Location based classification of Uber pickups in New York city from July to September 2014 on the basis of hour of pickup

Problem Definition and Solution Design

This bar graph provides the number of Uber rides taken in New York over that have been classified based on the hour of the ride. It considers two time periods – April to September 2014 and January to June 2015. The data has been obtained from FiveThirtyEight , an opinion poll analysis website owned by ESPN Inc. The graph does not consider the latitude and longitude values present in the data. The purpose of this exercise is to redesign the existing graph to include geolocation data and draw more useful conclusions from the existing data.

About the dataset and the existing report

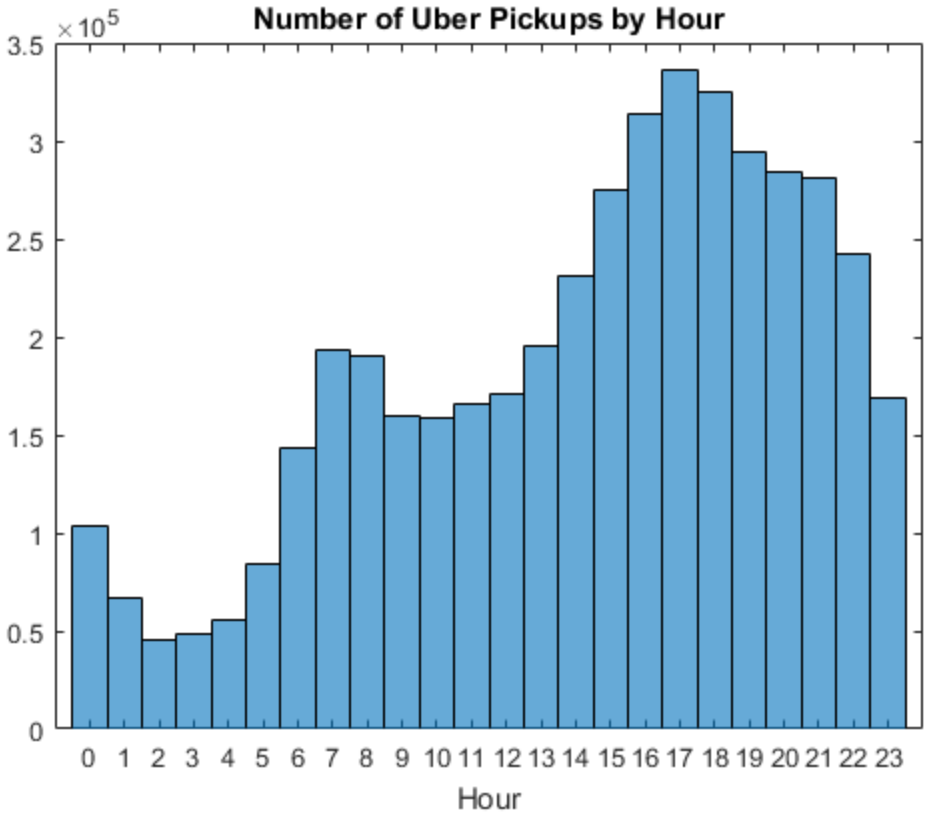
This dataset provides us with a list of Uber pickups in New York city; more specifically, the timestamp, latitude , longitude and the Taxi and Limousine commission (TLC) base code affiliated with the Uber pickup. The data is owned by FiveThirtyEight and has been obtained from their Github profile. The graph has been obtained from mathworks.com.

Here is a sample of the dataset :

```
"Date/Time","Lat","Lon","Base"
"4/1/2014 0:11:00",40.769,-73.9549,"B02512"
"4/1/2014 0:17:00",40.7267,-74.0345,"B02512"
"4/1/2014 0:21:00",40.7316,-73.9873,"B02512"
"4/1/2014 0:28:00",40.7588,-73.9776,"B02512"
"4/1/2014 0:33:00",40.7594,-73.9722,"B02512"
"4/1/2014 0:33:00",40.7383,-74.0403,"B02512"
"4/1/2014 0:39:00",40.7223,-73.9887,"B02512"
"4/1/2014 0:45:00",40.762,-73.979,"B02512"
"4/1/2014 0:55:00",40.7524,-73.996,"B02512"
"4/1/2014 1:01:00",40.7575,-73.9846,"B02512"
"4/1/2014 1:19:00",40.7256,-73.9869,"B02512"
"4/1/2014 1:48:00",40.7591,-73.9684,"B02512"
"4/1/2014 1:49:00",40.7271,-73.9803,"B02512"
"4/1/2014 2:11:00",40.6463,-73.7896,"B02512"
```

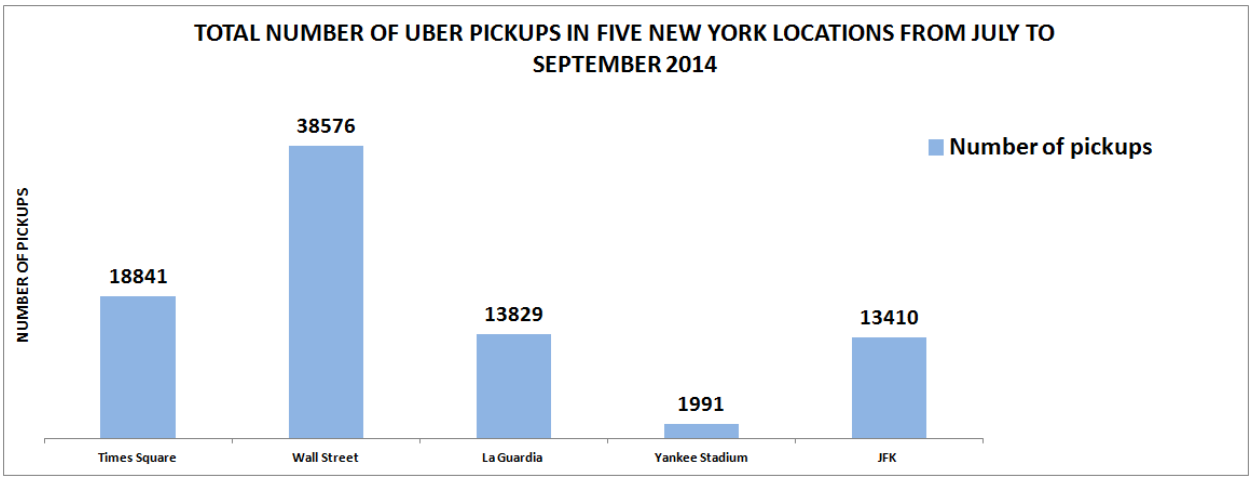
Disadvantages of the existing graph

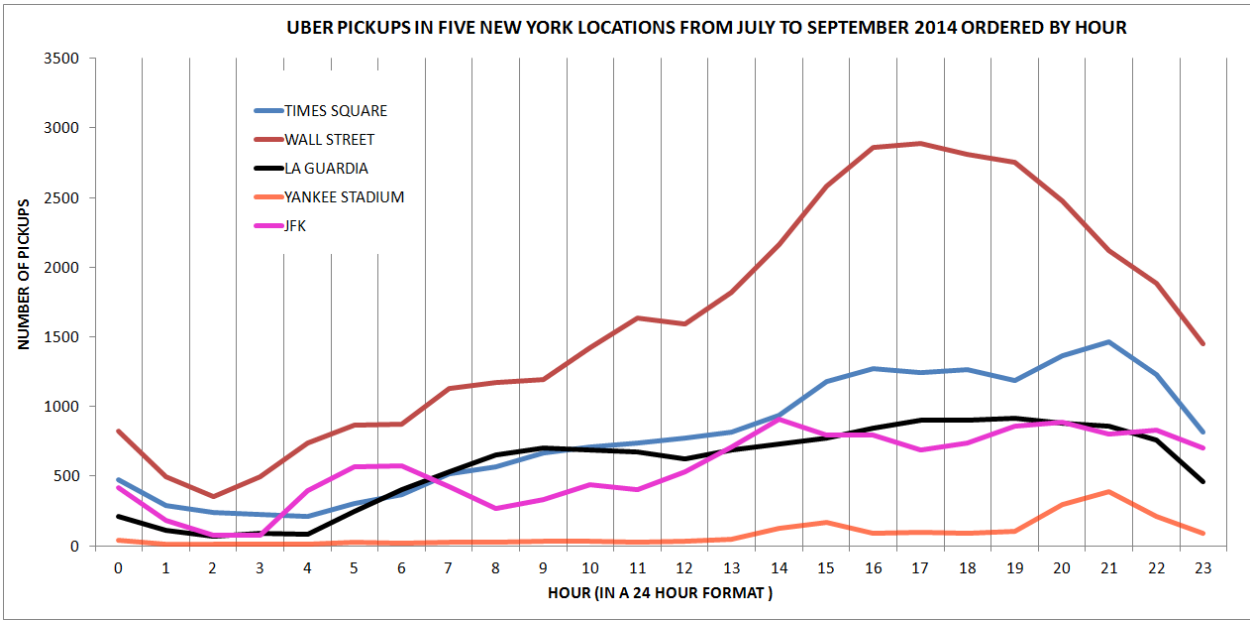
The existing graph does not take into consideration the latitude and longitude values that are present in the data, nor the TLC base code affiliation. It simply counts the number of Uber rides taken at each hour and maps them in a line graph format. Not a lot of insight can be gathered from the report, because the location data is not being used.



Redesigned graphs

I downloaded the data from FiveThirtyEight's Github profile for a three month period – July to September 2014. The plots were made using Microsoft Excel and Tableau. While it was possible to map every single ride ever taken in New York City for this time period, I focused on five specific locations to understand more about them – JFK Airport, LaGuardia Airport, Tiems Square, Wall Street and the Yankee Stadium. I was able to graph the number of rides taken in total in this specific time period for these locations. Additionally, I was also able to graph the manner in which the rides were divided based on the hour of the ride.





Some of the conclusions which can be drawn from the above two graphs :

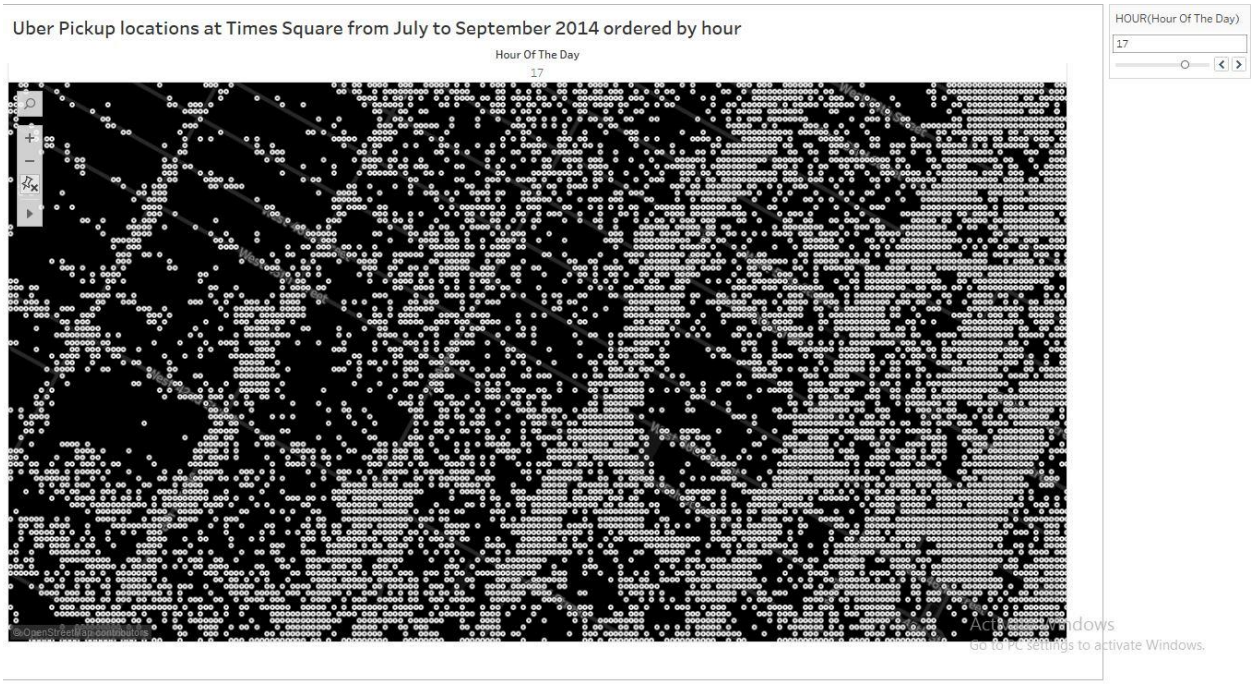
> Wall Street has the highest number of pickups, and the pickups peak between 4 and 5 PM.

> Yankee Stadium has the least number of pickups, and they occur during two specific time intervals – 1 PM to 4 PM and 7 PM to 11 PM.

> LaGuardia and JFK Airport follow very different pickup schedules, with the difference in the number of rides by hour significantly varying for JFK when compared to LaGuardia, where it is much more consistent.

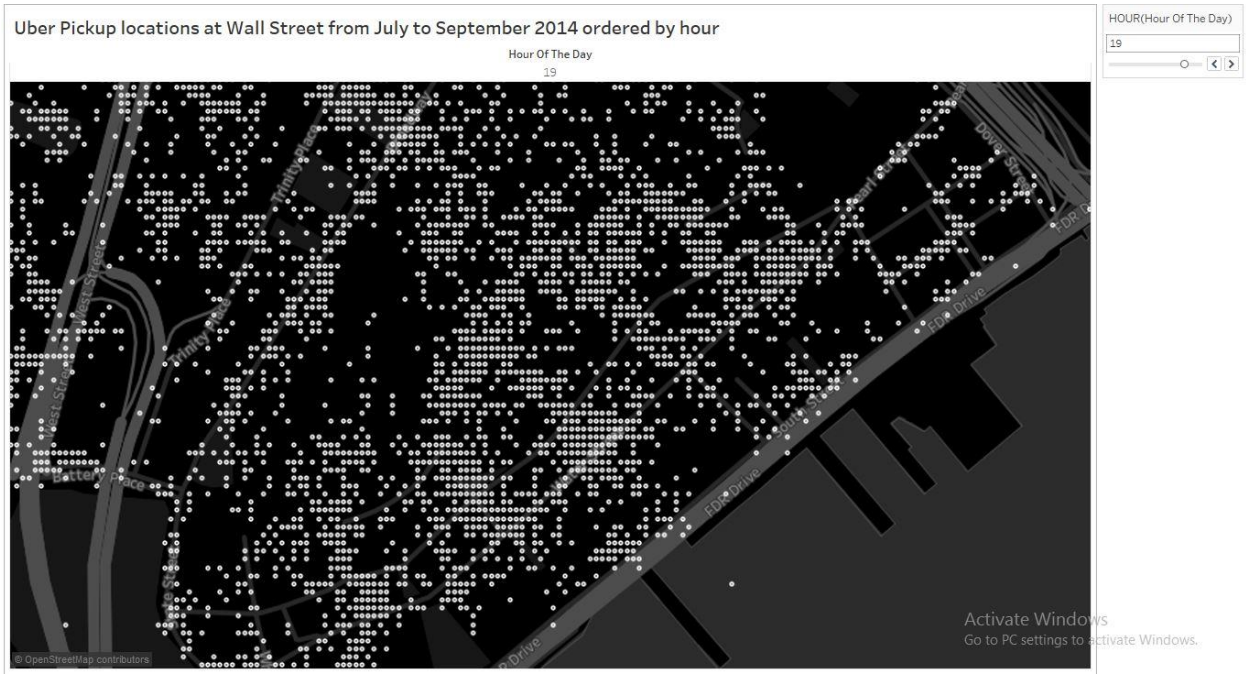
> Times Square starts gaining traction in the number of pickups at around 2 PM and it continues till around 9 PM.

The problem with these graphs is that while the viewer is able to understand how the number of rides are varying at specific locations based on the hour, it is still not possible to understand *where* the rides are being booked from. This is the reason why geolocation data has been used to produce interactive maps in Tableau where the viewer can view the locations of the rides being taken at each hour on a map, specific to each location.



>One very interesting conclusion that can be drawn from this map is that regardless of the hour , most of the pickups happen on streets which are called “avenue” (Ex : 7th Avenue , 6th Avenue etc) when compared to roads that are called “street” (Ex : West 46th Street , West 48th street etc)

> Additionally, most of the pickups do not happen *at* Times Square itself, but rather around it.



>At Wall Street, a pattern is observed which is quite distinct – the rides are not booked at the area around the New York Stock Exchange , but as one moves towards the FDR drive , the number of rides increase. Possible explanations for this include a larger presence of restaurants and bars closer to the FDR and the presence of higher traffic near the stock exchange.

> LaGuardia and JFK airport follow a very similar pattern with regard to the kind of flyers booking Uber rides. It is mostly flyers who flew Delta airlines that are using an Uber at both these airports, when compared to other flights. This information can be obtained from the fact that each major airline has its own separate entrance and exit at both these airports.

While the viewer is able to understand the time frequency for the most rides, it is still interesting to see where most of the rides are booked at the Yankee Stadium.

Conclusion

Overall, classifying the number of Uber rides by locations allows the viewer to gain more insight from the existing data. The line graphs help understand the frequency distribution for the rides at specific locations, while the maps help understand where the rides have been booked.

References

Data source : https://github.com/fivethirtyeight/uber-tlc-foil-response

Graph : http://blogs.mathworks.com/loren/2016/01/20/mapping-uber-pickups-in-new-york-city/