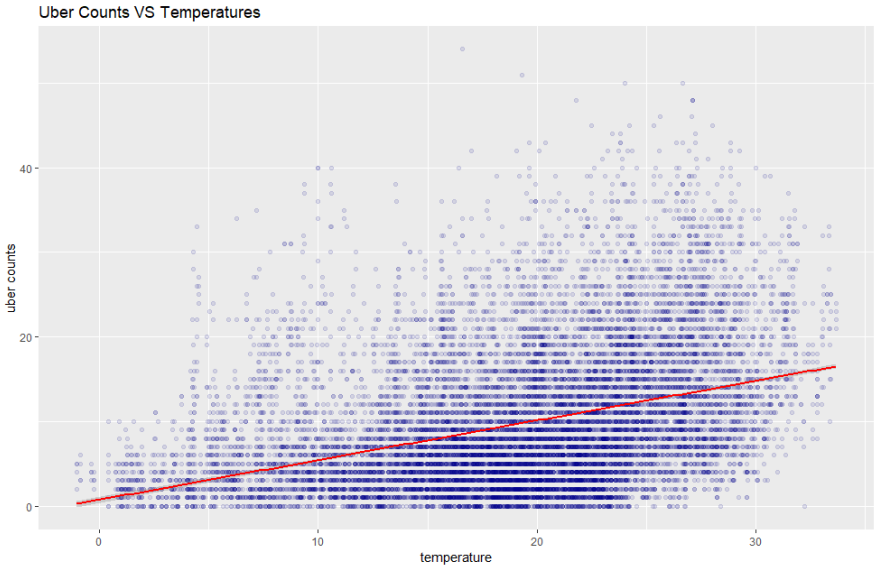
Final Project – Predicting Uber usage using external data

חן בן גל – 311337356 עמרי ז'אן – 302121140

Exploratory Analysis by Feature

Temperatures

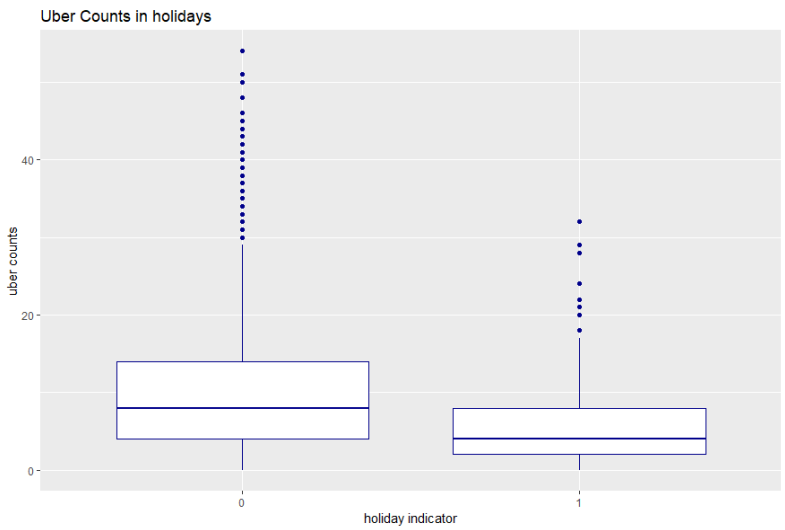
As the first external-data feature, we chose to add to the training set the observed hourly temperatures in NYC for the given dates. We then took the hourly temperatures we found and scaled them so that they seem continuous in the 15-minute intervals. For instance, if temperature at 13:00 was 24C and at 14:00 it was 28C then for 13:15, 13:30 and 13:45 we used 25C, 26C and 27C respectively. We then used those to predict the temperature for the testing set (more on that later). The following plot describes the relationship between the Uber counts for each interval and the observed temperatures (red line is the regression line):



In this plot we can see a positive relationship between the two variables, and if we will look at the summary of the regression we will see that the relationship is statistically significant.

Holidays

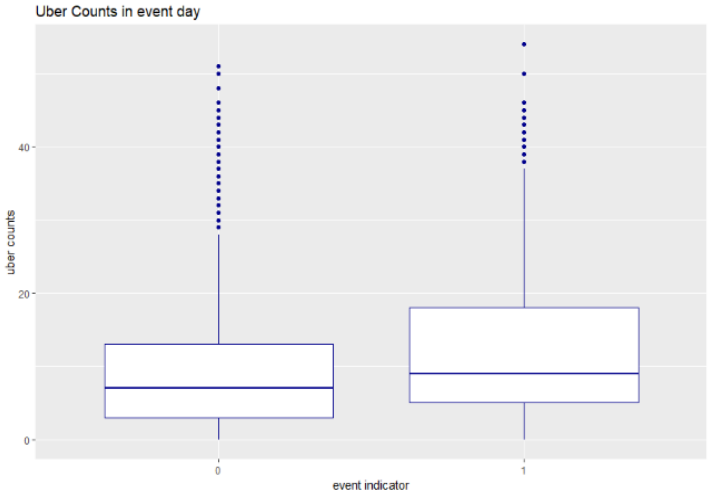
It seemed logical to us that during holidays the use of transportation, including Uber, varies from that of a standard weekday or even weekend. We marked all the national holidays that include a day off (in both the training set and testing set) as 1 and the rest of the days as 0.

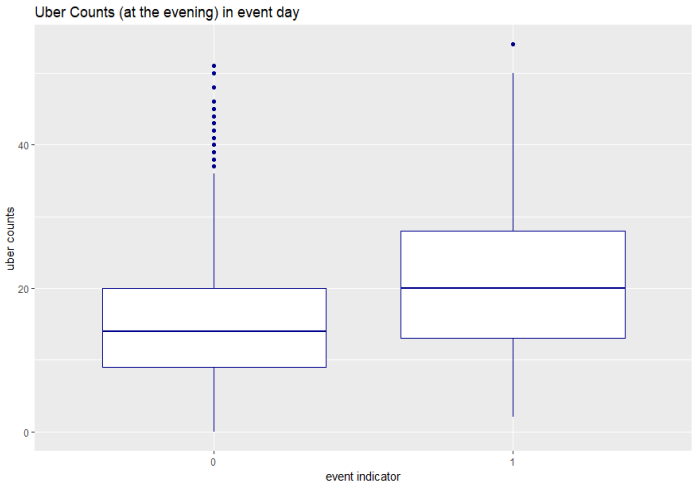


It is clear to see that Uber usage is much lower during national holidays.

Events

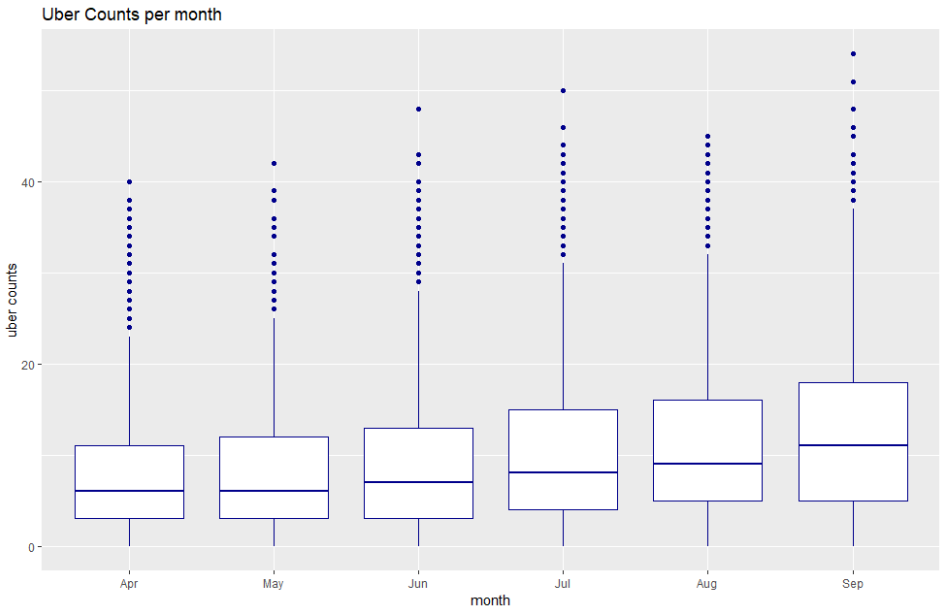
In our research we saw that around the area of the NYSE there are a couple of stadiums such as Madison Square Garden and Barclays Center that host concerts, sport event etc. We figured people will arrive to the events by some kind of transportation (such Uber) and leading up to the event we expected to see an increase in Uber counts. We marked all the event days as 1 and the rest of the days as 0. Our assumption was that in an event day a lot of people order Uber at around the same time (nearing the event) so the count in the intervals just before the event should be high. To validate our assumption, the next step was to look what happens to the counts of Uber in the evening of the event. The plot on the left shows Uber use during event days vs. non event days. The one on the right shows the same thing, except both boxplots are based on observations from 17:00 to 22:00 only.





Month

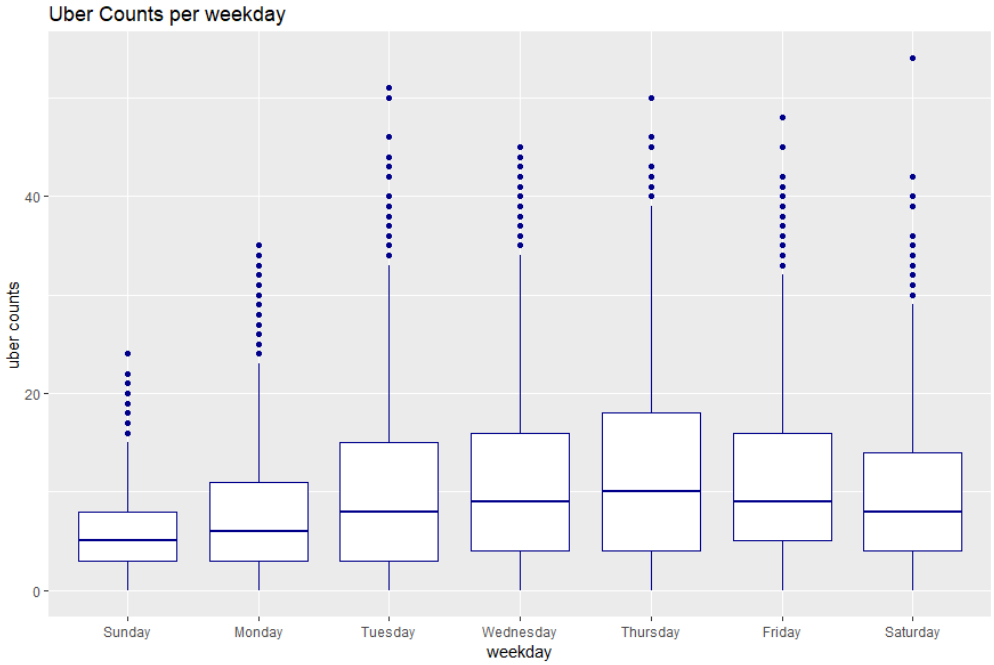
Uber usage is also influenced by the month of the year. This could be due to seasonality or even because of the growth of the company in each month. The following plot presents Uber count distribution by the months appearing in the dataset:



We can see a constant increase in the Uber count distribution by the following months.

Weekday

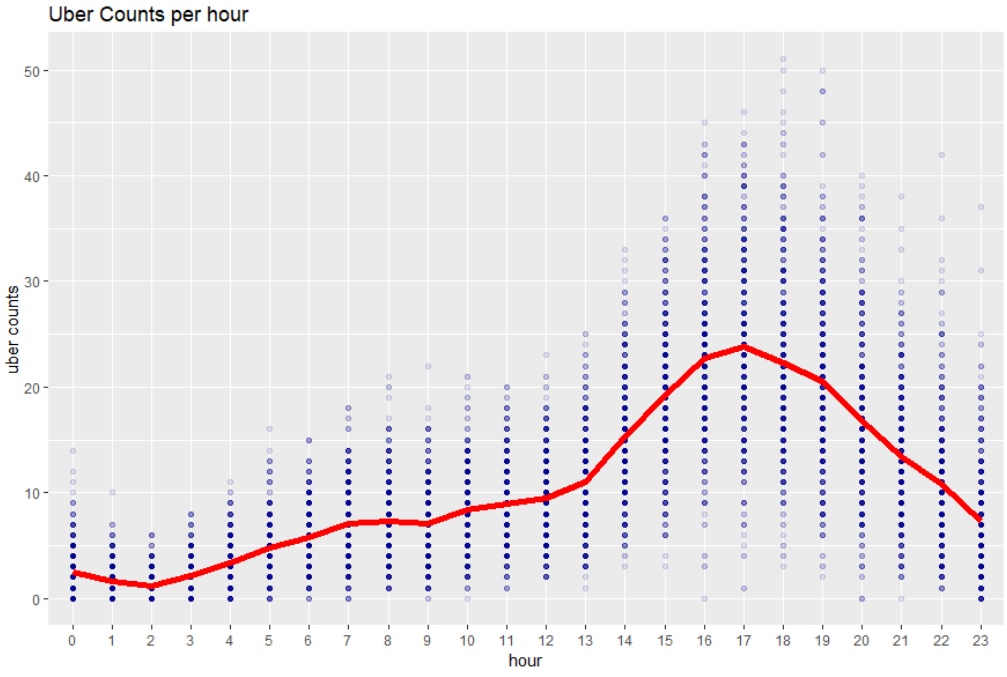
Uber use is also influenced by the day of the week. The following plot presents Uber count distribution by day of the week:



We can see that Sunday is the weakest day in terms of Uber counts, and the distribution of the Uber counts during the working day is increasing until the beginning of the weekend (Friday).

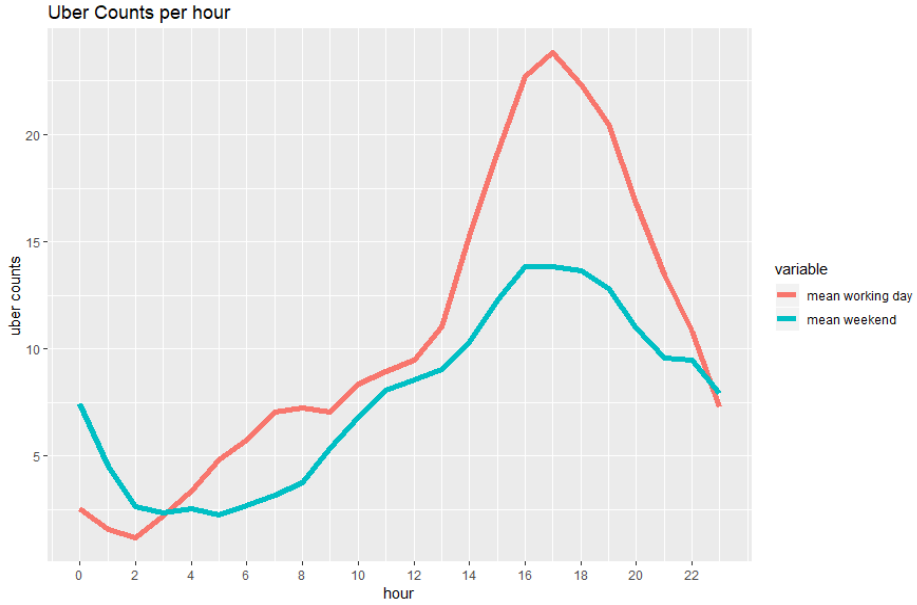
Hour

Uber usage depends on the hour of the day. For example, during sleeping hours the count of Uber orders will be low. The next plot shows the distribution of Uber counts for each hour. The red line is the connected means of Uber count for each hour.



In this plot we see that the higher counts in Uber orders happen around 15:00-20:00.

Additionally, each weekday possesses different transit "rush hours". For example, we expect to see a higher amount of Uber orders around 17:00 (as people leave their workplace) on working days than to weekends. In the next plot we show the difference in hourly use between working days and weekends:

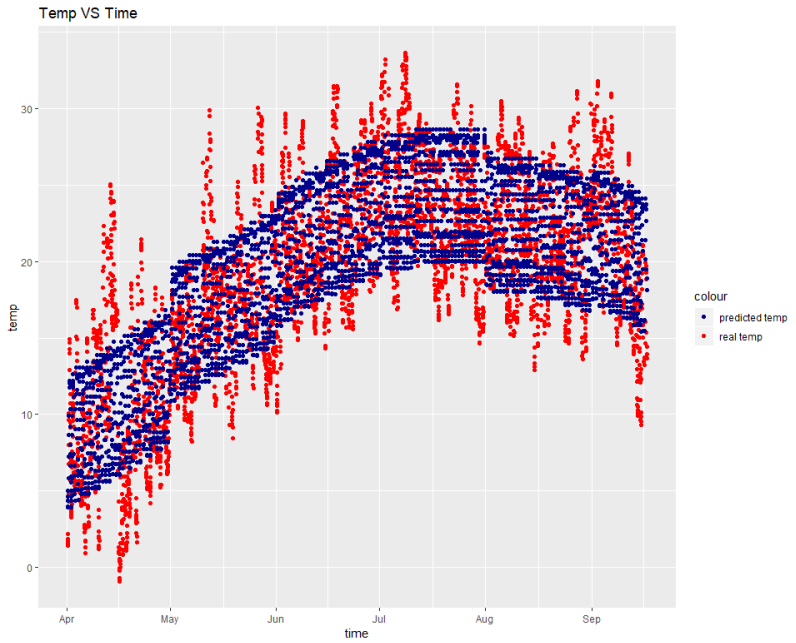


We can see that the hourly means vary between working days and weekends. For example, on weekends there are more Uber orders during the late hours (00:00-03:00) but less orders during the rest of the day, especially in the early evening as people leave the workplace.

Model Estimation

Weather Predictions

In order to use temperature as a feature in our Uber count prediction model, we needed to predict the temperature values for the testing set. Having read that weather forecasts 10 days ahead are usually no more accurate than historical averages, we chose to base our model on the latter. The model we used was a Linear Regression model based on 3 features: Daily historical averages for NYC (based on a data from 1891-2010), the hour of the day and the month of the year. The following plot presents the model's predictions (in blue) vs the actual weather (in red), for dates randomly sampled from the training set and constituting 30% of that data:

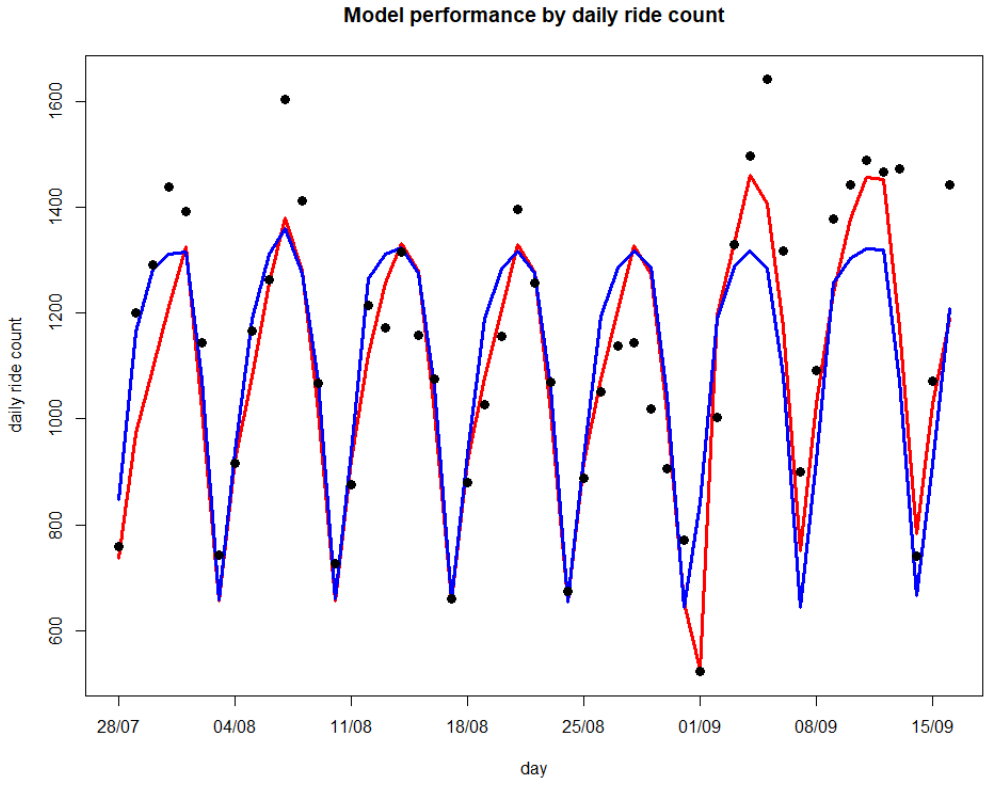


The results show that the model is pretty conservative (in that it avoids predicting extreme weather), which is expected when one considers the features used to build it.

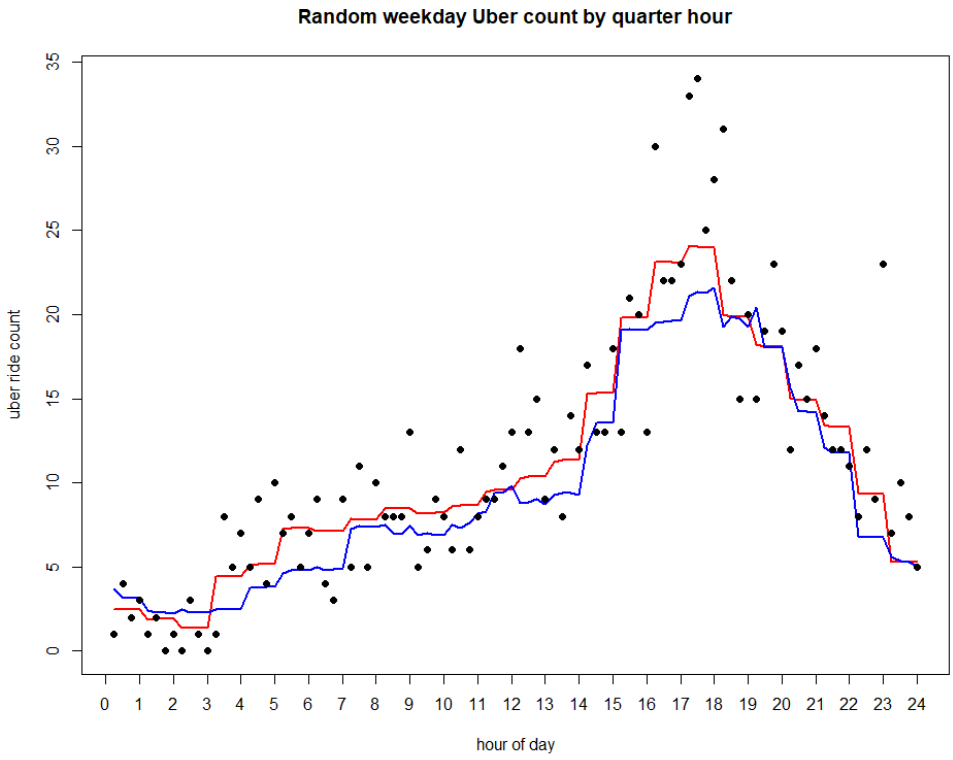
Model Design

For the final predictions, 2 different methods were tested: Linear Regression and Random Forests. Both models were built on the earlier 70% portion of the training set and validated on the remaining 30%. Feature selection for each of the models was made based on both, the knowledge gained during data exploration, and the features' influence on the R^2 values of predictions made on a validation set. For the Linear Regression model, we used the Box-Cox regression method (included in Appendix) to determine whether a transformation of the Y variable would be beneficial, and if so which one should be used. We decided, based on the method's output, to use the $\sqrt{y}$ transformation.

The next plot shows the daily Uber count for the 51 days that make up the validation set, and serves as an evaluation of the models' accuracy from a Macro perspective. The black dots indicate the actual Uber usage on a given day. The colored lines indicate the Linear Regression model's predictions (Red) and the Random Forest model's predictions (Blue):



Looking at the plot we notice that the Random Forest model follows a strict weekly pattern while the Linear Regression model is more "flexible" and usually gives better estimates for days that deviate from the expected pattern. To get a sense of things from a Micro perspective as well, we chose a random day from the validating set (in other words, a random black dot from the plot above) and plotted the 15-minute time interval counts for that day vs. Uber counts. As in the previous plot, black dots represent actual Uber counts and the red and blue lines represent the LM and RF models' predictions respectively:



It's important to remember that the plot above is particular to a specific date and that we would probably have somewhat different observations had we gotten a different date (We did take a look at most dates, but provided only one to meet the 6-page limit). With that being said, we can see that both models do relatively ok predicting the general pattern and not as well predicting variations within that pattern. More specifically, we can se that the Linear Regression model's changes in time resemble "stairs". The reason for that is that the values for the features within an hour (E.g. – 13:00,13:15,13:30 and 13:45) are the exact same with the exception of small changes in temperature (which were explained earlier). Therefore predictions within each hour are almost identical. Unfortunately, all the features we considered adding in order to explain within-hour variance (For example, Twitter tweet-count from our area for each 15-minute interval or Subway usage statistics) couldn't be known in advance (for the week of 24-30/09) and predictions for them would likely be based on the same features as our Uber count and therefore wouldn't solve the problem.

Final Model Selection:

In the end, we decided to go with the Linear Regression model. We chose that one partially because it did a better job predicting the unexpected surge in Uber use during the first 2 weeks of September (as seen it the daily count plot) but mostly due to the fact that it yielded a higher R^2 value for the validation set than its competitor (R^2 for LM predictions was 0.774, R^2 for RF predictions was 0.740).

Therefore, predictions for the test set are based on the following Linear Regression model:

$$\sqrt{Count}_i = Temp_i + Temp_i^2 + Temp_i^3 + Temp_i^4 + Hour_i + Month_i + Weekday_i + Weekday_i * Hour_i + Event_i * Hour_i + Holiday_i * Hour_i$$

With Hour and Weekday being used as categorical (factor), Event and Holiday being indicators and Temp and Month being continuous. The temperatures here are predictions based on the model explained in page 4.