ISyE 4031 Fall 2019 Project

NYC Uber Pickups

Team B9
Grant Cloud
Charles Howard
Harrison Lyle
Jake Porter
Ryan Zweber

Table of Contents

Introduction

Uber is a very popular ride sharing app that has seen significant growth over the past couple of years. Uber is also very popular in large cities, especially New York City. The most important part of the ride sharing experience is being able to provide a seamless process of pickup and transit. If there aren't enough drivers in a specific area, it will take longer for customers to get picked up and reach their destination, causing them to leave harsh reviews of Uber. Bad reviews for Uber and their drivers hurt both the company and the worker as neither is going to make as much money as it would with good reviews. Bad reviews affect Uber because they will cause potential new customers to leave for a different ride sharing app and they hurt the driver as they won't be able to pick up as many customers. Another important factor for Uber to consider is that customers will cancel their trip if it shows the pickup time wait to be too long. This causes Uber to be missing out on potential revenue as the customer was already ready to take the Uber.

Problem Statement

Due to Uber being in an environment that needs to have quick pickups and drop offs, they must be able to predict the number of pickups on a given day in a given location. New York City is a large city that requires Uber to have lots of drivers and have drivers that are able to meet demand on time. In order to predict total number of pickups per day, we will use weather data (wind speed, visibility, temperature, dew point, precipitation and snow depth), day type (weekdays, weekends, holidays, weekends and holidays), borough, and month. More information on the variables can be found in the variables section. The goal is to come up with a linear regression model to predict the total number of Uber pickups for a given day in a New York City borough to reduce cancellations and pickup wait times.

It is crucial for Uber to be able to meet the massive demands that arrives in large cities. Uber needs to be able to predict how many rides they expect to see in a given location during a given time period so they are able to meet up with the ever changing demands of their customers to maximize their potential revenue and expand their customer base in the fierce ride-share environment..

Data Description

The data was collected from Kaggle and is titled, *NYC Uber Pickups Enriched*. This dataset is an enriched form of the *uber-tlc-foil-response* dataset published by the website fivethirtyeight. The original Uber data originates from a freedom of information request to New York City's Taxi and Limousine Commission[1]. This data from fivethirtyeight includes over 4.5 million pickups from April to September 2014 and 14.3 million pickups from January to June 2015[1]. The original data set includes pickup date and time and pickup location within New York City. A user on Kaggle then took that information, aggregated the pickup dates on each day and each hour as well as aggregating the pickup location to be based on each borough in New York City[2]. The Kaggle data set also added in weather data for each borough in each hour for the pickups from the National Centers for Environmental Information[2]. The data set also includes a yes or no column that says whether or not it was a public holiday in New York City[2].

The downloaded *NYC Uber Pickups Enriched* dataset was then further modified with a self written python script. The original dataset had 29,101 data points, one point for every hour for every day for every borough in New York City[2]. The python script further modifies the dataset by calculating the means for the weather data for each day and totals for each day of data recording. The script also creates the month class column and dayType class column. The final dataset is named *uber_nyc_clean.csv*. See appendix B for full python script that further adjusts the data set. A preview of the cleaned data set can be seen in Figure B1.2 below.

	pickup_dt	borough	pickups	spd	vsb	temp	dewp	slp	pcp	sd	dayType	month
0	2015-01-01	Bronx	1075	6.086957	10.0	31.043478	7.869565	1019.795652	0.0	0.0	Hday	Jan
1	2015-01-01	Brooklyn	12528	6.086957	10.0	31.043478	7.869565	1019.795652	0.0	0.0	Hday	Jan
2	2015-01-01	Manhattan	35870	6.086957	10.0	31.043478	7.869565	1019.795652	0.0	0.0	Hday	Jan
3	2015-01-01	Queens	5108	6.086957	10.0	31.043478	7.869565	1019.795652	0.0	0.0	Hday	Jan
4	2015-01-01	Staten Island	53	6.086957	10.0	31.043478	7.869565	1019.795652	0.0	0.0	Hday	Jan

Figure B1.2

Variables:

1. borough: The Borough in New York City that the # of pickups occurred in
    - Qualitative (5 classes): Bronx, Brooklyn, Manhattan, Queens, Staten Island
2. pickups: Total number of uber pickups in that Borough on a given day
    - This is our response variable not a predictor (i.e. dependent variable)
    - Quantitative: Units is number of pickups
3. spd: Mean wind speed in New York City for the day the data point was recorded
    - Quantitative: Units are miles/hour
4. vsb: Mean visibility in New York City for the day the data point was recorded

    - Quantitative: Units are miles
5. temp: Mean temperature in New York City for the day the data point was recorded
    - Quantitative: Units are Fahrenheit
6. dewp: Mean dew point for New York City for the day the data point was recorded
    - Quantitative: Units are Fahrenheit
7. slp: Mean sea level pressure in New York City for the day the data point was recorded
    - Quantitative: Units are atmospheres
8. pcp: Total liquid precipitation in New York City for the day the data point was recorded
    - Quantitative: Units are inches
9. sd: Mean snow depth in New York City for the day the data point was recorded
    - Quantitative: Units are inches
10. dayType: Used to determine whether the day is during the week, on the weekend, and whether it is a holiday or not
    - Qualitative (4 classes): Wday (Weekday), Wend (Weekend), Hday (Holiday), WendHday (Weekend Holiday)
11. month: The month that the data point was recorded in
    - Qualitative (6 classes): Jan, Feb, Marc, Apr, May, Jun

Analysis

i. Initial Variables

Model: pickups ~ spd + vsb + temp + dewp + slp + pcp + sd

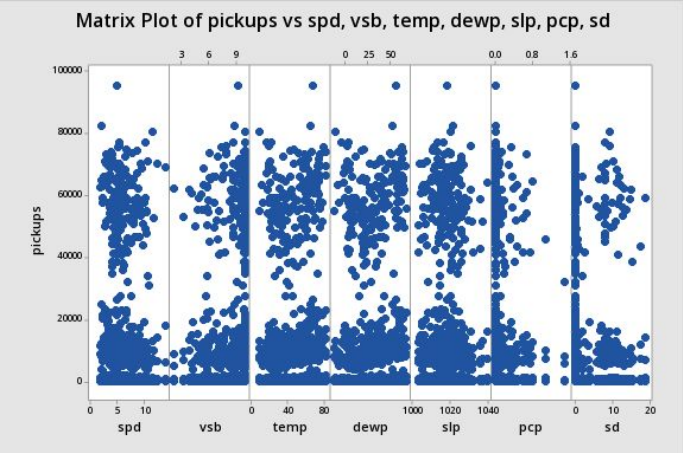

Figure A1.1

From Figure A1.1, we can see grouping within the independent, quantitative variables due to the qualitative variables not being displayed. From the figure, we also see that there is no signs of any patterns for a single independent variable vs the response variable which indicates that there is no need for any higher order variables.

ii. Correlation Analysis

Initially, we perform correlation analysis to check for correlation between independent variables to avoid correlation and multicollinearity. In figure A2.2 we observe that *temp* and *dewp* are highly positively correlated. Also notice that *temp* and *dewp* have correlation coefficients with magnitude greater than 0.5 with 3 different independent variables.

Correlations

	spd	vsb	temp	dewp	slp	pcp
vsb	0.122					
temp	-0.511	0.007				
dewp	-0.492	-0.236	0.937			
slp	-0.162	0.175	-0.212	-0.284		
pcp	0.005	0.022	-0.037	-0.031	-0.128	
sd	0.163	-0.077	-0.566	-0.507	0.132	0.073

Figure A2.2

In order to remove correlation between the independent variables and multicollinearity, several different interaction terms were created and tested involving *dewp* and *temp.* The effects of these different interaction terms that were tested can be viewed in Figure A2.3 and A2.4 in Appendix A2. Ultimately, the optimal interaction term was *temp*dewp* :

$$temp*dewp = temp \times dewp$$

Using the *temp*dewp* interaction term we observe from Figure A2.5 it can be seen that there are no correlations with magnitude greater than 0.5, this is an acceptable level of correlation in the model. It's nearly impossible to remove all correlation and multicollinearity from a dataset, but severe cases can have a consequential impact on the final model.

Correlations

	dewp*temp	spd	vsb	slp	pcp
spd	-0.482				
vsb	-0.124	0.122			
slp	-0.225	-0.162	0.175		
pcp	-0.043	0.005	0.022	-0.128	
sd	-0.487	0.163	-0.077	0.132	0.073

Figure A2.5

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	4.405	0.545	8.08	0.000	
vsb	0.00735	0.00264	2.78	0.006	2.14
pcp	-0.0571	0.0159	-3.60	0.000	1.14
sd	-0.00227	0.00107	-2.13	0.033	2.46
spd	-0.00472	0.00177	-2.67	0.008	1.71
temp	-0.001624	0.000780	-2.08	0.038	23.55
dewp	-0.000262	0.000652	-0.40	0.688	18.97
slp	-0.001291	0.000526	-2.45	0.014	1.52
borough					
Brooklyn	1.03119	0.00971	106.22	0.000	1.60
Manhattan	1.69158	0.00971	174.24	0.000	1.60
Queens	0.80086	0.00971	82.49	0.000	1.60
Staten Island	-1.50650	0.00971	-155.18	0.000	1.60
dayType					
Wday	-0.0093	0.0179	-0.52	0.602	7.38
Wend	0.0883	0.0185	4.78	0.000	7.33
WendHday	0.0550	0.0458	1.20	0.230	1.22
month					
Feb	-0.0483	0.0188	-2.57	0.010	4.88
Jan	-0.1630	0.0149	-10.91	0.000	3.36
June	0.1655	0.0139	11.89	0.000	2.84
Mar	-0.0414	0.0135	-3.06	0.002	2.75
May	0.1190	0.0132	9.01	0.000	2.63

Figure A2.6

Furthermore, from Figure A2.6 after creating a model with all variables, observe that the VIF values for *temp* and *dewp* is greater than 10, once again indicating multicollinearity with *temp* and *dewp*. In Figure A2.7, *temp* and *dewp* are removed from the model and *temp*dewp* is introduced into the model, and there are no VIF values larger than 10 indicating that *temp*dewp* do not have a high correlation to the other independent variables or high multicollinearity, this tells us that there would be an acceptable level of correlation and multicollinearity in the model if we used a model with every independent variable. The screening methods used to remove variables is discussed more in the *Analysis: iv. Screening* section.

iii. Transformations

The original model used pickups as the response variable. From Figure A3.1, the normal probability plot is clearly not normal, and the residual plot shows signs of telescoping. Transformations need to be applied to the model to fix these issues, and all bad attempts to fix the model can be found in the *Appendix A3. Model Modification / Transformation section.*

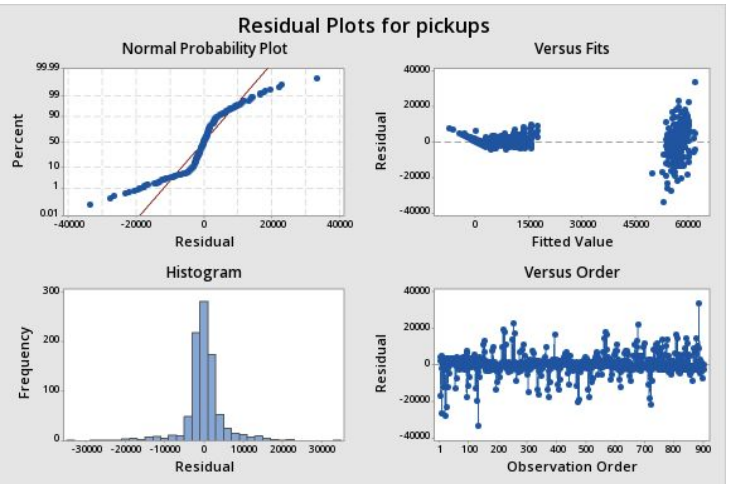

Figure A3.1

Because there is telescoping in the residual vs fit plot from Figure A3.1, the natural log transformation may be appropriate for this model. After applying a natural log transformation, the normal probability plot is much more linearized and the residual plot no longer is telescoping nor is any other pattern apparent. This can be seen in the normality plot from Figure A3.5.

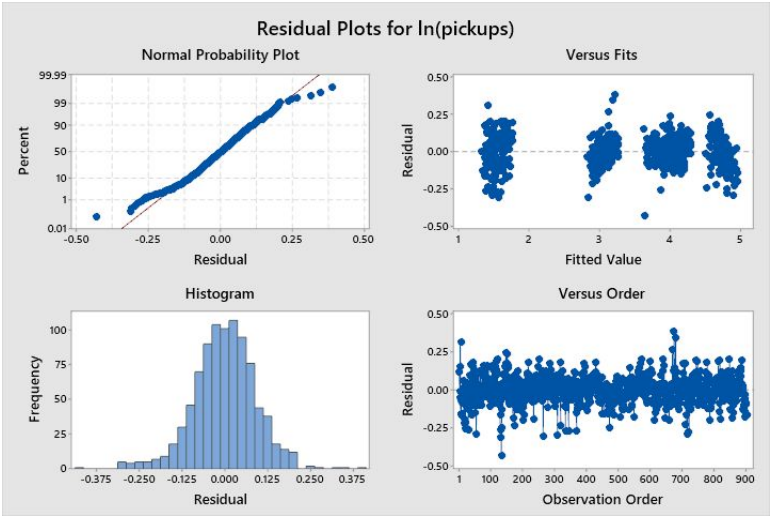

Figure A3.5

iv. Screening

To select the most appropriate model, two different screening techniques were used; backwards elimination and stepwise selection. Both techniques used an alpha of 0.1 to remove/enter to start. The step-by-step processes that resulted in the final models can be seen in Figures A4.1 and A4.4. The variables that were removed as a result of the screening include *spd*, *slp*, and the interaction term *dewp*temp*. Both screening techniques resulted in the same model that can be seen in Figures A4.2 and A4.5. This means that neither method is more accurate than the other - both have the same R-squared adjusted value of 99.29% that can be seen in Figure A4.3. The resulting model can be seen in Figure A4.7 below.

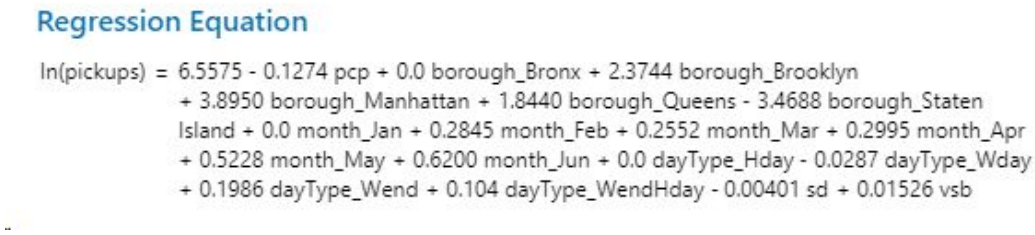

Figure A4.7

To be sure that this is the best model, the screening techniques were ran again, but instead of an alpha of 0.1, an alpha of 0.25 was used in an effort to include more variables in the model. A more complex model, including all of the variables, was the result of rerunning the backwards elimination, however the R-squared adjusted value stayed the same at 99.29% (Figures A4.10 and A4.11). A higher alpha value did not affect stepwise selection, as it resulted in the same model as when alpha was 0.1 (Figures A4.12 and A4.13). Due to the principle of parsimony, a simple, less complex model is more desired if it does not compromise the R-squared adjusted value. There is no need to include more variables if it does not increase the fit of the model. Therefore, the best model remains as the one found in Figure A4.7.

It is important to note that while some qualitative sub variables, particularly *Wday* and *WendHday* in *dayType* in Figure A4.2, have p-values greater than 0.1, those variables are still included in the model because at least one other qualitative in that class is significant, so we retain all qualitative variables in that class .

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	6.5575	0.0606	108.24	0.000	
pcp	-0.1274	0.0362	-3.52	0.000	1.10
borough					
Brooklyn	2.3744	0.0225	105.44	0.000	1.60
Manhattan	3.8950	0.0225	172.96	0.000	1.60
Queens	1.8440	0.0225	81.89	0.000	1.60
Staten Island	-3.4688	0.0225	-154.04	0.000	1.60
month					
Feb	0.2845	0.0317	8.97	0.000	2.59
Mar	0.2552	0.0262	9.73	0.000	1.93
Apr	0.2995	0.0251	11.94	0.000	1.71
May	0.5228	0.0251	20.80	0.000	1.77
Jun	0.6200	0.0248	25.00	0.000	1.68
dayType					
Wday	-0.0287	0.0414	-0.69	0.487	7.35
Wend	0.1986	0.0427	4.65	0.000	7.28
WendHday	0.104	0.106	0.99	0.325	1.21
sd	-0.00401	0.00241	-1.66	0.097	2.34
vsb	0.01526	0.00431	3.54	0.000	1.06

Figure A4.2

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.214232	99.30%	99.29%	99.28%

Figure A4.3

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	15	5828.56	388.57	8466.49	0.000
pcp	1	0.57	0.57	12.37	0.000
borough	4	5778.94	1444.73	31479.04	0.000
month	5	34.23	6.85	149.18	0.000
dayType	3	9.29	3.10	67.49	0.000
sd	1	0.13	0.13	2.76	0.097
vsb	1	0.57	0.57	12.51	0.000
Error	889	40.80	0.05		
Total	904	5869.36			

Figure A4.8

v. Assumptions

To check the independence assumption for the ln(pickups) model (error terms are independent of one another), the Durbin-Watson test statistic is used. From Figure A6.2, we see that the Durbin-Watson test statistic is lower than the low critical value for the Durbin-Watson test. This critical value was found from an online source to be 1.71755 [3]. This tells us that there *is* positive lag-1 autocorrelation between residuals.

Durbin-Watson Statistic

Durbin-Watson Statistic = 1.41551

Figure A6.2

In order to remove the positive autocorrelation, several transformations were made to the data. Information on these transformations can be found in Appendix A6. Ultimately, the model with a lag-2 difference applied was chosen and has a Durbin-Watson test statistic value of 1.7997, which is higher than the critical upper value of the Durbin-Watson Test. This tells us that the lag-2 differenced model has a low and acceptable level of positive autocorrelation.

Checking the error assumptions for the lag-2 differenced model we can see that all the assumptions hold. The other error assumptions that were checked and held for the model.

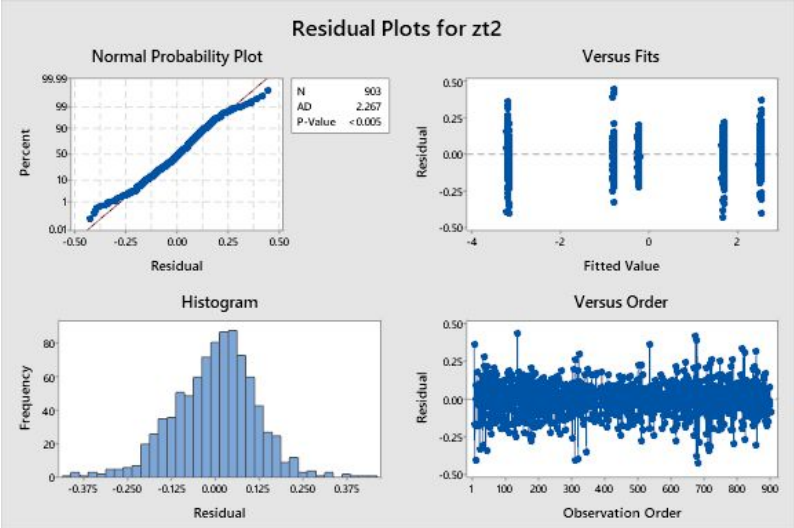

Figure A6.6

E(εi) = 0: The mean of the residuals is equal to 0

First we ensure that the expected value of each residual is equal to 0. From Figure A6.6 we see that the mean value of the residual is essentially equal to 0, so the first assumption holds.

Mean of RESIzt2 = -1.28604E-16

Figure A6.7

Normality

The Anderson Darling statistic that we got was less than 0.005. Even though the AD statistic is less than alpha, normality is assumed due to the eye test (the normality plot in Figure A6.6 appears to be mostly linear) and the Central Limit Theorem because n is so large (n = 905).

Identical distribution (constant variance)

The identical distribution assumption is not violated. Figure A6.6 shows that each of the five boroughs are centered with a mean around 0. The variance of each borough are about the same.

vi. Re-screening

Since a new transformation has been applied to the response variable, screening methods need to be applied again to select the best model. After screening the new lag-2 differenced model, both the stepwise and backward elimination methods resulted in the same model as seen in Figure A7.5 and A7.4. After screening the model, all of the error assumptions still hold, as seen in Figure A7.1 below.

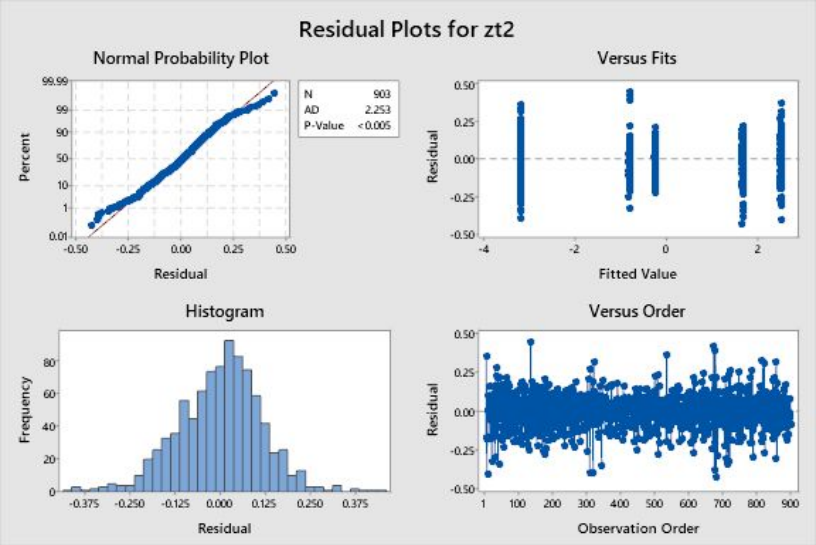

Figure A7.1

This re-screened model reaches a R-Squared Adjusted of 99.65 as seen in Figure A7.2 in Appendix A7. This model is better than the previous ln(pickups) model because it has a higher R-squared adjusted and this new model is simpler as it only has two variables, vsb and borough, as seen in Figure A7.x below.

vii. Unusual Observations

The re-screened model was fitted to include the unusual observations and their respectives HI, Cook's D, and DFITS. These can be found in Figure A5.1, the bolded values are high leverage values. All of the outlier observations were removed. The removed observations had studentized residual values greater than two, indicating that they are outlier observations. After removing the outlier observations, the model was refitted.

Regression Equation

borough	
Bronx	differences = -1.9125 + 0.00681 vsb
Brooklyn	differences = 5.7971 + 0.00681 vsb
Manhattan	differences = 3.8573 + 0.00681 vsb
Queens	differences = -0.5904 + 0.00681 vsb
Staten Island	differences = -7.4305 + 0.00681 vsb

Figure A5.2: Regression Equations

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
0.235254	99.74%	99.74%	48.2980	99.73%	-40.65	-7.43

Figure A5.3: Final Model Summary

The lag-2 differenced model was then compared to the lag-2 differenced model without the unusual observations. The model without the unusual observations had a larger adjusted R^2 value and a lower *S* value. Because both models have equal complexity, the model that performed better (the differenced model with outliers removed) was chosen to be the final model. The final model can be seen in Figure A5.2 and A5.3. Since there is a refitted model, the error assumptions were again checked for the model chosen. All the error assumptions stated earlier in *Analysis v. Assumption* hold for the final fitted model with the unusual observations removed. The error assumptions were checked using Figure A8.5.

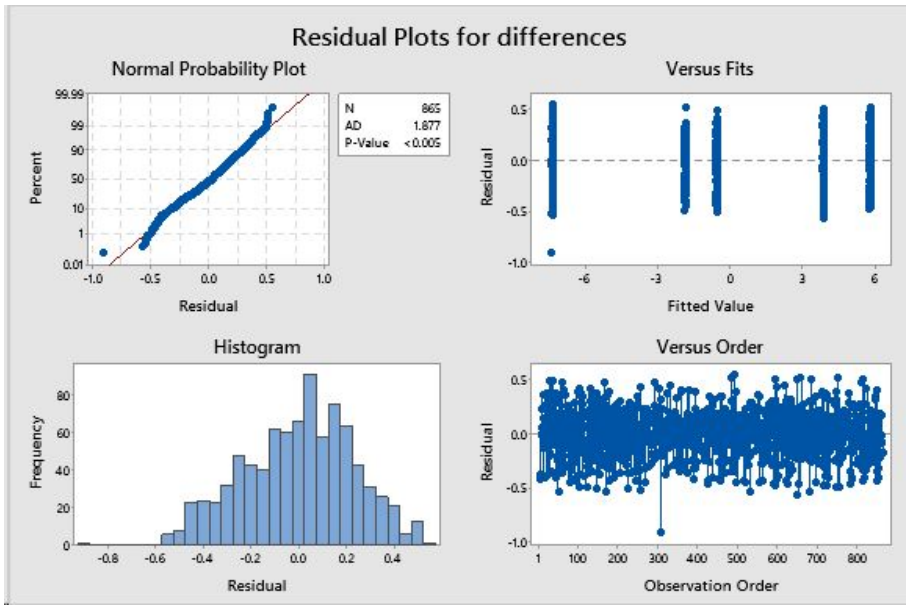

Figure A8.5: Residual Plots for Final Model

Conclusions and Recommendations

The results conclude that weather does not play a major role in how many uber pickups will occur on a given day. The borough, however, is a major factor in being able to determine the number of uber pickups in a day. This is to be expected though as the amount of traffic between boroughs and the population in the different boroughs will cause them to be significant. After transforming the model and removing the unusual observations, the final model contains the variables borough and visibility. The original prediction was that weather would be a huge role in the number of pickups because for example, it was assumed that on a rainy day there would be a lot more ubers called. The visibility variable includes more than just factors like precipitation. Rainy days are not going to be as bright as other days, but also the expected number of uber pickups on a clear sunny day is going to be less than a foggier or darker day. Another variable that was predicted to be important but was not significant was dayType. It was expected that the number of pickups on both holidays and weekends would be a lot more than weekdays and non-holidays. The reason holidays proved to not have any significance is because there are so many holidays that people do not celebrate such as President's Day and Columbus Day. The weekends are not significant because the expected number of people to go out over the weekend and call an uber is equivalent to the number of people who are going to call an uber to go to work throughout the week.

The R-squared adjusted value for the final model is 99.74%. Since the model is a lag-2 differenced model, the expected number of uber pickups for a given day is found by looking at the weather point for the visibility that day, the number of pickups that occured yesterday, and which borough the pickups is occurring in. Most importantly, Uber can use the regression model to know how to appropriately staff their available drivers to meet rider demand, which in turn will help Uber maximize their revenue and customer base in New York City. For example, Uber wants to know how many pickups to expect tomorrow in the Bronx so they can staff enough drivers to quickly pickup and deliver people to their destination. By checking any weather data source for projected visibility tomorrow, Uber can see that it's expected to be a clear sunny day (vsb = 10.00), Uber also knows that yesterday they had 7,469 pickups in Queens. By creating a prediction interval for z_t with the regression model, Uber can also know an expected range of pickups to expect. This enables Uber to know the minimum number of drivers they need for a day as well as what too many drivers is. Using the 95% prediction interval from Figure C1.1 for z_t values, Uber can apply the following equation to predict the range of total number of pickups:

$$y_t = y_{t-2}e^{z_t} = (7469)e^{-1.03093} = 2664$$

$$(7469)e^{-.561106} = 4261.64 = 4262$$

Thus, Uber knows with 95% confidence that the expected number of pickups in the Bronx tomorrow will be between 2664 and 4262 pickups.

Prediction

Fit	SE Fit	95% CI	95% PI
-0.796019	0.0093169	(-0.814304, -0.777733)	(-1.03093, -0.561106)

Figure C1.1

Another interesting point is that there were certain outlier points caused by events that could not be modeled. For example, New York City faced a blizzard in January of 2015, shutting down the New York City subway system. Naturally, people switched to Uber to get to work that day. This led to a massive boom in number of Uber pickups for that day and also had a large residual value, decreasing the adjusted R^2 value of the model as well. These freak occurrences that cannot be modeled led to increased loss in the original model, which is why the unusual observations were thrown out. After the unusual observations were thrown out, the adjusted R-squared increased from 99.65 to 99.74 .

It's also worth noting that all of the high-leverage observations that occurred were on holidays, and the prediction for each borough was significantly off for the holiday. It's interesting to note that there was a holiday qualitative variable that was removed during the screening process since the entire class was insignificant. This means that only some holidays have an impact on the number of total pickups for Uber, and others have little to no impact. It would be in Uber's best interest to, going forward, track which holidays have a surge of Uber pickups and create a binary qualitative variable for these surge holidays versus regular days.

Lastly, there are some flaws in the model. Mainly, applying the lag-2 difference to the model removes predictive power from the model. This occurs because of the way data is collected in the original dataset, there are five consecutive data points collected on one day, and each of these data points is identical with only borough and number of pickups changing. Naturally, points collected on the same day in different boroughs will be correlated, which is why the lag-2 difference transformation was applied. But this difference is not usable for several of the boroughs because in order to predict the number of pickups for a given borough (ex: Manhattan) then you need to know the number of pickups that occur in the Bronx for the same day. This is paradoxical and absolutely impairs the model from a lot of usability since we won't be able to predict a day until we have data on that day.

Despite this, the model can still form predictions for Bronx and Brooklyn because the ordering of their occurences in the dataset means their predictions depend on the previous day number of pickups in Queens and Staten Island respectively. Knowing this, we could reorder the original dataset to have any two boroughs (i.e. Manhattan and Queens) be the first occurrences for each day in the dataset and refit the model. This would enable predictions to be made on those two boroughs using the previous days number of pickups in the other boroughs.

Ultimately, going forward it would help in fitting and usability to create separate models for each borough to predict number of pickups. The qualitative variable borough has so much influence on the fit of the model and creates large clusters of points in the fitted line. Having multiple data points collected for the same day also significantly hampers the usability of the model as stated earlier (because of autocorrelated data).

Appendix

A1. Independent vs Dependent Plots

Model: pickups ~ spd + vsb + temp + dewp + slp + pcp + sd

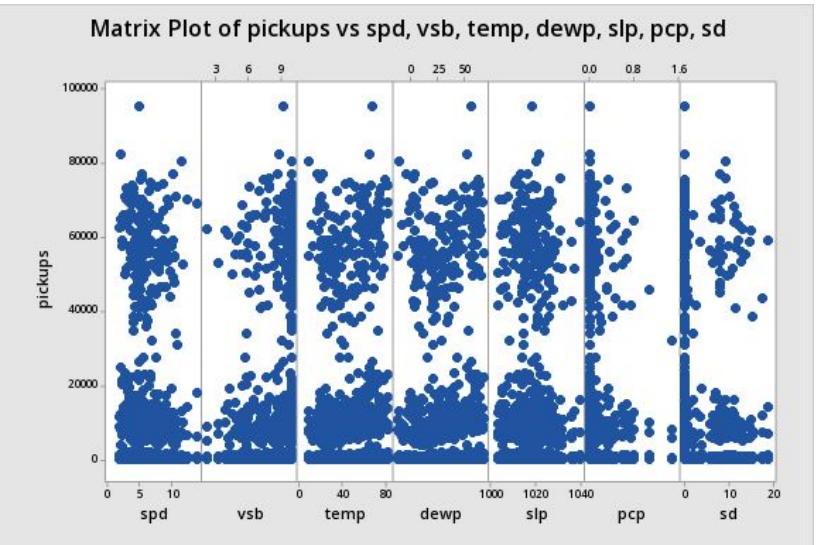

Figure A1.1: Matrix plot 1

*Model: pickups ~ dewp*temp + spd + vsb + temp + dewp + slp + pcp + sd*

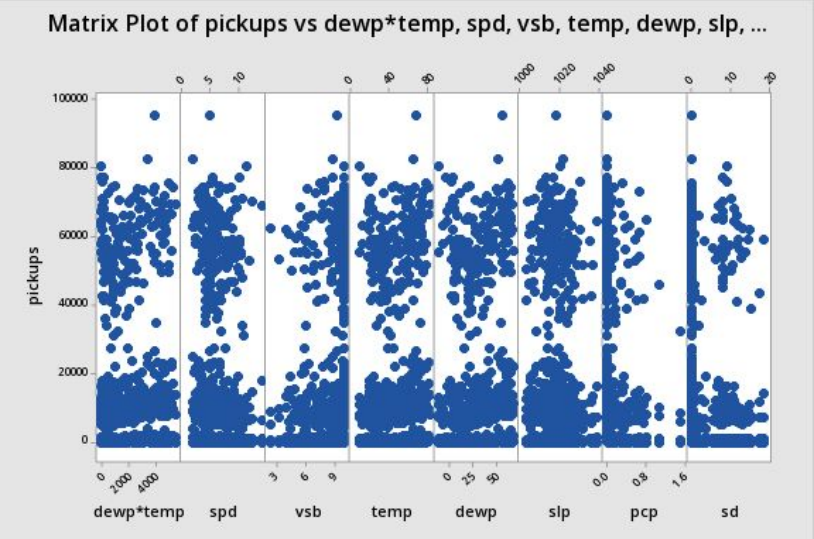

Figure A1.2: Matrix plot 2

*Model: ln(pickups) ~ dewp*temp + spd + vsb + temp + dewp + slp + pcp + sd*

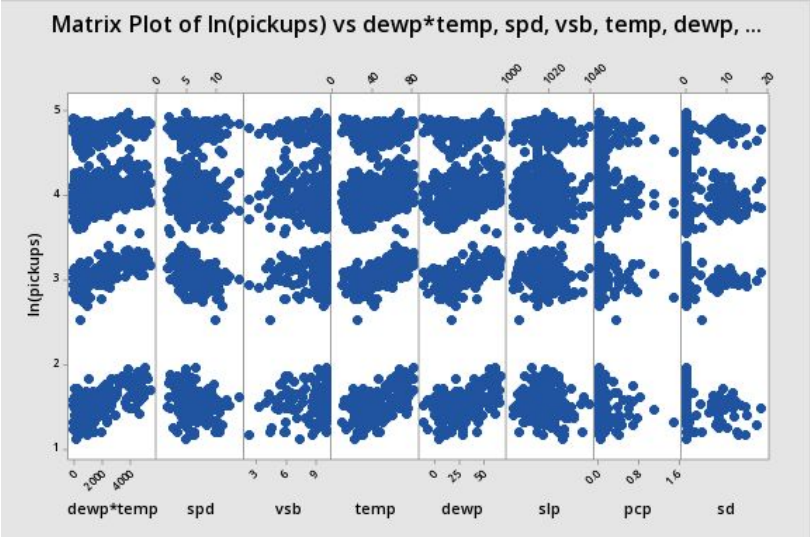

Figure A1.3: Matrix plot 3

Model: ln(pickups) ~ vsb + pcp + sd

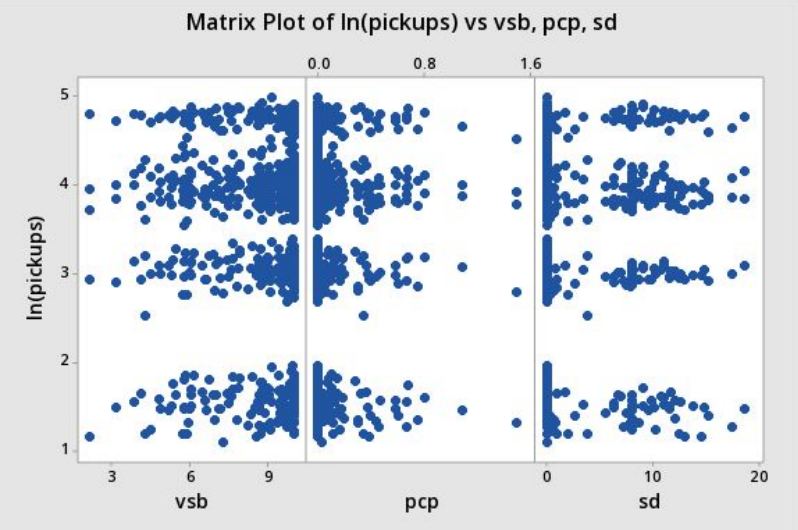

Figure A1.4: Matrix plot 4

Model: zt_2 ~ vsb + pcp + sd

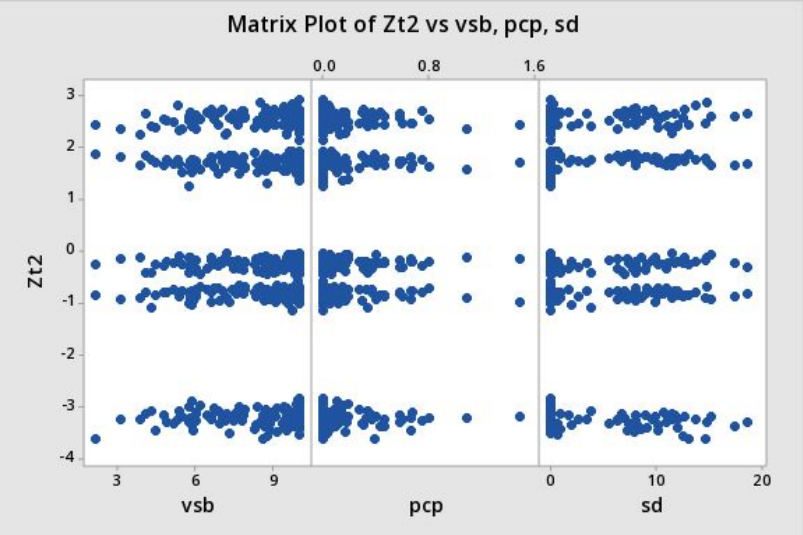

Figure A1.5: Matrix plot 5

A2. Correlation

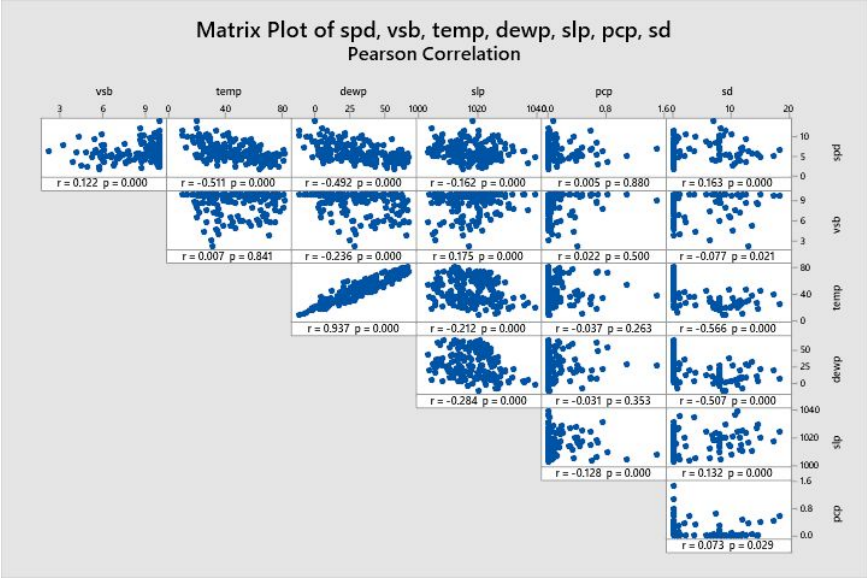

Figure A2.1: Matrix plot of correlation between quantitative independent variables

Correlations

	spd	vsb	temp	dewp	slp	pcp
vsb	0.122					
temp	-0.511	0.007				
dewp	-0.492	-0.236	0.937			
slp	-0.162	0.175	-0.212	-0.284		
pcp	0.005	0.022	-0.037	-0.031	-0.128	
sd	0.163	-0.077	-0.566	-0.507	0.132	0.073

Figure A2.2: Correlations between quantitative independent variables

Correlations of attempted transformations

Correlations

	ln(temp)	spd	vsb	dewp	slp	pcp
spd	-0.519					
vsb	-0.025	0.122				
dewp	0.913	-0.492	-0.236			
slp	-0.230	-0.162	0.175	-0.284		
pcp	-0.031	0.005	0.022	-0.031	-0.128	
sd	-0.583	0.163	-0.077	-0.507	0.132	0.073

Figure A2.3: Correlations after ln(temp)

Correlations

	temp^2	spd	vsb	dewp	slp	pcp
spd	-0.493					
vsb	0.023	0.122				
dewp	0.923	-0.492	-0.236			
slp	-0.191	-0.162	0.175	-0.284		
pcp	-0.038	0.005	0.022	-0.031	-0.128	
sd	-0.525	0.163	-0.077	-0.507	0.132	0.073

Figure A2.4: Correlations after temp^2

Correlations

	dewp*temp	spd	vsb	slp	pcp
spd	-0.482				
vsb	-0.124	0.122			
slp	-0.225	-0.162	0.175		
pcp	-0.043	0.005	0.022	-0.128	
sd	-0.487	0.163	-0.077	0.132	0.073

Figure A2.5: Correlations after dewp*temp

Initial model with all variables

pickups ~ vsb + pcp + sd + spd + temp + dewp + slp + borough + dayType + month

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	4.405	0.545	8.08	0.000	
vsb	0.00735	0.00264	2.78	0.006	2.14
pcp	-0.0571	0.0159	-3.60	0.000	1.14
sd	-0.00227	0.00107	-2.13	0.033	2.46
spd	-0.00472	0.00177	-2.67	0.008	1.71
temp	-0.001624	0.000780	-2.08	0.038	23.55
dewp	-0.000262	0.000652	-0.40	0.688	18.97
slp	-0.001291	0.000526	-2.45	0.014	1.52
borough					
Brooklyn	1.03119	0.00971	106.22	0.000	1.60
Manhattan	1.69158	0.00971	174.24	0.000	1.60
Queens	0.80086	0.00971	82.49	0.000	1.60
Staten Island	-1.50650	0.00971	-155.18	0.000	1.60
dayType					
Wday	-0.0093	0.0179	-0.52	0.602	7.38
Wend	0.0883	0.0185	4.78	0.000	7.33
WendHday	0.0550	0.0458	1.20	0.230	1.22
month					
Feb	-0.0483	0.0188	-2.57	0.010	4.88
Jan	-0.1630	0.0149	-10.91	0.000	3.36
June	0.1655	0.0139	11.89	0.000	2.84
Mar	-0.0414	0.0135	-3.06	0.002	2.75
May	0.1190	0.0132	9.01	0.000	2.63

Figure A2.6: Coefficients with all variables

*Model with the interaction term temp*dewp*

pickups ~ vsb + pcp + sd + spd + temp*dewp + slp + borough + dayType + month

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	8.30	1.20	6.93	0.000	
dewp*temp	-0.000017	0.000012	-1.36	0.174	7.60
pcp	-0.1352	0.0368	-3.68	0.000	1.14
borough					
Brooklyn	2.3744	0.0225	105.51	0.000	1.60
Manhattan	3.8950	0.0225	173.07	0.000	1.60
Queens	1.8440	0.0225	81.94	0.000	1.60
Staten Island	-3.4688	0.0225	-154.14	0.000	1.60
month					
Feb	0.2777	0.0322	8.63	0.000	2.67
Mar	0.2533	0.0268	9.46	0.000	2.01
Apr	0.3091	0.0295	10.48	0.000	2.37
May	0.5523	0.0432	12.80	0.000	5.22
Jun	0.6575	0.0483	13.61	0.000	6.37
dayType					
Wday	-0.0272	0.0413	-0.66	0.511	7.35
Wend	0.1980	0.0427	4.64	0.000	7.28
WendHday	0.118	0.106	1.11	0.268	1.23
spd	-0.00621	0.00390	-1.59	0.111	1.55
slp	-0.00166	0.00117	-1.42	0.156	1.40
sd	-0.00401	0.00245	-1.64	0.102	2.41
vsb	0.01561	0.00461	3.39	0.001	1.21

Figure A2.7: Coefficients after temp*dewp

H_0 : *no first order autocorrelation*

H_1 : *first order autocorrelation exits*

$d = \frac{\sum_{t=2}^{n}(\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=1}^{n}(\hat{\varepsilon}_t)^2}$, *Rejection region* : $d < d_{L,alpha/2} or\ (4-d) < d_{L,alpha/2}$

$Non-rejection\ region$: $d > d_{U,alpha/2} or\ (4-d) < d_{U,alpha/2}$, *Inconclusive region* : *Any other result*

Figure A2.8: Hypothesis testing criteria for the Durbin Watson test

A3. Model Modification / Transformation

pickups

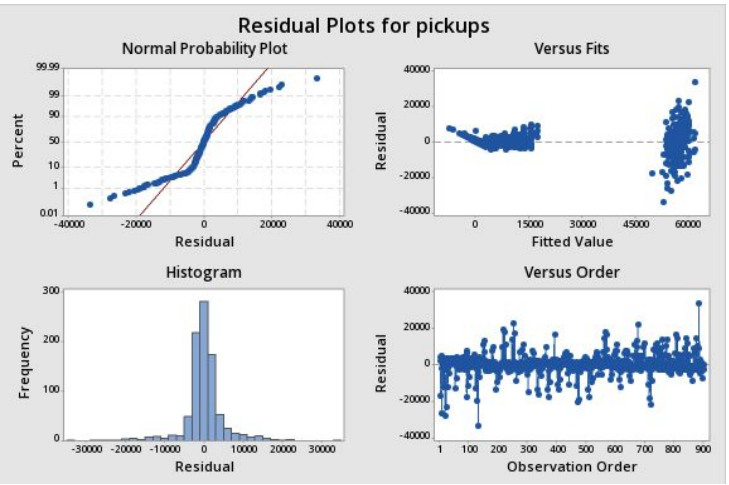

Figure A3.1: Residual plots for pickups

Attempts to deal with the linearization of normality probability plot and constant variance include taking the square root of pickups, squaring pickups, taking the cube root of pickups, and taking the natural log of our pickups.

$pickups^{1/2}$

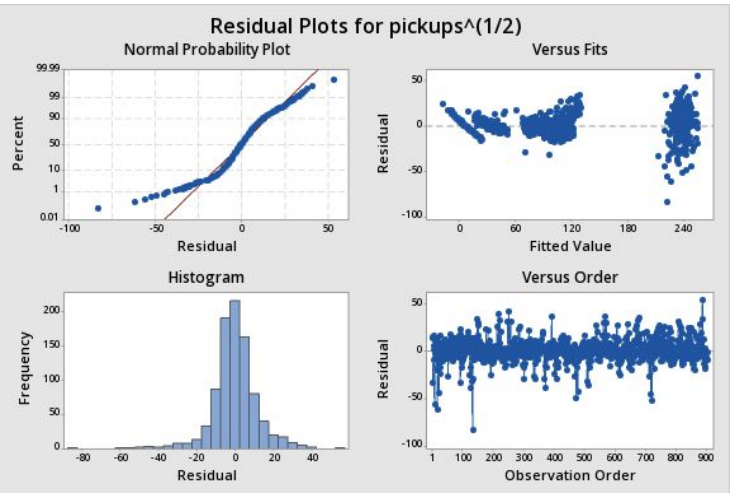

Figure A3.2: Residual plots for pickups^(½)

When data behaves like a Poisson distribution, taking the square root of the response variable tends to stabilize the variance for data; however, this is not the distribution we are dealing with. Taking the square root did straighten out the normal probability plot some, but there is still a S-shape and the residual plot still has an obvious telescoping pattern.

pickups^2

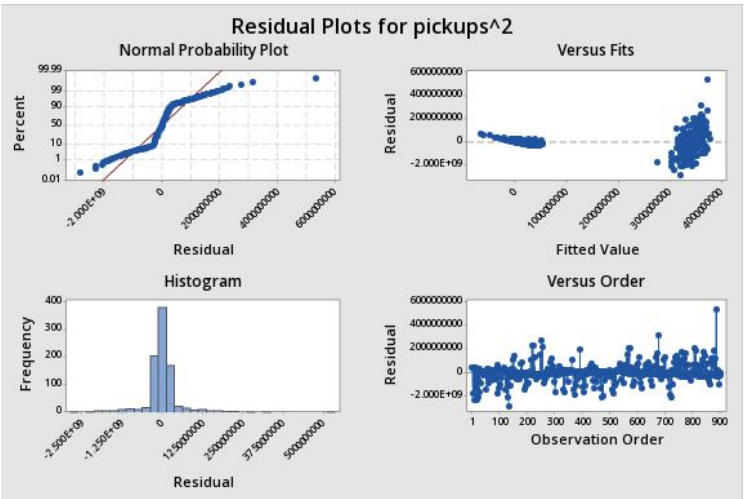

Figure A3.3: Residual plots for pickups^2

When the response variable (pickups) is squared, this only made the model worse. The S-shape on the normal probability plot and the telescoping on the residual plot are still there.

pickups$^{1/3}$

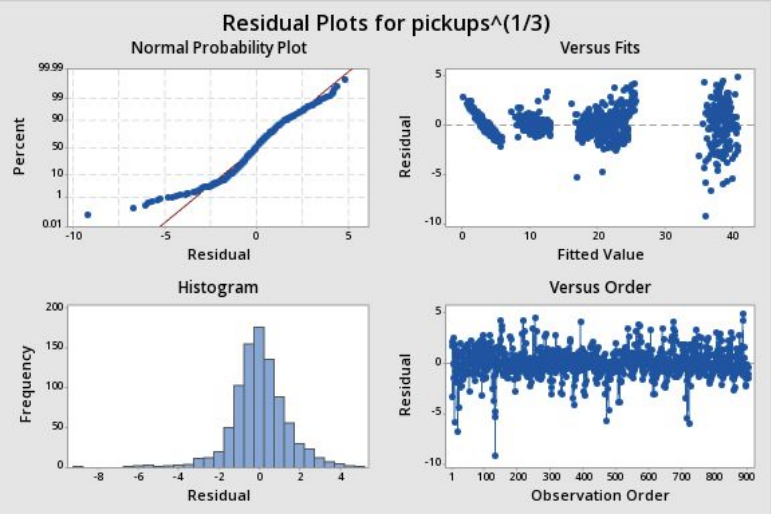

Figure A3.4: Residual plots for pickups^(⅓)

Taking the cube root of pickups did help minimize our issues, but the normal probability plot is not as strong of a fit as the natural log transformation, and the residual plot still show some signs of telescoping.

ln(pickups)$^{1/3}$

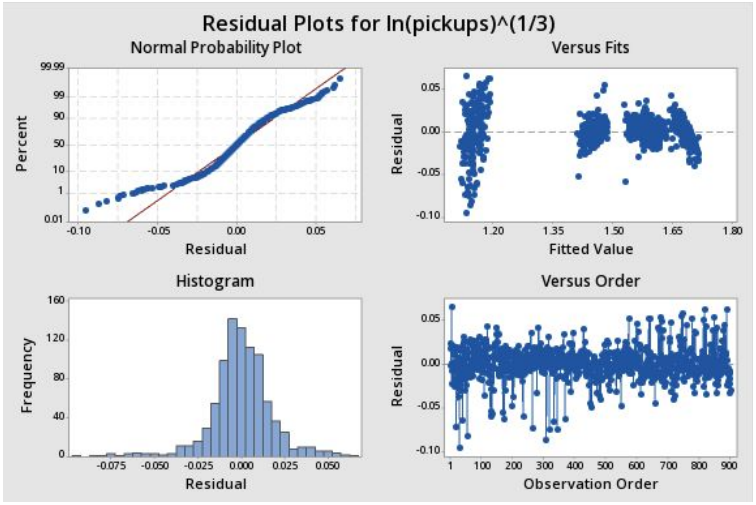

Figure A3.6: Residual plots for ln(pickups)^(⅓)

One last attempt was to try the cubic root of the natural log of our pickups. By looking at figure A3.5, this combination was not favorable. The normal probability plot is not as linear, and the residual plot reintroduces telescoping. After attempting all of these transformations, it was concluded that none of these were the right transformation for our model.

A4. Model Selection and Screening Techniques

Backwards elimination (alpha to remove = 0.1)

Backward Elimination of Terms

Candidate terms: dewp*temp, pcp, borough, month, dayType, spd, slp, sd, vsb

	Step 1		Step 2		Step 3	
	Coef	P	Coef	P	Coef	P
Constant	8.30		7.76		6.5823	
dewp*temp	-0.000017	0.174				
pcp	-0.1352	0.000	-0.1359	0.000	-0.1294	0.000
borough	3.8950	0.000	3.8950	0.000	3.8950	0.000
month	0.6575	0.000	0.6031	0.000	0.6099	0.000
dayType	0.1980	0.000	0.1981	0.000	0.1976	0.000
spd	-0.00621	0.111	-0.00499	0.189	-0.00408	0.270
slp	-0.00166	0.156	-0.00116	0.297		
sd	-0.00401	0.102	-0.00379	0.122	-0.00417	0.085
vsb	0.01561	0.001	0.01703	0.000	0.01583	0.000
S	0.214091		0.214194		0.214205	
R-sq	99.31%		99.31%		99.31%	
R-sq(adj)	99.29%		99.29%		99.29%	
Mallows' Cp	19.00		18.85		17.94	
AICc	-199.82		-200.02		-201.00	
BIC	-104.61		-109.53		-115.23	

	Step 4	
	Coef	P
Constant	6.5575	
dewp*temp		
pcp	-0.1274	0.000
borough	3.8950	0.000
month	0.6200	0.000
dayType	0.1986	0.000
spd		
slp		
sd	-0.00401	0.097
vsb	0.01526	0.000
S	0.214232	
R-sq	99.30%	
R-sq(adj)	99.29%	
Mallows' Cp	17.16	
AICc	-201.84	
BIC	-120.79	

α to remove = 0.1

If a term has more than one coefficient, the largest in magnitude is shown.

Figure A4.1: Step-by-step process for backwards elimination at alpha = 0.1 for ln(pickups) model

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	6.5575	0.0606	108.24	0.000	
pcp	-0.1274	0.0362	-3.52	0.000	1.10
borough					
Brooklyn	2.3744	0.0225	105.44	0.000	1.60
Manhattan	3.8950	0.0225	172.96	0.000	1.60
Queens	1.8440	0.0225	81.89	0.000	1.60
Staten Island	-3.4688	0.0225	-154.04	0.000	1.60
month					
Feb	0.2845	0.0317	8.97	0.000	2.59
Mar	0.2552	0.0262	9.73	0.000	1.93
Apr	0.2995	0.0251	11.94	0.000	1.71
May	0.5228	0.0251	20.80	0.000	1.77
Jun	0.6200	0.0248	25.00	0.000	1.68
dayType					
Wday	-0.0287	0.0414	-0.69	0.487	7.35
Wend	0.1986	0.0427	4.65	0.000	7.28
WendHday	0.104	0.106	0.99	0.325	1.21
sd	-0.00401	0.00241	-1.66	0.097	2.34
vsb	0.01526	0.00431	3.54	0.000	1.06

Figure A4.2: Coefficients for backwards elimination at alpha = 0.1 for ln(pickups) model

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.214232	99.30%	99.29%	99.28%

Figure A4.3: Model summary for backwards elimination at alpha = 0.1 for ln(pickups) model

Stepwise selection (alpha to enter/remove = 0.1)

Stepwise Selection of Terms

Candidate terms: pcp, borough, month, dayType, spd, slp, sd, vsb, dewp*temp

	Step 1		Step 2		Step 3		Step 4	
	Coef	P	Coef	P	Coef	P	Coef	P
Constant	7.0377		6.7066		6.6557		6.6995	
borough	3.8950	0.000	3.8950	0.000	3.8950	0.000	3.8950	0.000
month			0.6196	0.000	0.6254	0.000	0.6227	0.000
dayType					0.2097	0.000	0.1813	0.000
pcp							-0.1316	0.000
vsb								
sd								
S	0.316966		0.239587		0.217407		0.215931	
R-sq	98.46%		99.12%		99.28%		99.29%	
R-sq(adj)	98.45%		99.12%		99.27%		99.28%	
Mallows' Cp	1077.74		235.86		40.84		29.37	
AICc	495.74		-5.68		-178.37		-189.66	
BIC	524.50		46.91		-111.53		-118.08	

	Step 5		Step 6	
	Coef	P	Coef	P
Constant	6.5529		6.5575	
borough	3.8950	0.000	3.8950	0.000
month	0.6253	0.000	0.6200	0.000
dayType	0.1946	0.000	0.1986	0.000
pcp	-0.1353	0.000	-0.1274	0.000
vsb	0.01577	0.000	0.01526	0.000
sd			-0.00401	0.097
S	0.214443		0.214232	
R-sq	99.30%		99.30%	
R-sq(adj)	99.29%		99.29%	
Mallows' Cp	17.93		17.16	
AICc	-201.11		-201.84	
BIC	-124.80		-120.79	

α to enter = 0.1, α to remove = 0.1
If a term has more than one coefficient, the largest in magnitude is shown.

Figure A4.4: Step-by-step process for stepwise selection at alpha = 0.1 for ln(pickups) model

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	6.5575	0.0606	108.24	0.000	
pcp	-0.1274	0.0362	-3.52	0.000	1.10
borough					
Brooklyn	2.3744	0.0225	105.44	0.000	1.60
Manhattan	3.8950	0.0225	172.96	0.000	1.60
Queens	1.8440	0.0225	81.89	0.000	1.60
Staten Island	-3.4688	0.0225	-154.04	0.000	1.60
month					
Feb	0.2845	0.0317	8.97	0.000	2.59
Mar	0.2552	0.0262	9.73	0.000	1.93
Apr	0.2995	0.0251	11.94	0.000	1.71
May	0.5228	0.0251	20.80	0.000	1.77
Jun	0.6200	0.0248	25.00	0.000	1.68
dayType					
Wday	-0.0287	0.0414	-0.69	0.487	7.35
Wend	0.1986	0.0427	4.65	0.000	7.28
WendHday	0.104	0.106	0.99	0.325	1.21
sd	-0.00401	0.00241	-1.66	0.097	2.34
vsb	0.01526	0.00431	3.54	0.000	1.06

Figure A4.5: Coefficients for stepwise selection at alpha = 0.1 for ln(pickups) model

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.214232	99.30%	99.29%	99.28%

Figure A4.6: Model summary for stepwise selection at alpha = 0.1 for ln(pickups) model

After running both backwards elimination and stepwise selection screening techniques, the same model is produced, given that the alpha to remove/enter is the same at 0.1. The resulting model is below.

Regression Equation

ln(pickups) = 6.5575 - 0.1274 pcp + 0.0 borough_Bronx + 2.3744 borough_Brooklyn + 3.8950 borough_Manhattan + 1.8440 borough_Queens - 3.4688 borough_Staten Island + 0.0 month_Jan + 0.2845 month_Feb + 0.2552 month_Mar + 0.2995 month_Apr + 0.5228 month_May + 0.6200 month_Jun + 0.0 dayType_Hday - 0.0287 dayType_Wday + 0.1986 dayType_Wend + 0.104 dayType_WendHday - 0.00401 sd + 0.01526 vsb

Figure A4.7: Selected model at alpha = 0.1 for ln(pickups) model

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	15	5828.56	388.57	8466.49	0.000
pcp	1	0.57	0.57	12.37	0.000
borough	4	5778.94	1444.73	31479.04	0.000
month	5	34.23	6.85	149.18	0.000
dayType	3	9.29	3.10	67.49	0.000
sd	1	0.13	0.13	2.76	0.097
vsb	1	0.57	0.57	12.51	0.000
Error	889	40.80	0.05		
Total	904	5869.36			

Figure A4.8: Analysis of variance at alpha = 0.1 for ln(pickups) model

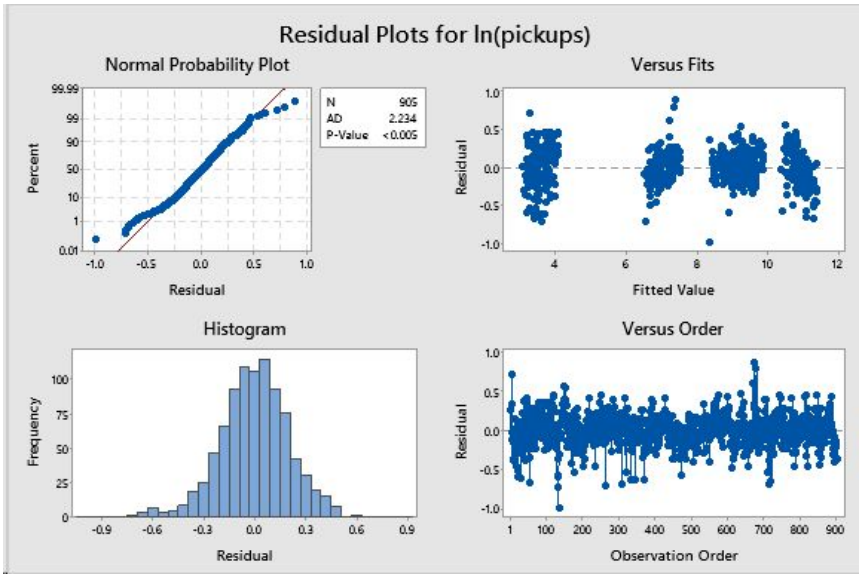

Figure A4.9: Residual plot for ln(pickups) at alpha = 0.1 for ln(pickups) model

Backwards elimination (alpha to remove = 0.25)

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.214091	99.31%	99.29%	99.28%

Figure A4.10: Model summary for backwards elimination at alpha = 0.25 for ln(pickups) model

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	18	5828.75	323.82	7064.87	0.000
dewp*temp	1	0.09	0.09	1.85	0.174
pcp	1	0.62	0.62	13.54	0.000
borough	4	5778.94	1444.73	31520.25	0.000
month	5	10.98	2.20	47.92	0.000
dayType	3	9.02	3.01	65.60	0.000
spd	1	0.12	0.12	2.54	0.111
slp	1	0.09	0.09	2.02	0.156
sd	1	0.12	0.12	2.68	0.102
vsb	1	0.53	0.53	11.46	0.001
Error	886	40.61	0.05		
Total	904	5869.36			

Figure A4.11: Analysis of variance for backwards elimination at alpha = 0.25 for ln(pickups) model

Stepwise selection (alpha to enter/remove = 0.25)

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.214232	99.30%	99.29%	99.28%

Figure A4.12: Model summary for forwards selection at alpha = 0.25 for ln(pickups) model

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	15	5828.56	388.57	8466.49	0.000
pcp	1	0.57	0.57	12.37	0.000
borough	4	5778.94	1444.73	31479.04	0.000
month	5	34.23	6.85	149.18	0.000
dayType	3	9.29	3.10	67.49	0.000
sd	1	0.13	0.13	2.76	0.097
vsb	1	0.57	0.57	12.51	0.000
Error	889	40.80	0.05		
Total	904	5869.36			

Figure A4.13: Analysis of variance for forwards selection at alpha = 0.25 for ln(pickups) model

A5. Unusual Observations

Fits and Diagnostics for Unusual Observations												
Obs	differences	Fit	SE Fit	95% CI	Resid	Std Resid	Del Resid	HI	Cook's D	DFITS		
5	-6.5174	-7.3534	0.0214	(-7.3954, -7.311	0.836	3.05	3.07	0.0060593	0.01	0.239402	R	
7	4.9268	5.8531	0.0215	(5.8110, 5.8952	-0.9262	-3.38	-3.4	0.0060959	0.01	-0.266359	R	
21	-2.5891	-1.8353	0.0211	(-1.8766, -1.794	-0.7538	-2.75	-2.76	0.0058773	0.01	-0.212352	R	
30	-8.0631	-7.3772	0.022	(-7.4204, -7.334	-0.6859	-2.5	-2.51	0.0063954	0.01	-0.201504	R	
32	6.5138	5.8507	0.0211	(5.8093, 5.8920	0.6631	2.42	2.43	0.0058773	0.01	0.186623	R	
40	-8.1304	-7.3534	0.0214	(-7.3954, -7.311	-0.777	-2.84	-2.85	0.0060593	0.01	-0.222338	R	
131	-2.4649	-1.8833	0.0317	(-1.9456, -1.821	-0.5816	-2.13	-2.14	0.013319	0.01	-0.248069	R	
136	-0.7942	-1.8329	0.0215	(-1.8750, -1.790	1.0387	3.79	3.82	0.0060959	0.01	0.299208	R	
170	-7.9563	-7.3632	0.0204	(-7.4033, -7.323	-0.5931	-2.16	-2.17	0.0055268	0	-0.161686	R	
175	-7.9961	-7.3611	0.0205	(-7.4013, -7.320	-0.6349	-2.32	-2.32	0.0055615	0.01	-0.173716	R	
302	5.2406	5.8282	0.0223	(5.7845, 5.8719	-0.5876	-2.15	-2.15	0.0065624	0.01	-0.174723	R	
310	-8.2756	-7.3658	0.0205	(-7.4060, -7.325	-0.9098	-3.32	-3.34	0.005544	0.01	-0.249301	R	
312	6.4404	5.8116	0.0278	(5.7571, 5.8661	0.6288	2.3	2.31	0.0102209	0.01	0.234313	R	
316	**-1.9439**	**-1.9023**	**0.0412**	**(-1.9831, -1.821**	**-0.0416**	**-0.15**	**-0.15**	**0.0224528**	**0**	**-0.023197**		**X**
317	**5.6046**	**5.7837**	**0.0412**	**(5.7029, 5.8645**	**-0.1791**	**-0.66**	**-0.66**	**0.0224528**	**0**	**-0.099882**		**X**
318	**4.2888**	**3.8361**	**0.0412**	**(3.7553, 3.9169**	**0.4527**	**1.67**	**1.67**	**0.0224555**	**0.01**	**0.252839**		**X**
319	**-0.5528**	**-0.5893**	**0.0412**	**(-0.6701, -0.508**	**0.0365**	**0.13**	**0.13**	**0.0224555**	**0**	**0.02035**		**X**
320	**-8.3295**	**-7.4228**	**0.0412**	**(-7.5036, -7.342**	**-0.9068**	**-3.34**	**-3.36**	**0.0224555**	**0.04**	**-0.508768**	**R**	**X**
322	6.5758	5.8397	0.0206	(5.7994, 5.8801	0.7361	2.69	2.7	0.0055944	0.01	0.202205	R	
345	-8.1636	-7.3645	0.0204	(-7.4045, -7.324	-0.7992	-2.92	-2.93	0.0055265	0.01	-0.218339	R	
450	-7.9253	-7.3575	0.0208	(-7.3983, -7.316	-0.5678	-2.07	-2.08	0.0057193	0	-0.157465	R	
512	5.2196	5.8531	0.0215	(5.8110, 5.8952	-0.6335	-2.31	-2.32	0.0060959	0.01	-0.181551	R	
537	6.7089	5.8531	0.0215	(5.8110, 5.8952	0.8558	3.12	3.14	0.0060959	0.01	0.245887	R	
617	5.2249	5.8487	0.0208	(5.8078, 5.8896	-0.6238	-2.28	-2.28	0.0057408	0	-0.173424	R	
645	-6.8101	-7.3873	0.0249	(-7.4361, -7.338	0.5771	2.11	2.11	0.0081964	0.01	0.192105	R	
676	-0.8641	-1.8437	0.0205	(-1.8839, -1.803	0.9796	3.58	3.6	0.0055558	0.01	0.268995	R	
678	3.0144	3.8947	0.0204	(3.8546, 3.9348	-0.8803	-3.21	-3.23	0.0055253	0.01	-0.240717	R	
681	-0.961	-1.8704	0.0262	(-1.9218, -1.818	0.9094	3.32	3.34	0.0091008	0.02	0.320431	R	
683	2.8812	3.868	0.0262	(3.8166, 3.9194	-0.9868	-3.61	-3.63	0.0090854	0.02	-0.347813	R	
687	5.1636	5.8277	0.0224	(5.7838, 5.8716	-0.6641	-2.42	-2.43	0.0066251	0.01	-0.198577	R	
713	3.2084	3.9055	0.0214	(3.8635, 3.9475	-0.6971	-2.54	-2.55	0.0060593	0.01	-0.199287	R	
718	3.1147	3.9055	0.0214	(3.8635, 3.9475	-0.7908	-2.89	-2.9	0.0060593	0.01	-0.226316	R	
720	-6.8009	-7.3534	0.0214	(-7.3954, -7.311	0.5525	2.02	2.02	0.0060593	0	0.157753	R	
723	3.1662	3.9055	0.0214	(3.8635, 3.9475	-0.7393	-2.7	-2.71	0.0060593	0.01	-0.211461	R	
725	-6.7449	-7.3534	0.0214	(-7.3954, -7.311	0.6084	2.22	2.23	0.0060593	0.01	0.173803	R	
750	-6.7515	-7.3541	0.0213	(-7.3958, -7.312	0.6025	2.2	2.2	0.0059935	0	0.171155	R	
790	-6.7288	-7.3534	0.0214	(-7.3954, -7.311	0.6246	2.28	2.29	0.0060593	0.01	0.178453	R	
792	5.1743	5.8527	0.0214	(5.8107, 5.8947	-0.6784	-2.48	-2.48	0.0060623	0.01	-0.193962	R	
820	-6.5783	-7.3537	0.0213	(-7.3956, -7.311	0.7754	2.83	2.84	0.0060258	0.01	0.221261	R	
825	-6.6315	-7.3534	0.0214	(-7.3954, -7.311	0.7219	2.64	2.64	0.0060593	0.01	0.206439	R	
827	5.1795	5.7986	0.0336	(5.7326, 5.8646	-0.6191	-2.27	-2.28	0.0149755	0.01	-0.280573	R	
860	-6.6395	-7.3901	0.0259	(-7.4409, -7.339	0.7506	2.74	2.75	0.0088827	0.01	0.26071	R	
862	5.1484	5.8488	0.0208	(5.8079, 5.8897	-0.7003	-2.56	-2.56	0.0057453	0.01	-0.194912	R	

Figure A5.1: Unusual observations for difference model, bolded rows were not removed from model

A6. Error Assumption Checks

Plots and statistics for the following model

ln(pickups) ~ vsb + pcp + sd + borough + dayType + month

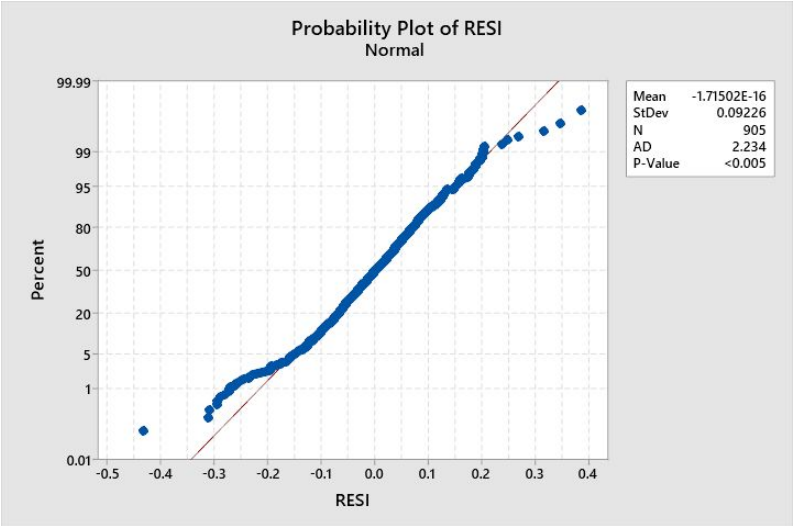

Figure A6.1: Probability plot for ln(pickups) model

Durbin-Watson Statistic

Durbin-Watson Statistic = 1.41551

Figure A6.2: Durbin-Watson test statistic for ln(pickups) model

Plots and statistics for the following time-series model

$z_t \sim vsb + pcp + sd + borough + dayType + month$

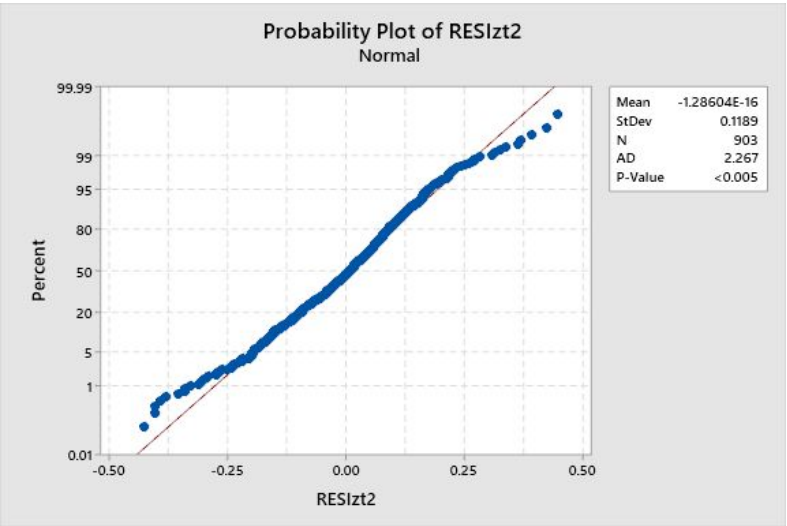

Figure A6.3: Probability plot for z_t model

Durbin-Watson Statistic

Durbin-Watson Statistic = 1.79777

Figure A6.4: Durbin-Watson test statistic for z_t model

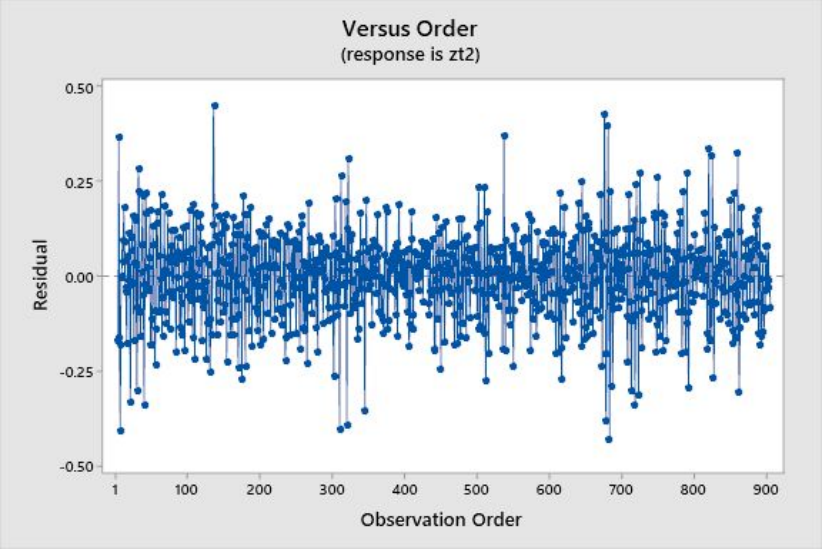

Figure A6.5: Versus order for z_t model

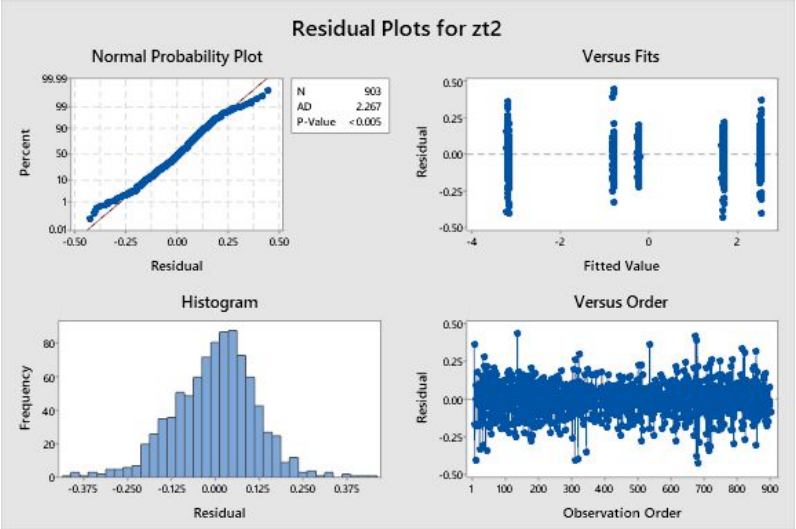

Figure A6.6: Residual plots for lag-2 differenced model

Mean of RESIzt2 = -1.28604E-16

Figure A6.7: Mean of the residuals for lag-2 differenced model

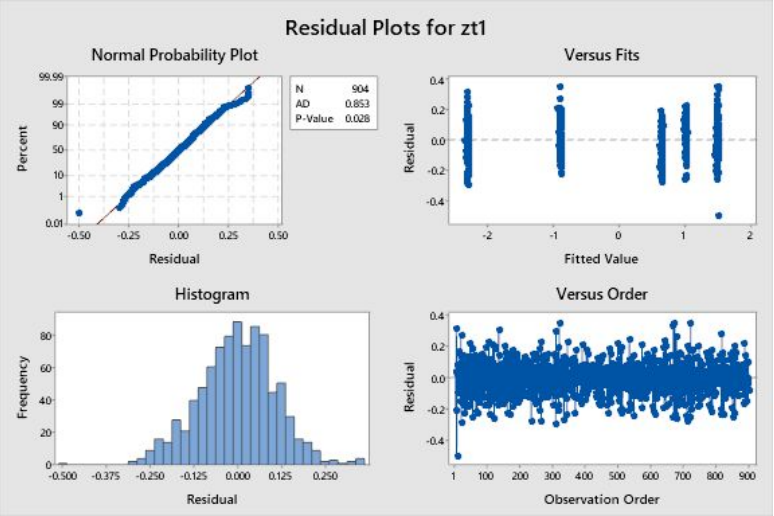

Figure A6.8: Residual plots for lag-1 differenced model

Figure A6.8: Durbin-Watson test statistic for lag-1 differenced model

A7: Rescreened Model

z_t ~ vsb + borough

Where $z_t = y_t - y_{t-2}$

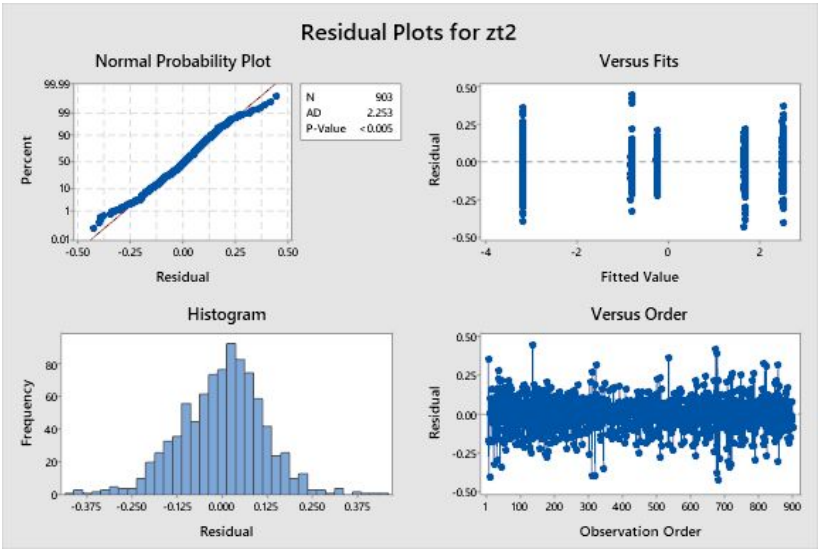

Figure A7.1: Rescreened model after lag-2 difference transformation applied

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.274769	99.65%	99.65%	99.65%

Figure A7.2: Adjusted R-squared for the rescreened model

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	5	19368.1	3873.62	51307.52	0.000
vsb	1	0.2	0.21	2.72	0.100
borough	4	19368.0	4841.99	64133.97	0.000
Error	897	67.7	0.08		
Lack-of-Fit	762	62.2	0.08	1.98	0.000
Pure Error	135	5.6	0.04		
Total	902	19435.8			

Figure A7.3: ANOVA for the rescreened model

Backward Elimination of Terms

Candidate terms: vsb, pcp, sd, borough, dayType, month

	------Step 1-----		------Step 2-----		------Step 3-----		------Step 4-----	
	Coef	P	Coef	P	Coef	P	Coef	P
Constant	-1.9277		-1.9335		-1.9217		-1.9187	
vsb	0.00895	0.108	0.00893	0.102	0.00916	0.091	0.00901	0.095
pcp	-0.0517	0.270	-0.0481	0.288	-0.0495	0.265	-0.0484	0.275
sd	0.00181	0.561	0.00067	0.743	0.00071	0.728		
borough	7.6860	0.000	7.6860	0.000	7.6860	0.000	7.6860	0.000
dayType	-0.045	0.959	-0.041	0.953				
month	-0.0202	0.998						
S	0.276012		0.275285		0.274874		0.274739	
R-sq	99.65%		99.65%		99.65%		99.65%	
R-sq(adj)	99.65%		99.65%		99.65%		99.65%	
Mallows' Cp	16.00		6.30		0.63		-1.25	
AICc	256.27		246.24		240.43		238.51	
BIC	337.28		303.55		283.48		276.79	

	------Step 5-----	
	Coef	P
Constant	-1.9216	
vsb	0.00887	0.100
pcp		
sd		
borough	7.6860	0.000
dayType		
month		
S	0.274769	
R-sq	99.65%	
R-sq(adj)	99.65%	
Mallows' Cp	-2.06	
AICc	237.67	
BIC	271.19	

α to remove = 0.1

If a term has more than one coefficient, the largest in magnitude is shown.

Figure A7.4: Step-by-step process for rescreened backwards elimination

Stepwise Selection of Terms

Candidate terms: temp*dewp, spd, vsb, temp, dewp, slp, pcp, sd, borough

	------Step 1------ Coef	P	------Step 2----- Coef	P
Constant	-0.80059		-0.8346	
borough	3.3380	0.000	3.3380	0.000
vsb			0.00385	0.100
S	0.119445		0.119331	
R-sq	99.65%		99.65%	
R-sq(adj)	99.65%		99.65%	
Mallows' Cp	1.30		0.60	
AICc	-1267.89		-1268.59	
BIC	-1239.15		-1235.07	

α to enter = 0.1, α to remove = 0.1
If a term has more than one coefficient, the largest in magnitude is shown.

Figure A7.5: Stepwise selection screening for a new model

A8: Final Model with Unusual Observations Removed

Regression Equation

borough	
Bronx	differences = -1.9125 + 0.00681 vsb
Brooklyn	differences = 5.7971 + 0.00681 vsb
Manhattan	differences = 3.8573 + 0.00681 vsb
Queens	differences = -0.5904 + 0.00681 vsb
Staten Island	differences = -7.4305 + 0.00681 vsb

Figure A8.1: Regression Equation

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
0.235254	99.74%	99.74%	48.2980	99.73%	-40.65	-7.43

Figure A8.2: Model Summary

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	-1.9125	0.0454	(-2.0016, -1.8233)	-42.12	0.000	
vsb	0.00681	0.00472	(-0.00246, 0.01608)	1.44	0.150	1.00
borough						
Brooklyn	7.7096	0.0253	(7.6599, 7.7593)	304.32	0.000	1.58
Manhattan	5.7698	0.0251	(5.7204, 5.8192)	229.42	0.000	1.60
Queens	1.3221	0.0249	(1.2731, 1.3710)	53.01	0.000	1.61
Staten Island	-5.5180	0.0255	(-5.5681, -5.4679)	-216.15	0.000	1.57

Figure A8.3: Anova Output for Final Model

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	5	18130.7	99.74%	18130.7	3626.14	65519.55	0.000
vsb	1	1.0	0.01%	0.1	0.11	2.08	0.150
borough	4	18129.7	99.73%	18129.7	4532.42	81894.99	0.000
Error	860	47.6	0.26%	47.6	0.06		
Lack-of-Fit	500	28.0	0.15%	28.0	0.06	1.03	0.394
Pure Error	360	19.6	0.11%	19.6	0.05		
Total	865	18178.3	100.00%				

Figure A8.4: Analysis of Variance of Final Model

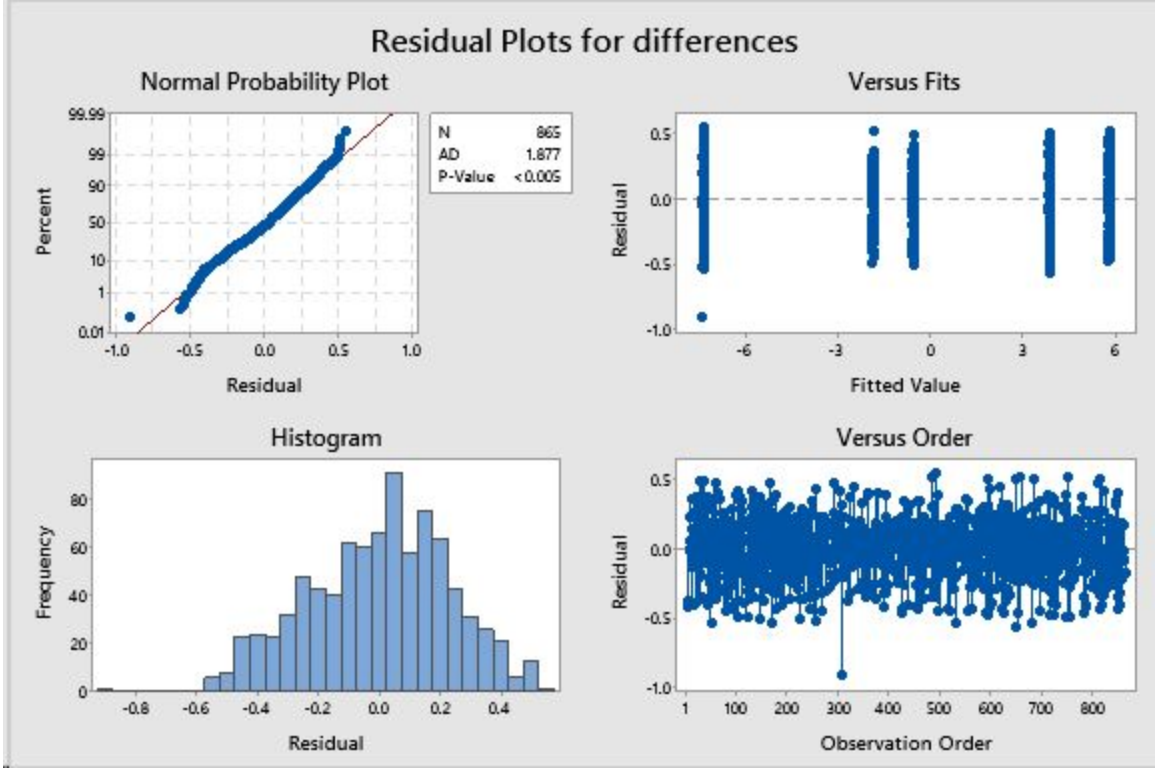

Figure A8.5: Residual Plots for Final Model

B. Data Description

```python
# RegProjPy

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

df = pd.read_csv('uber_nyc_enriched.csv')

# visualizing missing data in the dataset
sns.heatmap(df.isnull(), cbar=False, cmap="YlGnBu_r")
plt.show()

# removing n/a and non borough values
df = df[(df.borough != 'EWR')].dropna(how='any', axis=0) # drop airport and nan values

# compressing all the hourly datapoints into daily datapoints for the final df
edfList = [] # list with every row of data to be in the final df
for day in df.pickup_dt.str.slice(0,10).unique():
    for bur in df.borough.unique():
        # creating a df for each day by borough
        daydf = df[(df.pickup_dt.str.slice(0,10) == day) & (df.borough == bur)]
        # calculating averages and values for daily values
        daylist = [day, bur, np.sum(daydf.pickups), np.mean(daydf.spd),
                   np.mean(daydf.vsb), np.mean(daydf.temp), np.mean(daydf.dewp),
                   np.mean(daydf.slp), list(daydf.pcp24)[-1], np.mean(daydf.sd),
                   list(daydf.hday)[0]]
        edfList.append(daylist)

# creating the final dataframe that we will write to .csv
edf = pd.DataFrame(edfList, columns=df.columns.drop(['pcp06','pcp01'])).rename(columns={'pcp24':'pcp'})

# adding weekends into edf
weekendList = []
for i in range(0,905,5):
    if i == 10 or (i%35) == 10 or i == 15 or (i%35) == 15:
        for i in range(5):
            weekendList.append('Y')
    else:
        for i in range(5):
            weekendList.append('N')
edf['weekend'] = weekendList
```

```python
# creating column of dummy variable for dayType
dayType = []
for i in range(0,edf.shape[0]):
    if edf.hday[i] == 'Y' and edf.weekend[i] == 'Y':
        dayType.append('WendHday')
    elif edf.hday[i] == 'Y' and edf.weekend[i] == 'N':
        dayType.append('Hday')
    elif edf.hday[i] == 'N' and edf.weekend[i] == 'Y':
        dayType.append('Wend')
    elif edf.hday[i] == 'N' and edf.weekend[i] == 'N':
        dayType.append('Wday')
edf['dayType'] = dayType

# dropping interim columns from dayType calculations
edf = edf.drop(['hday','weekend'], axis = 1)

# adding column for qualitative month
edf['month'] = edf.pickup_dt.str.slice(5,7).replace({'01':'Jan','02':'Feb','03':'Mar','04':'Apr','05':'May','06':'Jun'})

# writing the final edf to .csv , purposely wont overwrite existing .csv
try:
    pd.read_csv('uber_nyc_clean.csv')
    print('uber_nyc_clean.csv already exists. Exiting RegProjPy')
except:
    edf.to_csv('uber_nyc_clean.csv')
```

Figure B1.1 Python script for aggregating raw data

	pickup_dt	borough	pickups	spd	vsb	temp	dewp	slp	pcp	sd	dayType	month
0	2015-01-01	Bronx	1075	6.086957	10.0	31.043478	7.869565	1019.795652	0.0	0.0	Hday	Jan
1	2015-01-01	Brooklyn	12528	6.086957	10.0	31.043478	7.869565	1019.795652	0.0	0.0	Hday	Jan
2	2015-01-01	Manhattan	35870	6.086957	10.0	31.043478	7.869565	1019.795652	0.0	0.0	Hday	Jan
3	2015-01-01	Queens	5108	6.086957	10.0	31.043478	7.869565	1019.795652	0.0	0.0	Hday	Jan
4	2015-01-01	Staten Island	53	6.086957	10.0	31.043478	7.869565	1019.795652	0.0	0.0	Hday	Jan

Figure B1.2: Preview of aggregated data set

C. Conclusion and Recommendations

Prediction

Fit	SE Fit	95% CI	95% PI
-0.796019	0.0093169	(-0.814304, -0.777733)	(-1.03093, -0.561106)

Figure C1.1: 95% prediction interval for z_t for Bronx and vsb = 10.00

References

[1] A. Flowers. “Uber TLC FOIL Response.” Internet: https://github.com/fivethirtyeight/uber-tlc-foil-response, Jan. 14, 2016 [Oct. 25, 2019].

[2] Y. Pappas. “NYC Uber Pickups with Weather and Holidays.” Internet: https://www.kaggle.com/yannisp/uber-pickups-enriched, 2016 [Oct. 25, 2019].

[3] “Technical Support by Phone or Online.” *Minitab, Inc.*, support.minitab.com/. https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/regression/supporting-topics/model-assumptions/test-for-autocorrelation-by-using-the-durbin-watson-statistic/

[4] Bowerman, Bruce L., et al. *Forecasting, Time Series, and Regression: an Applied Approach*. Thomson Brooks/Cole, 2005.