Group 2: Analyzing taxi trip data to uncover patterns in trip efficiency, demand dynamics, and customer behavior

Frankie Cao - 20341005, David Fabian - 20348655,
Omar Ibrahim - 20263583, Mohamed Hirsi - 20376846

Abstract—This study investigates patterns and anomalies within New York City taxi trip data to address three key research questions: (1) identifying common pickup/drop-off routes and their correlation with congestion surcharges, (2) determining predictors of high tip amounts, and (3) detecting anomalous trips using unsupervised methods. For the first question, dimensionality reduction (PCA) and HDBSCAN clustering revealed that clusters in high-traffic areas (e.g., label 2) exhibited weak but notable associations with congestion surcharges (Cramer's V = 0.11). For the second question, a Random Forest regressor identified fare amount, trip distance, and credit card payments as critical predictors of high tips, achieving an RMSE of 3.20. Feature importance analysis underscored fare amount as the strongest driver of tipping behavior. The third question employed an Isolation Forest model combined with lagged cost-per-mile features to detect anomalies, flagging irregularly high-cost or long-distance trips, though data sparsity and unlabeled anomalies posed challenges. Key contributions include a clustering framework for route analysis, a feature importance methodology for tipping behavior, and an anomaly detection pipeline for unlabeled taxi data. The study utilized the NYC Taxi dataset, preprocessed to handle missing values and outliers. Future work could integrate real-time traffic data and deep learning models to enhance prediction accuracy and anomaly detection granularity.

Index Terms—Dimensionality reduction, HDBSCAN clustering, Random Forest Regressor, Anomaly detection, Isolation Forest

1 Introduction

Taxi travel is a very common form of transportation in New York City. From looking at the number of data points in our dataset we can see that there are tens of millions of taxi rides happening over New York annually. The motivation of our problem is the belief that by identifying key features in the patterns of civilian transportation it is possible for a city to reduce traffic congestion and increase rider satisfaction. It is also possible to create a robust system to identify fraud in order to protect consumers and increase their confidence in the city transportation system.

We looked specifically at what are common pickup and drop off locations, indicators of high tip amounts, and anomaly detection. This was done through a variety of techniques such as HDBScan clustering, random forests, and isolation forests. For the common trip routes, we found that most of the trips can be clustered in roughly 3-4 clusters. For tip amounts, we found that the most indicative feature to determine high tip amounts is the fare amount and this makes sense because people generally tip based on a percentage of the total price. For the anomaly detection our progress was hindered by the quality of the data and found that the methodology of how this data is being collected could improve.

Overall, our methods provided a potential framework for others to build on in order to improve the transportation system and experience in New York. These ideas and methods can be further extended to apply to other major metropolitan areas that can be found in areas like Toronto or China.

2 Related Work

Omar: The paper "A Unified Neural Network Approach for Estimating Travel Time and Distance for a Taxi Trip" by Ishan Jindal et al. [3] discusses the problem of taxi travel time prediction and distance by using deep neural networks trained on the NYC taxi dataset. It utilizes the ST-NN (Spatio-Temporal Neural Network), which predicts both the travel time and distance at the same time. This research paper is directly related to our first research question since model's travel time estimation can be used for analyzing how common routes correlate with congestion surcharges; if a route between two locations takes unusually long, then that is a strong sign that there was high congestion during the trip.

David: The paper "Improving Prediction for taxi demand by using Machine Learning", by Mustafa Mahmoud Ibrahim and Foad Salem Mubarek [4], aims to improve the prediction of taxi demand. The research questions addressed are whether taxi demand prediction can be improved through data preprocessing, and which machine learning models perform best for taxi demand prediction. The paper is highly relevant since it evaluates and compares

- *Member 1, 2, 3, and 4 are with School of Computing at Queen's University E-mails: 21fc17@queensu.ca, 21dgf4@queensu.ca, 20omha@queensu.ca, 22xlb@queensu.ca*

4 distinct machine learning models (Linear Regression, RandomForest, ANN and LSTM) on taxi data. This gives us a better understanding of which models we could use for our own research questions since we are also using taxi data.

Mohamed: The paper *"Tippers and Stiffers: An Analysis of Tipping Behavior in Taxi Trips"* by David Elliott, Marcello Tomasini, Marcos Oliveira, and Ronaldo Menezes [1], examines tipping behavior in NYC taxi rides using a large-scale dataset. The research questions addressed include how tipping behavior correlates with socio-economic factors, whether tipping habits remain consistent across different times, and what factors influence whether a passenger tips. The paper is relevant since it empirically analyzes tipping patterns using statistical methods and large-scale data. This helps us understand id entify tipping behavior trends.

Frankie: The paper "Real-time taxi spatial anomaly detection based on vehicle trajectory prediction" By: Wenyan Hu, et al. [2] provides a proposed framework for handling anomaly detection for taxi data. This is done through a suggested architecture that they created called "TAPS" for "Taxi Anomaly detection framework on the basis of vehicle trajectory Prediction to detect taxi anomalous trajectories in the Spatial dimension". Their methodology is essentially an encoder and decoder model that uses LSTM to predict the expected paths which gives us a potential methodology to model our approach for the third research question that deals with anomaly detection.

3 Methodology

3.1 Research Question 1

In research question 1, we wanted to determine what the most common pickup/drop-off location pairs, and how these routes correlate with congestion surcharge frequency. To achieve this, we used dimensionality reduction, HDBSCAN clustering, and performed Cramer's V tests to see how strongly the different clusters were associated with the congestion surcharge variable.

First, we needed to obtain the coordinates of the pickup and drop-off location pairs. This was done using the taxi lookup table csv file which contained a dictionary of pickup/drop-off location IDs and their corresponding location names (e.g. location ID 1 is Newark Airport). Then, we used the Nominatim geocoder to get the longitudes and latitudes of each pickup/drop-off location pair. The utm library in python was used to turn the longitudes and latitudes into cartesian coordinates, which were then normalized using the Scikit-Learn StandardScaler class, and outliers were removed.

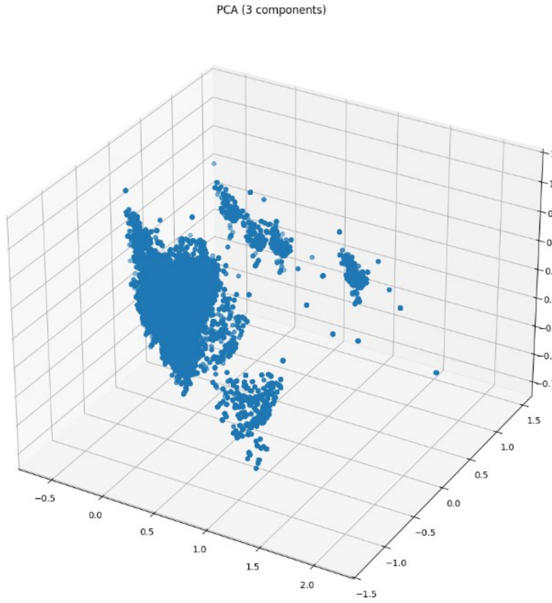

Fig. 1. Plot of 3D PCA reduced data

Since the data is 4D, a dimensionality reduction technique would be needed to not only visualize it, but to also improve the clustering. Although 2 PCA components explained around 85% of the data's variability, the outputted plot did not show any discernible clusters. We also tried UMAP, but it was both computationally expensive and time-consuming. We decided on doing PCA with 3 components (4D to 3D) due to its relatively low computational cost and the fact that the outputted plot displayed clear clusters.

Next, we needed to cluster the data to see which trips are similar to each other. K-means clustering was fast, but the results were not satisfactory, so we used HDBSCAN clustering which takes longer than K-means, but is more efficient for complex data and irregularly shaped clusters.

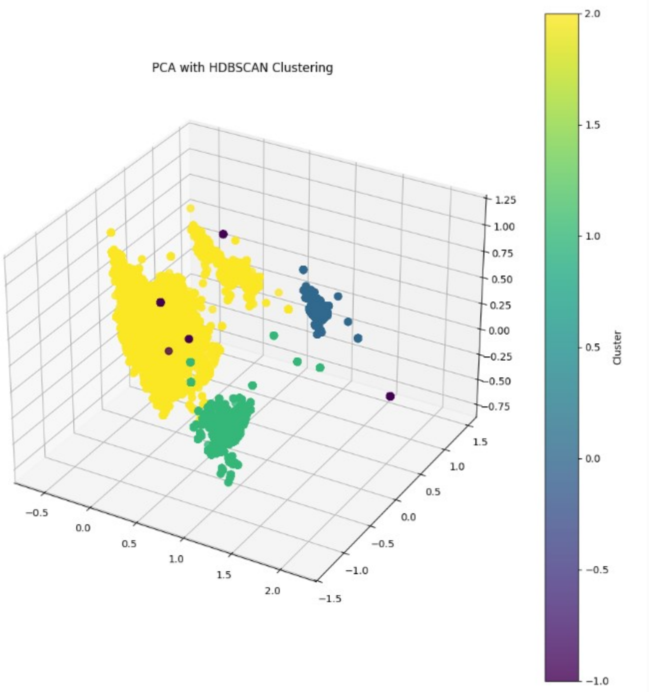

Fig. 2. Plot of HDBSCAN clusters, -1 is for outliers

Using these new labels for the data points, we performed Cramer's V tests to see how strongly each cluster was associated with the congestion surcharge.

3.2 Research Question 2

In Research Question 2, we wanted to determine what were key predictors of high tip amounts.

First, we used feature engineering and data processing to enhance the dataset. After removing rows with missing/erroneous values, we removed columns that weren't helpful in predicting the tip amount (e.g. total amount variable which already includes the tip amount) and one hot encoded the nominal category variables (payment type, RateCodeID, etc.). Then, correlation analysis was done to see which variables were highly correlated with the tip amount.

We needed to further determine which variables were the most important in predicting the tip amount, so we decided to train a Random Forest regressor. Random Forests provide an easy way to assess feature importance by using either an impurity-based method or a feature permutation method. Also, the paper ”Improving Prediction for taxi demand by using Machine Learning”, by Mustafa Mahmoud Ibrahim and Foad Salem Mubarek trains and compares a Linear regression model, Random Forest, ANN and LSTM on taxi data; the results show that the Random Forest model performed the best. We used 80% of the data to train the model and the remaining 20% for testing.

Selecting appropriate hyperparameters is critical for building effective machine learning models. Initially, we considered using GridSearchCV to find the best combination of hyperparameters in a predefined grid; however, this evaluates every single hyperparameter combination, which is very computationally and time expensive. Instead, we decided to use RandomSearchCV which evaluates a fixed number of random hyperparameter combinations from a given range or distribution, greatly reducing the number of computations while still finding high-performing models.

After RandomSearchCV was finished, we selected the best model and found the feature importances using Scikit-Learn's feature_importance_ property that all Random Forests have; this uses Mean Decrease Impurity (MDI) to determine which features are most relevant in predicting the target variable.

3.3 Research Question 3

In research question 3, we wanted to look at the data and try to identify anomalous trips. One of the major challenges of anomaly detection is that the nature of anomalies (they are rare) and in this data they are not labeled which means that we need to use an unsupervised approach.

We considered different approaches such as the TAPS architecture outlined in the research paper ”Real-time taxi spatial anomaly detection based on vehicle trajectory prediction” By: Wenyan Hu, et al. [2] and using the DeepOD [5] library's built-in anomaly detection models. The TAPs architecture relied on having specific data for a trip and analyzing the trip to see if it deviated from an expected path. This was done by the implementation of a RNN using LSTM. The data points that we were given only give the pick up and drop off locations which means that there was insufficient information to detect the fine grain details of potential anomalies that can be done using TAPS.

We considering implementing a library from the anomaly detection section on paperswithcode.com and leaned towards DeepOD because it contained many models for different approaches to unsupervised anomaly detection. We looked at the architectures that they used and came to the conclusion that for our specific use case it would not be effective because we would still need to implement mechanisms for determining which paths are close to one another. Hence, we chose to do a pipeline consisting of approximate nearest neighbors to figure out which routes were similar to each other and then using an isolation forest because it is an unsupervised technique that is capable of finding anomalies in time-series data.

Compared to the other research questions, the data prepossessing and cleaning was a little different for this step. We wanted to remove clear inconsistencies such as negative values but we were reluctant to remove sparse values because they would constitute as anomalies in some cases. Hence, for some instances it was difficult to consider what we wanted to remove and what we wanted to keep the same. For example, since our approach involved treating the data as a time series we opted to truncate the data so that the data points fall within the allotted time frame. However, for some other value like negatives instead of removing them we took their absolute value so some clerical errors would be mitigated while maintaining the characteristics of some anomalies.

Afterwards, we had to decide what features to use. When deciding which features to use, we needed to keep in mind that depending on what we are checking different things will be considered anomalies. For example, if we look at all the data, a vast majority of the trips will have a lower travel length which will cause the longer trips to be labeled as anomalies. Although this can be considered an outlier it is hard to tell if this is necessarily an anomaly. If this was during rush hour and the longer trips tend to be in the morning then it could be considered an anomaly but without adequate context it is hard to give a definitive answer. Hence, we opted for trying different features for our model; one that uses everything and one that focuses on the cost per mile.

We modeled the trips as time-dependent events (i.e. in the morning the trips around the same area should have similar data. To do this we first sorted the values by their pick up time then used an Approximate Nearest Neighbors algorithm to put similar routes close to each other. Due to the computational intensity of running the Approximate Nearest Neighbors algorithm we were forced to reduce the number of data points to 10 thousand since we could not load all the millions of points into memory. Then finally building and training the model was done by using scikit-learn's isolation forest algorithm.

4 Dataset

The dataset is large in size since it spans many years and is given in great detail. Hence, we will try to use as much of it as possible, and if certain techniques can not be computed in a reasonable amount of time, we will reduce the scope of the project.

The data is collected by the NYC taxi and limousine commission (TLC), specifically the Taxicab & Livery Passenger Enhancement Programs (TPEP/LPEP). It also includes for-hire vehicle trips, with submissions made by bases.

The TLC issues a disclaimer indicating the data may be incomplete or somewhat inaccurate, especially with regards to the for-hire vehicles.

The key features of each trip is as follows:

- VendorID: Identifier for the taxi service vendor.
- tpep_pickup_datetime: Timestamp when the ride starts.
- tpep_dropoff_datetime: Timestamp when the ride ends.
- passenger_count: Number of passengers in the ride.
- trip_distance: Distance of the trip.
- RatecodeID: The rate code used for fare calculation.
- PULocationID: Pickup location identifier.
- DOLocationID: Drop-off location identifier.
- payment_type: The method of payment used (cash, credit card, etc.).
- fare_amount: Base fare for the trip.
- extra: Additional charges (e.g., for time or distance).
- mta_tax: A special tax applied in New York City.
- tip_amount: The tip given by the passenger.
- tolls_amount: The total tolls incurred during the trip.
- improvement_surcharge: A surcharge applied for service improvements.
- total_amount: The total fare, including all charges and tips.
- congestion_surcharge: A surcharge applied during times of congestion or high demand.
- Airport_fee: A fee for trips to or from an airport.

The main features the project will focus on are PULocationID, DOLocationID, congestion surcharge, tpep pickup datetime, tpep dropoff datetime, trip distance, and payment type. PULocationID, DOLocationID, and congestion surcharge will be used for the first research question in order to analyze which routes correlate with the highest surcharges, tip amount will mainly be used for the second research question in order to analyze high tip amounts, but every other feature in the dataset will be considered here as well. For the third research question, the fare-related features will be considered the most, since these are the most relevant in determining fraudulent activity. Namely, these are: payment type, fare amount, extra, mta tax, tolls amount, improvement surcharge, congestion surcharge, Airport fee, and total amount.

5 Experiments and Results

5.1 Research Question 1

The cluster of trips with the strongest association was label 2 with a Cramer's V of 0.11, meaning trips belonging to this cluster are most affected by the congestion surcharge fee. Label 2 is the largest cluster, so this makes sense since there is a higher number of trips in a similar area. However, a Cramer's V of 0.11 is still relatively low so cluster labeling alone is not sufficient in determining the congestion surcharge. Trips belonging to label 0 and label 1 are less associated than label 2 with Cramer's V of 0.078 and 0.073 respectively. Trips belonging to label -1 (outliers) have negligible association, which is expected since they are unusual trips meaning they are less likely to be affected by congestion.

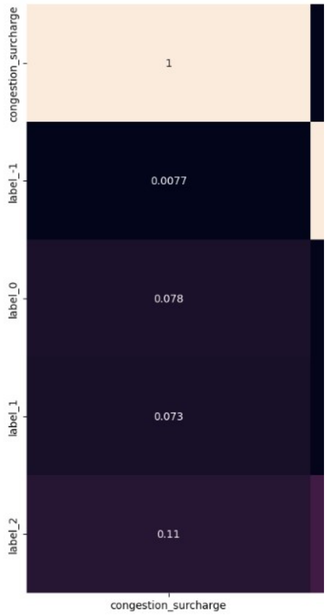

Fig. 3. Heatmap of the associations between the labels and the congestion surcharge

The main threat to the validity of our findings for research question 1 is our choice of hyperparameters for the HDBSCAN clustering algorithm. For example, we chose a minimum cluster size of 10,000 since we had more than 6 million data points; choosing a smaller number for this would result in a higher number of clusters, meaning the output of the Cramer's V test would be different.

5.2 Research Question 2

The correlation analysis showed that the variables with the highest correlations with tip amount were tolls amount, with a correlation of 0.45; payment type 1 (credit card) with a correlation of 0.4; Ratecode1D 2.0 (JFK rate), with a correlation of 0.35; Ratecode1D 1.0 (standard rate) with a correlation of -0.36; and payment type 2 (cash) with a correlation of -0.39.

After training, the Random Forest Regressor achieved an RMSE of 3.20; Mean Decrease Impurity showed that the fare amount was the primary indicator of a high tip amount. This makes sense because in general people tip based on a percentage of the total amount of fare, which is usually given as predetermined options on credit card readers. Other key features were the trip distance (the longer the trip is, the more likely people are to tip higher amounts), and payment_type_1 (credit card) meaning that if the customer is using a credit card, then there is a higher chance that they will also tip.

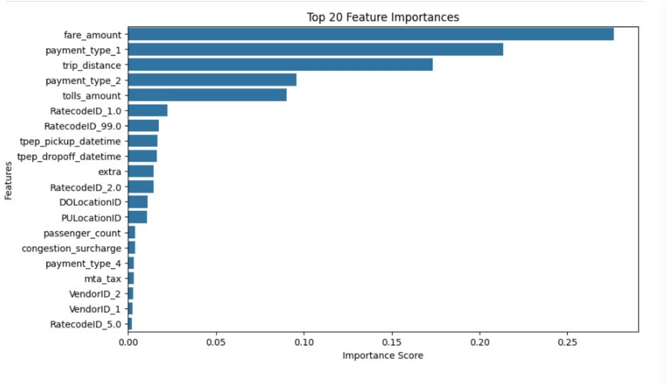

Fig. 4. Random Forest feature importance for predicting the tip amount

The main threat to the validity of our findings for research question 2 is the fact that RandomSearchCV tries random combinations of hyperparameters, its possible that the best combination it finds may still be suboptimal. Another threat to validity is the possibility that the Random Forest model overfits the data, leading to poor testing results. One hot encoding introduces sparsity which can increase the risk of overfitting.

5.3 Research Question 3

It is hard to quantify the performance of our model because the accuracies of its predictions depend on the parameters you are considering. As mentioned in the methodology portion, we performed the anomaly detection on a lag column comparing the previous cost per mile and rolling average. For the baseline approach the isolation forest was trained on all the data present except the location IDs because they are not necessary. Since it was difficult to evaluate the performance of anomaly detection numerically, to analyze them we just looked at the graphs.

It is expected that the total cost and cost per mile of all the data points should be close to each other (i.e. towards the left side of the graph). As can be seen from the two figures below, this is better captured when using the lag column (figure 5) because the primary feature that it is taking into account is the cost per mile. On the contrary, when using all the data, to run the model (figure 6) we can see that it fails to capture some of anomalous cost per mile data points where the rides had a cost per mile of over 400 and it classifies a majority of the points with a total cost over 60 as anomalies. Thus, for this specific instance when detecting anomalies for the cost per mile, using just the lag column is more accurate than all the data.

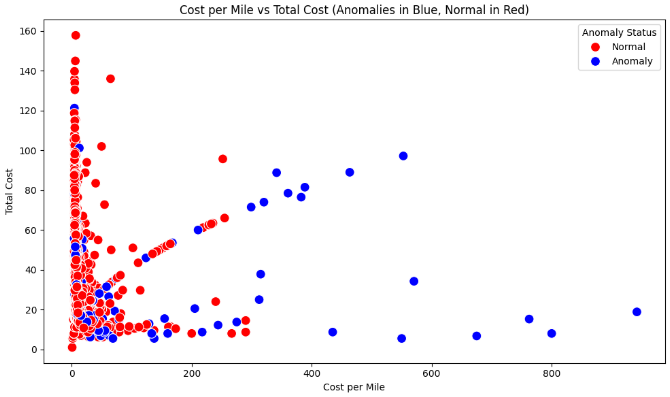

Fig. 5. Cost per mile vs Total Cost using the lag column

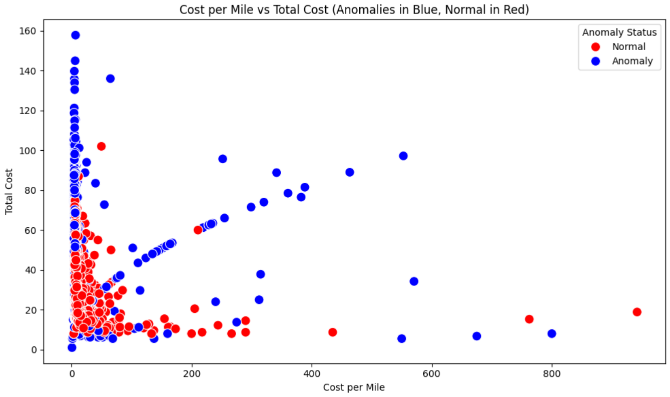

Fig. 6. Cost per mile vs Total Cost using all the data

But, if we instead graph the trip distance against the total cost, we will notice that the clear difference in how the anomalies are spread. As can be seen from the figures below, when using all the data the trips with larger distances are automatically flagged as anomalies. This makes sense when considering the contents of the data. A large majority the trips are short distance trips hence, all things considered a long trip will constitute as an anomaly despite the number being reliable.

The major limitation that could effect the validity of our results is, as mentioned before, it is not necessarily clear ”what is an anomaly” since it is not labeled. Also, due to resource constraints we were not able to train on a large number of data points which means that our model may not be able to generalize features very well. Similarly mentioned, the data we had to work with was very messy and was prone to various inconsistencies of many different form. This is the reason there is a straight line in the cost per mile vs total cost graphs from when we adjusted some of the missing values using the average cost per mile.

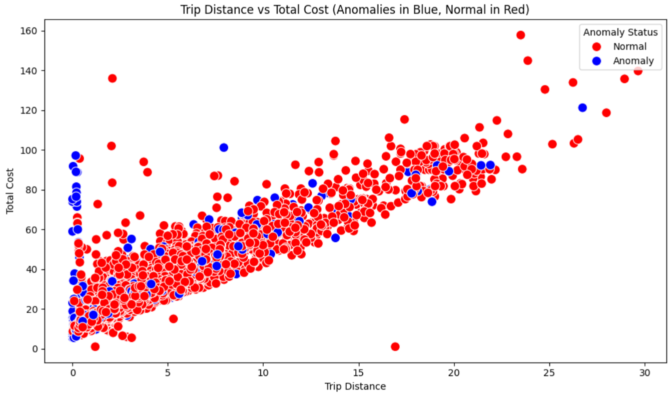

Fig. 7. Trip Distance vs Total Cost using the lag column

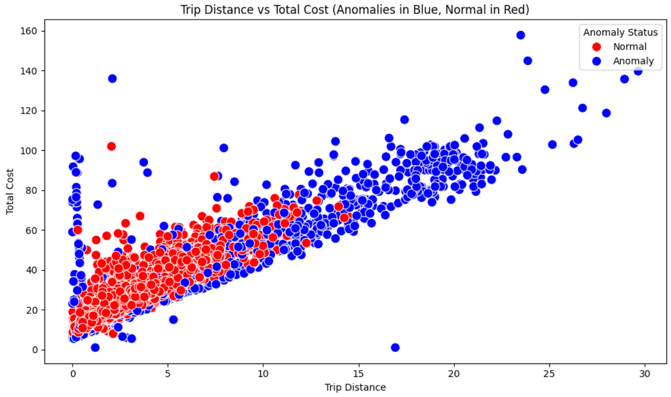

Fig. 8. Trip Distance vs Total Cost using all the data

6 Group Member Contributions

Omar: Data analysis/cleaning for research question 2, worked on research question 3 and presentation

David: Random forest model building/training and feature analysis for research question 2, worked on research question 1

Frankie: Random forest model building/training and feature analysis for research question 2, worked on research question 3

Mohamed: Data analysis/cleaning for research question 2, worked on research question 1

7 Replication Package

https://github.com/Omaro-IB/taxi-data-analysis/

8 Conclusion and Future Work

This study analyzed New York City taxi trip data to address three critical questions regarding route patterns, tipping behavior, and anomalous trips. For the first research question, dimensionality reduction (PCA) combined with HDBSCAN clustering identified high-traffic pickup/drop-off clusters, revealing a weak but notable association between cluster membership and congestion surcharges (Cramer's V up to 0.11). This suggests that routes in densely populated areas are more susceptible to congestion fees, though there is still a need for complementary data like real-time traffic conditions to strengthen these insights.

For the second question, a Random Forest regressor demonstrated that fare amount, trip distance, and credit card payments are key predictors of high tip amounts, achieving an RMSE of 3.20. The prominence of fare amount in feature importance highlights the practice of percentage-based tips, aligning with socioeconomic norms. These findings provide actionable insights for drivers and ride-hailing platforms to optimize service offerings and payment interfaces.

The third research question employed an Isolation Forest model with lagged cost-per-mile features to detect anomalies, successfully flagging irregularly expensive or lengthy trips. However, challenges such as data sparsity, computational constraints, and the lack of labeled anomalies limited the granularity of results. Visual analysis confirmed the model's ability to identify extreme outliers, though distinguishing genuine anomalies from rare but legitimate trips remains an open problem.

Future work could enhance these findings by integrating real-time traffic and socioeconomic data to refine congestion and tipping models. Deep learning architectures, such as LSTMs for trajectory prediction, could improve anomaly detection accuracy, while addressing data sparsity through sampling strategies or distributed computing. Validating results with labeled anomaly datasets and exploring causal relationships between features (e.g., payment type and tipping culture) would further strengthen the robustness of insights. Ultimately, this study is beneficial in aiding the decision making of Urban transportation agencies, city planners, taxi and ride-hailing companies, taxi drivers, data analysts, and regulatory bodies.

References

[1] M. O. R. M. David Elliott, Marcello Tomasini. Tippers and stiffers: An analysis of tipping behavior in taxi trips. In *2017 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computed, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pages 1–8. IEEE, 2017.

[2] W. Hu, M. Li, M.-P. Kwan, H. Luo, and B. Chen. Real-time taxi spatial anomaly detection based on vehicle trajectory prediction. *Travel Behaviour and Society*, 34, 2024.

[3] X. C. M. N. J. Y. Ishan Jindal, Zhiwei (Tony) Qin. A unified neural network approach for estimating travel time and distance for a taxi trip. 2017.

[4] F. S. M. Mustafa Mahmoud Ibrahim. Improving prediction for taxi demand by using machine learning. In *2023 15th International Conference on Developments in eSystems Engineering (DeSE)*, pages 451–456. IEEE, 2023.

[5] X. Zuo. Deepod: A deep learning framework for outlier detection. https://github.com/xuhongzuo/DeepOD, 2020. Accessed: 2025-04-15.