# Improve Subway Frequency by Understanding Weather and Travel Volume

*Ben Steers, Chun-Chieh Tsai, Isha Chaturvedi, Rachel Lim, Yi Xuan Tang*

*CUSP New York University*

## 1. ABSTRACT

**MTA subway commuter volume is independent of weather.**

Travel behavior are affected my multiple factors; weather is commonly identified as a key factor influencing the way people commute in the city. While existing studies have focused on the seasonality of weather on travel behaviour, our study aims to fill the gap in research by investigating the effect of variations in weather conditions on a more granular temporal scale by comparing intra-day weather conditions with subway commuter volumes across all subways stations in New York City. Using entry counts from the MTA subway turnstile dataset and weather data, regression models are applied to analyze the effect of weather on commuter volumes. Results from time series analysis reveal that there is a clear periodicity in commuter volumes with distinct clustering of entries. Results from modelling reveal that while all features are important, the predictive power of commuter volume solely on weather conditions is insufficient, as travel behaviour is influenced by a complex array of factors. Nevertheless, this study contributes to existing understanding of travel behaviour on public transit through a more intricate analysis of ridership volume at a finer temporal and spatial scale.

## 2. INTRODUCTION

The Metropolitan Transportation Authority (MTA) is the largest transportation network in North America. Its expansive subway network provides service across 4 boroughs, providing essential connectivity. With an annual ridership of 1.7 billion, it is crucial to understand the travel behavior and ridership patterns of subway commuters. Weather has been widely identified as one of the key factors influencing travel behavior on public transit. Unlike private modes of transport, commuters using public transport often have to deal with issues concerning first mile and last mile connectivity. The journey of the first and last mile exposes commuters to the physical environment where weather is a key component. As a result, modal choice and travel times may be affected by weather conditions. While a plethora of studies has been conducted on the impact of weather on travel behavior, majority of these studies are focused on evaluating the seasonality of weather in relation to private modes of transport, investigating the volumes of motorists and cyclists. This study aims to fill the gap by investigating the effects of intra-day weather conditions on New York City subway ridership. Using ridership counts from the MTA turnstile data and daily weather data, time series analysis is applied to understand variations in ridership across stations and over time. A series of regression models is then applied to investigate the relationship between weather and commuter volumes. The spatial and temporal scale of our study sets us apart from

existing research. Our findings will be useful in providing evidence to inform planning policy and station level operation and deployment strategies.

## 3. LITERATURE REVIEW

Existing literature on the effects of weather on transportation can be divided into two key themes: the effect of weather on the physical performance of the transport network, and that on commuter's travel behavior. An extensive amount of research has been conducted on the influence of weather on travel pattern in general. In his study on the effect of snow on traffic counts in western New York State, Call (2011) revealed that snow reduces overall traffic volume on urban highways and intra-urban roads. Tsapakis et al (2013) studied the impact of different weather conditions and varying intensities of weather on macroscopic urban travel times in the Greater London area. Their study found that the intensity of rain and snow has a positive relationship with travel time, while temperature has negligible effect on travel time. This finding is supported by similar studies conducted in Washington DC (Stern et al, 2003) and Nagoya City (Wang et al, 2006). Muller et al (2008) focused on assessing the impact of seasonal changes in temperature on travel pattern. Their research found that those choosing to cycle as the main mode of transport is three times higher in summer compared to winter. While extensive research has been conducted on assessing the effect of weather conditions on travel behavior, majority of the research is focused on evaluating the impact on private modes of transport, studying the impact on the motorists, and cyclist.

Compared to the literature assessing the effect of weather on private transport travel behavior, those studying the effect on public transport ridership is significantly less. Tao et al (2018) studied the effect of local weather conditions on bus ridership in Brisbane, Australia using data from transit smart card and weather measurements. Using time-series regression models, they that certain temperature and rainfall intensities have a significant effect on bus ridership. A similar study was conducted in Pierce County, Washington. Stover et al (2012) applied Ordinary Least Squares (OLS) regression models to assess the effect of weather on bus ridership. They considered four weather variables in their models: wind, temperature, rain and snow, using these measures to estimate ridership in different seasons. The OLS regression was also used to explore the relationship between weather and bus and rail ridership in a study in Chicago (Guo et al, 2007). Guo et al (2007) found that all variables have significant impacts on ridership, but have different effects on bus and rail. These studies have identified that weather has a clear impact on travel patterns on public transit. However, majority of these studies have focused on the seasonality of weather condition, evaluating travel pattern on a larger time scale. Our project aims to fill

the gap in existing research by evaluating the effect of weather on subway ridership on a more granular temporal scale, assessing intra-day variations in weather in New York City, in relation to subway commuter volumes in the corresponding time window. This is useful in informing wider planning policy and localized staff deployment and train frequency planning. Our study on local weather condition will also provide additional insight on travel mode choice and travel departure times.

## 4. DATA

Two datasets were used in this project which were merged into 1 dataset for analysis. Subway turnstile data was downloaded from the Metropolitan Transportation Authority (MTA) website and is sampled every 4 hours. The weather data was downloaded from an open source site, Weather Underground, and is given in mostly regular intervals, sampled every 1 hour with exceptions for times of heavy snow. In order to merge the datasets, the weather data was resampled to 4 hour intervals and the two datasets were merged along timestamps.

The raw data contains a column for 'condition' and columns with readings on temperature, precipitation, wind speed and direction, amounts of precipitation, air pressure, and other items. For the purposes of this research, only weather condition, temperature and humidity were considered. The weather condition is given as a categorical variable describing the weather condition in colloquial terms. To convert the column into a format that can be used in computational models, the 15 weather conditions were converted into dummy variables (listed in Table 1). Because of the redundancy in the weather conditions, for example, Heavy Rain, Light Rain, and Light Freezing Rain, the conditions were aggregated into 4 more general classes: clear, rain, snow, cloudy (Table 1). The condition dummy variables are a weighted average by duration of the conditions' one-hot vector encodings in each 4 hour time interval. The values represent the fraction of time that condition was in effect for in the 4 hour period and are normalized to sum to 1.

**Table 1.** *Classification of dummy variables*

| **Variable Name** | **Weather Conditions Covered** |
|---|---|
| **Rain** | Rain, Heavy Rain, Light Rain, Light Freezing Rain |
| **Snow** | Snow, Heavy Snow, Light Snow |
| **Cloudy** | Cloudy, Partly Cloudy, Mostly Cloudy, Overcast, Fog, Haze, Light Freezing Fog |
| **Clear** | Clear, Scattered Clouds |

This analysis requires that the MTA entry data be in counts of the number of entries in that given 4 hour window. The original dataset does not contain this information and instead contains a "register value" that acts as a cumulative counter for each individual turnstile. To get the number of entries for that hour, the previous register value was subtracted to give the difference. Occasionally, the turnstile register would reset and the number of entries at that time interval would register as negative. These missing values were interpolated linearly between the previous and subsequent points. The number of missing points is relatively small compared to the total sample count, only constituting approximately 0.1% of the total samples.

## 5. METHODOLOGY

### 5.1. Spectral Analysis

To better understand the subway commuter volume patterns, Fourier transformation were used to detect periodicity in the signal for the year, 2016. First, a hypothesis was raised that subway commuter volume is related to weather and has seasonal fluctuations. Under this hypothesis, 3 methods were designed to see the pattern of commuter volume: run the Fourier transformation station by station, run the Fourier transformation and average 367 stations' power, and select the 22 most prominent stations and plot by commuter volume.

Additionally, the results of the frequency spectrum analysis was used to identify and attenuate human behavioral patterns that aren't related to weather patterns, such as daily business traffic patterns and week/weekend traffic differences. This was performed by taking the average number of entries for each station, for each weekday and hour, and subtracting the weekday profile from every row in the dataset.

### 5.2. Modelling

Regression modelling was performed on the dataset to measure the predictive power of weather parameters on MTA commuter volume flows and to determine the effect of individual weather conditions. Due to its transparency, linear regression was used in order to see the individual model coefficients and their significance levels. Other models including decision trees regression, linear support vector regression, ridge regression, lasso, random forest regression and multilayer perceptron regression were run to evaluate the performance and accuracy of different models to find a suitable model. Decision tree

and random forest regression were particularly run to find important features (weather parameters) for our prediction model.

## 6. RESULTS

### 6.1. Spectral Analysis

Figure 1 shows the absolute value of the fourier transform, with several peaks that can be read from it.

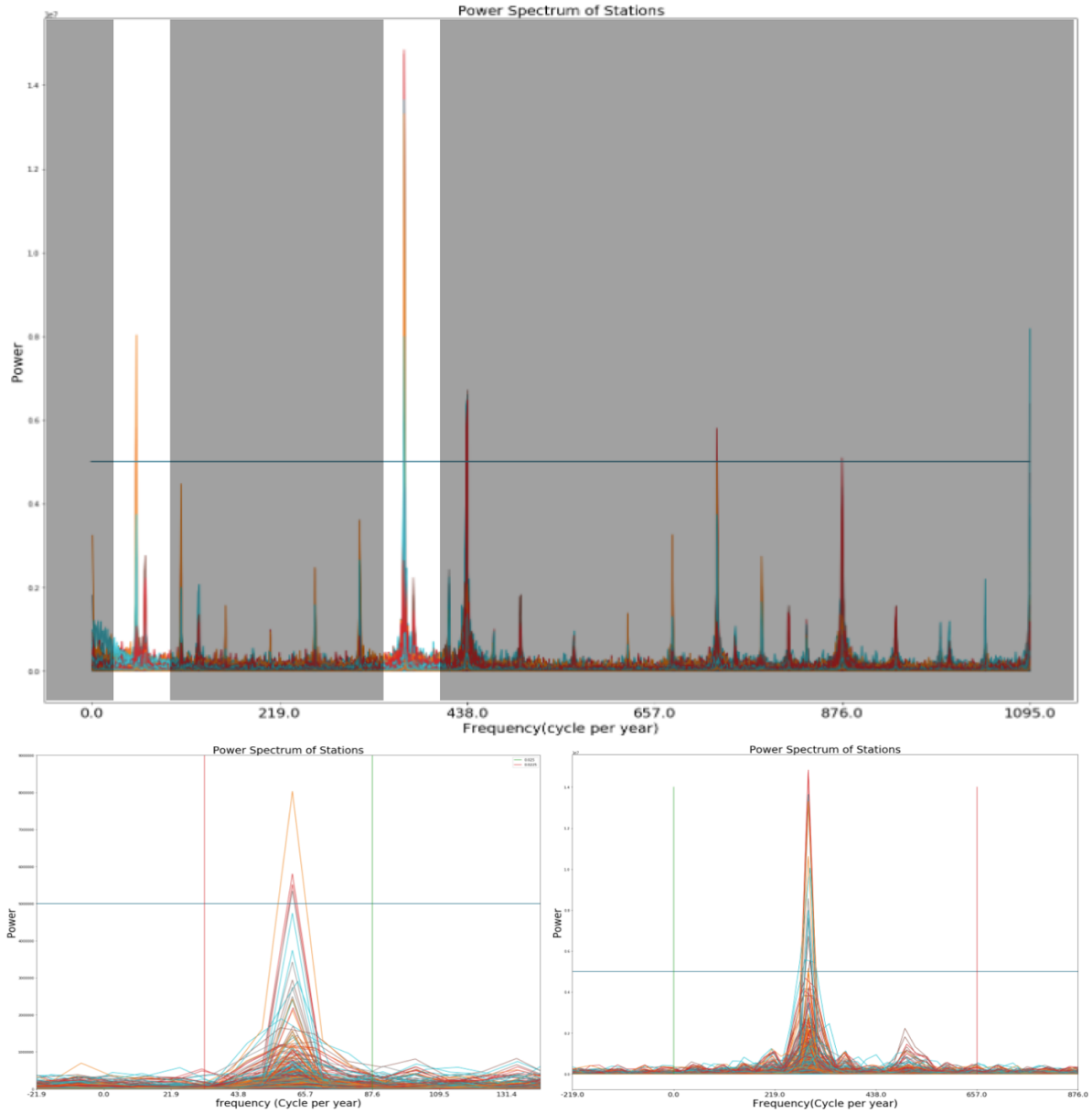


**Fig 1.** *a) Power spectrum of commuter volume of 367 stations, b) Highlight of the frequency between 43.8 and 60 cycles per year, c) Highlight of the frequency between 250.4 and 383.25 cycles per year*

In Figure 1b, the highest power shows between the frequency of 250.4 and 383.25 cycles per year, and it means most of the stations have periodicity between this 2 frequency. The time period falls between 22.9 hours and 25 hours, meaning that most of the stations follow a pattern by approximately 24 hours. In Figure 1c, there is a peak between the frequency of 43.8 and 60 cycles per year or a period between 160 hours and 177 hours, which is closed to 1 week. This can tell us that most of the stations also follow a pattern by week.

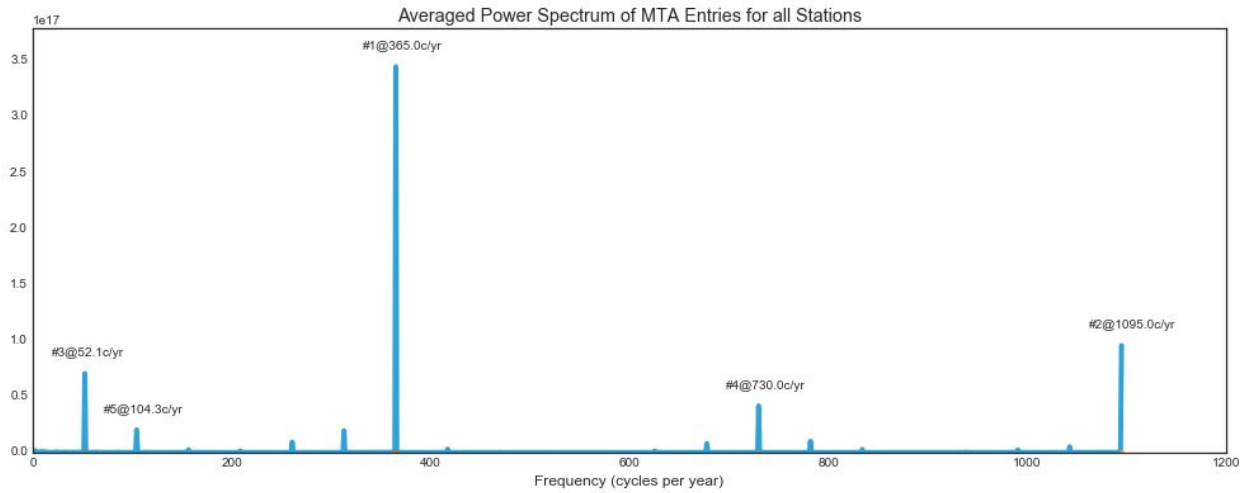


**Fig 4.** *Average Power Spectrum of Entries for all stations. The top 5 frequencies are labeled with their rank followed by their respective frequencies.*

The power spectrum averaged over all stations shows a slightly less ambiguous picture, showing only the peaks that are strong in a majority of the stations. The daily periodicity is by far the strongest oscillation. The second strongest oscillation is at 1095 cycles per year. This maps to a period of 8 hours, which is the approximate length of the work day and represents the peaks found at rush hour in the morning and evening. The third strongest peak is the weekly periodicity, at a frequency of 52.1 cycles per year.

Because the modelling method has no time dependence parameters, these periodicities that are caused by human sleep patterns only serve to add noise to the model and make prediction more difficult. The entry rate at midnight on Tuesday is going to be different from rush hour on Thursday for the same weather conditions. Therefore, it is important to attempt to remove some of the effect that behavior unrelated to weather has on our signal. The average daily profile for MTA commuter entries for all stations is shown for each weekday in Figure 5.

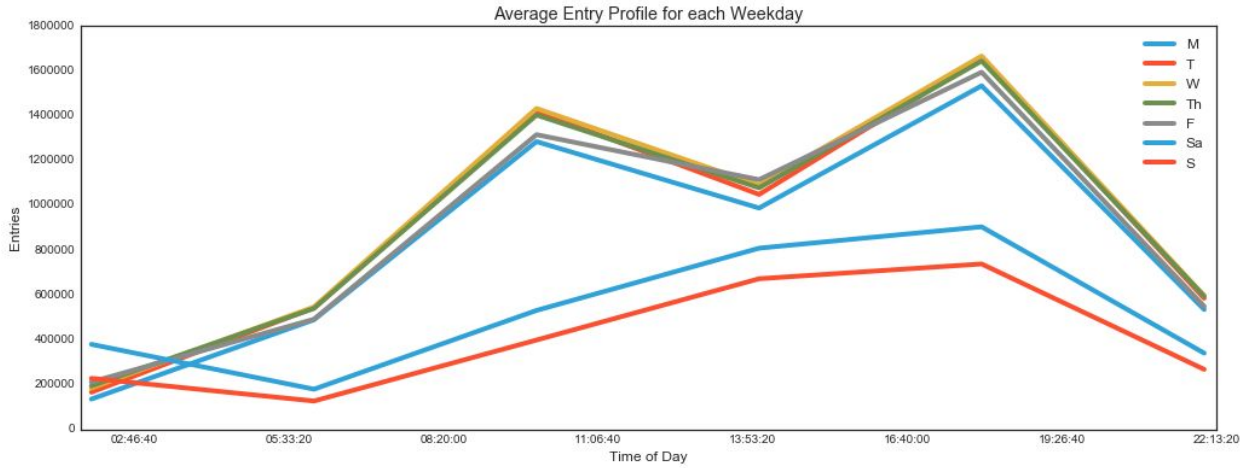


**Fig 5.** *Weekday profile averaged over all weeks and stations.*

The weekday profiles show distinct clustering, where weekday entries show peaks at 8-10am and 4-6pm, and weekends show a much more gradual increase over the course of the afternoon. This is a clear illustration of the business week. To reduce this effect, the average entry profile (like in Figure 5) are taken for the entire year for each station individually. The average profile is then subtracted from the number of entries to get the difference between the number of entries that day and the number of entries that is typical for that station, on that weekday, at that hour. The frequency spectrum of the entries after removing the weekday profiles can be seen in Figure 6.

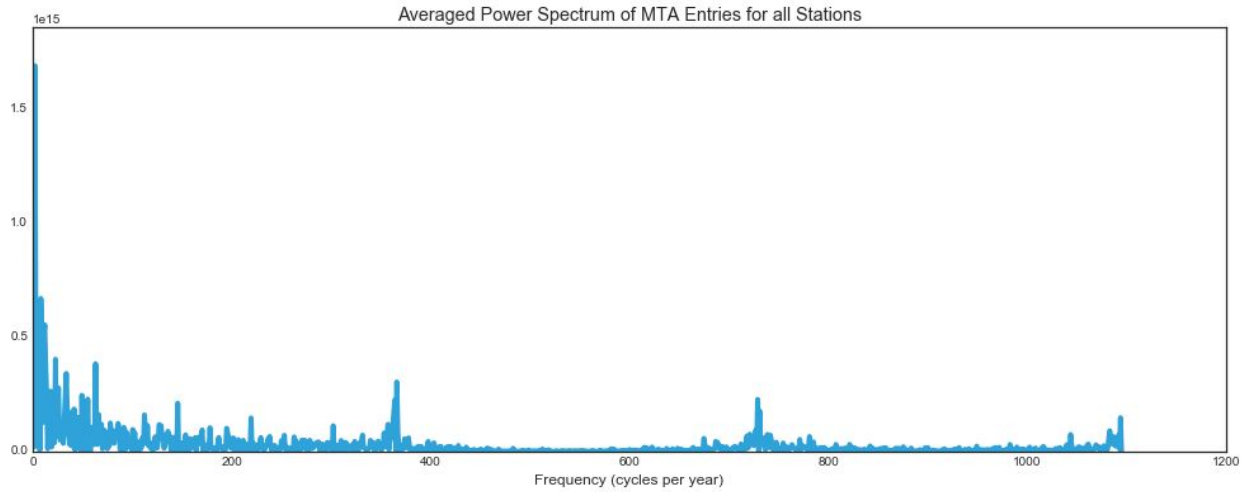


**Fig 6.** *Average power spectrum for all stations after removing the average weekday profile.*

After subtracting out the average weekday profiles, the daily, weekly, and rush hour periodicity all dropped in approximately 3 orders of magnitude, meaning that a large portion of the societal-driven

periodicity was attenuated. The peak frequency, the largest peak by a factor of 2, is located around 1.96 cycles per year. This is most likely a seasonal periodicity and should not be attenuated.

### 6.2. Modelling

The summary result of linear regression is given in Fig. 11 of Appendix. Although the coefficient of determination ($R^2$) is low (0.234), the estimate coefficients of the features have high significance values ($p<0.05$), showing that all the features are important for the model. Table 2 shows individual results of different models and the $R^2$ of the prediction. Here the $R^2$ is low for all the models with linear regression having the highest $R^2$ value. One of the reasons for low $R^2$ could be that the given features don't account for the complete picture for predicting commuter volume, and thus more factors should be taken in account. For linear regression, all the features except humidity have a positive impact on the commuter volume (Table 2). Random forest regression (next highest $R^2$) gives the important features for the model. As one can see (Table 2 - Random Forest Regression), snow has the least feature importance whereas humidity has the most feature importance.

The $R^2$ gives the score to judge the quality of the fit (prediction) on new data. In order to get a better measure of prediction accuracy (which could be use as a proxy for goodness of fit of the model), K-Fold cross validation was performed. Table 3 gives the K-Fold Cross Validation Results of different Models for number of splits = 10, and their running time performance. The results show that the scores were all low, with multilayer perceptron being the most expensive one in the terms of time performance. This is because it takes a lot of time to converge towards a minimum error and local minima stagnation state before finishing the learning of all training set samples (Gupta et al., 2016). Linear SVR also takes a lot of time because SVR by default has a huge computational cost for a large data set and thus need additional methods to ensemble it with for improving the time performance. Linear SVR still performs better than standard SVR in general for large datasets as the kernel is linear and implementation is in terms of liblinear rather than libsvm (scikit-learn modules), and thus have more flexibility in choosing loss function and penalties. Fig. 7 shows that random forest have the highest accuracy among all the other algorithms. This is because random forest works on an ensemble method wherein it takes different subset of observations and variables to construct multiple decision trees and merge them together to give more accurate and prediction results. Fig. 9 shows the decision tree constructed after running the decision tree regression. The figure shows that in general the ideal temperature for commuting is between 41 °F and 76°F approximately, with low humidity and less cloudy weather.

**Table 2.** *Individual Results of different Models and their coefficient of determination ($R^2$) of the prediction*

| Model | Coefficient of determination ($R^2$) | Other Information |
|---|---|---|
| Decision Tree Regression | 0.01642 | Feature Importance -<br>1. Humidity: 0.5874<br>2. Temperature: 0.2951<br>3. Cloudy: 0.0757<br>4. Rain: 0.0413<br>5. Clear: 0.0006<br>6. Snow: 0.0 |
| Linear Support Vector Regression (SVR) | -0.0868 | Estimated Coefficients -<br>1. Clear: -63.3848<br>2. Humidity: -17.4185<br>3. Rain: 719.8795<br>4. Snow: 317.1971<br>5. Temperature: 9.4659<br>6. Cloudy: 291.7373 |
| Linear Regression | 0.234 | Estimated Coefficients ($p<0.005$)<br>1. Clear: 1732.7414<br>2. Humidity: -17.5881<br>3. Rain: 2485.7491<br>4. Snow: 2129.0014<br>5. Temperature: 24.0281<br>6. Cloudy: 2062.8110 |
| Ridge Regression | 0.0132 | Estimated Coefficients -<br>1. Clear: -8.1717<br>2. Humidity: -28.8432<br>3. Rain: 1251.5166<br>4. Snow: 407.5046<br>5. Temperature: 13.8724<br>6. Cloudy: 505.6538 |
| Lasso | 0.0132 | Estimated Coefficients -<br>1. Clear: -13.5919<br>2. Humidity: -28.8013<br>3. Rain: 1240.6814<br>4. Snow: 384.9167<br>5. Temperature: 13.8481<br>6. Cloudy: 499.2250 |
| Random Forest Regression | 0.0178 | Feature Importance - |

| | | |
|---|---|---|
| | | 1. Humidity: 0.5695<br>2. Temperature: 0.3041<br>3. Cloudy: 0.0736<br>4. Rain: 0.0456<br>5. Clear: 0.0073<br>6. Snow: 0.0 |
| Multilayer Perceptron Regression | 0.0136 | - |

**Table 3.** *K-Fold Cross Validation Results of different Models and their performance*

| Model | K-Fold Cross Validation Results for no. of splits = 10 | | Performance (seconds) |
|---|---|---|---|
| | Mean | Standard Deviation | |
| Decision Tree Regression | -0.0910 | 0.1640 | 14.43 |
| Linear Support Vector Regression (SVR) | -0.0935 | 0.0682 | 274.45 |
| Linear Regression | -0.0928 | 0.1612 | 10.89 |
| Ridge Regression | -0.0928 | 0.1612 | 8.69 |
| Lasso | -0.0927 | 0.1609 | 11.15 |
| Random Forest Regression | -0.0108 | 0.2055 | 145.63 |
| Multilayer Perceptron Regression | -0.0930 | 0.1597 | 2492.16 |

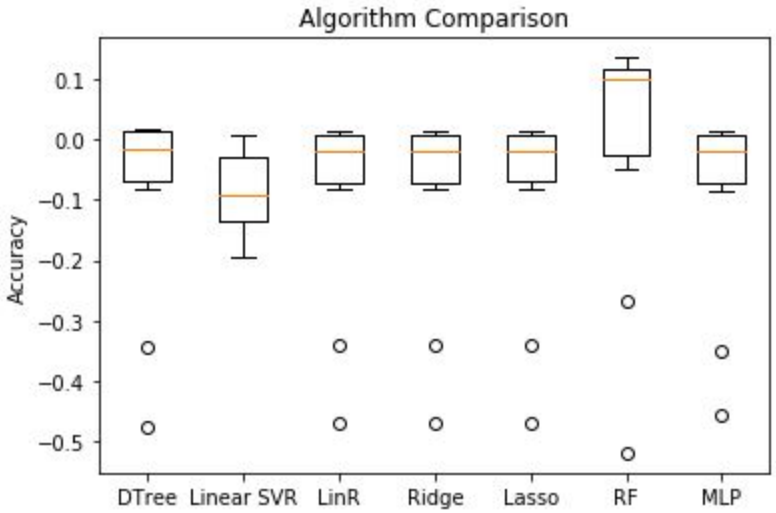

**Fig 7**. *Boxplot showing accuracy comparison of algorithms for cross validation results in Table 2.*

The summary result of linear regression for MTA entries with average weekday profile removed is given in Fig. 12 of Appendix. The coefficient of determination ($R^2$) is lower (0.002) than with the previous case (0.234). The estimate coefficients of the features have high significance values ($p<0.05$) for clear, humidity, rain and snow. Features like clear and humidity have a positive impact on the commuter volume (Table 3), whereas snow has a huge negative impact. Table 3 shows individual results of different models for dataset with average weekday profile removed and the $R^2$ of the prediction. In this case, the random forest regression has the highest $R^2$ among all the other regression models (Table 3). Contrary to the previous case, temperature and snow are the most important features and clear the least. Table 4 gives the K-Fold Cross Validation Results of different Models for number of splits = 10, and their running time performance for the dataset with weekday profile removed. The score results are higher as compared to the previous case with random forest regression having the highest accuracy (0.0630) of all the other models (Fig. 8). Although the running time efficiency is still low, it is better than multilayer perceptron regression and linear SVR. Fig. 10 shows the decision tree for the MTA entries with average weekday profile removed. The figure shows that in general people prefer less humid weather (less than 41% approx.) for commuting.

**Table 4.** *Individual Results of different Models and their coefficient of determination ($R^2$) of the prediction for MTA entries with average weekday profile removed*

| Model | Coefficient of determination ($R^2$) | Other Information |
|---|---|---|
| Decision Tree Regression | 0.0101 | Feature Importance -<br>1. Temperature: 0.5443<br>2. Snow: 0.3009<br>3. Humidity: 0.1547<br>4. Rain: 0.0<br>5. Clear: 0.0<br>6. Cloudy: 0.0 |
| Linear Support Vector Regression (SVR) | 0.0008 | Estimated Coefficients -<br>1. Clear: -3.7009e+00<br>2. Humidity: -5.1876e-01<br>3. Rain: -1.1211e+00<br>4. Snow: -1.2097e+02<br>5. Temperature: 1.9359e-02<br>6. Cloudy: -9.4227e+00 |

| | | |
|---|---|---|
| Linear Regression | 0.002 | Estimated Coefficients ($p<0.005$)<br>1. Clear: 68.4738<br>2. Humidity: -0.6241<br>3. Rain: 75.9894<br>4. Snow: -337.5063 |
| Ridge Regression | 0.0022 | Estimated Coefficients -<br>1. Clear: -35.7596<br>2. Humidity: -1.2962<br>3. Rain: 1.4584<br>4. Snow: -441.0935<br>5. Temperature: -0.6547<br>6. Cloudy: -90.4060 |
| Lasso | 0.0022 | Estimated Coefficients -<br>1. Clear: -29.6394<br>2. Humidity: -1.2966<br>3. Rain: 3.7192<br>4. Snow: -419.7801<br>5. Temperature: -0.6381<br>6. Cloudy: -84.0144 |
| Random Forest Regression | 0.01162 | Feature Importance -<br>1. Temperature: 0.5300<br>2. Snow: 0.2515<br>3. Humidity: 0.1556<br>4. Cloudy: 0.0616<br>5. Rain: 0.0013<br>6. Clear: 0.0 |
| Multilayer Perceptron Regression | 0.0049 | - |

**Table 5**. *K-Fold Cross Validation Results of different Models and their performance for MTA entries with average weekday profile removed*

| Model | K-Fold Cross Validation Results for no. of splits = 10 | | Performance (seconds) |
|---|---|---|---|
| | Mean | Standard Deviation | |
| Decision Tree Regression | 0.0079 | 0.0079 | 6.06 |
| Linear Support Vector Regression (SVR) | 0.0010 | 0.0004 | 1006.27 |
| Linear Regression | 0.0019 | 0.0013 | 3.59 |
| Ridge Regression | 0.0019 | 0.0013 | 2.37 |

| | | | |
|---|---|---|---|
| Lasso | 0.0019 | 0.0010 | 3.35 |
| Random Forest Regression | 0.0630 | 0.0385 | 145.85 |
| Multilayer Perceptron Regression | 0.0048 | 0.0047 | 2314.78 |

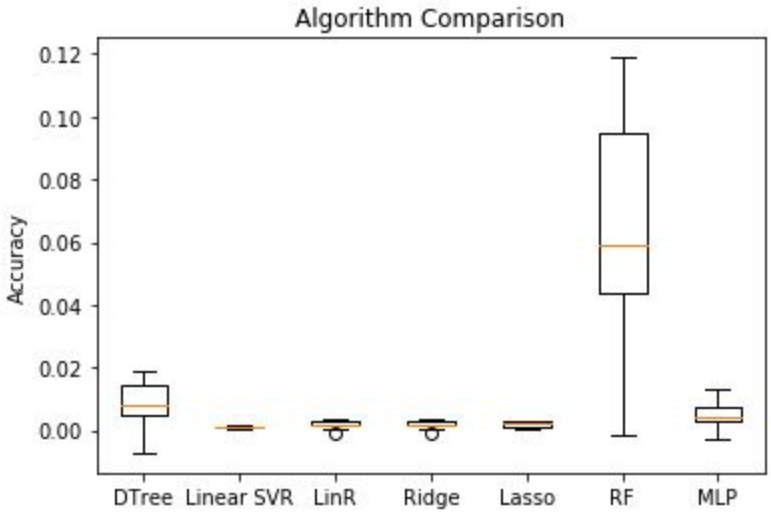


**Fig 8**. *Boxplot showing accuracy comparison of algorithms for cross validation results in Table 3.*

## 7. CONCLUSIONS

This research looked to determine the effect that weather has on MTA commuter volumes and if the chosen weather features could be used for prediction. It was found that all features used (temperature, humidity, and the weather conditions: cloudy, rain, clear, and snow) had a significant effect on commuter volumes, specifically for the case when MTA entries with average weekday profile were not removed . The predictive power was found to be low for all models, with random forest regression model performing the best based on the models' cross validation scores. This implies that there are other significant factors in determining commuter volume that are not accounted for in this report. The normalization by average weekday profile decreased the overall model performance but improved the accuracy of the models. Lastly, people generally prefer low humidity weather for commuting with ideal temperature to be between 41 °F and 76°F approximately.

## 8. LIMITATIONS AND FUTURE WORK

Weather is only one of many factors that influence commuter traffic, which means that the predictive power based solely on weather alone is quite low. In order to improve the predictive power of the model,

more features could be added to create a more informed model, such as wind speed, precipitation, or incorporating whether or not that day is a holiday. Additionally, running this over a longer timespan of results would remove some of the biases of 2016 and provide a more generalized model. With a longer timespan, temporally-aware models could be used, including hidden Markov models or LSTM neural networks.

It was seen during the time series analysis that much of the MTA traffic flow is generated right before and right after the typical workday (9am-5pm). Commuting to work is a necessity and it is hypothesised that people are likely to be less likely to adjust their travel plans when they're travelling to work. This should be investigated further by running the analysis separating the times were there are high volumes of people traveling to or from work to see if the effect is less pronounced. This might be able to add more fidelity to the model if this can be accounted for.

Further investigation could be performed to look into a multi-modal transportation model to see if shifts in transportation volume in one mode can be seen across others, like a dip in MTA traffic results in an increase in Uber or taxi usage. Lastly, ensemble modelling wherein relevant strong learning algorithms like random forest are stacked together, could be performed to obtain better predictive performance.

## CONTRIBUTIONS

The team most commonly met and worked in collaborative coding sessions, therefore it is difficult to assign roles specifically.

- *Ben Steers: data cleaning, time series, report*
- *Chun-Chieh Tsai: data acquisition, time series, report*
- *Isha Chaturvedi: data merging, modeling, report*
- *Rachel Lim: data acquisition, data cleaning, report*
- *Yi Xuan Tang: data merging, time series, report*

## REFERENCES


1. Böcker, L. Dijst, M. and Prillwitz, J (2013) Impact of Everyday Weather on Individual Daily Travel Behaviours in Perspective: A Literature Review, Transport Reviews, 33(1): 71-91.
2. Call, D.A. (2011). The effect of snow on traffic counts in western New York State. Weather Climate and Society, 3, 71–75.
3. Guo, Z., Wilson, N.H.M., & Rahbee, A. (2007). The impact of weather on transit ridership in Chicago, Illinois. Transportation Research Record: Journal of the Transportation Research Board, 2034, 3–10.

4. Gupta, M.K, Gupta, S. and Rawal, R.K (2016) Impact of artificial neural networks in QSAR and computational modelling. Artificial Neural Network for Drug Design, Delivery and Disposition, 152-179.
5. MTA (2016) Turnstile Data. Retrieved from: http://web.mta.info/developers/turnstile.html
6. MTA(2017) The MTA Network. Retrieved from: http://web.mta.info/mta/network.htm
7. Muller, S., Tscharaktschiew, S., Haase, K. (2008) Travel-to-school mode choice modelling and patterns of school choice in urban areas. J. Transp. Geogr. 16, 342–357.
8. Stern, A.D, Shah, V., Goodwin, L.C. and Pisano, P (2003) Analysis of weather impacts on traffic flow in metropolitan Washington DC. Federal Highway Administration (FHWA), Washington, DC, USA.
9. Stover, V.W. and McCormack E.D. (2012) The Impact of Weather on Bus Ridership in Pierce County, Washington. Journal of Public Transportation, 15 (1): 95-110
10. Tao, S, Corcoran, J., Rowe, F. and Hickmand, M (2018) To travel or not to travel: 'Weather' is the question. Modelling the effect of local weather conditions on bus ridership. Transportation Research, 86: 147–167.
11. The Weather Company (2017) Weather data. Retrieved from: http://wunderground.com
12. Tsapakis, I., Cheng, T. and Bolbol, A (2013) Impact of weather conditions on macroscopic urban travel times. Journal of Transport Geography, 24: 204-211.
13. Wang, L., Yamamoto, T., Miwa, T., Morikawa, T. (2006) An analysis of effects of rainfall on travel speed at signalized surface road network based on probe vehicle data. In: Proceedings of the Conference on Traffic and Transportation Studies, ICTTS, Xi'an, China, 2–4 August, pp. 615–624.
14. Zhou, M., Wang, D., Li, Q., Yang, Y., Tu, W., Cao, R (2017) Impacts of weather on public transport ridership: Results from mining data from different sources, Transportation Research, 75(3): 17–29

## APPENDIX

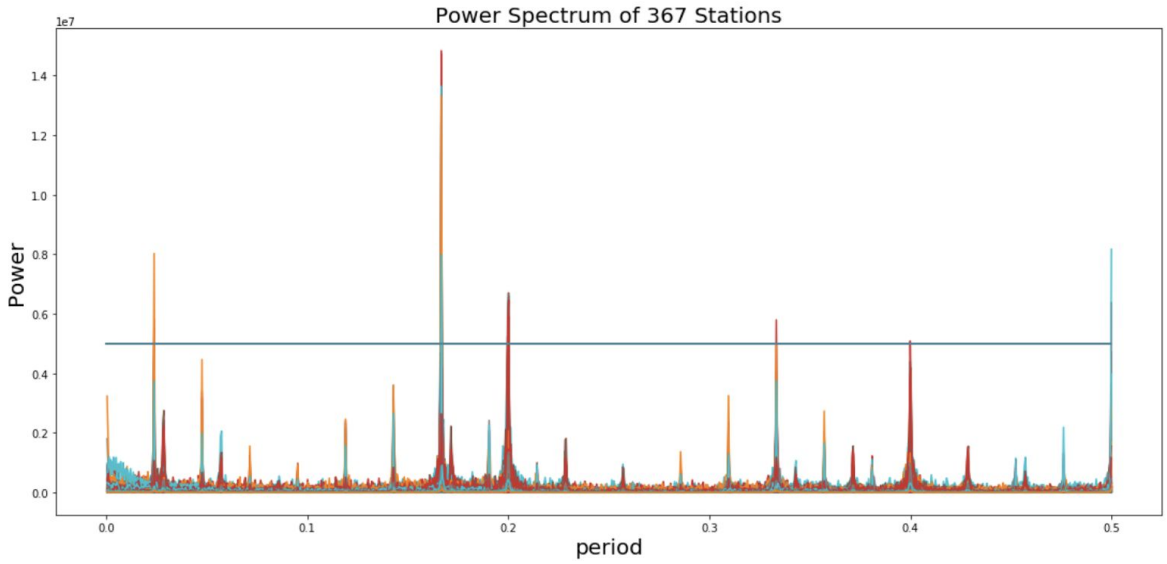


**Fig 2.** *Frequency spectrum for all stations.*

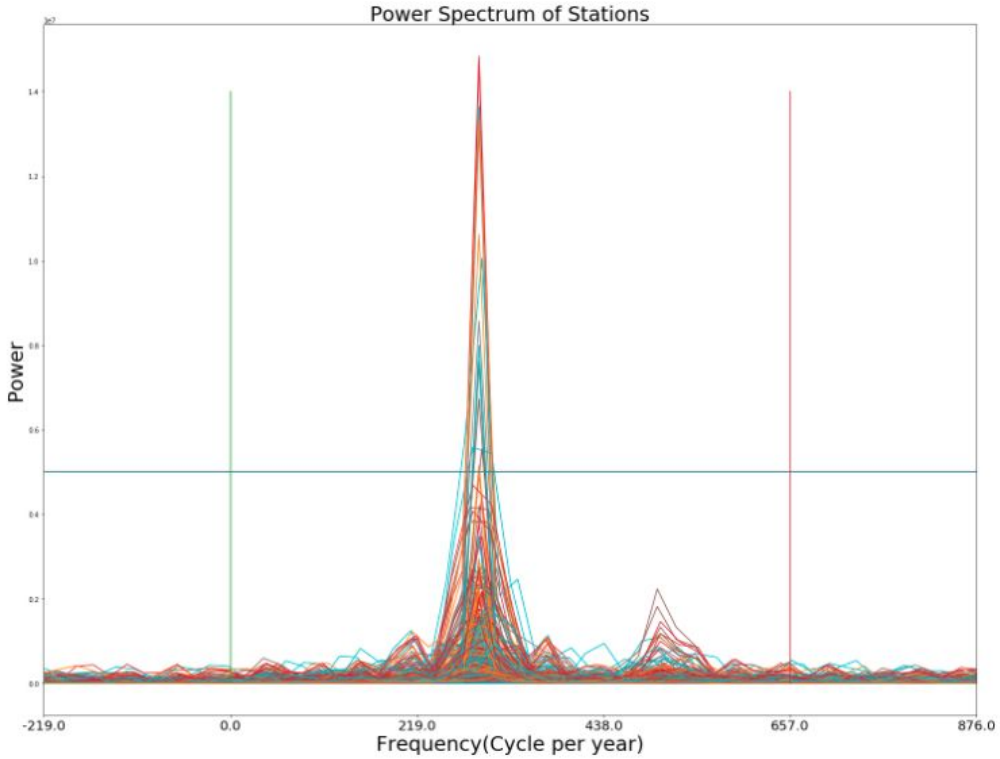


**Fig 2.** *Highlight of the frequency between 250.4 and 383.25 cycles per year*

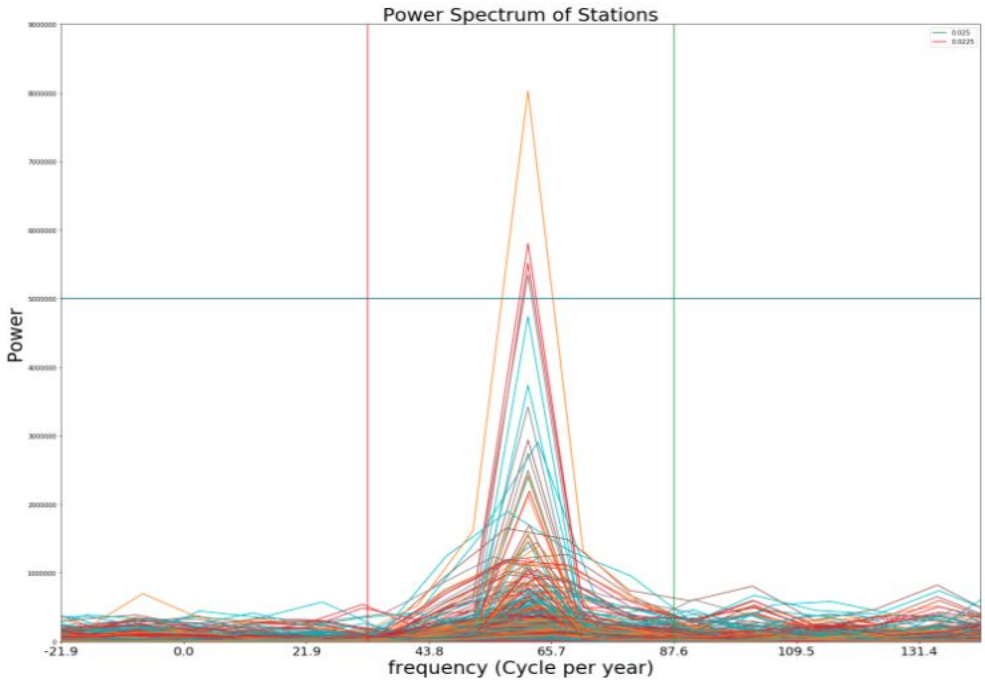


**Fig 3.** *Highlight of the frequency between 43.8 and 60 cycles per year*

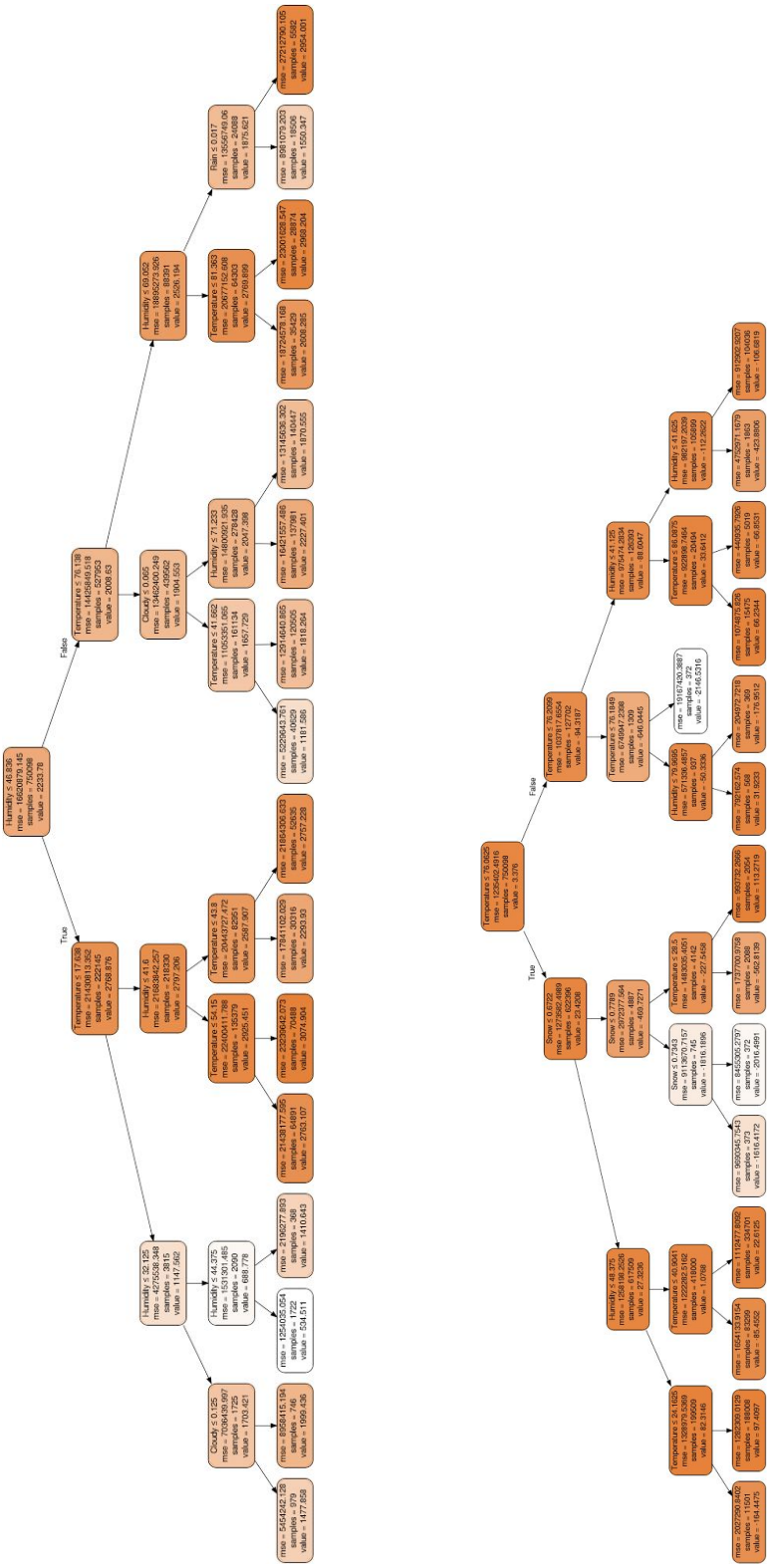

**Fig 9 & 10.** Decision Tree for Decision Tree Regression Model (right), Decision Tree for Decision Tree Regression Model for MTA entries with average weekday profile removed (left)

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | ENTRIES_hourly | R-squared: | 0.234 |
| Model: | OLS | Adj. R-squared: | 0.234 |
| Method: | Least Squares | F-statistic: | 3.813e+04 |
| Date: | Sat, 09 Dec 2017 | Prob (F-statistic): | 0.00 |
| Time: | 20:49:30 | Log-Likelihood: | -7.2986e+06 |
| No. Observations: | 750098 | AIC: | 1.460e+07 |
| Df Residuals: | 750092 | BIC: | 1.460e+07 |
| Df Model: | 6 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [95.0% Conf. Int.] | |
|---|---|---|---|---|---|---|
| Clear | 1732.7414 | 16.698 | 103.770 | 0.000 | 1700.014 | 1765.469 |
| Humidity | -17.5881 | 0.286 | -61.563 | 0.000 | -18.148 | -17.028 |
| Rain | 2485.7491 | 37.151 | 66.909 | 0.000 | 2412.934 | 2558.565 |
| Snow | 2129.0014 | 61.418 | 34.664 | 0.000 | 2008.624 | 2249.379 |
| Temperature | 24.0281 | 0.251 | 95.567 | 0.000 | 23.535 | 24.521 |
| Cloudy | 2062.8110 | 19.833 | 104.008 | 0.000 | 2023.938 | 2101.683 |

| | | | |
|---|---|---|---|
| Omnibus: | 752387.683 | Durbin-Watson: | 0.762 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 42480069.696 |
| Skew: | 5.031 | Prob(JB): | 0.00 |
| Kurtosis: | 38.467 | Cond. No. | 1.12e+03 |

**Fig 11.** Linear regression results

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | norm_entries | R-squared: | 0.002 |
| Model: | OLS | Adj. R-squared: | 0.002 |
| Method: | Least Squares | F-statistic: | 213.0 |
| Date: | Sat, 09 Dec 2017 | Prob (F-statistic): | 1.13e-272 |
| Time: | 22:40:32 | Log-Likelihood: | -6.3245e+06 |
| No. Observations: | 750098 | AIC: | 1.265e+07 |
| Df Residuals: | 750092 | BIC: | 1.265e+07 |
| Df Model: | 6 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Clear | 68.4738 | 4.555 | 15.034 | 0.000 | 59.547 | 77.401 |
| Humidity | -0.6241 | 0.079 | -7.940 | 0.000 | -0.778 | -0.470 |
| Rain | 75.9894 | 10.134 | 7.499 | 0.000 | 56.127 | 95.851 |
| Snow | -337.5063 | 16.765 | -20.132 | 0.000 | -370.364 | -304.648 |
| Temperature | -0.0500 | 0.069 | -0.724 | 0.469 | -0.185 | 0.085 |
| Cloudy | 3.3059 | 5.527 | 0.598 | 0.550 | -7.527 | 14.139 |

| | | | |
|---|---|---|---|
| Omnibus: | 376752.591 | Durbin-Watson: | 1.577 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 594472222.053 |
| Skew: | -0.830 | Prob(JB): | 0.00 |
| Kurtosis: | 140.905 | Cond. No. | 1.12e+03 |

**Fig 12**. Linear regression results for MTA entries with average weekday profile removed.