TAXI TRAJECTORIES: A PREDICTIVE APPROACH TO FORECASTING FARE, DURATION, AND DEMAND DYNAMICS IN NEW YORK CITY

DIVYA NEELAMEGAM, KEERTHANA RASKATLA, POOJITHA VENKATRAM,

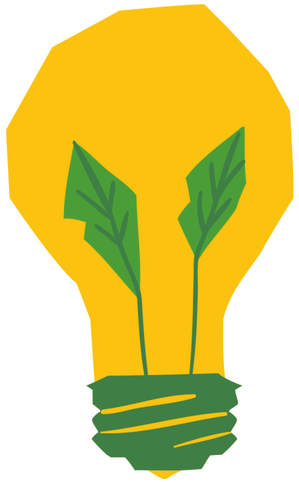
SREENIDHI POLINENI

GROUP-7

SAN JOSÉ STATE UNIVERSITY

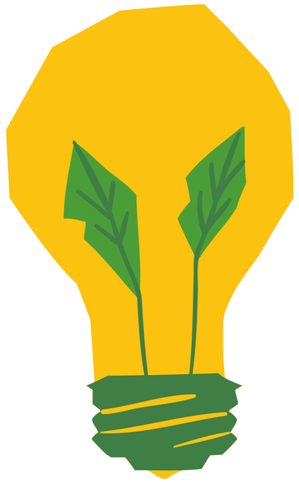
DR. SHAYAN SHAMS

MAY 7, 2024



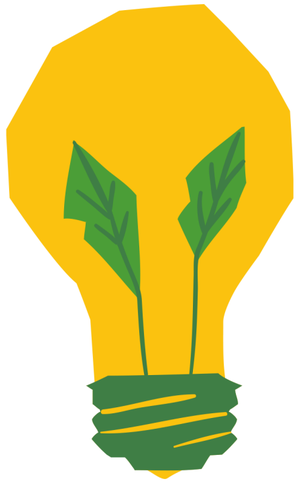
MOTIVATION

- **Reduce Wait Times**: Effective taxi dispatching can help drivers and passengers minimize the wait time to find each other.
- **Optimize Fleet Management:** Taxi operators can efficiently **allocate their fleet**, reducing idle time and ensuring a balanced distribution of taxis throughout the city.
- **Improve Customer Satisfaction:** With shorter wait times and greater taxi availability in key areas, passengers experience better service, leading to higher satisfaction and loyalty.
- **Traffic Congestion Reduction**: Accurate demand prediction allows for **better routing** and strategic fleet placement, potentially reducing unnecessary traffic congestion.
- **Increase Revenue**: By predicting areas and times of high demand, taxi drivers can position themselves where they are most likely to get fares, increasing their daily revenue while providing timely service.
- **Data-Driven Decision-Making**: Agencies can leverage demand prediction data in **policy-making**, ensuring efficient transportation systems that align with real-world needs. It's a window into the lifeblood of the city, offering endless possibilities for data-driven insights and urban planning.



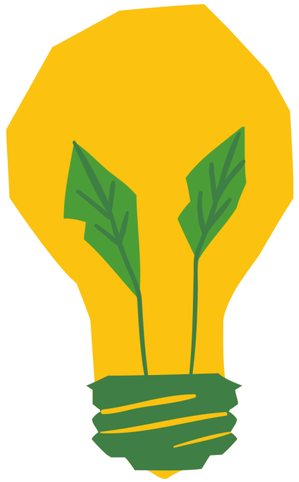
BACKGROUND

- Collected by the **New York City Taxi and Limousine Commission (TLC)**, covering millions of taxi rides over several years from 2009 to 2024.
- Includes detailed information such as **pickup/dropoff locations, trip distances, times, and fares.**
- Data sourced from GPS and payment systems, encompassing both yellow and green taxis.
- Offers rich insights into **urban traffic patterns, demand, and fare dynamics.**
- Valuable for analyzing transportation behaviour, **developing predictive models**, and enhancing urban mobility services.



LITERATURE REVIEW

Study Details	Dataset Used	Algorithms/Methods Used
Taxi Demand and Fare Prediction with Hybrid Models: Enhancing Efficiency and User Experience in City Transportation, Aug 2023 **Authors- Chou, Ka Seng, et al.**	New York City Taxi Data, over four months June to August 2018	A hybrid model combining LSTM-RNN with MDN forecasts taxi demand, while an ensemble model integrating multiple machine learning algorithms accurately predicts taxi fares.
Predicting the price of taxicabs using Artificial Intelligence: A hybrid approach based on clustering and ordinal regression models, May 2024 **Authors- Rathore, Bhawana, et al.**	New York City Taxi Data	A dynamic pricing mechanism by clustering cabs and converting fare predictions into quartiles, revealing differential predictor impacts.
New York City taxi trip duration prediction using MLP and XGBoost, July 2021 **Authors- Poongodi, M., et al.**	New York City Taxi Data	XGBoost and Multi-Layer Perceptron models for taxi trip duration-based predictions.
Back to the Past: Predicting Expected Ride Demand from Previous Dropoffs and Beyond, Nov 2023 **Authors- Hou, Emily, and Kimberly Hsueh**	New York City Taxi Data	Taxi demand forecasting by using key spatial and temporal features from NYC taxi data.
Taxi data in New York city: A network perspective, Feb 2016 **Authors- Deri, Joya A., and José MF Moura**	New York City Taxi Data, over a period of four years (2010-2013)	Explores movement patterns using graph Fourier transform for spectral analysis and identifies traffic co-behavior hotspots within the Manhattan road network.



METHOLODOLOGY

DATASET

NYC YELLOW TAXI TRIP RECORDS

- **Time Period:** January 2023 - December 2023
- **Count**: **The 2023 dataset describes 38,310,226 trips, averaging over 100,000 taxi rides across New York per day.**
- **Source:** **New York City Taxi & Limousine Commission (TLC)**
- **Purpose**: Provides comprehensive records of yellow taxi trips in New York City.
- **Content**: Includes details such as **trip timestamps, pickup and drop-off locations, trip distances, fare amounts**, and additional charges.
- **Usage**: Enables detailed analysis of taxi service trends, fare structures, and spatial-temporal demand patterns across NYC.

TAXI ZONE SHAPEFILE

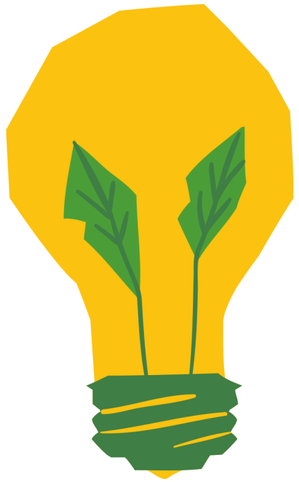
- **Purpose:** Defines geographical boundaries of designated taxi zones in New York City.
- **Content:** Includes unique identifiers for each zone, with attributes such as zone name and borough, facilitating precise spatial analysis.
- **Usage:** Essential for mapping and correlating taxi trip data to specific geographical locations, enhancing the analysis of traffic patterns and demand.

RAW DATASET

	VendorID	tpep_pickup_datetime	tpep_dropoff_datetime	passenger_count	trip_distance	store_and_fwd_flag	PULocationID	DOLocationID	payment_type	fare_amount	extra	mta_tax	tip_amo
0	2	2023-01-01 00:32:10	2023-01-01 00:40:36	1.0	0.97	N	161	141	2	9.3	1.00	0.5	0
1	2	2023-01-01 00:55:08	2023-01-01 01:01:27	1.0	1.10	N	43	237	1	7.9	1.00	0.5	4
2	2	2023-01-01 00:25:04	2023-01-01 00:37:49	1.0	2.51	N	48	238	1	14.9	1.00	0.5	15
3	1	2023-01-01 00:03:48	2023-01-01 00:13:25	0.0	1.90	N	138	7	1	12.1	7.25	0.5	0
4	2	2023-01-01 00:10:29	2023-01-01 00:21:19	1.0	1.43	N	107	79	1	11.4	1.00	0.5	3

```
full_df.columns
```

```
Index(['VendorID', 'tpep_pickup_datetime', 'tpep_dropoff_datetime',
       'passenger_count', 'trip_distance', 'RatecodeID', 'store_and_fwd_flag',
       'PULocationID', 'DOLocationID', 'payment_type', 'fare_amount', 'extra',
       'mta_tax', 'tip_amount', 'tolls_amount', 'improvement_surcharge',
       'total_amount', 'congestion_surcharge', 'airport_fee'],
      dtype='object')
```



DATA PREPROCESSING

- Implemented a threshold of $140 for the 'total_amount' feature, effectively filtering out extreme values while still allowing for inclusion of slightly higher fares, ensuring the dataset maintains a balance between capturing typical transactions and mitigating the influence of outliers.

- Processed “pickup_datetime” and “dropoff_datetime” data to derive key temporal attributes such as month, day of the month, day of the week, and hour, enhancing temporal analysis capabilities. Additionally, segmented pickup times into **four distinct time zones—Morning, Afternoon, Evening, and Late Night**—allowing for further refinement of time-based analysis.

- **Haversine Formula for Distance Calculation**: Utilizes the Haversine formula to accurately calculate the spherical distance between two points on the Earth’s surface, based on their latitude and longitude coordinates, which is critical for analyzing and optimizing route efficiency in the NYC dataset.

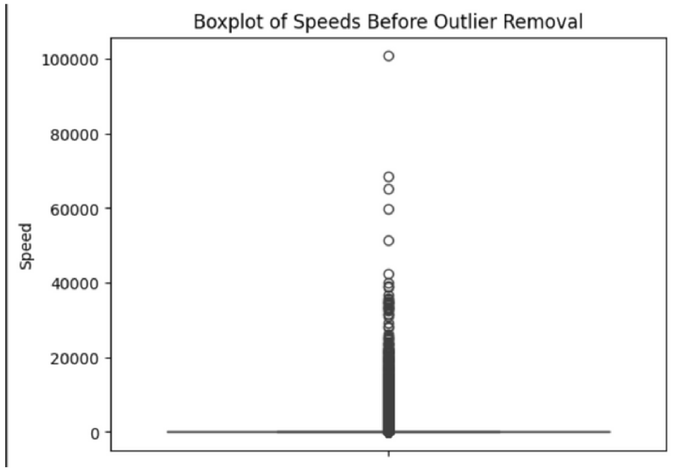
- Defined New York City's geographical boundaries with **latitude and longitude coordinates** of (40.5774, -74.15) for the southwest corner and (40.9176, -73.7004) for the northeast corner, ensuring that only pickup and dropoff coordinates falling within this range were included in the analysis, aligning with our focus on trips originating within New York City.

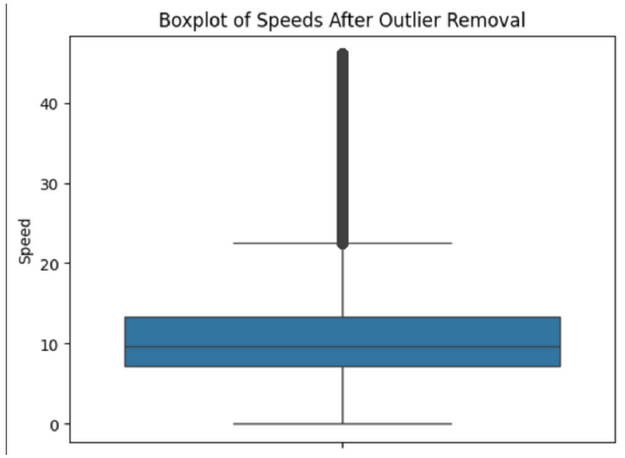
- Calculated the direction of travel from pickup to dropoff point using latitude and longitude. we have removed columns with highcorrelation,

- The outliers in the data was handled as follows,

passenger_count	
0.0	569663
1.0	27382732
2.0	5511322
3.0	1369416
4.0	771076
5.0	477904
6.0	313782
7.0	59
8.0	170
9.0	37

passenger_count	
1.0	27382732
2.0	5511322
3.0	1369416
4.0	771076
5.0	477904
6.0	313782

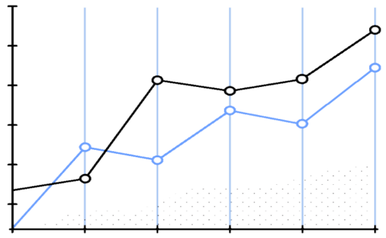




PREPROCESSED DATASET

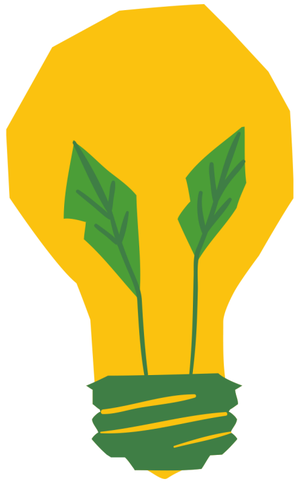
```
full_df.columns
```

```
Index(['VendorID', 'tpep_pickup_datetime', 'tpep_dropoff_datetime',
       'passenger_count', 'trip_distance', 'store_and_fwd_flag',
       'PULocationID', 'DOLocationID', 'payment_type', 'fare_amount', 'extra',
       'mta_tax', 'tip_amount', 'tolls_amount', 'improvement_surcharge',
       'total_amount', 'airport_fee', 'pickup_latitude', 'pickup_longitude',
       'dropoff_latitude', 'dropoff_longitude', 'direction', 'pickup_datetime',
       'dropoff_datetime', 'pickup_day', 'dropoff_day', 'pickup_day_no',
       'dropoff_day_no', 'pickup_hour', 'dropoff_hour', 'pickup_month',
       'dropoff_month', 'pickup_timeofday', 'dropoff_timeofday',
       'distance_miles', 'trip_duration'],
      dtype='object')
```



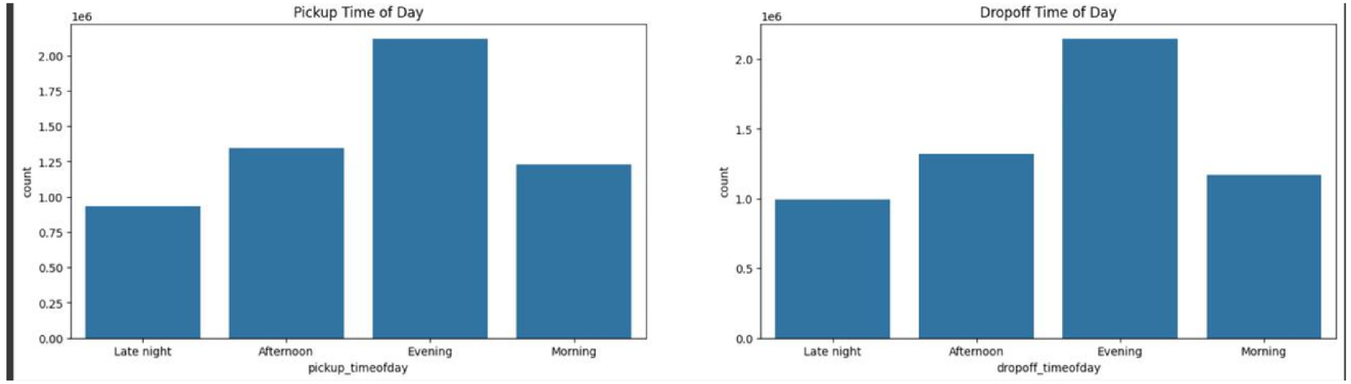
```
full_df.isnull().sum()
```

```
VendorID                 0
tpep_pickup_datetime     0
tpep_dropoff_datetime    0
passenger_count          0
trip_distance            0
store_and_fwd_flag       0
PULocationID             0
DOLocationID             0
payment_type             0
fare_amount              0
extra                    0
mta_tax                  0
tip_amount               0
tolls_amount             0
improvement_surcharge    0
total_amount             0
airport_fee              0
pickup_latitude          0
pickup_longitude         0
dropoff_latitude         0
dropoff_longitude        0
direction                0
pickup_datetime          0
dropoff_datetime         0
pickup_day               0
dropoff_day              0
pickup_day_no            0
dropoff_day_no           0
pickup_hour              0
dropoff_hour             0
pickup_month             0
dropoff_month            0
pickup_timeofday         0
dropoff_timeofday        0
distance_miles           0
trip_duration            0
dtype: int64
```

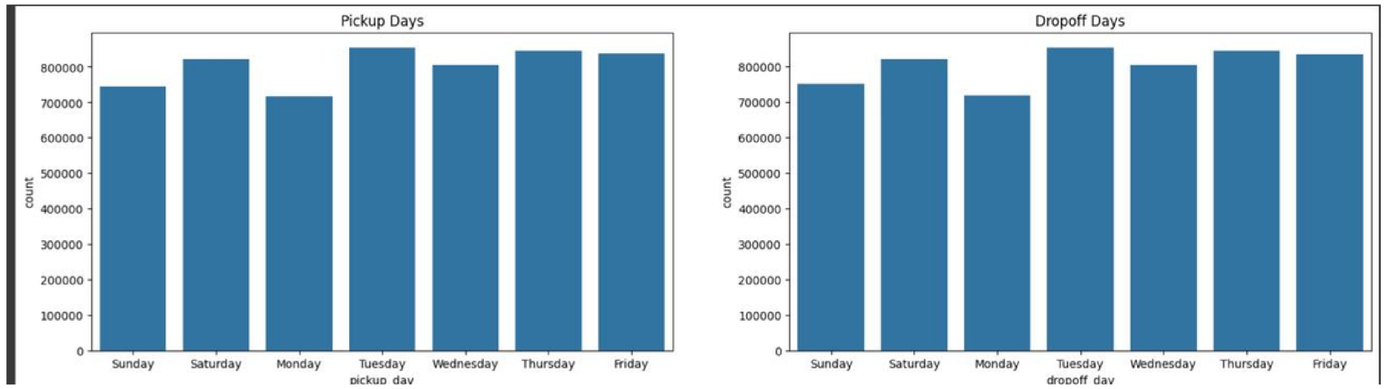


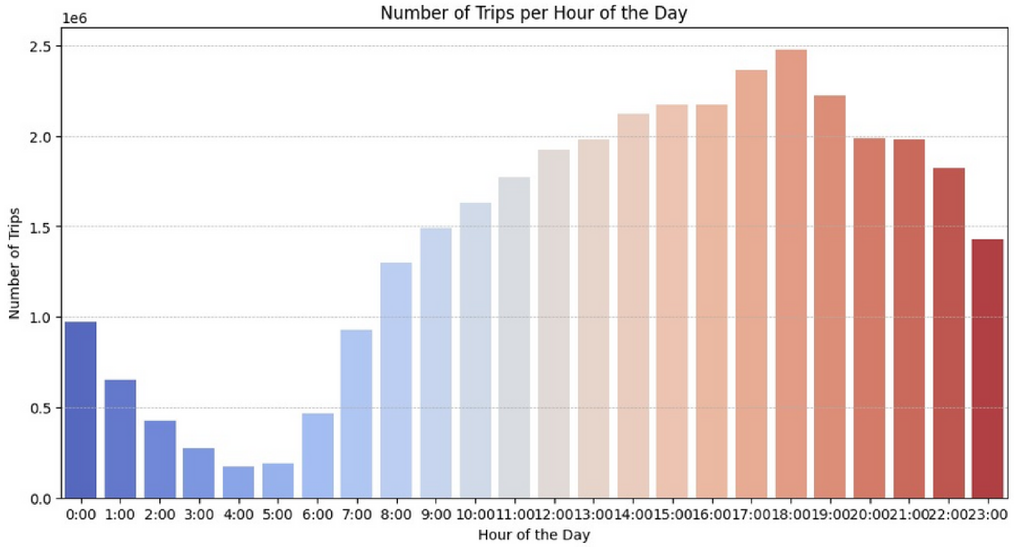
EXPLORATORY DATA ANALYSIS

- **Time of Day Analysis**: Evening periods experience the highest taxi activity, highlighting peak demand, while late night sees significantly reduced usage, suggesting off-peak hours could benefit from adjusted service strategies.



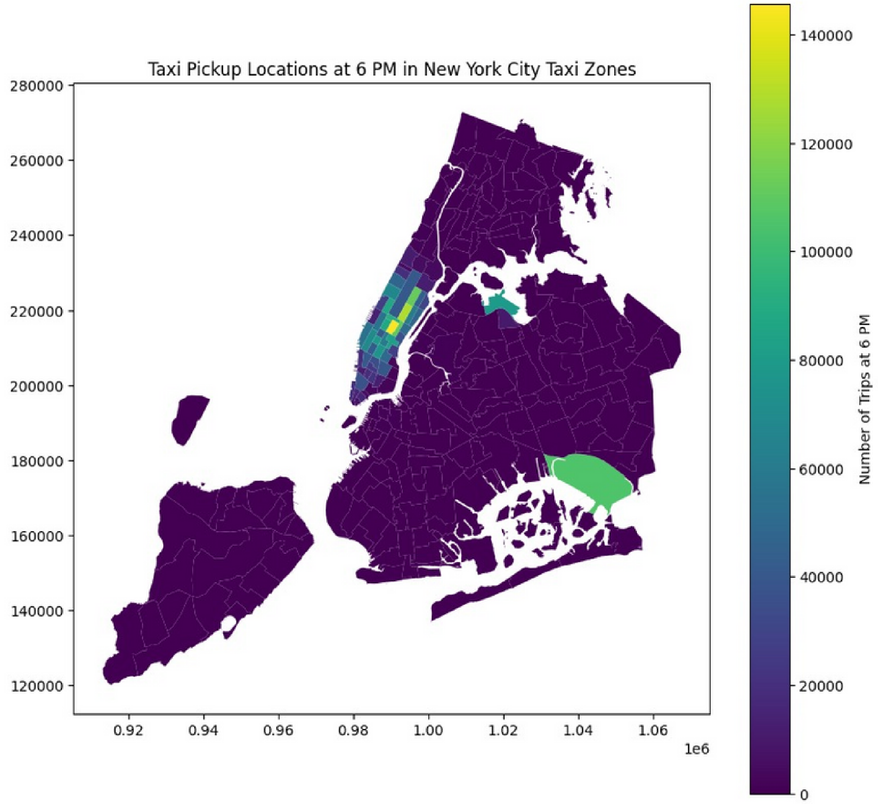
- **Day of the Week Analysis:** Weekdays show consistent taxi demand with a slight increase on Fridays, whereas weekends, especially Sundays, demonstrate a noticeable decline, indicating variable demand that could influence operational planning.





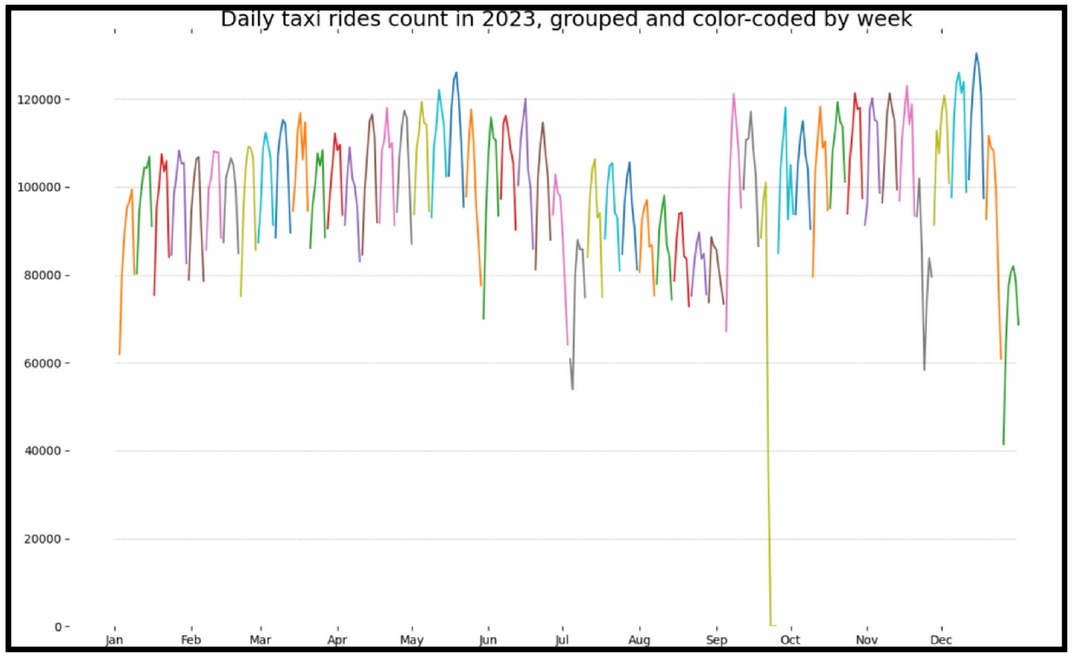
HOURLY DISTRIBUTION OF TAXI TRIPS

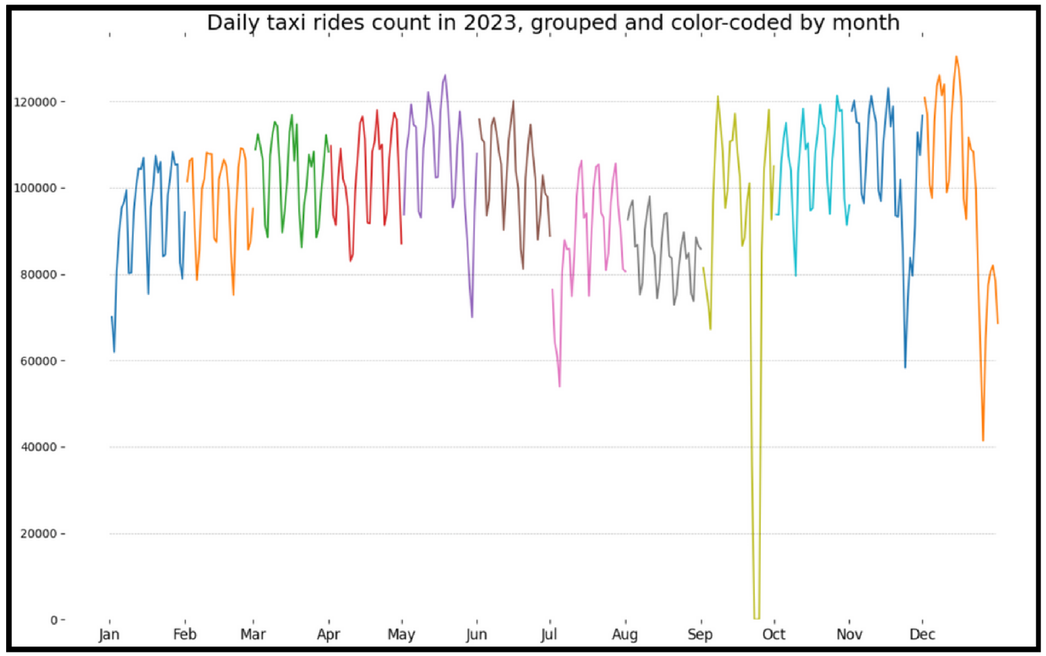
- **Peak Commute Times**: The number of taxi trips peaks in the late afternoon to early evening (around 6 PM), reflecting high commuter demand at the end of typical work hours.

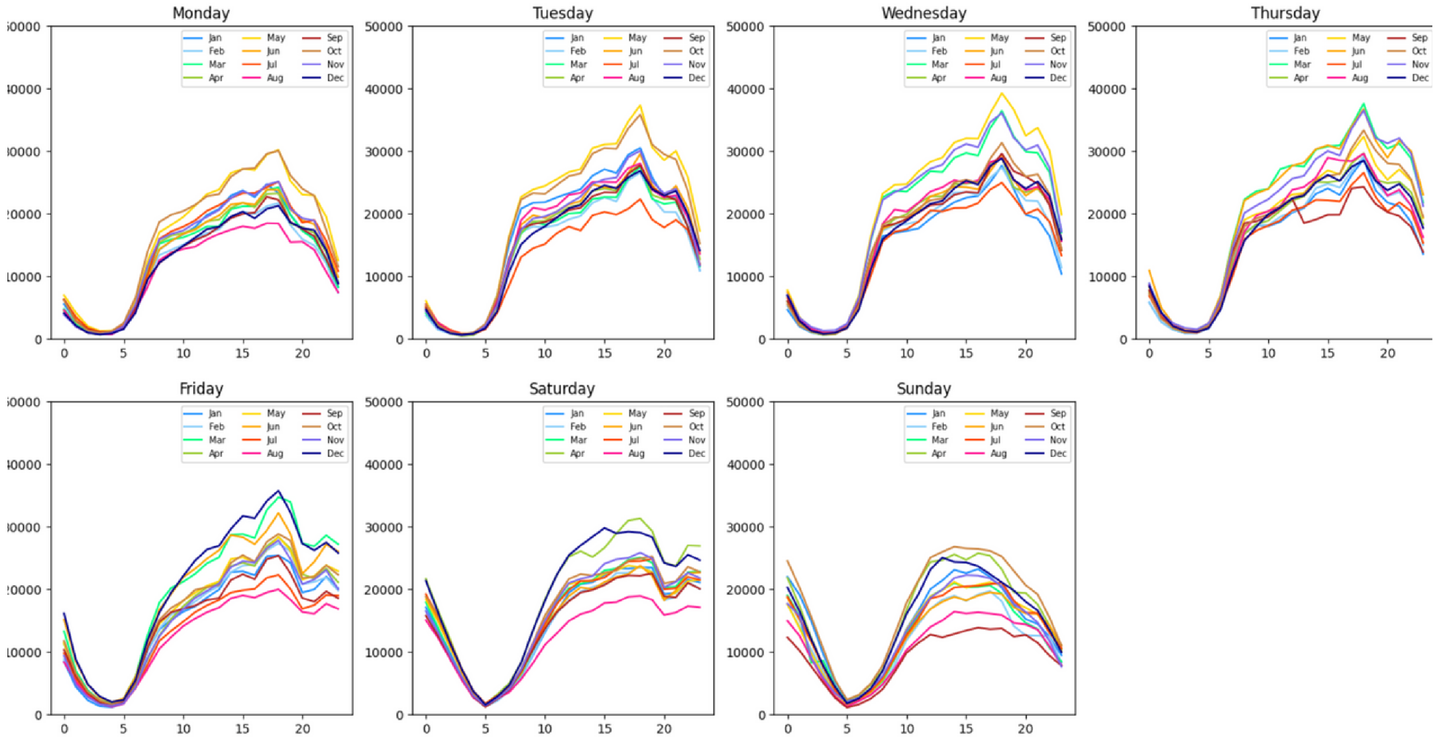


GEOGRAPHICAL DISTRIBUTION OF TAXI PICKUPS AT 6 PM

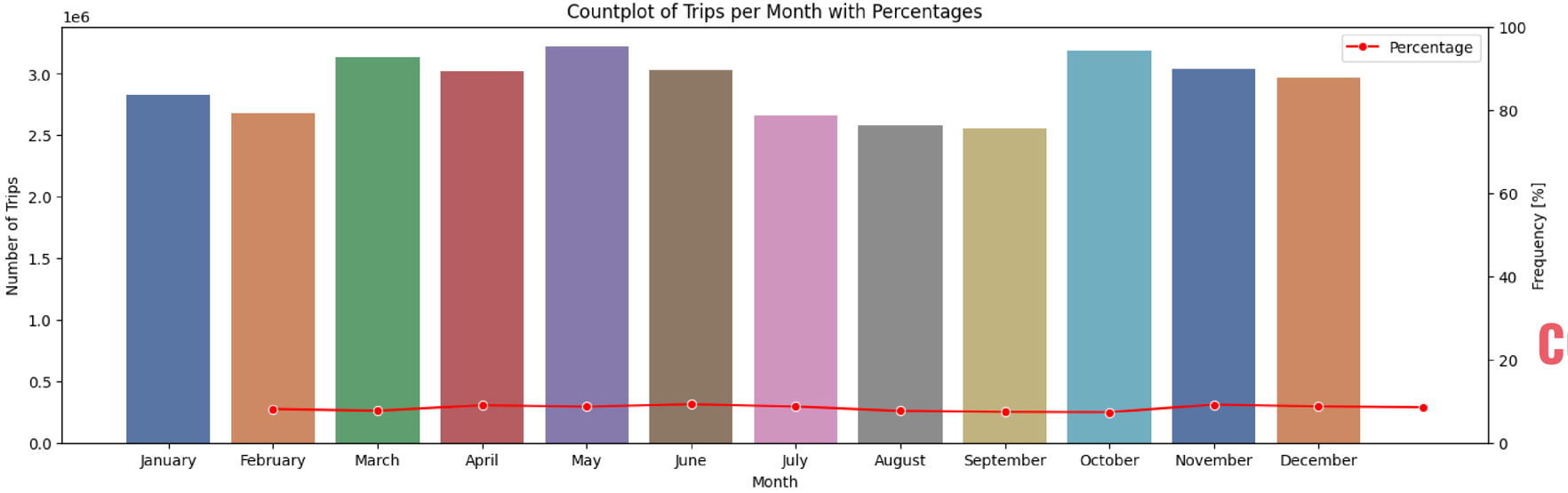
- **Concentration in Midtown Manhattan**: At 6 PM, there is a significant concentration of taxi pickups in midtown Manhattan, highlighting it as the primary hub of taxi activity during this peak time.





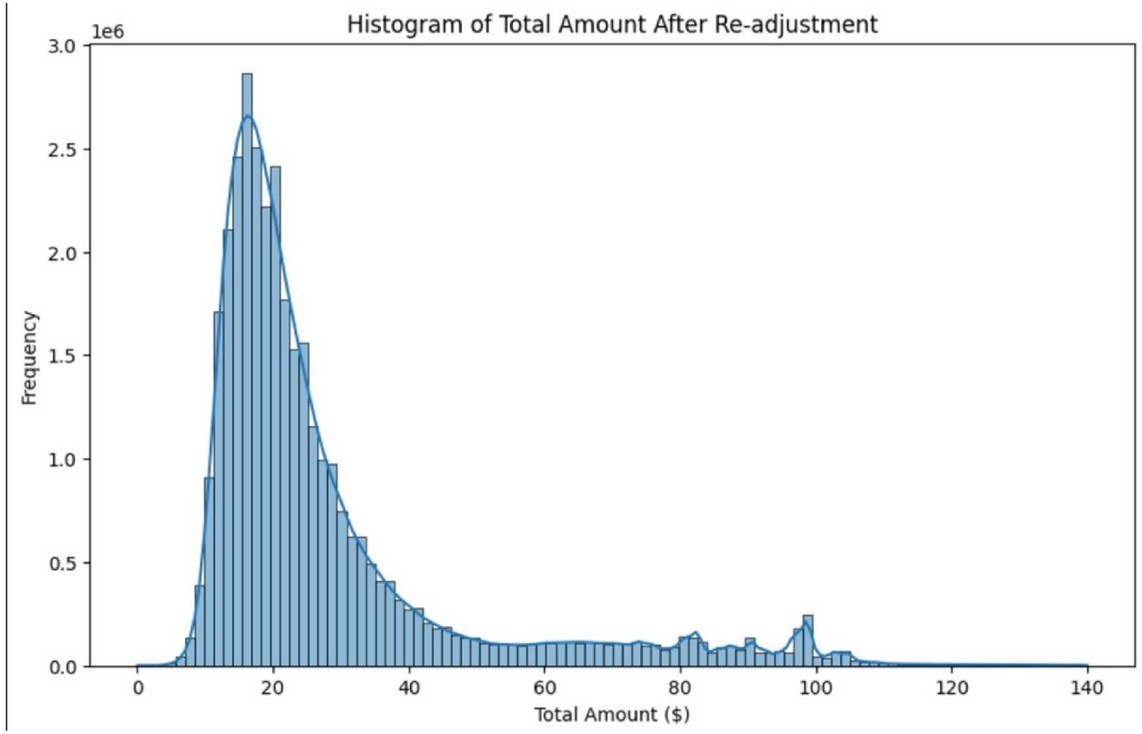


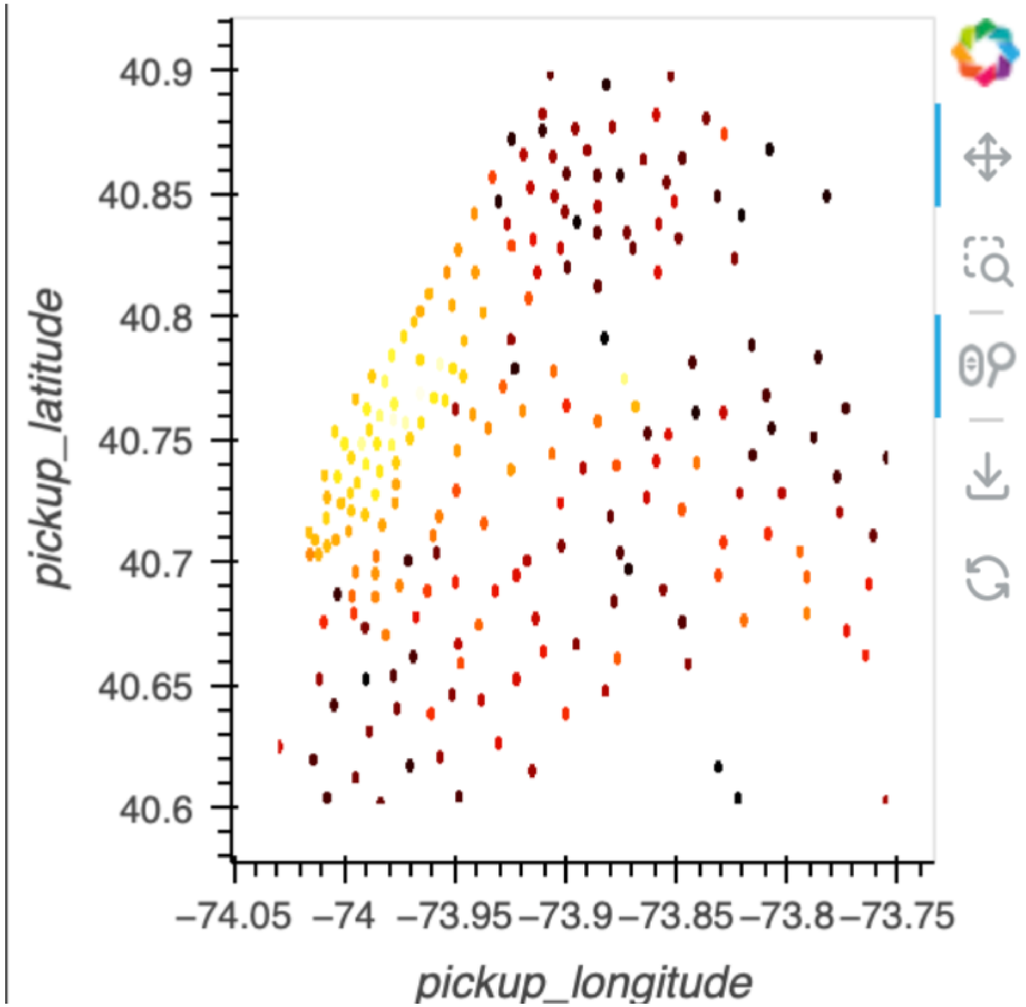
- Analyzing data by months, weeks, and days can uncover cyclical patterns and seasonal trends in taxi demand.
- It is easy to notice how each week follows a distinct pattern. Mondays start slow and the taxi demand peaks mid-week before falling back down for the weekend.
- For instance, identifying that mid-week peaks are consistent allows for strategic fleet positioning to meet demand efficiently.

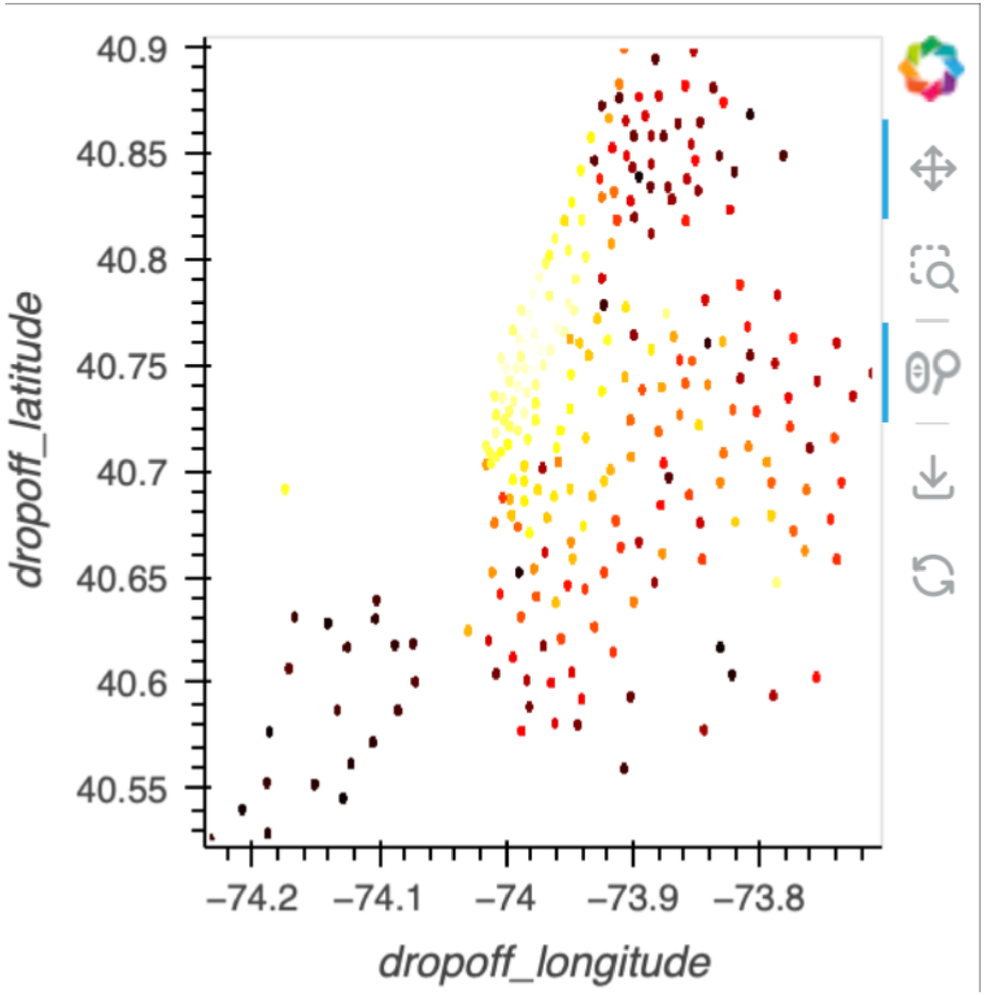


COUNTPLOT OF TRIPS PER MONTH

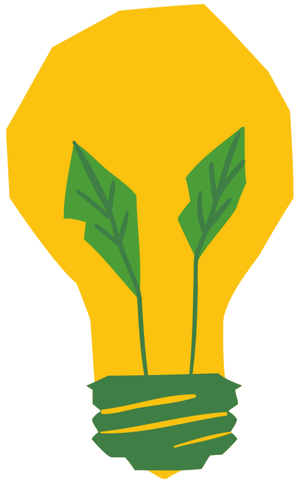
HISTOGRAM OF TOTAL AMOUNT AFTER RE-ADJUSTMENT







- These visualizations use Datashader to efficiently display the geographic distribution of NYC taxi pickups and dropoffs, highlighting demand hotspots.
- The `datashade` function aggregates points into a heatmap, while `dynspread` enhances visibility in dense areas.
- Thus, aiding in the strategic management of taxi fleet operations around these areas.



MODELS

Duration Prediction

- **Baseline Model: Linear Regression** - Establishes a simple baseline for performance comparison, achieving remarkably low validation and test Root Mean Square Error (RMSE) values of 6.83 and 9.27, respectively, indicating extremely high accuracy in predictive modeling.

- **XGBoost:** Implements the XGBoost algorithm, renowned for handling large datasets efficiently. The model uses a DMatrix data structure optimized for memory efficiency and speed, with hyperparameters like objective set to reg:squarederror for regression tasks, ensuring accurate predictions of trip durations.

- Employs systematic hyperparameter tuning to optimize the model, focusing on parameters such as eval_metric set to rmse (Root Mean Square Error) to evaluate performance. This approach enhances model accuracy by iteratively refining the learning process based on the validation set, leading to precise and reliable duration predictions.

Fare Amount Prediction

- Utilized Random Forest to construct advanced models surpassing the baseline Linear Regression fare predictor.

  Linear Regression Baseline: MAE: 0.8, RMSE: 1.2

- Achieved the lowest RMSE of 0.36 with Random Forest post PCA, showcasing superior predictive precision post-feature engineering and model optimization.

  Random Forest (after PCA): MAE: 0.06, RMSE: 0.36

- Highlighted pivotal features like trip_distance, pickup and dropoff time of day, distance_miles, pivotal in refining models and enhancing accuracy.

- Realized a substantial 70% reduction in RMSE compared to the baseline, affirming the efficacy of advanced modeling techniques and feature engineering in enhancing predictive accuracy.

ARIMA RESULTS

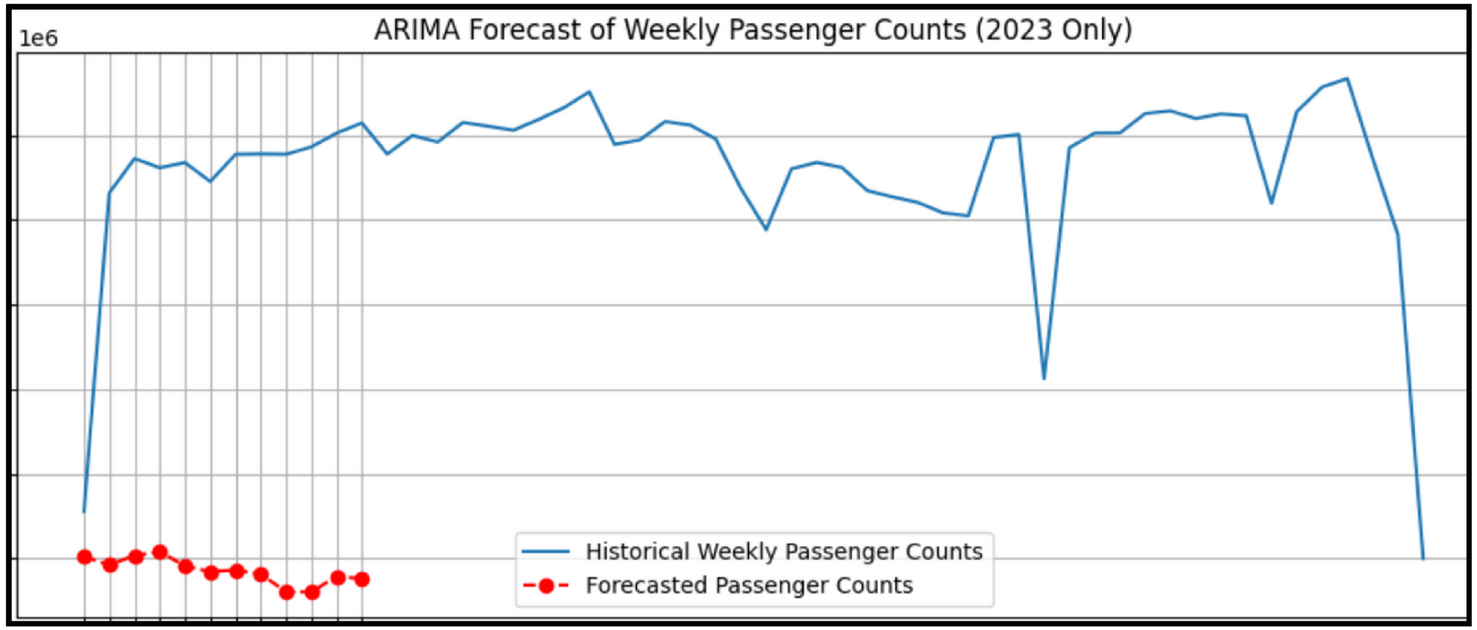
SARIMAX Results

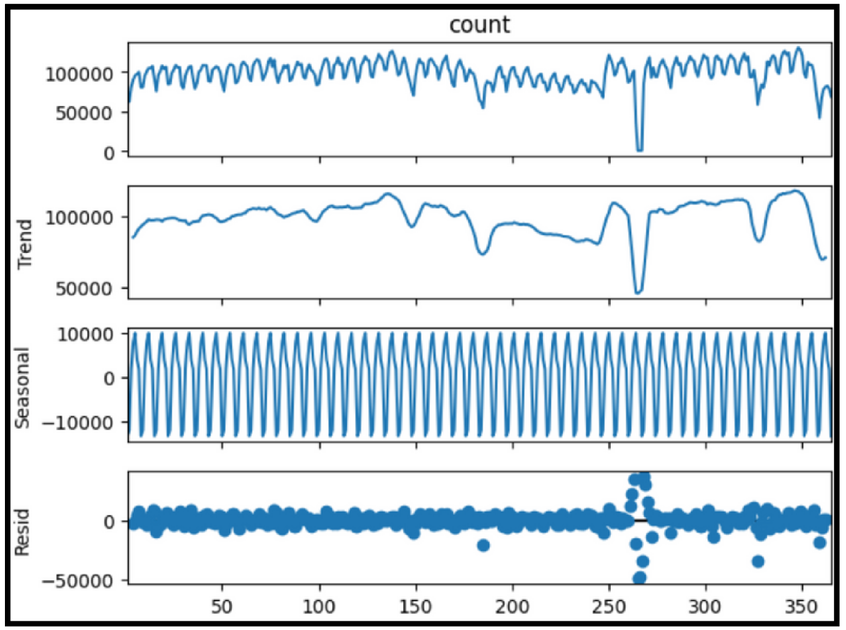
Dep. Variable:	y	No. Observations:	1201
Model:	SARIMAX(4, 1, 2)x(2, 0, [], 52)	Log Likelihood	-14278.730
Date:	Tue, 07 May 2024	AIC	28575.459
Time:	00:27:52	BIC	28621.270
Sample:	01-07-2001	HQIC	28592.716
	- 01-07-2024		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-2.0384	0.012	-169.635	0.000	-2.062	-2.015
ar.L2	-1.3041	0.024	-55.454	0.000	-1.350	-1.258
ar.L3	-0.3751	0.033	-11.508	0.000	-0.439	-0.311
ar.L4	-0.1393	0.018	-7.850	0.000	-0.174	-0.104
ma.L1	1.9625	0.010	198.713	0.000	1.943	1.982
ma.L2	0.9771	0.010	94.990	0.000	0.957	0.997
ar.S.L52	-0.8719	0.026	-33.655	0.000	-0.923	-0.821
ar.S.L104	-0.2899	0.307	-0.945	0.345	-0.891	0.311
sigma2	1.722e+09	2.78e-11	6.2e+19	0.000	1.72e+09	1.72e+09

Ljung-Box (L1) (Q):	0.05	Jarque-Bera (JB):	3925952.07
Prob(Q):	0.82	Prob(JB):	0.00
Heteroskedasticity (H):	959977617.46	Skew:	9.20
Prob(H) (two-sided):	0.00	Kurtosis:	282.61

- The ARIMA results indicate statistically significant model coefficients and reveal clear seasonal trends in weekly passenger counts.
- The forecast graph for 2023 demonstrates a strong alignment of predicted counts with historical patterns, despite anomalies likely from special events or irregularities.
- The diagnostic plots confirm well-calibrated residuals, affirming the model's reliability in predicting future demand.

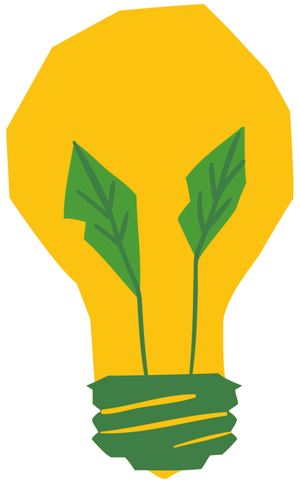




TEMPORAL DEMAND FORECASTING

Temporal Forecasting Approach

- **Models Used**: Implemented Linear Regression and Random Forest to predict hourly taxi demand based on temporal data.
- **Feature Engineering**: Utilized temporal features such as hour of the day and day of the week, which are critical for capturing patterns in taxi usage.
- **Linear Regression Model Performance**
  - Exhibits strong generalization capabilities, performing consistently across both training and testing datasets. This model effectively captures the underlying trends without fitting excessively to noise or outliers in the data.
  - Offers straightforward interpretability, allowing for easier understanding and explanation of the factors influencing demand forecasts.
- **Random Forest Model Performance**
  - Demonstrates a high capacity for learning complex patterns and relationships in the training data, suggesting robustness in handling diverse scenarios.
  - However, it tends to overfit the training data, leading to less effective performance on unseen test data, which underscores the need for tuning model parameters and possibly integrating regularization techniques to enhance generalization.



EXPERIMENTS AND RESULTS

MODEL RESULTS

DURATION PREDICTION

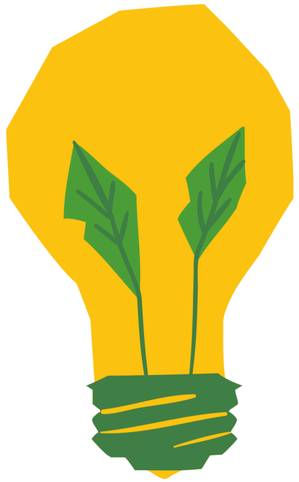
Model	Performance Evaluation
Linear Regression	Validation Root Mean Square Error (RMSE) : 4.312 Test Root Mean Square Error (RMSE) : 4.045
XGBoost	Validation Root Mean Square Error (RMSE) : 23.778 Test Root Mean Square Error (RMSE) : 18.67

FARE PREDICTION

Model	Performance Evaluation
Baseline model: Linear Regression	Validation Root Mean Square Error (RMSE) : 1.207 Test Root Mean Square Error (RMSE) : 1.209
Random Forest	Validation Root Mean Square Error (RMSE) : 0.362 Test Root Mean Square Error (RMSE) : 0.364

DEMAND FORECASTING

Model	Performance Evaluation
Baseline model: Linear Regression	Train Root Mean Square Error (RMSE) : 51.65 Train Mean Absolute Error (MAE) : 29.49 Test Root Mean Square Error (RMSE) : 39.93 Test Mean Absolute Error (MAE) : 31.32
Random Forest	Train Root Mean Square Error (RMSE) : 46.20 Train Mean Absolute Error (MAE) : 28.41 Test Root Mean Square Error (RMSE) : 123.03 Test Mean Absolute Error (MAE) : 82.67

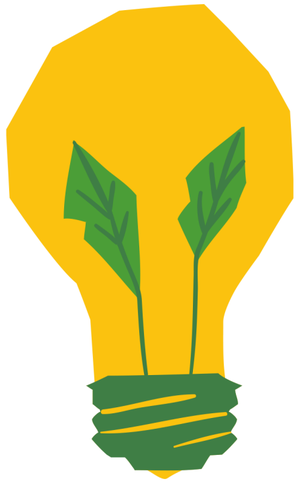


FUTURE WORK

- **Dynamic Ride Sharing Integration :** This could involve predicting optimal matches between passengers with similar destinations or travel routes to facilitate cost-sharing and reduce vehicle emissions.

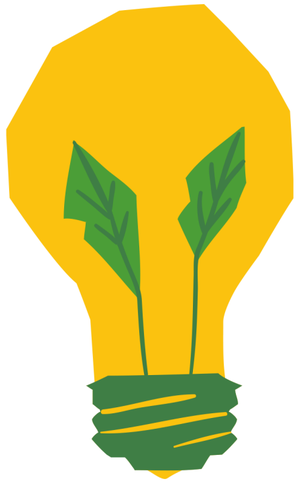
- **Autonomous Taxi Fleet Management:** By leveraging real-time data on traffic conditions, passenger demand, and fare dynamics, autonomous taxis could be dispatched efficiently to meet passenger needs while optimizing energy usage and reducing congestion.

- **Augmented Reality Taxi Navigation:** Develop augmented reality (AR) navigation systems for taxi drivers that incorporate predictive information on upcoming demand hotspots, traffic congestion, and optimal routing.



TEAM MEMBER CONTRIBUTION

Task	Divya	Keerthana	Poojitha	Sreenidhi
Background Research	Yes	Yes	Yes	Yes
Literature review	Yes	Yes	Yes	Yes
Data Collection	Yes	Yes	Yes	Yes
Data Processing	Yes	Yes	Yes	Yes
Modeling	Duration Prediction(Linear Regression, XGBoost)	Temporal Forecasting- Linear Regression, Random Forest	Clustering- K-Means, Time-Series Forecasting- ARIMA	Fare Prediction (Linear Regression Random Forest XGBoost)
Documentation	Yes	Yes	Yes	Yes



THANK YOU

HAVE ANY QUESTIONS?