Project Report

Forecasting demand for pickups per hour in New York City for Uber

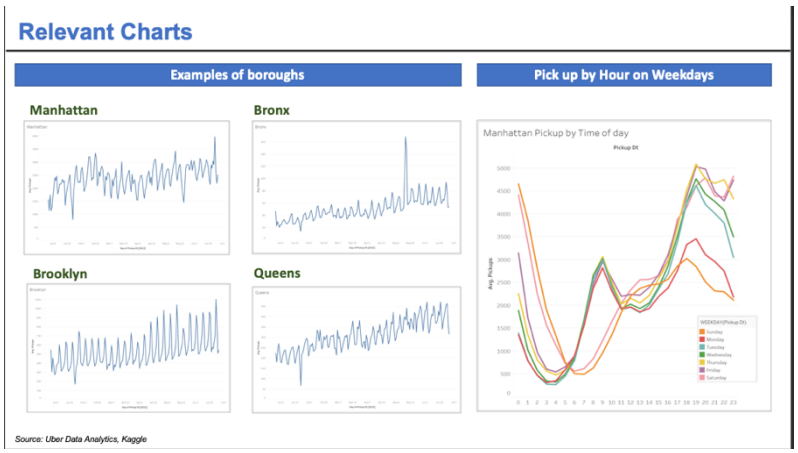
Team: A-04
Aniket Jain (61910075),
Rachit Nagalia (61910325),
Nakul Singhal (61910179)
Ayush Anand (61910393),
Priyakansha Paul (61910800),
Prakhar Megotia (61910054)

Executive summary

Uber is a ride hailing company which was founded in San Francisco, California. Since its inception, it has expanded into multiple other businesses like ride sharing, food delivery etc and according to estimates, Uber has close to 100 million customers and operations in as many as 800 metropolitan areas. Given its ever expanding scale, Uber continuously manages the gap between supply and demand through surge pricing, incentivising drivers and charging riders. There has, however, been a lot of backlash as surge pricing has gone above 20x at times. This project focuses only on city of New York and its six boroughs. More specifically, we intend to solve below problems for Uber:

- Manage demand by optimizing driver location across six boroughs which is expected to result in increased driver pick up efficiency and ultimately guest experience.
- Accurate demand forecasts combined with driver supply data, which Uber has all the time, can be used to get a better idea of surge pricing.

Data for this project was pulled from Kaggle for the timeline 01/01/2015 to 06/30/2015. Data contained hourly pick up data for New York city and its six boroughs namely Newark, Manhattan, Bronx, Queens, Staten Island and Brooklyn. Data for Newark was limited so we combined it with data which had NA as its borough. We found out that there is seasonality within a week also within the day itself, the below charts show pick up patterns in boroughs and pickup by time of the day.



With an intention to forecast for 2 weeks, we partitioned our data with a validation period of 2 weeks. Given the nature of the data we decided to focus on linear regression as our forecasting method and benchmarked the output against Seasonal Naive forecasts. We found that for some boroughs, Naive performed than linear regression. So, our final forecasts have been arrived at using both Naive and linear regression. The table highlights performance metrics for different boroughs.

Borough	Naïve model	Regression (with dummies)	with external variables	with lag
Manhattan	0.0989	0.3333	0.3112	0.0887
Newark	0.8060 (MAAPE)	0.8959 (MAAPE)	-	No autocorrelation
Bronx	0.3454	0.2955	0.2891	0.2256
Brooklyn	0.1051	0.2083	0.2099	0.0968
Queens	0.1434	0.2523	0.3040	0.1380
Staten Island	0.6033 (MAAPE)	0.4162 (MAAPE)	-	No autocorrelation

Technical Summary

Staten Island

Data for Staten Is. contained a trend with multiplicative seasonality. However, lots of pickups were zero and on average the pick-up volume was low. This presented us with a challenge of measuring the error in our forecast. To overcome this issue, we decided to use MAAPE instead of MAPE to avoid errors arising due to division by zero. To come up with final model, below methods were deployed:

- Model 1 (Naïve forecast): A seasonal naïve forecast for validation was developed using data a week prior to forecast periods. We obtained a MAAPE, for Naïve model, of ~0.6 radians on the validation period.
- Model 2 (Linear regression): To improve upon Naïve model, we used linear regression. Since seasonality was found to be of multiplicative kind, output variable was log(pickup') where pickup'=pickup+1 to avoid errors due to zero values in log. Predictor variables were trend, holiday dummy, day of week & hour of the day. Using forecast generated by this model, MAAPE came down to ~0.4 radians, (0.76 on the training data).

Including furthermore variables like cold months (using rain, snow depth) didn't improve MAAPE by much so we went with Model 2 for Staten Island. Lag analysis was also done, but no significant correlation was found so the model wasn't adjusted for any lag. The table compares MAAPE for both the models.

Model	MAAPE
Naïve	0.60329
Linear Regression	0.416214

Bronx

The data for Bronx contained a trend with multiplicative seasonality. The pickup volume was generally low, but there were very few 0 values. Given multiple levels of seasonality, smoothing methods were not viable.

Methods used:

- Seasonal Naïve Forecast for 1 week
- Linear regression with dummies for Day of the week, Time of the day and Holiday
- Lag analysis on the errors to check for autocorrelations. Lag of 1 period was observed.
- External variables: Ran regression with external variables including snow, precipitation, wind speed. The result wasn't significant and thus not included.

Eventually, we chose the lag analysis model for forecasts as the MAPE was the lowest. It should however be pointed out that using a lag analysis would limit the

Model	MAPE
Seasonal Naïve	0.3454
Regression with lag Analysis	0.2256

quality of the forecast in the future, therefore it is not probably ideal. However, we believe that using the forecasts as a proxy for actual pickups can work for short horizons.

Brooklyn

Brooklyn aggregate daily data had weekly multiplicative seasonality and the trend was upward linear. There was slight noise however overall the pattern was easily identifiable. Other details of the data were same as mentioned above (i.e., weather factors and few pickups with zero value). Also, it is important to note that the data had hourly seasonality, hence the methods such moving average, Holts winter etc were not applied. Following process was adopted to forecast the data.

- The regression was carried out including external factor and separately without external factors (weather parameters). We found that there was hardly any change in MAPE.
- We went with the regression with dummy variables and considered variables – trend, weekday, hour of the day, holiday to predict pick-ups.
- We also conducted lag analysis, on the residuals and found that there is lag 1, 2and 3. Post this we forecasted error with the above lag. Despite MAPE of the validation set being reduced to ~9.68% the forecasted values seemed absurd.
- Finally, we did the seasonal Naive forecast (one week) and found out that MAPE was 10.5% which was marginally higher than the lag analysis and 50% better than regression with dummies and external variable. The table shows the errors for different method.

Naïve model	Regression (with dummies)	with external variables	with lag
0.1051	0.2083	0.2099	0.0968

Newark

Newark data had at least one-third of 0 pickups and majority of the remaining pickups were either limited. The demand was low because in 2015, Uber was banned[1] from picking up passengers from EWR airport[2].

Methods used:

- Seasonal Naive forecast - Using data for the prior week to forecast. Also, given the large number of 0 pickups, we used MAAPE instead of MAPE to calculate the error in the validation period which turned out to be 0.8060 radians.

[1] https://www.nj.com/essex/2016/04/newark_releases_new_details_of_tentative_agreeemen.html
[2] https://www.nj.com/union/index.ssf/2016/05/uber_ordinance_raised_again_at_elizabeth_city_coun.html

- Linear Regression - We tried using a linear regression with dummy variables but the MAAPE for Regression model was worse than the Naïve. The time in a day or day in a week or any weather variables had no impact on pickups from Newark.

We also performed lag analysis but did not find any autocorrelation with the previous lags. The table here shows the computer errors.

Model	MAAPE
Naive forecast	0.8060
Linear Regression	0.8959

Queens

The Queens time series exhibited seasonality with trend. So, the below methods were deployed:

- Model 1 (Naïve forecast): Naïve forecast for validation and forecasting horizon were done by using data a week prior to forecast periods. The data for queens was good and had only a few values that were 0, hence MAPE was used as a measure and was around 0.1434.
- Model 2 (Linear regression): To improve upon Naïve model, we used linear regression. Since it was a normal seasonality, pickups were used as output variable and time of day and day of week were mainly used as dummy predictors to capture seasonality. Here the MAPE was obtained to be around 0.2523.
- Model 3 (Linear regression including Lag analysis): We conducted a lag analysis to detect any autocorrelation in the residuals. It was found out that Lag 1 had a correlation. The model was then plotted including the lag1 residual component and forecast was made, the MAPE obtained was 0.138.

The table compares MAPE for the three models:

Model	MAPE
Naïve	0.1434
Linear Regression	0.2523
Linear Reg with Lag analysis	0.1380

Manhattan

Manhattan had by far the largest number of pick-ups and the data exhibited intraday as well as intraweek seasonality with trend. The complexity was assumed to be beyond smoothening and linear regression models were built:

- Model 1 (Naïve forecast): Naïve forecast was used to benchmark performance of all other models, and a validation MAPE of 0.0989 was observed.
- Model 2 (Linear regression): The linear regression model with dummies to capture both levels of seasonality was used. Given the possible decline in ridership in Feb (possibly due to snow), a dummy variable for Feb was used as well. The MAPE of the model was 0.3333.
- Model 3 (Linear regression including Lag analysis): We conducted a lag analysis to detect any autocorrelation in the residuals. It was found out that Lag1 and Lag2 were correlated. Thus, the whole model was re-run using the aforementioned variables as well as the 2

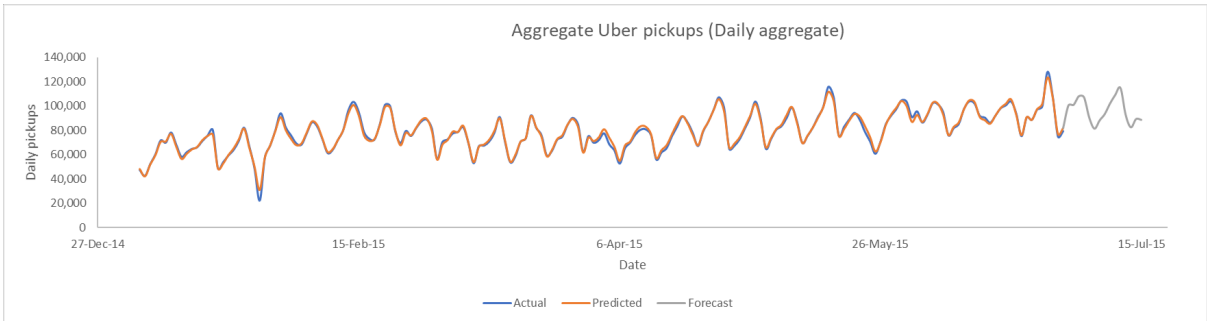
lagged variables. The MAPE was found to be 0.0887. While this model was a very slight improvement on the Naïve, it was found preferable, since the forecast would be more accurate for shorter forecast horizons.

Naïve model	Regression (with dummies)	with lag
0.0968	0.3333	0.0887

The table compares MAPE for the three models:

Conclusion and Recommendations

The Naïve forecast proved tough to beat. On the whole, our aggregated model had a MAPE of 8.6% which we felt was a good outcome, given the complexity of the data.



What started as an exercise to forecast for demand for 2 weeks, seems an issue given the auto-correlation in the data. We therefore can provide a forecast for the 2 weeks but would only be comfortable with the accuracy of a shorter horizon (maybe 24 hours).

While this may not be ideal, we feel Uber would still benefit from giving drivers even a minor heads up on where to expect surge pricing. Using this model in roll forward mode to give predictions for the next 24 hours, every hour is something that Uber can easily automate, sending updates to drivers on the potential demand surges.

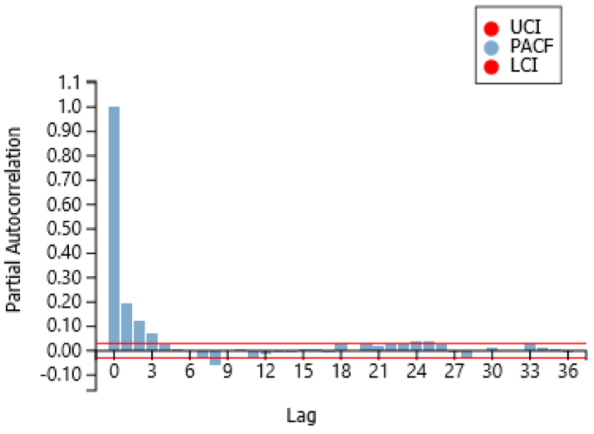
Possible Reservations

There are other issues that may need to be addressed. The pickup data is only a proxy for the demand, and as is likely under times of high surge pricing, the demand gets artificially suppressed. With a solution that might reduce the extent of surge pricing, the demand may no longer be stationary, requiring calibration of this model every 2-3 weeks. Again, this is something which Uber can easily automate.

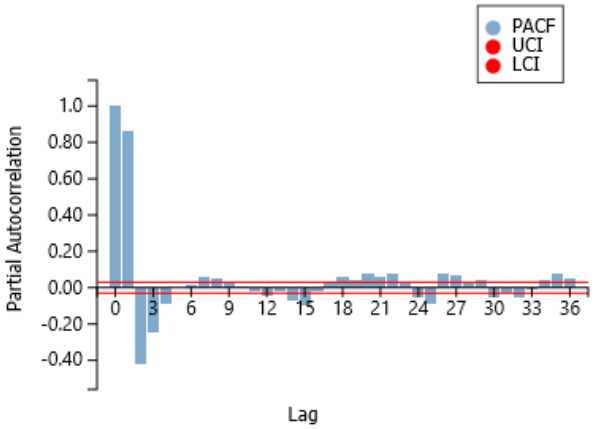
Appendix

Autocorrelation plots (PACF):

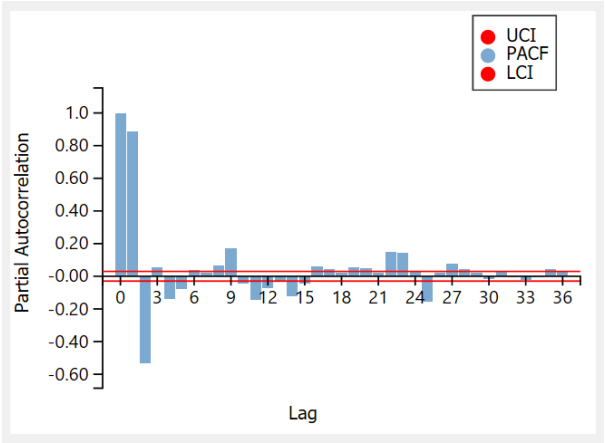
Staten



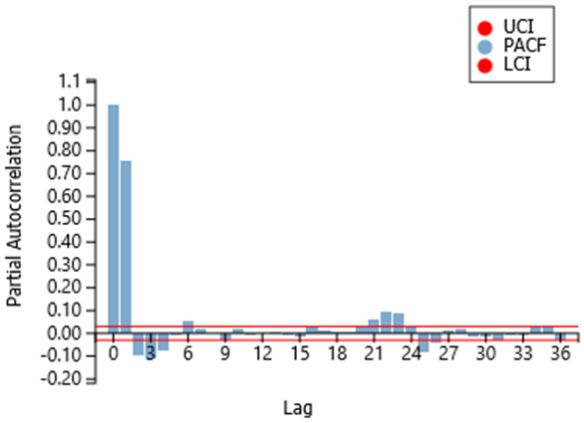
Brooklyn



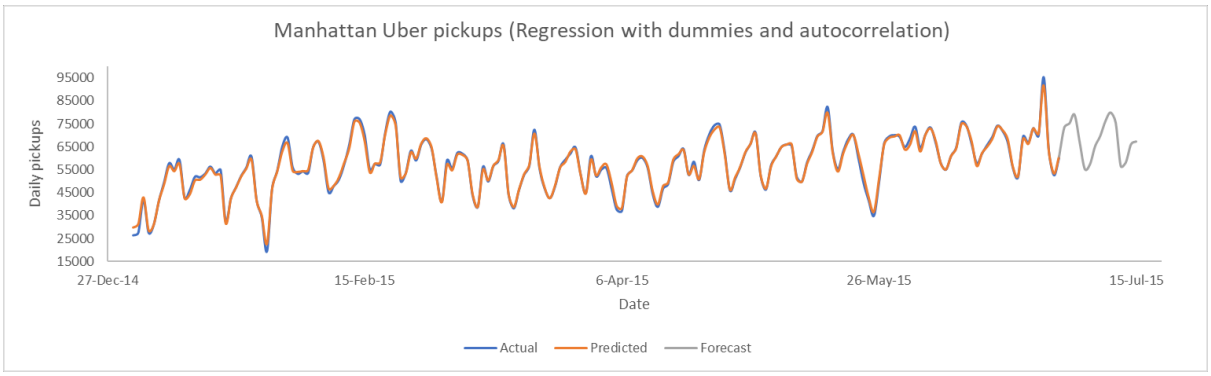
Manhattan

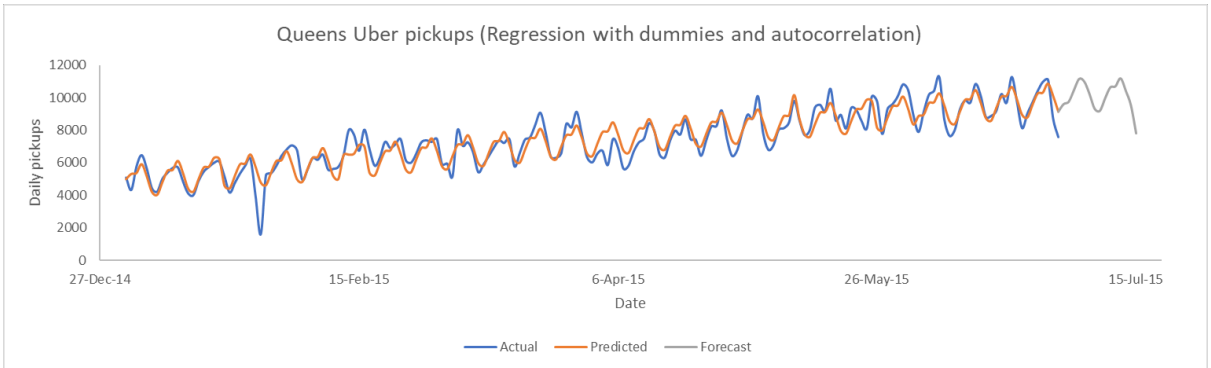


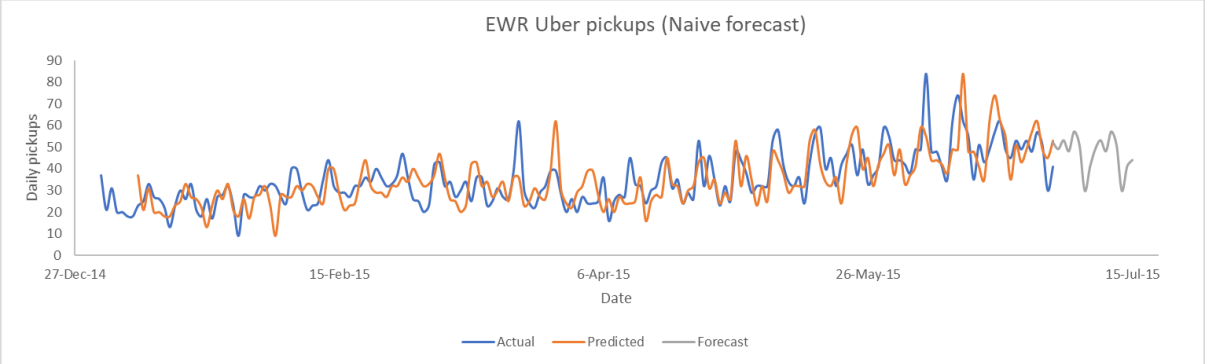
Queens

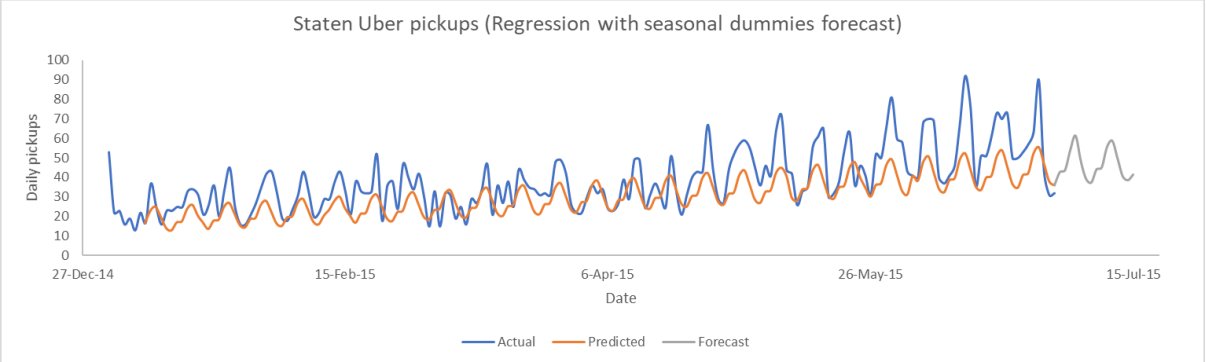


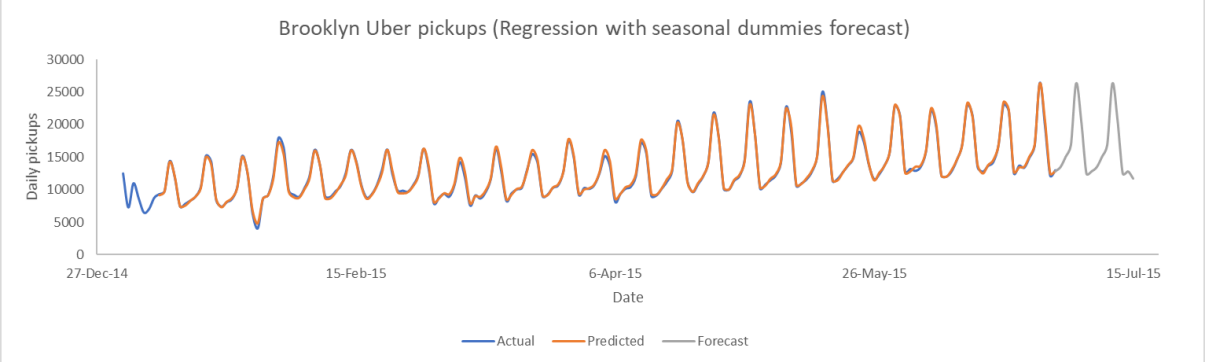
Individual actual vs prediction vs forecast plots:

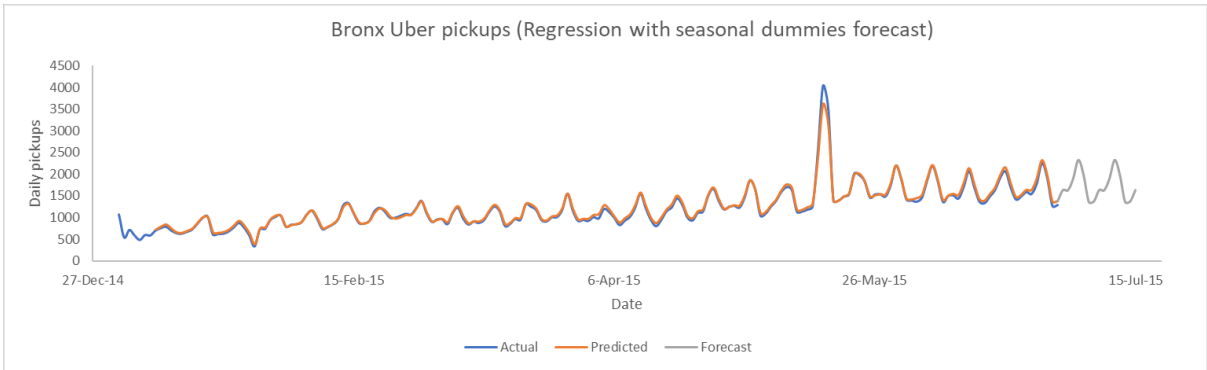



Queens Uber pickups (Regression with dummies and autocorrelation)
Daily pickups
12000
10000
8000
6000
4000
2000
0
27-Dec-14
15-Feb-15
6-Apr-15
26-May-15
15-Jul-15
Date
Actual
Predicted
Forecast


EWR Uber pickups (Naive forecast)
Daily pickups
90
80
70
60
50
40
30
20
10
0
27-Dec-14
15-Feb-15
6-Apr-15
26-May-15
15-Jul-15
Date
Actual
Predicted
Forecast


Staten Uber pickups (Regression with seasonal dummies forecast)
Daily pickups
100
90
80
70
60
50
40
30
20
10
0
27-Dec-14
15-Feb-15
6-Apr-15
26-May-15
15-Jul-15
Date
Actual
Predicted
Forecast


Brooklyn Uber pickups (Regression with seasonal dummies forecast)
Daily pickups
30000
25000
20000
15000
10000
5000
0
27-Dec-14
15-Feb-15
6-Apr-15
26-May-15
15-Jul-15
Date
Actual
Predicted
Forecast


Bronx Uber pickups (Regression with seasonal dummies forecast)
Daily pickups
4500
4000
3500
3000
2500
2000
1500
1000
500
0
27-Dec-14
15-Feb-15
6-Apr-15
26-May-15
15-Jul-15
Date
Actual
Predicted
Forecast