MAST30034. Applied Data Science

Y2019S1 Assignment 2

Yin Zhou Zheng (911261)

ABSTRACT

The overall purpose of this project is to understand "what constitutes a profitable taxi driver in New York". In this report, we analyse the degree to which major sporting events such as the US Tennis Open affect overall tip proportions (with respect to the total amount paid by the hirer). We also attempt to identify other significant predictors of tip proportion.

We believe events such as the US Open should encourage higher tips, and therefore higher tip proportions. Evidence of such a relationship will encourage taxi drivers to work longer hours during the US Open to maximise their tip proportions and consequently, their overall profits.

We completed the data pre-processing in Python; and followed up with statistical tests and variable selection in R. We built a Linear Regression model with tip proportion as the response variable, selecting predictor variables with stepwise selection (starting from the full model) using the Akaike's Information Criterion (AIC) as our goodness-of-fit measure.

To test our hypothesis, we compared the tip proportions of trips made on the days of the US Open with typical business days and non-business days. Although we found the day type to be a significant predictor in the resultant tip proportion, the corresponding effects were both small in scale and contradictory to our hypothesis. Trips made on an event day instead resulted in smaller tip proportions.

Lastly, our final Linear Regression model was a very poor fit on the data. This undermines the utility of the model as a predictor of tip proportions; and consequently, the taxi driver's ability to use this model to seek higher tip proportions.

DATA PREPARATION

Dataset Selection

We chose to only use the yellow taxi trip datasets. Unlike green taxis, yellow taxis were permitted to respond to street hails anywhere within New York City [1]. Consequently, this provided us with more data to form our analysis. Similarly, the yellow taxi trip datasets were preferred over the for-hire-vehicle datasets due to the former's greater availability of data.

To test the effects of the 2015 US Tennis Open in determining tip proportions, we chose to compare the trip records of two event days, two business days (weekdays), and two non-business days (weekend). The business and non-business days were chosen specifically to not feature any major sporting events; allowing us to inspect the effects of the US Open.

The two event days were the 12th and 13th of September, the days featuring the Women's and Men's Singles Finals respectively. The two weekdays were the 21st and 22nd of September; and the weekend was the 26th and 27th of September. Since the datasets were separated by month, we only required analysis to be done on the September 2015 dataset. This particular dataset was a 1.8 gigabyte CSV file containing more than 11 million trip records.

Attribute Selection

Our main aim of this project is to create a Linear Regression model that can predict tip proportions, a continuous quantity. To keep things simple, we only include one categorical variable in our model: a newly defined variable which will indicate if the trip record was completed on an event day, a weekday, or a weekend.

The attributes we initially selected included:

- 'tpep_pickup_datetime': The date and time of pickup.
  This attribute will allow us to create our categorical variable which indicates the type of day (event, business or non-business) of the trip.
- 'trip_distance': The distance of the trip in miles.
  A candidate predictor attribute for our model. Further analysis will be performed to decide whether or not this attribute is included in our final model.
- 'pickup_longitude' & 'pickup_latitude': The longitude and latitude values of the pickup location.
  It is important to note that these attributes are separated when fitting the model. However, if either or both appear to be significant predictors of the tip proportion, then this points to the same conclusion: the pickup location (or more specifically, just the latitude or longitude value) is a good indicator of the tip proportion

- ‘dropoff_longitude’ & ‘dropoff_latitude’: The longitude and latitude values of the drop-off location.
  Similarly, significance of these variables would suggest that the drop-off location is a good indicator of the tip proportion.
- ‘payment_type’: A numeric code that represents the payment method.
  This will be used in the pre-processing section to remove any trips with non-card payments; since tips are not recorded for such payments [2].
- ‘tip_amount’ & ‘total_amount’: The tip amount and total payment amount for the trip in USD.
  These attributes will be used to determine the tip proportion. The tip proportion will communicate the proportion of tips with respect to the total payment amount. Taxi drivers will seek to maximise this proportion to be more profitable.

Data Cleansing

We initially loaded our entire dataset into a Python Pandas data-frame. We then extracted our selected attributes listed above; thereby removing half of the original attributes. With a quick examination of the new dataset, we could identify several errors:

- ‘pickup_longitude’, ‘pickup_latitude’, ‘dropoff_longitude’, ‘dropoff_frequency’: Several trips recorded zero for these attributes. This would not make sense since New York City is located at an approximate longitude of -74 and latitude of 40.5; and the dataset should only contain records with values near these numbers. We assumed these to be GPS tracking issues.
- ‘trip_distance’: Many records featured negative or zero trip distances. Negative distances should be impossible, while zero distances suggest no movement. These may have occurred due to tracking errors.
- ‘tip_amount’ & ‘total_amount’: We also observed several negative tip amounts and total amounts which were assumed to be input errors.

We chose to remove all the records containing these errors. This involved using logical statements to identify the records with errors and removing them from our dataset. These logical statements identified:

- Records with zero values for longitude or latitude attributes.
- Records with negative or zero trip distances.
- Records with negative tip amounts.
- Records with negative total amounts.

After this initial attribute selection and cleansing process, the original dataset of more than 11 million records was reduced to less than seven million. In terms of file size, this reduced the original 1.8 gigabyte dataset down to 0.8 gigabytes.

Data Pre-Processing – Tip Proportion Attribute

We began pre-processing by creating the 'tip proportion' attribute. This involved dividing the 'tip amount' by the 'total amount'. This was completed instantly by vectorizing the division operation. The total amount also includes the tip amount [2]; therefore, our new attribute communicates the proportion of the total amount paid by the hirer that comes from tips.

We also chose to remove any outliers in the tip proportion attribute. Since we wish to fit a Linear Regression, outliers have a lot of potential to influence the gradient of the line of best fit. Consequently, this could result in a model that cannot predict a 'realistic' tip proportion. The same logic also applies for the other potential predictors which will be addressed in the next section.

We initially plot a histogram of the trip proportion values in *Figure 1*. At first glance, it is evident that there are outliers with a trip proportion above 0.4. The scales auto-adjust to fit all the data; therefore, from 0.4 to 1.0, although we do not see any columns, they still exist. They are simply insignificant relative to the frequencies of the visible columns; therefore, invisible (appear to be zero in frequency) due to the figure size limitations of the histogram.

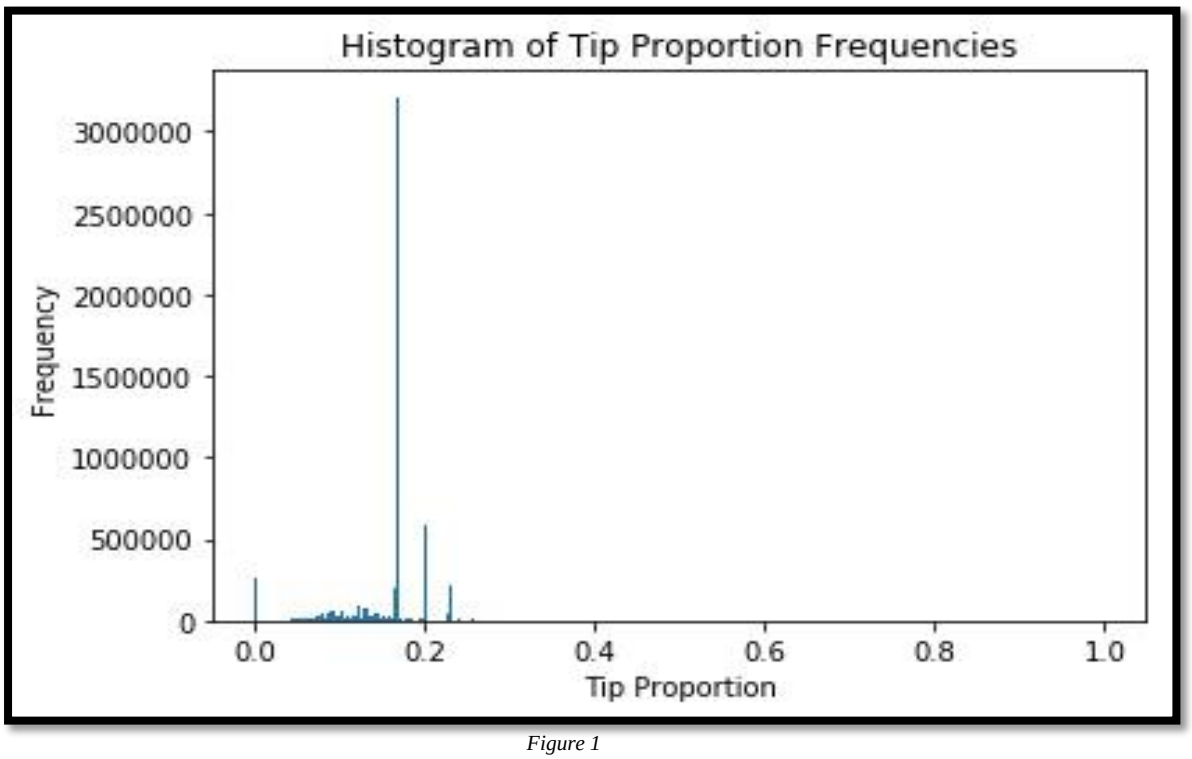

Figure 1

We proceeded to restrict the dataset to only contain records with trip proportions less than 0.4. However, we encountered a similar situation, now with outliers between 0.3 and 0.4. This led to our

final decision to keep records with trip proportions less than 0.3; producing the following histogram in *Figure 2*. Evidently, we can see a clear peak at a tip proportion of approximately 0.165. The restriction on the tip proportion removed 29,905 records. However, this is rather insignificant to the pre-removal size of 6,897,179 records. We note that more than half of these records involve a tip proportion of around 0.165.

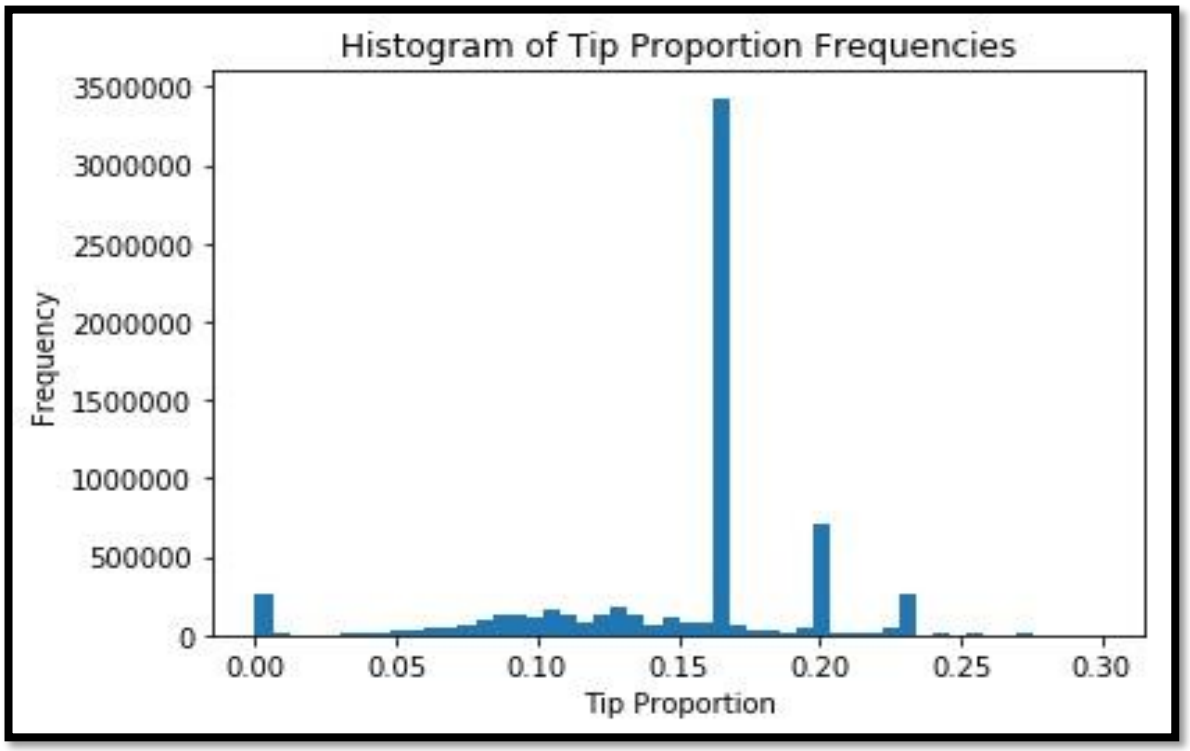

Figure 2: Histogram with Tip Proportion < 0.3

Data Pre-Processing – Predictor Variables & Restrictions

The predictor variables that will be used in the initial fitting of the Linear Regression model are the following:

- 'trip_distance'
- 'pickup_longitude'
- 'pickup_latitude'
- 'dropoff_longitude'
- 'dropoff_latitude'

We note that neither 'tip_amount' nor 'total_amount' will be used to fit our Linear Regression model. This is because they were used to directly calculate our tip proportion values. These two attributes alone determine the tip proportion; a model would be unnecessary. The purpose of this model is to inform taxi drivers how they can possibly maximise their tip proportions. Neither of these attributes would be known until after the trip occurs. On the other hand, taxi drivers will know the other attributes (listed above) prior to completing the trip. If an attribute is a significant

predictor of tip proportion, then it can be used to help determine a method of maximising the resultant tip proportion.

Following the same reasoning as for tip proportions, outliers must be removed from these attributes to avoid a misleading model. We follow the same process as above and obtain the following histograms after applying several restrictions (see *Figure 3 & 4*).

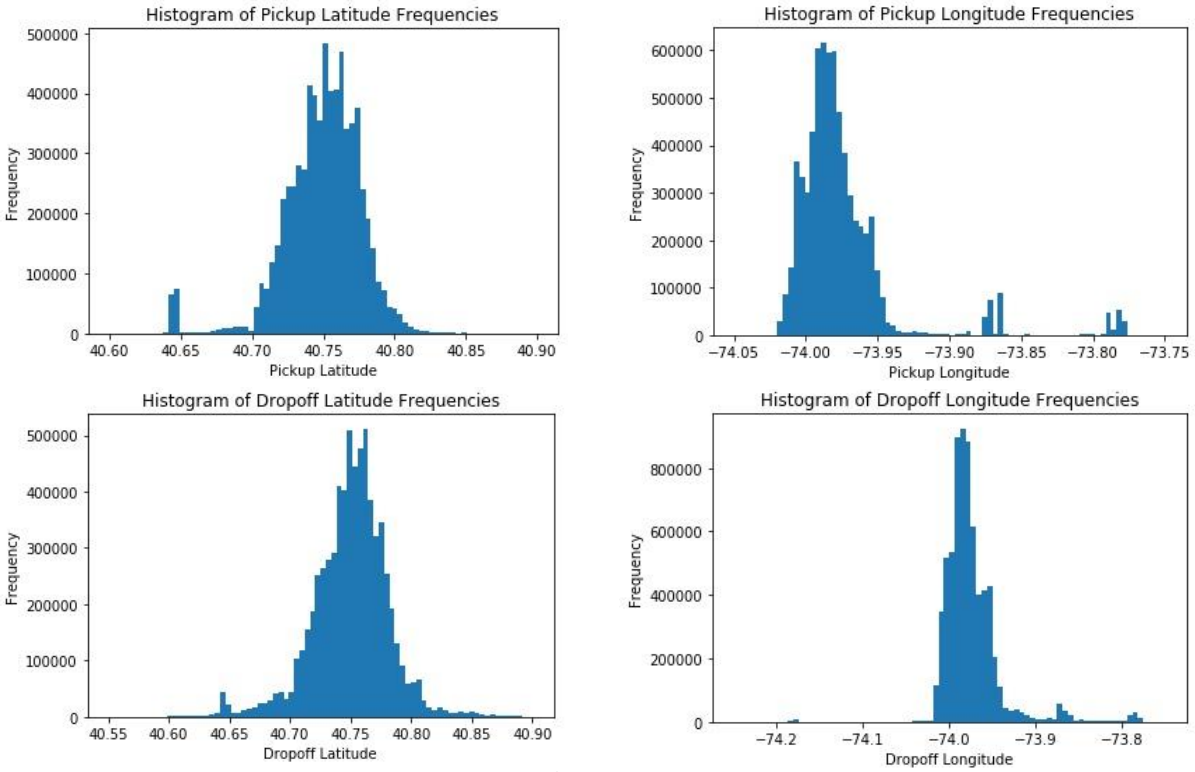

Figure 3: Histograms of Longitude & Latitude Attributes

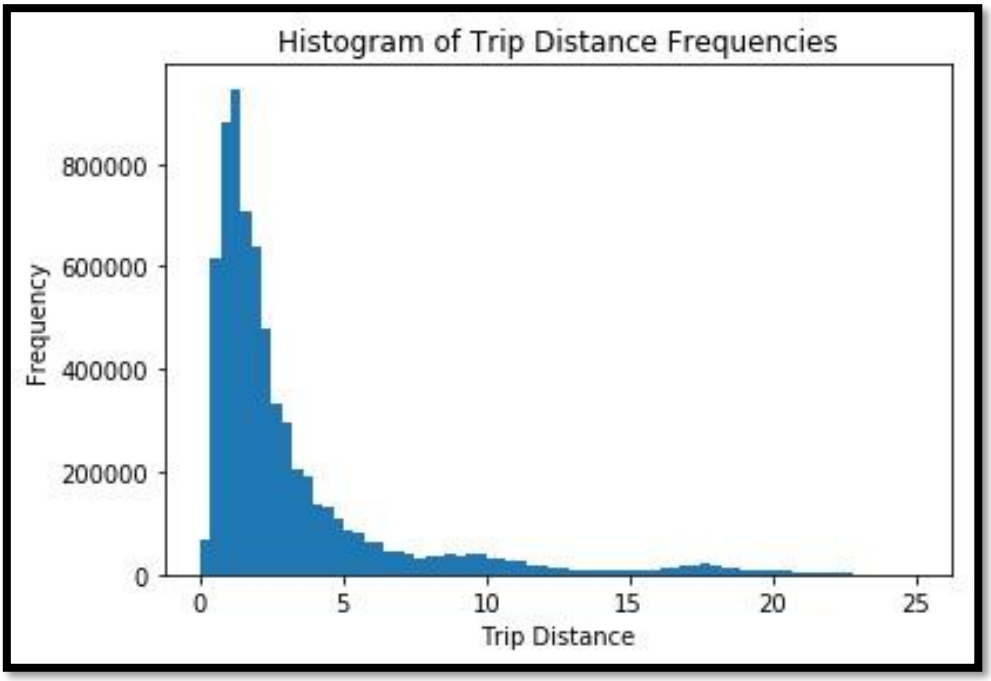

Figure 4: Histogram of Trip Distance

In *Figure 3*, we can see that both the longitude histograms share a common, single peak. And similarly, both our latitude histograms share a common, single peak. The peaks most likely occur in Manhattan, where we observed the most activity from our visualisations in Assignment 1.

In *Figure 4*, we see a heavily right skewed histogram. The majority of the taxi trips are less than five miles in distance. This should be expected since we observed common peaks in both our longitude and latitude histograms; suggesting passengers did not make long taxi trips.

We removed the outliers by applying the following restrictions:

- 'trip_distance' restricted to (0, 25).
- 'pickup_longitude' restricted to (-74.05, -73.75).
- 'pickup_latitude' restricted to (40.6, 40.9).
- 'dropoff_longitude' restricted to (-74.25, -73.75).
- 'dropoff_latitude' restricted to (40.55, 40.9).

As a consequence of these restrictions (including the restriction on tip proportion), a total of 45,588 records were removed from the original 6,897,179 records.

Data Pre-Processing – Date Specific Datasets

Next, we extracted all the trip records that occurred during our dates of interest (12th, 13th, 21st, 22nd, 26th and 27th of September). By iterating through our dataset, we recorded the indexes with pickup datetimes matching our dates of interest. We then created a separate Python Pandas data-frame for each of our dates by indexing their relevant rows from the original dataset.

Once we obtained our six date-specific data-frames, we saved them as CSV files. Our previous attribute selection and cleansing resulted in a 0.8 gigabyte CSV file with less than seven million records. For our date-specific CSV files, the smallest was 26 megabytes, while the largest was 32 megabytes; additionally, each file contained 220,000 to 260,000 records; a significant reduction in size.

Data Pre-Processing – 'Day Type' Attribute

In order to perform our analysis in R, we required the combined dataset of the date-specific datasets. However, before combining our datasets, we first created a "day_type" attribute for each dataset. This was a categorical attribute that indicated if the trip occurred during an event day, business day, or a non-business day. Since the datasets were separated by date, we could manually append a column with the corresponding day type of the date. The values of the "day_type" attribute included:

- "Event": This indicated that the trip occurred during the 12th or 13th of September, the days of the Women's and Men's Singles Finals respectively.
- "Weekday": The trip occurred during the 21st or 22nd of September, our two weekdays or business days.
- "Weekend": The trip occurred during the 26th or 27th of September, our weekend or two non-business days.

After creating the "day_type" attribute, we could combine the datasets by simply appending them to one another (since they shared the same attributes). The resultant dataset was a 155 megabyte CSV file with 1,427,323 trip records.

ANALYSIS AND ATTRIBUTE SIGNIFICANCE

Day-Type Averages

One of our main goals for this project is to determine the degree to which major sporting events such as the US Tennis Open affect overall tip proportions. To gain an initial idea of this effect, we can look at the average values for each type of day. We observe the results in *Table 1*:

Day Type	Mean Average	Median Average
Event	0.1493	0.1664
Weekday / Business Day	0.1536	0.1667
Weekend / Non-Business Day	0.1497	0.1664

Table 1: Mean & Median Averages for Different Day Types

And although we seem to observe a very slight effect on the average tip rates in both the mean and median cases, the effects contradict our hypothesis that the US Open would result in overall higher tip proportions. Instead, we see that an event day results in an overall lower tip proportion compared to both business and non-business days.

ANOVA F-Test – Significance of Day Type

We now begin to complete several statistical tests in R. We do an initial Analysis of Variance (ANOVA) F-Test to test the following hypothesis:

H_0: Effects of Day-Type are Equal H_1: Effects of Day-Type are Different

In this statistical test, we compare the variances of two different Linear Regression models. We have the null model which only contains an intercept parameter value. This is essentially a horizontal line that will predict a constant tip proportion. Then we include our 'day type' categorical variable, which will create a model which uses three different horizontal lines corresponding to the type of day (event, weekday or weekend).

The null model essentially assumes that the effects of day-type are equal; equivalent to making no change to the tip proportion that will be predicted. The other model creates different horizontal lines reflecting the effects of each day type on the resultant tip proportion prediction. The ANOVA F-Test obtains the overall variance values of each of these fitted models under the assumption of H_0 and calculates the ratio between them.

This ratio is our test statistic that is compared to the corresponding F-distribution, allowing us to determine the p-value: the probability of observing a test statistic more extreme than the current one. A smaller p-value indicates how extreme our current test statistic is under H_0; and a sufficiently extreme value (typically with p-value below 0.05) encourages rejection of the currently assumed

hypothesis H_0 since we are observing a value not 'typical' or 'normal' of what we would expect from assuming H_0.

The ANOVA F-Test R code and output are below in *Figure 4*:

```
# f-test for h0: classes have the same effect
nullmodel <- lm(tip_prop ~ 1, data = taxi)
model <- lm(tip_prop ~ day_type, data = taxi)
anova(nullmodel, model)
summary(model)
```

```
> anova(nullmodel, model)
Analysis of Variance Table

Model 1: tip_prop ~ 1
Model 2: tip_prop ~ day_type
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1 1426322 3460.9
2 1426320 3455.6  2     5.275 1088.6 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Under *Pr(>F)*, we have our p-value of *less than 2.2 * 10^{-16}*. Therefore, we have very strong evidence to reject our null hypothesis H_0. In other words, we have evidence to believe that the day type does influence the tip proportion to some degree. We further examine our model with only the day type variable in *Figure 5*:

```
> summary(model)

Call:
lm(formula = tip_prop ~ day_type, data = taxi)

Residuals:
     Min       1Q   Median       3Q      Max
-0.15364 -0.01895  0.01605  0.01734  0.15065

Coefficients:
                 Estimate Std. Error  t value Pr(>|t|)
(Intercept)     1.493e-01  7.013e-05 2129.089  < 2e-16 ***
day_typeWeekday 4.317e-03  1.014e-04   42.578  < 2e-16 ***
day_typeWeekend 4.022e-04  9.974e-05    4.033 5.51e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04922 on 1426320 degrees of freedom
Multiple R-squared:  0.001524,  Adjusted R-squared:  0.001523
F-statistic:  1089 on 2 and 1426320 DF,  p-value: < 2.2e-16
```

Figure 5

Under the '*Estimates*', we obtain the tip proportion values, or the horizontal lines of our model corresponding to each day type. The estimates are constructed such that:

- 'Event' day type corresponds to a tip proportion value of $1.493 * 10^{-1}$.
- 'Weekday' day type corresponds to a tip proportion value of $(1.493 * 10^{-1}) + (4.317 * 10^{-3})$.

- 'Weekend' day type corresponds to a tip proportion value of $(1.493 * 10^{-1}) + (4.022 * 10^{-4})$.

These are equivalent to the mean averages from *Table 1*. Again, we observe that the effects contradict our hypothesis. Instead of event day (US Open) resulting in higher tip proportions, we observe lower tip proportions.

Variable Correlation

We proceed to observe the correlations of our candidate predictor variables (listed on page 5) with the response variable in *Figure 6*, which are highlighted in yellow.

```
# correlation of attributes
cont_var <- c("trip_distance", "pickup_longitude", "pickup_latitude", "dropoff_longitude", "dropoff_latitude", "tip_prop")
cor(taxi[,cont_var])
```

```
> cor(taxi[,cont_var])
                trip_distance pickup_longitude pickup_latitude dropoff_longitude dropoff_latitude     tip_prop
trip_distance     1.00000000      0.564721690     -0.29781303       0.37697438     -0.185963416 -0.062321869
pickup_longitude  0.56472169      1.000000000     -0.11069596       0.16554620      0.053791740 -0.004434891
pickup_latitude  -0.29781303     -0.110695958      1.00000000       0.08798595      0.369265287  0.015040475
dropoff_longitude 0.37697438      0.165546201      0.08798595       1.00000000      0.136657600 -0.023053380
dropoff_latitude -0.18596342      0.053791740      0.36926529       0.13665760      1.000000000 -0.002388037
tip_prop         -0.06232187     -0.004434891      0.01504048      -0.02305338     -0.002388037  1.000000000
```

Figure 6: Correlation Matrix

Evidently, all our candidate predictor variables have very low correlation with respect to the response variable; all with a correlation magnitude less than 0.1. This already suggests that none of our continuous variables are important predictors of the tip proportion.

ANOVA F-Test – Additive Model vs. Interactive Model

Despite poor correlation values, it is still possible that our predictor variables interact with each other to better predict the response variable. Interaction occurs when variables affect one another. For our model, interaction would mean that our three fitted lines of best fit (for each day type) have different gradients. On the other hand, an additive model has no interaction between variables; resulting in three fitted lines of best fit with the same gradient (parallel) but different intercepts.

Once again, we observe the ratio of the variances generated by each model testing the hypotheses:

H_0: No Interaction (Additive Model) H_1: Interaction (Interaction Model)

In *Figure 7*, we observe a very low p-value of *less than* $2.2 * 10^{-16}$. This gives us strong evidence to reject our null hypothesis, preferring our model with interaction. The interaction between variables is captured in additional parameters that define the lines of best fit. Despite a more complex model from the extra parameters, the fitted Linear Regression model (with interaction) still ultimately consists of three straight lines of best fit; however, this time with different gradients.

```
# (a)dditive and (i)nterative models - anova test
amodel <- lm(tip_prop ~ day_type + trip_distance + pickup_longitude + pickup_latitude + dropoff_longitude + dropoff_latitude, data = taxi)
imodel <- lm(tip_prop ~ (day_type + trip_distance + pickup_longitude + pickup_latitude + dropoff_longitude + dropoff_latitude)^2, data = taxi)
anova(amodel, model)
```

```
> anova(amodel, model)
Analysis of Variance Table

Model 1: tip_prop ~ day_type + trip_distance + pickup_longitude + pickup_latitude +
    dropoff_longitude + dropoff_latitude
Model 2: tip_prop ~ day_type
   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1 1426315 3435.8
2 1426320 3455.6 -5   -19.748 1639.6 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 7: ANOVA Additive vs. Interaction

It also makes sense to fit a model with interaction simply because we have several longitude and latitude variables. Individually, these values represent lines across the Earth; but together they interact to represent point coordinates.

Model Refinement with Stepwise Variable Selection

We then proceed to further refine our interaction model. To refine the interaction model, we chose to implement stepwise variable selection starting from our full model (including all our parameters). The algorithm evaluates the current model's goodness-of-fit (how well the model fits the data). It then evaluates the goodness-of-fit of all other models that involve either one more or one less parameter. From these models, it selects the model with the best goodness-of-fit value to become the current model. It proceeds to do this until the current model has the best goodness-of-fit value.

To use stepwise variable selection, we also needed to define our goodness-of-fit measure. We chose to use the Akaike's Information Criterion (AIC). This measure is based on the likelihood of the data fitting the model. A smaller value would be indicative of a better model. In *Figure 8*, we perform the stepwise variable selection with our full interaction model.

```
# aic stepwise selection starting with full interaction model
fmodel <- step(imodel)
```

```
> fmodel <- step(imodel)
Start:  AIC=-8603378
tip_prop ~ (day_type + trip_distance + pickup_longitude + pickup_latitude +
    dropoff_longitude + dropoff_latitude)^2

                                    Df Sum of Sq    RSS      AIC
<none>                                          3424.5 -8603378
- day_type:pickup_latitude           2    0.0242 3424.5 -8603372
- day_type:dropoff_latitude          2    0.0268 3424.5 -8603371
- day_type:dropoff_longitude         2    0.1040 3424.6 -8603339
- trip_distance:pickup_latitude      1    0.1694 3424.7 -8603309
- pickup_latitude:dropoff_latitude   1    0.2115 3424.7 -8603292
- pickup_longitude:dropoff_latitude  1    0.3415 3424.8 -8603238
- pickup_longitude:pickup_latitude   1    0.3721 3424.9 -8603225
- day_type:pickup_longitude          2    0.4284 3424.9 -8603203
- dropoff_longitude:dropoff_latitude 1    0.4285 3424.9 -8603201
- pickup_latitude:dropoff_longitude  1    0.4761 3425.0 -8603182
- day_type:trip_distance             2    0.4918 3425.0 -8603177
- trip_distance:dropoff_longitude    1    0.7018 3425.2 -8603088
- trip_distance:dropoff_latitude     1    1.6450 3426.2 -8602695
- pickup_longitude:dropoff_longitude 1    2.0553 3426.6 -8602524
- trip_distance:pickup_longitude     1    3.3476 3427.9 -8601986
> summary(fmodel)
```

Figure 8: Stepwise Variable Selection on Full

To our surprise, the current full interaction model is the best model based on the AIC measure. In the first column, we see the action that can be performed. Since we start with the full model, we can either remove a parameter such as the interaction between ‘day_type’ and ‘pickup_latitude’ (day_type:pickup_latitude), or do nothing (<none>). We cannot initially add any new parameters since the current model is the full model.

Goodness of Fit – R-Squared

Although we used the AIC as our goodness-of-fit measure for the variable selection procedure, this does not give us an idea of how well our model fits the data. Instead, it can only be used to compare how well models fit the data *relative* to one another. On the other hand, R-Squared is bounded from zero to one, with a higher value indicative of a better fit.

Our final fitted model is the full interaction model. Unfortunately, our model has an R-Squared value of 0.0105; suggesting that our model fits the data very poorly. Despite an improvement from the R-Squared value of 0.001523 for our model with only the day-type variable (see *Figure 5 Adjusted R-Squared*), it is still objectively, very low. A very poor fitting model suggests that our model has very little ability to predict the tip proportion, at least for the dataset we used.

CONCLUSION & REFLECTION

Conclusion

In summary, we found that our day type attribute (and equivalently, the US Open), although it had a significant effect on the tip proportion, this effect was both very small and contradictory to our hypothesis. Instead of our hypothesised increase in tip proportions for event days, we saw a decrease.

We also explored several other attributes which were potentially strong predictors of tip proportion. These consisted of the longitude and latitude values, as well as the trip distances. From our correlation matrix in *Figure 6*, these attributes showed very little correlation with the response variable: trip proportion; suggesting our candidate predictor variables were not important predictors of the tip proportion.

Through our variable selection procedure, we obtained our locally best fitting model. The procedure did not discard any of our original variables despite the low correlation observed. However, it was evident that the resultant model was a very poor fit on the data from the very low goodness-of-fit (R-Squared) value.

If, however, the model was a good predictor of the tip proportion, further analysis could have been conducted to optimise each variable to maximise tip proportions for the taxi driver. For our current predictor variables, this could result in identifying the optimal day types, pickup locations, drop-off locations, and trip distances to maximise tip proportions; and consequently, overall profit.

Reflection

Our Linear Regression model may have been a very poor fit as a result of non-linear data. Although efforts were made to observe the relationship between the response variable and each predictor variable (using scatterplots), the sample size limited the ability to observe such relationships. The resultant scatterplots were densely packed with points appearing everywhere.

Perhaps a better approach to observing such plots would be to initially subsample the dataset, creating plots with a reasonable number of data points. This may have allowed us to observe non-linear relationships which could have then been addressed using transformations. However, the process of sub-sampling may also result in a smaller subset of data points that are unrepresentative of the entire dataset; producing misleading relationships, which is more likely when sampling a constant amount from *larger* datasets.

Additionally, it is possible that the dataset is unsuitable for a Linear Regression model. Other more suitable models for predicting the tip proportion could have been explored, at least with the attributes featured in this datasets.

RESOURCES

[1] New York City Taxi and Limousine Commission. (2019). *TLC Trip Records User Guide*. Retrieved from https://www1.nyc.gov/assets/tlc/downloads/pdf/trip_record_user_guide.pdf

[2] New York City Taxi and Limousine Commission. (2018). *Data Dictionary – Yellow Taxi Trip Records*. Retrieved from https://www1.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf