Report on the Analysis of Seoul Bike Sharing Dataset
Aman Bhatt
DSC 323: Data Analysis and Regression

Introduction

The SeoulBikeSharing dataset offers valuable insights into bike sharing patterns in Seoul, covering a range of factors that potentially influence the count of rented bikes. The primary objective of our analysis is to predict the number of rented bikes through regression modeling. The dataset comprises a total of 14 variables, with Date, Hour, Temperature, Humidity, Windspeed, Visibility, Dew Point Temperature, Solar Radiation, Rainfall, Snowfall, Seasons (Winter, Spring, Summer, Autumn), and Functional Day (Non-Functional Hour and Functional Hour) as independent variables. The dependent variable is the count of rented bikes, labeled as "Rented Bike Counts." (Figure A1)

I chose to work on this particular dataset because it stood out as the most intriguing among the options available. Its appeal lies in its realistic representation, allowing for the exploration of real-world problems. Engaging with this dataset not only enhances my understanding of data science but also provides an opportunity to contribute to solving practical challenges and making meaningful improvements to the world through data analysis.

To make it easier to use in regression modeling, I introduced dummy variables for certain categories in the dataset, such as 'Seasons,' 'Holiday,' and 'Functioning_Day.' Moreover, a binary dummy variable was specifically crafted for the 'Snowfall' category. I transformed

the Snowfall variables into 'SnowfallDummy,' assigning 1 for Snowfall and 0 for No-Snowfall situations. Similarly, the 'Holiday' variable became 'IsHoliday,' with values 1 for Holiday and 0 for No Holiday. The 'Functioning_Day' variable was transformed into 'IsFunctionalDay,' where 1 represents a Yes for Functioning Day, and 0 stands for No Functioning Day. As for the 'Season' variables, they were converted into 'Dum_Winter,' 'Dum_Spring,' and 'Dum_Summer,' each having 1 for its respective season and 0 for the others (Figure A2). This conversion simplifies these categorical features, making them suitable for inclusion in regression analyses.

The data was summarized using PROC MEANS (Figure A3), which provided essential statistics for the entire dataset. These statistics included the mean, standard deviation, minimum, maximum, and quartiles for each numerical variable, offering an overview of the data's distribution. The ranges of values in the numerical columns seem reasonable too, so we may not have to do much data cleaning. The “Wind speed”, “Dew point temperature(°C)”, “Solar Radiation”, “Rainfall” and “Snowfall” column seems to be significantly skewed however, as the median (50 percentile) is much lower than the maximum value.

Exploratory Data Analysis

So based on the primary objective of this dataset, I decided to do a couple of histograms and boxplot to for a basic representation of the distribution of the data. So, the first histogram I was for Rental Bikes compared by Hour (Figure B1), with Hours ranging from 1 to 24 on x axis and bike share count on y axis. From the figure we can clearly see that there

is a high demand at 8th and 18th hour or at 8AM or 6PM with bike shares increasing to limits of 1000 and 2500 bikes shares. And talking about the histogram it is bimodal, and somewhat left skewed.

The next data exploration involved a scatter plot that compared Temperature to Bike Rentals (Figure B2), plotting temperature on the x-axis and bike rentals on the y-axis. Examining the figure, a clear pattern emerges there's a strong correlation between temperature and bike shares. When the temperature rises, the total bike shares also increase and the maximum range is reached when the temperature ranges from 20°C to 40°C, and as the temperature drops, the bike share count decreases accordingly. Interestingly, once the temperature falls below 0 degrees Celsius, the bike share count nearly reaches zero, indicating a significant drop in demand during extremely cold temperatures.

The next exploration involved a boxplot (Figure B 3.1) and a frequency table (Figure B 3.2) to compare bike shares on Functional and Non-Functional Day. To confirm the frequency of the functional day vs nonfunctional days I decided to do a frequency table which will return the total count of bike shares on the particular variable. And from that we can confirm that Total bikes shared on functioning days are 8465 (96.63%) vs total bikes shared on non-functioning days are 295(3.37%). And form the boxplot we can observe that The Q1, Median, and Q3 (interquartile range) encompass a higher range of bike shares on functioning days than the non-functioning days.

The next exploration involved a boxplot (Figure B 4.1) and a frequency table (Figure B 4.2) to compare bike shares on Days with Snowfall and Days with No Snowfall. To confirm the frequency of the day with snowfall vs days with no snowfall I decided to do a frequency table which will return the total count of bike shares on the particular variable. And from that we can confirm that Total bikes shared on Days with No Snowfall are 8721 vs total bikes shared on days with Snowfall are 39. And form the boxplot we can observe that The Q1, Median, and Q3 (interquartile range) encompass a higher range of bike shares on days with no snowfall than the days with Snowfall.

For an overall overview of the data, I opted for a scatterplot matrix (Figure B5) to grasp how different variables are distributed. Upon analysis, it appears there aren't any clear linear relationships between Rented_Bike_Count and the other variables (Hour, Temperature, Humidity, Visibility, Dew_Point_Temperature, Solar_Radiation, Rainfall). The connections seem more scattered than following a linear pattern.

Observations from the scatterplot matrix reveal interesting trends:

- Reduced demand during colder periods or winter seasons.
- Heightened demand on holidays.
- Minimal demand on non-functional days.
- Increased demand around 8 am and 6 pm daily.
- Rainfall substantially decreases bike share demand.

- Solar radiation peaks around midday, yet it doesn't notably impact bike share at that time.
- Higher visibility aligns with increased bike shares.

In wrapping up my data exploration, I looked at some key stats like Pearson Correlation Coefficients, VIF, and Tolerance for the complete model. (Figure B6) Checking out the Pearson Correlation Coefficients, I noticed a strong connection of about 0.91 between temperature and Dew point temperature. This high correlation signaled an issue called multicollinearity. To tackle this, I decided to drop Dew point temperature from our analysis. Because its link to our target variable was weaker (around 0.4) compared to temperature's connection, which was 0.56.

VIF and TOL numbers for the full model (Figure C1). It turned out that 'Visibility' had a p-value higher than 0.05. Also, Dew point temperature had a VIF greater than 10. So, after removing these variables, the others in the model didn't affect with each other due to multicollinearity. Each variable ended up having a p-value below 0.05 and a VIF lower than 10 (Figure C2), showing they're individually important and don't cause trouble when put together.

This detailed checkup helped refine our model. By handling multicollinearity issues and dropping some variables, we made sure our analysis is on solid ground. It gave us a clearer picture of how each variable affects our main target. Removing highly linked variables and those causing multicollinearity made our model stronger and more reliable for future analyses.

Linear Regression Analysis

To initiate the linear regression analysis, I partitioned the dataset into training and testing subsets (Refer to Figure C2.1 and Figure C2.2). This partition was executed using an 80-20 split ratio, with 80% of the data allocated for training the model and the remaining 20% for assessing its accuracy. The split was achieved using a sampling rate of 0.8 and a random seed value of 592587.

Moving into the model selection phase, I employed three distinct techniques: forward selection, backward elimination, and stepwise selection. These methods aim to identify the most suitable model for our data. Forward Selection (Figure C4) involves starting with an empty set of features and progressively adding one feature at a time. It selects the feature that significantly enhances the model's performance. The resulting model from this method highlighted Rented Bike Count = Hour + Temperature + Humidity + Wind_Speed _ Solar_Raditation + Rainfall. Contrarily, Backward Elimination (Figure D5) initiates with all features and iteratively removes the one contributing the least to the model's performance. The derived model from this approach showcased Rented Bike Count = Hour + Temperature + Humidity + Wind_Speed _ Solar_Raditation + Rainfall. Stepwise Selection (Figure D6) combines elements of both forward and backward methods. It begins with an empty set and adds features gradually while assessing if removing previously added features could enhance the model's performance. The resulting model from this method was Rented Bike Count = Hour + Temperature + Humidity + Wind_Speed _ Solar_Raditation + Rainfall.

Each of these approaches aims to refine the model by either including or excluding features based on their impact on the model's overall performance. These methodologies play a crucial role in constructing an optimal model that effectively represents the inherent relationships within the dataset.

Following that, I implemented a procedure to eliminate outliers and influential data points that might disrupt the dataset and lead to uncertainties in our predictive values and overall model accuracy. To achieve this, I utilized multiple linear regression models on the training dataset, focusing on understanding how the predictor variables relate to the target variable ("Rented_Bike_Count"). Concurrently, I generated diagnostic plots to assess the model's assumptions and performance.

Subsequently, upon obtaining the table of results, I scrutinized specific criteria to identify outliers and influential points. Specifically, I looked for values where rStudent ≥ +- 3, hii > 2p/n, diffits > 2 root p/n, and dfbeta > 2/ root n. I ran four rounds of Cooks Distance graphs (Figure C3.3) removing outliers and influential points until the adjusted r-square stopped improving. I did this because I did not want to overfit the data. The number of observations were cut down from 8760 to 8156 and the adjusted r-square improved from 0.5500 to 0.5117. (Refer to Figures D1, D2, D3, D4 for visual representations of the plotted influential points and outliers.)

Next, I checked how well the model could predict new data by using two methods: backward and stepwise selection. Surprisingly, both methods gave the same results—they picked the same variables. When I compared the performance of the models on the

training and test data, they turned out to be identical. The numbers for things like RSME, R-squared, and others were the same for both the training and test data. This made me confident that this model is the best at predicting new data outside of what we've already seen.

To test the model's response to new data, I ran three predictions using the final model, which includes variables like Hour, Temperature, Humidity, and others. This test helps ensure the model works well with data it hasn't seen before, giving me more trust in its ability to predict bike rental counts beyond our current dataset.

I computed three predictions to see how the model will respond with outside data using the Final Model → Rented Bike Count = Hour + Temperature + Humidity + Wind_Speed _ Solar_Raditation + Rainfall + Dum_Winter +Dum_Spring + Dum_Summer +IsHoliday + IsFunctionalDay. The first data line is 1 -5.5 38 0.8 2000 -17.6 0 0 0 Winter Holiday Yes, 3 20.5 38 1.0 1000 -21.1 0 0 0 Summer Holiday No .

Conclusion

Upon meticulous analysis and model refinement, the final model for predicting Rented Bike Count emerges as:

Rented Bike Count = -188.39 + 27.259 Hour +26.108 Temperature -7.9786 Humidity +18.085 Wind_Speed -79.748 Solar_Raditation -59.924 Rainfall -366.1 Dum_Winter +144.83 Dum_Spring +154.57 Dum_Summer -120.09 IsHoliday +931.43 IsFunctionalDay

Our observations yield valuable insights into the dynamics of bike demand based on various factors:

- **Time of Day Influence**: The peak demand for bikes occurs at 8 AM and 6 PM, with a noticeable rise in demand starting approximately an hour before these times.
- **Weekday vs. Weekend Demand**: Weekdays exhibit higher bike demand compared to weekends, indicating varying usage patterns based on the day of the week.
- **Weather Impact**: A correlation is evident between weather conditions and bike demand. For instance, a rise in wind speed by one unit corresponds to an increase in bike demand by 18 units. Conversely, a decrease in humidity by one unit results in a decrease in demand by 8 bikes.
- **Solar Radiation Effect**: Surprisingly, higher solar radiation translates to a decrease in bike demand, indicating potential behavioral shifts during sunnier periods.
- **Seasonal Variations**: Seasonal changes significantly affect bike demand. Winter sees a substantial decrease in demand by 366 bikes, whereas spring and summer witness increases by 145 and 155 bikes, respectively.
- **Functional Day Impact**: On functional days, the presence of a holiday diminishes demand by 120 bikes. However, on regular functional days, demand drastically increases by 931 bikes.

In short, these discoveries show that many different things affect how much people use bikes. Knowing these details is really important for putting resources where they're needed and providing good service, especially when bike demand changes during the day, week,

and different seasons. This helps make bike-sharing systems better, making users happier and operations smoother.

Appendix 1: Introduction

Figure A1: Initial Dataset

Importing Data: Seoul Bike Sharing Dataset

Obs	Date	Rented_Bike_Count	Hour	Temperature	Humidity	Wind_speed	Visibility	Dew_point_temperature	Solar_Radiation	Rainfall	Snowfall	Seasons	Holiday	Functioning_Day
1	01/12/20	254	0	-5.2	37	2.2	2000	-17.6	0.00	0.0	0	Winter	No Holid	Yes
2	01/12/20	204	1	-5.5	38	0.8	2000	-17.6	0.00	0.0	0	Winter	No Holid	Yes
3	01/12/20	173	2	-6.0	39	1.0	2000	-17.7	0.00	0.0	0	Winter	No Holid	Yes
4	01/12/20	107	3	-6.2	40	0.9	2000	-17.6	0.00	0.0	0	Winter	No Holid	Yes
5	01/12/20	78	4	-6.0	36	2.3	2000	-18.6	0.00	0.0	0	Winter	No Holid	Yes
6	01/12/20	100	5	-6.4	37	1.5	2000	-18.7	0.00	0.0	0	Winter	No Holid	Yes
7	01/12/20	181	6	-6.6	35	1.3	2000	-19.5	0.00	0.0	0	Winter	No Holid	Yes
8	01/12/20	460	7	-7.4	38	0.9	2000	-19.3	0.00	0.0	0	Winter	No Holid	Yes
9	01/12/20	930	8	-7.6	37	1.1	2000	-19.8	0.01	0.0	0	Winter	No Holid	Yes
10	01/12/20	490	9	-6.5	27	0.5	1928	-22.4	0.23	0.0	0	Winter	No Holid	Yes
11	01/12/20	339	10	-3.5	24	1.2	1996	-21.2	0.65	0.0	0	Winter	No Holid	Yes
12	01/12/20	360	11	-0.5	21	1.3	1936	-20.2	0.94	0.0	0	Winter	No Holid	Yes

Figure A2: After Dummy Variables

Creating Dummy Variables

Count	Hour	Temperature	Humidity	Wind_speed	Visibility	Dew_point_temperature	Solar_Radiation	Rainfall	Dum_Winter	Dum_Spring	Dum_Summer	IsHoliday	IsFunctionalDay	SnowfallDummy
254	0	-5.2	37	2.2	2000	-17.6	0.00	0.0	1	0	0	0	1	0
204	1	-5.5	38	0.8	2000	-17.6	0.00	0.0	1	0	0	0	1	0
173	2	-6.0	39	1.0	2000	-17.7	0.00	0.0	1	0	0	0	1	0
107	3	-6.2	40	0.9	2000	-17.6	0.00	0.0	1	0	0	0	1	0
78	4	-6.0	36	2.3	2000	-18.6	0.00	0.0	1	0	0	0	1	0
100	5	-6.4	37	1.5	2000	-18.7	0.00	0.0	1	0	0	0	1	0
181	6	-6.6	35	1.3	2000	-19.5	0.00	0.0	1	0	0	0	1	0
460	7	-7.4	38	0.9	2000	-19.3	0.00	0.0	1	0	0	0	1	0
930	8	-7.6	37	1.1	2000	-19.8	0.01	0.0	1	0	0	0	1	0
490	9	-6.5	27	0.5	1928	-22.4	0.23	0.0	1	0	0	0	1	0
339	10	-3.5	24	1.2	1996	-21.2	0.65	0.0	1	0	0	0	1	0

Figure A3: Summarizing the Data

Summarizing Data

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	25th Pctl	Median	75th Pctl	Maximum
Rented_Bike_Count	8760	704.6020548	644.9974677	0	191.0000000	504.5000000	1065.50	3556.00
Hour	8760	11.5000000	6.9225817	0	5.5000000	11.5000000	17.5000000	23.0000000
Temperature	8760	12.8829224	11.9448252	-17.8000000	3.5000000	13.7000000	22.5000000	39.4000000
Humidity	8760	58.2262557	20.3624133	0	42.0000000	57.0000000	74.0000000	98.0000000
Wind_speed	8760	1.7249087	1.0363000	0	0.9000000	1.5000000	2.3000000	7.4000000
Visibility	8760	1436.83	608.2987120	27.0000000	940.0000000	1698.00	2000.00	2000.00
Dew_point_temperature	8760	4.0738128	13.0603693	-30.6000000	-4.7000000	5.1000000	14.8000000	27.2000000
Solar_Radiation	8760	0.5691107	0.8687462	0	0	0.0100000	0.9300000	3.5200000
Rainfall	8760	0.1486872	1.1281930	0	0	0	0	35.0000000
Dum_Winter	8760	0.2465753	0.4310419	0	0	0	0	1.0000000
Dum_Spring	8760	0.2520548	0.4342173	0	0	0	1.0000000	1.0000000
Dum_Summer	8760	0.2520548	0.4342173	0	0	0	1.0000000	1.0000000
IsHoliday	8760	0.0493151	0.2165374	0	0	0	0	1.0000000
IsFunctionalDay	8760	0.9663242	0.1804036	0	1.0000000	1.0000000	1.0000000	1.0000000
SnowfallDummy	8760	0.0044521	0.0665788	0	0	0	0	1.0000000

Appendix 2: Exploratory Data Analysis

Figure B1: Hour vs Rented Bike Count Histogram

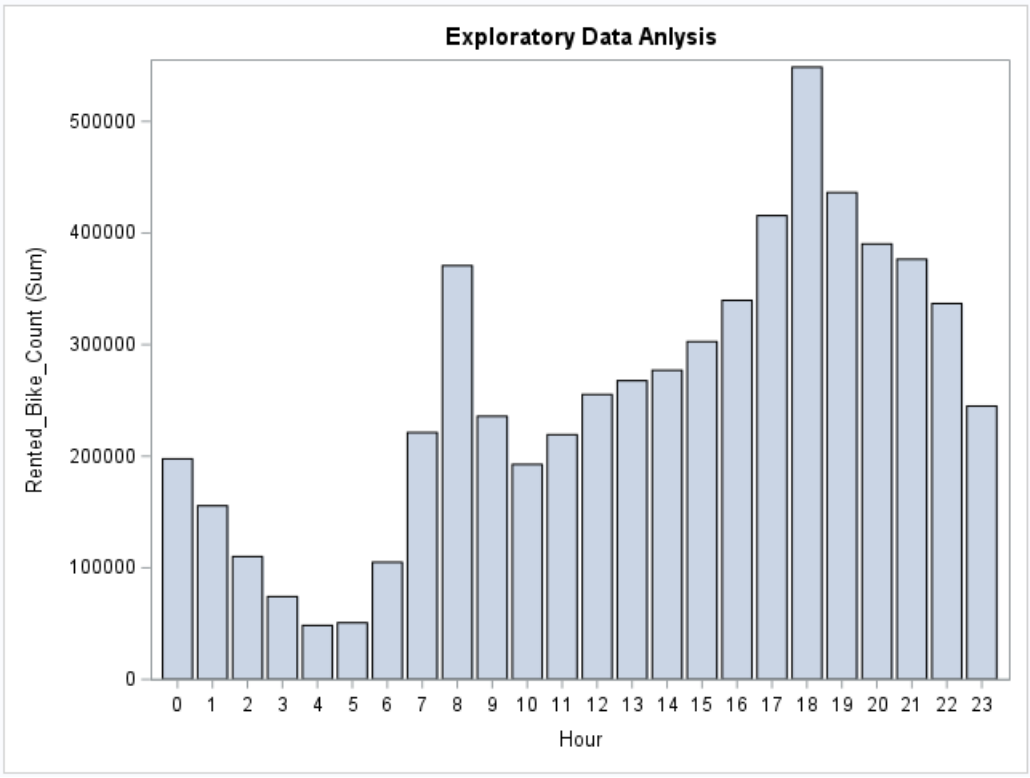


Figure B2: Temperature vs Rented Bike Count Scatterplot

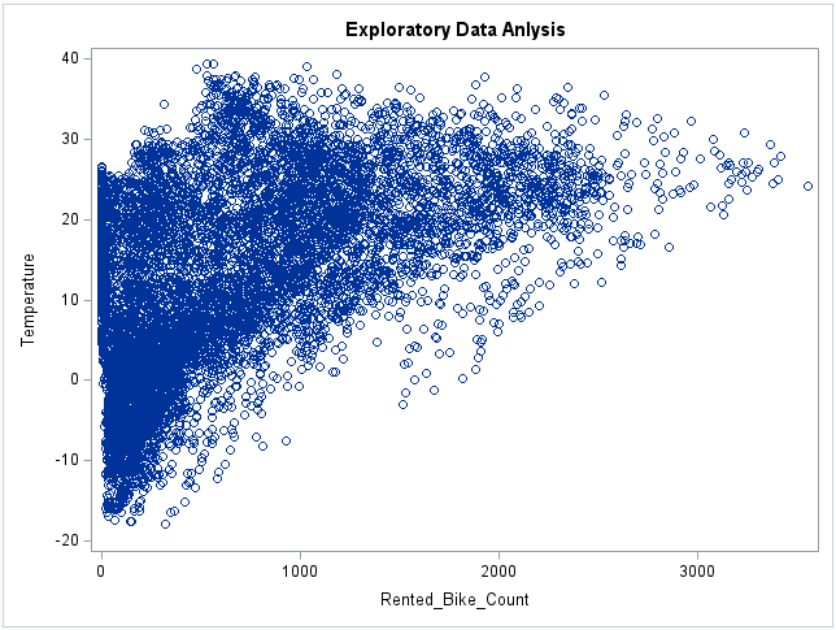


Figure B3.1: Boxplot for Bike Share on Functional Day vs Non-Functional Day

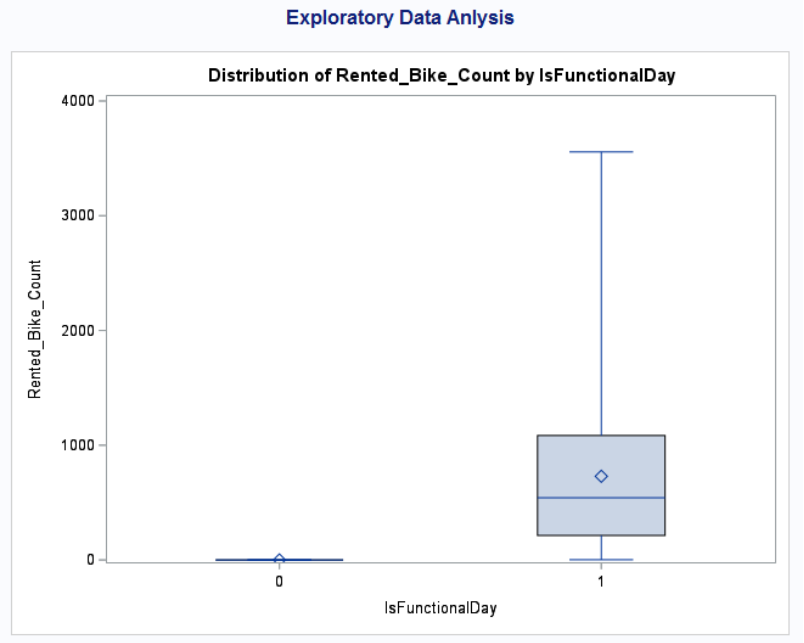


Figure B3.2: Frequency Table for Bike Share on Functional Day vs Non-Functional Day

3298	3309	3365	3380	3384	3404	3418	3556	Total
0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	295 3.37
1 0.01 0.01 100.00	1 0.01 0.01 100.00	1 0.01 0.01 100.00	1 0.01 0.01 100.00	1 0.01 0.01 100.00	1 0.01 0.01 100.00	1 0.01 0.01 100.00	1 0.01 0.01 100.00	8465 96.63
1 0.01	1 0.01	1 0.01	1 0.01	1 0.01	1 0.01	1 0.01	1 0.01	8760 100.00

Figure B4.1: Boxplot for Bike Share on Snowfall Day vs Non-Snowfall Day

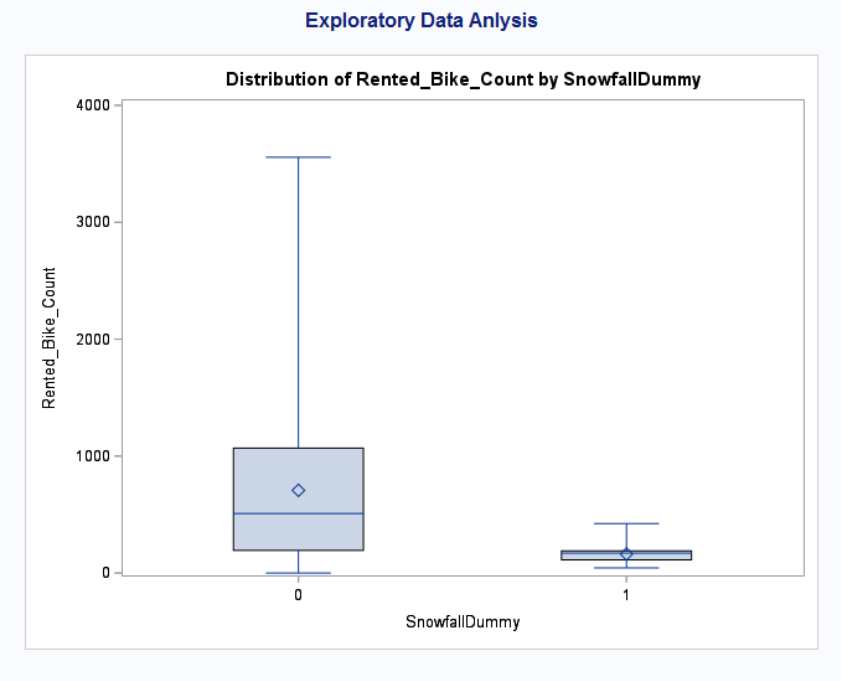


Figure B4.1: Frequency Table for Bike Share on Snowfall Day vs Non-Snowfall Day

Exploratory Data Anlysis

The MEANS Procedure

Analysis Variable : Rented_Bike_Count		
SnowfallDummy	N Obs	Sum
0	8721	6165957.00
1	39	6357.00

Figure B5: Scatterplot Matrix

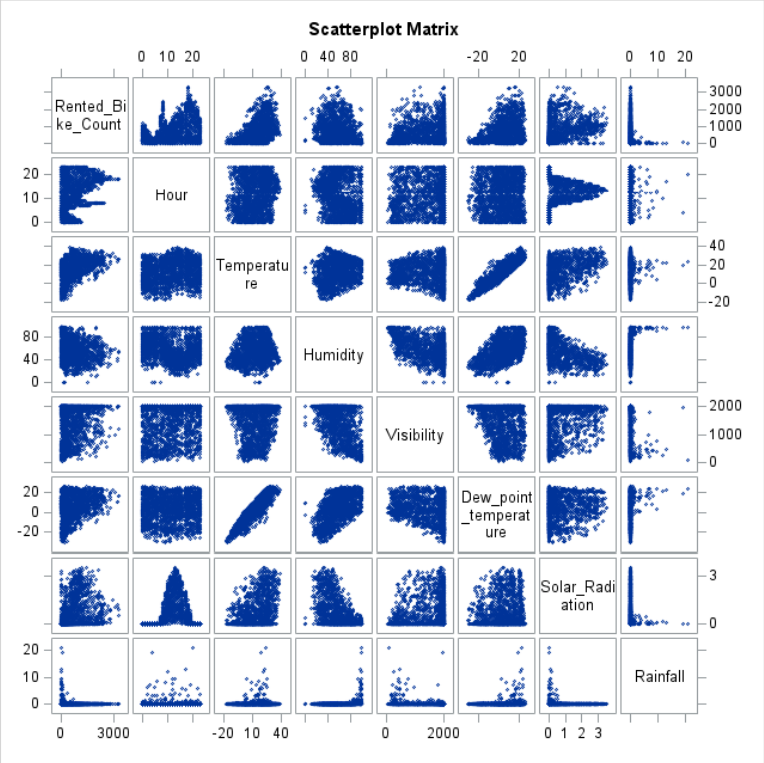


Figure B6: Pearson Correlation Coefficients

The CORR Procedure

8 Variables:	Rented_Bike_Count Hour Temperature Humidity Visibility Dew_point_temperature Solar_Radiation Rainfall

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Rented_Bike_Count	1752	710.66324	648.86087	1245082	0	3309
Hour	1752	11.46518	6.91657	20087	0	23.00000
Temperature	1752	12.55188	12.10403	21991	-17.50000	38.70000
Humidity	1752	57.06735	20.19061	99982	0	98.00000
Visibility	1752	1451	608.68586	2542129	38.00000	2000
Dew_point_temperature	1752	3.47534	13.11471	6089	-30.50000	26.10000
Solar_Radiation	1752	0.59174	0.88449	1037	0	3.52000
Rainfall	1752	0.14264	1.05904	249.90000	0	21.00000

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

	Rented_Bike_Count	Hour	Temperature	Humidity	Visibility	Dew_point_temperature	Solar_Radiation	Rainfall
Rented_Bike_Count	1.00000	0.41441 <.0001	0.54998 <.0001	-0.18135 <.0001	0.18860 <.0001	0.39937 <.0001	0.27439 <.0001	-0.13439 <.0001
Hour	0.41441 <.0001	1.00000	0.14636 <.0001	-0.22746 <.0001	0.12183 <.0001	0.02881 0.2281	0.15696 <.0001	-0.00041 0.9863
Temperature	0.54998 <.0001	0.14636 <.0001	1.00000	0.14620 <.0001	0.03900 0.1027	0.91317 <.0001	0.37636 <.0001	0.05323 0.0259
Humidity	-0.18135 <.0001	-0.22746 <.0001	0.14620 <.0001	1.00000	-0.53199 <.0001	0.52401 <.0001	-0.46452 <.0001	0.25467 <.0001
Visibility	0.18860 <.0001	0.12183 <.0001	0.03900 0.1027	-0.53199 <.0001	1.00000	-0.16619 <.0001	0.14457 <.0001	-0.17775 <.0001
Dew_point_temperature	0.39937 <.0001	0.02881 0.2281	0.91317 <.0001	0.52401 <.0001	-0.16619 <.0001	1.00000	0.11273 <.0001	0.13384 <.0001
Solar_Radiation	0.27439 <.0001	0.15696 <.0001	0.37636 <.0001	-0.46452 <.0001	0.14457 <.0001	0.11273 <.0001	1.00000	-0.07780 0.0011
Rainfall	-0.13439 <.0001	-0.00041 0.9863	0.05323 0.0259	0.25467 <.0001	-0.17775 <.0001	0.13384 <.0001	-0.07780 0.0011	1.00000

Appendix 3: Linear Regression Analysis

Figure C1: Influential points and VIFs

Multiple Regression and VIF

The REG Procedure
Model: MODEL1
Dependent Variable: Rented_Bike_Count

Number of Observations Read	8760
Number of Observations Used	8760

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	2004173836	154167218	822.28	<.0001
Error	8746	1639760527	187487		
Corrected Total	8759	3643934363			

Root MSE	432.99759	R-Square	0.5500
Dependent Mean	704.60205	Adj R-Sq	0.5493
Coeff Var	61.45279		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	10.08233	96.18448	0.10	0.9165	0
Hour	1	27.61298	0.73449	37.59	<.0001	1.20777
Temperature	1	16.59975	3.65933	4.54	<.0001	89.25760
Humidity	1	-10.43131	1.02192	-10.21	<.0001	20.22905
Wind_speed	1	19.31135	5.09731	3.79	0.0002	1.30358
Visibility	1	0.00970	0.00988	0.98	0.3265	1.68843
Dew_point_temperature	1	10.31851	3.82564	2.70	0.0070	116.62782
Solar_Radiation	1	-75.45111	7.56960	-9.97	<.0001	2.02029
Rainfall	1	-58.88760	4.26991	-13.79	<.0001	1.08415
Dum_Winter	1	-362.37915	19.67649	-18.42	<.0001	3.36061
Dum_Spring	1	-138.24854	13.84919	-9.98	<.0001	1.68945
Dum_Summer	1	-154.51586	17.21802	-8.97	<.0001	2.61134
IsHoliday	1	-119.35433	21.60543	-5.52	<.0001	1.02253
IsFunctionalDay	1	933.37414	26.66037	35.01	<.0001	1.08070

Figure C2: Influential points and VIFs

Number of Observations Read	8760
Number of Observations Used	8760

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	2003993317	166999443	890.73	<.0001
Error	8747	1639941045	187486		
Corrected Total	8759	3643934363			

Root MSE	432.99667	R-Square	0.5500
Dependent Mean	704.60205	Adj R-Sq	0.5493
Coeff Var	61.45265		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	43.19878	90.06877	0.48	0.6315	0
Hour	1	27.56318	0.73273	37.62	<.0001	1.20201
Temperature	1	16.39517	3.65337	4.49	<.0001	88.96787
Humidity	1	-10.69002	0.98732	-10.83	<.0001	18.88251
Wind_speed	1	19.78163	5.07472	3.90	<.0001	1.29205
Dew_point_temperature	1	10.54079	3.81892	2.76	0.0058	116.21890
Solar_Radiation	1	-76.94661	7.41456	-10.38	<.0001	1.93839
Rainfall	1	-59.02476	4.26761	-13.83	<.0001	1.08298
Dum_Winter	1	-365.88923	19.34855	-18.91	<.0001	3.24953
Dum_Spring	1	-141.26176	13.50441	-10.46	<.0001	1.60639
Dum_Summer	1	-153.61635	17.19356	-8.93	<.0001	2.60394
IsHoliday	1	-118.95628	21.60158	-5.51	<.0001	1.02217
IsFunctionalDay	1	932.98346	26.65734	35.00	<.0001	1.08046

Figure C3: Influential points and VIFs

Dependent Variable: Rented_Bike_Count

Number of Observations Read	8760
Number of Observations Used	8760

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	2002564965	182051360	970.28	<.0001
Error	8748	1641369397	187628		
Corrected Total	8759	3643934363			

Root MSE	433.16043	R-Square	0.5496
Dependent Mean	704.60205	Adj R-Sq	0.5490
Coeff Var	61.47590		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-184.69236	36.00604	-5.13	<.0001	0
Hour	1	27.48078	0.73240	37.52	<.0001	1.20001
Temperature	1	26.19223	0.86547	30.26	<.0001	4.98902
Humidity	1	-8.09285	0.29913	-27.05	<.0001	1.73196
Wind_speed	1	19.31556	5.07383	3.81	0.0001	1.29062
Solar_Radiation	1	-81.71191	7.21349	-11.33	<.0001	1.83330
Rainfall	1	-60.38737	4.24057	-14.24	<.0001	1.06849
Dum_Winter	1	-367.62035	19.34569	-19.00	<.0001	3.24612
Dum_Spring	1	-142.81440	13.49780	-10.58	<.0001	1.60360
Dum_Summer	1	-149.30870	17.12906	-8.72	<.0001	2.58249
IsHoliday	1	-118.23412	21.60816	-5.47	<.0001	1.02202
IsFunctionalDay	1	930.66254	26.65415	34.92	<.0001	1.07939

Figure C3.1: Training Dataset

Split

Obs	Selected	Date	Rented_Bike_Count	Hour	Temperature	Humidity	Wind_speed	Visibility	Dew_point_temperature	Solar_Radiation	Rainfall	Dum_Winter	Dum_Spring	Dum_Summer	IsHol
1	0	01/12/20	204	1	-5.5	38	0.8	2000	-17.6	0.00	0.0	1	0	0	
2	0	01/12/20	100	5	-6.4	37	1.5	2000	-18.7	0.00	0.0	1	0	0	
3	0	01/12/20	449	12	1.7	23	1.4	2000	-17.2	1.11	0.0	1	0	0	
4	0	01/12/20	463	15	2.1	36	3.2	2000	-11.4	0.54	0.0	1	0	0	
5	0	02/12/20	219	8	-4.2	79	2.1	1436	-7.3	0.01	0.0	1	0	0	
6	0	02/12/20	479	12	4.3	41	1.3	1666	-7.8	1.09	0.0	1	0	0	
7	0	02/12/20	385	19	5.0	52	2.3	1666	-4.0	0.00	0.0	1	0	0	
8	0	02/12/20	359	20	4.6	51	1.2	1585	-4.6	0.00	0.0	1	0	0	

Figure C3.2 : Testing Dataset

Split

Obs	Selected	Date	Rented_Bike_Count	Hour	Temperature	Humidity	Wind_speed	Visibility	Dew_point_temperature	Solar_Radiation	Rainfall	Dum_Winter	Dum_Spring	Dum_Summer	IsHol
1	1	01/12/20	254	0	-5.2	37	2.2	2000	-17.6	0.00	0.0	1	0	0	
2	0	01/12/20	204	1	-5.5	38	0.8	2000	-17.6	0.00	0.0	1	0	0	
3	1	01/12/20	173	2	-6.0	39	1.0	2000	-17.7	0.00	0.0	1	0	0	
4	1	01/12/20	107	3	-6.2	40	0.9	2000	-17.6	0.00	0.0	1	0	0	
5	1	01/12/20	78	4	-6.0	36	2.3	2000	-18.6	0.00	0.0	1	0	0	
6	0	01/12/20	100	5	-6.4	37	1.5	2000	-18.7	0.00	0.0	1	0	0	
7	1	01/12/20	181	6	-6.6	35	1.3	2000	-19.5	0.00	0.0	1	0	0	
8	1	01/12/20	460	7	-7.4	38	0.9	2000	-19.3	0.00	0.0	1	0	0	
9	1	01/12/20	930	8	-7.6	37	1.1	2000	-19.8	0.01	0.0	1	0	0	
10	1	01/12/20	490	9	-6.5	27	0.5	1928	-22.4	0.23	0.0	1	0	0	
11	1	01/12/20	339	10	-3.5	24	1.2	1996	-21.2	0.65	0.0	1	0	0	
12	1	01/12/20	360	11	-0.5	21	1.3	1936	-20.2	0.94	0.0	1	0	0	
13	0	01/12/20	449	12	1.7	23	1.4	2000	-17.2	1.11	0.0	1	0	0	
14	1	01/12/20	451	13	2.4	25	1.6	2000	-15.6	1.16	0.0	1	0	0	
15	1	01/12/20	447	14	3.0	26	2.0	2000	-14.6	1.01	0.0	1	0	0	
16	0	01/12/20	463	15	2.1	36	3.2	2000	-11.4	0.54	0.0	1	0	0	
17	1	01/12/20	484	16	1.2	54	4.2	793	-7.0	0.24	0.0	1	0	0	
18	1	01/12/20	555	17	0.8	58	1.6	2000	-6.5	0.08	0.0	1	0	0	

Figure C3.2 Influential Points and Outliers

Outliers & Influential Points

The REG Procedure
Model: MODEL1
Dependent Variable: Rented_Bike_Count

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	-2-1 0 1 2	Cook's D	RStudent	Hat Diag H	Cov Ratio	DFFITS	DFBETAS Intercept	DFBETAS Hour	DFBETAS Temperature	DFBETAS Wind_speed	DFBETAS Solar_Radiation	DFBETAS Rainfall	DFBETAS Dum_Winter	DFBETAS Dum_Spring
1	254	-180.5397	13.4329	434.5397	450.7	0.964	| |* |	0.000	0.9642	0.0009	1.0010	0.0287	0.0028	-0.0183	0.0011	0.0091	-0.0041	-0.0004	0.0096	-0.0011
2	204	-205.1310	12.7633	409.1310	450.7	0.908	| |* |	0.000	0.9077	0.0008	1.0010	0.0257	0.0056	-0.0117	-0.0010	-0.0067	0.0014	-0.0002	0.0097	0.0007
3	173	-178.2615	12.2161	351.2615	450.7	0.779	| |* |	0.000	0.7793	0.0007	1.0012	0.0211	0.0046	-0.0090	-0.0018	-0.0043	0.0009	-0.0002	0.0075	0.0004
4	107	-154.9306	12.0646	261.9306	450.7	0.581	| |* |	0.000	0.5811	0.0007	1.0016	0.0156	0.0036	-0.0056	-0.0018	-0.0042	0.0010	-0.0001	0.0054	0.0004
5	78	-69.6798	11.9508	147.6798	450.7	0.328	| | |	0.000	0.3276	0.0007	1.0018	0.0087	0.0008	-0.0040	-0.0006	0.0029	-0.0012	-0.0002	0.0026	-0.0004
6	100	-75.5552	11.0684	175.5552	450.8	0.389	| | |	0.000	0.3894	0.0006	1.0017	0.0096	0.0018	-0.0031	-0.0014	-0.0004	-0.0001	-0.0002	0.0031	-0.0001
7	181	-55.7315	10.9982	236.7315	450.8	0.525	| |* |	0.000	0.5252	0.0006	1.0015	0.0128	0.0026	-0.0030	-0.0024	-0.0021	0.0003	-0.0002	0.0040	0.0001
8	460	-56.8888	11.5896	516.8888	450.7	1.147	| |** |	0.000	1.1468	0.0007	1.0003	0.0295	0.0074	-0.0028	-0.0080	-0.0105	0.0033	-0.0003	0.0077	0.0007
9	930	-22.9057	11.2530	952.9057	450.8	2.114	| |**** |	0.000	2.1145	0.0006	0.9963	0.0528	0.0126	-0.0027	-0.0161	-0.0155	0.0051	-0.0007	0.0128	0.0008
10	490	16.0185	12.6105	473.9815	450.7	1.052	| |** |	0.000	1.0516	0.0008	1.0006	0.0294	0.0071	0.0020	-0.0079	-0.0165	0.0076	0.0000	0.0078	0.0013
11	339	147.3722	11.1447	191.6278	450.8	0.425	| | |	0.000	0.4251	0.0006	1.0016	0.0105	0.0015	0.0001	-0.0015	-0.0041	0.0036	0.0001	0.0041	0.0002
12	360	255.9869	11.5013	104.0131	450.7	0.231	| | |	0.000	0.2307	0.0007	1.0018	0.0059	0.0003	0.0001	0.0001	-0.0023	0.0023	0.0001	0.0030	0.0001
13	449	344.4148	11.9368	104.5852	450.7	0.232	| | |	0.000	0.2320	0.0007	1.0019	0.0061	-0.0001	0.0002	0.0009	-0.0022	0.0025	0.0002	0.0035	0.0001

Figure C4: Forward Method

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	10.08233	96.18448	2060.07453	0.01	0.9165
Hour	27.61298	0.73449	264990993	1413.38	<.0001
Temperature	16.59975	3.65933	3858087	20.58	<.0001
Humidity	-10.43131	1.02192	19535077	104.19	<.0001
Wind_speed	19.31135	5.09731	2691004	14.35	0.0002
Visibility	0.00970	0.00988	180518	0.96	0.3265
Dew_point_temperature	10.31851	3.82564	1363947	7.27	0.0070
Solar_Radiation	-75.45111	7.56960	18627602	99.35	<.0001
Rainfall	-58.88760	4.26991	35659988	190.20	<.0001
Dum_Winter	-362.37915	19.67649	63591960	339.18	<.0001
Dum_Spring	-138.24854	13.84919	18682867	99.65	<.0001
Dum_Summer	-154.51586	17.21802	15099092	80.53	<.0001
IsHoliday	-119.35433	21.60543	5721651	30.52	<.0001
IsFunctionalDay	933.37414	26.66037	229800030	1225.69	<.0001

Bounds on condition number: 116.63, 3161.4

All variables have been entered into the model.

Summary of Forward Selection

Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Temperature	1	0.2900	0.2900	5042.46	3577.99	<.0001
2	Hour	2	0.1198	0.4098	2716.48	1777.27	<.0001
3	IsFunctionalDay	3	0.0516	0.4614	1715.26	839.21	<.0001
4	Humidity	4	0.0408	0.5023	923.830	718.07	<.0001
5	Dum_Winter	5	0.0209	0.5232	519.049	384.26	<.0001

Figure C5: Stepwise Method

Dum_Winter	-365.88923	19.34855	67045810	357.60	<.0001
Dum_Spring	-141.26176	13.50441	20514783	109.42	<.0001
Dum_Summer	-153.61635	17.19356	14966229	79.83	<.0001
IsHoliday	-118.95628	21.60158	5685555	30.33	<.0001
IsFunctionalDay	932.98346	26.65734	229658917	1224.94	<.0001

Bounds on condition number: 116.22, 2869.8

All variables left in the model are significant at the 0.1500 level.

No other variable met the 0.1500 significance level for entry into the model.

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Temperature		1	0.2900	0.2900	5042.46	3577.99	<.0001
2	Hour		2	0.1198	0.4098	2716.48	1777.27	<.0001
3	IsFunctionalDay		3	0.0516	0.4614	1715.26	839.21	<.0001
4	Humidity		4	0.0408	0.5023	923.830	718.07	<.0001
5	Dum_Winter		5	0.0209	0.5232	519.049	384.26	<.0001
6	Rainfall		6	0.0113	0.5345	301.774	212.13	<.0001
7	Solar_Radiation		7	0.0065	0.5410	177.822	123.56	<.0001
8	Dum_Spring		8	0.0030	0.5440	121.596	57.49	<.0001
9	Dum_Summer		9	0.0033	0.5473	58.5302	64.71	<.0001
10	IsHoliday		10	0.0015	0.5488	31.0846	29.38	<.0001
11	Wind_speed		11	0.0007	0.5496	18.5812	14.49	0.0001
12	Dew_point_temperature		12	0.0004	0.5500	12.9628	7.62	0.0058

Figure C6: Backward Method

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	43.19878	90.06877	43128	0.23	0.6315
Hour	27.56318	0.73273	265302615	1415.05	<.0001
Temperature	16.39517	3.65337	3775835	20.14	<.0001
Humidity	-10.69002	0.98732	21979125	117.23	<.0001
Wind_speed	19.78163	5.07472	2848849	15.19	<.0001
Dew_point_temperature	10.54079	3.81892	1428352	7.62	0.0058
Solar_Radiation	-76.94661	7.41456	20191901	107.70	<.0001
Rainfall	-59.02476	4.26761	35864724	191.29	<.0001
Dum_Winter	-365.88923	19.34855	67045810	357.60	<.0001
Dum_Spring	-141.26176	13.50441	20514783	109.42	<.0001
Dum_Summer	-153.61635	17.19356	14966229	79.83	<.0001
IsHoliday	-118.95628	21.60158	5685555	30.33	<.0001
IsFunctionalDay	932.98346	26.65734	229658917	1224.94	<.0001

Bounds on condition number: 116.22, 2869.8

All variables left in the model are significant at the 0.1000 level.

Summary of Backward Elimination							
Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Visibility	12	0.0000	0.5500	12.9628	0.96	0.3265

Figure D1: Model 1 (Removing Outliers and Influential Points)

Outliers & Influential Points

The REG Procedure
Model: MODEL1
Dependent Variable: Rented_Bike_Count

Number of Observations Read	8623
Number of Observations Used	8623

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	1698276221	169827622	1017.82	<.0001
Error	8612	1436945941	166854		
Corrected Total	8622	3135222162			

Root MSE	408.47756	R-Square	0.5417
Dependent Mean	675.61498	Adj R-Sq	0.5411
Coeff Var	60.46011		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-726.25848	27.74669	-26.17	<.0001
Hour	1	30.64848	0.67856	45.17	<.0001
Temperature	1	22.26600	0.81699	27.25	<.0001
Wind_speed	1	28.93627	4.77977	6.05	<.0001
Solar_Radiation	1	17.86944	5.96752	2.99	0.0028
Rainfall	1	-93.12001	4.19980	-22.17	<.0001
Dum_Winter	1	-311.37417	18.29567	-17.02	<.0001
Dum_Spring	1	-157.12309	12.82473	-12.25	<.0001
Dum_Summer	1	-177.55102	16.34349	-10.86	<.0001
IsHoliday	1	-85.18805	20.37853	-4.18	<.0001
IsFunctionalDay	1	916.94745	25.19826	36.39	<.0001

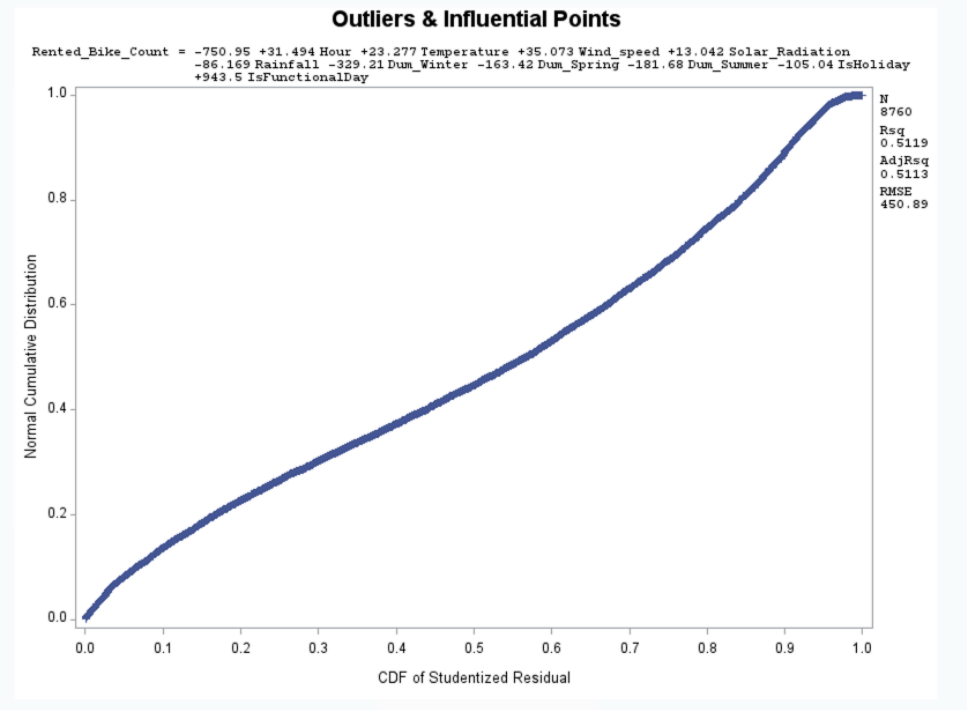


Figure D2: Model 2 (Removing some Outliers and Influential Points)

Outliers & Influential Points

The REG Procedure
Model: MODEL1
Dependent Variable: Rented_Bike_Count

Number of Observations Read	8684
Number of Observations Used	8684

Analysis of Variance

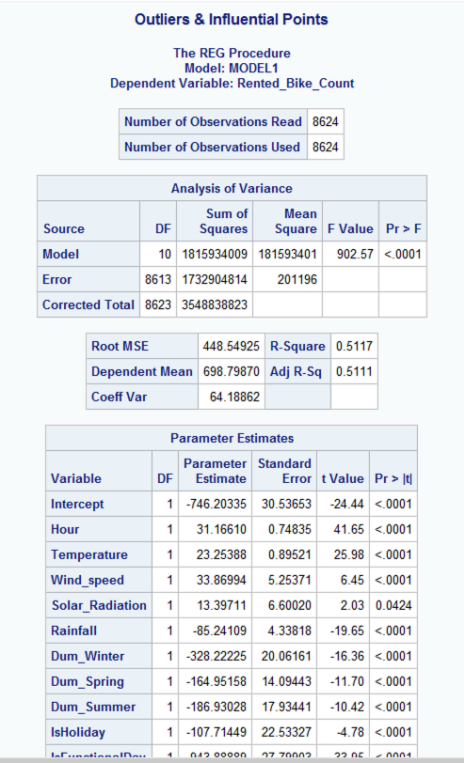
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	1837329893	183732989	912.35	<.0001
Error	8673	1746614394	201385		
Corrected Total	8683	3583944287			

Root MSE	448.75970	R-Square	0.5127
Dependent Mean	700.94784	Adj R-Sq	0.5121
Coeff Var	64.02184		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-747.17580	30.42453	-24.56	<.0001
Hour	1	31.31007	0.74552	42.00	<.0001
Temperature	1	23.30063	0.89264	26.10	<.0001
Wind_speed	1	35.03365	5.24007	6.69	<.0001
Solar_Radiation	1	12.17077	6.56051	1.86	0.0636
Rainfall	1	-86.07810	4.36043	-19.74	<.0001
Dum_Winter	1	-329.16564	20.03234	-16.43	<.0001
Dum_Spring	1	-167.43694	14.03220	-11.93	<.0001
Dum_Summer	1	-185.99878	17.82912	-10.43	<.0001
IsHoliday	1	-106.65362	22.46239	-4.75	<.0001
IsFunctionalDay	1	942.32927	27.66441	34.06	<.0001

Figure D3: Model 3 (Removing some more)outliers and Influential Points)



Outliers & Influential Points

The REG Procedure
Model: MODEL1
Dependent Variable: Rented_Bike_Count

Number of Observations Read	8624
Number of Observations Used	8624

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	1815934009	181593401	902.57	<.0001
Error	8613	1732904814	201196		
Corrected Total	8623	3548838823			

Root MSE	448.54925	R-Square	0.5117
Dependent Mean	698.79870	Adj R-Sq	0.5111
Coeff Var	64.18862		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-746.20335	30.53653	-24.44	<.0001
Hour	1	31.16610	0.74835	41.65	<.0001
Temperature	1	23.25388	0.89521	25.98	<.0001
Wind_speed	1	33.86994	5.25371	6.45	<.0001
Solar_Radiation	1	13.39711	6.60020	2.03	0.0424
Rainfall	1	-85.24109	4.33818	-19.65	<.0001
Dum_Winter	1	-328.22225	20.06161	-16.36	<.0001
Dum_Spring	1	-164.95158	14.09443	-11.70	<.0001
Dum_Summer	1	-186.93028	17.93441	-10.42	<.0001
IsHoliday	1	-107.71449	22.53327	-4.78	<.0001

Figure E1: Test Model

N	Mean	Std Dev	Sum	Minimum	Maximum	Label
388	5.67526	0.76601	2202	4.00000	8.00000	
388	5.66294	0.46930	2197	4.59114	6.97084	Predicted Value of new_y

quality	yhat
1.00000	0.66409 <.0001
0.66409 <.0001	1.00000

Figure D4: Model 4 (Removing some more)outliers and Influential Points)

Dependent Variable: Rented_Bike_Count

Number of Observations Read	8624
Number of Observations Used	8624

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	1815934009	181593401	902.57	<.0001
Error	8613	1732904814	201196		
Corrected Total	8623	3548838823			

Root MSE	448.54925	R-Square	0.5117
Dependent Mean	698.79870	Adj R-Sq	0.5111
Coeff Var	64.18862		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-746.20335	30.53653	-24.44	<.0001
Hour	1	31.16610	0.74835	41.65	<.0001
Temperature	1	23.25388	0.89521	25.98	<.0001
Wind_speed	1	33.86994	5.25371	6.45	<.0001
Solar_Radiation	1	13.39711	6.60020	2.03	0.0424
Rainfall	1	-85.24109	4.33818	-19.65	<.0001
Dum_Winter	1	-328.22225	20.06161	-16.36	<.0001
Dum_Spring	1	-164.95158	14.09443	-11.70	<.0001
Dum_Summer	1	-186.93028	17.93441	-10.42	<.0001
IsHoliday	1	-107.71449	22.53327	-4.78	<.0001
IsFunctionalDay	1	943.88889	27.79903	33.95	<.0001