**Analysis Problem and Hypothesis**

Using taxi ride data from the city of Chicago Data Portal, can correlation be established between trip miles, trip duration, tolls, taxi fare, payment type and customers leaving a tip for their driver using multiple linear regression?

**Null Hypothesis:** Customers tipping their taxi drivers in the city of Chicago is *not* statistically correlated with trip miles, trip duration, tolls, taxi fare, or payment method.

**Alternative Hypothesis:** Customers tipping their taxi drivers in the city of Chicago is statistically correlated with trip miles, trip duration, tolls, taxi fare, or payment method.

Chicago cab companies are losing drivers to Uber and Lyft. Discovering what variables are correlated with customer’s tipping their drivers will enable taxi companies to offer bonuses to their drivers when they take on customers who do not tip. This will improve driver satisfaction and mitigate driver loss for the cab companies.

**Data Analysis Process**

Columns containing timestamp data and geolocation data were omitted from the dataset since they were not pertinent to the analysis and contained large amounts of null values. The median was imputed for the remaining null values in the dataset. Outliers were addressed by either dropping columns with the largest values or by dropping columns outside quantile ranges by flooring and capping the data. One hot encoding was used to re-express the Payment Method column into numerical values.

An initial multiple linear regression was created with Tips as the target variable and TripSeconds, TripMiles, Fare, Tolls, Extras, TripTotal, PaymentCreditCard, PaymentDIspute, PaymentMobile, PaymentNoCharge, Paymentprcard, and PaymentUnknown were set as the explanatory variables. The initial model had an adjusted R-squared of 0.994, and P-values for the explanatory variables ranged from 0.000 to 0.227. The notes section listed multicollinearity issues. The residual standard error of the initial model was calculated at 0.317.

A screenshot of a computer

Description automatically generated

Variance inflation factor was used to solve multicollinearity issues. All explanatory variables with a VIF above 5 were removed one at a time. TripTotal, Fare, and TripSeconds were dropped from the model. After solving multicollinearity issues, I eliminated explanatory variables that were not statistically significant based on P-values. PaymentDispute, PaymentUnknown, and PaymentNoCharge were dropped from the model.

The final model had TripMiles, Tolls, Extras, PaymentCreditCard, PaymentMobile, and Paymentprcard as the explanatory variables. The final model had an adjusted R-squared of 0.598 and a residual standard error of 2.580. Residual plots of the final model showed heteroskedasticity issues in TripMiles, Tolls, Extras, and PaymentCreditCard.

A screenshot of a computer

Description automatically generated

A red and blue dotted graph

Description automatically generatedA graph showing the difference between the same number of numbers

Description automatically generated with medium confidence

A graph showing different types of data

Description automatically generated with medium confidenceA graph showing the difference between a credit card and a credit card

Description automatically generated

**Study Findings**

The initial model was unreliable due to multicollinearity issues. The final model was unreliable due to the low adjusted R-squared score and signs of heteroskedasticity in the residual plots, breaking the homoscedasticity assumption of multiple linear regression. Exploratory data analysis revealed customers were more likely to tip their drivers on shorter rides with fewer miles. Exploratory data analysis also revealed customers who paid by credit card were more likely to tip (and tip higher amounts) than those who used other payment methods.

**Limitations of Techniques and Tools**

Multiple Linear regression is sensitive to outliers and variables that are correlated. It also does not handle data with wide ranges of values well. The Chicago taxi dataset included legitimate outliers that would compromise the integrity of the data if removed. Many explanatory variables, such as Extras, had wide ranges of value.

**Proposed Actions**

Although the multiple linear regression model was not accurate, action can be taken on the insights gleaned during data exploration. Chicago taxi companies can give their taxi drivers a $0.50 bonus every time they accept non-credit card payment to make up for any lack of tips.

If Chicago cab companies wish to do further analysis to see if Tips and the explanatory variables are statistically correlated, I suggest redoing the analysis and transforming the data with a log transformation during data preparation. A log transformation would coerce the data into a normal bell curve distribution. Since multiple linear regression models handle normally distributed data well, the resulting model may be more accurate.

**Expected Benefits of Study**

Offering a $0.50 bonus for drivers who accept non-credit card form of payment could increase a driver’s salary up to $200 a month, assuming they give rides to 20 customers day and the driver works five days a week. This bonus would increase driver retention and incentivize drivers to increase their daily rides, increasing revenue for Chicago cab companies.

**Sources**

No external sources were used in the creation of this executive summary.