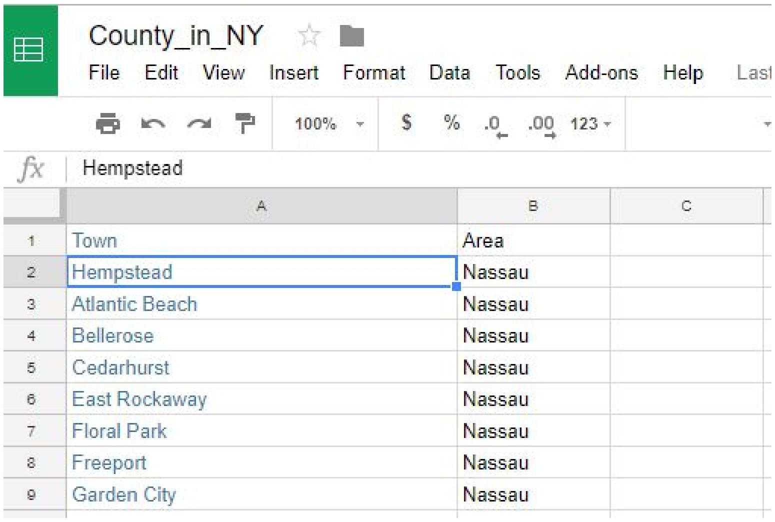
Data Story

Uber is used quite frequently in NYC, however it does not seem to be popular in all parts of the city. Instead many non-Uber cab companies are preferred. The data for non-Uber companies - American, Dial7 and Carmel has been used for this analysis over a time period from July 2014 - September 2014. The Uber data-set used also is from July 2014 - September 2014. This analysis can help Uber to determine which parts of the city and state have a high demand for cabs which is being serviced currently by non-Uber companies. Also, timing wise Uber can understand time slots when Uber services have slowed down and other cab companies are providing more service.

Cleaning the Non-Uber data set -

There were three non-Uber companies - American, Dial7 and Carmel which had the pick-up times and addresses for a time period of July'14 to September'14. Based on the addresses given for each pick-up an Area field was derived. The Area field was calculated based on simple replace functions for the major areas known in New York City. For eg. If the addresses contained either Brooklyn, Bronx, Manhattan, Queens or Staten Island then the Area field would be replaced by these regions. However with the more widespread areas across the NY State a mapping spreadsheet was made which consisted of two columns - Town and Area. Every Town was categorized into a county. For eg. a town - Hempstead belongs to the Nassau county. If Hempstead is found in the pick-up Address then the Area would be Nassau county. Since the number of rows which had missing addresses was less than 5% in the overall dataset those rows were ignored. The mapping spreadsheet was created based off the towns listed in each of the county websites.

Here's a screenshot of the spreadsheet with the mappings -

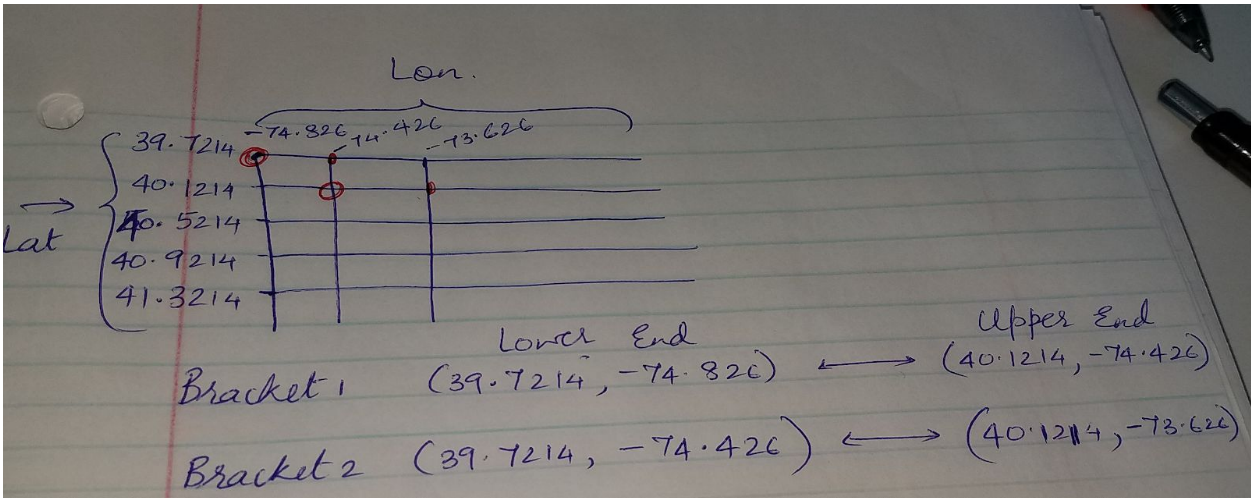


	A	B	C
1	Town	Area	
2	Hempstead	Nassau	
3	Atlantic Beach	Nassau	
4	Bellerose	Nassau	
5	Cedarhurst	Nassau	
6	East Rockaway	Nassau	
7	Floral Park	Nassau	
8	Freeport	Nassau	
9	Garden City	Nassau	

This mapping spreadsheet was applied against the datasets and the Area fields were derived based on these mappings.

Cleaning Uber Data sets -
The Uber data sets provided were also over a time period of July'14 to September'14. The pick-up address in these datasets was presented in latitude longitude format. In order to derive the area associate with each location, the location points had to be converted to addresses. A program was written to reverse geocode each pickup point. The program can look-up every pick-up point using the revgeocode function which uses the Google API to return the addresses. The issue with this approach was that Google API has a daily limit of 2500 lookups and if a paid version of this API is used then the program can look-up all addresses presented in the data set. In order to get around this problem, an approximation approach was used. The minimum and maximum longitude and latitude points were calculated from the dataset. Then based on these small brackets were created with small increments to the latitude and longitude values. So every bracket would have a min Lat, Lon pair and a max Lat,Lon pair. The idea was that if any pick-up point lies between these brackets then it would be assigned a particular area which was designated to each bracket.

Here's a screenshot of the way the brackets were created -



Bracket 1 (39.7214, -74.826) ⟷ (40.1214, -74.426)

Bracket 2 (39.7214, -74.426) ⟷ (40.1214, -73.626)

The brackets were created as such that the entire area between the latitude and longitude values provided in the dataset would be covered. Once the brackets were created then Areas were assigned based on the lookups for the min end of the bracket and the max end of the bracket. These look-ups were done from an external website which would take an input of lat lon combinations and return an address. If a pick-up point falls below the mid-point of the bracket then the Area would be the Area associated with the lower end of the bracket and if the lookup point is above the midpoint of the bracket then the Area associated with the higher end of the bracket would be assigned to the pick-up point.

Data Analysis

Once the datasets were cleaned up there 3 graphs were plotted for the Uber and non-Uber datasets.

1. Bar plot by Area
2. Bar plot by individual day
3. Bar plot by the weekdays

When the non-Uber datasets were plotted, the general consensus was that these cab companies were in lot more demand in the NYC as compared to the counties in NY state. Timing wise across the three cab companies, they seem to be used at the peak on weekends - Friday, Saturday and Sunday.

Cab Company	**Time Frame**	**Maximum Demand in Area**	**Medium to Low demand in Areas**	**Maximum demand on weekdays**	**Medium to Lowest demand on Weekdays**
American	July’14 - Sep'14	Bronx,Orange, Suffolk	Manhattan, EWR, Pennsylvania	Saturday, Sunday, Friday in this order	Monday, Tuesday, Wednesday Thursday can be comparable
Dial 7	July’14 - Sep'14	Manhattan, JFK Airport, LaGuardia Airport, EWR Airport	New Jersey, Queens, Nassau, Long Island(Lowest)	Thursday, Friday	Monday, Tuesday, Wednesday Lowest on Saturday and Sunday
Carmel	July’14 - Sep'14	Connecticut, New Jersey, JFK Airport, Laguardia, EWR	Manhattan,Bronx, Queens, Columbia(Lowest)	Thursday,Tuesday,	Monday, Wednesday Lowest on Saturday and Sunday

Uber	July’14	New Jersey, Brooklyn, Nassau, Suffolk, Westchester	Orange(Lowest)	Tuesday, Wednesday, Thursday, Friday	Monday, Saturday Lowest on Sunday
Uber	Aug’14	Bronx, Queens, Richmond	Nassau, Sussex(Lowest)	Friday, Saturday	Tuesday, Monday(Lowest)
Uber	Sep’14	Bronx, Kings, Richmond	Suffolk, Monmouth, Fairfield(Lowest)	Tuesday, Saturday	Wednesday, Sunday(Lowest)