasons Holiday functioning_day /Selection=Backward SLE=0.05;
run;

/* ROC curve for Multiple Logistic regression - stepwise selection */
Proc Logistic Data=PROJECT.GROUP_7_TRAIN plot(only)=roc;
    Class Seasons Holiday Functioning_Day;
    model Rented_Bike_Count(event="1") = Hour Rainfall Snowfall Temperature Humidity Wind_Speed Visibility Dew_Point_Temperature
          Solar_Radiation Seasons Holiday functioning_day /Selection=Stepwise SLE=0.05;
run;
```