**Data Analytics Capstone Topic Approval Form**

**Student Name:** Kimberly Hubacek

**Student ID:** 001249836

**Capstone Project Name:** Multiple Linear Regression on 2024 Chicago Taxi Data

**Project Topic**: This project will use Chicago taxi ride data to determine what variables are statistically associated with customers tipping their cab driver in Chicago in the month of January 2024.

**This project does not involve human subjects research and is exempt from WGU IRB review.**

**Research Question:** Are trip miles, trip duration, tolls and taxi fare statistically correlated with customers leaving a tip for their cab driver in the city of Chicago?

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

**Context:** In 2015, the city of Chicago granted transportation network provider licenses to all drivers operating under the rideshare companies Uber and Lyft. Ever since, Chicago cab companies have faced harsh competition from rideshare services. Not only have their customer shares declined, but many of their drivers are also quitting due to driver dissatisfaction and low pay (Knowles, 2017). This data analysis will reveal if customers who travel fewer miles are more or less likely to tip their drivers compared to customers with higher fares. To improve driver satisfaction and keep their drivers from quitting, Chicago taxi companies may choose to provide their drivers extra bonuses for taking on customer trips that statistically tip less.

**Data:** The Chicago Taxi Trip 2024 dataset will be used to generate a Multiple Linear Regression Model in this analysis. The data contains information on all reported taxicab rides in the city of Chicago during 2024. The dataset contains 425,230 rows and 23 columns. The dataset is owned by the City of Chicago’s Department of Business Affairs & Consumer Protection. The data is openly available for use on Chicago’s Data Portal (<https://data.cityofchicago.org/>). No other research has been conducted on this data.

One limitation with this data is the trip start time and trip end times have been redacted to ensure customer and driver privacy. This information would have been useful to determine if customers tip more at certain times of the day. Although the data spans a little over a month, there are over 425,000 taxi trips recorded. This will enable me to provide the most up to date customer tip data available.

**Data Gathering:** Data containing information about each taxi trip in the month of January and the first week of February 2024 is published on Chicago’s Data Portal. The dataset is available for download in a .CSV format. The data is updated monthly and is up to date as of February 7th, 2024. The data is generally clean, with some missing values in the pickup census data and drop off census data. This analysis will not utilize the pickup and drop off locations, so this is not a large concern. All the financial data including the customer fare, tolls, extras, tips, and trip total cost contain no missing values. This is the data the analysis will utilize.

**Data Analytics Tools and Techniques**: Exploratory data analysis will be performed to look at several views of the financial data, such as the statistics of the customer fare, tolls, extras, tips, and the trip total cost. After this exploration has been performed, the data will be cleaned. Missing and null values will be detected and treated, along with duplicate rows, and outliers. Zero values will also be evaluated to see if they are relevant to the data integrity. If not, they will be addressed. After the data has been cleaned, I will perform summary statistics on the dataset. Then I will perform a multiple linear regression will be used to identify if there is a relation between the customer tip amount and other variables in the dataset. The Adjusted R-Squared score of the variables in the regression models OLS Report will be used to evaluate the model. After the model has been evaluated, I will use backwards-stepwise reduction with a P-Value score threshold of 0.05 or less to select the features that will be included in my final model.

The following tools will be used during the analysis:

* Jupyter Notebook ver. 6.3 launched from the Anaconda environment.
* Python programming language ver. 3.8.8

The following Python libraries will be used during the analysis:

* Pandas
* NumPy
* Matplotlib
* Seaborn
* Statsmodels

**Justification of Tools/Techniques:**The Anaconda environment provides an all-in-one solution that allows data scientists and data analysts to access all their programming environments, languages, and tools in one convenient package. Jupyter Notebooks, as a part of that environment, allows users to easily moves back and forth between Python script and narrative output in the same place.

Python programming language is easy to read and edit due to its simple syntax and intuitive layout. Python’s readability has made it one of the most popular programming languages in the world. Python also has many modules, tools, and libraries that support machine learning and data science tasks (Payne, 2022).

The data frame will be loaded, viewed, and manipulated with Pandas. Pandas is a data frame structure in Python used for data science and machine learning projects. It has similar functions as Excel spreadsheets and SQL tables, but Pandas is much more powerful and can manage larger files (Stojiljković, 2023).

NumPy will be used to create and manipulate arrays and matrices. Complex mathematical equations can be performed quickly without using excessive resources using NumPy. The code produced from these calculations is concise and easy to read (What Is NumPy? — NumPy v1.25 Manual, n.d.).

Matplotlib and Seaborn will be used to create visualizations and graphs. Matplotlib was created off NumPy and allows users to easily create basic to highly detailed graphs quickly without using excessive computing resources. Seaborn acts as an extension to Matplotlib and offers more customization options and statistical functions (GeeksforGeeks, 2022).

Statsmodels will be used to perform the multiple linear regression and feature elimination. Statsmodels is a Python module used for running statistical tests and data explorations. The Statsmodels OLS Report provides statistical measurements of the data that will be used to evaluate the multiple linear regression model (Sutton, 2022).

Multiple linear regression demonstrates if a correlation between a single target variable and multiple explanatory variables is present. The target variable must be numerical, and the explanatory variables must also be numerical or converted to numerical data from categorical data. The target variable, tip amount, is numerical along with all the other variables that will be used in the analysis.

TheAdjusted R-squared score accurately reflects how dependable the correlation was between the target variable, tip amounts, and the explanatory variables (Potters, 2023).

Backwards Stepwise elimination is a wrapper method, meaning each grouping of variables is evaluated as one set. It is crucial to eliminate only one variable at a time and rerun the model to perform backwards stepwise elimination a second time. Dropping a variable will turn the remaining variables into a new set. The variable with the highest P-value and lowest statistical significance each round will be eliminated until all the variables are statistically significant with P-values of 0.05 or less. P-values determine if a variable has statistical significance in the model. Variables with P-Values of 0.05 or less will be retained. P-value scores range from 0.00 to 1.0. A lower P-value indicates stronger statistical significance while a larger P-Value rating indicates little or no statistical significance (Katari, 2021).

**Project Outcomes**: The project will generate a multiple linear regression model that will show if any variables are statistically correlated with customers tipping their taxi drivers in Chicago from January 2024 through early February 2024 as measured by adjusted R-Square score and P-values. At least one variable will need to be correlated with customer tipping to reject the null hypothesis.

**Projected Project End Date**: March 31st, 2024

**Sources**:

Data source -<https://data.cityofchicago.org/>

**Academic Sources:**

GeeksforGeeks. (2022). Data Visualizations in Python using Matplotlib and Seaborn. GeeksforGeeks. [https://www.geeksforgeeks.org/data-visualisation-in-python-using-matplotlib-and-seaborn/#](https://www.geeksforgeeks.org/data-visualisation-in-python-using-matplotlib-and-seaborn/)

Katari, K. (2021, December 16). Multiple Linear Regression model using Python: Machine Learning. Medium. <https://towardsdatascience.com/multiple-linear-regression-model-using-python-machine-learning-d00c78f1172a>

Knowles, J. (2017, June 2). *Cab crash: Is Chicago’s cab industry on the verge of collapse?* ABC7 Chicago. https://abc7chicago.com/chicago-taxis-cabs-taxi-medallions-lyft/2061655/

Payne, J. (2022). Benefits of Python Programming Language. Developer.com. <https://www.developer.com/languages/python/python-benefits/>

Potters, C. (2023). R-Squared vs. Adjusted R-Squared: What’s the Difference? Investopedia. https://www.investopedia.com/ask/answers/012615/whats-difference-between-rsquared-and-adjusted-rsquared.asp#:~:text=The%20most%20obvious%20difference%20between,and%20R%2Dsquared%20does%20not

[Stojiljković](https://realpython.com/pandas-dataframe/#author), M. (2023). The pandas DataFrame: Make Working With Data Delightful. realpython.com. <https://realpython.com/pandas-dataframe/>

Sutton, B. (2022). Scikit-learn vs. StatsModels: Which, why, and how? The Data Incubator. https://www.thedataincubator.com/blog/2022/12/01/scikit-learn-vs-statsmodels/

What is NumPy? — NumPy v1.25 Manual. (n.d.). <https://numpy.org/doc/stable/user/whatisnumpy.html>

**Course Instructor Signature/Date:**

The research is exempt from an IRB Review.

An IRB approval is in place (provide proof in appendix B).

Course Instructor’s Approval Status: Approved

Date: Click here to enter a date.

Reviewed by:

Comments: Click here to enter text.