NYC Transportation - Study of the Impact of Uber of Yellow Taxi

Waterloo Correlation One Datathon 2017

Analysis Report

Team: 25

Contents

1 *Overview*

This this report, we visualized and analyzed New York City's yellow taxi, green taxi, subway and uber trip data from April 2014 to June 2015 to understand how do different travel modes impact each other.

First, some visualizations and travel pattern analysis is conducted to study the effect of time, demographic variables on different modes of travel.

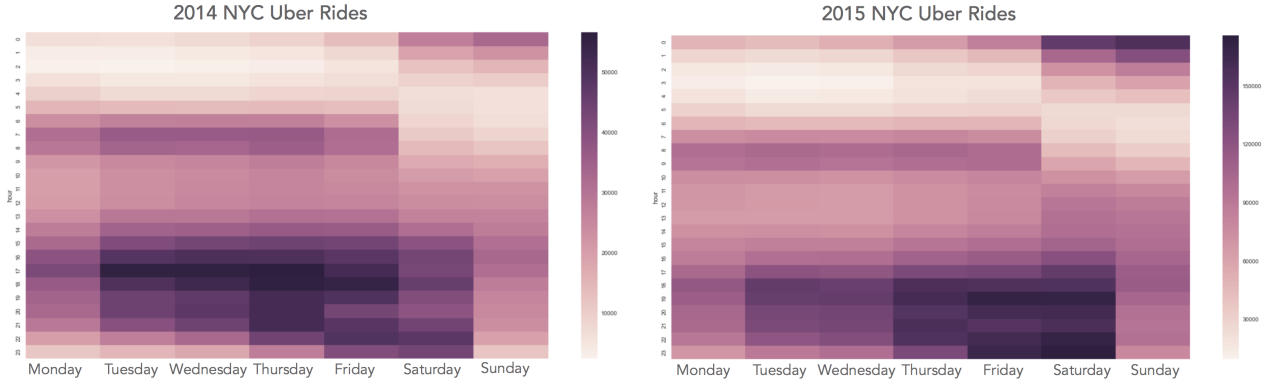
Then, a detailed time series analysis is conducted to study the traffic pattern and predict the future. linear regression model is made to study how other traffic affect the trip count of yellow taxi.

2 *Visualization - Time Study*

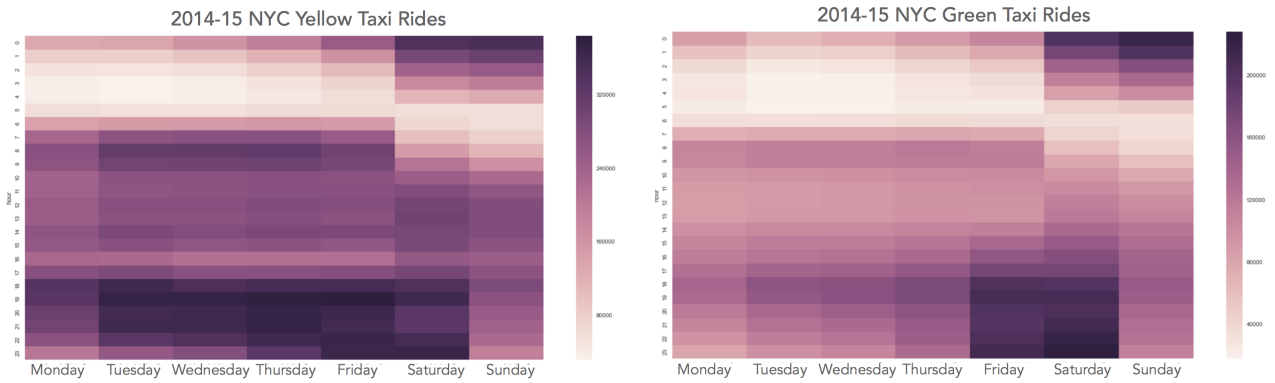
[Used python with seaborn to create heatmap]

In this section, we study how time related variables affect the trip counts and travel patterns by passengers.

This is the 2014 and 2015 Uber data segmented by weekday and time of day (0 to 23). We can see that the most busy time for Uber drivers are weekday evenings (from 4pm to 9pm) which is when people finish work. We can also see that Wednesday is the time when more people use Uber at later night (possible guess: people likely to work longer on Wednesday). Friday and Saturday nights also a busy time (guess: people go out playing, partying), the traffic even continues until the midnight of next day. This pattern is stronger in 2015 than in 2014. We cannot quite confidently explain why the traffic is low on Monday night, maybe it is the begining of the week and people tends to work less and thus take mta to go home? Moreover, overall, the traffic on weekend also increaased in 2015 compared to 2014 (can be seen from the darkness of the color).



The following two heatmaps represent the yellow and green taxi data. We can see that the traffic of yellow taxi looks more stationary and are less time-affected. This might be the reason that drivers of yellow taxi are full time drivers as oppose to for Uber drivers, some of them work during the day and don't have time to run Uber. For green taxi, we can see a clear drop in the traffic. Green taxis are mostly used during Friday and Saturday night.

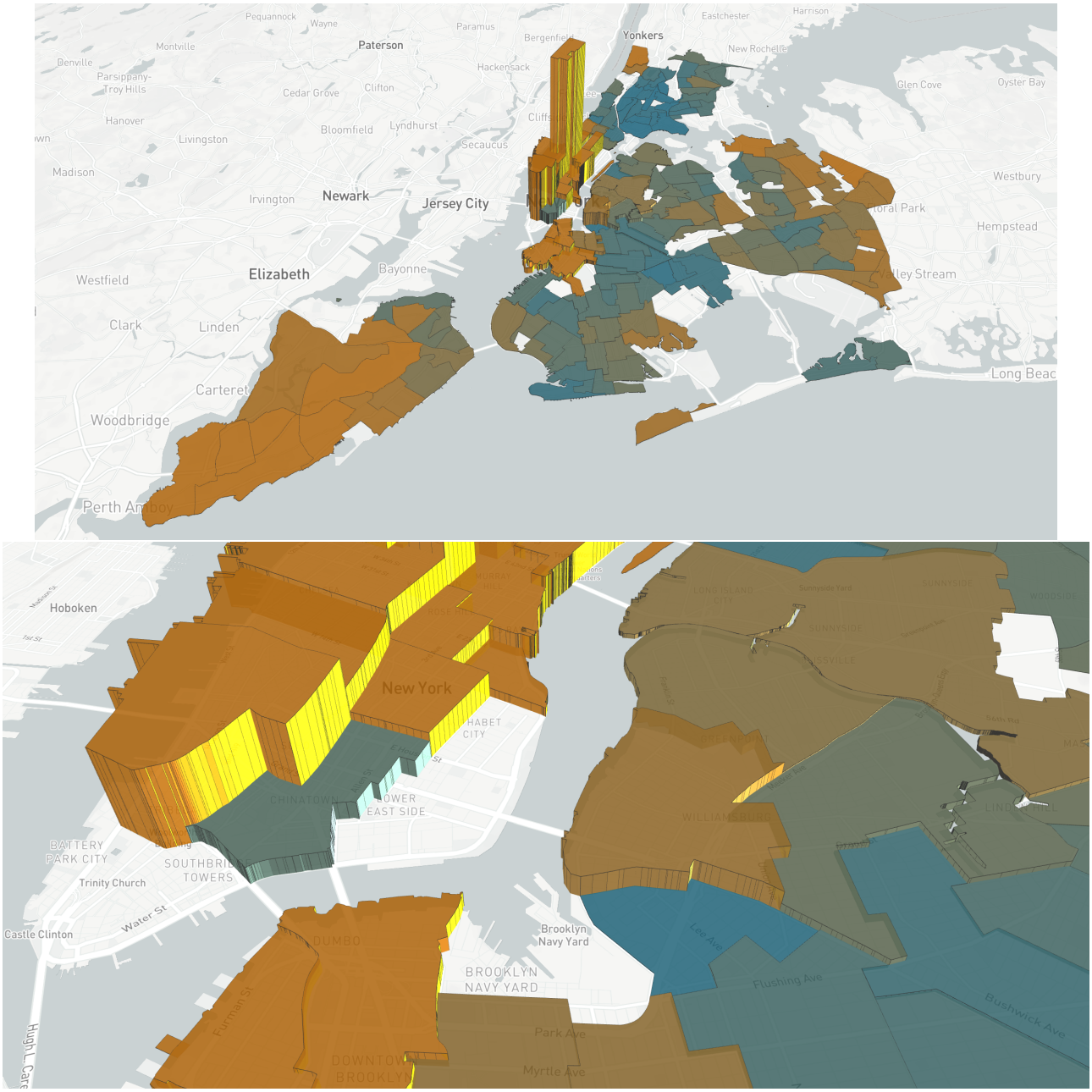


3 *Visualization - Demographic and Space Study*

This section studies how different modes of travel are affected and related to the demographic variables and loation within NYC.

3.1 Income Level

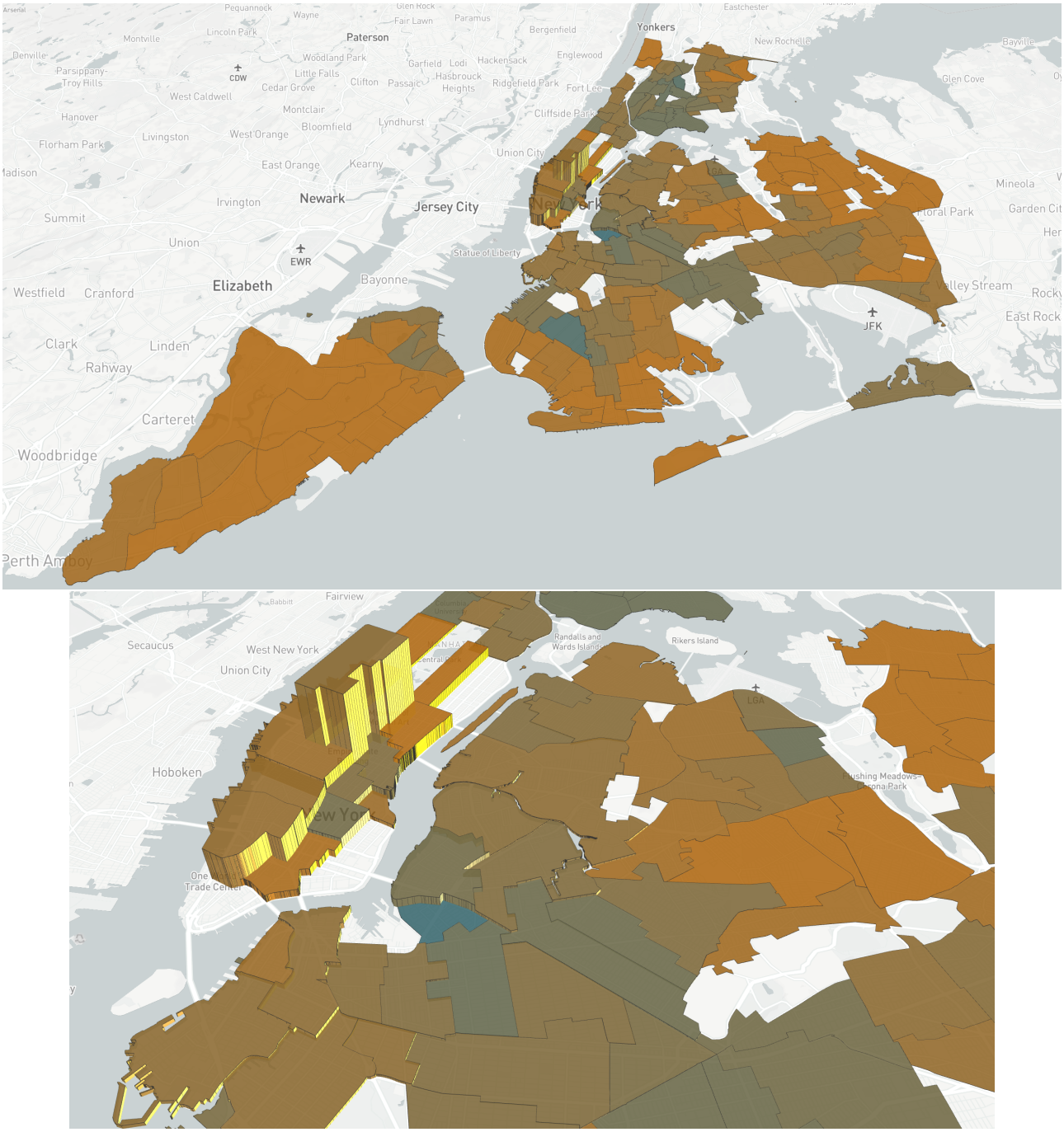
The following plot is a map of NYC. The color represents the income level (orange:high, blue:low). The heigh of the bar represents the Uber trip count per person. We can see that ingeneral, the trip count within the center of NYC is highly correlated to the income level. The higher the income, the more likely a person is to take Uber than other travel modes.



We also included a zoomed in map of the Manhattan region, which is the heart of NYC. We can see that even within Manhattan, the correlation pattern is clear. The orange regions have a much higher Uber trip per person than the more blue region.

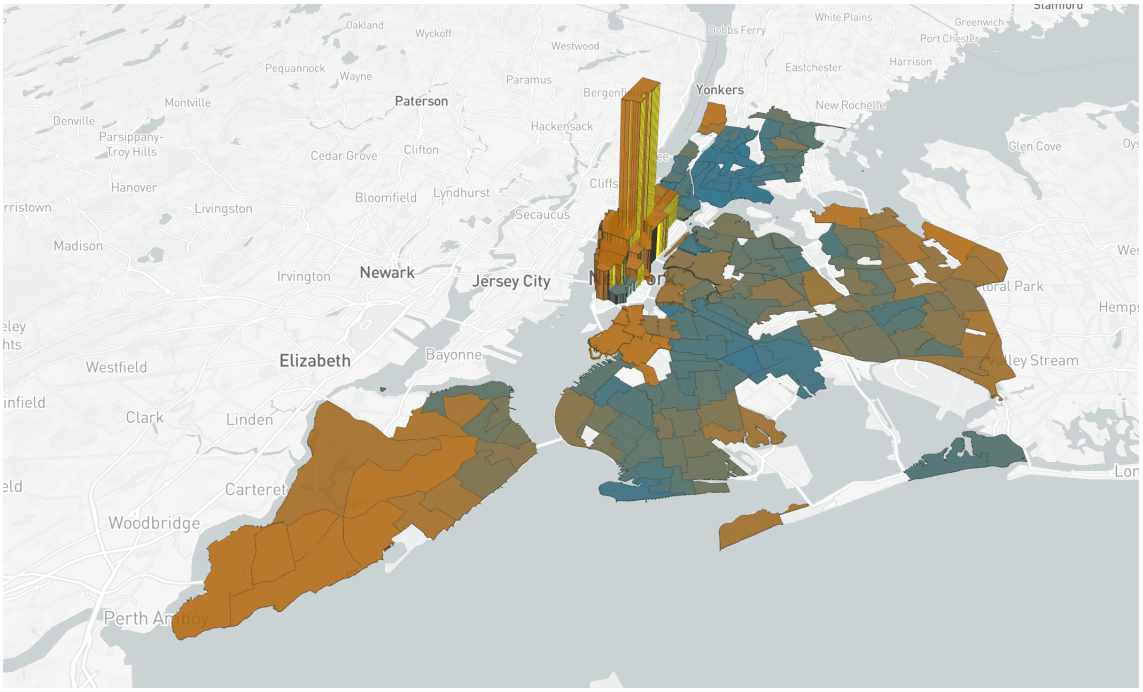
3.2 Age

This section studies how people at different ages have varing travel preferences. The median age per region is used in constructing the plot with a zoom in of Manhattan region.



We can see that for each region, there is less difference in the median ages (both old and young live in those areas), and we cannot see much color differences. The Uber trips are concenterated in the Manhattan region. Also, we can see how age is distributed across NYC. Younger people are likely to live in the centers of NYC and older people tend to live more in the further counties.

The following plot is the yellow taxi's trips per capital in each NTA. We can see that yellow taxi is very concentrated within manhattan city. If we compare this plot with the Uber plot, Uber's area is more spread to the neighbours of Manhattan.

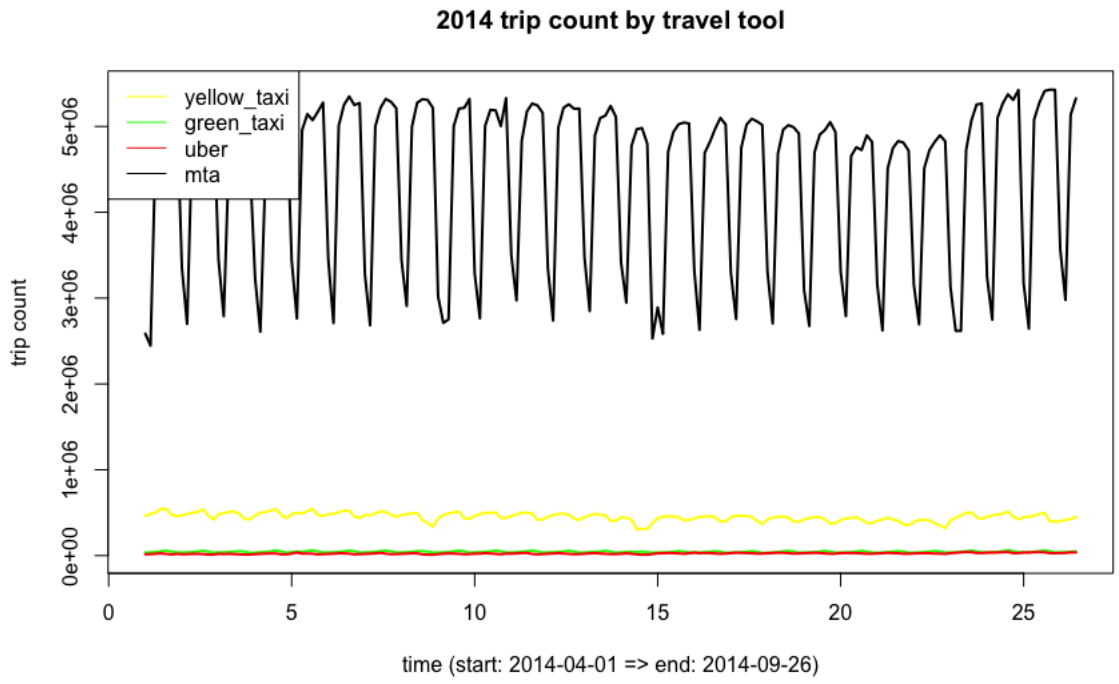


4 *Time Series Analysis*

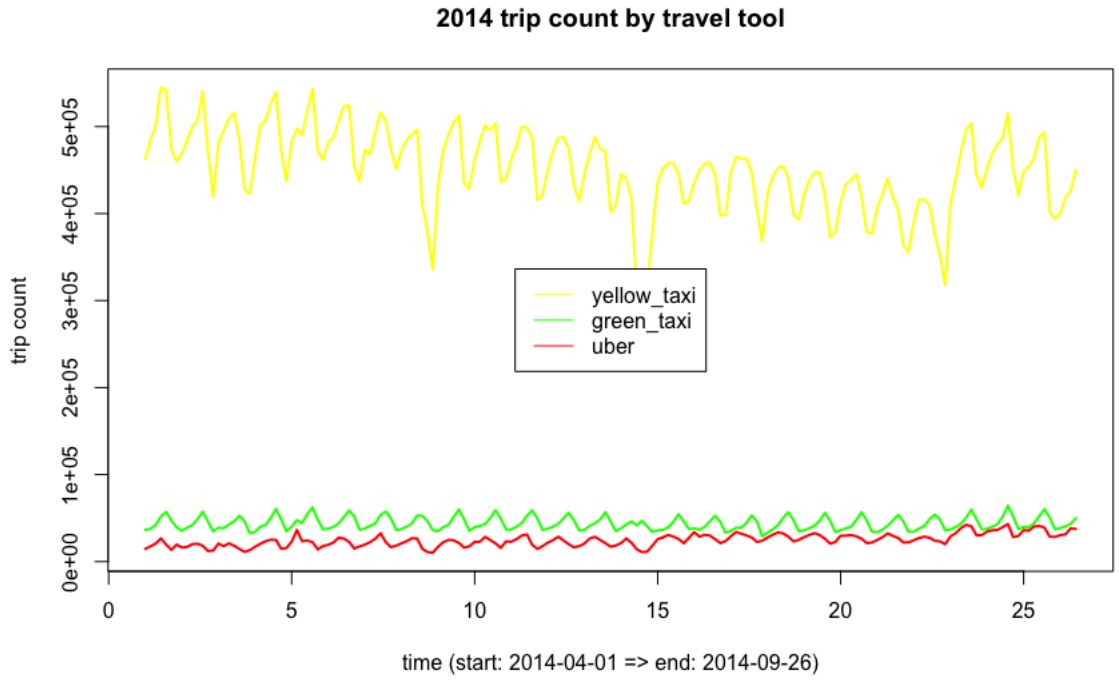
[Used R for time series analysis]

This section summarizes the time series analysis we performed to predict future trip count and the impact of Uber on yellow taxi. We mainly have four time series data: yellow taxi, green taxi, mta and Uber. In order to reduce the data same as well as having meaningful prediction, the data is aggregated by "day" and summed. For taxi and Uber data, the trip count is used instead of number of passengers (which we have in taxi data but not Uber data). For mta data, we used the "new_entries" column to get the total number of passengers and summed over by day.

First, simple visualization of the time series is made to study the general trend and relationship between four sources of data. We can see that mat has multiple times of traffic than other transportation modes, which makes sense as most people do travel by train during work days. It can be also seen that the time series data is highly seasonal, and there are different patterns for weekday vs. weekend. People take less trains on weekends because they don't need to go to work. This is also proved by visualizations from previous sections.



If we look closely to taxi and uber trips, yellow taxi have much higher traffic than gree taxi and uber. Similar weekly seasonal patterns can also be detected. it is also interesting to see that green taxi traffic is quite stationary and are less affected by other travel models.



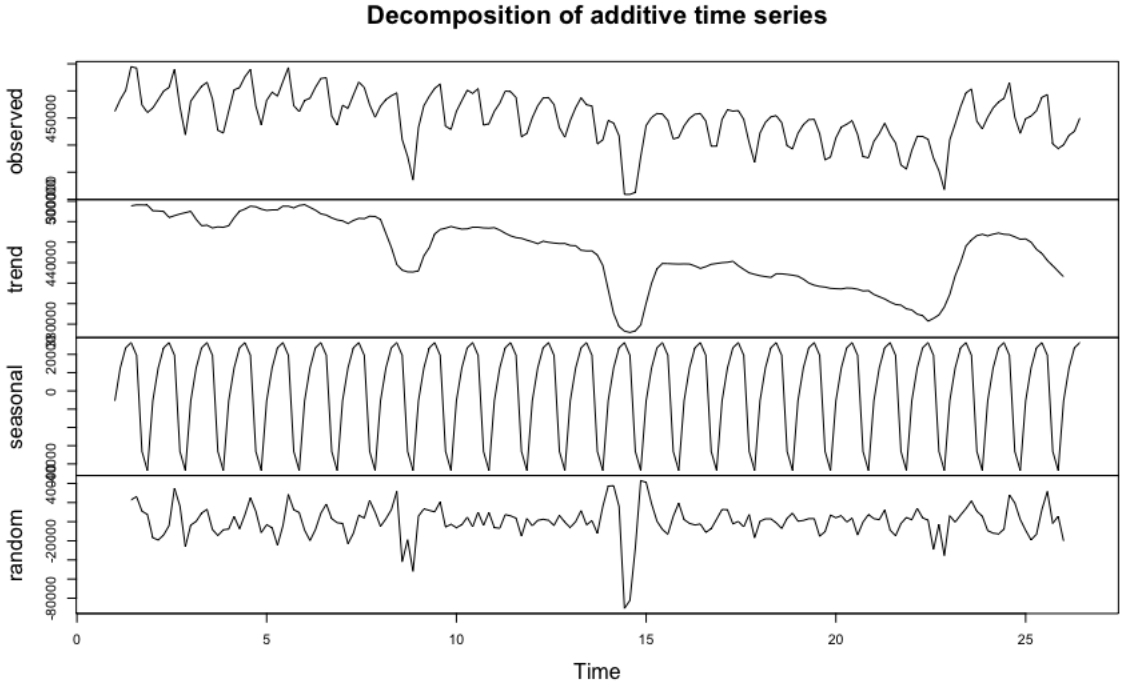
The next section models yellow taxi and uber traffic independently using Box-Jenkins models. The next next section models yellow taxi traffic data using time series data to study the impact and affect of uber and other travel modes on yellow taxi.

4.1 *Box-Jenkins Model (ARIMA) to Predict Yellow Taxi and Uber Traffic*

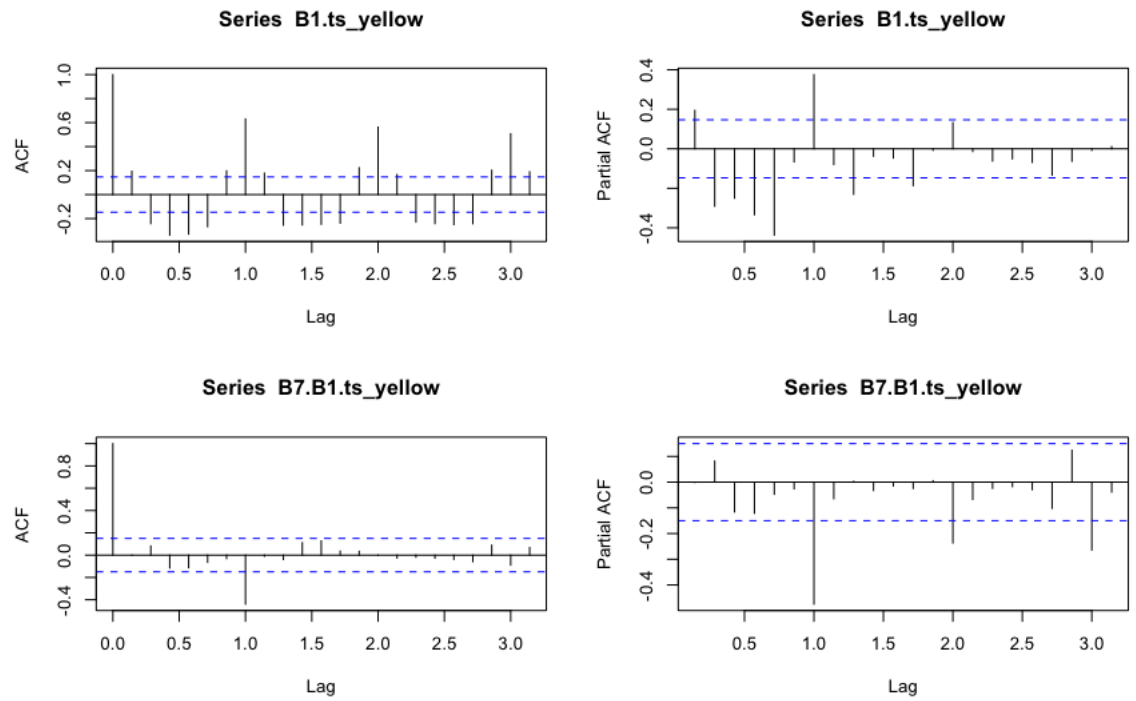
Since the given data is missing observations from October 2014 to December 2014 as well as last 4 days in September for mta data, we decided to use days with all four sources have overlapping information, which is from 2014-04-01 to 2014-09-26. Data from this period is used as "train" and then the prediction is made 270 days to the future so that we can compare the prediction result to the given 2015 data (which is from 2015-01-01 to 2015-06-30). Box-Jenkins models are fit and the best parameters are found using visual inspection of ACF(autocorrelation function) and PACF (partial autocorrelation function) plots.

4.1.1 Yellow Taxi ARIMA

We want to model First we decomposed the time series data and plotted ACF and PACF to understand the decomposition and the correlation between observations, so that we can find the parameters for ARIMA. From the decomposition, we can see that there is a in general declining trend and clear seasonal pattern. The ramdom component after removing trend and seasonality seems random noise and don't have clear patterns.

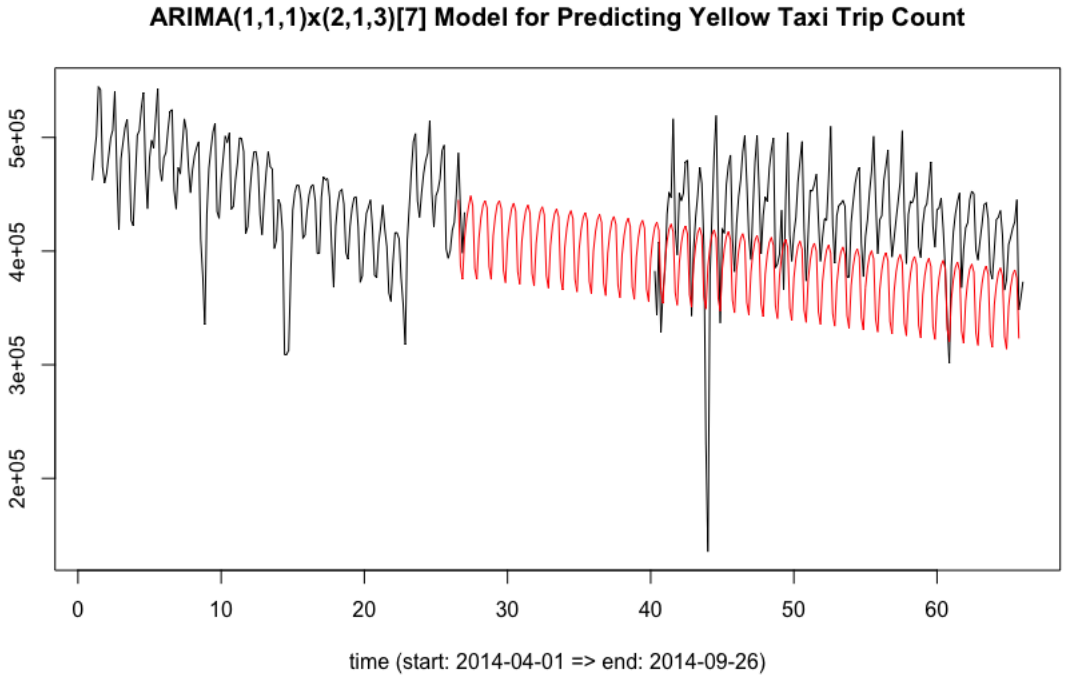


To remove the trend and weekly seasonality pattern, we difference the original time series with lag 1 and lag 7. Then, we yield the following ACF and PACF plots.



By visual inspection of the spikes at off-period time and on-period time, we decided to have the following model: ARIMA(1,1,1)x(2,1,3)[7] (reason: from bottom plots, for seasonality, there is 1 lag in ACF , and 3 lags in PCAF [only look at the spikes at integer lag points]; if we check spikes at non-integer lag points, all of them are within the denoted dotted threshold lines).

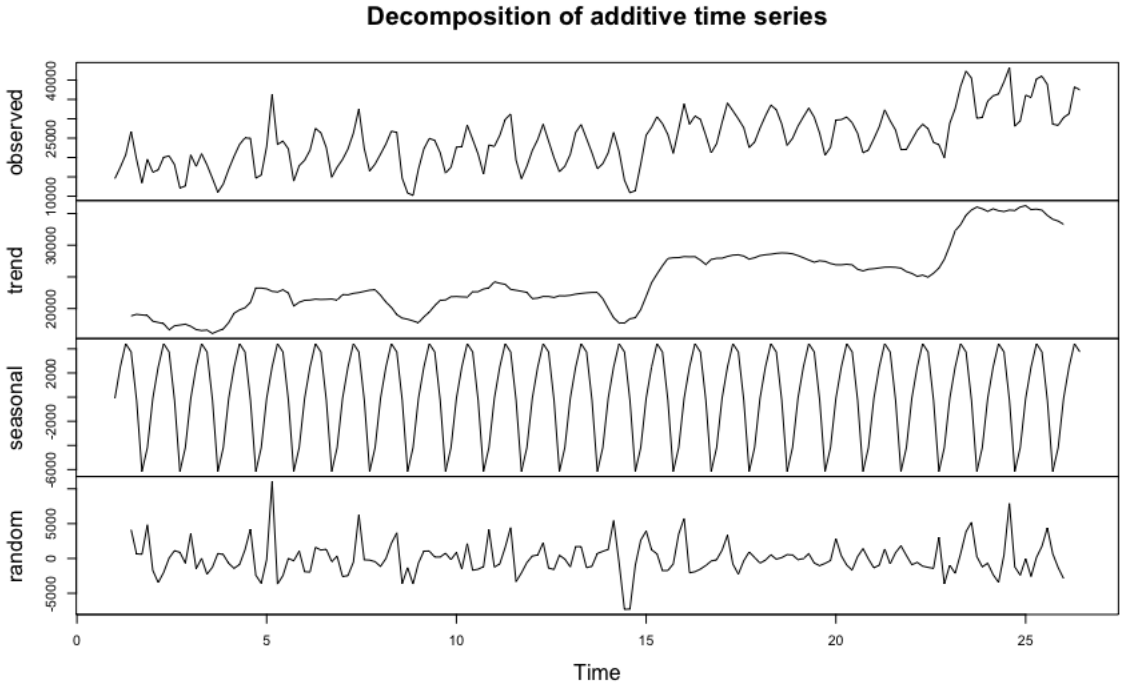
Thus, the specified model above is fit and the predictions are madeand the AIC value is 3916.13, which is checked to be the lowst among a few other possible parameter options. To compare with the true data we have in 2015, we made predictions 275 days into the future. We can see that we don't have a good prediction as the model just tried to repreat the observed trend+seasonal pattern from the train data. Although it catches the overall declining trend but it failed to predict the correct "relatively stationary trend" in 2015.



See section 4.2 for modelling the same data by accounting for other time series data.

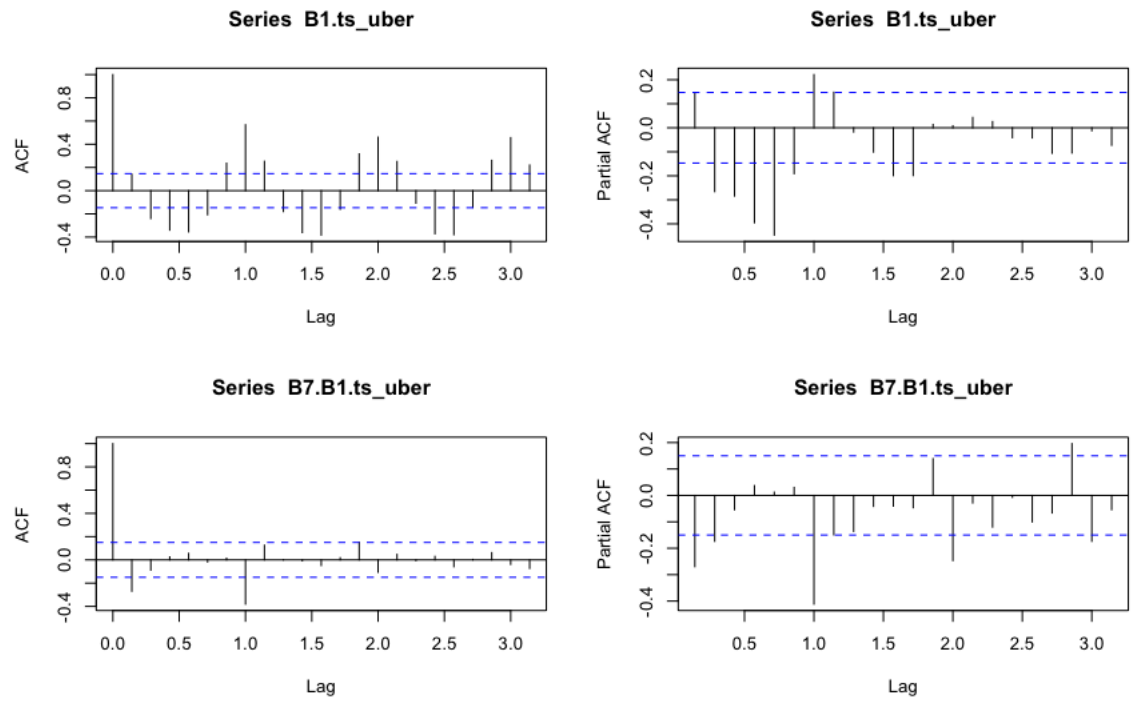
4.1.2 Uber ARIMA

Exact same analysis procesure as the previous analysis. From the decomposition, we can see that opposite from yellow taxi data, Uber traffic has an increasing trend and clear seasonality pattern.

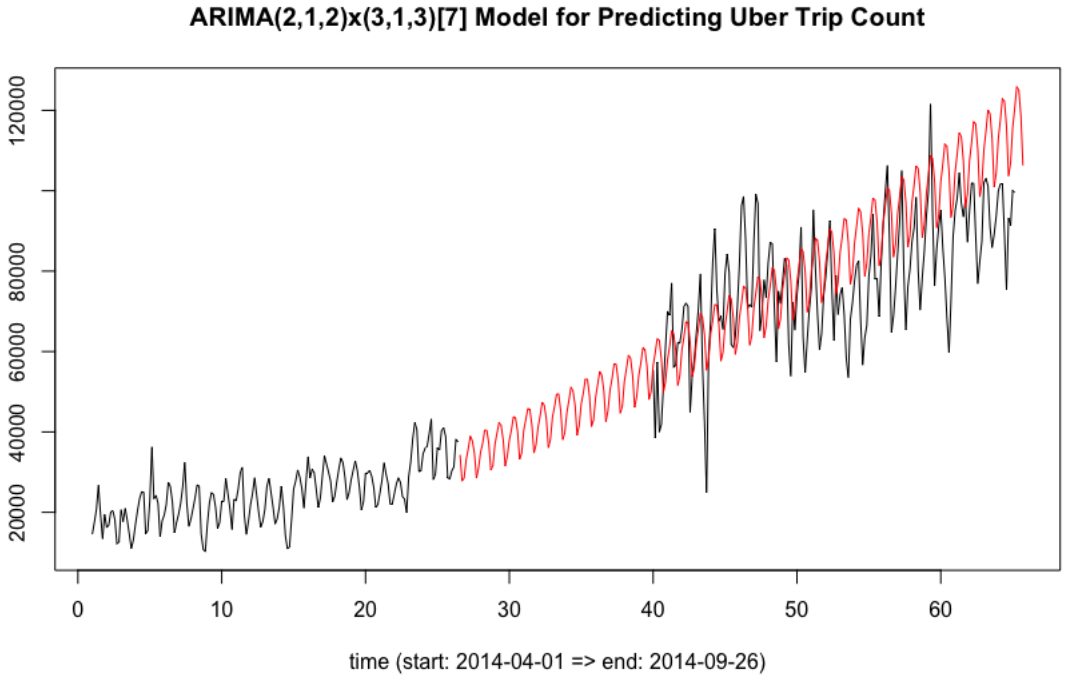


After back difference at lag 1 and lag 7, we have the following ACF and PACF plots. Again, we visual

inspect and determined the following parameters for ARIMA(2,1,2)x(3,1,3)[7].



After the prediction using the fitted model, we made a strong assumption here that is different from the yellow taxi data. **Since Uber started in NYC in 2011, and it is still at the relative early stage, we assumed the increase in uber drivers to double within prediction period. We don't have this assumption in yellow taxi data because they have been existing for a long time and by considering the developing strategy, they are not likely to have significant capacity change in the future.** Thus yielded the following prediction and it seems to match the real data quite well.



4.2 Linear Regression to Study the Impact

In reality, it is unlikely that taxi, mta and uber traffic data are independent to each other but rather in a competing situation. This section uses linear regression to model yellow taxi traffic data using date related data as well as the traffic of other three (green taxi, uber and mta) time series data we have.

Different from the ARIMA model case, since for linear regression, it does not make sense to build model using ONLY 2014 data and the predict on 2015 data, we decided to borrow first 3 months of 2015 data to build the model. **This will make the model more consistent**.

We also decomposed "2014-04-01" to: year=2014, month=4, day=1, weekday=remainder(day/7) to generate time related variables. The following is a snapshot of the dataframe we use.

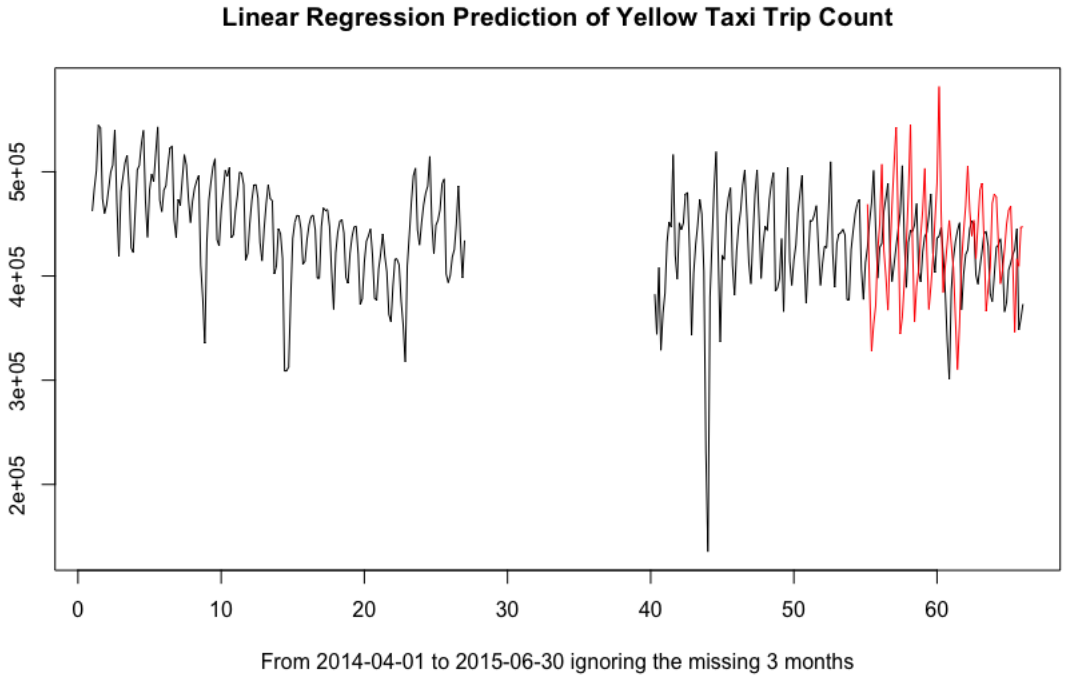
	year	month	day	weekday	yellow	green	uber	mta
1	2014	4	1	1	462220	36512	14546	2585278
2	2014	4	2	2	483976	37709	17474	2446197
3	2014	4	3	3	500892	41881	20701	4819725
4	2014	4	4	4	544712	51783	26714	5195297
5	2014	4	5	5	541976	57164	19521	5244804
6	2014	4	6	6	474152	47171	13445	5312009
7	2014	4	7	0	460104	39271	19550	5278349
8	2014	4	8	1	468400	35554	16188	3354427
9	2014	4	9	2	483812	38788	16843	2700006
10	2014	4	10	3	499592	41799	20041	4935410
11	2014	4	11	4	506424	48702	20420	5104670

Linear regression is fit on Y=yellow and X=all other variables. The intercept is excluded. The model output is as follows:

```
Call:
lm(formula = yellow ~ . - 1, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-105565  -13453    5780   18124   55972

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
year      -2.848e+05  1.334e+04 -21.342  < 2e-16 ***
month     -2.116e+04  1.289e+03 -16.413  < 2e-16 ***
day       -8.790e+02  1.908e+02  -4.607 6.33e-06 ***
weekday0   5.739e+08  2.688e+07  21.354  < 2e-16 ***
weekday1   5.739e+08  2.688e+07  21.354  < 2e-16 ***
weekday2   5.739e+08  2.688e+07  21.354  < 2e-16 ***
weekday3   5.739e+08  2.688e+07  21.354  < 2e-16 ***
weekday4   5.739e+08  2.688e+07  21.354  < 2e-16 ***
weekday5   5.739e+08  2.688e+07  21.354  < 2e-16 ***
weekday6   5.739e+08  2.688e+07  21.354  < 2e-16 ***
green      1.534e+00  2.352e-01   6.523 3.46e-10 ***
uber       3.466e+00  2.262e-01  15.318  < 2e-16 ***
mta        2.182e-03  1.623e-03   1.345     0.18
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1          1

Residual standard error: 27120 on 266 degrees of freedom
Multiple R-squared:  0.9965,  Adjusted R-squared:  0.9963
F-statistic:  5772 on 13 and 266 DF,  p-value: < 2.2e-16
```

We can see that all variables are significant except for mta. It is interesting to see that uber and green taxi have positive effects on the yellow taxi traffic (which might be the opposite of what we expect). **This might be due to the situation that when there is an increase in the demand, both the them tend to increase. It is not likely to study the effect of Uber on yellow taxi using the given 2014-2015 data since it has been 3-4 years since the launch of Uber at NYC and it is likely that the traffic has reached a reasonably stationary stage. To see the impact, we can use the yellow taxi data from 2010-2012 which is from 1 year before Uber launch to 1 year after Uber launch. Although data can be found on line but due to the limited time and large data size, we were not able to study that.**

The prediction then is made on the remaining 2015 data (after borrowing the first 3 months to train) and we have the following plot.



The 95% prediction interval can also be obtained but is rather messy to be included in this prediction so we ignored.

We can see that it predicts the actual trend really well but with a slightly larger fluctuations. It can be shown that the traffic of yellow taxi have close relation with other traffic modes. For later work, we might be able to include other demographic variables in the model to make the prediction more reasonable and powerful.