cabDataProject

Anthony Saita, Juna Iafelice, Joelle Bracco, Nicholas Weidner

5/24/2020

### Purpose of the Project

Yellow Cabs have become synonymous with New York City. Though struggling with the onslaught on new entrants from shared ride services, New York Yellow Cabs remain an immensely convenient and accessible means of transportation in a city of 8 million inhabitants. This project will examine the raw data associated with Yellow Cabs in New York from the perspective of a cab driver. Given the 130 million rows of data, how can a cab driver parse usable information without getting lost in a forest of data? This project seeks to provide cab driver with useable information to maximize their tipping profits and maximize their days to provide the highest profitability possible in a city inundated with competition.

### Overview of Data

The data on Yellow Cab trips was obtained from <https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page>, a New York City organization dedicated to tracking cab data for various purposes that are outlined in the above link. For more granular analysis, the data is divided into the 12 months of the year. After combining all the data sets, the Taxi Zone Lookup Table was added to the combined data file to provide borough and neighborhood information for each pickup and dropoff location. Without the inner joined table, the pickup and dropoff locations would be simple numeric identifiers.  
After this inner join, the taxiCabData2018 variable was created, but 130 million rows were far too onerous for the group’s systems to work with in a quick and reliable manner. Use sample\_n of 1,000,000 rows from each month, a 12 million row data set entitled nycTaxiData2018 was created. To further enhance usability further, this new dataset of 12 million was converted into a tibble, with the datetime columns separated into date and time (by seconds). After the tibble creation, work began on the cleaning of the data.

### Data Dictionary for Columns

To better understand the data set, it’s important to follow the data dictionary and variables the group decided were important to the success of the project. Accordingly, the following link: <https://www1.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf> details the column names and their descriptions. Nothing was edited from the source material except for the following: tpep\_pick\_datetime and tpep\_dropoff\_datetime were changed divided into Pickup\_Date/Pickup\_Time and Dropoff\_Date/Dropoff\_Time. The \_date variables are formatted as dates in format of yyyymmdd. The times are in hms format in seconds, which were converted into minutes for the purposes of analysis. In addition, Pickup\_Location\_ID and Dropoff\_Location\_ID were divided into PULocationID, PU\_Borough, PU\_Neighbourhood, and PU\_ServiceZone. The Borough and Neighborhood provide character variables with the name of the boroughs and neighborhoods in NYC.

### Data Cleaning

The data set was cleaned in the following ways:

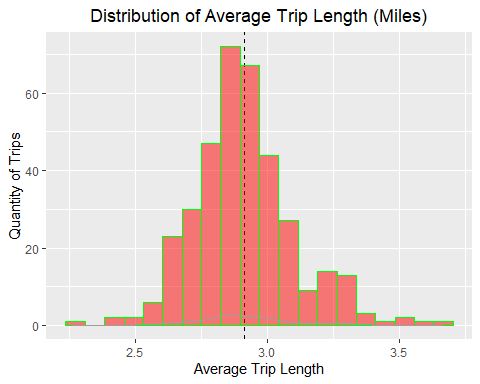
1. All trips that had negative time values were excluded from the dataset. (i.e. Dropoff\_Time-Pickup\_Time must be greater than 0)
2. The Pickup\_Date must be between 2018-01-01 and 2018-12-31
3. Trip length is over 0 miles.
4. Trip length is less than 32 miles (The distance between Newark and LaGuardia Airports)

The resulting data set, taxiDataClean, contained 11,785,260 out of 12,000,000 original rows of data. This equates to a loss of 1.8% of the data.

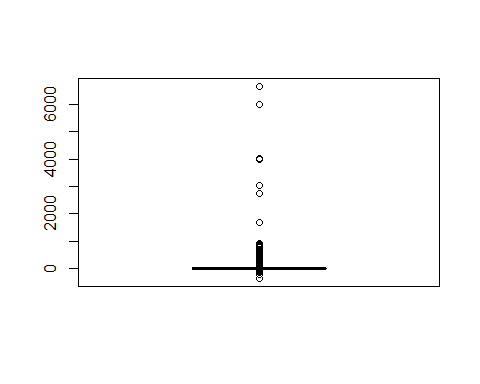
### Testing Data for Centrality

In addition to the cleaning, an additional column was added for Average Trip Length by pickup date to visualize the distribution. The following graphs below some tests. When filtering the data for average trip length and summarizing the columns by day (356 days), a historgram is produced that illustrates a normal distribution with a mean slighly higher than the median, indicating a positive skewing of the data. This is illustrated when outliers are included in the attached box plot, below the histogram, which shows the effect of outliers on the greater majority of data points.

### Test Plot  
ggplot(data=AverageTripLength, aes(x=avgTripLength))+  
 geom\_vline(aes(xintercept = mean(avgTripLength)), linetype = 2)+  
 geom\_histogram(bins=20, fill='red', color="green", alpha=0.5)+  
 geom\_density(col=200) +  
 labs(x="Average Trip Length", y = "Quantity of Trips") +   
 ggtitle("Distribution of Average Trip Length (Miles)") +  
 theme(plot.title = element\_text(hjust = 0.5))



### Test Plot  
boxplot(taxiDataClean$Total\_Charge\_to\_Passengers)



### Predictions and Approach

Our group is approaching our theory from the lens of maximizing daily income for a proposective Yellow Cab driver. For many, a large part of their income that they have “control” over is the tips that they are able to generate, both daily and per ride. Our group theorizes that not only do the two direct inputs to calculating fare: Trip Duration and distance factor into the total tip, but Time of Day and Day of the week do as well. Payment types other than credit cards have been excluded, as Tip Data is not tracked for non-credit card tips. In order to analyze potential models of these variables, they will need to be added to the dataframe, and then we can run a random forest as well as a conventional lm() model and check for significance.

### Random Forest and Model

Our initial model as previously mentioned will consist of running a Random Forest on the variables Trip Distance, Trip Duration, Hour, Day of Week and Month. We will then also run a conventional lm() model and run ANOVA tests to determine which variables are significant to determining Tip Percentage. We will then run addional models to determine if there is variance between nominal and percent tip signficant variables.

fitted\_forest <- randomForest(formula = Tip\_Percent ~ Trip\_Distance\_Miles + Trip\_Duration + hour + wday + month,  
 data =taxiTipData,  
 ntree = 80, sampsize = 10000)  
  
print(fitted\_forest)

##   
## Call:  
## randomForest(formula = Tip\_Percent ~ Trip\_Distance\_Miles + Trip\_Duration + hour + wday + month, data = taxiTipData, ntree = 80, sampsize = 10000)   
## Type of random forest: regression  
## Number of trees: 80  
## No. of variables tried at each split: 1  
##   
## Mean of squared residuals: 4664.027  
## % Var explained: 0.06

importance(fitted\_forest)

## IncNodePurity  
## Trip\_Distance\_Miles 121230.14  
## Trip\_Duration 165908.27  
## hour 42440.62  
## wday 72653.53  
## month 24460.27

As we can see from the output, the model does a relatively poor job of explaining the variance in tip percentage, accounting for only 12% of variance. What is somewhat surprising though is how much more important Trip Duration appears to be than Trip Distance. The month, day, hour importance is somewhat inconclusive by this technique, so further analysis will be pursued.

TipModel <- lm(Tip\_Percent~ Trip\_Distance\_Miles + Trip\_Duration + hour + wday + month, taxiTipData)  
summary(TipModel)

##   
## Call:  
## lm(formula = Tip\_Percent ~ Trip\_Distance\_Miles + Trip\_Duration +   
## hour + wday + month, data = taxiTipData)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20 -2 0 1 133313   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 20.431102 0.066949 305.175 < 2e-16 \*\*\*  
## Trip\_Distance\_Miles -0.007787 0.009541 -0.816 0.414365   
## Trip\_Duration -0.038964 0.002745 -14.193 < 2e-16 \*\*\*  
## hour -0.013493 0.004057 -3.326 0.000881 \*\*\*  
## wday.L -0.035215 0.067594 -0.521 0.602382   
## wday.Q -0.148865 0.067265 -2.213 0.026891 \*   
## wday.C 0.025424 0.065919 0.386 0.699727   
## wday^4 0.053580 0.065255 0.821 0.411596   
## wday^5 -0.083065 0.064277 -1.292 0.196259   
## wday^6 -0.007313 0.063562 -0.115 0.908406   
## month.L -0.037308 0.085358 -0.437 0.662052   
## month.Q 0.074694 0.085633 0.872 0.383070   
## month.C 0.055384 0.085505 0.648 0.517165   
## month^4 0.176835 0.085413 2.070 0.038420 \*   
## month^5 -0.047502 0.085348 -0.557 0.577822   
## month^6 0.004464 0.085208 0.052 0.958222   
## month^7 -0.078327 0.085351 -0.918 0.358776   
## month^8 0.072747 0.085385 0.852 0.394220   
## month^9 -0.018080 0.085482 -0.212 0.832489   
## month^10 0.139058 0.085753 1.622 0.104886   
## month^11 -0.036022 0.085946 -0.419 0.675123   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 68.31 on 7676344 degrees of freedom  
## Multiple R-squared: 6.131e-05, Adjusted R-squared: 5.87e-05   
## F-statistic: 23.53 on 20 and 7676344 DF, p-value: < 2.2e-16

Anova(TipModel)

## Anova Table (Type II tests)  
##   
## Response: Tip\_Percent  
## Sum Sq Df F value Pr(>F)   
## Trip\_Distance\_Miles 3.1090e+03 1 0.6662 0.4143653   
## Trip\_Duration 9.4014e+05 1 201.4541 < 2.2e-16 \*\*\*  
## hour 5.1623e+04 1 11.0619 0.0008812 \*\*\*  
## wday 3.6906e+04 6 1.3180 0.2449015   
## month 4.8837e+04 11 0.9513 0.4891239   
## Residuals 3.5824e+10 7676344   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

### Model Validation

This was a surprising result in our groups opinion. A traditional linear model determines that for Tip Percentage, the Trip Distance is statisticaly insignificant, but that Month, Day of Week and Hour are all statistically significant. The Model summary further identifies specific days of the week, times and months which significantly impact Tip Percentage.

Now for Nominal Tips:

fitted\_forest2 <- randomForest(formula = Tip\_Amount ~ Trip\_Distance\_Miles + Trip\_Duration + hour + wday + month,  
 data = taxiTipData,  
 ntree = 80, sampsize = 10000)  
print(fitted\_forest2)

##   
## Call:  
## randomForest(formula = Tip\_Amount ~ Trip\_Distance\_Miles + Trip\_Duration + hour + wday + month, data = taxiTipData, ntree = 80, sampsize = 10000)   
## Type of random forest: regression  
## Number of trees: 80  
## No. of variables tried at each split: 1  
##   
## Mean of squared residuals: 2.11147  
## % Var explained: 67.63

importance(fitted\_forest2)

## IncNodePurity  
## Trip\_Distance\_Miles 26490.7129  
## Trip\_Duration 21075.0507  
## hour 2421.1747  
## wday 970.6508  
## month 1049.0737

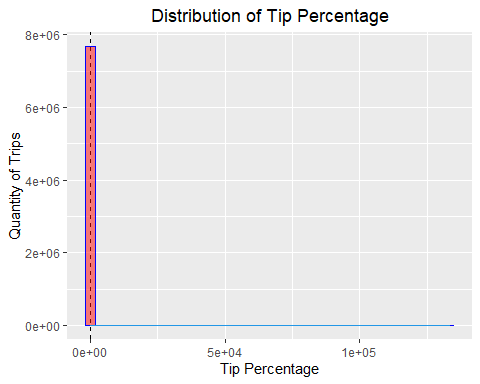
This output is not surprising, as the Nominal Tip amount is directly a function of Trip Distance and Trip Duration. It is further not surprising to see that there is little if any impact on Tip Amount depending on the Month, Day of the Week or Hour of day. With over 68% of variance in Tip Amount explained by the variables above, the model for Tip Amount could be trimmed to just Trip Distance and Trip Duration, and explore location data to improve variance explanation.

Because Tip Amount is a function of standard trip metrics that the driver has little to no control over, we will divert our focus to Tip Percentage. Accounting for the variability, minimal though it may be, a prospective driver should look to maximize their tip percentage, if possible.

### Tip Percentage

Yellow Cabs have pre-set tip selections available for anyone electing to pay by card. The common presets are visible in the histogram below.

ggplot(data=taxiTipData, aes(x=Tip\_Percent))+  
geom\_vline(aes(xintercept = mean(Tip\_Percent)), linetype = 2)+  
geom\_histogram(bins=40, fill='red', color="blue", alpha=0.5)+  
geom\_density(col=100) +  
labs(x="Tip Percentage", y = "Quantity of Trips") +   
ggtitle("Distribution of Tip Percentage") +  
theme(plot.title = element\_text(hjust = 0.5))

 It is quite evident that there is a strong preference for tipping 20%, and that 20% is one of the aforementioned presets. It is also notable that though there is a noticeable tail towards higher tip percentages, they are dwarfed in impact by tips below 20%, resulting in an average tip of 19.43% for the dataset.

In an effort to pursue tips which are above average, we can now incorporate location data to our previous model in an effort to improve it.

fitted\_forest3 <- randomForest(formula = Tip\_Percent ~ PU\_Lat + PU\_Long + Trip\_Distance\_Miles + Trip\_Duration + hour + wday + month,  
 data = taxiTipData2,  
 ntree = 80, sampsize = 10000)  
  
print(fitted\_forest3)

##   
## Call:  
## randomForest(formula = Tip\_Percent ~ PU\_Lat + PU\_Long + Trip\_Distance\_Miles + Trip\_Duration + hour + wday + month, data = taxiTipData2, ntree = 80, sampsize = 10000)   
## Type of random forest: regression  
## Number of trees: 80  
## No. of variables tried at each split: 2  
##   
## Mean of squared residuals: 4715.001  
## % Var explained: -1.81

importance(fitted\_forest3)

## IncNodePurity  
## PU\_Lat 68942.95  
## PU\_Long 49399604.24  
## Trip\_Distance\_Miles 6012071.83  
## Trip\_Duration 8304966.94  
## hour 61892.55  
## wday 27331.31  
## month 66325.44

Unfortunately, using location data does not appear to improve our model in this case. We may be attempting to use the wrong model or have the data in the incorrect format to test in this way. Despite the failure to improve the model, we are still able to utilize the neighborhood data in order to identify neighborhoods that tip above average. The resulting dataframe provides 3877 observations of pairings of Pickup and Dropoff with average and volume statistics. Trips which occured in too limited a quantity were excluded, as a trip which occurs once per year is not useful for planning where to focus your driving if you are attempting to maximize tips. Maybe you get lucky and land the single occurence trip and get excellent tips for that day but the low volume means that trip and the resulting tip are likely extremely infrequent. As a result, a minimum number of trips of 150 was selected, as this represents trips which should occur with some degree of frequency even if it’s not daily.

## Best Tip Routes  
bestTipRoute <- taxiTipData %>%  
 filter(Tip\_Amount != 0, Pickup\_Location\_ID < 264, Dropoff\_Location\_ID < 264) %>%  
 group\_by(PU\_Borough, PU\_Neighbourhood, DO\_Borough, DO\_Neighbourhood) %>%  
 summarise(avgTipPercentage = mean(Tip\_Amount/Fare\_Pre\_Tip\*100), AvgTipAmount = mean(Tip\_Amount), PU\_DO\_Count = n())%>%  
 filter(PU\_DO\_Count > 150) %>%  
 arrange(desc(avgTipPercentage))  
  
head(bestTipRoute, n=10)

## # A tibble: 10 x 7  
## # Groups: PU\_Borough, PU\_Neighbourhood, DO\_Borough [9]  
## PU\_Borough PU\_Neighbourhood DO\_Borough DO\_Neighbourhood avgTipPercentage  
## <chr> <chr> <chr> <chr> <dbl>  
## 1 Queens Astoria Queens LaGuardia Airpo~ 128.   
## 2 Brooklyn Carroll Gardens Brooklyn Carroll Gardens 117.   
## 3 Manhattan Penn Station/Ma~ Manhattan Newark Airport 101.   
## 4 Manhattan Two Bridges/Sew~ Manhattan Two Bridges/Sew~ 99.2  
## 5 Manhattan Upper East Side~ Queens JFK Airport 48.3  
## 6 Manhattan Flatiron Manhattan Midtown South 47.9  
## 7 Manhattan Chinatown Manhattan Chinatown 45.4  
## 8 Queens LaGuardia Airpo~ Queens LaGuardia Airpo~ 40.4  
## 9 Manhattan Financial Distr~ Manhattan Financial Distr~ 39.2  
## 10 Queens LaGuardia Airpo~ Queens East Elmhurst 37.1  
## # ... with 2 more variables: AvgTipAmount <dbl>, PU\_DO\_Count <int>

When sorting by the largest tip percentage, a few common Neighborhoods pop up - The Financial District in particular - as well as both Airports in Queens. Those familiar with NYC will also be quick to point out that all of these trips are quite close, either within the same neighborhood or into the adjacent neighborhood. Additionally, while all the tip amounts are above average, the actual Tip Amount is relatively small. It is feasible that with enough volume, this could be a valid strategy for maximizing daily tips by trying to mainly operate out of these locations - the financial district in particular featuring three times in the top 10 for tip percentage seems a likely place to begin if this is the chosen strategy. The volume of trips however, indicated by PU\_DU\_Count are relatively low given the size of our data set.

What if we were to however, only consider the pairings where the average tip Percentage is greater than the sample average and look at the volume of trips pairings. This still should yield higher than average tips, however we can identify neighborhoods with much larger volume numbers and identify a location with more frequent trips if a volume strategy is sought.

sample\_AvgTipPct <- mean(bestTipRoute$avgTipPercentage)  
bestTipRoute2 <- taxiTipData %>%  
 filter(Tip\_Amount != 0, Pickup\_Location\_ID < 264, Dropoff\_Location\_ID < 264) %>%  
 group\_by(PU\_Borough, PU\_Neighbourhood, DO\_Borough, DO\_Neighbourhood) %>%  
 summarise(avgTipPercentage = mean(Tip\_Amount/Fare\_Pre\_Tip\*100), AvgTipAmount = mean(Tip\_Amount), PU\_DO\_Count = n())%>%  
 filter(avgTipPercentage>sample\_AvgTipPct) %>%  
 arrange(desc(PU\_DO\_Count))  
  
head(bestTipRoute2, n=10)

## # A tibble: 10 x 7  
## # Groups: PU\_Borough, PU\_Neighbourhood, DO\_Borough [6]  
## PU\_Borough PU\_Neighbourhood DO\_Borough DO\_Neighbourhood avgTipPercentage  
## <chr> <chr> <chr> <chr> <dbl>  
## 1 Manhattan Upper East Side~ Manhattan Upper East Side~ 20.0  
## 2 Manhattan Upper East Side~ Manhattan Upper East Side~ 19.4  
## 3 Manhattan Upper East Side~ Manhattan Upper East Side~ 20.6  
## 4 Manhattan Upper East Side~ Manhattan Upper East Side~ 20.8  
## 5 Manhattan Upper West Side~ Manhattan Upper West Side~ 20.3  
## 6 Manhattan Upper West Side~ Manhattan Lincoln Square ~ 20.3  
## 7 Manhattan Lincoln Square ~ Manhattan Upper West Side~ 20.2  
## 8 Manhattan Upper East Side~ Manhattan Midtown East 19.8  
## 9 Manhattan Upper West Side~ Manhattan Upper West Side~ 20.2  
## 10 Manhattan Lenox Hill West Manhattan Upper East Side~ 19.4  
## # ... with 2 more variables: AvgTipAmount <dbl>, PU\_DO\_Count <int>

In this instance, we see lower average Tip Percentage, but substantially higher volume, while still earning above average tips by percent. Additionally, these locations are geographically distinct from what we saw in our previous case, though they have the same phenomenon of predominantly representing relatively short trips within one or two neighborhoods. If a cab driver were to focus on using a high volume strategy with frequent short trips earning above average tips, it seems that the neighborhoods surrounding central park would be a good place to start.

sample\_AvgTipPct <- mean(taxiTipData$Tip\_Percent)  
  
taxiTipData3 <- taxiTipData %>%  
 filter(Tip\_Amount != 0, Pickup\_Location\_ID < 264, Dropoff\_Location\_ID < 264) %>%  
 group\_by(wday, hour) %>%  
 summarise(avgTipPercentage = mean(Tip\_Amount/Fare\_Pre\_Tip\*100), AvgTipAmount = mean(Tip\_Amount), PU\_DO\_Count = n())%>%  
 filter(avgTipPercentage>sample\_AvgTipPct, PU\_DO\_Count > 150) %>%  
 arrange(desc(avgTipPercentage))  
  
head(taxiTipData3, n=20)

## # A tibble: 20 x 5  
## # Groups: wday [7]  
## wday hour avgTipPercentage AvgTipAmount PU\_DO\_Count  
## <ord> <int> <dbl> <dbl> <int>  
## 1 Thu 3 25.7 2.85 5222  
## 2 Wed 12 22.5 2.89 55589  
## 3 Fri 6 22.3 2.74 30345  
## 4 Tue 15 21.6 2.93 58686  
## 5 Tue 2 21.5 2.83 5096  
## 6 Sat 6 21.1 3.59 10382  
## 7 Sun 8 21.1 2.76 20668  
## 8 Wed 3 21.0 2.93 4305  
## 9 Sun 10 20.9 2.48 43620  
## 10 Mon 2 20.8 2.90 6342  
## 11 Fri 5 20.8 3.82 11070  
## 12 Wed 4 20.7 3.86 4493  
## 13 Fri 2 20.7 2.68 13938  
## 14 Thu 2 20.7 2.70 8710  
## 15 Thu 4 20.6 3.82 5021  
## 16 Tue 12 20.5 2.87 54271  
## 17 Wed 2 20.5 2.76 7360  
## 18 Tue 1 20.5 3.11 9379  
## 19 Sat 5 20.5 4.13 6484  
## 20 Mon 3 20.4 3.01 4259

Looking at hour and day of the week when sorted by average Tip Percentage, it appears that late night and very early morning hours are the highest tipping times of day, with friday between 2 am and 5 am appearing as 4 of the best tipping hours during the week.

### Drawing Conclusions

Our initial Random Forest found that Trip Duration, Hour, Day of the Week and Month were all factors when considering how to maximize Tip Percentage. The Random Forest additionally factored in Trip Distance; however, Linear Models and model validation via ANOVA rejected Trip Distance as not statistically significant. If the data set is Linear in Nature it would make sense to leave distance out of the equation. If the Random Forest is a better model, then it is prudent to keep Trip Distance as a part of the model.

Additional testing is needed regardless, as the % of variance explained by the model is insufficient for predictive purposes. At this time, a prospective yellow cab driver would do best with the knowledge that late nights, in neighborhoods around Central Park are likely the best way to maximize Tip Percentage while still having a high volume of customers.