**Seoul Bike Share: Rental Prediction**

**Desiree Valdez**

**DSC 323: Data Analysis and Regression**

**Professor Nandhini Gulasingam**

**November 16, 2022**

Table of Contents:

Introduction……………………………………………………………………………………….1

* + Introduction……………………………………………………………………………...1

Dataset Information ……………………………………………………………………………....2

* + Dataset Information……………………………………………………………………...2
  + Summary Statistics………………………………………………………………………2

Data Visualization………………………………………………………………………………...3

* + Data Visualization……………………………………………………………………….3
  + Checking Normality: Histograms and Q-Q plots………………………………………..4
  + Checking Linearity: Scatterplot Matrix………………………………………………….5
  + Checking Independence: Residual Plots…………………………………………………5
  + Outliers and Influential Points…………………………………………………………...6

Analysis, Results & Findings……………………………………………………………………...6

* + Exploratory Analysis: Transformed Full Model…………………………………………6
  + Checking for Collinearity: VIF and TOL………………………………………………..7

Improving the Model……………………………………………………………………………...8

* + Variable Selection Methods: Stepwise and ADRSQ………………………………….....8

Predictive Power…………………………………………………………………………………..8

* + Training and Test Set: Model Performance……………………………………………...8
  + Computing Predictions: Prediction 1 and Prediction 2………………………………......9

Conclusion..……………………………………………………………………………………...10

Appendices……………………………………………………………………………………….11

* + Appendix A. Dataset Information……………………………………………………....11
  + Appendix B. Summary Statistics…………………….………………………………....12
  + Appendix C. Pre-Transformation Histograms……….………………………………....13
  + Appendix D. Transformed Histograms………………………………………………....14
  + Appendix E. Pre-Transformed Q-Q Plots……………………………………………....15
  + Appendix F. Transformed Q-Q Plots………………...………………………………....16
  + Appendix G. Full Model Transformed Scatter Plot Matrix.…………………………....17
  + Appendix H. Transformed Residual Plots…………...………………………………....18
  + Appendix I. Cook D and hii………………………….………………………………....19
  + Appendix J. Transformed Full Model……………….……………………………….....20
  + Appendix K. VIF and TOL…………..…………….……………………………….......21
  + Appendix L. Selection Methods: Stepwise and ADJR…...………………………….....22
  + Appendix M. Predicted Power of Fitted Model………..…………………………….....23
  + Appendix N. Prediction 1 and Prediction 2…….…….…………………..………….....24

Introduction

Introduction

Currently, bike-sharing services are being adopted globally as a means of affordable, un-congested, and environmentally friendly transport measures. As of 2015, Seoul, S. Korea joined the popularizing public bike-sharing service industry, quickly becoming the most popular transportation system of Seoulites1. Uninterested in Seoul’s crowded trains and buses, the demand for bike rentals rapidly increased, growing by at least 200% yearly. To ensure stable availability at any given time, we will build a linear regression model, visualize our data, clean our data, and select and compare models to produce the optimal predictive model to predict how many rental bikes will be in use at any given point.

Dataset Information

**Dataset Information**

Our data set will utilize fourteen features to predict a rental bike count at any given predictor value. These features are one Date data type, Date. Four Integer data types, Hour, Humidity (%), Visibility (10m), and Rented Bike Count. Six float data types, Temperature (°C), Wind Speed (m/s), Dew Point Temperature (°C), Radiation (MJ/m2), Rainfall (mm) and Snow (cm), and three qualitative data types, Seasons (winter/spring/summer/autumn), Holiday (Holiday/No Holiday), and Functioning Day (Yes/No). To properly use the qualitative data types, we will create dummy variables that will represent a sub-group within the qualitative data types in numerical form such that d\_season1\_w represents ‘Winter,’ d\_season2\_sp represents ‘Spring,’ d\_season3\_a represents ‘Autumn,’ d\_holiday represents ‘Holiday,’ and d\_func represents ‘Yes.’ Therefore, the features we will utilize for our full model are Date, Hour, Humidity, Visibility, Rented Bike Count, Temperature, Wind Speed, Dew Point Temperature, Radiation, Rainfall, Snow, d\_season1\_w, d\_season2\_sp, d\_season3\_a, d\_func and d\_holiday. Our full model will take Bike\_Count as our dependent variable (DV) and the remaining features as independent variables (IV) or, in other words, our predictor variables.

**Summary Statistics**

In reference to Appendix B. Summary Statistics, we can determine Bike\_Count’s central tendency by viewing its mean, averaging at approximately 692 bikes rented, a median of 499 bikes rented, separating the lower and higher half of bikes rented, and a mode of 0, noting that there were no days with the same number of bikes rented. Similarly, we can analyze its measure of spread, such that Bike\_Count’s minimum value is 0, signifying that there was a day on which no bikes were rented. A maximum of 3556, such that there was a day when 3556 bikes were rented. A standard deviation of approximately 627, such that one standard deviation from the average, 692, is ± 627. A standard error of approximately ± 6.73 indicates the difference between the population mean from the sample mean and an IQR of 190 to 1052 rented bikes, representing the middle 50% of which most observation values lie.

Data Visualization

**Data Visualization**

Data visualization allows us to better graphical represent the ‘inter-relation’ of data that goes beyond the numeric representation of our summary statistics. We can visualize our data using various methods such as scatterplots, histograms, boxplots, etc. The key to visualizing our data is checking the four assumptions of linear regression: linear relationship, independence, homoscedasticity, and normality. If these assumptions aren’t realized, we may use non-linear transformations to readjust our variables. To terms of linearity, there should exist a linear relationship between the independent variable x and the dependent variable y. We can determine linearity by creating scatterplots and evaluating the correlation coefficient “r,” which ranges from -1 to 1; the larger the absolute value of r, the stronger the linear relationship. Independence, the residuals in our residual plots should be independent of each other and depict no patterns. Homoscedasticity assumes an equal or similar variance in our plots; if this assumption isn’t recognized, it will signify a skew in our results symbolizing a lack of linearity. To prove homoscedasticity, we can create a fitted values vs. residual plot. Finally, similar to the previous assumptions, normality assumes the data is normally distributed, which can be proven visually using histograms and Q-Q plots.

**Checking Normality: Histograms and Q-Q plots**

As noted in Appendix C: Pre-Transformation Histograms, our variables are not normally distributed, excluding Humidity, despite its slight left-skew. Due to the lack of linearity in our data, we will transform the appropriate variables for readjustment: bike count, windspeed, visibility, rain, and snow, which has been transformed by the cube root, square root, log, cube root, and square root, respectively.

Appendix D. Transformed Histograms show that the data has been readjusted to increase normality. Although variables, visibility, rain, snow, and bike count are not entirely normal, their normality has been optimized compared to pre-transformation. After reviewing their correlation and fit statistics, we can make a more conclusive decision about their reliability in our fitted model.

We can also see the changes in normality by comparing Appendix E: Pre-Transformed Q-Q Plots and Appendix F. Transformed Q-Q Plots. Q-Q plots allow us to check normality similar to checking linearity; if the residuals follow a roughly linear path, the variable is normally distributed. Before the transformation, our variables had a lot of curvatures, deviating from the linear ideal. Via transformation, the shape of our variables will appear more normally distributed, as we saw in Appendix E. Transformed Histograms.

**Checking Linearity: Scatterplot Matrix**

In the Seoul Bike Share dataset, we will create a scatterplot matrix between the dependent y-variable Bike\_Count per independent x-variable to ensure the assumption of linearity isn’t violated. If one or more assumptions are violated, we may need to reconsider whether the x-variable can exist in the model. Note due to the qualitative nature of our dummy variables, we will not be checking them against the four assumptions.

Regarding Figure G. Full Model Transformed Scatter Plot Matrix, the form of the variables, humidity, and windspeed, have a less prominent shape with no definitive direction. Variables, visibility, dewpoint, radiation, and temperature appear to have a weak to moderate positive pattern, while rain and snow seem to have a weak to moderate negative pattern. In the case of variables humidity and windspeed, these may be signs of a ‘weak fit’ for our model, and we may need to consider removal.

**Checking Independence and Homoscedasticity: Residual Plots**

Considering Appendix H. Transformed Residual Plots we can note irregularities in residual models of variables humidity, wind speed, visibility, dewpoint, temperature, and radiation which have been circled. However, the overall shape of these models appears independent, excluding the funnel shapes of wind speed, visibility, and dewpoint. When a variable’s residual plot begins to distribute irregularly such that the residuals do not have a constant random scatter around the zero-line, the variable lacks independence and constant variance. We may be able to procure more conclusive results by removing outliers to prevent unrepresentative skewing.

**Outliers and Influential Points**

Visually we can perceive outliers and influential points in our model by making a residual plot such as those in Appendix H. Transformed Residual Plots or a Q-Q Plot such as Appendix. Transformed Q-Q Plots. However, we can further pinpoint these observations by utilizing influential point detection statistics such as leverage statistic (hii), cov ratio, dffits, dfbetas, and cook’s distance d. In the case of this dataset, we will focus on finding and removing outliers and influential points via the cook’s distance d formula such that where n is equal to the used number of observations in our data set.

Considering Appendix I. Cook D and hii, we can view labels of the numbered observations that are either outliers, influential points, or both. These observations should and will be removed due to their power to distort the slope of the regression line.

Analysis, Results & Findings

**Exploratory Analysis: Transformed Full Model**

Once we have completed the cleaning and investigation of our data, we can begin to explore our variables' correlation and statistical significance by utilizing a full model regression. Regarding our full model in Appendix J. Transformed Full Model, we can determine a variable as significant to our model if the variable's p-value is less than 0.05, which can be found in the Pr > |t| column shown in Appendix J. Transformed Full Model. The variables that prove to be significant are Date, Hour, temperature, dewpoint, humidity, wind speed, radiation, rain, snow, function day, winter, and autumn. Variables that prove insignificant are visibility, holiday, and spring; due to the insignificance of these variables, they are likely to be removed.

Additionally, we can rank the standardized parameter estimate of each variable, weighing which variable has the greatest significance for our prediction. In this case, from most to least significance, dew point, function day, humidity, temperature, hour, rain, winter, autumn, radiation, date, wind speed, snow, visibility, spring, and at least significance holiday. Considering the significance rank, we are advised to remove variables visibility, spring, and holiday. Therefore, we hypothesize that our optimal fitted model may consist of the predictors, dew point, function day, humidity, temperature, hour, rain, winter, autumn, radiation, date, wind speed, and snow.

**Checking for Collinearity: VIF and TOL**

To ensure the independence between our x-variables, we should check whether collinearity exists in our model. Considering Appendix K. VIF and TOL, we can note severe collinearity between x-variables, temperature, and dewpoint. We will remove one of the collinear variables to fix the collinearity in our model. There are various potential solutions to solve multicollinearity, such as variable removal, standardizing predictors, adding additional observations, or transforming our data. In this case, we will remove temperature since it is less statistically significant than dewpoint, as noted in Appendix J. Transformed Full Model.

Improving the Model

**Variable Selection Methods**

Regarding Appendix L. Selection Methods: Stepwise and ADJRSQ, we chose to review two variable selection methods, stepwise and adjrsq, which produced differing fitted models. To select the most optimal fitted model, we can compare each selection method’s F-value, value, value, and p-value. F-values are utilized to measure a model’s goodness-of-fit; a higher f-value typically signifies the model with the best fit. For example, comparing the stepwise model and adjrsq model stepwise has a higher f-value at 1213.62 compared to adrsq’s 1112.82. However, in terms of value and value, the stepwise model’s value is insignificantly smaller than adrsq’s value at 0.8022 compared to adrsq’s 0.8023, while both model’s value is the same at 0.8016. Note, despite adrsq’s more significant value, is less statistically crucial than a model’s value since the value can only increase if a model’s predictors are statistically significant.

On the other hand, the value increases with any additional predictors to the model, regardless of statistical significance. Further, the adjrsq model contains the variable d\_holiday, which is statistically insignificant with a p-value greater than 0.05. Therefore, considering the above results, the stepwise variable selection model is optimal.

Predictive Power

**Training and Test Set**

Regarding Appendix M. Predictive Power of Fitted Model, our model’s predictive power for our dependent variable, bike count, is approximately 90%. The objective of the training and test set is to validate our model and test how well the model predicts new data. In the case of our validation model, the data has been split 80/20, respectively. It is essential to validate a model to ensure it performs how it is intended. Otherwise, the model would be unusable. Therefore, to create a training and test set, we must use a dataset with a large number of observations statistically significant enough to establish meaningful results. In addition, the observations being chosen should be randomized to represent the population adequately.

**Computing Predictions: Prediction 1 and Prediction 2**

To test the predictive power of the validation model, we can create a random scenario in which we can predict how many bikes will be rented based on the variables in the fitted model. As seen in Appendix N. Prediction 1 and Prediction 2, we have two predicted values, their standard error mean prediction, and their 95% confidence level prediction, which represents the highest and lowest possible range of values the predicted value can exist within.

The first prediction is based on the following values: date: 01/01/2018, hour: 9, dewpoint -17.6 (°C), humidity: 40 (%), wind speed: 0.8 (m/s), radiation: 0.05 (MJ/m2), rain: 0 (mm), snow: 0 (cm), functioning day: Yes, winter: Yes, autumn: No. Therefore, the predictive value equates to 5.5909, the lower confidence level equates to 2.7347, and the upper confidence level equates to 8.4470. However, the y-variable bike count was transformed by the non-linear transformation: cube root; therefore, the prediction values must be cubed such that the values are approximatelly174.76 and 20.45-602.70, respectively. Thus, our prediction indicates that with the above values for our variables, the predicted bikes rented are 174 with a lower and upper confidence interval of 20-602, significant at the 95% level.

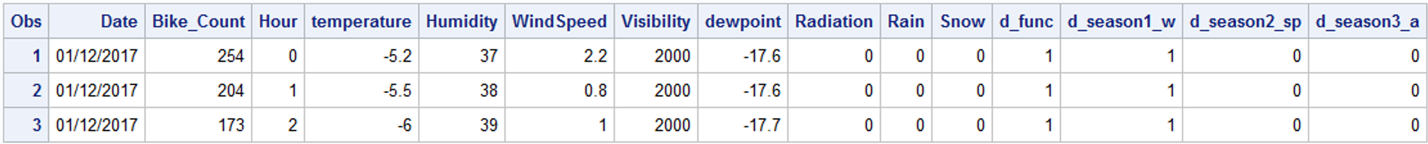
The second prediction is based on the following values: date: 17/03/2017, hour: 13, dewpoint -10.6 (°C), humidity: 45 (%), wind speed: 1.1 (m/s), radiation: 0.07 (MJ/m2), rain: 0.9 (mm), snow: 0 (cm), functioning day: Yes, winter: No, autumn: Yes. Therefore, the predictive value equates to 6.0295, the lower confidence level equates to 3.1634, and the upper confidence level equates to 8.8955. Again, since the cube root transformed the y-value, the values must be cubed to be approximately 219.20 and 31.66-703.90, respectively. Therefore, our prediction indicates that with the above values for our variables, the predicted bikes rented are 219 with a lower and upper confidence interval of 31.66-703.90, significant at the 95% level.

Conclusion

In conclusion, by analyzing, visualizing, transforming, cleaning, training, and testing our dataset, we could create a model with approximately 90% valid predictive capabilities. In other words, we can predict the rented bike count at any given point with 90% accuracy and an error rate of 10%. To advance the model towards greater accuracy, we should consider expanding our dataset with more observations, adding new predictor variables, or removing additional outliers and influential points. All-in-all, the model performs exceptionally with high satisfaction.

Appendices

**Appendix A. Dataset Information**



By understanding the dataset, one works with, we are better capable of making the decisions necessary to pick the proper statistical tests and build the most optimized models. Depending on the data types of our variables, whether variables are continuous or discrete, quantitative, or qualitative, will affect the decisions we make to our dataset.

**Appendix B. Summary Statistics**

Table

Description automatically generated

The summary statistics will consist of the measures of central tendency: mean, median, and mode, as well as the measures of spread: mix, max, standard deviation, standard error, and interquartile range (IQR). The measures of central tendency and spread pertain to a value that attempts to describe data by identifying the central position or spread, respectively2.

**Appendix C. Pre-Transformation Histograms**

Chart, histogram

Description automatically generated

First, for continuous data, we should check the normality of our features via Histograms and Q-Q plots; these visualization tools will allow us to understand the distribution of a variable. Distribution is essential since linear regression assumes variables are picked from a normal distribution. When a variable with a non-linear distribution is plotted on a histogram, the plot will appear skewed either to the left or right. However, bimodal or plots with more than one peak, is also a sign of non-linearity.

**Appendix D. Transformed Histograms**

Graphical user interface, histogram

Description automatically generated

When normality is violated, we can transform our variables to readjust into a linear form. Examples of transformation may include taking the log, square root, cube root, inverse, etc. However, it should be noted that only square root and cube root transformations can transform variables with observations greater than or equal to zero. A log transformation requires variables with observations greater than zero, while an inverse transformation requires variables to have observations greater or less than zero. When a model is transformed, a histogram should have a bell curve, or in other words, a normal distribution.

**Appendix E. Pre-Transformed Q-Q Plots**

Graphical user interface, chart, line chart

Description automatically generated

Q-Q plots allow us to explore whether our data is normally distributed between a y-variable and an x-variable. A normal distribution of a Q-Q plot will plot each variable's quantiles against each other to check whether there exists a linear relationship. A linear relationship exists if the plot appears linearly negative or positive. On the other hand, if the plot seems curved in a 'U' shape or an 'S' shape, the relationship is not linear, indicating a need for a non-linear transformation to readjust the data.

**Appendix F. Transformed Q-Q Plots**

Graphical user interface, chart, line chart

Description automatically generated

A transformed Q-Q plot is the result of transforming the data variables that appeared non-linear. Transformation is typically utilized as solution to fix the format, structure, or values in a dataset that lack the desired outcome. In the case of linear regression models, transformation encourages normality which will allow the full utilization of the dataset, enhance data quality, and reduce the number of mistakes.

**Appendix G. Full Model Transformed Scatterplot Matrix**

Graphical user interface, application, Word

Description automatically generated

To interpret a scatterplot, we should observe an overall pattern recognized by its direction, form, and strength of the relationship between the y and x variables. We can determine these aspects of the scatterplot by finding the correlation value r and analyzing the x-variables goodness-of-fit (p-value and f-value). At the same time, the form of the variables can be deduced via the scatterplot’s visualization of the data. Note we will be checking the assumptions of the transformed model from this point onward.

**Appendix H. Transformed Residual Plots**

Graphical user interface, application, Word

Description automatically generated

A residual plot is often used to prove the constant variance and independence of the variable. Split by the zero-line, a residual plot plots values of the variable on the x-axis while plotting the residual on the y-axis. The residual is the difference between predicted values and observed values. Therefore, the greater the constant variance of a residual from the zero-line, the less accurate the predicted observation. In addition, if the spread of the residuals is pattern-like, the variable is unlikely to be independent.

**Appendix I. Cook D and hii**

Graphical user interface, chart

Description automatically generated

We can determine a plot's outliers and influential points by utilizing detection statistics such as leverage statistic (hii), cov ratio, dffits, dfbetas, and cook's distance d. Outliers are observations in the dataset that are abnormally different from other observations and can typically be found outside the 3rd standard deviation of a variable's dataset. Influential points tend to have an undue influence over the regression line and can typically be found by utilizing the cook D formula √(4/n) or the leverage formula (hii) (k+1)/n. Where n is the total number of observations used in the dataset and k is the total number of independent variables.

**Appendix J. Transformed Full Model**

Table

Description automatically generated

We can determine a variables statistical significance in a model by reviewing their Pr > |t| (p-value) and standardized estimate. If a variable’s p-value is greater than 0.05 the variable is statistically insignificant, but if the p-value is less than 0.05 the variable is significant to the model. On the other hand, the standardized estimate is measuring the change in the x-variable per standard deviation of the y-variable. The greater the standard estimate, the more significant the variable is to the model. Additionally, statistics such as Root MSE,, and allow us to both determine the strength of our model, but compare which model is a better fit.

**Appendix K. VIF and TOL**

Table

Description automatically generated

To detect collinearity, we should regard statistical techniques such as variance inflation factors (VIF), Tolerance values (TOL), Standard error values (SE), R-squared values (), or adjusted R-squared values (Adj-). In this case, we will review each x-variable’s VIF and TOL values; note for our dataset, a VIF greater than 10 and a TOL less than or equal to 0.10 signifies collinearity. Furthermore, each statistical method is based on the value in which indicating a correlation in the results of a variables VIF value and its TOL value, such that if one test suggests collinearity, the second test will also suggest collinearity.

**Appendix L. Selection Methods: Stepwise and ADJRSQ**

Table

Description automatically generated

Variable selection methods allow us to fit our full model computationally to create the most optimal fitted model. We can utilize various selection methods such as forward, backward, adjrsq, stepwise, etc. The forward selection method starts with the intercept. It then sequentially adds an x-variable that most improves the model's fit, terminating once there is no significant improvement to the model. The backward model works reversely to the forward method, starting with the full model and removing insignificant variables sequentially to increase the model's fit. The adrsq selection method selects the model with the greatest overall . In contrast, the stepwise selection method works similarly to the forward method but removes or adds x-variables throughout the selection process to establish the best overall model.

**Appendix M. Predictive Power of Fitted Model**

**Graphical user interface, application

Description automatically generated**

Once we’ve selected the most optimal model via the variable selection method, we can create a training and test set to evaluate the performance of our model. In other words, a training set is a subset of the dataset used to create a model, while a test set is a subset of the dataset to ‘validate’ the trained model. The training and test set is typically split in 60-80% and 20-40% ratios.

**Appendix N. Prediction 1 and Prediction 2**

Table, Excel

Description automatically generated

To test the predictive power of the validation model, we can create a random scenario in which we can predict the value of our response variable based on the explanatory variables in the fitted model. Our prediction is found in the predicted value column and may need to be readjusted if the response variable is transformed. The prediction should be considered a strong estimation and not 100% certainty, falling within the 95% confidence interval. The 95% confidence interval is the range between the lower and upper confidence interval, so there is a 95% chance that the actual value exists within the interval.