Database Management Systems
COP 5725
Fall 2022
(Instructor – Dr. Markus Schneider)

Project Deliverable 1
Group 10

Title: New York City Taxi Trend Analysis

Team Members

Budham, Keanu	“kbudham@ufl.edu”
Thumsi, Bhushan	“thumsib@ufl.edu”
Vyas, Dishant	“vyasdishant@ufl.edu”
Rangaraju, Deepakraju	“d.rangaraju@ufl.edu”

Table of Contents

1.) Introduction

1.1) Overview

With the introduction of the Internet in the 80s, the amount of information available has increased tremendously. This led to an increasing need to store and organize data. This saw the rise of Databases and Database Management Systems(DBSM), which allowed the user to extract data from that stored in a database. In a study conducted, it was discovered that 5 exabytes of data had been created at Google by 2003. By the end of 2010, the same amount of data was being created every 2 days and by 2022, it is being believed to be created every 30 minutes. The total amount of data being captured by industries is doubled every 1.2 years, but only an approx of 0.5% of data created is ever used or analyzed. We can easily see the need for a more reliable way of storing and analyzing the data available to us. Database system overcomes all the shortcomings and helps us better utilize the enormous data we have access to.

1.2) Description of the application

With the introduction of mobile based taxis like Uber and Lyft, Yellow Cabs in NYC are facing a downward trend in terms of number of trips and passengers. Covid'19 didn't help their case, as passengers looking for safer options preferred mobile based taxi services. Our project is meant to help the drivers as well as passengers decide if operating or taking a taxi is the best possible option from location A to location B at specific or random time intervals or particular periods across a calendar year. The user has the flexibility to choose the parameters like time, location, payment method, trip fare etc. in order to make an informed decision whether it be a taxi driver, a tourist or tech giants like Google Pay looking for their user demographic.

We are planning to build a user interface which would allow various users to see and analyze trends from data consisting of records from early 2005 to 2022.

2.) Database needs of the application and the potential user interest

2.1) Motivation of the database needs of the application

The data about Yellow Cabs in NYC is easily available on the internet. But there is no computerized application which acts as a one-stop for everything related to trip data based on Yellow Cabs. Building an effective system that could manage such data requires significant time and investment. For an average user, looking at the Excel files is counterproductive as they can see what the data consists of, but deriving useful conclusions for over a million of records requires a useful tool that can comprehend, visualize data and give users a reliable architecture that could be accessed from anywhere at anytime. Our goal is to fill this gap and develop an application based on a database management system which is not just storing and retrieving data, but also represent the data in order to generate more effective trends.

The dataset consists of over 20 attributes and over 2+ millions records every month, which gives us ample data to derive trends from and show a graphical representation of that particular trend, enabling users to better visualize the vast amount of data. Hence, a Database System that can handle this data requires performing millions of computations to generate results.

2.2.) Potential User interest in the application

Taxi Drivers:

- There are approximately 1,80,000 NYC drivers. This App provides the trend of passengers in the requested interval at the nearest areas where the taxi driver can park his fleet to maximize his pickups. This query can maximize the profit of taxi drivers by reducing their waiting time. From the trend, they can decide whether the requested interval will yield them a sufficient amount of profit or not, and will be very conducive to decision-making. Uncertainty virtually results in the opposite, so we ensure that the results we yield are proved with probabilistic certainty through past trends. The chances of the trends falling opposite are minimal, as we take years of data that is worth more than millions of points. The interactive dashboard also provides the driver to see the patterns of the price and trips on special days or seasons through time.

Epidemiologists:

- The app helps the analysts to view the trends of the passengers before and after COVID-19 to get the influence of taxis on the pandemic. This trend would have helped them to reduce the burden of this vicious disease from permeating the environment.

Analysts:

- Cashless payments became common across metropolitan areas. This app helps us find the trend of how the payment mode has changed over time across various areas. From the data, we can find areas where people use cash payments across different clusters in New York City. This will help the bank identify its target people and shift them to use cashless payments.

- There is a strict decline in Yellow Cab due to the rise of Uber and Lyft. This application provides a solid trend that shows the decline of yellow cabs with the growth of mobility services like Uber and Lyft. This will be helpful to find features like price, the flexibility of timings, etc. that have ensued the downfall.

3.) Description of the web based User Interface Functionality

For the user to interact with our database, there will need to be a front end user interface (UI) which the user can use. For our applications, we will be going with a graphical user interface (GUI) with the idea that even a layman can easily use the application and view meaningful information. Our goal is to create a simple and straightforward design where the user gets a graph as an out that shows the change over a time of particular parameters, based on their input.

We will have different parameters based on the queries. When the user clicks on a particular query, the parameters with respect to time will be displayed on the screen. The user would also have different dropdowns and filters based on the parameters mentioned earlier, like fare, location or even tip amount. These user specified parameters correspond to the queries we make and the graphs we then display to the user. We display the change in these parameters over time.

For example, for a query related to the effects of COVID-19 on the taxi industry, the user would click on a button for that query, and it would show parameters focussing on the time period that was under the effects of COVID. This would include parameters, like total average customers per day, average passenger count, etc. From there, with a drop-down checkbox, the user could add or delete parameters from the graph or “zoom out” and look at a further stretch of time. Basically, each query would correspond to a button push with certain parameters being displayed on the graph, and that graph could be further customized through user inputted filters.

4.) Description of the application goals regarding trend analysis

The goals that this application intends to accomplish through trend analysis are catered to both the drivers and passengers of NYC taxicabs as well as data analysts.

Taxicab Drivers: this application will in determining the best ways to maximize profit by analyzing past trends in the taxi records from observations such as when would be the best day and time to do trips in a certain area of NYC to attain the most passengers.

Taxicab Passengers: This application will help passengers determine how to make their trip cost is minimized dependent on the day, time, and place through trend analysis of various factors from the taxi records.

Data Analysts: This application will display complex trends that data analysts can use for various purposes. For instance, the effects that ride-sharing apps have contributed to the possible decline of taxi cabs in New York City.

5.) Description of the real-world data forming the basis of the application and the complex trend queries

For the real-world data surrounding the basis of our application, we will be utilizing the real-world datasets provided by the New York City Taxi & Limousine Commission website that holds information on New York City yellow and green taxicab trip records from the years 2009 to 2022. Each month has more than 2 million records of taxi trip records. Therefore, there are more than 336 million records in total to choose from and operate on.

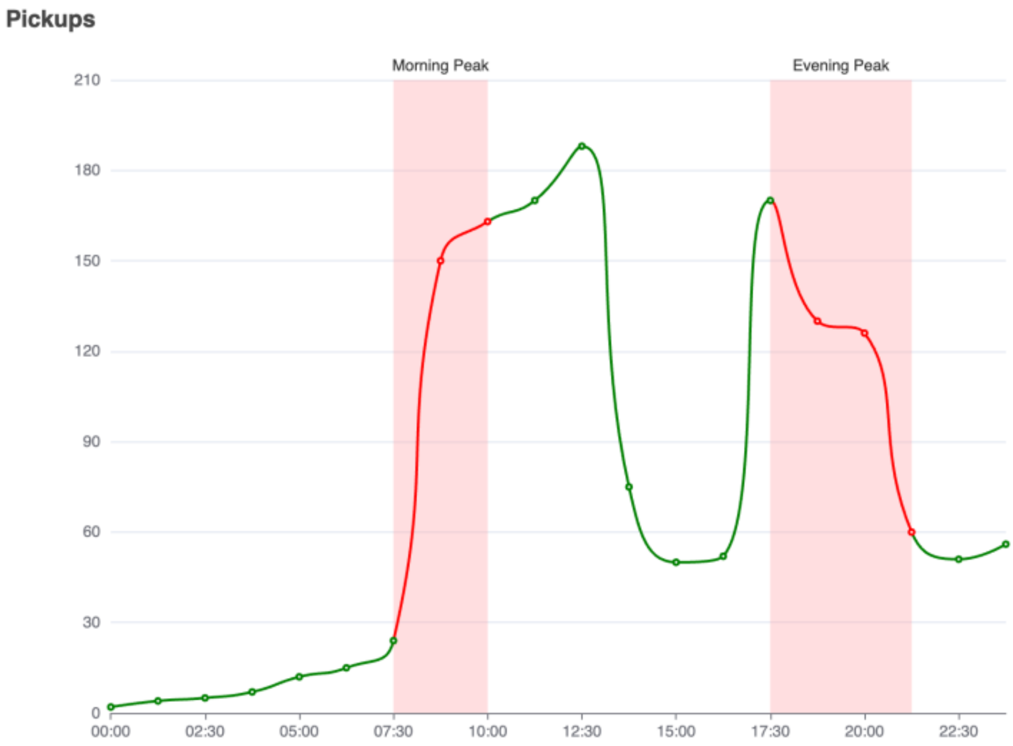
Each record will include 24 attributes of each taxicab's trip such as the location that the taxi meter was disengaged, location where the taxi meter was engaged, extra fees and surcharges, fare amount, improvement surcharge, MTA tax, rate code, store and forward flag, trip distance, date and time when the meter was engaged, date and time when the meter was disengaged, month, year, passenger count, payment method, tip amount, toll amount, total amount, vendor ID, start longitude and latitude, and end longitude and latitude.

Utilizing this data, we will be able to formulate complex queries to filter and analyze any interesting trends that can appeal to multiple types of users. The queries will all focus on the various data attributes listed above and observe their changes that occurred over time. The real-world data will all be raw data and not be manipulated by any algorithms in any way to maximize the efficiency of the trend analysis.

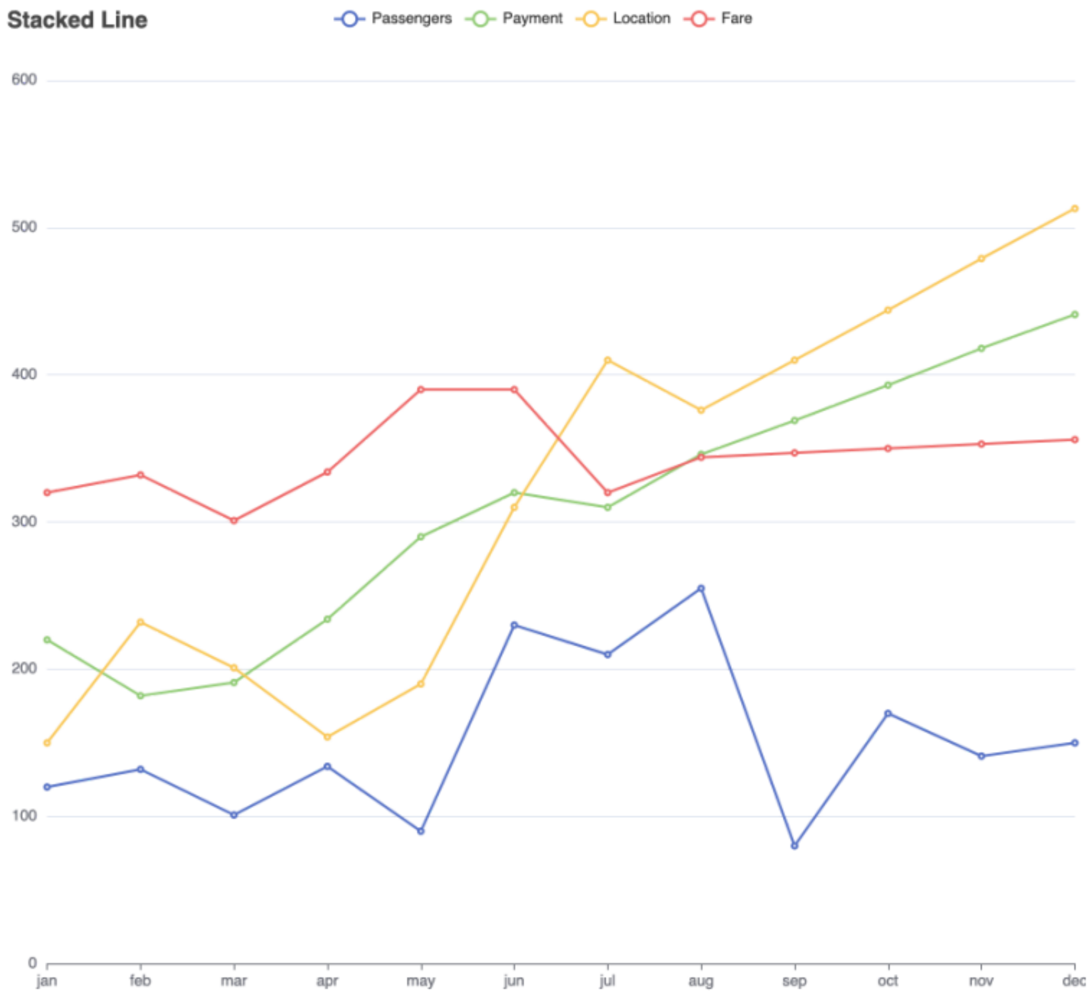
6.) Complex Trend Queries and their explanations

Query #1: Find the trend of passengers over time for the given time-bin at the current and k - nearest clusters to maximize the pickups.

Taxi is the most suitable or popular means of commuting worldwide. Precisely and empirically estimating the demand for taxi passengers is critical for any ride-sharing companies like Uber, OLA, and Lyft. Taxis efficaciously assigns their fleet to some pre-defined stands and minimizes passengers' waiting time, thus increasing their overall satisfaction and customer retention. Nowadays, trip information is available in the database, which we can use to analyze the patterns and trends in passenger demands in specific areas, maximizing the taxi driver's profit. The query will derive the past trends based on the parameters retrieved from the client. We will also fit a simple linear model, giving an appropriate estimation of the number of pickups in the nearby clusters of the client.

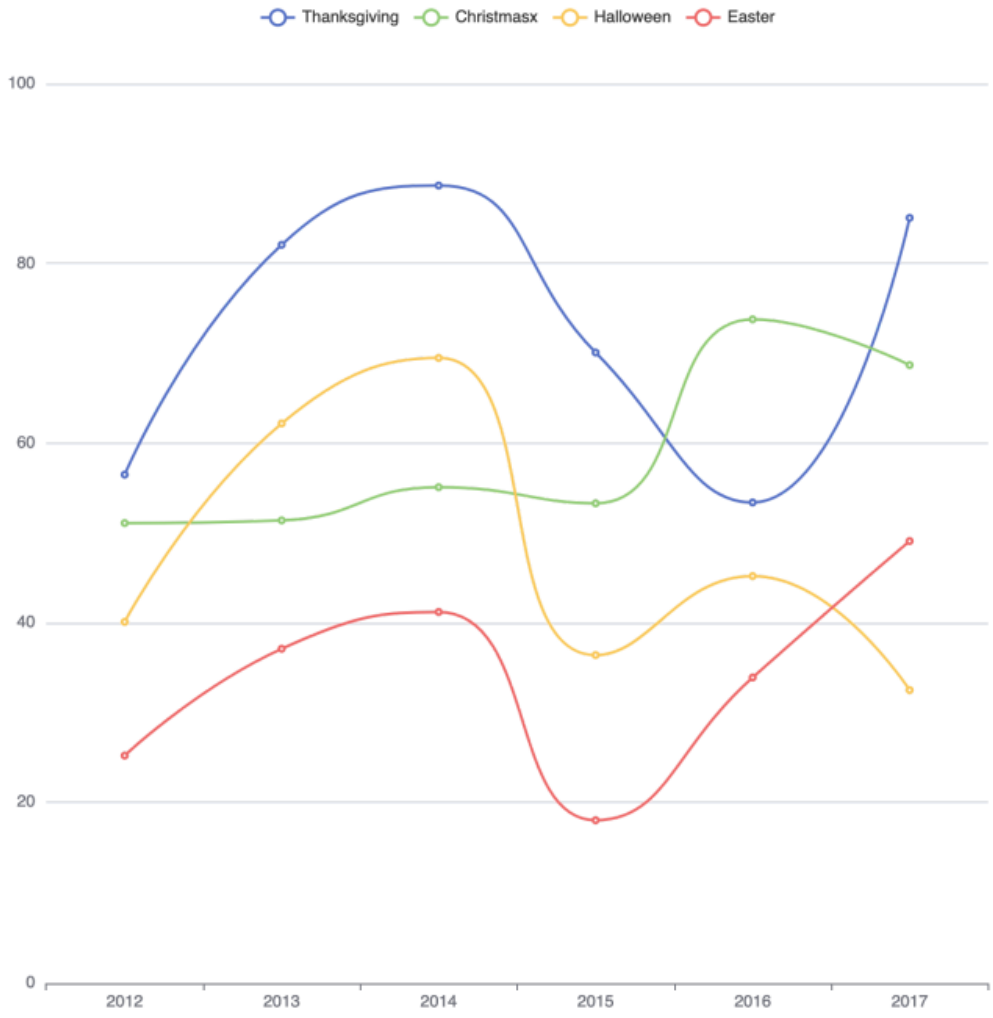


Query #2: Find the effect of Uber and Lyft on Yellow Cabs in NYC over time.
Uber and Lyft's growth in NYC saw a decline in the value of Yellow Cab medallions, which ended up ruining investments once seen as solid as a bar of gold. With passing time, a lot of cab drivers lost their medallions as they lost riders to Uber and Lyft. With the data available to us, we can easily find a trend to visualize the decline of yellow cabs with the growth of mobile based taxi services like Uber and Lyft. This query will focus on comparing the decline over years for a combination of parameters which will be based on user input and will include location, ride fares, tips and number of rides. This will give users a comprehensive overview and hopefully a better insight into this particular trend.



Query #3. Find the overall trend of cab trips during major holidays across a calendar year.

Across a calendar year, we can easily observe from the data available to us that the average population in NYC changes from holiday to holiday and how that affects an average cab driver. At holidays like Thanksgiving and Christmas where all the major tourist locations, shops and mall closed, people might not prefer to go out at all. This could help drivers and riders recognize the hot spots across the city during the holidays, if they ever decide to go out or drive a cab to earn some extra bucks.



Query #4: Track the number of passengers over time before and after COVID to see the effects that COVID may have had on people taking a taxi.

Another potential database query that can be used for our application would be tracking the number of passengers over time before and after the COVID-19 pandemic to observe the effects that the virus may have had on people taking a taxi.

Time is by far the most important variable to consider for our query and data analysis because it helps us determine the trends that we can draw conclusions from. We will have three main time periods based on the data we have: before the pandemic, during the pandemic, and after the pandemic. The years 2009-2019 will comprise the "before the pandemic" timeframe, years 2020 and 2021 will represent “during the pandemic”, and 2022 will be considered “after the pandemic”.

Variables to consider: Number of passengers: One of the primary variables we will consider in this query is the number of passengers that took rides on the NYC taxicabs. With the data given to us, we can find trends in how the number of passengers changed due to potential factors such as the COVID-19 pandemic by fitting that information into a linear graph. Number of COVID-19 cases in NYC : Another potential variable we can consider is the number of COVID-19 cases in NYC for the time that the pandemic took place. By representing the number of cases in a linear graph, we can make connections between how the number of cases could affect peoples’ reactions to the virus’ severity and their unwillingness to take a taxicab causing a decrease in the number of passengers.

By representing both the datasets on the number of COVID cases in NYC and the number of passengers against these timeframes, we can make astute observations on any cause effect relationships that may have occurred.

Query #5: Different payment methods used over time by passengers

One complex database query that our application can utilize is tracking the amount of different payment methods over time used by passengers for taxis to determine the effects the introduction of cash-less payments has had on the modern world. It is known that as time progresses, so does human innovation and inventions to better suit the needs of people and improve convenience.

The primary parameters that we plan on following are the different kinds of payment methods used by passengers in the transactions for the taxi as well as how many were used during our observed timeframe. The payment types recorded in the dataset include credit cards, cash, no charge, dispute, unknown, and voided trip.

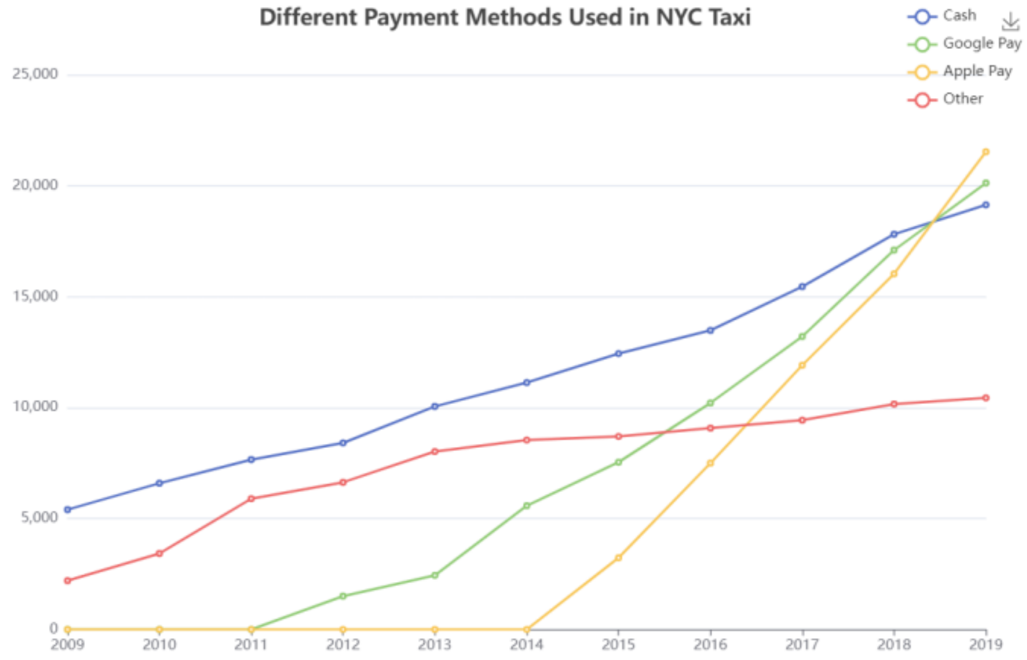


Figure 1: Sample Diagram of Potential Graph of NYC Taxi Payment Methods

In the case of our dataset, we have a considerable amount of taxi data from the years 2009 to 2022, so we'd be able to utilize the total numbers of the different payment methods used and plot them by year. As indicated by the sample plot, Figure 1, above, each payment method would be plotted linearly on a line graph showing the change of their counts over time.

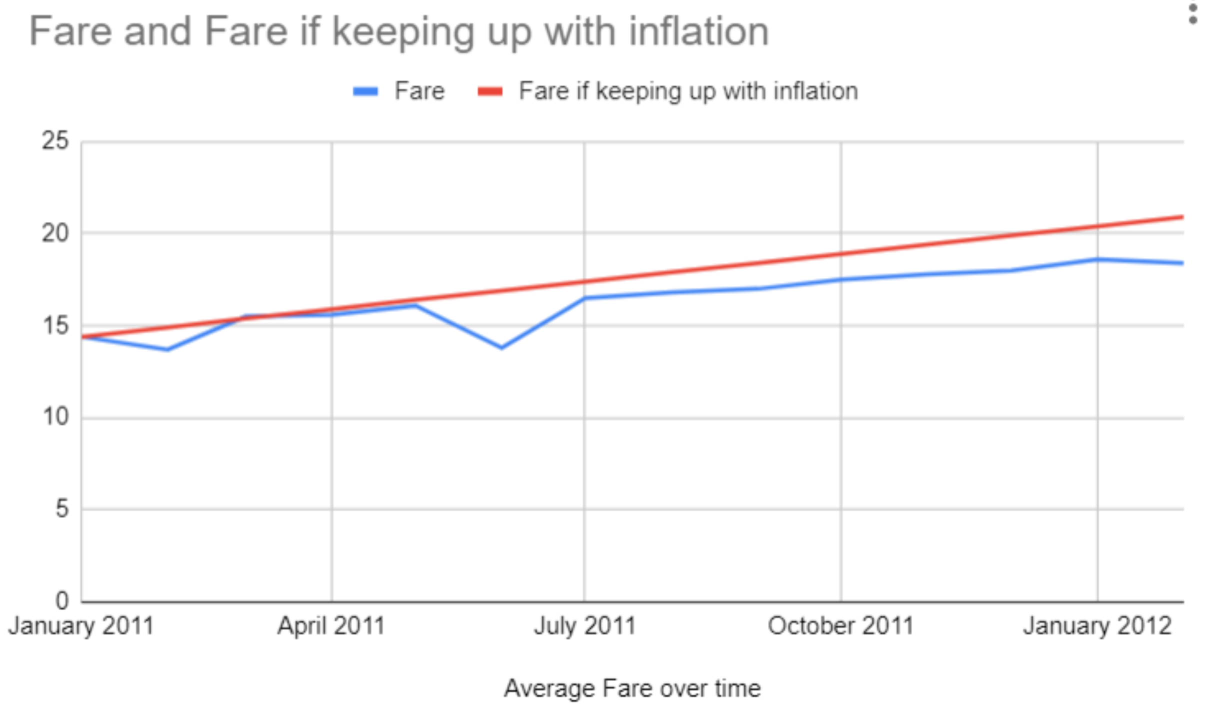
Query #6: Find the trend of taxi driver income compared to that of inflation

https://www.bloomberg.com/news/articles/2022-05-23/nyc-taxi-drivers-call-for-first-fare-increase-in-a-decade?leadSource=uverify%20wall

The fares of taxi drivers have become a hot topic, with many unions fighting for increased fares. This is especially so with the rise of inflation. We must look at the change in income and prices over time, benchmarked to inflation. Taxi wages depend on a number of things, from the average fare they get per trip, and how many trips they can take per day. It also depends on how much they get tipped. Tipping culture for taxis isn't always a percentage of the fare, sometimes it can take the form of an informal "keep the change" style tip. Do surcharges happen more or less often than before? What about extra fees? All things considered, when you add up all the sources of income, are taxi wages keeping up with inflation?
There are several things we need to do to create our final change over time comparison:

- Change of average fare amount over time (main source of income)
- Change of average number of trips per day over time (main source of income)
- Change of average tip over time (Do tips increase with inflation, has tipping culture itself changed. Tipping culture for taxi’s isn’t always a percentage based value, often times it takes the form of “keep the change”)
- Change of average use of Extra Fees over time. E.G a driver may get the extra fee on average 1.2 times a day. How has this average extra fees per day changed over time.
- Change of improvement surcharge over time. E.G a driver may get the improvement Surcharge on average 1.2 times a day. How has this average extra fees per day changed over time.

Finally, combine these totals to show the change in average daily wage and compare it to the average daily wage needed to keep up with inflation.



7.) The intended use of Public Domain/ Proprietary Software

The User Interface, Middleware, the Database System are the primary segments of the application. The front-end web application contains pertinent graphs and filters that the user can use to visualize the trends across various parameters effectively; via this interface, the user interacts with the application. The backend will form the query based on the parameters the user submits, fetch the data from the database and send the response back to the user. The database system is the cardinal component of the application where entities are stored. This system is responsible for executing the query and sending back the responses to the requested backend. We prefer using the following frameworks. We will use ReactJS for developing the front end of the web application.

ReactJS uplifts productivity because of the flexibility of libraries that ReactJS offers. Many companies prefer ReactJS for its ease. We will use Spring Boot for developing our backend. No backend framework is more reliable and efficient in performance than the Spring framework. We will be storing and querying data using Oracle database management systems.