The Daily NYC Taxi Crisis – An Urban Legend?

INTRODUCTION

This is an analysis piece on the urban legend around taxis disappearing in the afternoon in New York City. No specific background is required other than basic high-school math and a general knowledge of statistics to understand the analysis performed. It might also help to understand the geography of New York, as Staten Island and south of it have been restricted from the analysis, as it tends to skew the results of the behavior observed within the main confines of the metropolitan area. I've chose this problem because, like many New Yorkers, have found it hard to take cabs at this critical hour, when many people leave their jobs to head home. Until recently, it's just been an urban legend that cabs disappear at this time, earning the name "ghost" taxis. The objective of this blog is to examine if this phenomenon is true.

METHODS

It is important to note, that we are using drop-off coordinates, because it seems that cabs will either allow or deny service based on where they have to be AFTER the trip. This would make sense, since cabs won't pick up passengers if they all have to be somewhere at a certain time. The code is structured to restrict certain areas by latitude, like Staten Island and even more south, so that we can focus on the more significant four other boroughs. Once that is in the RStudio disk image, you can run the main program again with full analysis. There are two classes of functions I use in this script: Importing and Analysis. In this section, I will focus on the Analysis functions used in the script.

The first method I used to get a good sense of how the cab pickups are distributed throughout the day was a simple histogram, running at half hour intervals from midnight to the next. The data used had dates in text format, so I broke them down into separate variables, like "hour" and "minute". I then transformed them into a new unit, called "minhours", which I'll cover later. In each bar of the histogram is the number of cabs that picked up passenger(s) at that half hour interval. Using the locator() tool, I could designate the beginning and end of the drop for each month, represented in the histogram. I then decided to run a regression in this "dropzone" period, when cabs begin to drop until they set back out. This is where having the minhour variable is critical. Minhour is essentially the timestamp, quantized. For example, 16:30 (or 4:30 PM) would be 16.5, 9:15 would be 9.25, and so on. This makes it easier because it made regressing very easy. As my dependent variable, I decided to use GPS coordinates, latitude and then longitude. My intuition was that the cabs are not "disappearing" completely, but are rather all moving out of view. If there was a shift change at 5:30 PM, which seems to be the case, then cabs would start to converge in a direction. By regressing latitude on minhour (in the dropzone) we can see how latitude changes as minhour increases, and the same applies with longitude. Over the same time period, regressing longitude on minhour in the same time period captures the horizontal element of shift in their movement. By combining the two, we can get a directional vector to see in which direction the cabs are moving in.

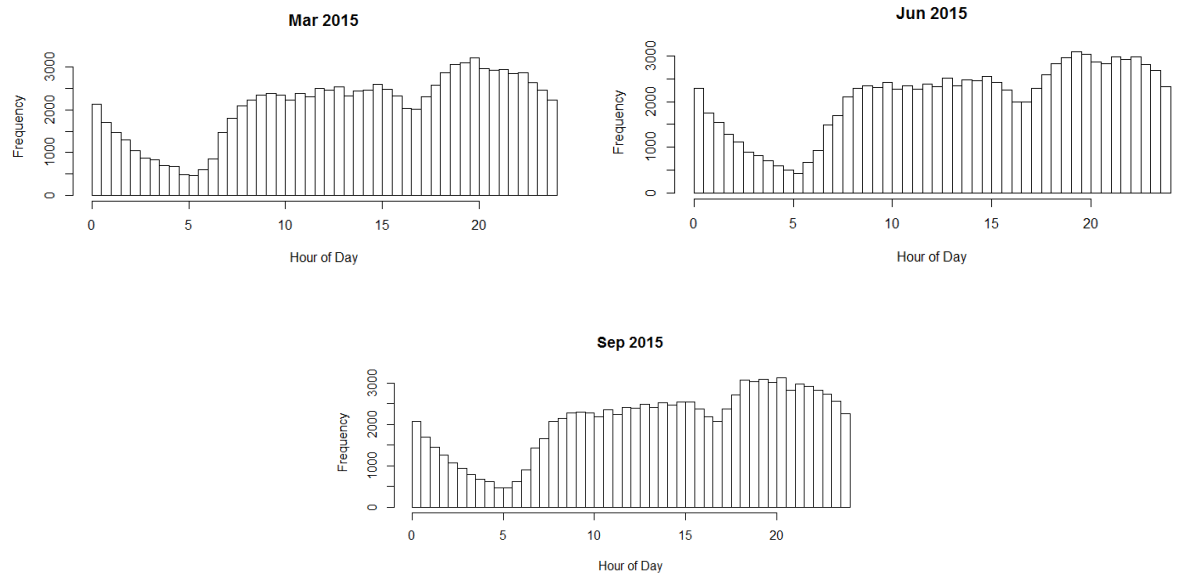
RESULTS

In conclusion, my results correspond almost precisely with the New York Times article written on this phenomenon. First, that there is in fact a drop in cab supply in the afternoon hours. Second, that the drop in cab supply is approximately 20%, give or take about 3%. For example, here is the analysis printout log of the month of June.

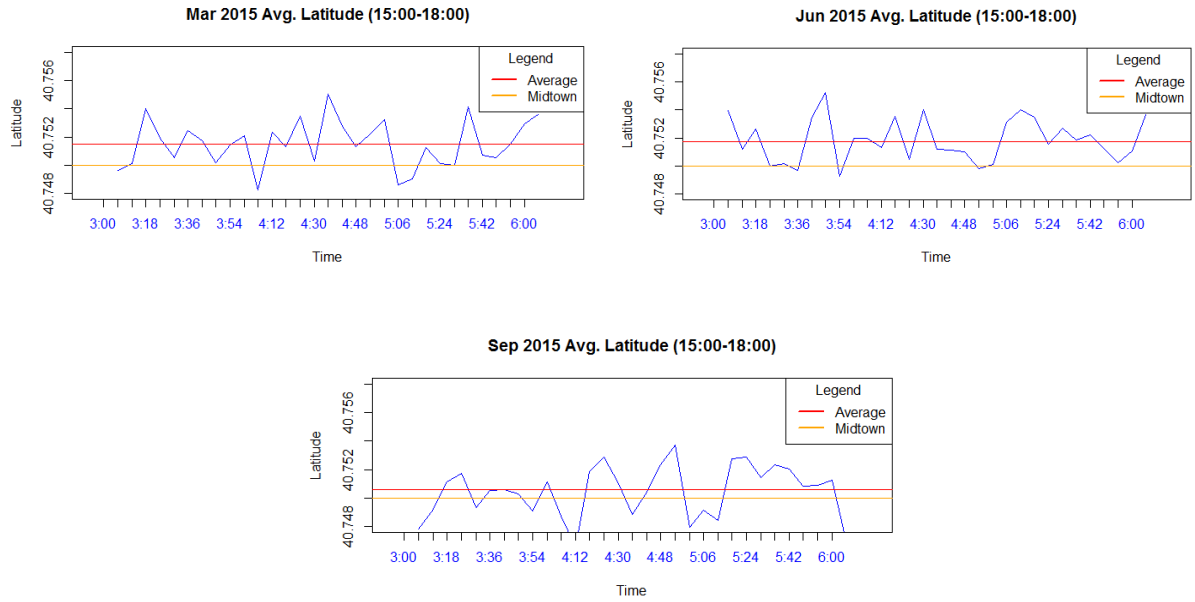
```
------------------------------------------------------------------------'
Jun 2015"
VECTOR DIR (LAT, LONG): 0.00119387094536796, 0.000108227343438344"
SLOPE: 11.0311397049868"
CAB SUPPLY PERCENT CHANGE: -21.2552255822792%"
------------------------------------------------------------------------'
```

In this print out we can see the slope direction (which means it is moving north, quite fast), which is computed with the latitude and longitude betas from the regressions stated earlier. And secondly, the cab supply percent change is derived from the histogram, to denote how many cabs "disappear", which is very close to the NYT article's estimate of 20%.

Here are some of the figures to capture overall distribution over the day. First, we have the histograms, representing the frequency over half hour intervals:



And secondly, here is how the latitude varies during the designated "dropzone" period, with the average and midtown latitudes overlaid:



From this analysis, I was able to learn that this phenomenon, while is definitely existent, is far more complex. The vector direction is very volatile, given what your designated dropzone is. While there is a definite shift upwards – there doesn't seem to be a converging location for all the cabs. This is most likely because the distribution of cab garages is wide throughout the city. There are garages un the Far West Side, Queens, and even the Bronx. However, while the converging location of the cabs is undefined, we know that the cab supply drop is real, and cabs are moving to these more northern locations for a presumed shift change – which is enough to confirm that the urban legend of ghost taxis, is in fact, real.