Project-1

Section 1. Statistical Test

1.1 Which statistical test did you use to analyse the NYC subway data? Did you use a one-tail or a two-tail P value? What is the null hypothesis? What is your p-critical value?

*For the purpose of studying the data and exploring the data set I have used the **Mann-Whitney U test**, as this statistical test doesn't assume that the data is normalized. Also, as part of this test I used the 2 tailed P value. The null hypothesis is that the subway ridership doesn't wary based on weather conditions i.e. the subway ridership is consistent across rainy and non-rainy days*.

Null & Alternate Hypothesis:

H0 – Data is drawn from a normally distributed population
Ha – Data is not drawn from a normally distributed population

Two Tailed hypothesis test was done for this analysis with p-critical value of 0.025.

1.2 Why is this statistical test applicable to the dataset? In particular, consider the assumptions that the test is making about the distribution of ridership in the two samples.

We can determine the likelihood that our samples are drawn from normally distributed population by following method

1. Shape of the distribution

Below two histograms depict the distribution of hourly entries in our NYC subway data when it's raining and not raining. Shape of the both distributions is positively skewed. Hence we can prove that **data is not normally distributed**.

The below histograms tests prove that the data is drawn from the non-normal distribution. Hence we have to use a non-parametric test to compare our data. One such test is the **Mann-Whitney U test** which was used for the data analysis.

Legends:

Not Raining

Raining

1.3 What results did you get from this statistical test? These should include the following numerical values: p-values, as well as the means for each of the two samples under test.

- *Mean Ridership on rainy days - 1105.4463767458733*
- *Mean Ridership on non-rainy days - 1090.278780151855*
- *One sided p-value - 0.024999912793489721*
- *Two sided p-value – 2*0.024 = 0.048*

1.4 What is the significance and interpretation of these results?

The P value of two sided is less than 0.05 which signifies that the 95% confidence level the null hypothesis that is the ridership is same for both rainy days and non-rainy days can be rejected. Meaning there is a significance difference in ridership between rainy and non-rainy days.

Section 2. Linear Regression

2.1 What approach did you use to compute the coefficients theta and produce prediction for ENTRIESn_hourly in your regression model:

2 Gradient descent (as implemented in exercise 3.5)
3 OLS using Statsmodels
4 Or something different?

OLS using statasmodel produced a higher coefficient value than gradient descent, which implies that the OLS model is more predictable than the gradient descent.

2.2 What features (input variables) did you use in your model? Did you use any dummy variables as part of your features?

Features used in this model are: rain, precipi, hour, fog, meantempi

Dummy variables used in the model: UNIT, day_week, conds, TIMEn

2.3 Why did you select these features in your model? We are looking for specific reasons that lead you to believe that the selected features will contribute to the predictive power of your model. Your reasons might be based on intuition. For example, response for fog might be: "I decided to use fog because I thought that when it is very foggy outside people might decide to use the subway more often." Your reasons might also be based on data exploration and experimentation, for example: "I used feature X because as soon as I included it in my model, it drastically improved my R2 value."

Features:

- **Hour -** By plotting the total number of hourly entries and exits grouped by hour I found that there were spikes during several hours in a day
- **Meantempi –** If the temperature is high I assumed that more people will use subway
- **Rain –** I assumed that when it is raining more people will use subways
- **Fog -** I assumed that when it is foggy more people will use subways
- **Precipi –** With higher precipitation the chances of rain is higher therefore precipication and rain have a direct relationship and therefore I used it as one of my variables

Dummy Variables:

- **UNIT** – Unit is improving the R^2 value drastically. Hence included it in my model.
- **day_week** – Normally traffic is more during weekdays compared to weekends.
- **Conds**: Normally looking at the conditions of the whether people will decide whether to use subways or not.
- **TIMEn**: I used this dummy variable which helps to understand how people will use subway at different times.

2.4 The weights for the different features are as below

Variables	weights				
fog	-580.803	183.541	-3.164	0.001555	**
hour	-106.230	4.534	-23.428	< 2e-16	***
meantempi	-26.187	1.650	-15.874	< 2e-16	***
precipi	-2374.727	755.188	-3.145	0.001665	**
rain	157.310	30.482	5.161	2.47e-07	***
UNITR011	7172.381	215.869	33.226	< 2e-16	***
UNITR012	8523.071	215.321	39.583	< 2e-16	***
UNITR013	2421.512	215.321	11.246	< 2e-16	***
UNITR016	603.647	215.865	2.796	0.005170	**
UNITR017	4036.507	215.321	18.746	< 2e-16	***
UNITR018	7700.213	215.221	35.778	< 2e-16	***
UNITR019	3174.300	215.011	14.763	< 2e-16	***
UNITR020	6212.603	215.321	28.853	< 2e-16	***
UNITR021	4523.510	215.874	20.954	< 2e-16	***
UNITR022	9356.996	215.321	43.456	< 2e-16	***
UNITR023	5992.173	215.321	27.829	< 2e-16	***
UNITR024	3132.685	215.285	14.551	< 2e-16	***
UNITR025	5271.192	215.011	24.516	< 2e-16	***
UNITR027	2806.609	215.321	13.035	< 2e-16	***
UNITR029	7068.582	215.321	32.828	< 2e-16	***
UNITR030	2938.776	215.321	13.648	< 2e-16	***
UNITR031	4190.544	215.321	19.462	< 2e-16	***
UNITR032	4283.138	215.592	19.867	< 2e-16	***
UNITR033	8073.555	215.321	37.496	< 2e-16	***
UNITR034	958.515	220.060	4.356	1.33e-05	***
UNITR035	2639.361	215.865	12.227	< 2e-16	***
UNITR036	642.086	217.514	2.952	0.003160	**
UNITR037	761.395	215.779	3.529	0.000418	***
UNITR039	649.021	220.996	2.937	0.003318	**
UNITR040	1169.242	215.498	5.426	5.80e-08	***
UNITR041	2938.587	215.321	13.647	< 2e-16	***
UNITR043	2725.722	215.321	12.659	< 2e-16	***
UNITR044	4517.130	215.321	20.979	< 2e-16	***
UNITR046	8183.582	215.321	38.007	< 2e-16	***
UNITR049	2611.394	215.321	12.128	< 2e-16	***
UNITR050	3867.211	215.875	17.914	< 2e-16	***
UNITR051	4973.136	215.321	23.096	< 2e-16	***

UNITR052	1129.751	219.051	5.157	2.51e-07	***
UNITR053	3096.593	215.498	14.370	< 2e-16	***
UNITR054	1297.148	215.873	6.009	1.88e-09	***
UNITR055	8258.184	214.947	38.420	< 2e-16	***
UNITR056	1285.446	215.869	5.955	2.63e-09	***
UNITR057	4721.523	215.321	21.928	< 2e-16	***
UNITR059	1014.803	217.895	4.657	3.21e-06	***
UNITR060	630.648	217.004	2.906	0.003661	**
UNITR062	2575.383	215.321	11.961	< 2e-16	***
UNITR063	957.164	220.704	4.337	1.45e-05	***
UNITR064	670.714	218.177	3.074	0.002112	**
UNITR065	661.421	219.411	3.015	0.002575	**
UNITR067	767.162	220.662	3.477	0.000508	***
UNITR069	846.199	218.736	3.869	0.000110	***
UNITR070	1626.362	215.321	7.553	4.33e-14	***
UNITR080	3450.690	215.321	16.026	< 2e-16	***
UNITR081	3399.790	215.869	15.749	< 2e-16	***
UNITR082	1342.025	215.874	6.217	5.12e-10	***
UNITR083	2965.098	215.321	13.771	< 2e-16	***
UNITR084	9869.372	215.321	45.836	< 2e-16	***
UNITR085	2442.961	215.876	11.317	< 2e-16	***
UNITR086	2438.690	215.321	11.326	< 2e-16	***
UNITR087	1067.223	216.434	4.931	8.21e-07	***
UNITR091	1037.939	219.685	4.725	2.31e-06	***
UNITR092	1919.428	217.822	8.812	< 2e-16	***
UNITR093	1962.588	218.429	8.985	< 2e-16	***
UNITR094	1693.684	215.498	7.859	3.95e-15	***
UNITR095	2110.398	216.347	9.755	< 2e-16	***
UNITR096	2283.782	215.221	10.611	< 2e-16	***
UNITR097	2909.251	215.221	13.518	< 2e-16	***
UNITR098	1673.249	215.321	7.771	7.96e-15	***
UNITR099	2230.276	215.321	10.358	< 2e-16	***
UNITR101	2664.668	215.321	12.375	< 2e-16	***
UNITR102	3553.507	215.321	16.503	< 2e-16	***
UNITR103	1300.746	219.048	5.938	2.90e-09	***
UNITR104	1281.305	215.498	5.946	2.77e-09	***
UNITR105	3203.206	215.321	14.876	< 2e-16	***
UNITR106	993.536	220.661	4.503	6.73e-06	***
UNITR108	5092.372	215.321	23.650	< 2e-16	***
UNITR111	3089.496	215.321	14.348	< 2e-16	***
UNITR112	1625.469	215.221	7.553	4.35e-14	***
UNITR114	848.600	215.287	3.942	8.10e-05	***
UNITR115	1221.789	215.011	5.682	1.34e-08	***

UNITR116	3068.641	215.321	14.251	< 2e-16	***
UNITR117	818.820	220.333	3.716	0.000202	***
UNITR119	1796.855	216.923	8.283	< 2e-16	***
UNITR120	1439.701	218.427	6.591	4.41e-11	***
UNITR121	1317.235	218.183	6.037	1.58e-09	***
UNITR122	2517.596	216.352	11.637	< 2e-16	***
UNITR123	1494.923	217.001	6.889	5.70e-12	***
UNITR126	1730.550	215.321	8.037	9.44e-16	***
UNITR127	4671.259	215.321	21.694	< 2e-16	***
UNITR137	2404.092	214.947	11.185	< 2e-16	***
UNITR139	2408.068	215.591	11.170	< 2e-16	***
UNITR163	3206.996	215.321	14.894	< 2e-16	***
UNITR172	1773.034	215.321	8.234	< 2e-16	***
UNITR179	6653.808	215.321	30.902	< 2e-16	***
UNITR181	1630.223	217.302	7.502	6.40e-14	***
UNITR183	700.217	220.997	3.168	0.001534	**
UNITR184	854.346	220.051	3.882	0.000104	***
UNITR186	947.877	216.724	4.374	1.22e-05	***
UNITR188	2204.841	215.592	10.227	< 2e-16	***
UNITR189	1267.739	217.892	5.818	5.99e-09	***
UNITR194	1872.066	217.290	8.615	< 2e-16	***
UNITR196	1203.467	216.154	5.568	2.60e-08	***
UNITR198	1988.914	215.874	9.213	< 2e-16	***
UNITR199	593.831	217.291	2.733	0.006281	**
UNITR200	1025.588	216.127	4.745	2.09e-06	***
UNITR202	2188.839	215.500	10.157	< 2e-16	***
UNITR203	1633.480	218.489	7.476	7.79e-14	***
UNITR204	1334.534	215.321	6.198	5.78e-10	***
UNITR205	1466.623	216.061	6.788	1.15e-11	***
UNITR207	1881.531	215.601	8.727	< 2e-16	***
UNITR208	2469.219	216.061	11.428	< 2e-16	***
UNITR209	647.809	220.700	2.935	0.003335	**
UNITR211	2291.964	215.321	10.644	< 2e-16	***
UNITR212	1574.869	215.591	7.305	2.82e-13	***
UNITR213	1023.162	217.595	4.702	2.58e-06	***
UNITR215	1477.824	215.876	6.846	7.71e-12	***
UNITR216	647.888	215.868	3.001	0.002690	**
UNITR217	854.324	220.682	3.871	0.000108	***
UNITR218	1947.598	215.221	9.049	< 2e-16	***
UNITR219	1231.418	215.500	5.714	1.11e-08	***
UNITR220	1366.690	215.321	6.347	2.21e-10	***
UNITR221	1277.789	221.032	5.781	7.48e-09	***
UNITR223	2087.728	215.221	9.700	< 2e-16	***

UNITR224	603.650	218.183	2.767	0.005665	**
UNITR226	668.491	219.684	3.043	0.002344	**
UNITR227	985.953	215.321	4.579	4.69e-06	***
UNITR228	984.536	220.374	4.468	7.93e-06	***
UNITR231	844.461	216.719	3.897	9.77e-05	***
UNITR232	890.810	219.417	4.060	4.92e-05	***
UNITR233	956.544	221.698	4.315	1.60e-05	***
UNITR235	2457.326	215.591	11.398	< 2e-16	***
UNITR236	1490.038	216.059	6.896	5.41e-12	***
UNITR237	598.540	220.007	2.721	0.006520	**
UNITR238	2071.117	215.222	9.623	< 2e-16	***
UNITR239	783.114	215.321	3.637	0.000276	***
UNITR240	2541.759	216.153	11.759	< 2e-16	***
UNITR243	1309.191	218.735	5.985	2.18e-09	***
UNITR244	1575.551	218.745	7.203	6.00e-13	***
UNITR248	2990.398	215.591	13.871	< 2e-16	***
UNITR249	1239.894	217.000	5.714	1.11e-08	***
UNITR250	850.258	217.593	3.908	9.34e-05	***
UNITR251	1163.493	216.144	5.383	7.37e-08	***
UNITR252	842.055	215.874	3.901	9.61e-05	***
UNITR253	665.014	219.059	3.036	0.002401	**
UNITR254	2558.450	215.223	11.887	< 2e-16	***
UNITR255	730.605	216.124	3.380	0.000724	***
UNITR256	942.742	215.879	4.367	1.26e-05	***
UNITR257	1770.286	215.321	8.222	< 2e-16	***
UNITR258	1411.256	216.153	6.529	6.70e-11	***
UNITR259	802.513	217.013	3.698	0.000218	***
UNITR260	932.207	224.866	4.146	3.40e-05	***
UNITR261	1526.536	217.516	7.018	2.28e-12	***
UNITR265	741.628	219.727	3.375	0.000738	***
UNITR266	870.859	215.221	4.046	5.21e-05	***
UNITR269	766.902	215.871	3.553	0.000382	***
UNITR273	1214.726	220.728	5.503	3.75e-08	***
UNITR274	812.180	219.417	3.702	0.000215	***
UNITR275	851.609	217.517	3.915	9.05e-05	***
UNITR276	1292.824	215.321	6.004	1.94e-09	***
UNITR279	716.210	216.923	3.302	0.000962	***
UNITR281	1161.316	217.006	5.352	8.77e-08	***
UNITR282	1462.169	215.874	6.773	1.28e-11	***
UNITR284	650.452	215.874	3.013	0.002587	**
UNITR291	1686.055	215.321	7.830	4.97e-15	***
UNITR294	801.592	217.892	3.679	0.000235	***
UNITR300	2095.120	215.321	9.730	< 2e-16	***

UNITR303	1098.094	217.001	5.060	4.20e-07	***
UNITR304	958.113	215.874	4.438	9.09e-06	***
UNITR308	721.778	218.736	3.300	0.000968	***
UNITR309	746.908	218.428	3.419	0.000628	***
UNITR310	1219.627	219.684	5.552	2.85e-08	***
UNITR319	1207.790	217.290	5.558	2.74e-08	***
UNITR321	957.716	215.321	4.448	8.69e-06	***
UNITR322	1641.483	217.296	7.554	4.30e-14	***
UNITR323	1124.163	219.405	5.124	3.01e-07	***
UNITR330	842.913	217.886	3.869	0.000110	***
UNITR346	1099.316	218.804	5.024	5.08e-07	***
UNITR356	1020.210	217.892	4.682	2.85e-06	***
UNITR382	823.185	217.582	3.783	0.000155	***
UNITR429	937.205	215.778	4.343	1.41e-05	***
UNITR453	1593.361	222.731	7.154	8.58e-13	***
day_week	-159.005	4.861	-32.714	< 2e-16	***
condsFog	1588.316	349.999	4.538	5.69e-06	***
condsHaze	447.687	59.555	7.517	5.71e-14	***
condsLight Drizzle	-1072.976	116.490	-9.211	< 2e-16	***
condsMist	1172.223	414.965	2.825	0.004732	**
condsMostly Cloudy	-433.327	34.703	-12.487	< 2e-16	***
condsOvercast	-396.258	30.320	-13.069	< 2e-16	***
TIMEn12:00:00	2886.508	48.602	59.391	< 2e-16	***
TIMEn16:00:00	2602.622	64.016	40.656	< 2e-16	***
TIMEn20:00:00	3955.908	80.787	48.967	< 2e-16	***
TIMEn4:00:00	-646.710	29.852	-21.664	< 2e-16	***

2.5 What is your model's R2 (coefficients of determination) value?

R-squared : 0.535
Adjusted R-squared : 0.5322

2.6 What does this R2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R2 value?

R-squared explains 53.5% of variability of the dependent variable. As R-squared value is low this is not the perfect model for this dataset, we need to explore other methods or algorithms to fit this data. Other option would be transform the variables in our data set to improve R-squared.

Section 3. Visualization

Please include two visualizations that show the relationships between two or more variables in the NYC subway data. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots, or histograms) or attempt to implement something more advanced if you'd like.

Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

3.1. Ridership by time-of-day or day-of-week

Legends:

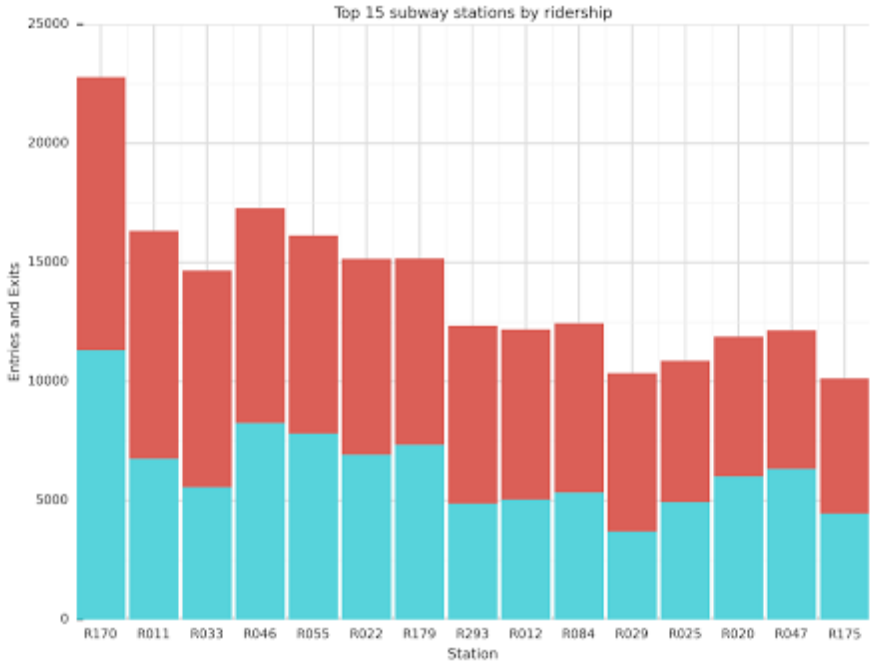
Raining

Not Raining

The above graph shows different ranges in the x-axis which constitute a range to represent volume of ridership and y-axis represents the frequency of people using the subway for these given ranges. The blue colored histograms are the frequency of ridership when it is raining and the orange are the ones when it is not raining.

3.2. One visualization can be more freeform, some suggestions are:

2. Which stations have more exits or entries at different times of day



Legends:

Mean Entries

Mean Exits

The above graph shows the top 15 subway stations based on the hourly entries and exits from each of the subways stations. The above graph represents the mean value of ridership per subway station and not the actual count of entries and exits. The red coloured bars represent mean exits per subway station and the blue coloured bars represent mean Entries per subway stations.

Section 4. Conclusion

Please address the following questions in detail. Your answers should be 1-2 paragraphs long.

4.1 From your analysis and interpretation of the data, do more people ride the NYC subway when it is raining or when it is not raining?

By analyzing the data visually and by performing statistical tests it was observed that more people used the subway when it was raining compared to when it was not raining.

4.2 What analyses lead you to this conclusion? You should use results from both your statistical tests and your linear regression to support your analysis.

Looking at the below table by doing some simple data exploration it can be observed that the median and mean number of entries and exits per hour are higher when it is raining than when it is not raining.

			ENTRIESn_hourly	EXITSn_hourly
not raining	count	87847	87847	87847
	mean	0	1090.27878	883.25961
	std	0	2320.004938	1998.516762
	min	0	0	0
	25%	0	38	31
	50%	0	278	231
	75%	0	1111	846
	max	0	43199	45249
raining	count	44104	44104	44104
	mean	1	1105.446377	894.123572
	std	0	2370.527674	2028.552487
	min	1	0	0
	25%	1	41	33
	50%	1	282	235
	75%	1	1103.25	849
	max	1	51839	41503

Also, on performing the Mann Whitney U-Test on the sample dataset it was observed that the null hypothesis i.e. the ridership remains the same for both rainy and non-rainy days can be rejected as the p-value was less than 0.05. The OLS model that was built predicted the same outcome with a 46.11% stability that more people use the subway when it raining compared to when it is not raining.

Section 5. Reflection

Please address the following questions in detail. Your answers should be 1-2 paragraphs long.

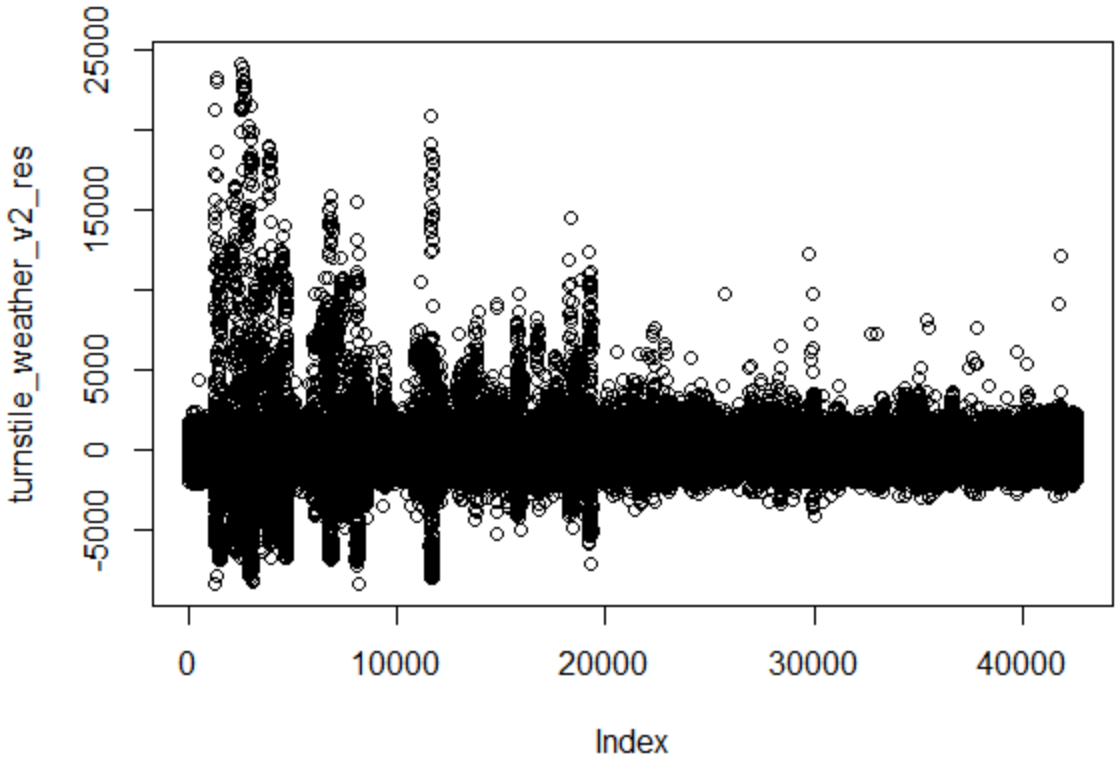
5.1 Please discuss potential shortcomings of the methods of your analysis, including:

Some of the potential shortcomings of the dataset are:-

- There should be data present per subway station for the same time but collected for both rainy and non-rainy days which are not present in the current data set. This is with the assumption that the same set of people travel on the same routes.
- The subways stations can have an additional geographic dimension to show whether the subway station is near a residential neighborhood or near to workplace
- There might be omitted variables like public holidays, festivity or event dates, closed dates for maintenance, road works, traffic signal problems, any accidents that might have happened etc.

Some of the potential shortcomings of the analysis method are:-

- My analysis is purely based on a liner model. I think applying a nonlinear model with composite features would lead to a better fit over the sample data.
- When I plotted the residual plot the data points what I observed that the data was distributed around the zero that shows the model is fit but our R-squared is low hence we need to explore other models to improve the R-Squared value.



5.2 (Optional) Do you have any other insight about the dataset that you would like to share with us?

References:

Udacity website: https://www.udacity.com/nanodegrees

Blogs: https://piazza.com/class

Nanodegree Phython course : https://www.udacity.com/course/ud036

Sample Project:
http://nbviewer.ipython.org/url/www.asimihsan.com/articles/Intro%20to%20Data%20Science%20-%20Final%20Project.ipynb

Python Course: https://developers.google.com/edu/python/introduction

Pandas: http://www.gregreda.com/2013/10/26/intro-to-pandas-data-structures/

ggplot : https://pypi.python.org/pypi/ggplot/

Statistical Learning http://www-bcf.usc.edu/~gareth/ISL/book.html

Pandas tutorial : http://pandas.pydata.org/pandas-docs/stable/tutorials.html

machine learning:
http://openclassroom.stanford.edu/MainFolder/CoursePage.php?course=MachineLearning

(http://openclassroom.stanford.edu/MainFolder/CoursePage.php?course=MachineLearning)

Statistics
https://www.youtube.com/watch?v=JvS2triCgOY&src_vid=zPG4NjIkCjc&feature=iv&annotation_id=annotation_742878

(https://www.youtube.com/watch?v=JvS2triCgOY&src_vid=zPG4NjIkCjc&feature=iv&annotation_id=annotation_742878)

Learn Python hard way: http://learnpythonthehardway.org/book/

Internet searching through google.

Data Analysis in google.

topic: http://www.statsoft.com/Textbook/Multiple-Regression#cresidual

Here is another helpful blog post on R squared

values: http://blog.minitab.com/blog/adventures-in-statistics/how-high-should-r-squared-be-in-regression-analysis