**NYC Taxi Data: New Year’s Week**

**Overview**

When I began searching for interesting datasets for this project, I came across this blog post: <http://chriswhong.com/open-data/foil_nyc_taxi/> which had tons of csv files on NYC taxi’s that was retrieved via a FOIL request. Despite listing torrent downloads for the zipped files, I found that there was only 1 seeder and that the zip files that are downloaded may be corrupted. But the files are also available for download as individual zip folders containing csv files at this link: <http://www.andresmh.com/nyctaxitrips/>. For this analysis, only trip\_data\_1.csv and trip\_fare\_1.csv were used. The two were combined into one excel file to make it easier to load into pandas as a data frame. Column names include:

["medallion", "hack\_license", "vendor\_id", "pickup\_datetime", "payment\_type", "fare\_amount", "surcharge", "mta\_tax", "tip\_amount", "tolls\_amount", "total\_amount", "dropoff\_datetime", "passenger\_count", "trip\_time\_in\_secs", "rate\_code", "trip\_distance", "pickup\_longitude", "pickup\_latitude", "dropoff\_longitude", "dropoff\_latitude"]

Since I couldn’t find a read-me file for this data, I had to infer or lookup column variable meanings. The only variable set that was truly puzzling to me was payment type, but thanks to this person for their analysis and explanation of the payment type abbreviations: <http://www.claygervaisgibson.com/nyc-taxi-data/>.

Questions guiding analyses:

* Performing tip analysis, are there any trends?
* Is there a way to categorize taxi trips?
* Do fare and tip vary depending on number of passengers?

After loading the data frame, stripping whitespace from the 19 column names, and then dropping unnecessary columns as needed, I ended up with 1,048,575 rows of clean data (each representing a taxi ride). The date time ranges from 1/1/13 to 1/9/13 and includes a total of 9,383 unique medallion cabs and 17,610 unique cab drivers.

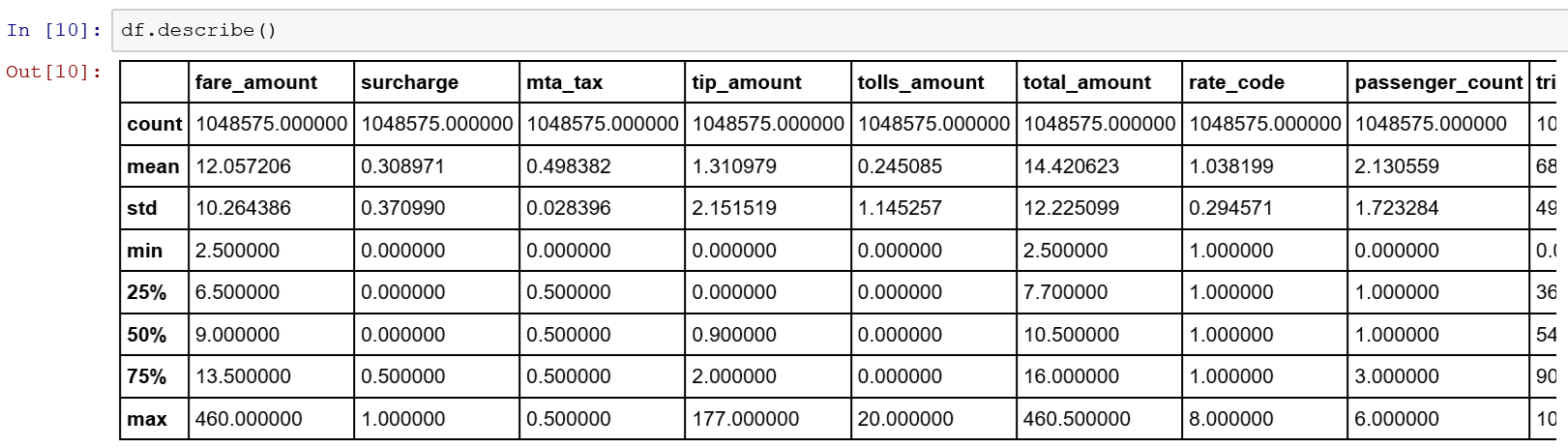


Table 1: Data frame descriptive statistic sample

**Analyses**

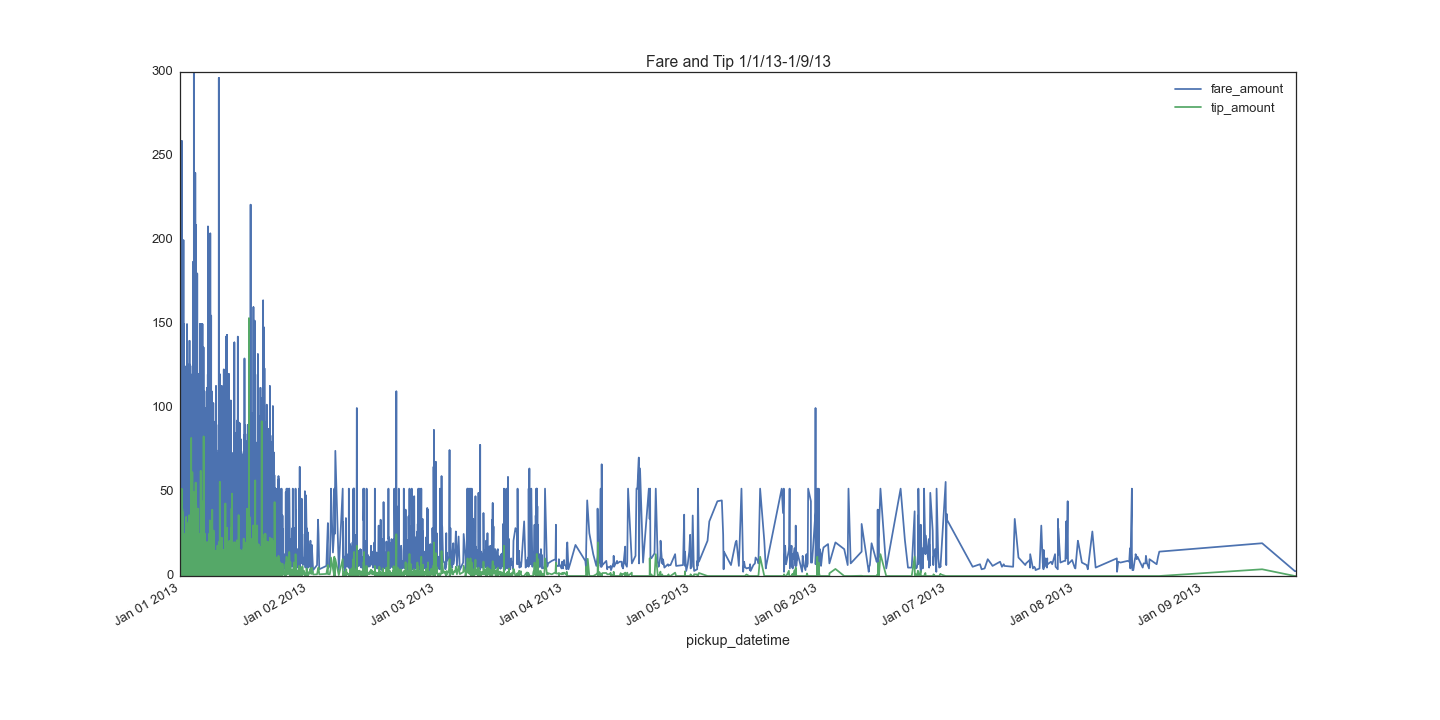
I applied the following analyses to my dataset:

* Distributional analyses
* Regression
* Correlation analysis
* Clustering (and related evaluations)
* Dimensionality reduction
* Time series
* Multidimensional visualization techniques

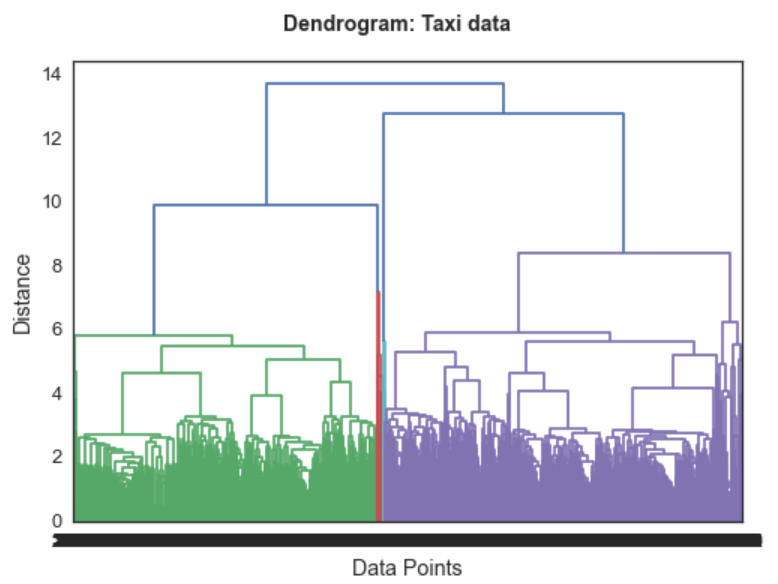
Out of these concepts, the time series analysis and clustering provided some of the most interesting outcomes.

For the time series, I plotted tip amount and fare amount over eight days between 1/1/13 and 1/9/13. I didn’t quite grasp the sheer volume of rides given on New Year’s Day until it was plotted over time.

See time series plot on next page:

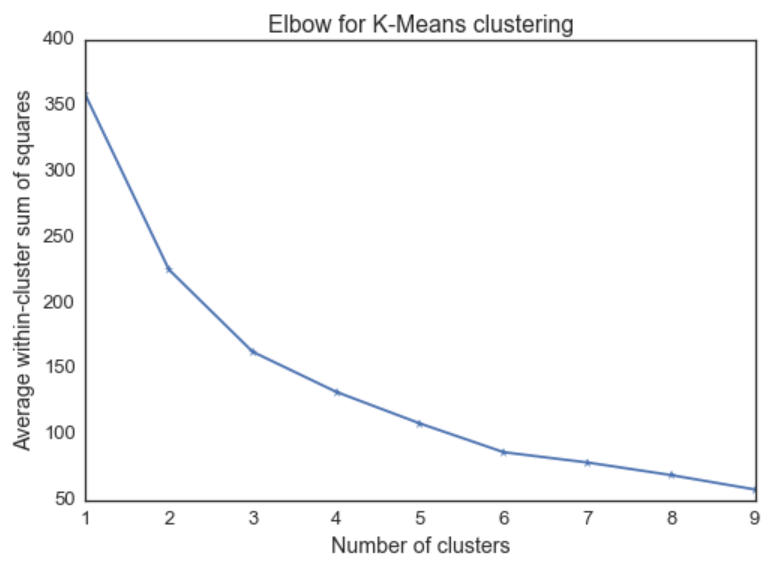
Fare amounts and tip amounts have a massive spike on New Year’s day, but quickly taper off after Tuesday, January 2nd. Intuitively this makes sense because millions of people travel to New York to celebrate New Year’s and they presumably are all still in the city the next day. But as people leave, the frequency of rides drastically reduces, and with it so does fare prices and tip amounts. I would have liked to have done some autocorrelation or checked for seasonal decomposition, but my computer doesn’t have enough processing power to run anything more than the existing observations.

Cluster analysis allowed me to dig into consumer behavior as well as taxi trip patterns. Due to computational limits, the following analyses were run using a sample size of 4,000 taxi trips. Using hierarchical clustering and k-means, I was able identify four different categories of taxi rides:

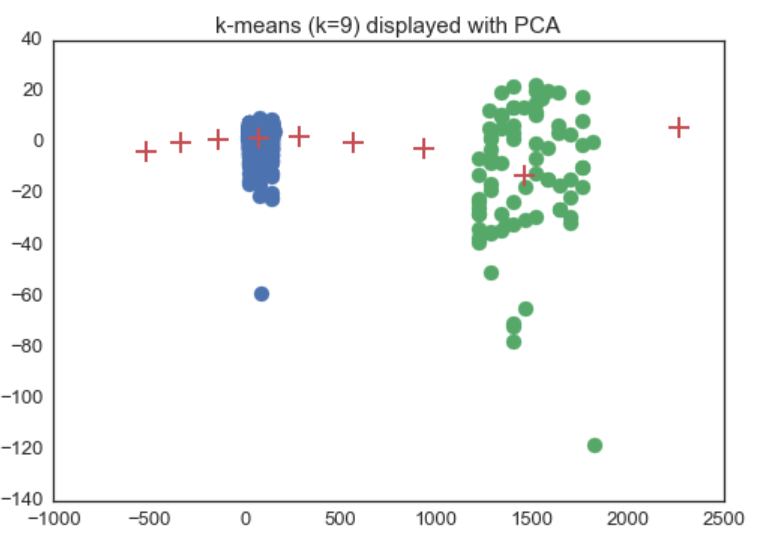


From the dendrogram we can see that there are two large, very distinct groups from the green and purple clusters. Right in the middle, there two other small groups can be seen that are red and blue. Just from looking at the diagram, I decided to cut the clusters at 9 so we would only get a few well defined categories.

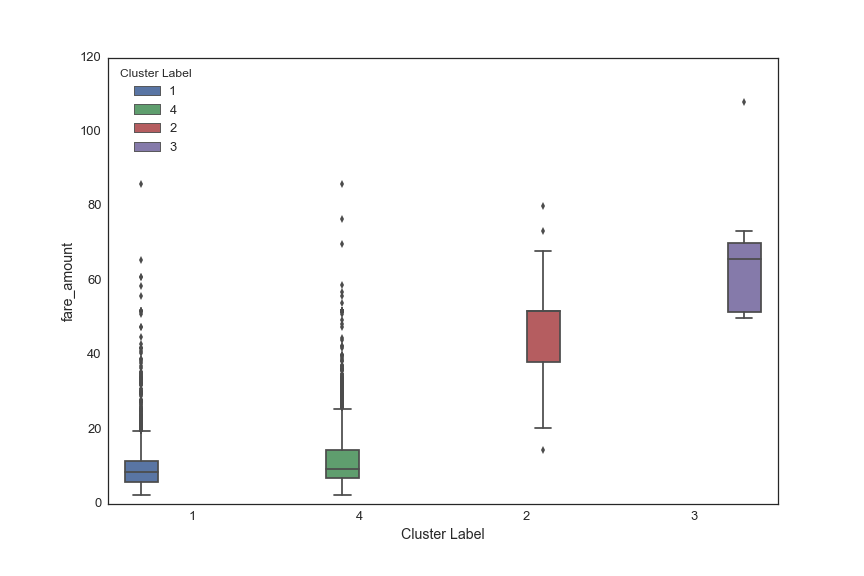
As confirmed in the Elbow plot below, 9 is a pretty good number of components for running the k-means clustering model.

****

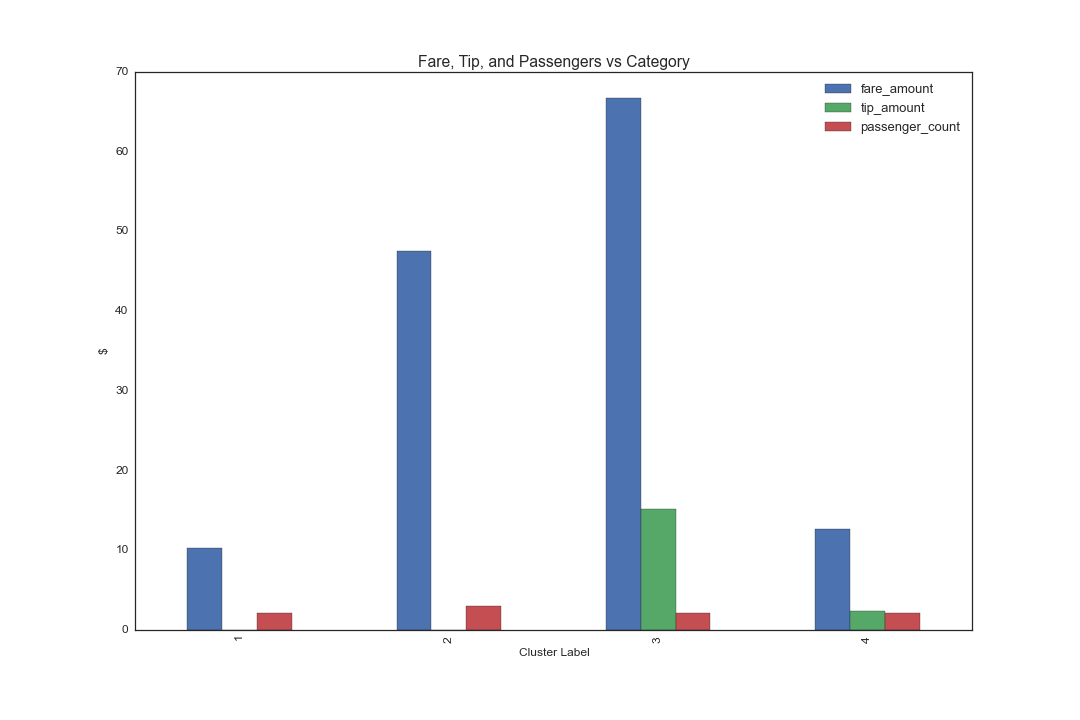
Which, after running the k-means model with 9 components, gives us this k-means cluster plot (on the next page) which displays two distinct groups:

****

Once I had the cluster labels, four categories overall, I added a category label column back to the denormalized data frame. The reason for this is that the models require normalized data to run, but denormalized data is more intuitive to view and interpret. So, once I had the cluster labels, I grouped the data frame on the four cluster labels and took the mean of each column so that I could hone in on individual factors that distinguish each category.

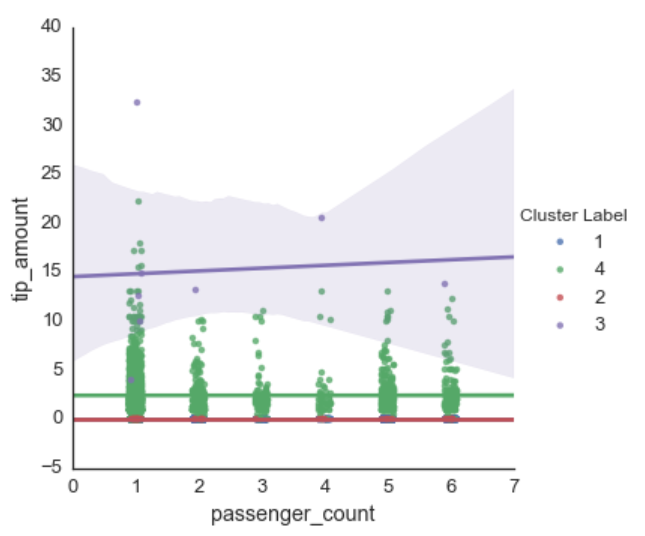


As illustrated in the boxplot above, the main clusters 3 and 4 reveal distinct fare costs. And clusters 1 and 2 closely mirror the behavior of 4 and 3 respectively. To dig into this further, I created a bar plot of fare amount, tip amount, and passenger count:

****

Here, the cluster distinctions become clearer. Cluster 3 is characterized by high fare amounts, longer distance trips, low passenger counts, and high tips. Cluster 4 has lower fares, shorter distance trips, low passenger counts, and low tips. Cluster 3 exhibits a more stable tip pattern and frequency, while cluster 4 shows lots of outliers and variance amongst tips. Clusters 1 and 2 share most of these trip characteristics, but the main difference is that these customers provide no tip, with cluster 2 having slightly higher passenger rates. Luckily for cab drivers, clusters 3 and 4 represent the majority of taxi trips, which means that they would hopefully receive a tip for their services most of the time.

The regression plot below shows just how much more tip cluster 3 provides compared to the shorter, more common trip type of cluster 4. While this information could be valuable for predicting consumer and cabbie behavior, more work can and should be done to further analyze taxi trip factors and customer behaviors. The cluster model can be improved and outlier detection could reveal that the data is being skewed by a few high tips or a certain group of taxi passengers. As seen by the green cluster 4 plots, tips are higher and more frequent when there are fewer than two passengers.



**Future Steps**

For future analysis, I would examine outlier detection, refine the clustering models, break down cluster factor composition even further, and use GIS and time series components to map taxi routes around the city.

Based on some of the scatterplots, it’s clear that some clusters have outliers regarding how much they tip with relation to their overall fare cost. It would be interesting to see what might be causing these outliers, and it could tell us more about how to predict types of taxi customer or cab driver behavior. Given that we have the cab number and cabbie license for each trip, there could be some cool insights on how particular drivers behave or influence passenger tipping behavior.

The clustering model could probably be improved. I chose a k-means cutoff that would provide 4 cluster labels so that it would be easier to analyze and display the clusters for the purposes of this project, but the optimal number of cluster labels is probably higher. More clusters would mean more factors to analyze, and this could reveal finer trip details that are glossed over by aggregating the data into limited cluster labels.

Finally, it would be super fun to use the GIS data to map cab trips. Studying route patterns could help consumers have a more pleasant experience with taking cabs in NYC and it could help cab drivers optimize their fares and tips.