TAXI DATA ANALYSIS AND PREDICTION

KAVYA VEMPATI
MSBA, CALIFORNIA STATE UNIVERSITY
kvempati@horizon.csueastbay.edu , JB7720

KRISHNA SAHITYA DEVINENI
MSBA, CALIFORNIA STATE UNIVERSITY
kdevineni@horizon.csueastbay.edu , NN9061

LEELA KALI MANASA MANDADI
MSBA, CALIFORNIA STATE UNIVERSITY
lmandadi@horizon.csueastbay.edu , TH7515

AJITH NARASIMHA
MSBA, CALIFORNIA STATE UNIVERSITY
anarasimha@horizon.csueastbay.edu , SS1084

INTRODUCTION

Transportation stands as an indispensable cornerstone of services within sprawling urban areas. A plethora oftransportation options are at one's disposal. Across major cities in the United States and globally, taxis hold a paramount position as a mode of conveyance, emerging as a preferred alternative for the public to fulfill their commuting needs. Taxi rides within New York City constitute the central pulse of its urban traffic. The numerous daily rides undertaken by city residents offer valuable insights into traffic patterns, potential road obstructions, and more. Within this context, the dataset encompasses diverse details concerning taxi journeys and their durations within New York City.

Being a prominent global financial hub, the transportation system within New York City (NYC) has been a subject of extensive examination from multiple perspectives. Starting in 2009, the NYC Taxi and Limousine Commission began releasing data regarding taxi operations in the city, providing a valuable opportunity for in-depth analysis. Consequently, the primary aim of this study is to scrutinize the determinants of taxi demand, identify the locations with the highest pickup and drop-off frequencies for the public, pinpoint peak traffic times, and explore strategies to meet the public's transportation needs more effectively.

BUSINESS PROBLEM

The challenge here is to fully grasp and forecast when and where people will need taxis in cities. To do this, we're looking at lots of things that affect taxi use, like weather, where businesses are, who lives in different areas, how people get around, and how much rides cost. We want to create a strong plan using tools like Tableau to visualize data and special models to guess when and where more taxis might be needed. Ultimately, our goal is to improve taxi services by understanding these patterns and being ready for changes in demand.

Weather Dynamics and Travel Patterns:

Weather conditions wield a substantial influence on commuter behavior, shaping travel choices and preferences. Variances in temperature, precipitation, seasonal shifts, and climatic nuances significantly impact individuals' inclination towards taxi services. The correlation between weather fluctuations and taxi demand at

different locations and times unveils invaluable insights.

Business Locations:

Identifying key business hubs, their operational hours, and workforce commuting patterns are critical elements in forecasting peak demand periods. Understanding the ebb and flow of taxi demand concerning business locations facilitates strategic resource allocation and service optimization.

Transportation Facilities and Complementary Services:

The accessibility, reliability, and efficiency of public transportation systems intricately intertwine with taxi demand. Analyzing the relationship between public transportation routes, their utilization patterns, and taxi demand illuminates the complementary or competitive nature between these transit modes. Understanding these interactions aids in optimizing transportation services and fostering synergies between different transit options.

Demographic Segmentation and Preferences:

Delving into demographic compositions, including population density, age distributions, employment profiles, and commuting behaviors, to discern distinct demand patterns across different neighborhoods and demographic segments.

DATA COLLECTION

The 1st data Set for the project is taken from New York City Taxi and Limousine Commission (TLC)
https://www.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf .

The dataset was imported and decoded from its native parquet format into a csv format. Here we have collected taxi data of three types – Green taxi, Yellow Taxi and For-Hire Vehicle Taxis.

The 2nd dataset for the project is the Weather data across each zone collected from https://home.openweathermap.org/history_bulks/new.

Analyzing the integrated dataset to identify correlations and patterns between weather variables and taxi demand. By blending these datasets, we gain a deeper understanding of how different weather conditions impact the need for taxi services

leading to more accurate predictions and optimized resource allocation for taxi services.

The 3rd dataset is demographics and Transportation data collected from

Census Data- Demographics

Including taxi data, demographics, and transportation information, we gain a holistic understanding of taxi needs among various demographic segments. This approach enables more targeted and efficient service offerings tailored to the specific needs of different groups within a population.

The 4th dataset is Business Locations - NYC Business Locations.

This understanding can help taxi services optimize their operations, tailor services, and strategically allocate resources to meet the transportation needs of different business-related demographics.

DATA PREPROCESSING & CLEANING

Identifying and prioritizing columns crucial to the business objectives ensures that the data warehouse aligns with the organization's goals. This involves understanding which data attributes are essential for analysis, decision-making, and reporting.

Knowing the significance of each column helps in maintaining data accuracy and consistency. It allows for validation, ensuring that the information stored is reliable and valid for its intended use.

Understanding the importance of columns aids in optimizing storage space within the data warehouse. It prevents unnecessary duplication or storage of irrelevant data, ensuring efficient use of resources. Essential columns facilitate effective analysis and reporting by providing the necessary data elements required for generating insights and making informed business decisions.

Taxi Data:

Yellow and green taxi trip records include fields capturing 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. For-Hire Vehicle (“FHV”) trip records include fields capturing the dispatching base license number and the pick-up date, time, and taxi zone location ID.

The data granularity was set at the census tract level, allowing for a detailed analysis within this specific geographic unit. To achieve this, we mapped each zone and location ID to encompass all corresponding census tracts falling beneath it.

- Each taxi dataset is reviewed to understand their structure, including columns, data types, and unique identifiers.
- Identified common columns across datasets that hold similar information (e.g., pickup time, drop-off time, location, fare, etc.).
- Standardized column names and formats to ensure consistency. For instance, "pickup_datetime" should have the same format in all datasets.
- Addressed missing values, inconsistencies, or discrepancies in columns across datasets.
- Converted data types to match if necessary (e.g., ensuring consistency in date formats or numeric representations).
- Handled duplicate or mismatched records that might occur during the merge process. Ensure the dataset integrity by resolving any conflicts.
- Validated the merged dataset to ensure all necessary columns are present and aligned correctly.

Weather Data:

The raw weather dataset was initially examined to understand its structure, columns, and contents. Based on the project's requirements, specific columns such as temperature, humidity, weather condition, wind speed, and timestamp were selected for further analysis. Granularity of data is considered at zone level here

- Any identified missing data entries were carefully addressed. Imputation techniques were employed to ensure data completeness without compromising the dataset's integrity.
- Temperature units were standardized to a consistent format to maintain uniformity across the dataset.
- Date and time columns were parsed and formatted into a standardized, human-readable format, facilitating ease of analysis and interpretation.
- Rigorous quality checks were performed to verify data consistency, accuracy, and adherence to predefined standards.

Demographics and Transportation Data:

The dataset underwent a meticulous data cleaning process, wherein only essential columns were retained. However, due to the absence of zones in the original dataset, GIS (Geographic Information System) was employed to establish the interrelationship between zones, boroughs, and census tracts

- Identified and extracted necessary columns vital for analysis, ensuring relevance and eliminating redundant information.
- Addressed missing values, inconsistencies, and outliers within the dataset, maintaining data accuracy and integrity.
- Utilized GIS techniques to establish spatial connections between zones, boroughs, and census tracts explained in detailed in the next section.

Business Location Data:

Business location data was sourced from diverse database encompassing information on active businesses within designated localities. The dataset comprised details such as business names, addresses, categories, and licensing specifics.

- Identified and selected columns relevant to the analysis, including business names, categories, addresses, and licensing information.
- Eliminated duplicate entries and addressed any errors or inconsistencies present in the dataset to ensure data accuracy.
- Dealt with missing values by employing strategies such as imputation or exclusion, preserving data integrity.
- Standardized formats for categorical data and addresses to ensure uniformity and ease of analysis.

Geographic Information System (GIS)

GIS stands for Geographic Information System. It's a system designed to capture, store, manipulate, analyze, manage, and present spatial or geographic data. It allows users to visualize, interpret, and understand data patterns and relationships in a geographic context.

Its like a digital atlas with superpowers. It's a computer system that captures, stores, and presents data tied to specific locations on Earth. Essentially, GIS lets you

explore and analyze a wealth of data layered onto a map, making it simpler to understand how different things relate in the world around us.

With GIS technology, people can compare the locations of different things to discover how they relate to each other. It enables spatial analysis, allowing for better decision-making by providing insights into geographic patterns, relationships, and trends. It aids in understanding complex spatial data, identifying correlations, and predicting outcomes, crucial for urban planning, disaster management, environmental conservation, and more.

When it comes to identifying which census tract belongs to which zone, GIS plays a crucial role -

Spatial Analysis: GIS helps in analyzing the spatial relationships between census tracts and different zones. It allows for overlaying various geographic layers (such as census tract boundaries and zoning maps) to determine the intersection or relationship between them.

Data Integration: GIS can integrate diverse datasets, including census tract boundaries, zoning information, demographic data, land use data, etc. This integration allows for a comprehensive understanding of how different zones relate to specific census tracts based on various parameters.

Decision Making: When making decisions related to urban planning, resource allocation, infrastructure development, or policy implementation, GIS helps by providing spatial insights. Understanding which census tracts belong to specific zones assists in making informed decisions about development, resource distribution, or policy changes.

Ensuring that both datasets (Census Tract and Taxi Zones) contain geometry columns representing their respective geographic shapes is very important.

In GIS, a projection refers to the method used to represent the Earth's curved surface on a flat map. Since the Earth is three-dimensional and maps are two-dimensional, a projection involves mathematically transforming the Earth's spherical or ellipsoidal surface onto a flat plane.

When bringing two different datasets into the same projection in GIS, you aim to ensure that they align properly for accurate analysis and visualization.

- **Identify Current Projections:** Check the projection information for each dataset. This information is often stored within the metadata of the data file.

Common projections include Mercator, Albers Equal Area, Lambert Conformal Conic, etc.

- **Reprojecting Data:** Use GIS software or libraries like GeoPandas to reproject one or both datasets to a common projection. The goal is to have both datasets in the same coordinate reference system (CRS). GeoPandas, for instance, provides the to_crs method to reproject GeoDataFrames.
- **Check Alignment:** After reprojecting, check if both datasets align properly.
- **Perform Analysis:** Once both datasets are in the same projection and aligned correctly, you can perform spatial operations, joins, or analyses on them confidently.

In GIS analysis, understanding which census tract belongs to which zone often involves spatial relationships between different geographic units, such as census tracts and zoning areas. The centroid method can be used.

- The centroid of a census tract can be used to determine its spatial relationship with zoning areas. By calculating the centroid of each census tract, you create a point that represents its approximate center. Then, by checking which zone polygon contains each centroid point, you can infer which zone a census tract belongs to based on the zone containing its centroid.
- Census tracts and zoning areas might have intricate shapes, but their centroids are single points, making spatial comparisons and computations more straightforward. Centroids help handle irregularly shaped polygons. For instance, if a census tract overlaps multiple zoning areas or has an irregular shape, its centroid can provide a single point for spatial analysis, simplifying the identification process.
- Centroids of single-part geometries are straightforward to calculate as they represent the center point of that single shape.
- For multi-part geometries the centroid might be calculated for the entire multi-part geometry, potentially resulting in a centroid that doesn't accurately represent each individual part of the geometry.

Single-Part Geometries:

- Single-part geometries have a single centroid for each individual shape. Calculating centroids for single-part features is straightforward because each feature represents a single entity.

Multi-Part Geometries:

- Multi-part geometries, however, contain multiple disconnected shapes within a single feature. These shapes might represent distinct geographic entities but are grouped together.
- When calculating a centroid for a multi-part geometry, the software would compute a single centroid for the entire multi-part feature, which might not accurately represent the centroids of individual parts.
- To get separate centroids for each individual part within a multi-part geometry, it's necessary to convert the multi-part feature into separate single-part features. This separation allows each individual shape to be treated independently, enabling the calculation of accurate centroids for each part separately.

Example:

- A single-part geometry could be a single polygon representing a city block or a single line representing a river.
- A multi-part geometry could represent a district made up of multiple non-contiguous areas or a complex polygon representing a land parcel with internal cut-outs or holes (like a park within a larger area).

Spatial joins in GIS involve combining information from different spatial datasets based on their spatial relationships. The terms "one-to-many" and "many-to-one" refer to the cardinality of the relationship between the datasets being joined.

One-to-Many Spatial Join:

- In a one-to-many spatial join, one feature from the first dataset (let's call it the "left" dataset) can be associated with multiple features from the second dataset (the "right" dataset) based on their spatial relationship.
- For example, if you have a set of points representing cities (left) and a set of polygons representing states (right), a one-to-many spatial join could associate multiple cities with each state polygon based on which state each city falls within.

Many-to-One Spatial Join:

- Conversely, in a many-to-one spatial join, multiple features from the first dataset can be associated with a single feature from the second dataset based on their spatial relationship.

- Using the same example of cities (left) and states (right), a many-to-one spatial join might associate multiple states with each city, determining which state each city falls within.

In our project but for the Demographics Data and Business Data we have used GIS to gather at the data at census tract granularity. The code for the both datasets is as follows-

DATA WAREOUSING:

Data warehousing serves as a crucial component in modern-day businesses, facilitating efficient data management, analysis, and decision-making. Understanding and identifying fact tables and dimension tables within a data warehouse are pivotal aspects that contribute significantly to its functionality. Here's why they're important:

Importance of Data Warehousing:

Centralized Data Repository: Data warehouses act as centralized repositories, amalgamating data from various sources across an organization. This consolidation enables a single source of truth for analytical purposes.

Historical Data Storage: They store historical data, allowing for trend analysis, performance evaluation, and forecasting, providing valuable insights into long-term patterns and trends.

Enhanced Decision-Making: By providing structured, cleansed, and integrated data, data warehouses empower decision-makers with accurate and timely information, fostering informed and strategic decision-making processes.

Support for Business Intelligence (BI) and Analytics: Data warehouses serve as the backbone for business intelligence and analytics, enabling users to generate reports, perform complex queries, and derive actionable insights.

Understanding Fact Tables:

Fact tables store quantitative, numeric data known as measures or metrics. They contain the core data around which business analysis revolves, such as sales figures,

quantities sold, or financial metrics.

Fact tables are typically associated with events or transactions and often contain foreign keys that link to dimension tables.

Identifying Dimension Tables:

Dimension tables contain descriptive attributes or contextual information related to the measures stored in fact tables. They provide context to the data in the fact tables.

Dimension tables categorize and organize data into hierarchical structures, allowing for drill-down analysis and segmentation. For instance, time dimensions may include hierarchies like year, quarter, month, etc.

Fact tables are linked to multiple dimension tables, establishing relationships that facilitate multidimensional analysis.

The segregation between fact and dimension tables allows for optimized query performance by enabling users to focus on necessary dimensions without accessing unnecessary data.

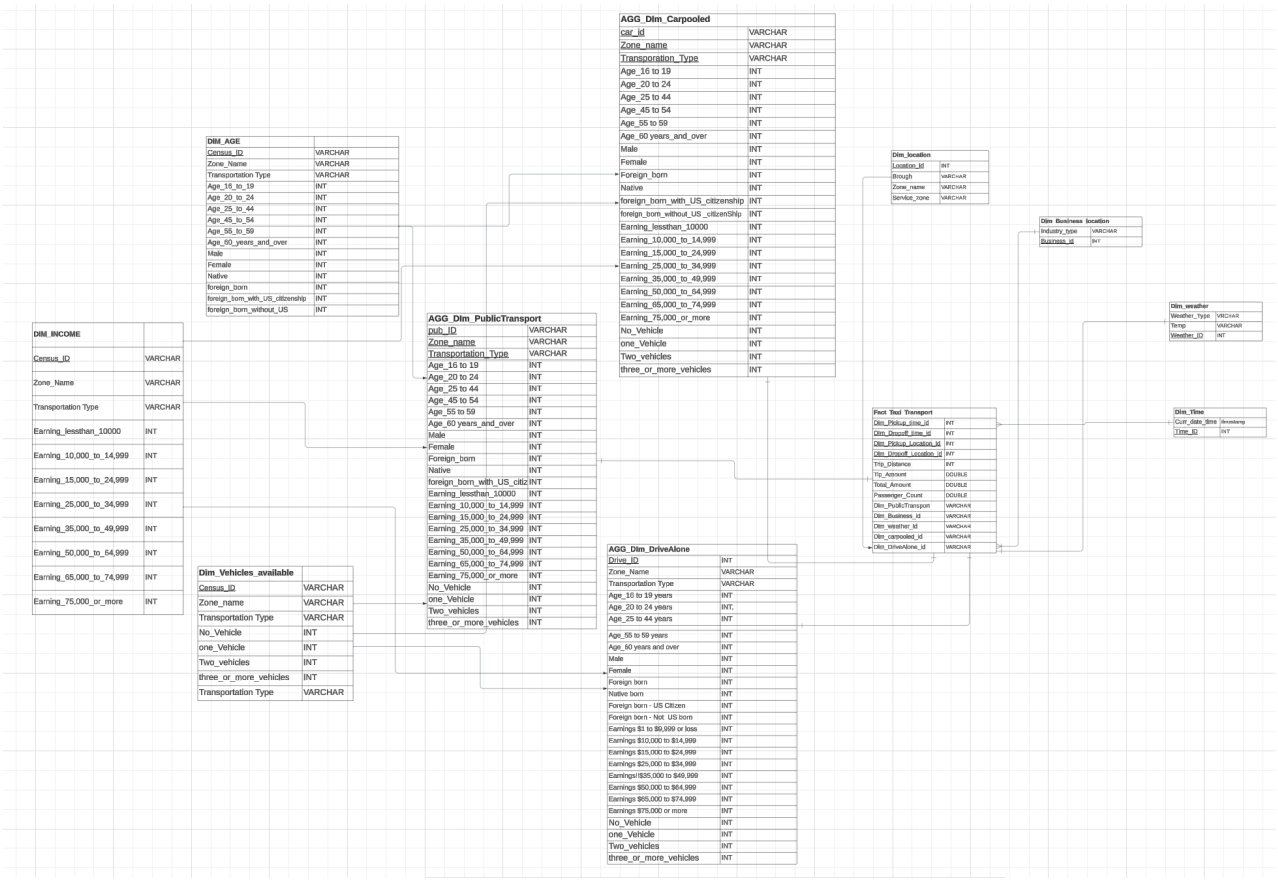
By separating measures from descriptive attributes, fact and dimension tables streamline data access, making it more manageable for analysis.

Granularity:

Level of Detail: Granularity refers to the level of detail or aggregation within data. It defines how fine or coarse the data is stored and analyzed.

Granularity influences the level of insights derived from analyses. Fine granularity offers detailed insights but might impact query performance, while coarse granularity provides summarized information but faster query execution.

FACT AND DIMENSION TABLES:



FACT_DEPT_TRANSPORT:

Fact_dept_transport		**Description**
Location_id	INT	This column contains unique identifier
Customer_id	VARCHAR	This column contains unique identifier
Transportation_Type	VARCHAR	Type Of Transportation

FACT_TAXI_TRANSPORT:

Fact_Taxi_Transport		**Description**
Dim_Pickup_time_id	INT	A dimension or identifier representing the pickup time of the taxi ride
Dim_Dropoff_time_id	INT	A dimension or identifier representing the dropoff time of the taxi ride
Dim_Pickup_Location_id	INT	A dimension or identifier indicating the location where the taxi pickup occurred.
Dim_Dropoff_Location_id	INT	A dimension or identifier indicating the location where the taxi dropped off the passenger.
Trip_Distance	INT	The distance traveled during the taxi ride
Trip_type	DOUBLE	A categorization or identifier representing the type of trip
Tip_amount	DOUBLE	The amount of tip given by the passenger for the taxi ride
Total_amount	DOUBLE	The total fare or cost of the taxi ride, including the base fare, additional charges, and tips
Passenger_count	DOUBLE	The number of passengers in the taxi during ride
Dim_Business_id	DOUBLE	A dimension or identifier associated with business related details
Dim_weather_id	VARCHAR	A dimension or identifier linking to weather-related details during the taxi ride

DIM_BUSINESS_LOCATION:

Dim_BusinessLocation		**Description**
Business_id	INT	This column contains unique identifier
Industry_type	VARCHAR	Type of Industry

DIM_WEATHER:

Dim_weather		**Description**
Weather_id	INT	This column contains unique identifier
Temperature	VARCHAR	Temperature
Weather Type	VARCHAR	Type Of Weather

DIM_LOCATION:

Dim_location		**Description**
Location_Id	INT	This column contains unique identifiers or codes assigned to different locations within New York
Brough	VARCHAR	One among the five boroughs
Zone_name	VARCHAR	The name or label assigned to specific zones within each borough
Service_zone	VARCHAR	The service zone associated with each location.

DIM_TIME:

Dim_time		**Description**
Time_id	INT	This column contains unique identifier
Date	DATE	Date of taxi booked
Time	TIME	Time at which taxi is booked

DIM_DRIVEALONE:

Dim_DriveAlone	**Data Type**	**Description**
Customer_id	VARCHAR	Unique identifier for each individual who drives alone
Age_group	VARCHAR	Categorization of individuals who drive alone into different age groups
sex	VARCHAR	Gender of the individuals who drives alone
Birth_Area	VARCHAR	Place or region where the customer who drives alone was born
Employment_type	VARCHAR	Categorization of the job type of individuals who drive alone
Department_Type	VARCHAR	The division or industry that the individual who drives alone is associated with
Income_range	VARCHAR	The range of incomes corresponding to individuals who drive alone
Poverty_level	VARCHAR	The level of poverty associated with the customer who drives alone
No_of_Vechicles_Available	VARCHAR	The number of vehicles that the individual who drives alone has access to

DIM_PUBLICTRANSPORT:

DIM_PUBLIC TRANSPORT	Data Type	Description
Customer_id	VARCHAR	Unique identifier for each customer
Age_group	VARCHAR	Categorization of customers into different age groups of public transport users
sex	VARCHAR	Gender of the customer using public

		transport
Birth_Area	VARCHAR	Place or region where the customer was born
Employment_type	VARCHAR	Categorization of the job type of commuters using public transportation
Department_Type	VARCHAR	The division or industry that the commuter is associated with
Income_range	VARCHAR	The range of incomes corresponding to each passenger using public transportation
Poverty_level	VARCHAR	The level of poverty associated with the customer using public transport
No_of_Vechicles_Available	VARCHAR	The number of vehicles that the customer has access to

DIM_CARPOOLED:

DIM_CARPOOLED	**Data Type**	**Description**
Customer_id	VARCHAR	Customer's unique ID from the carpool reservation
Age_group	VARCHAR	Customer age group that made the carpool reservation
Sex	VARCHAR	The customer's gender who made the carpool reservation
Birth_Area	VARCHAR	Place or region where the customer was born
Employment_type	VARCHAR	Classification of the customer's employment status
Department_Type	VARCHAR	The department or sector to which the customer belongs
Income_range	VARCHAR	The range of income associated with each customer

Poverty_level	VARCHAR	The level of poverty associated with the customer
No_of_Vechicles_Available	VARCHAR	The number of vehicles that the customer has access to


Actual tables.txt


Aggregated Tables.txt


dim tables.txt


insert commands.txt

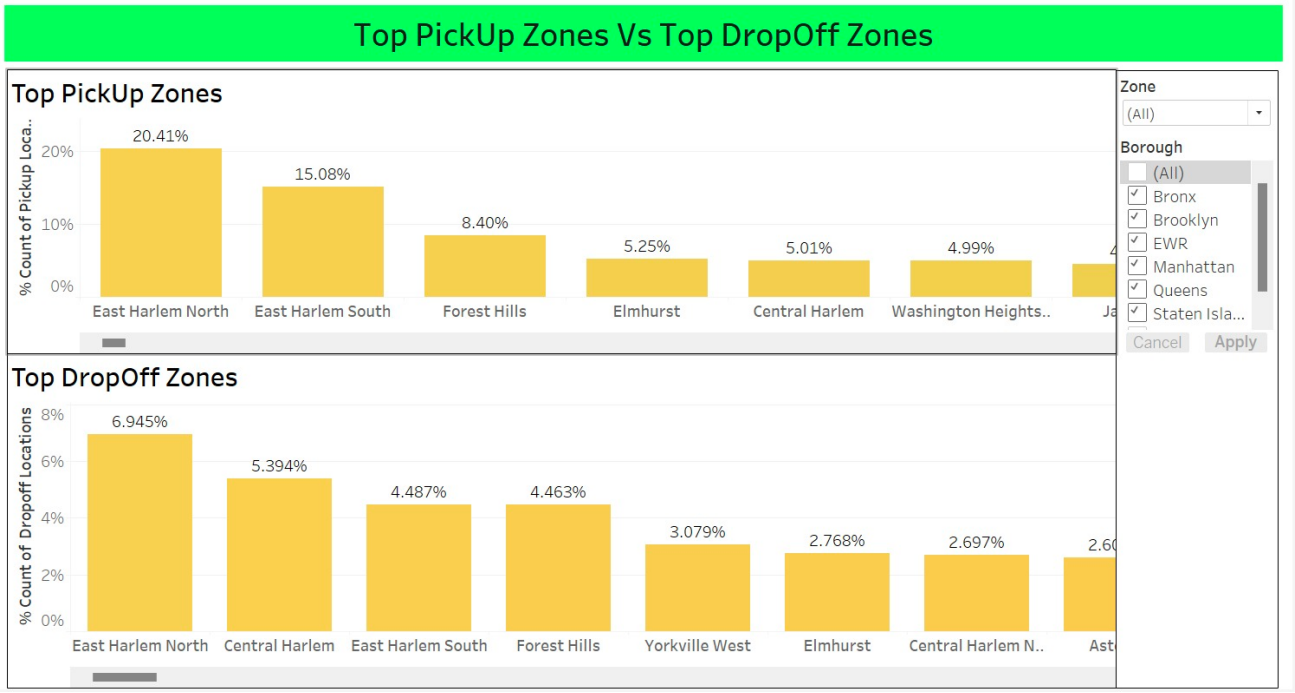

Modified Tables.txt

DATA VISUALIZATION:

Data visualization and exploratory data analysis are indispensable processes in understanding and extracting insights from raw datasets. Through visual representations and exploratory data analysis we can unveil patterns, trends, and relationships within the data. It allows for the detection of outliers, the exploration of variable distributions, and the examination of correlations, providing a foundational understanding before more advanced analyses. On the other hand, data visualization transforms these insights into compelling visuals that aid in conveying complex information in an easily understandable format. It's the bridge that connects the analytical discoveries to actionable insights, enabling stakeholders to grasp the significance of the data at a glance and make informed decisions.

Pick Up Zones Vs Drop Off Zones

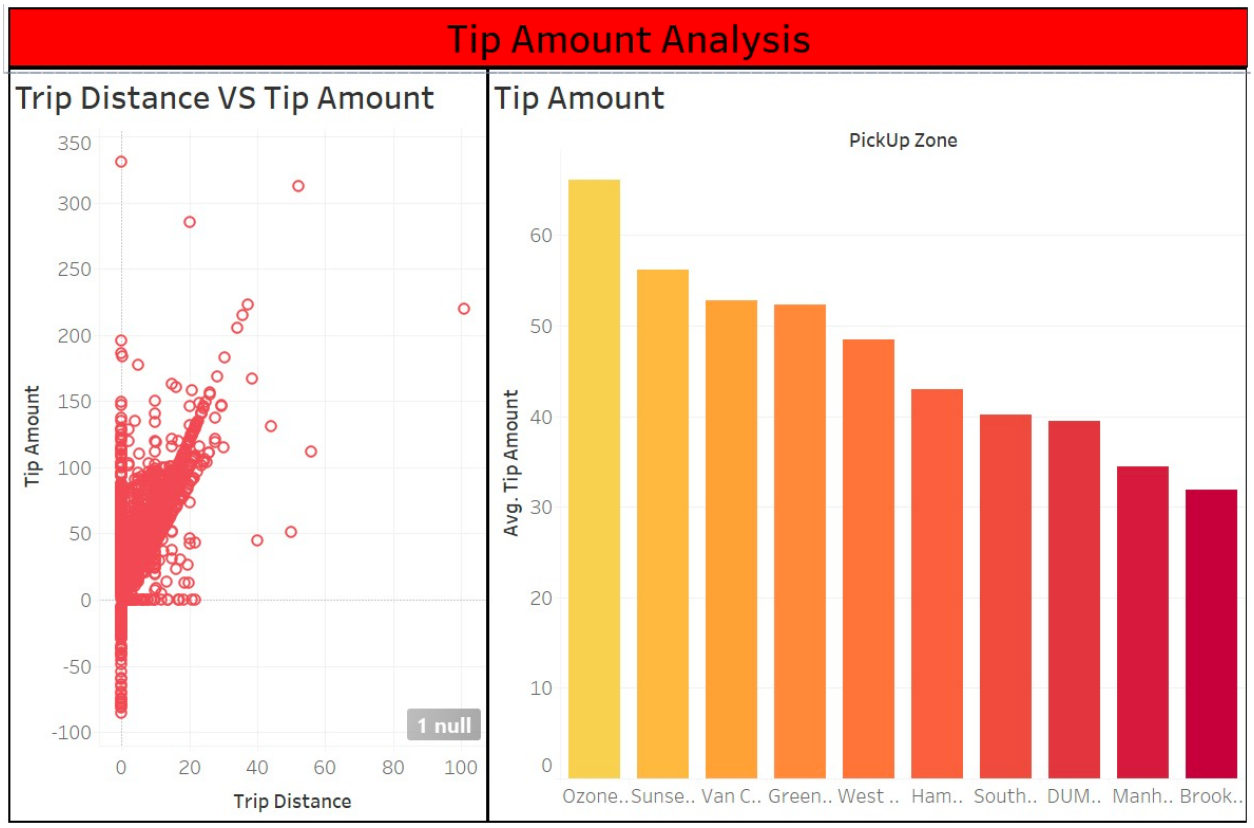
This dashboard presents a clear view of the prime locations where taxi pickups and drop-offs predominantly occur. By prioritizing these key zones, taxi owners can strategically position their fleet to maximize the influx of requests. Concentrating efforts on these high-traffic areas holds the potential to significantly increase taxi demand and optimize operational efficiency, ensuring a more responsive and lucrative service within the most sought-after areas.



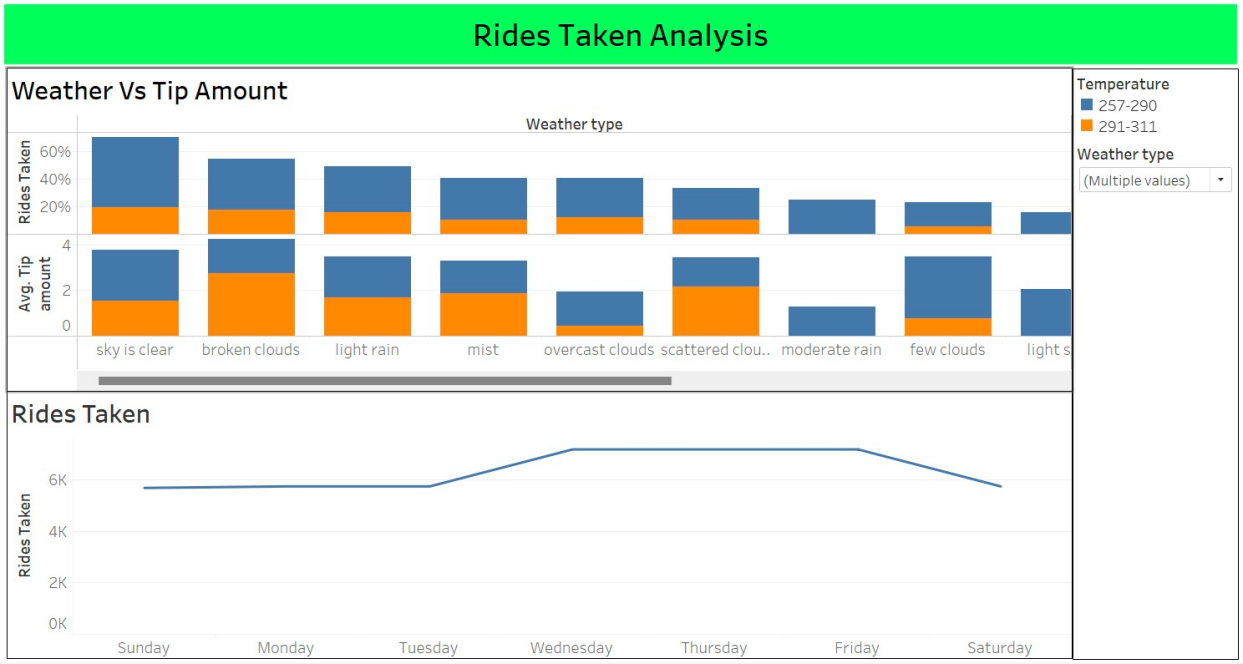
Tip Amount Analysis

The visualization is a scatter plot that shows the relationship between trip distance and tip amount. The x-axis represents the trip distance, and the y-axis represents the tip amount. As the trip distances extend, a notable trend emerges: an apparent rise in tip amounts. This correlation suggests that longer journeys are often associated with increased tipping behavior from customers.

The bar chart highlights specific pickup zones that stand out as potential areas where higher tips are more likely. Understanding this connection between trip duration and tipping habits can assist taxi service providers in identifying opportunities to enhance customer satisfaction and potentially increase earnings by focusing on longer-distance trips and these prominent pickup zones.



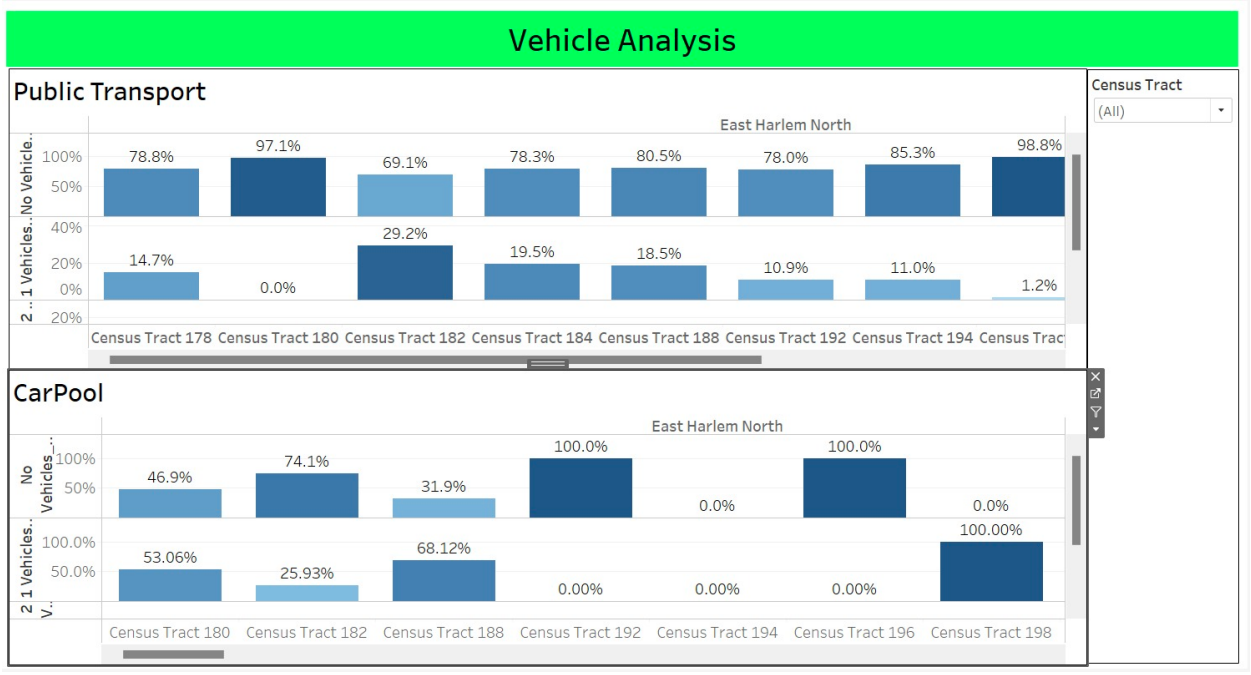
Rides Taken Analysis:



The above plot depicts average tip amounts against the number of taxi pickups under various weather conditions presents a valuable correlation, shedding light on how weather impacts both tipping behavior and the frequency of taxi requests. Understanding these patterns across boroughs, industry presence, and weather conditions is pivotal in devising targeted strategies to optimize taxi services and cater more effectively to varying demands in different zones and weather scenarios. Also understanding peak demand days allows for better service anticipation in turn helping taxi companies enhance their customer experience by providing better service levels and ensuring availability during these busy periods.

Here we observed that tips tend to increase during cloudy weather, while clearer skies correspond to a higher frequency of rides. Conversely, the fewest rides occur during snowy, hazy, and foggy weather conditions. Ride demand peaks on Wednesday, Thursday, and Friday, experiencing a noticeable decline on Saturdays.

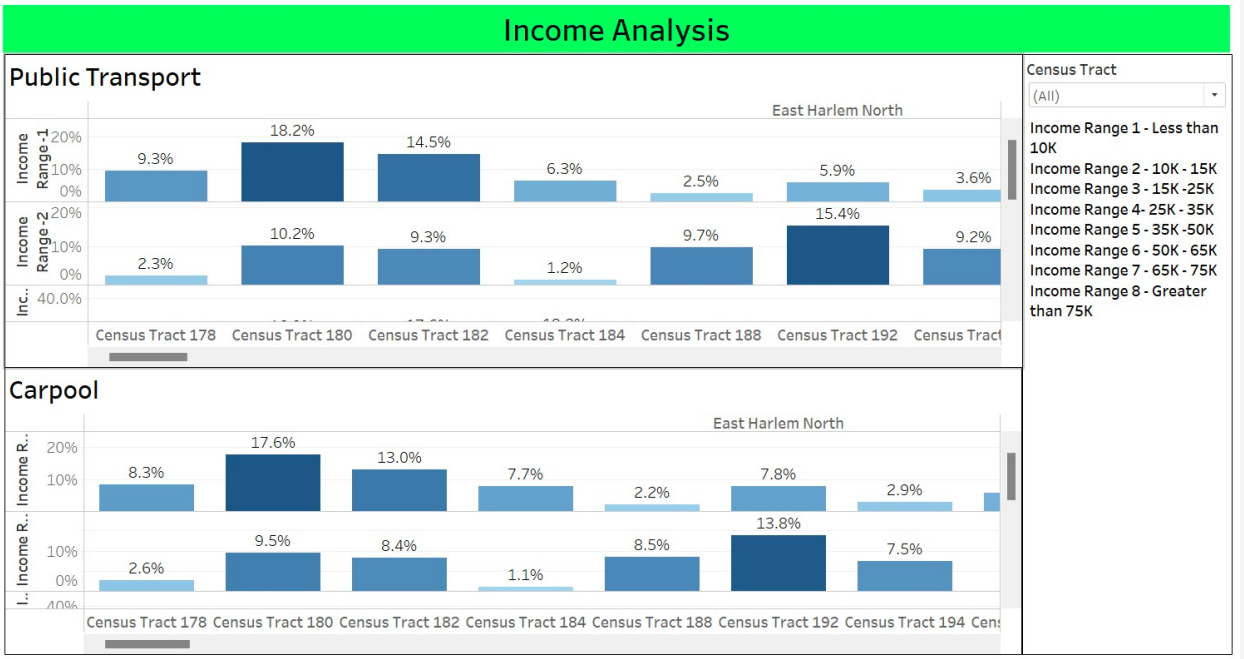
Vehicles Analysis at Zone/Census Tract:



The dashboard provides a comprehensive breakdown of zones and census tracts where individuals predominantly utilize carpooling and public transport, offering insights on their vehicle ownership. By cross referencing this information with the

dashboard listing the zones with the highest taxi pickups, taxi drivers gain a strategic advantage. They can leverage this data to pinpoint specific zones where carpooling and public transport are prevalent, potentially identifying opportunities to cater to commuter preferences or specific demographics. This targeted approach enables taxi drivers to optimize their service offerings, aligning them more closely with the commuting habits and preferences observed in these zones with high taxi demand.

Income Analysis at Zone/Census Tract:



Income levels often influence transportation choices. Identifying areas with varying income brackets helps tailor services to cater to diverse customer needs. For instance, areas with higher income might demand premium services, while others might prefer cost-effective options. Higher-income areas may exhibit different ride behaviors, such as preference for longer rides, specific destinations, or particular times of day. Understanding these patterns helps optimize driver routes and availability.

Also, income levels affect how sensitive riders are to price changes. Areas with lower income might be more price-sensitive, influencing demand based on fare affordability. Analyzing this aids in pricing strategies and service offerings.

DATA MODELLING:

The initiative aimed to address a significant concern among taxi drivers regarding the uncertainty surrounding tip amounts until the completion of a ride. This challenge stemmed from the intricate calculation involved in determining tips, where distance covered was merely one aspect among several contributing factors. To mitigate this issue, a robust model was developed using the wealth of historical taxi data amassed by NYC taxi Data along with Weather and Business Data. The model's objective was to offer real-time tip estimations for drivers upon passenger entry into the taxi.

The project encompassed a comprehensive pipeline, starting with data ingestion and rigorous cleansing processes to ensure the integrity and quality of the dataset. Subsequent stages included intricate feature engineering and judicious feature selection to distill the most relevant variables affecting tip amounts. Finally, various regression models were constructed leveraging this curated dataset, incorporating crucial elements such as time of day, pickup and dropoff locations, and other pertinent ride-specific information.

By amalgamating advanced machine learning techniques with the wealth of data at our disposal, the project aimed to empower drivers with a predictive tool capable of offering insightful tip estimations, thus enhancing their overall experience and financial predictability.

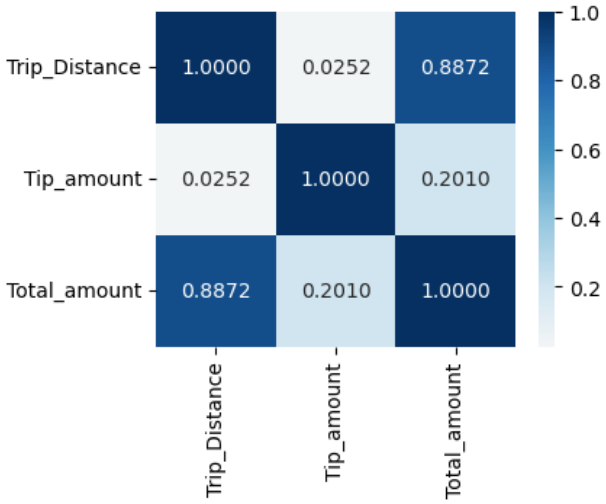
Feature Engineering: During the data cleaning phase, attention was devoted to feature engineering, particularly in processing datetime columns. By converting the DOTimestamp and PUTimestamp columns into datetime format during data ingestion, the analysis of temporal data became more efficient. This transformation enabled the extraction of new features such as 'PU_weekday_num', 'PU_month', 'PU_pickup_hour', 'DO_weekday_num', 'DO_month', and 'DO_pickup_hour'.

Converting the data types as appropriate

```
taxi_df['DOTimestamp'] = pd.to_datetime(taxi_df['DOTimestamp'], format='%d-%m-%Y %H:%M')
taxi_df['PUTimestamp'] = pd.to_datetime(taxi_df['PUTimestamp'], format='%d-%m-%Y %H:%M')
taxi_df.DOZone = taxi_df.DOZone.astype('category')
taxi_df.PUZone = taxi_df.PUZone.astype('category')
taxi_df.weather_type = taxi_df.weather_type.astype('category')
taxi_df.Temperature = taxi_df.Temperature.astype('category')
taxi_df.Industry = taxi_df.Industry.astype('category')
```

Machine Learning and Model Development: The prepared dataset was then utilized for tip amount predictions, employing supervised regression models such as linear regression and Gradient Boosting. Evaluation metrics including RMSE and r2 were computed for each model to gauge their performance.

Feature Selection: To identify relevant features, a heatmap analysis was conducted. Notably, due to high correlation between trip_distance and total_amount, total_amount was omitted as a predictor to mitigate issues related to multicollinearity.



Further Steps: The subsequent stage involved the creation of a subset comprising the selected features alongside the target variable. Feature scaling was implemented to ensure uniformity in feature magnitudes. Additionally, the dataset was split into a 70:30 ratio for training and testing the models respectively.

```
predictors = ['zTrip_Distance', 'weather_type', 'Temperature', 'DOZone','PUZone','PU_weekday_num', 'PU_pickup_hour','DO_weekday_r
outcome = 'Tip_amount'


# partition data
X = pd.get_dummies(taxiNorm[predictors], drop_first=True)
y = taxiNorm[outcome]
train_X, valid_X, train_y, valid_y = train_test_split(X, y, test_size=0.3, random_state=100)
```

Models Building:

Linear Regression Model:

The Linear Regression model was the backbone of our predictive analysis for estimating tip amounts in taxi rides. Its premise rests on establishing a linear connection between input features and the target variable—predicting tip amounts based on factors like time of day, pickup/dropoff locations, and other ride specifics.

By fitting a line that best represents these relationships, the model seeks coefficients for each feature, minimizing the disparity between predicted and actual tip amounts from our dataset.

Leveraging this model in our project involved creating a predictive function using historical taxi data. Through training, the model learned patterns and correlations between various ride features and tip amounts. This empowered our drivers by providing a tool to anticipate potential tip earnings, enabling better planning and work optimization based on these predictions.

Linear Regression Model

```
taxi_lm = LinearRegression()
taxi_lm.fit(train_X, train_y)

# print coefficients
print('intercept ', taxi_lm.intercept_)
print(pd.DataFrame({'Predictor': X.columns, 'coefficient': taxi_lm.coef_}))

# print performance measures
regressionSummary(train_y, taxi_lm.predict(train_X))
```

```
intercept  4.627801461423881
                   Predictor   coefficient
0             zTrip_Distance  1.005767e+00
1             PU_pickup_hour -3.140394e+10
2             DO_pickup_hour  3.140394e+10
3    weather_type_few clouds  1.136840e+00
4           weather_type_fog  9.465912e-01
```

```
# Use predict() to make predictions on a new set
taxi_lm_pred = taxi_lm.predict(valid_X)

result = pd.DataFrame({'Predicted': taxi_lm_pred, 'Actual': valid_y,
                       'Residual': valid_y - taxi_lm_pred})
print(result.head(20))

# Compute common accuracy measures
regressionSummary(valid_y, taxi_lm_pred)
```

```
      Predicted  Actual  Residual
5441   3.639682    3.64  0.000318
1244   0.353315    0.00 -0.353315
3648   0.244768    0.00 -0.244768
2982   1.219827    3.00  1.780173
709    1.366665    2.80  1.433335
```

Assessing the model's accuracy relied on metrics like Root Mean Squared Error

(RMSE) and r-squared (r2). These metrics gauged how well the model predicted tip amounts, forming a crucial component of our analytical framework. Ultimately, the Linear Regression model laid the groundwork for empowering our drivers with practical insights, representing a pivotal step in our efforts to enhance their experience and earnings.

When Linear regression Model is used, RMSE of 1.3823 and r2 of 0.55 was obtained

```
Regression statistics

               Mean Error (ME) : 0.0509
Root Mean Squared Error (RMSE) : 1.3823
     Mean Absolute Error (MAE) : 0.9020
```

```
]: print('r2 : ', r2_score(valid_y, taxi_lm_pred))
   print('adjusted r2 : ', adjusted_r2_score(valid_y, taxi_lm_pred, taxi_lm))
   print('AIC : ', AIC_score(valid_y, taxi_lm_pred, taxi_lm))
   print('BIC : ', BIC_score(valid_y, taxi_lm_pred, taxi_lm))
```

```
r2 :  0.545388929727596
adjusted r2 :  0.48630814768125075
AIC :  6145.312973137317
BIC :  7184.099615918954
```

Gradient Boosting Regressor model:

The Gradient Boosting Regressor model played a pivotal role in our project's predictive analysis for estimating tip amounts in taxi rides. Unlike Linear Regression, this model operates by combining multiple weak learners, typically decision trees, to create a stronger predictive model. It sequentially builds upon the errors of its predecessors, minimizing these errors with each successive iteration to improve predictions.

At its core, Gradient Boosting Regressor aims to optimize predictive accuracy. It constructs a series of decision trees, each one focusing on areas where the previous tree's predictions were lacking. By continuously adjusting and fine-tuning its predictions, this model effectively learns from its mistakes, resulting in a robust and accurate prediction of tip amounts.

XGBoost Model

```
import xgboost as xgb
```

```
# XGBoost regression model
xgb_reg = xgb.XGBRegressor(n_estimators=100, learning_rate=0.1)

# Train
xgb_reg.fit(train_X, train_y)

# Predict
y_pred = xgb_reg.predict(valid_X)
```

In our project, the Gradient Boosting Regressor was employed alongside Linear Regression to harness its ability to capture complex relationships between various ride-specific features and tip amounts. By training on historical taxi data, this model excelled in identifying intricate patterns and nuances within the dataset, providing our drivers with nuanced and precise estimations of potential tip earnings. The model's performance was evaluated using metrics like RMSE and r2 to ensure its effectiveness in predicting tip amounts accurately.

```
# Evaluate RMSE
from sklearn.metrics import mean_squared_error
rmse = np.sqrt(mean_squared_error(valid_y, y_pred))
print("RMSE: %f" % rmse)

r2 = r2_score(valid_y, y_pred)


print("R^2:", r2)
```

```
RMSE: 0.782562
R^2: 0.8542946805197011
```

This model served as a sophisticated complement to our predictive framework, contributing significantly to our goal of empowering drivers with reliable insights for their work optimization.

Capstone-project_modeling_Python Code

LEARNINGS:

- Design for scalability to accommodate growing data volumes and developed solutions to address complexities in merging diverse datasets with varying formats.
- Continuously monitor and optimize the warehouse for query performance and data retrieval.
- Regularly revisit and refine data models and processes based on feedback and changing requirements.
- Ensure data consistency and same granularity across sources to maintain integrity and reliability.

FUTURE RECOMMENDATIONS:

- Explore geospatial analytics to understand taxi movement patterns concerning demographics, weather conditions, and public transportation availability. This could uncover valuable insights about where and when taxis are most in demand.
- Explore avenues for even deeper integration of additional relevant datasets. Consider incorporating socioeconomic data, traffic patterns, or events happening in the city to provide a more comprehensive understanding of taxi usage and its correlation with various factors.

REFERENCES:

design-data-model-for-ride-sharing-or-taxi-service-5fc20d8eb424

https://www.kaggle.com/code/jasonlimcx/data- science-project-predictive-analytics-nyc-taxi

Whong, C. (2013) NYC Taxis: A Day in the Life.http://chriswhong.github.io/nyctaxi/#

Schneider, T. (2016) Taxi, Uber, and Lyft Usage inNew York City. http://toddwschneider.com/posts/taxi-uber-lyft-usage-new-york-city/

http://toddwschneider.com/posts/analyzing-1-1-billion-nyc-taxi-and-uber-trips-with-a-vengeance/

https://mavenanalytics.io/project/237

An Empirical Data Analytics and Visualization for UBER Services: A Data Analysis Based Web Search Engine | IEEE Conference Publication | IEEE Xplore

https://www.scirp.org/journal/paperinformation.aspx?paperid=94087

researchgate.net/profile/UmangPatel/publication/287205718_NYC_Taxi_Trip_and_Fare_Data_Analytics_using_BigData/links/567318ab08ae1557cf49472f/NYC-Taxi-Trip-and-Fare-Data-Analytics-using-BigData.pdf

https://www.kaggle.com/code/breemen/nyc-taxi-fare-data-exploration
https://www.kaggle.com/datasets/yasserh/uber-fares-dataset

https://www.nyc.gov/site/tlc/about/tlc-trip-record- data.page

https://medium.com/analytics-vidhya/building-a-linear-regression-model-on-the-new-york-taxi-trip-duration-dataset-using-python-2857027c54f3

https://medium.com/@adam.hajjej.ah/nyc-taxi-tip-amount-prediction-1bacf9eac920

https://python.plainenglish.io/new-york-city-taxi-and-limousine-commission-project-a57d3c0369b8