A Peloton of Data

Predicting bike share demand based on weather and day information

FINAL PROJECT

DSC 323/423 DATA ANALYSIS & REGRESSION

Kristine Biedenstein · Sameer Ishaq · Robert Kaszubski

Daniel Kwan · Mauro Mujica-Parodi

Table of Contents

Introduction 4

Methodology 6

- Variables 6
- Data Cleansing 7
- Data Exploration 8
- Model Approach 11

Analysis, Results, & Findings 13

- Full Model 13
  - Model Assumptions 13
  - Outliers & Influential Points 14

Model Selection & Testing 15

Validation of Final Model 17

Predictions & Conclusions 20

Future Work 21

Research References 22

Appendix 23

- Methodology 23
- Analysis, Results, & Findings 31
- Model Selection & Testing 37
- Validation of Final Model 43
- Predictions & Conclusions 49

Introduction

Bike sharing is an alternative mode of transportation that has become more prominent over the past few years. Bike sharing systems (BSS) allow riders to rent a bicycle at the location of their choosing and then return the bicycle when they are done at any other location in the city. More and more urban areas are installing bike sharing stalls to allow for short term bicycle rentals. The benefits of bike sharing have been made abundantly clear. A study conducted in Beijing, China, another city that has widely adopted bike sharing found these programs to have "positive externalities on the economy, energy use, the environment, and public health" (Qiu L-Y) with an estimated increase in GDP by "RMB1.20 billion" and decreased health costs by about "RMB 2420.57 million yuan" (Qui L-Y). Not only does it offer a fun alternative while enhancing city mobility freeing up street congestion, but it is also pollution free, provides riders with a healthier lifestyle, and reduces the financial burden of other forms of transport. There is no commitment to purchase your own bicycle, protect your property from theft nor perform annual maintenance.

Bike sharing can be used for a variety of reasons on an as-needed basis. Many utilize it for recreational reasons, some for exercise, and others for their daily commutes. Some may use it every day, others only once when they're visiting from out of town. Bike sharing sees different demands on a daily basis. There are multiple factors that contribute to the number of bike rentals seen daily or even hourly. The companies and local governments operating these bike sharing programs need to adjust to meet those varying demands, particularly since program popularity tends to increase year after year (Hosford). One downside to bike sharing is arriving at a station and finding it empty, or full. "The size of the bike fleet, i.e., the number of bikes within the BSS has an impact on the availability of both bikes and free bike racks. If the BSS has too many bikes, bike racks will be blocked more often. If the BSS has too few bikes, bikes will be less available" (Brinkmann). We believe that by analyzing the circumstances surrounding bike sharing, we can figure out the factors that influence usage allowing for a relatively accurate prediction of the bikes required at a given time.

Our dataset includes weather and date/time information from Seoul, South Korea. We believe both of these will play a role in accurately predicting bike sharing usage. Our initial hypothesis points to temperature and hour of the day being key variables affecting bicycle rentals, with heavy expectations toward rainfall and snowfall as well. It's fairly reasonable to assume that more people will utilize bike sharing services on a relatively warm day, without any rain or snow rather than in the middle of a freezing

winter night. However, we want to see exactly how correlated these variables are, and be able to see the effects they have on one another. Ultimately our model will allow us, and others, to predict the number of bikes expected to be needed, thus improving the service. Specifically, through this analysis we hope to identify ideal situations, based on consistent recurring decreases in demand, for general bike maintenance. Our work here will be highly valuable to the many bike sharing companies not only in Seoul but in other cities around the world. It will allow them to best serve their customers and ensure bicycles are always around to provide a healthy and clean transportation alternative.

Methodology

Our dataset is titled Seoul Bike Sharing Demand and was sourced from the UCI Machine Learning Repository.

https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand

The UCI Machine Learning Repository offers a conglomeration of several different datasets of diverse topics tailored to various machine learning techniques such as classification, clustering, and regression.

Our dataset was created using government data from Seoul, South Korea and features 14 variables and 8,760 observations.

VARIABLES

- Date - in year-month-day format
- Rented Bike Count - the number of bikes rented at a given hour
- Hour - the hour of the day
- Temperature - the temperature at that hour in Celsius
- Humidity - the concentration of water vapor in the air, expressed as a percentage
- Wind Speed - in meters per second
- Visibility - how well one can see an object 10 miles away
- Dew Point Temperature - the temperature at which the air is saturated in moisture, expressed in Celsius
- Solar Radiation - expressed in MJ/m2 (megajoules per square meter)
- Rainfall - expressed in millimeters
- Snowfall - expressed in centimeters
- Seasons - Winter, Spring, Summer, Autumn
- Holiday - Holiday/No Holiday
- Functional Day - Yes/No (Refers to workdays)

While we didn't feel like the specific date was relevant to our analysis, there was a hypothesis that the day of the week may be a significant variable. As such, we enriched the data by standardizing the date format, created a new column, and then mapped each observation date to the associated day of week using excel. The result was the Variable "Day_of_Week" where: 1 → Monday, 2 → Tuesday, 3 → Wednesday, 4 → Thursday, 5 → Friday, 6 → Saturday, 7 → Sunday

DATA CLEANSING

The data was first inspected for any discrepancies such as missing values of which there were none. Due to the size of the dataset, random sampling with a specific seed ('123456789') was used to select only 3000 observations. This helped with running the dataset reliably on SAS using Remote Desktop. However, it must be noted that although we are using the same seed, results may slightly vary if the selection process chooses different observations. We could not simply select the first or last 3000 observations as then we would be stuck with data all from the same season!

Next the distribution of the dependent variable (y) *Rented_Bike_Count* was inspected. We found a right skewed distribution (M.a) that had to be transformed. The mean, 705.8, was far to the right of the median, 516.5. About 20% of the data was in the first bin alone, and over 50% in the first three bins. The best result came from applying a square root transformation. Applying a log transformation resulted in several undefined values due to some hours having zero bike rentals, therefore a log transformation could not be used. Applying square or cubic transformations did not alter the distribution either. By using square root transformation, we found a much closer to normal distribution, although one that was still not perfect (M.b). The mean and the median were now much closer together at 23.49 and 22.73. The other Quantiles were far closer to what is expected from a normal distribution as well. The newly created variable is labeled *sqrt_rented*.

Following this, dummy variables were created for the categorical variables in our dataset. Three dummy variables were created for seasons. One for winter, one for spring, and one for summer. Each of these is binary, with a 1 corresponding to the proper season for each observation. When all three dummies equal 0, then the season is autumn.

Holiday and *Functioning_Day* were likewise turned into dummy variables with 1 corresponding to a Holiday/Functioning Day and 0 corresponding to a Non-Holiday/Non-Functioning Day.

Finally, our created *Day_of_Week* variable required six dummy variables to be created. Dummies were created corresponding to Tuesday-Sunday while Monday became our base when all six variables were set to 0.

After the dummy variables were created, the old variables were dropped as we would not be requiring them to construct our models. *Date* and the *parsed_Date* were also

removed, as the first would not be used and the latter was used to create *Day_of_Week*. Some further tidying was done by renaming a few of the variables that SAS did not interpret well while importing the dataset. For instance, Wind Speed m/s was interpreted as *Wind_speed__m_s_*, so this was simply renamed to *Wind_Speed*. The same was done for several other variables. This was later removed upon noticing errors during the import process relating to the *Date* variable which was interpreted as mmddyyyy instead of ddmmyyyy. `Proc Import` was changed to `infile` followed by immediately specifying variable names rather than renaming them later. This prevented any import errors.

DATA EXPLORATION

The data exploration process already began by looking at the distribution of *Rented_Bike_Count* when dealing with transforming the variable (M.a). We saw that during most hours, bike rentals fell on the lesser side. 20% of the data fell in the first bin and the majority of the data fell in the first three which consisted of about 0 to 400 bike rentals. However, the remainder of the data saw a fairly slow decline and was fairly spread out ranging from 400 all the way to over 3000 bike rentals. Those days seeing over 2500 bike rentals are likely going to be outliers but there is still a lot of data represented in that margin from 400 to 2500 bike rentals.

We created a correlation table to see the association the different variables had to *sqrt_rented* (M.c). None of our independent variables had very strong or strong correlation to *sqrt_rented*. The closest came from *Temperature* with a correlation coefficient of 0.546 indicating moderate positive correlation which was not as strong as we had hypothesized. This indicates an expected trend of *sqrt_rented* increasing as the temperature increases. Next came *Hour* with a low to moderate correlation value of 0.403, again we had predicted hour being influential but its correlation value is lower than expected. *Dew_Point_Temperature* is at 0.376. The most noteworthy thing regarding *Dew_Point_Temperature* is that we may have a possible issue with multicollinearity as its association to temperature comes at a very strong 0.913. We address this after building our first model and finding the VIF and tolerance values. Solar radiation also shows some correlation at 0.315. The rest of the variables fall even closer to 0 indicating low to no correlation. *Rainfall*, *Snowfall*, and *Humidity* all show negative correlation, implying less bike sharing rentals with higher humidity, more rainfall, or more snowfall.

Box Plots (M.d) were created using *sqrt_rented* and the categorical variables in our dataset. These included holidays, functioning days, hours, and seasons. These offer some insight into the relationship between these variables and *sqrt_rented*.

Looking at the *d_func_day* vs *sqrt_rented* plot, we can see perhaps a lack of data for non-functional days as the entire boxplot is set at 0. This may also mean that any time we have a functional day, there are in fact 0 bike rentals. We'll have to investigate this further but it does give an early indication that functional day (*d_func_day*) likely won't be a significant predictor in our final model. For functional days, we see a mean and median around 25, indicating about 625 bike rentals on average when re-transformed.

The *d_holiday* vs *sqrt_rented* chart offers a bit more information. Overall, the two plots are similar but we do see fewer bike rentals on holidays than on non-holidays. Comparing the medians, on non-holidays it is around 24, while on holidays it is around 18. Re-transforming our values, this proves to be a fairly substantial difference of around 200 bikes. The means are a bit closer together but still distant enough to confirm this observation. This contradicts our expectations regarding holidays and perhaps implies that a significant portion of bike share rentals are for commuting purposes. One would think that bike sharing would increase on days most employees have off, but that is not the case.

Looking at the *hour* vs *sqrt_rented* plot, we can confirm our theory about bike sharing used heavily for commuting. We see relatively low numbers in the early morning hours. Impressively we can still see about 100 bikes rented on average at 4am. But this pales to the spike we can see at 8am with around 800 (~28 square) bikes rented, around the time one would commute to their 9-to-5 job, or to school. We see fairly consistent use from then until another massive spike around 6pm, when many are commuting home from work or school. "Commuter usage is denoted by a request peak in the morning, when users leave their homes and cycle to work, and by a request peak in the evening, when users cycle back home. The authors of all studies agree that commuter usage is common for most [Bike Sharing Systems] "(Brinkmann). We see a *sqrt_rented* value of around 37, which when squared transforms back to 1369 bikes rented on average. The median at that time was even higher at 40, or 1600 bikes rented. Interesting, we see only a gradual decline until midnight, implying heavy recreational usage after working hours. Looking at the boxplots we also see the wide range that can be expected at every hour. For instance, the IQR of 8am, is from about 15 to 40, anywhere from 225 to 1600 bikes. On some days, usage drops all the way to 0, or rises past 2000 bicycles. Therefore,

while hour does seem to give us some indication of how many bike rentals to expect, it doesn't tell the whole story and other variables need to be accounted for.

Looking at the *seasons* vs *sqrt_rented* plot, we don't see anything too surprising. Usage is at its highest in the summer, about the same in autumn and spring, and falls severely in the winter. "The amount of trips in the summer months is significantly higher than in the winter months due to weather conditions such as the temperature and precipitation (Brinkmann)." The average in the summer is just over *sqrt_rented* 30 or 900 bicycles. In the autumn it's about 25 or 626 bicycles, likewise in the spring except the median is slightly lower. In the winter, the average plummets to about 15, or 225 bicycles rented.

Finally, we don't see anything too significant on the day of the week plot. Every day sees rather consistent bike sharing usage with the exception of Day 7 (Sunday) which is noticeably lesser than the other days of the week. Saturday remains consistent so it isn't because Sunday is part of the weekend.

Scatterplots (M.e) were also created depicting the numerical variables versus *sqrt_rented*.

As *Temperature* had the highest correlation value in relation to *sqrt_rented*, it ended up having the clearest scatterplot. We see a faint positive trend upwards, the larger the temperature becomes. It isn't very clear because the data remains very scattered but looking at the darker shading from the overlap, we can see the trend and could fit a line. However, temperature is no guarantee of bicycle rentals. We can see that even at moderate/higher temperatures, there are still several instances of 0 rentals or very low rental numbers. Nonetheless temperature will be a key predictor in the final model.

The *sqrt_rented* vs Humidity plot shows no clear pattern of any sort; the data is scattered in practically every direction possible. We do see a large overlap in data where humidity is nearing 100 percent and bike rentals are very low.

Looking at *sqrt_rented* vs *Wind_Speed*, we again see no clear linear association between the two. *Wind_Speed* does not seem to affect the number of bicycles rented, and the data is very scattered.

Likewise, the *sqrt_rented* vs *Visibility* plot shows no clear pattern. We do see that the majority of days saw perfect visibility (2000), so likely this won't be a good predictor.

Dew_Point_Temperature is similar to *Temperature* as there is a faint linear trend. Lower temperatures have lower rental numbers, however higher temperatures have a mix of low rentals and high rentals, but the trend is higher rentals. This is likely due to the data

being hourly, so even on days that see very high rental numbers in the afternoon, they'll still see lower numbers at night despite the same temperature.

Solar_radiation is very similar to the *Visibility* plot although it does seem that days with hours with high solar radiation do tend to always have a high number of rentals. However, as most of the data is clumped at 0 solar radiation, there isn't a linear trend.

Looking at *sqrt_rented* vs *Rainfall*, we can see that days with no rainfall are the ones where peak bicycle rentals are reached. However, there are likewise many days with no rainfall that only have some or low rental usage. There isn't a linear trend here but we can at least see that higher rainfall leads to lower rentals. This is also true regarding the snowfall chart.

MODEL APPROACH

Our approach began with data exploration. Box plots and scatter plots were generated and examined, noting the relationship between the independent variables and sqrt_rented. Likewise, a correlation matrix was created. Keeping these findings in mind, an initial model was created with all of the variables included.

From there, we conducted analysis on the full model beginning with a review of a global F-test, the Adj R^2, and the RMSE for an assessment on overall model adequacy. We then removed insignificant variables that had a P-Value that were greater than 0.05. VIF statistics and Tolerance levels were computed and analyzed to ensure there were no issues with multicollinearity and we conducted an analysis to assess whether the four assumptions associated with a linear regression model, namely that of constant variance, linearity, independence, and normality, held. The last step prior to splitting the data into training and testing sets was to identify and remove outliers and influential data points.

Once the initial model was created, we split the data into training and testing sets, then ran multiple model selection techniques on the training section. After analyzing the results from Forward, Backward, Stepwise, CP, and ADJRSQ selections, we chose one CP model and one ADJRSQ model for additional testing on the remaining data. In the end, the predictive power of the two models were roughly equal, but ADJRSQ had fewer predictors and therefore was chosen as the final model to validate.

In order to confirm that the final model chosen is in fact the optimal representation of how many bikes are rented at any given time, we chose to repeat the model fit diagnostics that were performed at the time of the full regression model. Specifically checking for outliers in the dataset, multicollinearity between variables, and reviewing the residual and normal probability plots. With the verification of each of these steps, we can be confident that our final model will make accurate predictions in the future.

Analysis, Results, & Findings

FULL MODEL

The following is an analysis of the full model (with all independent variables) to predict *sqrt_rented*. In analyzing the global F-test for overall model adequacy we can conclude:

> H_0: $B_k = 0$ (None of the *X* variables included in the model have any association with Price)
>
> H_a: $B_j \neq 0$ (At least one of the *X* variables included in the model has a significant effect on changes in Price)
>
> F Value = 285.56 with a P Value less than 0.0001. As such, we can reject H_0 and conclude that there is at least one *X* variable included in the model that has a significant effect on changes in *sqrt_rented*

The model has an Adjusted R^2 of 0.6549, which is quite good and indicates the model already explains ~65% of the variance, but it also has RMSE of 7.29698 indicating there are some rather large estimation errors (A.a). However, as there are insignificant independent variables, as discussed below, the model performance will likely be improved upon them being removed.

The full model had several insignificant values (A.b) as their respective P-Value is > 0.05: *Wind_Speed*, *Visibility*, *Snowfall*, *d_thursday*, and *d_saturday*. As such, these independent variables will be removed from the model and rerun.

After computing and analyzing the VIF statistics (A.b), There looks to be several issues of multicollinearity given as the following independent variables have a VIF > 10 or Tolerance < 0.1: *Temperature* (VIF), *Humidity* (VIF), and *Dew_Point_Temperature* (VIF and TOL). As *Dew_Point_Temperature* had the highest value, it was removed and the model was rerun (A.c), resulting in no independent variables indicating issues of multicollinearity.

Model Assumptions

Constant Variance (A.d)

- Based on the residuals plotted against X-variables, of the relevant VAR's residuals, *Hour* and *Humidity*, look to have homoscedasticity, although *Solar_Radiation* only looks to have a slight funnel to its shape and therefore may not be an issue. *Temperature* residuals look to increase in error towards middle values and decrease for small and large values. *Rainfall* residuals have a funnel

shape. As such, the Constant Variance assumption does not hold. This is an issue as violations of this assumption may make it difficult to gauge the true standard deviation of the forecast errors and could result in confidence intervals that are too wide or too narrow.

- Based on the residuals plotted against predicted values, it looks like the Constant Variance assumption is violated as the residuals do not seem random.

Linearity (A.e)

- Based on the scatterplots, *sqrt_rented* against *Hour*, *Temperature*, *Humidity*, *Solar_Radiation* and *Rainfall*, only *Temperature* and *Solar_Radiation* indicate a somewhat linear pattern. With *Hour* the association between it and the dependent variable is negative, positive, and then negative. With *Humidity*, the association looks to be positive and then negative. With *Rainfall*, the association is extremely negative for low values and then levels off as the values increase. As such, linearity assumption does seem violated.

Independence (A.d)

- Based on the residuals plotted against X-variables, of the relevant VAR's residuals, *Temperature, Humidity*, and *Solar_Radiation* looks to be randomly plotted. Neither *Hour* nor *Rainfall* look to be randomly plotted.

- Based on the residuals plotted against predicted values, it looks like the Independence assumption is violated as the residuals do not seem random.

Normality (A.f)

- Based on the NPP plot, it looks like the Normality assumption holds as it closely mirrors a 45° line

Outliers & Influential Points

The following is an analysis of outliers and influential points (A.g).

Influential & Outlier Observations: 2256

Influential Observations: 838, 1405, 1406, 1418, 1419, 1420, 1422, 1423, 1424, 1425, 1463, 1468, 1612, 1613, 1716, 2261, 2567, 2568, 2818,

Outlier Observations: None

Given the above, we removed all observations indicated above. After doing so, no new Influential or Outlier Observations were identified. Adjusted R^2 improved from the original 0.6549 to 0.6739 and the original RMSE from 7.29698 to 7.07711 (A.h).

MODEL SELECTION & TESTING

After processing our data and fitting our first full model, we proceeded with model selection. We began by splitting 60% of our dataset into a training section, with the remaining 40% reserved for testing. We were able to use this ratio since we had more than 30 observations for each of our 20 predictors.

Once the data were split, we ran five different model selection methods: Forward, Backward, Stepwise, Mallow's CP, and Adjusted R^2. We decided against using CV PRESS since we had a suitable number of observations on which to base our analysis. The Forward and Backward methods produced models with similar overall health: Comparable Adjusted R^2 values of 0.6584 and 0.6580, respectively, with only a 0.005 difference in Root MSE (S.a). The Backward model, however, had a superior F-Value (287 vs 247) and fewer predictors, leading us to discard the Forward model as an option.

Stepwise selection yielded somewhat better results than Forward (fewer predictors and 283 F-Value), but the Adjusted R^2 was somewhat lower at 0.655, and was therefore disregarded (S.b). Additionally, these three models all included predictors not significant at the 0.05 level, leading to further scrutiny.

Mallow's CP selection returned many results, but only two were appropriate to test, the rest having C(p) values too divergent from DF + 1 to warrant consideration (S.c). The first model on the list below performed better, due to its lower number of predictors and better F-Value (there being very little difference in Adjusted R^2 between the two) (S.d). In addition, the second model included two insignificant predictors, whereas the first only included one.

ADJRSQ selection also returned many options, however these models were all very similar in both Adjusted R^2 and number of predictors. For further testing, we selected only models with 11 or fewer predictions from the top 15 models offered, since we saw very little Adjusted R^2 drop-off in the list. This led us to assume the dummy variables for individual days of the week were not offering much predictive power.

Next, we examined four ADJRSQ models in more detail, and selected the one with 10 predictors (Model 4) for further testing (S.e). Its overall metrics of Root MSE and Adjusted R^2 were extremely close in performance to the other three, but its F-Value was significantly higher (342.85, versus the next highest of 313.05), all its predictors were significant, and it had the advantage of the fewest overall predictors of any comparable model.

After analyzing all models produced with different model selection methods, we selected the best CP model and the best ADJRSQ model to compare to one another using the unseen test data. The CP model offered the highest Adjusted R^2 while still offering a good F-Value and a better CP value compared to the Backward selection model. ADJRSQ was our strongest overall candidate, with the fewest predictors and highest F-Value, and just small changes to Root MSE and Adjusted R^2.

Running validation for the ADJRSQ model showed an improvement on the test data, with Root MSE decreasing from 7.26 to 6.85 (S.f). Adjusted R^2 also improved from 0.6567 to 0.6943, which was a positive sign. Validation for the CP model showed similar changes: Root MSE decreased from 7.24 to 6.85, and Adjusted R^2 improved from 0.6583 to 0.6946. Adjusted R^2 values were calculated by squaring the correlation between *yhat* and *sqrt_rented*.

Both models ended up improving on the test data, which is a sign of strength. In the end, though the CP model had a slight advantage in Root MSE and Adjusted R^2, we selected the ADJRSQ model for final validation, since the CP model still had an insignificant predictor, and the ADJRSQ model had fewer overall predictors, with 10 compared to 13.

VALIDATION OF FINAL MODEL

Now that we have used various selection methods to find the optimal variables to put into our final regression model, we must look back at our dataset to make any necessary adjustments and ensure that the model can reach its full potential accuracy when it is used for computing predictions. The starting values of each diagnostic statistic are:

F-Value [P-Value]	6.1112 [< 0.0001]	
Root MSE	7.09885	
R2	0.6730 or 67.30%	
Adj-R2	0.6719 or 67.19%	(V.a)

The first test performed on the Bike sharing dataset was to produce a graphical representation of the studentized residuals and Cook's distance of every single observation. As was done earlier in the model construction, every point was checked for a residual with an absolute value larger than three and marked as an outlier. Three such points were found, 1640, 1646, and 2798. On the other side, the Cook's distance chart was also used again to find influential points using the result of 4/n that in this case came to be 0.001. Twelve observations had a distance greater than 0.001 and were flagged as influential. Those data points were 836, 1373, 1740, 1761, 1826, 2113, 2240, 2435, 2550, 2797, 2798, and 2799. An abbreviated Studentized Residuals and Cook's Distance graph can be found in the appendix at V.b. It has been edited to only include the observations mentioned above. In only one case, observation 2798, on the visual was a point marked as both influential and an outlier. However, upon further inspection of the other two outliers and their corresponding segments of the Cook's distance graph, we found that both had a Cook's distance of 0.002. For unknown reasons the illustrations in the graph did not convey that these observations are indeed influential. Additionally, two of the influential points, 2797 and 2799, were found to have studentized residuals greater than three but were also lacking the correct image. In total, there were five observations that were outliers in the dataset while also being influential to the final model; 1640, 1646, 2797, 2798, and 2799. All five of these points were removed.

Before the last set of outliers was removed the diagnostic statistics of our final model were adequate, yet not without room for improvement (V.a). Afterwards there were a number of positive changes. The root mean square error decreased by 0.06323. R^2 and Adjusted R^2 both increased by 0.0042. Finally, the F-Values increased by 10.67 while the P-Values remained less than 0.0001. All of these indicate a more accurate model, and can be seen in the appendix (V.c).

The next step taken in pursuit of the best possible model was checking again for any multicollinearity between the final set of variables. To locate any of these pairs both the tolerance and the variance inflation was calculated. Likely collinearity can be spotted by a tolerance smaller than 0.1 or a variance inflation greater than ten. As can be seen in the Parameter Estimate table (V.d), there was no variable that exceeded either of those conditions and therefore no issues of collinearity.

We now know that all the observations fit within, and all the independent variables contribute separately to the final model. We can now check the assumptions of constant variance, independence and linearity for each variable along with the model itself. We can also check the assumption that the model can follow a normal distribution. In our final regression model we have ten independent variables. Of those, five contain numerical data and five contain binary data. To check the assumptions, we plotted the studentized residuals against each of those variables, along with the predicted values for the square root of rented bikes. The least useful of these visuals were the binary, which clumped results on only the extremes of the X-axis. Regardless, the plots of that data will be included in the appendix (V.e).

The variables that contain numerical values were more useful in revealing the overall health of the model. As can be seen in the additional plots (V.f), these contain points all along the X-axis and are able to be assessed visually. In our opinion, the distribution of hour, temperature and humidity all fit within the archetype of points randomly scattered around the central line. Hence the assumptions of constant variance, independence, and linearity hold in all those cases. The plot on the bottom left of appendix V.f shows the spread of the residuals of solar radiation. Even though there is a concentrated area of points nearing 0.0 megajoules per square meter, the overall shape is still consistent enough to confirm all the assumptions. The last graph shows the studentized residuals versus millimeters of rainfall. We concluded that this plot represents a variable that does not confirm anything. Given that we have already removed outliers from the data set, all the observations on the plot must be valid. Therefore, the odd decreasing and increasing shape of the spread indicates a transformation of the variable containing rainfall may be beneficial.

A new dataset was created to investigate transformations performed on the values of rainfall. First it was determined that neither a logarithm or an inverse would be helpful in this situation as the amount of rain on a given day could potentially be none, and those calculations are invalid when the value given is zero. Furthermore, any exponential transformations seemed to increase the distances between the points of the residual plot in an unhelpful manner. For these reasons, a square root was used.

Another model was computed, leaving us with a new residual plot of the square root of rainfall, shown in V.g in comparison with the original. This transformation changed the shape of the spread to be closer to randomly scattered around the zero line than it did initially.

This new iteration of the final regression model produced another positive jump in diagnostic statistics and can be seen in the appendix at V.h. This time the root mean square error decreased by 0.17933, the R^2 and Adjusted R^2 increased by 0.0162 and 0.0163 respectively, and the P-Value once again held under 0.0001 while the F-Value increased by 48.66. Those are all good indicators but the residuals must be consulted again to corroborate the efficacy of the square root of rainfall.

Something that was checked earlier but not shown until now is the Studentized Residuals versus Predicted Values and Normal Probability plots from the second running of the regression model. We thought that while they both upheld the assumptions they were intended to, they could possibly be improved further. When shown alongside the same plots from this third running of the model, the similarities and differences are much easier to spot (V.i). For reference: the model corresponding to the top graphs is before the transformation, and the model on the bottom is after. On the right one can see that the normal probability in both cases is an identical diagonal line across the graph, which is exactly the distribution needed to confirm normality. However, on the right there is some variation in the plots. Our assessment is that the slightly stretched shape of the second plot is a better representation of a linear model with constant variation.

In consideration of all the validation checks that have been executed, we have determined that our final linear regression model contains the variables containing values of hour, temperature, humidity, solar radiation, the square root of the rainfall collected, along with the binary variables containing the answers to the questions of if it is winter, spring, summer, a functioning day, or Sunday, and can be used to calculate the square root of the number of rented bikes. The final fit of the model and the final diagnostic values are as follows:

```
F-Value [P-Value]      6.7045 [< 0.0001]
Root MSE               6.85629
R2                     0.6934 or 69.34%
Adj-R2                 0.6924 or 69.24%          (V.j)
```

PREDICTIONS

For our first prediction we created a typical pleasant day in Seoul: 12:00PM, 23°C (73.4°F), 34% Humidity, 2.4 MJ/m2 of Solar Radiation, with no rain, in the summer on a work day. Looking at other observations in our dataset, we can see these conditions usually lead to moderate-to-high bike rentals, so we expect something similar from our prediction. After running the calculations, the model predicts an average day will have between 301 – 415 bike rentals (post transformation) 95% of the time (P.a). The prediction interval for a specific day with these characteristics is between 28 and 1,050. If we compare our actual predicted *sqrt_rented* value of 18.876 to other observations, this appears to be a reasonable prediction.

For our second prediction, we did the opposite and picked a winter day in Seoul at 8:00AM, -7.4°C (18.68°F), 40% Humidity, 1.1 MJ/m2 of Solar Radiation, with no rain, in the winter on Sunday. We thought we might examine the conditions for someone who wakes up early and does any outdoor activities. Temperature, Day of the Week and Season are the main factors in this prediction and immediately we see the number of predicted bike rentals has taken a significant dip to -1156.9317 - -47.5975 (P.b). However, for the number of bikes rented, any negative values would in actuality be 0, since you can't have negative bike rentals. It's shown that regardless of this specific day chosen, we can see that there are still bike rentals happening during 'Obs #6-9'. Comparing our first prediction to our second it's evident that Temperature & Season had a significant impact, being as those two variables were the only ones to change significantly. Even though statistically it states there will be no bike rentals on this Sunday, this isn't necessarily true. This goes to show the statistical limitations of a model versus what could happen in reality.

CONCLUSION

$$\begin{aligned} sqrt\_rented = \; & -5.461 + 0.537\ Hour + 0.466\ Temperature - 0.119\ Humidity \\ & - 0.652\ Solar\_Radiation - 9.233\ sqrt\_rain - 7.735\ d\_winter \\ & - 2.506\ d\_spring - 2.644\ d\_summer + 29.284\ d\_func\_day \\ & - 2.613\ d\_sunday \end{aligned}$$

Where:

```
d_winter=(Seasons="Winter");
d_spring=(Seasons="Spring");
d_summer=(Seasons="Summer");
d_holiday=(Holiday="Holiday");
d_func_day=(Functioning_Day="Yes");
d_tuesday=(Day_of_Week="2");
d_wednesday=(Day_of_Week="3");
d_thursday=(Day_of_Week="4");
d_friday=(Day_of_Week="5");
d_saturday=(Day_of_Week="6");
d_sunday=(Day_of_Week="7");
```

With our predictions finished, we can examine the final model in full. While *Hour* was important, it turns out that *Temperature* has the largest effect once sorted by standardized coefficients (P.x). The fact that this relationship is positive indicates people do not enjoy riding bikes in hotter weather, which makes intuitive sense. Predictably, *sqrt_rain* and *Humidity* have negative relationships instead—biking is less pleasant in the rain or high humidity. However, these are somewhat lesser effects than *Temperature*.

The second largest effect was the dummy variable *d_func_day*, which is whether or not it was a workday. This indicates a large portion of bike demand is commuters, which is something we had anticipated. A workday in Seoul increases rentals by over 850 per hour (post transformation). This number is even more striking than it first appears since the effect is averaged over the course of an entire day; actual commute times must see extreme spikes.

Seasonal variation in bike rentals was much stronger for winter (-0.269) than spring (-0.088) or summer (-0.094), but all three have a negative relationship compared to the fall. These effects are likely to be region-specific, but make sense for Seoul, which features muggy, wet summers and somewhat cooler springs (Weather Spark).

d_sunday is tougher to explain. It could be due to the country's growing share of Christians (Connor), or because Korea's demanding work culture (DHR International)

necessitates working on Saturdays. In either case, Sundays see a moderate drop of 6.83 fewer bikes rented per hour (post transformation).

Finally, *Solar_Radiation* has a small (-0.047) but statistically significant negative effect on rentals. Solar radiation is a measure of how much sunlight is striking a given area, and it could be that increased glare and warmer-feeling temperatures are, like higher temperatures, deterrents to would-be bikers.

Overall, our model has improved from 0.6549 Adjusted R^2 in the full model to 0.6924 in the final, and our Root MSE has dropped from 7.297 to 6.856—both positive indicators. The F-Value has also increased from 285.56 to 670.45, a more than 100% improvement. Through our efforts we have likely reached the limits of predictive power from this dataset for our given approach, and research shows there are other factors that also influence bike sharing numbers, such as station location and number of bikes available. We discuss future avenues of investigation below.

Future Work

There are several other avenues worth exploring. We completed our analysis with a limited number of observations and with a focus on weather and time. Running our model with our full dataset or an even larger dataset containing several years' worth of data, may have improved our analysis. Additionally, we would like to look at other factors unrelated to weather and time, that may have a strong link to bike sharing usage--such as station location and number of bikes available (*Science Daily*).

Looking at our initial exploratory work, we saw immediately a lack of clear linear association between most of our variables. There are several other datasets that offer other variables which may have a clearer connection. "Bike Share Research (BSR) aims to facilitate the curation of BSS data through a collaborative and open data platform while making it API accessible (*Open Bike Share Data*). This is a website we found during our research, that offers up to date bike sharing data from hundreds of bike sharing services located around the world. These datasets can be sourced using the services API and contain data for each individual bike sharing station. This in itself would open up several possibilities for further analysis as we could compare different stations of the same bike sharing network unlike our present dataset which only showcases overall figures.

Another potential avenue to explore would be utilizing several bike sharing datasets from different cities located in vastly differing climates. It's important to remember that our model was created using data from Seoul. It heavily focuses on the weather, and the weather isn't going to be the same in other cities thereby limiting our model.

Finally, we found research that argued there were weaknesses associated with symmetric modeling using continuous variables and null hypothesis statistical testing (NHST) and proposed use of somewhat precise outcome testing (SPOT) procedures for more effective data analysis (Woodside). Regardless of the specific argument as stated in the research, further analysis would be focused on testing competing algorithms, such as a random forest or an XGBoost algorithm, to improve model performance.

Research References

Brinkmann J. (2020) Bike Sharing Systems. In: Active Balancing of Bike Sharing Systems. Lecture Notes in Mobility. Springer, Cham. https://doi-org.ezproxy.depaul.edu/10.1007/978-3-030-35012-3_2

DHR International. Korean Productivity: Work & Life Balance. *DHR International*. Retrieved July 14, 2021 from https://www.dhrinternational.com/insights/korean-productivity-work-life-balance/

Hosford, K., Winters, M., Gauvin, L. et al. Evaluating the impact of implementing public bicycle share programs on cycling: the International Bikeshare Impacts on Cycling and Collisions Study (IBICCS). *Int J Behav Nutr Phys Act* 16, 107 (2019). https://doi.org/10.1186/s12966-019-0871-9

Institute for Operations Research and the Management Sciences. (2020, January 6). Maximizing bike-share ridership: New research says it's all about location: 10% increase in bike-availability levels increases ridership by 12%. *ScienceDaily*. Retrieved July 14, 2021 from www.sciencedaily.com/releases/2020/01/200106141616.htm

Open Bike Share Data [β]. Bike Share Research. (n.d.). https://bikeshare-research.org/

Connor, P. 6 facts about South Korea's growing Christian population. *Pew Research Center*. Retrieved July 14, 2021 from https://www.pewresearch.org/fact-tank/2014/08/12/6-facts-about-christianity-in-south-korea/

Qiu L-Y, He L-Y. Bike Sharing and the Economy, the Environment, and Health-Related Externalities. *Sustainability*. 2018; 10(4):1145. https://doi.org/10.3390/su10041145

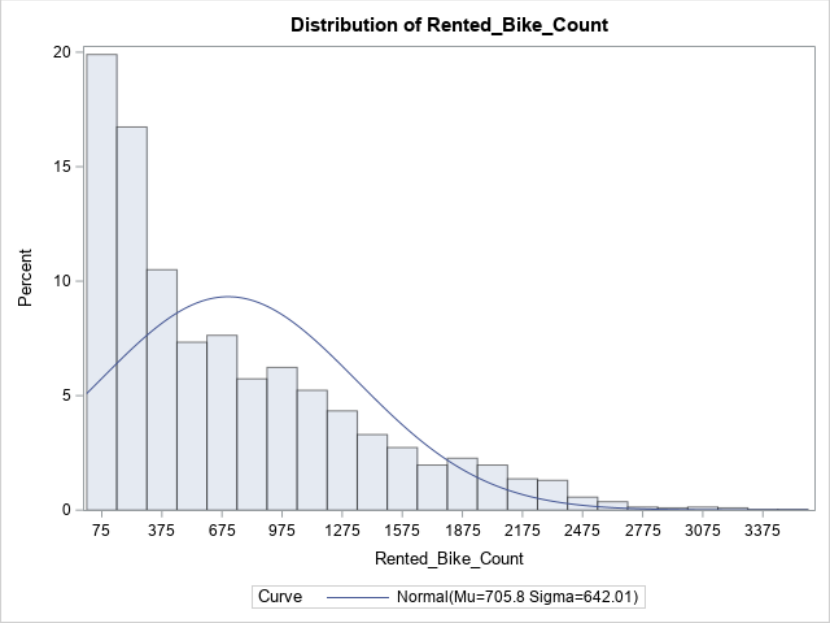
Weather Spark. Average Weather in Seoul. Retrieved July 14, 2021 from https://weatherspark.com/y/142033/Average-Weather-in-Seoul-South-Korea-Year-Round

Woodside, Arch. (2017). Releasing the death-grip of null hypothesis statistical testing (p < .05): Applying complexity theory and somewhat precise outcome testing (SPOT). Journal of Global Scholars of Marketing Science. 27. 1-15. 10.1080/21639159.2016.1265323. http://dx.doi.org/10.1080/21639159.2016.1265323

Appendix

METHODOLOGY

M.a



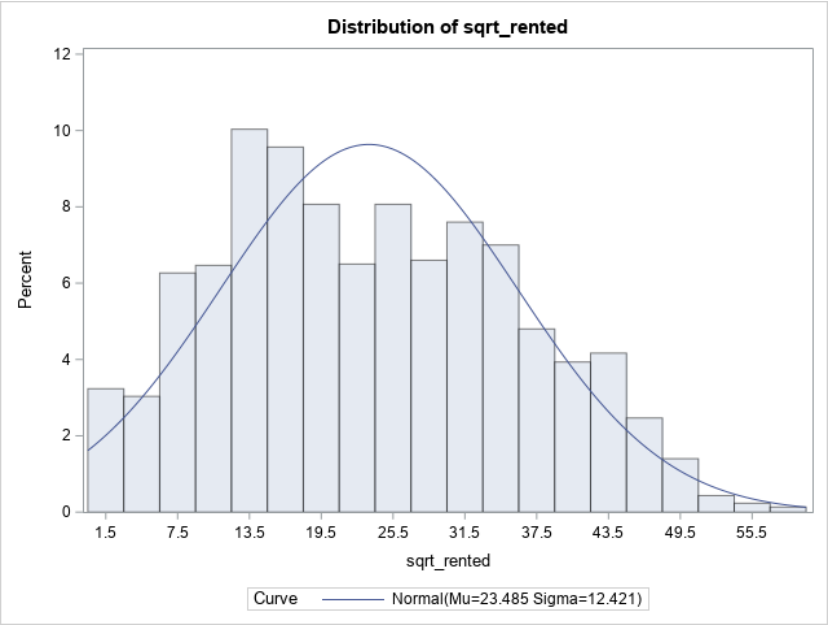
The UNIVARIATE Procedure
Fitted Normal Distribution for Rented_Bike_Count

Parameters for Normal Distribution		
Parameter	Symbol	Estimate
Mean	Mu	705.803
Std Dev	Sigma	642.0111

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.135805	Pr > D	<0.010
Cramer-von Mises	W-Sq	17.136729	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	104.670883	Pr > A-Sq	<0.005

Quantiles for Normal Distribution		
	Quantile	
Percent	Observed	Estimated
1.0	0.00	-787.738
5.0	25.50	-350.211
10.0	63.00	-116.967
25.0	189.00	272.773
50.0	516.50	705.803
75.0	1076.50	1138.833
90.0	1666.50	1528.573
95.0	2009.50	1761.817
99.0	2506.50	2199.344

M.b



Distribution of sqrt_rented

The UNIVARIATE Procedure
Fitted Normal Distribution for sqrt_rented

Parameters for Normal Distribution		
Parameter	Symbol	Estimate
Mean	Mu	23.48543
Std Dev	Sigma	12.42131

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.0595562	Pr > D	<0.010
Cramer-von Mises	W-Sq	2.3962747	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	14.2717875	Pr > A-Sq	<0.005

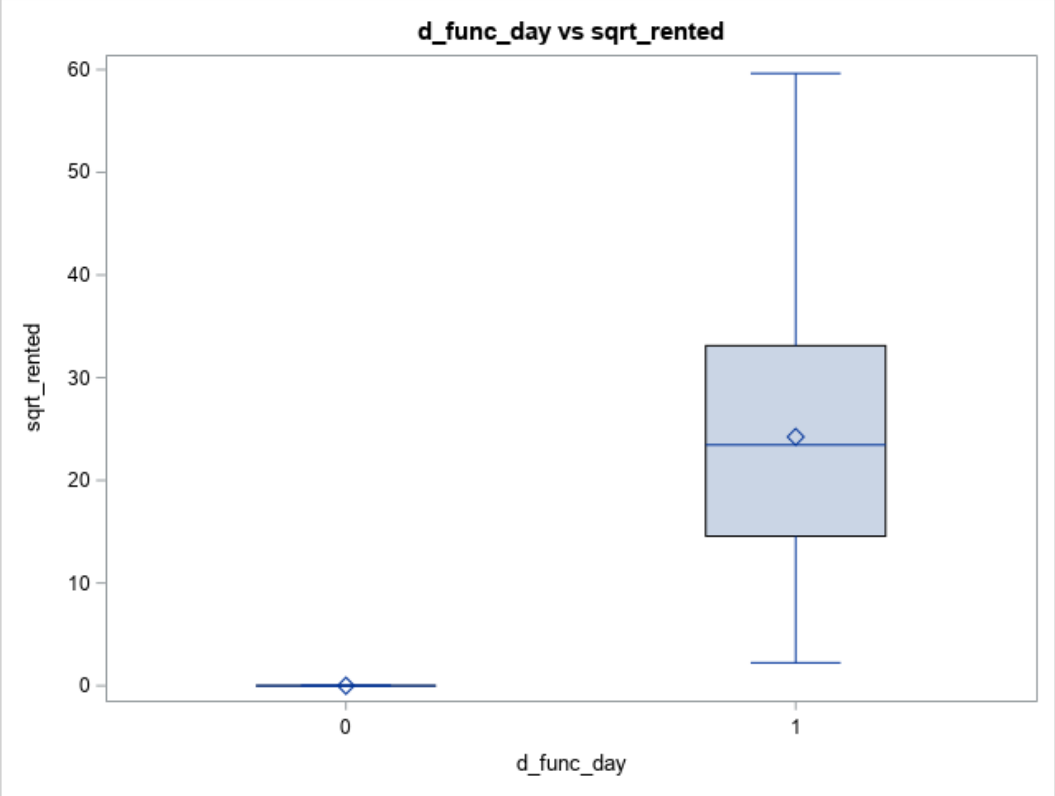
Quantiles for Normal Distribution		
	Quantile	
Percent	Observed	Estimated
1.0	0.0000	-5.41086
5.0	5.0495	3.05419
10.0	7.9373	7.56688
25.0	13.7477	15.10738
50.0	22.7266	23.48543
75.0	32.8101	31.86348
90.0	40.8228	39.40398
95.0	44.8274	43.91667
99.0	50.0650	52.36172

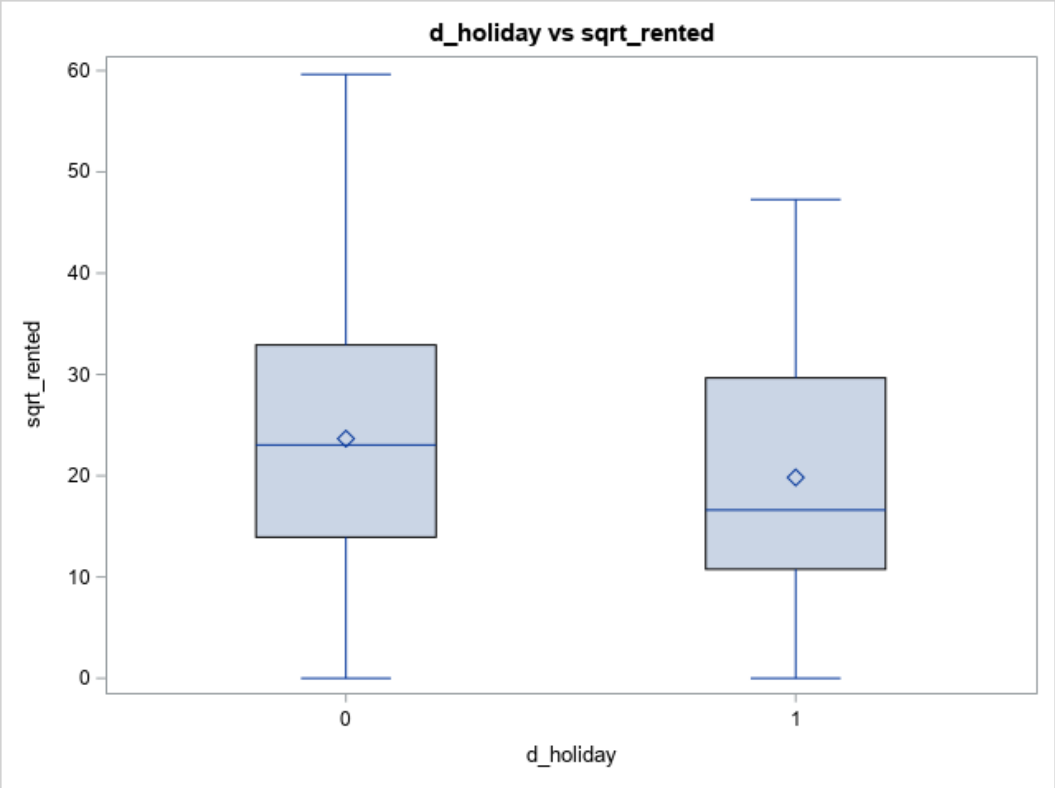
M.c

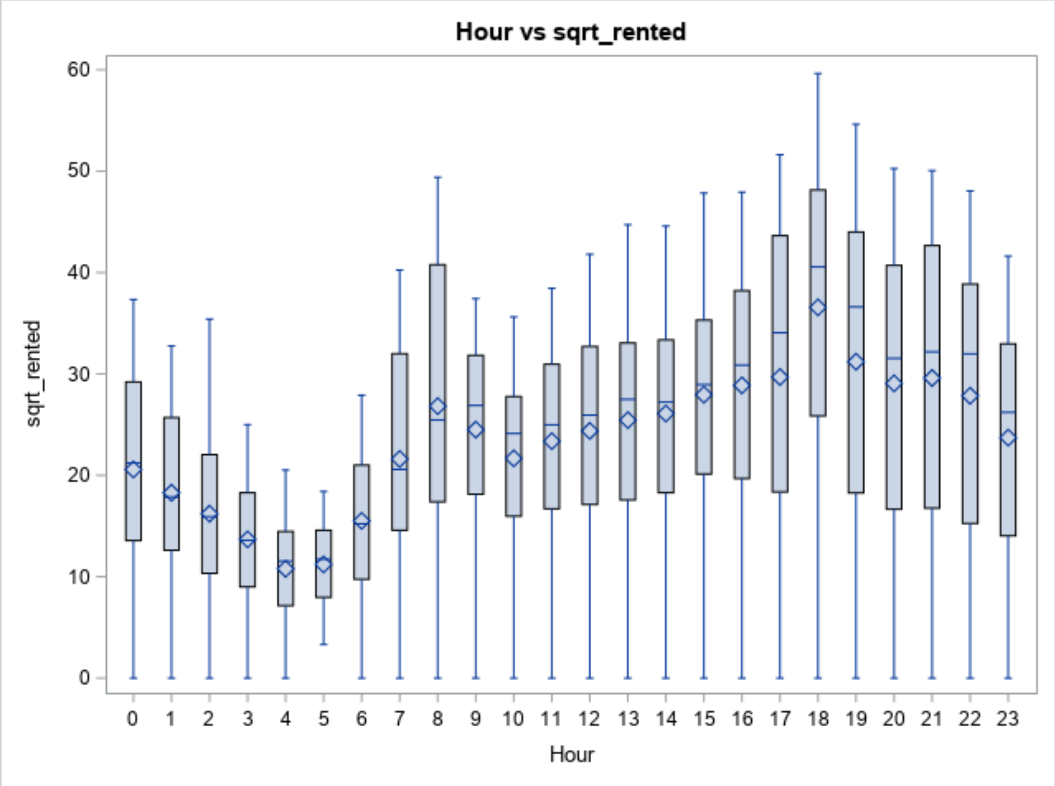
Pearson Correlation Coefficients, N = 3000
Prob > |r| under H0: Rho=0

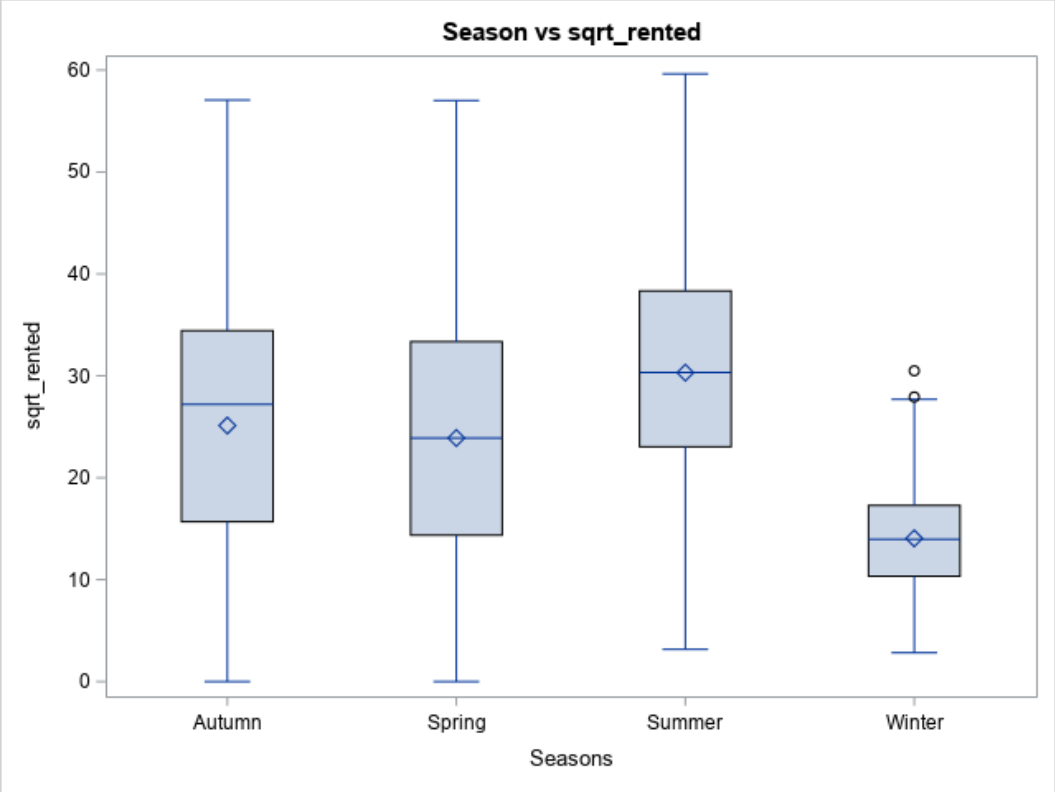
	sqrt_rented	Hour	Temperature	Humidity	Wind_Speed	Visibility	Dew_Point_Temperature	Solar_Radiation	Rainfall	Snowfall
sqrt_rented	1.00000	0.40280 <.0001	0.54558 <.0001	-0.23326 <.0001	0.13113 <.0001	0.21244 <.0001	0.37622 <.0001	0.31522 <.0001	-0.17079 <.0001	-0.16577 <.0001
Hour	0.40280 <.0001	1.00000	0.15104 <.0001	-0.21750 <.0001	0.29846 <.0001	0.08861 <.0001	0.03533 0.0530	0.15188 <.0001	0.01718 0.3468	-0.03266 0.0737
Temperature	0.54558 <.0001	0.15104 <.0001	1.00000	0.14329 <.0001	-0.01904 0.2972	0.04333 0.0176	0.91267 <.0001	0.36479 <.0001	0.05916 0.0012	-0.21914 <.0001
Humidity	-0.23326 <.0001	-0.21750 <.0001	0.14329 <.0001	1.00000	-0.32501 <.0001	-0.54740 <.0001	0.52165 <.0001	-0.46593 <.0001	0.25172 <.0001	0.13623 <.0001
Wind_Speed	0.13113 <.0001	0.29846 <.0001	-0.01904 0.2972	-0.32501 <.0001	1.00000	0.15883 <.0001	-0.15846 <.0001	0.34434 <.0001	-0.03731 0.0410	-0.00756 0.6791
Visibility	0.21244 <.0001	0.08861 <.0001	0.04333 0.0176	-0.54740 <.0001	0.15883 <.0001	1.00000	-0.17040 <.0001	0.16364 <.0001	-0.19149 <.0001	-0.14447 <.0001
Dew_Point_Temperature	0.37622 <.0001	0.03533 0.0530	0.91267 <.0001	0.52165 <.0001	-0.15846 <.0001	-0.17040 <.0001	1.00000	0.10273 <.0001	0.13945 <.0001	-0.14404 <.0001
Solar_Radiation	0.31522 <.0001	0.15188 <.0001	0.36479 <.0001	-0.46593 <.0001	0.34434 <.0001	0.16364 <.0001	0.10273 <.0001	1.00000	-0.07770 <.0001	-0.07167 <.0001
Rainfall	-0.17079 <.0001	0.01718 0.3468	0.05916 0.0012	0.25172 <.0001	-0.03731 0.0410	-0.19149 <.0001	0.13945 <.0001	-0.07770 <.0001	1.00000	-0.00816 0.6552
Snowfall	-0.16577 <.0001	-0.03266 0.0737	-0.21914 <.0001	0.13623 <.0001	-0.00756 0.6791	-0.14447 <.0001	-0.14404 <.0001	-0.07167 <.0001	-0.00816 0.6552	1.00000

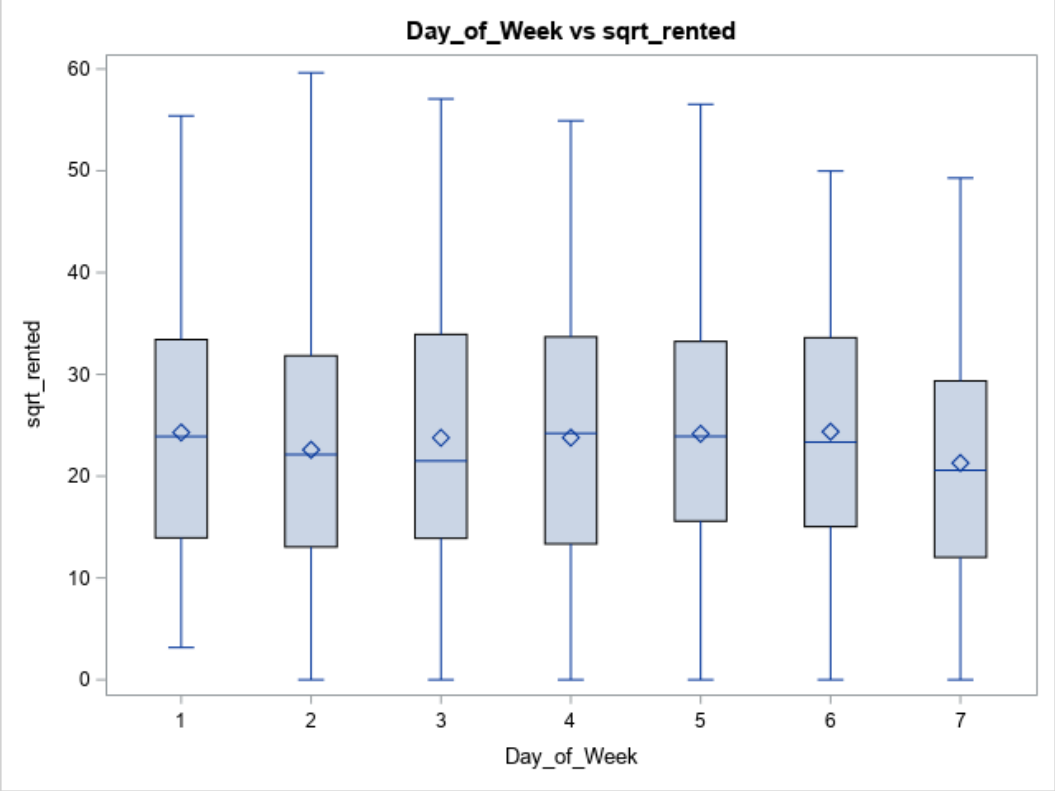
M.d



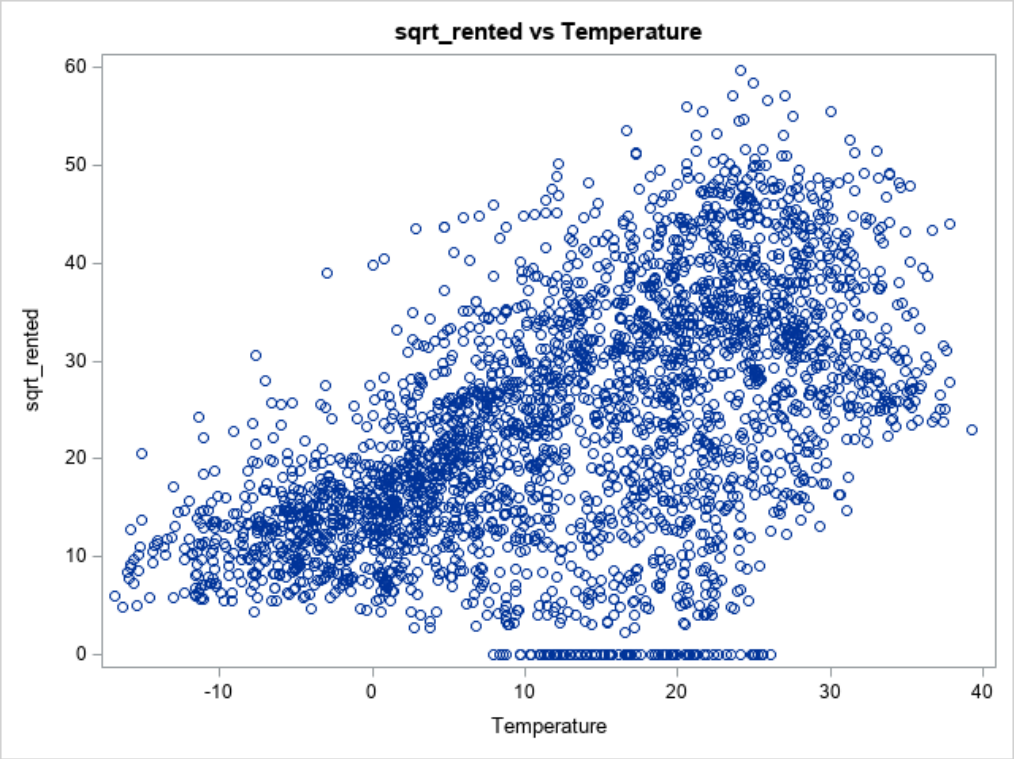

d_holiday vs sqrt_rented
sqrt_rented
60
50
40
30
20
10
0
0
1
d_holiday

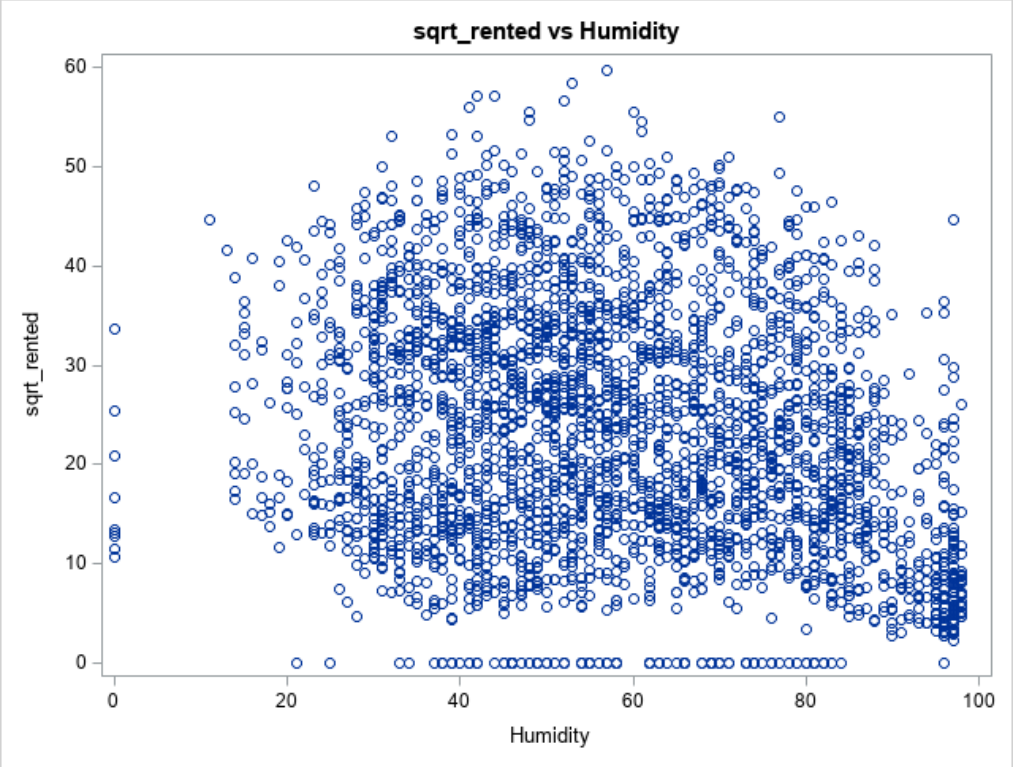

Hour vs sqrt_rented
sqrt_rented
60
50
40
30
20
10
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Hour

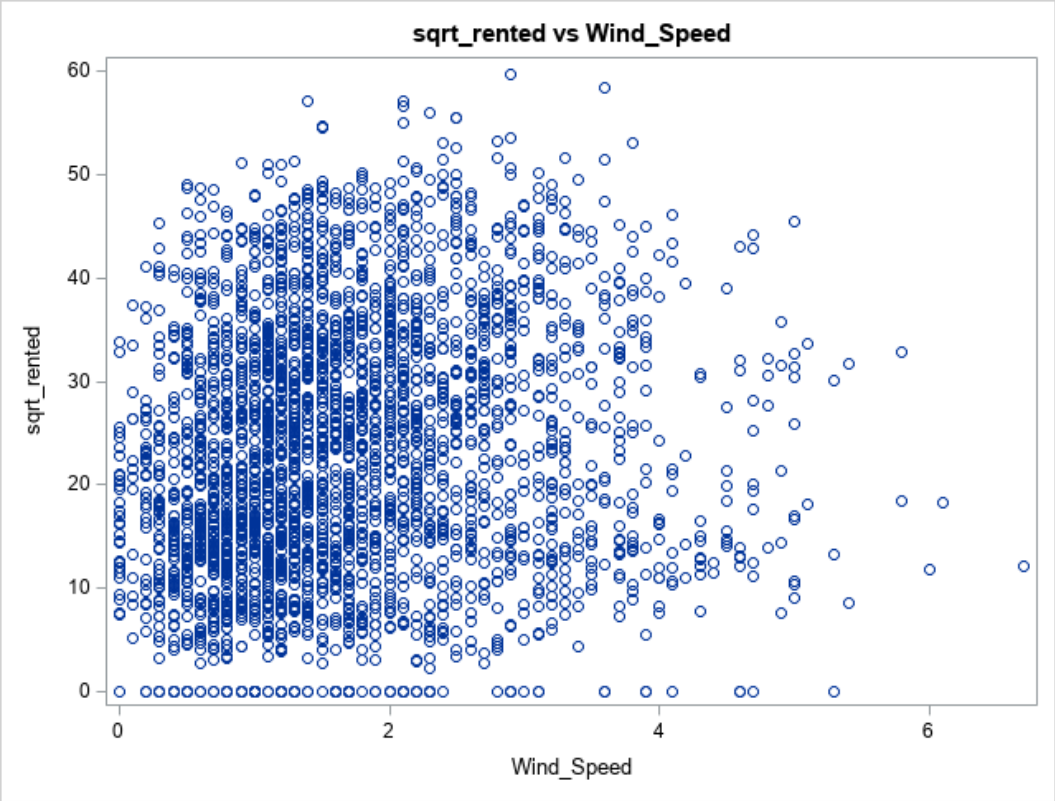

Season vs sqrt_rented
sqrt_rented
60
50
40
30
20
10
0
Autumn
Spring
Summer
Winter
Seasons

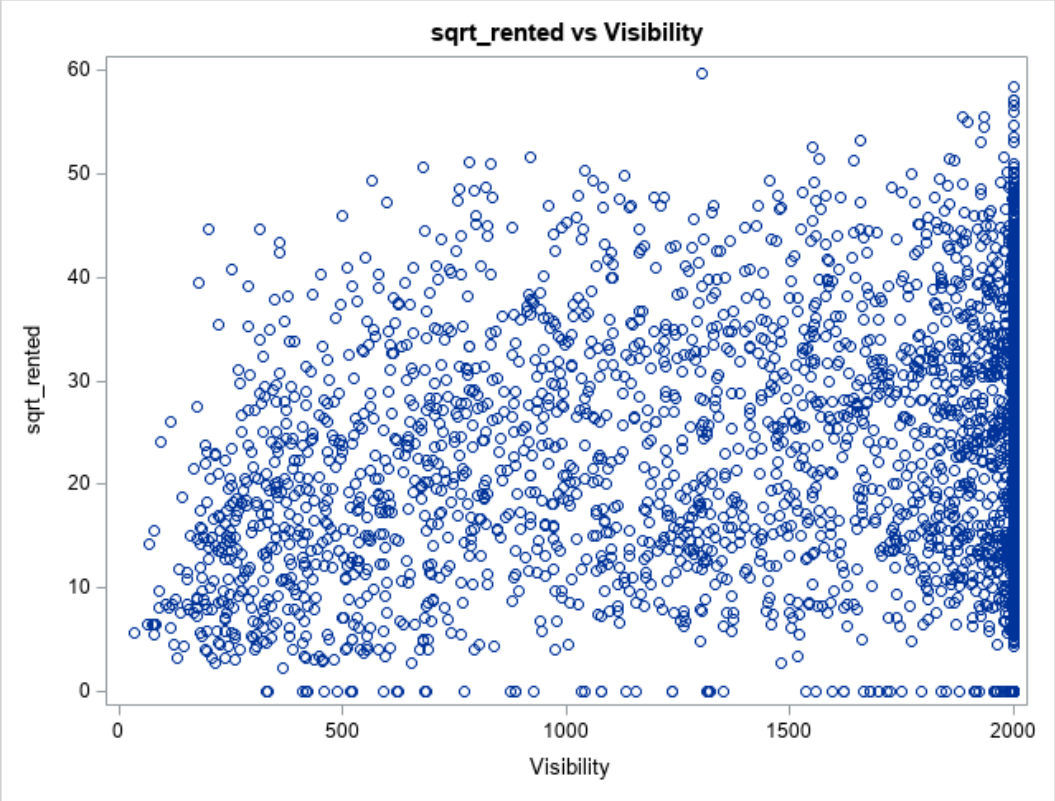

Day_of_Week vs sqrt_rented
sqrt_rented
60
50
40
30
20
10
0
1
2
3
4
5
6
7
Day_of_Week

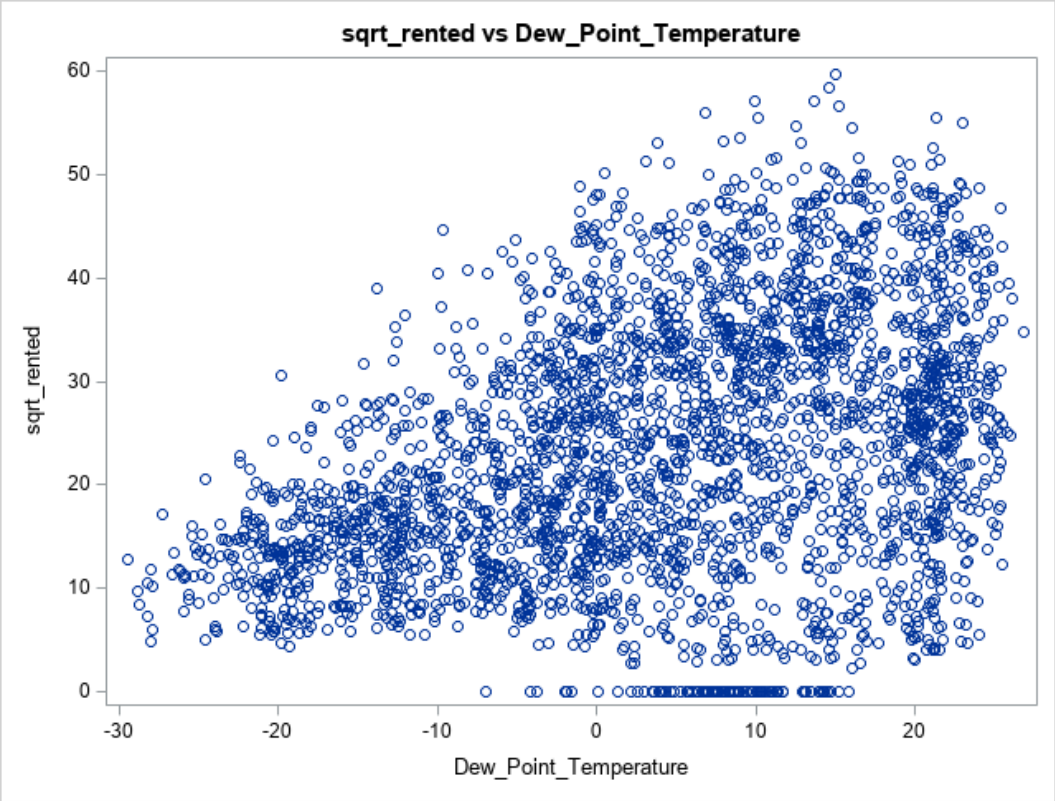
M.e

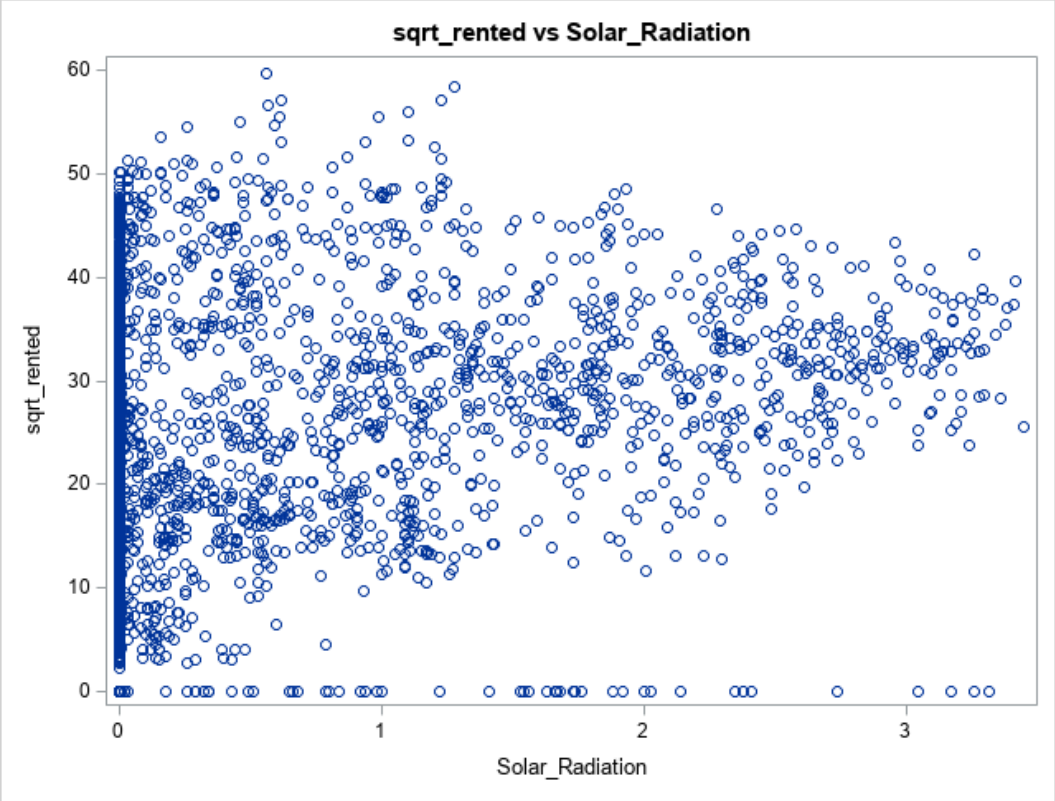

sqrt_rented vs Temperature
sqrt_rented
Temperature
60
50
40
30
20
10
0
-10
0
10
20
30
40

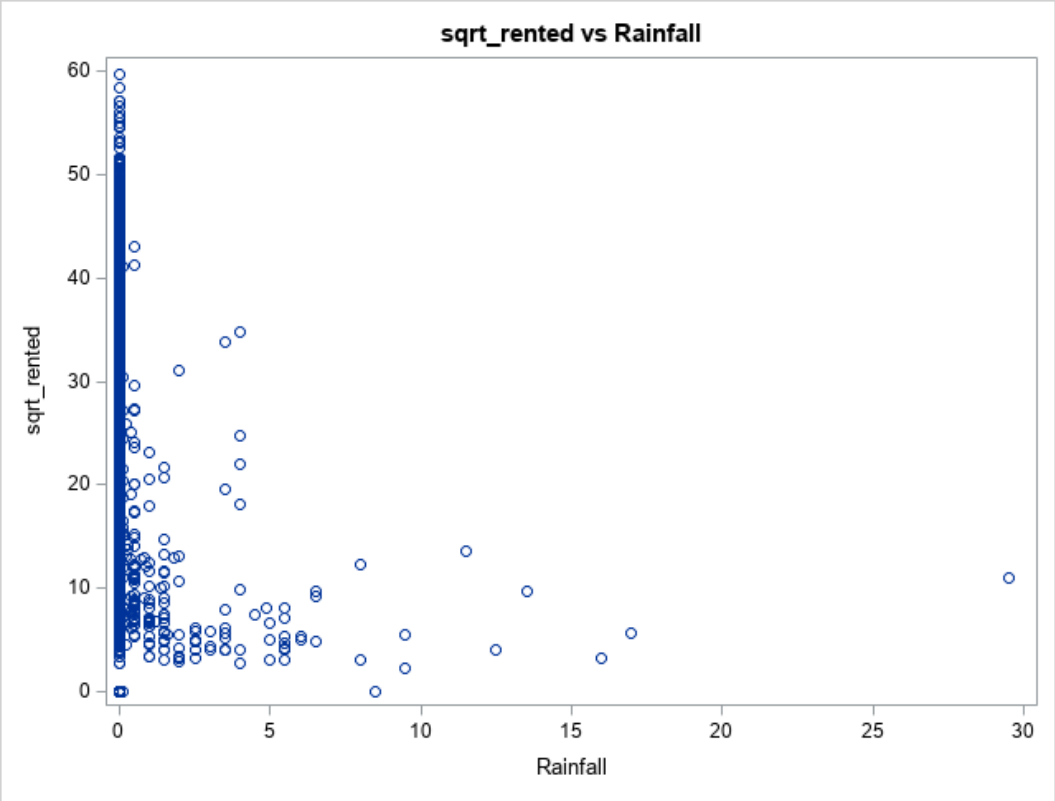

sqrt_rented vs Humidity
sqrt_rented
Humidity
60
50
40
30
20
10
0
0
20
40
60
80
100

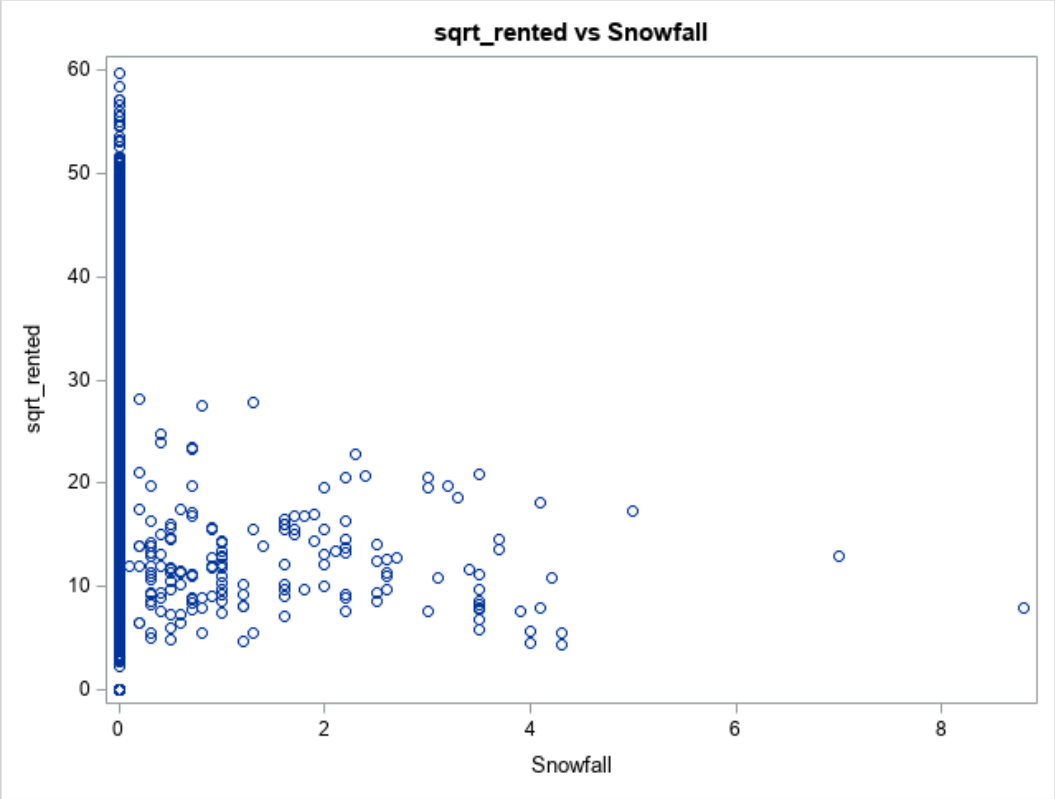

sqrt_rented vs Wind_Speed
sqrt_rented
Wind_Speed
0
10
20
30
40
50
60
0
2
4
6


sqrt_rented vs Visibility
sqrt_rented
Visibility
0
10
20
30
40
50
60
0
500
1000
1500
2000


sqrt_rented vs Dew_Point_Temperature
sqrt_rented
Dew_Point_Temperature
60
50
40
30
20
10
0
-30
-20
-10
0
10
20


sqrt_rented vs Solar_Radiation
sqrt_rented
Solar_Radiation
60
50
40
30
20
10
0
0
1
2
3


sqrt_rented vs Rainfall
sqrt_rented
Rainfall
0
10
20
30
40
50
60
0
5
10
15
20
25
30


sqrt_rented vs Snowfall
sqrt_rented
Snowfall
0
10
20
30
40
50
60
0
2
4
6
8

A.a

Analysis of Variance Output from full regression model

The REG Procedure
Model: MODEL1
Dependent Variable: sqrt_rented

Number of Observations Read	3000
Number of Observations Used	3000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	20	304093	15205	285.56	<.0001
Error	2979	158620	53.24595		
Corrected Total	2999	462713			

Root MSE	7.29698	R-Square	0.6572
Dependent Mean	23.48543	Adj R-Sq	0.6549
Coeff Var	31.07025		

A.b

Parameter Estimate table with Tolerance and Variance Inflation included

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Tolerance	Variance Inflation
Intercept	1	1.82044	2.62016	0.69	0.4872	0	.	0
Hour	1	0.50689	0.02127	23.84	<.0001	0.28160	0.82440	1.21300
Temperature	1	0.21968	0.09984	2.20	0.0279	0.21284	0.01230	81.31502
Humidity	1	-0.21971	0.02763	-7.95	<.0001	-0.36070	0.05593	17.87970
Wind_Speed	1	0.14159	0.14817	0.96	0.3394	0.01180	0.75414	1.32602
Visibility	1	0.00003206	0.00028333	0.11	0.9099	0.00158	0.58773	1.70147
Dew_Point_Temperature	1	0.26512	0.10398	2.55	0.0108	0.27834	0.00966	103.56237
Solar_Radiation	1	-0.60704	0.21957	-2.76	0.0057	-0.04300	0.47569	2.10220
Rainfall	1	-1.86596	0.13619	-13.70	<.0001	-0.15407	0.90999	1.09892
Snowfall	1	0.01838	0.29485	0.06	0.9503	0.00071471	0.87505	1.14279
d_winter	1	-7.65917	0.57231	-13.38	<.0001	-0.26401	0.29568	3.38204
d_spring	1	-2.97981	0.40028	-7.44	<.0001	-0.10485	0.58013	1.72376
d_summer	1	-2.59853	0.49709	-5.23	<.0001	-0.09189	0.37240	2.68528
d_holiday	1	-2.80496	0.65890	-4.26	<.0001	-0.04649	0.96491	1.03637
d_func_day	1	28.79770	0.80093	35.96	<.0001	0.40189	0.92107	1.08570
d_tuesday	1	1.21401	0.51059	2.38	0.0175	0.03330	0.58667	1.70454
d_wednesday	1	1.56170	0.49749	3.14	0.0017	0.04461	0.56978	1.75505

A.c

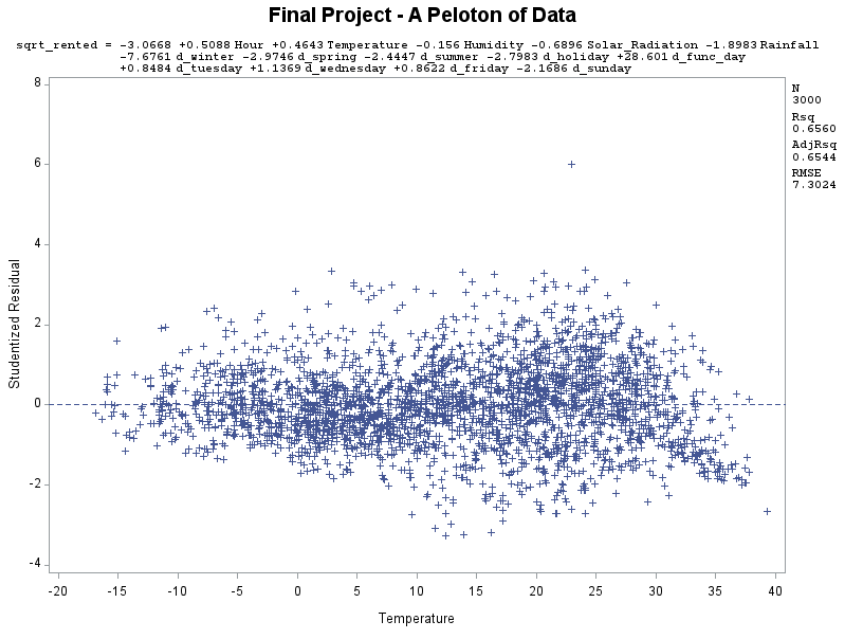
Parameter Estimate table with Tolerance and Variance Inflation included after removing independent variable driving multicollinearity issues

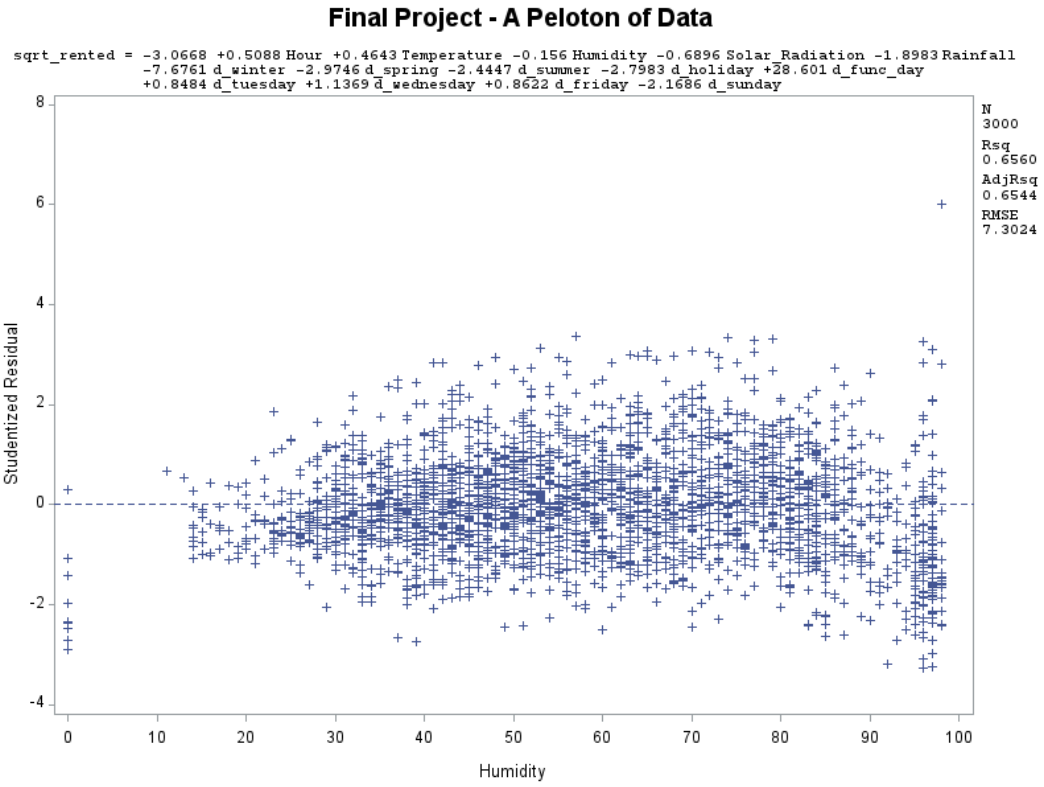
Parameter Estimates

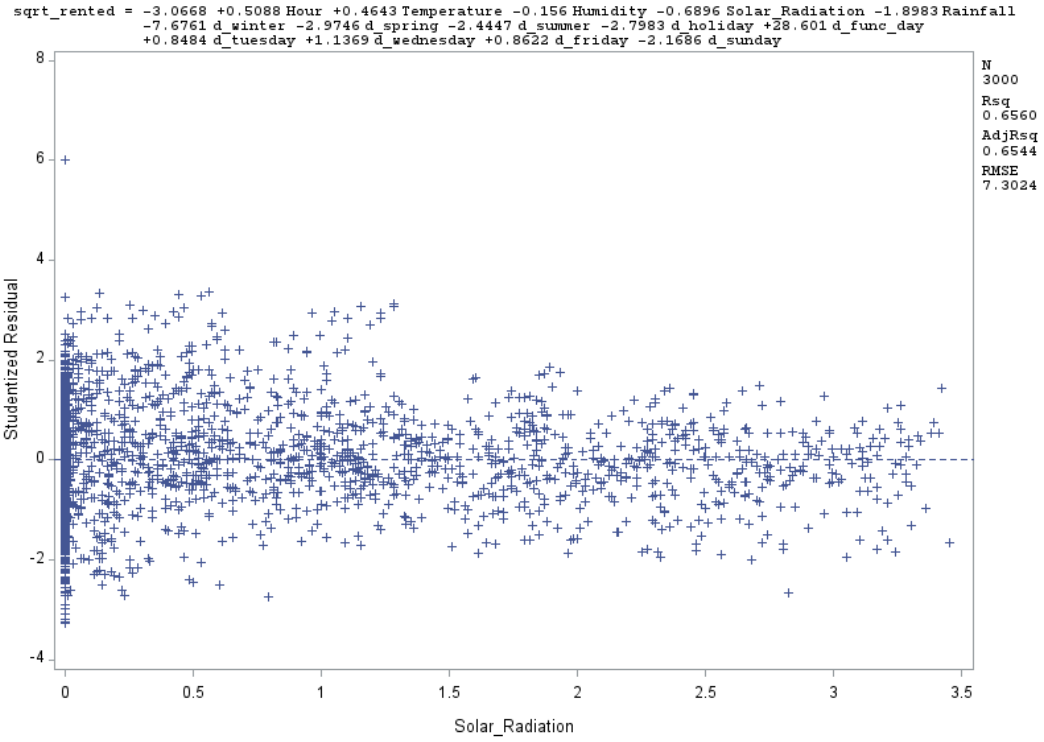
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Tolerance	Variance Inflation
Intercept	1	-3.06679	1.05233	-2.91	0.0036	0	.	0
Hour	1	0.50883	0.02046	24.87	<.0001	0.28268	0.89206	1.12100
Temperature	1	0.46428	0.02499	18.58	<.0001	0.44984	0.19653	5.08831
Humidity	1	-0.15596	0.00850	-18.36	<.0001	-0.25605	0.59225	1.68846
Solar_Radiation	1	-0.68958	0.20047	-3.44	0.0006	-0.04885	0.57149	1.74981
Rainfall	1	-1.89827	0.13506	-14.05	<.0001	-0.15674	0.92663	1.07918
d_winter	1	-7.67615	0.55932	-13.72	<.0001	-0.26460	0.31004	3.22543
d_spring	1	-2.97457	0.38520	-7.72	<.0001	-0.10466	0.62736	1.59399
d_summer	1	-2.44469	0.49335	-4.96	<.0001	-0.08645	0.37863	2.64108
d_holiday	1	-2.79830	0.65786	-4.25	<.0001	-0.04638	0.96939	1.03158
d_func_day	1	28.60147	0.79858	35.82	<.0001	0.39915	0.92787	1.07774
d_tuesday	1	0.84838	0.42037	2.02	0.0437	0.02327	0.86682	1.15365
d_wednesday	1	1.13687	0.40253	2.82	0.0048	0.03248	0.87161	1.14730
d_friday	1	0.86218	0.41200	2.09	0.0365	0.02394	0.88027	1.13601
d_sunday	1	-2.16865	0.41149	-5.27	<.0001	-0.06035	0.87893	1.13775

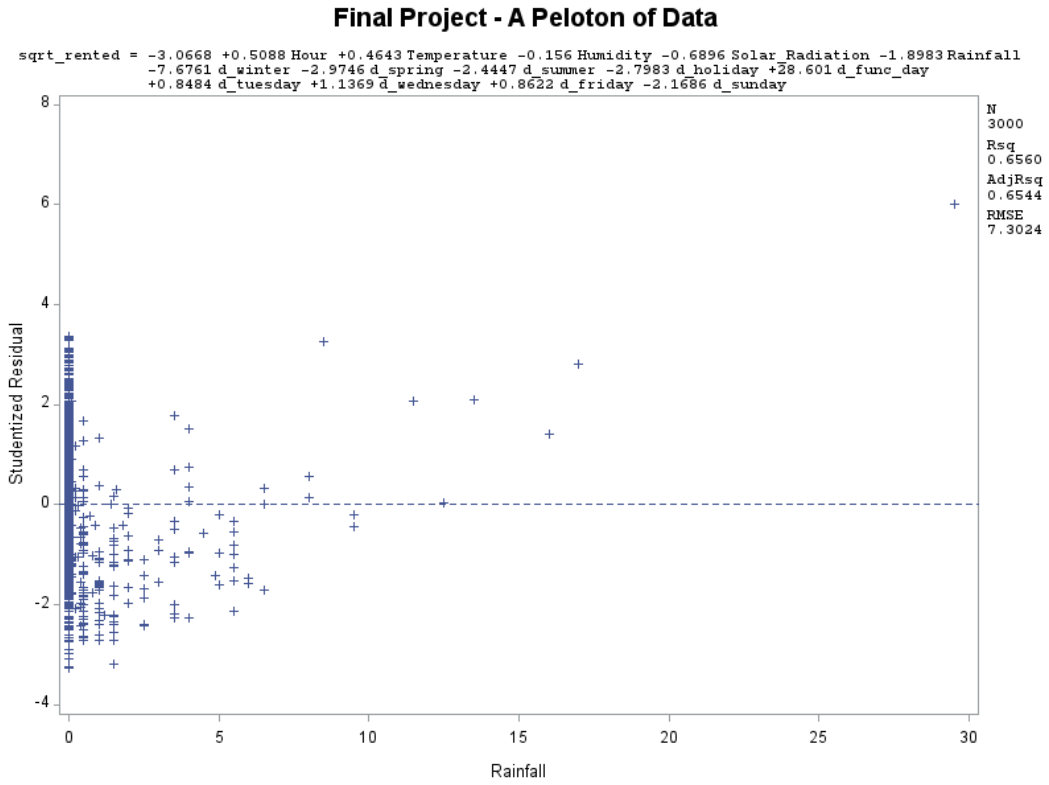
A.d

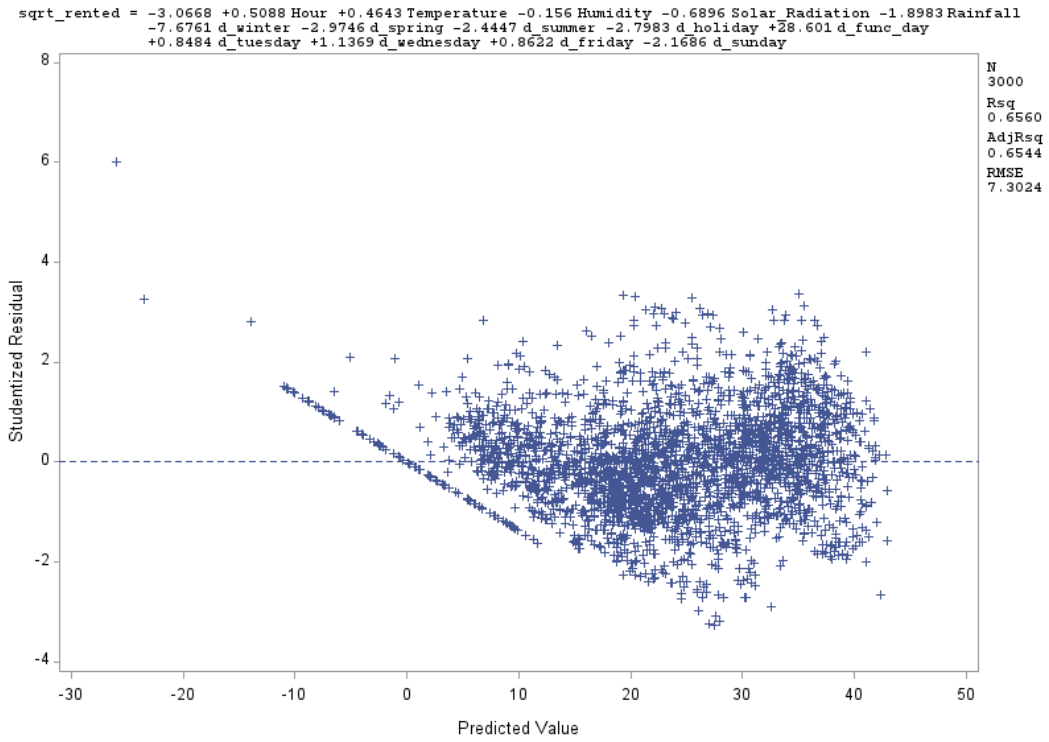
Studentized Residuals vs. each numerical independent variable and Residuals vs. Predicted Values




Final Project - A Peloton of Data
sqrt_rented = -3.0668 +0.5088 Hour +0.4643 Temperature -0.156 Humidity -0.6896 Solar_Radiation -1.8983 Rainfall
-7.6761 d_winter -2.9746 d_spring -2.4447 d_summer -2.7983 d_holiday +28.601 d_func_day
+0.8484 d_tuesday +1.1369 d_wednesday +0.8622 d_friday -2.1686 d_sunday
N
3000
Rsq
0.6560
AdjRsq
0.6544
RMSE
7.3024
Studentized Residual
Humidity

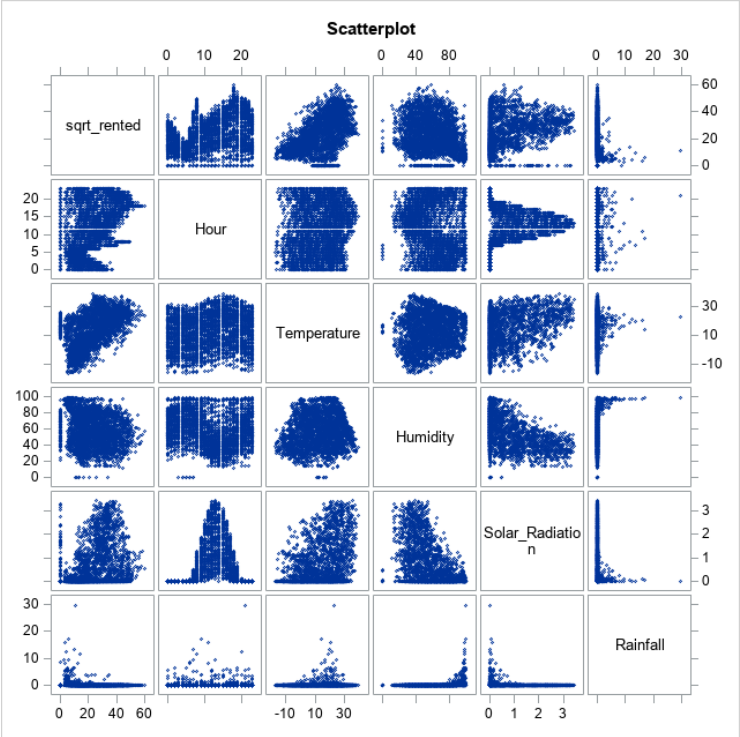

Final Project - A Peloton of Data
sqrt_rented = -3.0668 +0.5088 Hour +0.4643 Temperature -0.156 Humidity -0.6896 Solar_Radiation -1.8983 Rainfall
-7.6761 d_winter -2.9746 d_spring -2.4447 d_summer -2.7983 d_holiday +28.601 d_func_day
+0.8484 d_tuesday +1.1369 d_wednesday +0.8622 d_friday -2.1686 d_sunday
N
3000
Rsq
0.6560
AdjRsq
0.6544
RMSE
7.3024
Studentized Residual
Solar_Radiation


Final Project - A Peloton of Data
sqrt_rented = -3.0668 +0.5088 Hour +0.4643 Temperature -0.156 Humidity -0.6896 Solar_Radiation -1.8983 Rainfall
-7.6761 d_winter -2.9746 d_spring -2.4447 d_summer -2.7983 d_holiday +28.601 d_func_day
+0.8484 d_tuesday +1.1369 d_wednesday +0.8622 d_friday -2.1686 d_sunday
N
3000
Rsq
0.6560
AdjRsq
0.6544
RMSE
7.3024
Studentized Residual
Rainfall


Final Project - A Peloton of Data
sqrt_rented = -3.0668 +0.5088 Hour +0.4643 Temperature -0.156 Humidity -0.6896 Solar_Radiation -1.8983 Rainfall
-7.6761 d_winter -2.9746 d_spring -2.4447 d_summer -2.7983 d_holiday +28.601 d_func_day
+0.8484 d_tuesday +1.1369 d_wednesday +0.8622 d_friday -2.1686 d_sunday
N
3000
Rsq
0.6560
AdjRsq
0.6544
RMSE
7.3024
Studentized Residual
Predicted Value

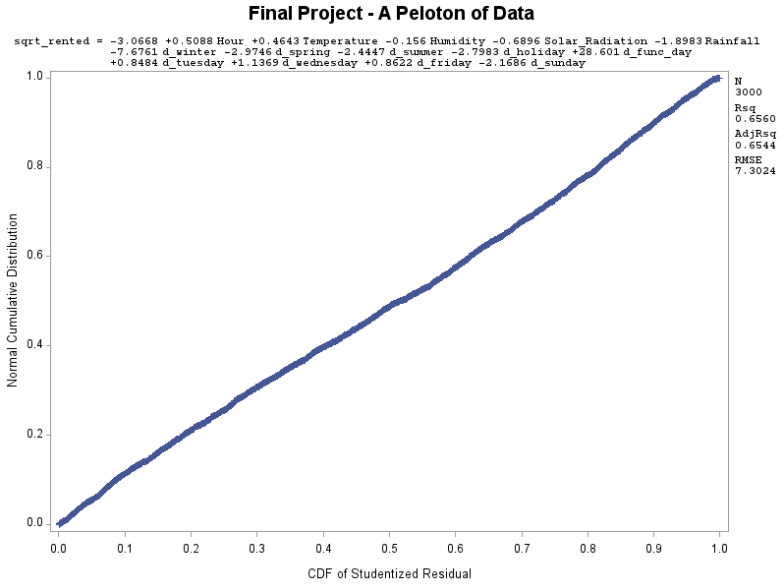
A.e

Scatterplot of dependent variables against continuous independent variables to test linearity assumption



A.f

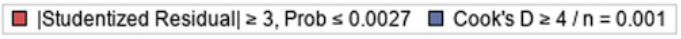
Normal Probability Plots



A.g

Selected points from Studentized Residuals & Cook's Distance plot of full model.

Includes some data points that weren't considered outliers when accounting for Prob <= 0.0027



Obs	Rented_Bike_Count	Hour	Temperature	Humidity	Wind_Speed	Visibility	Dew_Point_Temperature	Solar_Radiation	Rainfall	Snowfall	sqrt_rented	d_winter	d_spring	d_summer	d_holiday	d_func_day	d_tuesday	d_wednesday	d_thursday	d_friday
1	1997	8	13.8	79	0.9	313	10.2	0.44	0	0	44.6878	0	1	0	0	1	0	0	0	0
2	182	5	15.7	0	0.6	1610	10.6	0	0	0	13.4907	0	1	0	1	1	1	0	0	0
3	2035	8	14.6	70	3.1	2000	9.1	1.28	0	0	45.1110	0	1	0	0	1	0	1	0	0
4	174	5	17.6	0	0.8	1304	9.7	0	0	0	13.1909	0	1	0	0	1	0	0	0	0
5	134	4	17.2	0	0.4	1619	8.8	0	0	0	11.5758	0	1	0	0	1	0	0	0	0
6	3404	18	24.9	53	3.6	2000	14.6	1.28	0	0	58.3438	0	0	1	0	1	1	0	0	0
7	2357	8	22.7	65	0.7	759	15.7	1.05	0	0	48.5489	0	0	1	0	1	0	0	0	0
8	3556	18	24.1	57	2.9	1301	15	0.56	0	0	59.6322	0	0	1	0	1	1	0	0	0
9	2440	8	20.6	77	1.2	566	16.4	0.53	0	0	49.3964	0	0	1	0	1	0	1	0	0
10	2370	8	21.4	65	1.9	821	14.5	1.15	0	0	48.6826	0	0	1	0	1	0	0	1	0
11	1995	8	19.7	97	1.5	200	19.2	0.25	0	0	44.6654	0	0	1	0	1	0	1	0	0
12	122	21	22.9	98	2.1	1146	22.5	0	29.5	0	11.0454	0	0	1	0	1	1	0	0	0
13	23	22	16.2	92	2.6	1771	14.8	0	1.5	0	4.7958	0	0	0	0	1	0	0	0	1
14	0	6	16.5	96	2.3	417	15.8	0	8.5	0	0.0000	0	0	0	0	0	0	0	0	0
15	2113	8	7.9	63	0.6	2000	1.2	0.3	0	0	45.9674	0	0	0	0	1	0	0	0	1
16	1899	8	2.8	74	0.9	1366	-1.3	0.13	0	0	43.5775	0	0	0	0	1	0	0	1	0
17	1906	8	4.7	72	1	1306	0	0.12	0	0	43.6578	0	0	0	0	1	0	0	0	1
18	12	20	13.9	97	2.5	594	13.4	0	0	0	3.4641	0	0	0	0	1	0	0	1	0
19	14	22	12.4	96	2.7	545	11.7	0	0	0	3.7417	0	0	0	0	1	0	0	1	0
20	26	23	11.5	96	3	691	10.8	0	0	0	5.0990	0	0	0	0	1	0	0	1	0
21	1910	8	4.7	68	0.5	1032	-0.7	0.06	0	0	43.7035	0	0	0	0	1	1	0	0	0

A.h

Analysis of Variance Output after full model analysis and adjustments

The REG Procedure
Model: MODEL1
Dependent Variable: sqrt_rented

Number of Observations Read	2980
Number of Observations Used	2980

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	309082	22077	440.79	<.0001
Error	2965	148503	50.08544		
Corrected Total	2979	457586			

Root MSE	7.07711	R-Square	0.6755
Dependent Mean	23.57414	Adj R-Sq	0.6739
Coeff Var	30.02064		

S.a

Analysis of Variance and Parameter Estimates for Forward and Backward Selection

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	181568	12969	247.06	<.0001
Error	1773	93072	52.49403		
Corrected Total	1787	274640			

Root MSE	7.24528	R-Square	0.6611
Dependent Mean	23.57053	Adj R-Sq	0.6584
Coeff Var	30.73871		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-5.04295	1.39744	-3.61	0.0003
Hour	1	0.50328	0.02667	18.87	<.0001
Temperature	1	0.50908	0.03224	15.79	<.0001
Humidity	1	-0.14602	0.01130	-12.92	<.0001
Solar_Radiation	1	-0.97025	0.26405	-3.67	0.0002
Rainfall	1	-3.56384	0.27880	-12.78	<.0001
d_winter	1	-6.71131	0.72124	-9.31	<.0001
d_spring	1	-2.54921	0.49087	-5.19	<.0001
d_summer	1	-2.71293	0.63263	-4.29	<.0001
d_holiday	1	-2.26284	0.90567	-2.50	0.0126
d_func_day	1	29.62476	1.04179	28.44	<.0001
d_tuesday	1	0.73440	0.54199	1.36	0.1756
d_wednesday	1	1.15758	0.51491	2.25	0.0247
d_friday	1	0.99695	0.53581	1.86	0.0630
d_sunday	1	-2.38153	0.51755	-4.60	<.0001

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	181340	15112	287.50	<.0001
Error	1775	93300	52.56327		
Corrected Total	1787	274640			

Root MSE	7.25005	R-Square	0.6603
Dependent Mean	23.57053	Adj R-Sq	0.6580
Coeff Var	30.75898		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-4.49337	1.37137	-3.28	0.0011
Hour	1	0.50283	0.02668	18.85	<.0001
Temperature	1	0.50553	0.03217	15.72	<.0001
Humidity	1	-0.14694	0.01130	-13.00	<.0001
Solar_Radiation	1	-0.97536	0.26421	-3.69	0.0002
Rainfall	1	-3.57528	0.27892	-12.82	<.0001
d_winter	1	-6.73220	0.71960	-9.36	<.0001
d_spring	1	-2.48884	0.49021	-5.08	<.0001
d_summer	1	-2.62576	0.63148	-4.16	<.0001
d_holiday	1	-2.22909	0.90611	-2.46	0.0140
d_func_day	1	29.47737	1.03727	28.42	<.0001
d_wednesday	1	0.82517	0.48875	1.69	0.0915
d_sunday	1	-2.71342	0.49178	-5.52	<.0001

S.b

Analysis of Variance and Parameter Estimates for Stepwise Selection

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	180521	15043	283.71	<.0001
Error	1775	94119	53.02481		
Corrected Total	1787	274640			

Root MSE	7.28181	R-Square	0.6573
Dependent Mean	23.57053	Adj R-Sq	0.6550
Coeff Var	30.89372		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-3.71360	1.36481	-2.72	0.0066
Hour	1	0.51319	0.02668	19.24	<.0001
Temperature	1	0.42686	0.02587	16.50	<.0001
Humidity	1	-0.14897	0.01134	-13.14	<.0001
Solar_Radiation	1	-0.82783	0.26317	-3.15	0.0017
Rainfall	1	-3.59967	0.28008	-12.85	<.0001
d_winter	1	-7.17156	0.71405	-10.04	<.0001
d_spring	1	-1.74451	0.45802	-3.81	0.0001
d_holiday	1	-1.79647	0.90397	-1.99	0.0470
d_func_day	1	28.69914	1.02520	27.99	<.0001
d_wednesday	1	1.00275	0.50093	2.00	0.0455
d_friday	1	0.67817	0.52181	1.30	0.1939
d_sunday	1	-2.59572	0.50411	-5.15	<.0001

S.c

Models & CP Values from Mallow's CP Selection

Number in Model	C(p)	R-Square	Variables in Model
13	14.8360	0.6608	Hour Temperature Humidity Solar_Radiation Rainfall d_winter d_spring d_summer d_holiday d_func_day d_wednesday d_friday d_sunday
14	15.0000	0.6611	Hour Temperature Humidity Solar_Radiation Rainfall d_winter d_spring d_summer d_holiday d_func_day d_tuesday d_wednesday d_friday d_sunday
12	15.3415	0.6603	Hour Temperature Humidity Solar_Radiation Rainfall d_winter d_spring d_summer d_holiday d_func_day d_wednesday d_sunday
11	16.1957	0.6597	Hour Temperature Humidity Solar_Radiation Rainfall d_winter d_spring d_summer d_holiday d_func_day d_sunday
13	16.4620	0.6605	Hour Temperature Humidity Solar_Radiation Rainfall d_winter d_spring d_summer d_holiday d_func_day d_tuesday d_wednesday d_sunday

S.d

Analysis of Variance and Parameter Estimates for CP Models

CP Selection models

The REG Procedure
Model: MODEL1
Dependent Variable: sq_rented_train

Number of Observations Read	2980
Number of Observations Used	1788
Number of Observations with Missing Values	1192

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	181472	13959	265.80	<.0001
Error	1774	93168	52.51877		
Corrected Total	1787	274640			

Root MSE	7.24698	R-Square	0.6608
Dependent Mean	23.57053	Adj R-Sq	0.6583
Coeff Var	30.74595		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-4.73713	1.37942	-3.43	0.0006
Hour	1	0.50264	0.02667	18.85	<.0001
Temperature	1	0.50944	0.03225	15.80	<.0001
Humidity	1	-0.14634	0.01130	-12.95	<.0001
Solar_Radiation	1	-0.97251	0.26411	-3.68	0.0002
Rainfall	1	-3.57055	0.27882	-12.81	<.0001
d_winter	1	-6.66418	0.72057	-9.25	<.0001
d_spring	1	-2.51195	0.49022	-5.12	<.0001
d_summer	1	-2.69149	0.63258	-4.25	<.0001
d_holiday	1	-2.24420	0.90578	-2.48	0.0133
d_func_day	1	29.48392	1.03684	28.44	<.0001
d_wednesday	1	0.98253	0.49856	1.97	0.0489
d_friday	1	0.82358	0.52043	1.58	0.1137
d_sunday	1	-2.55388	0.50180	-5.09	<.0001

CP Selection models

The REG Procedure
Model: MODEL2
Dependent Variable: sq_rented_train

Number of Observations Read	2980
Number of Observations Used	1788
Number of Observations with Missing Values	1192

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	181568	12969	247.06	<.0001
Error	1773	93072	52.49403		
Corrected Total	1787	274640			

Root MSE	7.24528	R-Square	0.6611
Dependent Mean	23.57053	Adj R-Sq	0.6584
Coeff Var	30.73871		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-5.04295	1.39744	-3.61	0.0003
Hour	1	0.50328	0.02667	18.87	<.0001
Temperature	1	0.50908	0.03224	15.79	<.0001
Humidity	1	-0.14602	0.01130	-12.92	<.0001
Solar_Radiation	1	-0.97025	0.26405	-3.67	0.0002
Rainfall	1	-3.56384	0.27880	-12.78	<.0001
d_winter	1	-6.71131	0.72124	-9.31	<.0001
d_spring	1	-2.54921	0.49087	-5.19	<.0001
d_summer	1	-2.71293	0.63263	-4.29	<.0001
d_holiday	1	-2.26284	0.90567	-2.50	0.0126
d_func_day	1	29.62476	1.04179	28.44	<.0001
d_tuesday	1	0.73440	0.54199	1.36	0.1756
d_wednesday	1	1.15758	0.51491	2.25	0.0247
d_friday	1	0.99695	0.53581	1.86	0.0630
d_sunday	1	-2.38153	0.51755	-4.60	<.0001

S.e

Analysis of Variance and Parameter Estimates for ADJRSQ Models

Adj-R2 Selection Models

The REG Procedure
Model: MODEL1
Dependent Variable: sq_rented_train

Number of Observations Read	2980
Number of Observations Used	1788
Number of Observations with Missing Values	1192

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	181340	15112	287.50	<.0001
Error	1775	93300	52.56327		
Corrected Total	1787	274640			

Root MSE	7.25005	R-Square	0.6603
Dependent Mean	23.57053	Adj R-Sq	0.6580
Coeff Var	30.75898		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-4.49337	1.37137	-3.28	0.0011
Hour	1	0.50283	0.02668	18.85	<.0001
Temperature	1	0.50553	0.03217	15.72	<.0001
Humidity	1	-0.14694	0.01130	-13.00	<.0001
Solar_Radiation	1	-0.97536	0.26421	-3.69	0.0002
Rainfall	1	-3.57528	0.27892	-12.82	<.0001
d_winter	1	-6.73220	0.71960	-9.36	<.0001
d_spring	1	-2.48884	0.49021	-5.08	<.0001
d_summer	1	-2.62576	0.63148	-4.16	<.0001
d_holiday	1	-2.22909	0.90611	-2.46	0.0140
d_func_day	1	29.47737	1.03727	28.42	<.0001
d_wednesday	1	0.82517	0.48875	1.69	0.0915
d_sunday	1	-2.71342	0.49178	-5.52	<.0001

Adj-R2 Selection Models

The REG Procedure
Model: MODEL2
Dependent Variable: sq_rented_train

Number of Observations Read	2980
Number of Observations Used	1788
Number of Observations with Missing Values	1192

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	181268	15106	287.16	<.0001
Error	1775	93372	52.60409		
Corrected Total	1787	274640			

Root MSE	7.25287	R-Square	0.6600
Dependent Mean	23.57053	Adj R-Sq	0.6577
Coeff Var	30.77092		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-4.53677	1.37678	-3.30	0.0010
Hour	1	0.50079	0.02667	18.78	<.0001
Temperature	1	0.51005	0.03227	15.80	<.0001
Humidity	1	-0.14614	0.01131	-12.92	<.0001
Solar_Radiation	1	-0.95855	0.26422	-3.63	0.0003
Rainfall	1	-3.57245	0.27904	-12.80	<.0001
d_winter	1	-6.62561	0.72089	-9.19	<.0001
d_spring	1	-2.48582	0.49044	-5.07	<.0001
d_summer	1	-2.70338	0.63307	-4.27	<.0001
d_holiday	1	-2.18146	0.90596	-2.41	0.0161
d_func_day	1	29.46812	1.03765	28.40	<.0001
d_friday	1	0.61902	0.51039	1.21	0.2254
d_sunday	1	-2.76323	0.49083	-5.63	<.0001

Adj-R2 Selection Models

The REG Procedure
Model: MODEL3
Dependent Variable: sq_rented_train

Number of Observations Read	2980
Number of Observations Used	1788
Number of Observations with Missing Values	1192

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	181190	16472	313.05	<.0001
Error	1776	93450	52.61804		
Corrected Total	1787	274640			

Root MSE	7.25383	R-Square	0.6597
Dependent Mean	23.57053	Adj R-Sq	0.6576
Coeff Var	30.77500		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-4.37109	1.37017	-3.19	0.0014
Hour	1	0.50117	0.02667	18.79	<.0001
Temperature	1	0.50691	0.03217	15.76	<.0001
Humidity	1	-0.14663	0.01131	-12.97	<.0001
Solar_Radiation	1	-0.96253	0.26424	-3.64	0.0003
Rainfall	1	-3.57591	0.27907	-12.81	<.0001
d_winter	1	-6.68369	0.71940	-9.29	<.0001
d_spring	1	-2.47100	0.49035	-5.04	<.0001
d_summer	1	-2.65043	0.63164	-4.20	<.0001
d_holiday	1	-2.17749	0.90607	-2.40	0.0164
d_func_day	1	29.46498	1.03778	28.39	<.0001
d_sunday	1	-2.86187	0.48410	-5.91	<.0001

Adj-R2 Selection Models

The REG Procedure
Model: MODEL4
Dependent Variable: sq_rented_train

Number of Observations Read	2980
Number of Observations Used	1788
Number of Observations with Missing Values	1192

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	180887	18089	342.85	<.0001
Error	1777	93754	52.75945		
Corrected Total	1787	274640			

Root MSE	7.26357	R-Square	0.6586
Dependent Mean	23.57053	Adj R-Sq	0.6567
Coeff Var	30.81632		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-4.27191	1.37139	-3.12	0.0019
Hour	1	0.50166	0.02671	18.78	<.0001
Temperature	1	0.49816	0.03201	15.56	<.0001
Humidity	1	-0.14701	0.01132	-12.99	<.0001
Solar_Radiation	1	-0.95331	0.26457	-3.60	0.0003
Rainfall	1	-3.57484	0.27944	-12.79	<.0001
d_winter	1	-6.89378	0.71502	-9.64	<.0001
d_spring	1	-2.41377	0.49043	-4.92	<.0001
d_summer	1	-2.47341	0.62818	-3.94	<.0001
d_func_day	1	29.40432	1.03887	28.30	<.0001
d_sunday	1	-2.92898	0.48395	-6.05	<.0001

S.f

Validation and Correlation for Tested Models

Validation for AdjR2

Obs	_TYPE_	_FREQ_	rmse	mae
1	0	1192	6.85288	5.38657

Validation for AdjR2

The CORR Procedure

2 Variables:	sqrt_rented yhat

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
sqrt_rented	1192	23.57955	12.39382	28107	0	56.53318	
yhat	1192	23.61250	10.32270	28146	-11.56624	42.68489	Predicted Value of sq_rented_train

Pearson Correlation Coefficients, N = 1192
Prob > |r| under H0: Rho=0

	sqrt_rented	yhat
sqrt_rented	1.00000	0.83323 <.0001
yhat Predicted Value of sq_rented_train	0.83323 <.0001	1.00000

Validation for CP

Obs	_TYPE_	_FREQ_	rmse	mae
1	0	1192	6.84922	5.39389

Validation for CP

The CORR Procedure

2 Variables:	sqrt_rented yhat

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
sqrt_rented	1192	23.57955	12.39382	28107	0	56.53318	
yhat	1192	23.60499	10.32370	28137	-11.44445	43.40404	Predicted Value of sq_rented_train

Pearson Correlation Coefficients, N = 1192
Prob > |r| under H0: Rho=0

	sqrt_rented	yhat
sqrt_rented	1.00000	0.83343 <.0001
yhat Predicted Value of sq_rented_train	0.83343 <.0001	1.00000

VALIDATION OF FINAL MODEL

V.a

Analysis of Variance Output from first time running final regression model

Check for Outliers/Influential Points

The REG Procedure
Model: MODEL1
Dependent Variable: sqrt_rented

Number of Observations Read	2980
Number of Observations Used	2980

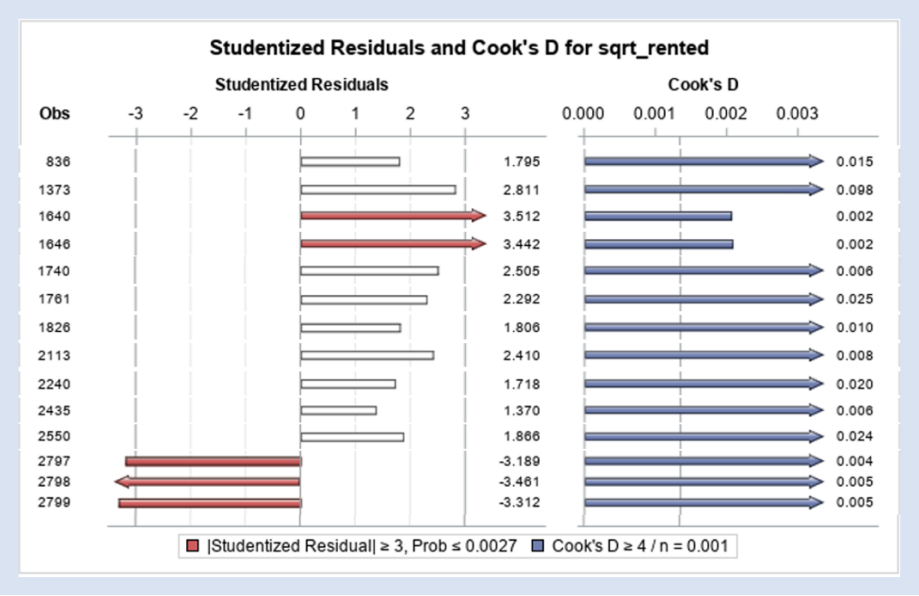
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	307967	30797	611.12	<.0001
Error	2969	149619	50.39365		
Corrected Total	2979	457586			

Root MSE	7.09885	R-Square	0.6730
Dependent Mean	23.57414	Adj R-Sq	0.6719
Coeff Var	30.11287		

V.b

Selected points from Studentized Residuals & Cook's Distance plot of final model



V.c

Analysis of Variance Output from second time running final regression mode

Checks for Mulicollinearity and Model Assumptions

The REG Procedure
Model: MODEL1
Dependent Variable: sqrt_rented

Number of Observations Read	2975
Number of Observations Used	2975

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	307786	30779	621.79	<.0001
Error	2964	146718	49.49993		
Corrected Total	2974	454504			

Root MSE	7.03562	R-Square	0.6772
Dependent Mean	23.57277	Adj R-Sq	0.6761
Coeff Var	29.84638		

V.d

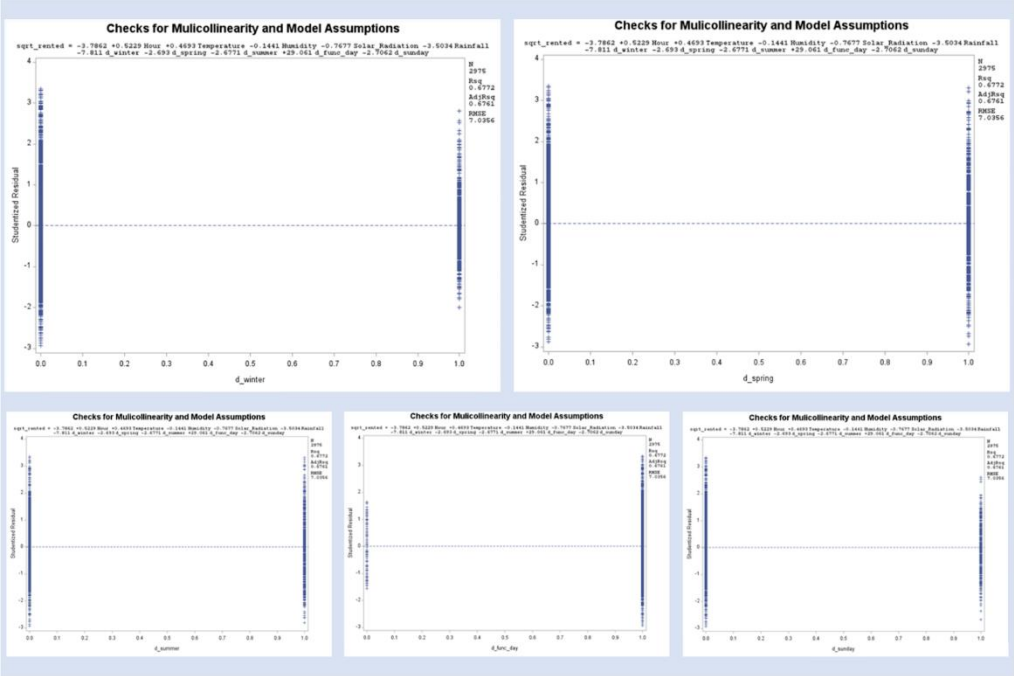
Parameter Estimate table with Tolerance and Variance Inflation Included

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Tolerance	Variance Inflation
Intercept	1	-3.78622	1.01062	-3.75	0.0002	0	.	0
Hour	1	0.52288	0.01992	26.25	<.0001	0.29173	0.88183	1.13400
Temperature	1	0.46929	0.02402	19.53	<.0001	0.45837	0.19779	5.05585
Humidity	1	-0.14413	0.00852	-16.92	<.0001	-0.23472	0.56570	1.76772
Solar_Radiation	1	-0.76770	0.19533	-3.93	<.0001	-0.05479	0.56048	1.78417
Rainfall	1	-3.50339	0.20211	-17.33	<.0001	-0.19032	0.90343	1.10690
d_winter	1	-7.81103	0.53632	-14.56	<.0001	-0.27130	0.31385	3.18620
d_spring	1	-2.69295	0.37278	-7.22	<.0001	-0.09498	0.63004	1.58720
d_summer	1	-2.67708	0.47421	-5.65	<.0001	-0.09510	0.38376	2.60577
d_func_day	1	29.06103	0.76881	37.80	<.0001	0.40702	0.93935	1.06457
d_sunday	1	-2.70617	0.37274	-7.26	<.0001	-0.07586	0.99767	1.00234

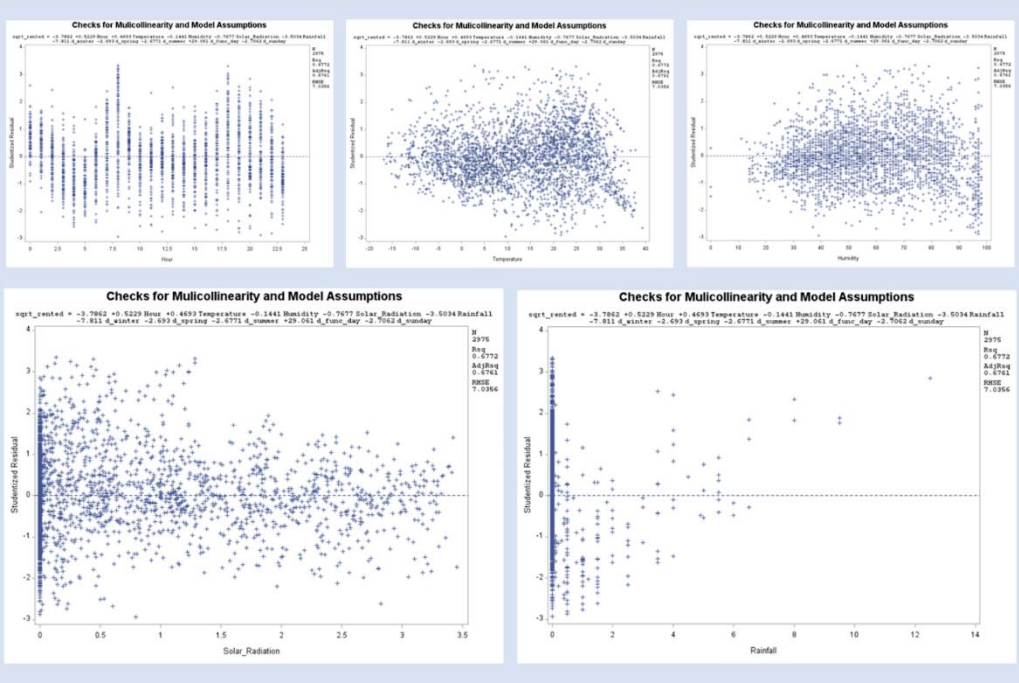
V.e

Studentized Residuals vs. each binary independent variable



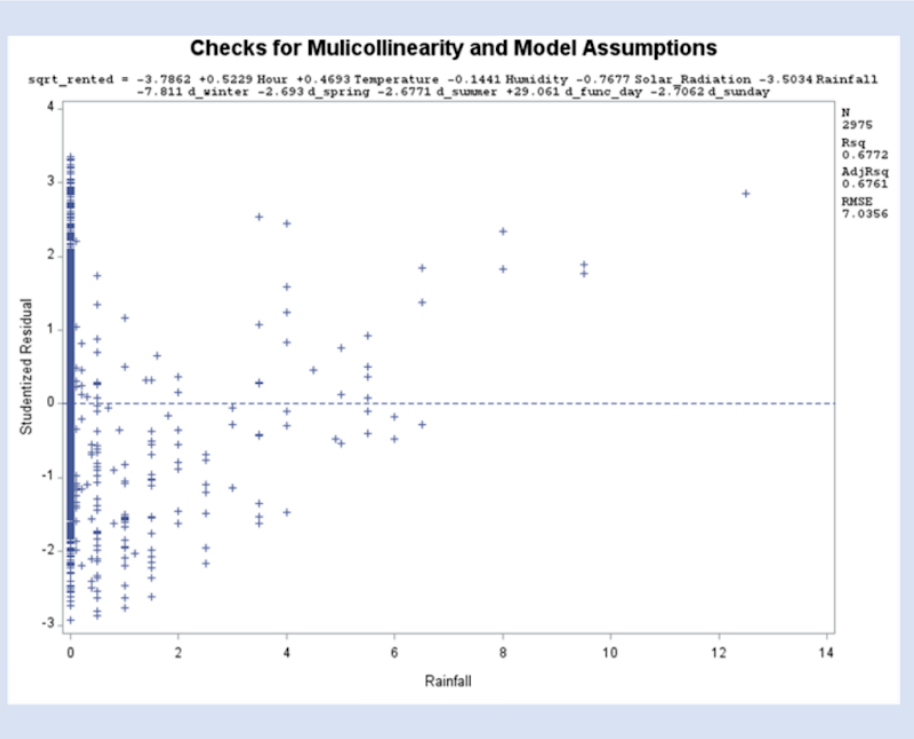
V.f

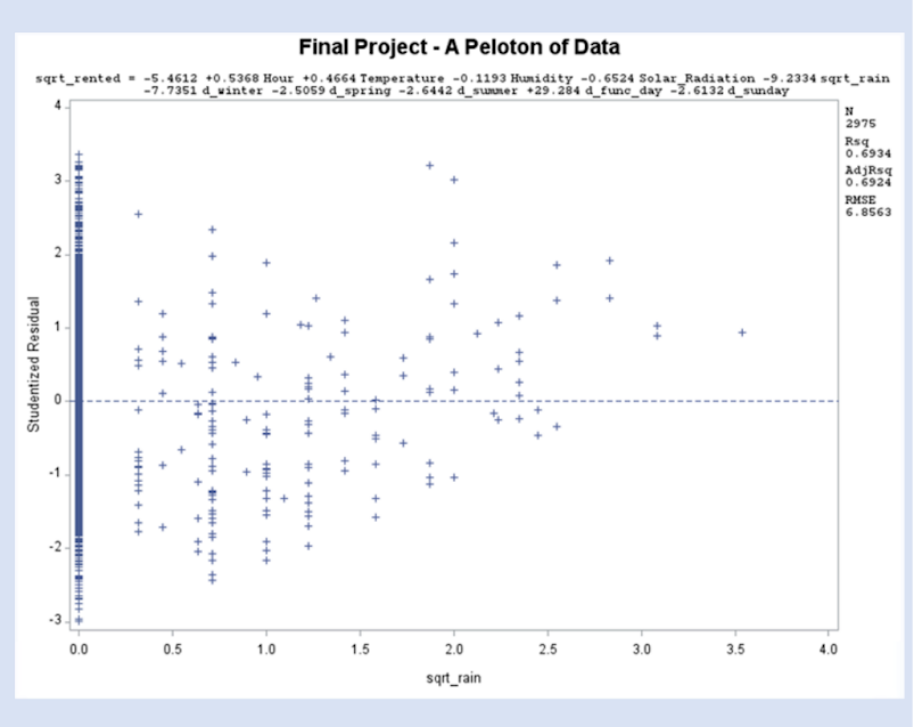
Studentized Residuals vs. each numerical independent variable



V.g

Studentized Residuals vs. Rainfall & Studentized Residuals vs. square root of Rainfall





V.h

Analysis of Variance Output from third time running final regression model

Checks for Mulicollinearity and Model Assumptions

The REG Procedure
Model: MODEL1
Dependent Variable: sqrt_rented

Number of Observations Read	2975
Number of Observations Used	2975

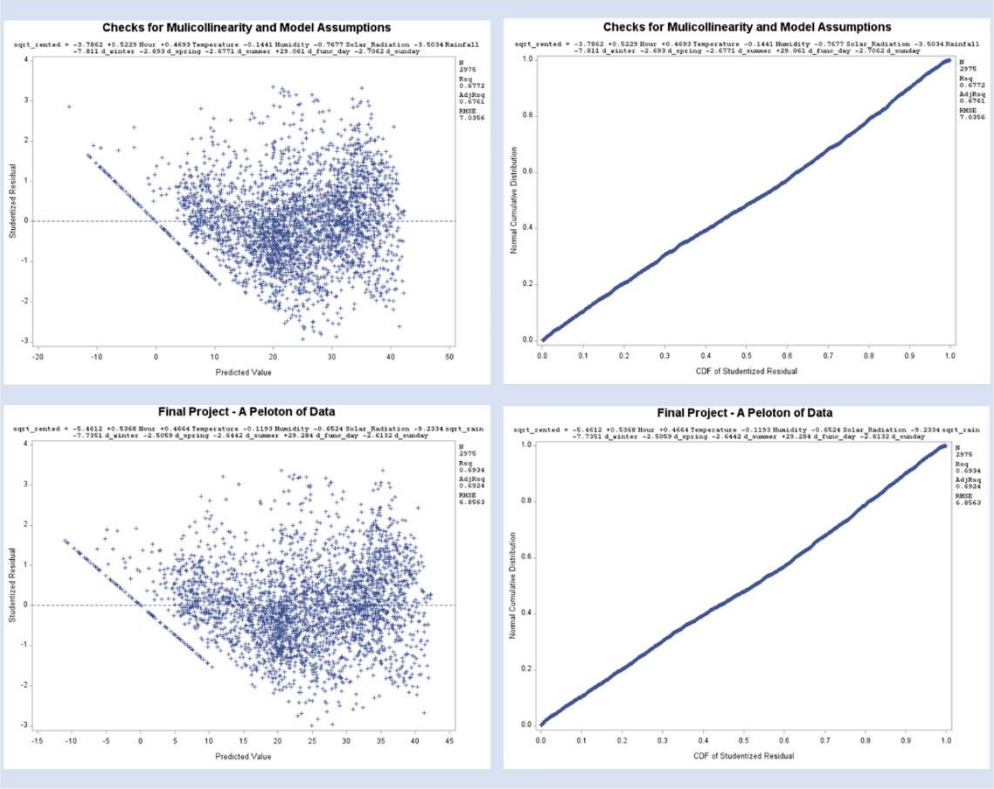
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	315170	31517	670.45	<.0001
Error	2964	139334	47.00878		
Corrected Total	2974	454504			

Root MSE	6.85629	R-Square	0.6934
Dependent Mean	23.57277	Adj R-Sq	0.6924
Coeff Var	29.08566		

V.i

Residuals vs. Predicted Values & Normal Probability Plots from before/after `sqrt_rain`



V.j

Analysis of Variance Output and Parameter Estimates from last time running final model

Verified Final Model

The REG Procedure
Model: MODEL1
Dependent Variable: sqrt_rented

Number of Observations Read	2975
Number of Observations Used	2975

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	315170	31517	670.45	<.0001
Error	2964	139334	47.00878		
Corrected Total	2974	454504			

Root MSE	6.85629	R-Square	0.6934
Dependent Mean	23.57277	Adj R-Sq	0.6924
Coeff Var	29.08566		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-5.46121	0.99375	-5.50	<.0001
Hour	1	0.53682	0.01944	27.61	<.0001
Temperature	1	0.46643	0.02341	19.92	<.0001
Humidity	1	-0.11931	0.00853	-13.99	<.0001
Solar_Radiation	1	-0.65241	0.19056	-3.42	0.0006
sqrt_rain	1	-9.23338	0.42434	-21.76	<.0001
d_winter	1	-7.73507	0.52267	-14.80	<.0001
d_spring	1	-2.50591	0.36353	-6.89	<.0001
d_summer	1	-2.64415	0.46211	-5.72	<.0001
d_func_day	1	29.28367	0.74942	39.08	<.0001
d_sunday	1	-2.61318	0.36326	-7.19	<.0001

PREDICTIONS & CONCLUSIONS

P.a

Verified Final Model

The REG Procedure
Model: MODEL1
Dependent Variable: sqrt_rented

Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	.	18.8757	0.7722	17.3616	20.3898	5.3471	32.4043	.
2	14.28	9.5252	0.3619	8.8155	10.2348	-3.9371	22.9875	4.7577
3	10.00	11.3720	0.3277	10.7295	12.0145	-2.0870	24.8309	-1.3720
4	13.45	12.0541	0.3293	11.4085	12.6998	-1.4050	25.5132	1.3995
5	30.50	12.4162	0.3132	11.8022	13.0302	-1.0414	25.8738	18.0797
6	22.14	14.5156	0.3388	13.8514	15.1799	1.0557	27.9756	7.6203
7	21.14	21.2412	0.3228	20.6083	21.8742	7.7828	34.6997	-0.0989
8	23.56	18.6145	0.3012	18.0239	19.2051	5.1579	32.0710	4.9440
9	24.49	17.1003	0.3854	16.3446	17.8560	3.6355	30.5651	7.3946
10	9.43	7.0369	0.3466	6.3573	7.7166	-6.4238	20.4977	2.3970
11	14.80	8.9911	0.3397	8.3251	9.6572	-4.4689	22.4512	5.8075
12	18.11	11.2575	0.3001	10.6691	11.8459	-2.1990	24.7140	6.8533

P.b

Verified Final Model

The REG Procedure
Model: MODEL1
Dependent Variable: sqrt_rented

Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	.	-20.4564	0.8936	-22.2086	-18.7042	-34.0137	-6.8991	.
2	14.28	9.5252	0.3619	8.8155	10.2348	-3.9371	22.9875	4.7577
3	10.00	11.3720	0.3277	10.7295	12.0145	-2.0870	24.8309	-1.3720
4	13.45	12.0541	0.3293	11.4085	12.6998	-1.4050	25.5132	1.3995
5	30.50	12.4162	0.3132	11.8022	13.0302	-1.0414	25.8738	18.0797
6	22.14	14.5156	0.3388	13.8514	15.1799	1.0557	27.9756	7.6203
7	21.14	21.2412	0.3228	20.6083	21.8742	7.7828	34.6997	-0.0989
8	23.56	18.6145	0.3012	18.0239	19.2051	5.1579	32.0710	4.9440
9	24.49	17.1003	0.3854	16.3446	17.8560	3.6355	30.5651	7.3946
10	9.43	7.0369	0.3466	6.3573	7.7166	-6.4238	20.4977	2.3970
11	14.80	8.9911	0.3397	8.3251	9.6572	-4.4689	22.4512	5.8075
12	18.11	11.2575	0.3001	10.6691	11.8459	-2.1990	24.7140	6.8533

P.c

Verified Final Model

The REG Procedure
Model: MODEL1
Dependent Variable: sqrt_rented

Number of Observations Read	2975
Number of Observations Used	2975

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	315170	31517	670.45	<.0001
Error	2964	139334	47.00878		
Corrected Total	2974	454504			

Root MSE	6.85629	R-Square	0.6934
Dependent Mean	23.57277	Adj R-Sq	0.6924
Coeff Var	29.08566		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	-5.46121	0.99375	-5.50	<.0001	0
Hour	1	0.53682	0.01944	27.61	<.0001	0.29951
Temperature	1	0.46643	0.02341	19.92	<.0001	0.45558
Humidity	1	-0.11931	0.00853	-13.99	<.0001	-0.19429
Solar_Radiation	1	-0.65241	0.19056	-3.42	0.0006	-0.04656
sqrt_rain	1	-9.23338	0.42434	-21.76	<.0001	-0.24088
d_winter	1	-7.73507	0.52267	-14.80	<.0001	-0.26867
d_spring	1	-2.50591	0.36353	-6.89	<.0001	-0.08838
d_summer	1	-2.64415	0.46211	-5.72	<.0001	-0.09393
d_func_day	1	29.28367	0.74942	39.08	<.0001	0.41014
d_sunday	1	-2.61318	0.36326	-7.19	<.0001	-0.07325