Proposal: Analysis of cab services in NYC

1 Introduction

In recent years, a noticeable shift in transportation preferences has emerged within metropolitan areas, with cities like New York serving as a prominent example. Despite the extensive and well-established subway system, an increasing number of commuters are opting for cab services. This trend raises a crucial question: What are the driving factors behind this preference for cabs over the highly accessible subway network? This inquiry forms the cornerstone of our research endeavor, aimed at unearthing the underlying motivations and socio-economic dynamics that have contributed to this significant shift in transportation habits.

For understanding the trends and shift in the preference of the commuter, in metropolitan areas, particularly cities like New York, a significant challenge lies in addressing the socio-economic dynamics that underlie this transportation shift. To uncover the motivations driving this change, we must consider factors such as income levels, employment patterns, and urban demographics. This necessitates the examination of data on household incomes, employment rates, and the geographical distribution of commuters. Furthermore, it is imperative to understand the specific needs and preferences of various demographic groups, including the working class, students, and tourists. Additionally, we must account for external factors, such as changes in traffic conditions, weather, crime records, and special events, as they can influence transportation choices.

In our pursuit to comprehend the reasons behind the increasing preference for cab services over the established subway system, we must acknowledge the complexities and challenges inherent in conducting this analysis. One of the primary challenges we encounter pertains to data handling. To comprehensively examine the shift in transportation preferences, we must collect and integrate data from diverse sources, often contending with differences in data formats, quality, and granularity. Ensuring the accuracy and consistency of this data is paramount to derive meaningful insights.

Our investigation is centered on a comprehensive analysis of this evolving phenomenon, offering a nuanced understanding of the contemporary urban commuter's choices. To achieve our objective, we have structured our approach to perform a comparative analysis of the adoption of public transportation modes, such as buses and subways, and private transportation options, including FHVs, yellow taxis, Uber, and Lyft, among others. This analysis will encompass various analytical components, including safety measures, population growth, and congestion in busy areas, among other factors. [7.][5.]

The subsequent sections will expound on our proposed analysis and methodology, detailing the datasets we will employ to achieve our research objectives. The document is structured as follows: **Proposed Analysis and Methods, Dataset Description, and References.**

2 Dataset Description

- **NYC TLC Dataset:** The taxi dataset used in this project is yellow taxi trip data covering the year 2018, which records attributes such as pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts. Yellow taxis are taxis that are allowed to respond to streets hailed from a passenger in all five boroughs. [7.]

- **Taxi Zone Dataset:** The pick-up and drop-off locations are populated by numbers ranging from 1 to 263. These numbers correspond to taxi zones also contains geometric information of each taxi zone and a list of TLC taxi zone location IDs, location names, and corresponding boroughs of each zone. [8.]

- **NYC Weather Dataset:** Weather data, sourced from the National Centers for Environmental Information

(NCEI), provides daily observations for weather in NYC, including temperature, precipitation, snowfall, and wind speed. This data is used to study how weather conditions impact taxi usage in NYC. [3.]

- **NYPD Crime Statistics Data:** We will utilize a pivotal dataset from NYPD crime statistics, encompassing crimes in subway stations, subway cars, green taxis, yellow taxis, and FHVs. This data is central to our study of crime and transportation trends in New York City. [1.]

3 Proposed Method

Our project proposal centers on the comprehensive comparative study of two predominant modes of transportation within the context of New York City: taxis (cabs) and public transportation (subways). Our primary objective is to conduct an in-depth analysis that delves into various factors, thereby providing a comprehensive understanding of the dynamics underpinning the transportation choices made by New York City residents and visitors.

The project plan includes a series of proposed analyses, with the aim of formulating statements and hypotheses regarding transportation preferences in the city. Furthermore, our investigation seeks to identify patterns that may be of particular interest to cab drivers, specifically pertaining to tipping behavior. To accomplish this, we propose a range of analyses, including the development of a statistical model designed to predict tips based on various trip parameters, which will be detailed in the subsequent sections. These analyses are intended to shed light on the factors influencing people's preference for cab services.

- **Comparative Analysis**: In pursuit of our research objectives, we propose a detailed analysis that entails a comparative examination of cab and subway services. This investigation will scrutinize the influence of diverse factors, encompassing weather conditions, special events, and disruptions in public transportation services, on ridership patterns. Through this analysis, we aim to elucidate the intricate dynamics of urban transportation choices within the context of New York City. The primary components of this analysis are structured into the following four parts:

1. **Geospatial Analysis:** Our analysis examines pickup and drop-off patterns of yellow and green taxis to confirm regulatory distinctions. Using data from The New York City Taxi and Limousine Commission (TLC), we gain insights into transportation trends and spatial dynamics. [8.][7.]
2. **Demographic Mapping:** Incorporating demographic data, encompassing population, age groups, gender, income, and education from sources like the census, this component allows us to represent the distribution of city residents across neighborhoods. By scrutinizing spatial distribution, we can identify trends in the residential patterns of individuals with varying economic backgrounds and age groups. This data forms a crucial foundation for contrasting taxi service preferences in different geographical areas, unveiling correlations between specific neighborhoods and taxi pick-up and drop-off patterns. [2.][8.]
3. **Weather and special events**: Weather and special events wield a notable influence on cab users' transportation choices. Inclement weather often leads people to favor cabs for their shelter and convenience. Likewise, special events heighten the demand for cab services due to increased traffic and limited public transportation options. These factors play a pivotal role in cab users' decision-making, impacting their preference over the subway. To delve into this dynamic, we'll utilize a Kaggle dataset with Central Park weather records, providing essential insights into how weather and special events shape people's travel preferences. [3.]
4. **Safety Analysis** : In New York City, residents often prefer taxis to subways due to concerns about potential crime incidents. Taxis provide a reassuring and less stressful way to explore the city, fostering a sense of security and comfort among passengers. The increase in subway crime has led commuters to seek safer transportation options. Our initial analysis involves examining crime data and safety perceptions

in various neighborhoods and areas to assess the significance of safety concerns. We plan to utilize the NYPD crime statistics dataset, which includes crimes in subway stations, subway cars, as well as incidents in green taxis, yellow taxis, and FHVs. [1.][8.]

- **Statistical Modelling** : We aim to understand the traffic patterns and evolving preferences of urban commuters in New York City. Additionally, we aim to provide cab drivers with insights to enhance their trip profitability. To achieve this, we propose building a statistical model that predicts the tips given by passengers. We will consider various trip parameters, such as time, location, drop-off, and pick-up, among others. As our foundation, we will employ Linear Regression, and we will further enhance our model with advanced algorithms like XGBoost and Random Forest. Our approach begins with an initial analysis to explore the correlations between these parameters, with the intention of subsequently incorporating additional variables to refine and expand our model.

  In order to assess the effectiveness of our models, we will employ predefined metrics outlined in Table 1.

Models	Accuracy	Precision	RMSE	Other
LinearRegression				
XGBoost				
RandomForest				

Table 1: Evaluation

4 References

1. https://data.cityofnewyork.us/Public-Safety/Harassment-on-NYC-Subway/6zpn-pdex
2. https://www.nyc.gov/assets/planning/download/pdf/data-maps/nyc-economy/employment-patterns-nyc.pdf
3. https://www.kaggle.com/datasets/ecboxer/nyc-weather
4. https://www.researchgate.net/publication/335504977_Exploring_the_Taxi_and_Uber_Demand_in_New_York_City_An_Empirical_Analysis_and_Spatial_Modeling
5. https://medium.com/@haonanzhong/new-york-city-taxi-data-analysis-286e08b174a1
6. https://www.nyc.gov/assets/tlc/downloads/pdf/fhv_congestion_study_report.pdf
7. https://catalog.data.gov/dataset?q=High+Volume+FHV+trip+records&sort=score+desc
8. https://data.cityofnewyork.us/Transportation/NYC-Taxi-Zones/d3c5-ddgc