**Title:** **Emissions Forecasting and Traffic Route Optimization to Reduce Emissions in New York City Using Uber Data**

**Team:**

Aakanksha Maheshwari - Aakanksha-Maheshwari

Aman Kumar - akumar63-wq

Bhavya Chheda - bhavyachheda2

Diksha Raina - draina27

Eniya Kulshreshtha -eniya19

**Introduction**

Urban transportation is a major contributor to greenhouse gas (GHG) emissions, particularly during congested peak office hours when slower vehicle speeds, prolonged idling, and stop-and-go movement exacerbate environmental harm. In cities like New York, addressing traffic congestion is critical for reducing emissions and improving urban mobility. This project focuses on Uber trip data during **7 AM to 9 AM**, a period representing the peak of office commutes, to analyze traffic patterns and propose strategies that can significantly reduce emissions.

The **primary stakeholders** for this project are **policymakers and transportation authorities** who require data-driven insights to design effective congestion management strategies and minimize environmental impact. **Secondary stakeholders** include **ride-hailing companies** like Uber, which benefit from improved operational efficiency through optimized routes, and **environmental agencies** working toward targeted emission reduction initiatives.

Our solution addresses these needs through a three-pronged approach:

1. **Trip-Level Analysis:** Identifying high-emission zones by analyzing Uber trip data.
2. **Emissions Forecasting:** Applying advanced time-series models to predict future emissions and provide actionable insights.
3. **Route Optimization:** Proposing optimized routes and assessing the impact of speed improvements on emission levels to highlight strategies for smoother traffic flow.

While this project focuses specifically on morning peak hours, it represents an **incremental step** toward addressing broader challenges of urban transportation. Future extensions could analyze evening rush hours or weekend travel patterns to provide a holistic solution for reducing congestion and mitigating environmental harm in urban areas.

.

**Literature Review**

Urban congestion and ride-hailing services have been identified as significant contributors to greenhouse gas (GHG) emissions. According to a 2020 report from the **Union of Concerned Scientists**, ride-hailing trips generate **69% more emissions than private car trips** due to deadheading—when drivers travel without passengers. This issue highlights the need for **route optimization** and **congestion management** to reduce unnecessary emissions and improve environmental outcomes.

Prior studies emphasize the critical role of **vehicle speeds** in fuel efficiency and emissions. The **U.S. Department of Energy** reports that lower average speeds, particularly in congested urban environments, result in frequent braking, idling, and increased fuel consumption, which directly exacerbate GHG emissions. These findings indicate that improving vehicle speeds through congestion mitigation strategies, such as optimized routing, can have a notable impact on reducing emissions.

Key insights derived from existing literature include:

1. **Urban congestion** remains a primary source of emissions due to prolonged idling and stop-and-go traffic.
2. **Speed optimization** enhances fuel efficiency and reduces emissions, particularly in urban environments.
3. **Route optimization** has been underexplored as a strategy in the context of ride-hailing emissions studies, despite its potential for emission reductions and traffic flow improvement.

Building on these findings, this project focuses on bridging the identified gaps by:

* Leveraging **ride-hailing trip data** to analyze emissions and pinpoint high-emission zones.
* Applying **emission forecasting models**, such as SARIMA and a hybrid SARIMA-XGBoost approach, to predict emissions trends and seasonal variations.
* Exploring **route optimization** combined with speed improvements to propose actionable strategies that address congestion and emissions simultaneously.

By integrating these methods, this project not only builds upon existing research but also fills a critical gap by combining emissions forecasting with route optimization. The focus on publicly available datasets ensures scalability and applicability, making the findings relevant to **policymakers**, **environmental agencies**, and **ride-hailing companies**. Policymakers gain actionable insights to improve urban traffic flow, environmental agencies receive targeted solutions for emissions reduction, and ride-hailing companies can optimize routes to enhance fuel efficiency and operational performance.

**Data:**

The dataset for this project was obtained from the **NYC Taxi and Limousine Commission (TLC) Trip Record Data** (<https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page>), a credible source that provides comprehensive and government-regulated trip records. The data includes monthly Uber trip records from **January 2020 to August 2024**, capturing critical attributes such as trip distances, timestamps, and location information. These attributes align directly with the project’s objective of analyzing traffic emissions during peak office hours.

Given the large size of the data, individual monthly files were first merged into yearly datasets and then combined into a **comprehensive working dataset** containing approximately **56 million rows**. To align with the focus on urban congestion during peak office hours, trips between **7 AM and 9 AM** were specifically filtered, resulting in a reduced and manageable subset of the data. ([AML\_Project](https://sumailsyr-my.sharepoint.com/:f:/r/personal/aamahesh_syr_edu/Documents/Syracuse%20sem%203/AML/AML_Project?csf=1&web=1&e=HcpVtm))

**Data Cleaning and Preprocessing**

To ensure the quality and consistency of the data, the following preprocessing steps were performed:

1. **Missing Value Removal:** Rows with null or incomplete values were excluded.
2. **Outlier Handling:** Outliers were treated using the **Interquartile Range (IQR)** method to retain meaningful trip-level data while removing extreme values.
3. **Duplicate Removal:** Duplicate trip records were identified and removed to avoid redundancy.
4. **Date Standardization:** Timestamps for pickup and drop-off were formatted to a consistent **datetime** format, ensuring temporal analysis accuracy.

**Feature Engineering**

To address the lack of direct emissions data, additional features were derived to facilitate analysis and modeling:

1. **Estimated Emissions:** Calculated using the formula:  
   Estimated Emissions=Trip Miles×440\text{Estimated Emissions} = \text{Trip Miles} \times 440Estimated Emissions=Trip Miles×440  
   The constant **440 grams of CO₂ per mile** is based on average emissions reported by the **U.S. Environmental Protection Agency (EPA)** for gasoline-powered vehicles. This derived feature serves as the **dependent variable** for forecasting models.
2. **Average Speed:** Trip-level speed was calculated as:  
   Average Speed=Trip MilesTrip Time (in hours)\text{Average Speed} = \frac{\text{Trip Miles}}{\text{Trip Time (in hours)}}Average Speed=Trip Time (in hours)Trip Miles​  
   This feature was crucial for analyzing the impact of congestion (low speeds) on emissions.
3. **Emission Levels:** Estimated emissions were categorized into three levels: **Low, Medium, and High**, based on defined thresholds to facilitate classification analysis.
4. **Speed Categories:** Trips were classified as **Congested**, **Moderate**, or **Efficient** based on average speed thresholds to support congestion analysis.

**Relevant Features**

The following features were retained as critical variables for emissions forecasting and route optimization:

* **Trip Miles:** Total distance traveled during a trip.
* **Trip Time:** Total duration of the trip (in minutes).
* **Pickup and Drop-off Datetime:** Timestamps for trip start and end.
* **Estimated Emissions:** Derived CO₂ emissions (grams), serving as the target variable.
* **Average Speed:** Derived trip speed for congestion analysis.
* **Estimated Levels:** Categorized emissions levels (Low, Medium, High).
* **Speed Category:** Categorized trip speeds to study congestion patterns.

**Data Insights from Visualizations**

1. **Hourly Trip Distribution:** Trip activity peaks during **5 PM–7 PM** and rises sharply between **7 AM–9 AM**, validating the focus on peak office hours for emissions analysis.
2. **Distance vs. Emissions Relationship:** A positive correlation was observed, where longer trips contribute disproportionately to emissions, exceeding **1100 grams** for trips over 2.4 miles.
3. **Emissions by Day of the Week:** Weekdays exhibited stable emission levels, while weekends showed slightly higher emissions, likely due to increased leisure activity.
4. **Impact of Speed on Emissions:** Emissions were highest when **average speeds fell below 14 mph**, emphasizing the role of congestion in increasing emissions.
5. **Correlation Heatmap:** Strong positive correlations were identified between **trip distance**, **trip time**, and **estimated emissions**, validating these as key predictors for modeling.

A graph and chart of a graph

Description automatically generated with medium confidence

A graph showing the growth of a company

Description automatically generated with medium confidence

A graph showing different colored squares

Description automatically generated

A graph showing the impact of average emission

Description automatically generated

A diagram of a heatmap

Description automatically generated

**Methods:**

**Exploratory Data Analysis (EDA)**

To understand the factors driving greenhouse gas (GHG) emissions and traffic congestion, a comprehensive **Exploratory Data Analysis (EDA)** was conducted. The analysis was focused on identifying high-emission zones, evaluating daily traffic patterns, and studying the relationships between trip characteristics and emissions.

Initially, **pickup and drop-off locations** were analyzed to pinpoint zones with the highest emissions. Airports emerged as dominant contributors, with **JFK Airport** and **LaGuardia Airport** in Queens consistently ranking as the top pickup and drop-off zones. These locations act as major travel hubs, generating significant emissions due to high passenger turnover and longer travel distances. This finding is particularly valuable for policymakers who can prioritize these areas for traffic interventions. For Uber, optimizing routes to reduce idling at airports can enhance efficiency and minimize emissions. In addition to airports, **Brooklyn** neighborhoods such as **East New York**, **Crown Heights North**, and **Canarsie** were identified as high-emission zones for pickups. Similarly, in **Manhattan**, neighborhoods like **Central Harlem North**, **TriBeCa/Civic Center**, and **Washington Heights South** experienced high pickup activity. For drop-offs, **non-NYC zones** (outside the city) emerged as a significant factor, indicating emissions from long-distance suburban or intercity trips. Additionally, busy urban zones such as **Midtown Manhattan** (including Times Square and Theater District) contributed heavily to drop-off emissions due to traffic congestion from business and tourism activities.

To evaluate the economic and environmental trade-offs, a **cost-benefit analysis** of the congestion surcharge was conducted. The analysis revealed that the surcharge generated approximately **$48.5 million** in revenue. However, the environmental cost due to additional emissions caused by congestion amounted to **$399,649**, with further costs stemming from lost time for drivers ($1.58 million) and passengers ($1.31 million). Despite these costs, the surcharge yielded a **net benefit** of nearly **$46 million**, highlighting its financial success while emphasizing the need to address associated environmental impacts.

Hourly traffic dynamics were analyzed to identify **peak hours** for trip activity. The findings indicated that evening hours, specifically **5 PM to 7 PM**, recorded the highest number of trips, with over 3.7 million trips at **5 PM** alone. Morning hours, particularly **7 AM to 9 AM**, also showed sharp increases in trip volumes, aligning with office commute hours. Conversely, trip activity remained minimal during early morning hours (12 AM–5 AM), reflecting low demand. This analysis validated the decision to focus on **7 AM to 9 AM** as a critical time window for emissions reduction, as targeting these hours would directly address peak office-hour congestion.

The analysis of **trip distances, fares, and emissions** revealed a positive relationship between these variables. Short trips (0.55–1.03 miles) were associated with lower emissions, typically under **500 grams**, while longer trips (2.4–2.9 miles) produced emissions exceeding **1100 grams**. This relationship suggests that longer trips disproportionately contribute to total emissions, emphasizing the importance of optimizing longer routes and encouraging pooled rides to reduce environmental impacts without compromising operational efficiency.

Further examination of emissions **by day of the week** showed stable patterns during weekdays, driven primarily by routine work commutes. However, emissions were slightly higher on weekends, particularly **Saturdays and Sundays**, likely due to leisure and recreational trips. This trend highlights opportunities for optimization outside of office hours, as weekend traffic remains an overlooked area for emissions reduction efforts.

The relationship between **average speed and emissions** was analyzed to assess the impact of congestion. Analyzing speed impacts was necessary because prior studies indicate that low average speeds exacerbate fuel inefficiencies, increasing emissions. This aligns with our focus on identifying actionable traffic flow interventions. Emissions were observed to be highest at **low average speeds** (below 14 mph), reflecting the effects of stop-and-go traffic during congestion. In contrast, as average speeds increased beyond **14 mph**, emissions decreased significantly due to smoother traffic flow and reduced idling times. This analysis underscores the importance of traffic flow optimization strategies, such as improving signal timings, managing congestion-prone areas, and increasing vehicle speeds, even by small margins, to achieve notable reductions in emissions.

Lastly, a **correlation heatmap** was generated to identify the relationships between key features in the dataset. Strong positive correlations were observed between **trip distance**, **trip time**, and **estimated emissions** (~0.87), confirming that longer trips are primary contributors to emissions. Similarly, base passenger fares showed strong correlations with trip distance and time (~0.91), reflecting fare structures tied to trip lengths. In contrast, **pickup and drop-off location IDs** exhibited weak correlations with emissions, indicating that location alone is not a strong predictor. These findings validated the selection of key features—such as trip distance, average speed, and trip duration—as primary drivers of emissions, while also highlighting the need for efficient route planning to reduce travel distances and emissions.

In summary, the EDA provided valuable insights into emission hotspots, peak congestion hours, and the relationships between trip characteristics and emissions. The identification of high-emission zones, peak activity periods, and speed-emission relationships informed subsequent modeling and optimization strategies, ensuring the study addressed emissions reduction in a targeted and data-driven manner.

## **Data Modelling**

To address the goals of forecasting greenhouse gas (GHG) emissions for 2025 and identifying high-emission zones for optimized traffic management, a multi-step modeling and analytical approach was implemented. The two key components of the project were emissions forecasting and route optimization. Several techniques were explored, starting with preprocessing and progressing to advanced models to ensure the results aligned with stakeholder needs.

The first stage involved **data preprocessing** to ensure the emissions data was suitable for time-series forecasting and clustering analysis. Emissions data was aggregated monthly, as this format best captured long-term trends and seasonal variations. A **stationarity check** was performed using the **Augmented Dickey-Fuller (ADF) test**, which confirmed that the data exhibited non-stationary behavior. To address this, first-order differencing was applied, transforming the data into a stationary form necessary for time-series models. Additional features, such as pickup location IDs, trip distances, and average speeds, were extracted for clustering analysis. These features were scaled using **Min-Max Scaling** to normalize values across the dataset. Missing values were addressed through forward fill for time-series data and median imputation for other features, ensuring data quality and consistency.

For **emissions forecasting**, multiple models were tested to determine the best-performing method. Initially, **XGBoost** was applied as a baseline due to its ability to model complex non-linear relationships. However, XGBoost struggled with the sequential nature of time-series data, producing a high RMSE of **60.398**. Its failure stemmed from the lack of explicit mechanisms to capture trends and seasonality, both of which are fundamental in emissions data. Recognizing this limitation, the **ARIMA (Auto-Regressive Integrated Moving Average)** model was introduced. ARIMA models time-series data by accounting for trends and autoregressive relationships. After differencing the data to achieve stationarity, ARIMA parameters (p, d, q) were optimized using the **Akaike Information Criterion (AIC)**. ARIMA significantly improved the RMSE to **10.841** but failed to capture recurring seasonal patterns, leaving room for further refinement.

To overcome ARIMA’s limitations, **SARIMA (Seasonal ARIMA)** was implemented, as it explicitly models seasonal components in time-series data. SARIMA was chosen due to its ability to model both seasonal trends and long-term patterns, which were evident during the EDA. This method is well-documented in time-series forecasting literature for handling emissions data with recurring patterns. SARIMA parameters were configured to include non-seasonal and seasonal components, with the seasonal period set to **12 months**. Model diagnostics, including AIC (195.526) and BIC (202.362), confirmed a good model fit with balanced complexity and accuracy. Residual analysis demonstrated that the errors were uncorrelated and homoscedastic, validating SARIMA’s performance. SARIMA achieved an RMSE of **6.674**, effectively capturing trends and seasonal patterns in the data. However, while SARIMA excelled in modeling linear patterns, it could not address residual non-linear relationships present in the data.

To further improve forecasting accuracy, a **SARIMA-XGBoost hybrid model** was developed. The hybrid approach combined SARIMA’s strength in capturing linear trends and seasonality with XGBoost’s ability to model non-linear residuals. The hybrid approach addresses limitations of standalone models. SARIMA models linear seasonality, while XGBoost complements it by capturing complex non-linear relationships. This aligns with methods suggested in prior research on combining statistical and machine learning models for enhanced forecasting accuracy. SARIMA was first applied to predict emissions, and the residuals (errors) were extracted. These residuals were then passed as input to XGBoost, which modeled the complex, non-linear relationships not captured by SARIMA. The final predictions were obtained by adding XGBoost’s residual corrections to SARIMA’s outputs. This hybrid approach provided the most accurate results, achieving an RMSE of **3.263** for 2024. When extended to forecast emissions for 2025, the RMSE increased to **84.04**, reflecting the inherent uncertainty in long-term predictions due to external factors such as traffic policies or infrastructure changes.

The second stage of the project focused on **route optimization** to identify high-emission zones and explore potential strategies for emission reduction. Identifying these high-emission zones allows policymakers to prioritize infrastructure improvements and helps Uber optimize routes to reduce congestion-related delays, thereby lowering operational costs and emissions. **K-Means clustering** was applied to group pickup locations based on their emissions and additional features such as trip distance and average speed. The dataset was divided into two clusters: **Cluster 0**, representing high-emission zones requiring immediate attention, and **Cluster 1**, comprising low-emission zones with smoother traffic flow. Locations such as **JFK Airport** and **East New York** emerged as significant contributors to emissions, making them priority areas for targeted interventions.

Further analysis was conducted to understand the **impact of vehicle speed on emissions** in high-emission zones (Cluster 0). The relationship between average speeds and emissions revealed that vehicles traveling at speeds below **14 mph** experienced significantly higher emissions due to prolonged travel times and stop-and-go traffic. Simulations were performed to evaluate the impact of speed improvements: a **10% increase in average speed** led to approximately **15% reduction in emissions**, while a **20% increase** resulted in **30% reduction**. These results emphasized the importance of traffic flow improvements, such as signal optimization, congestion management strategies, and infrastructure upgrades, to reduce emissions effectively.

Finally, **alternate route identification** was carried out for high-emission zones to explore options for diverting traffic. For pickup and drop-off locations, nearby alternate routes were identified where feasible. Zones such as **Brooklyn – East New York** and **Queens – Astoria** had viable alternatives, while locations in **Manhattan** lacked feasible alternate routes. The lack of feasible alternate routes in Manhattan highlights infrastructure limitations, suggesting that policymakers need to focus on expanding traffic diversion options to address emissions in high-congestion zones. This highlighted the need for additional infrastructure development in heavily congested urban areas to support emissions reduction strategies.

**Results Summary**

The project results directly align with the needs of **policymakers**, **environmental organizations**, and **Uber**. The **emissions forecasting** for 2024–2025 highlights seasonal peaks, allowing policymakers to implement targeted traffic policies and environmental organizations to plan sustainability interventions. The SARIMA-XGBoost hybrid model achieved an RMSE of **3.263** for 2024 and **84.04** for 2025. This highlights the model's ability to balance seasonal trends and residual non-linear relationships, providing reliable forecasts for emissions reduction planning. The **route optimization analysis** demonstrates that increasing average vehicle speeds by **10–20%** reduces emissions by **15–30%**, benefiting both environmental goals and Uber's operational efficiency. Additionally, alternate route identification for high-emission zones like **Brooklyn – East New York** and **Queens – Astoria** offers actionable traffic diversion strategies. However, areas lacking feasible alternatives, such as **Manhattan – Civic Center**, emphasize the need for traffic flow improvements and infrastructure planning.

|  |  |  |
| --- | --- | --- |
| **Stakeholder** | **Need** | **Results Addressing Need** |
| **Policymakers** | Reduce congestion and improve traffic flow. | - Emissions forecasts pinpoint peak periods for intervention. |
|  | Plan effective infrastructure improvements. | - Route optimization identifies areas needing infrastructure upgrades. |
| **Environmental Organizations** | Lower CO₂ emissions for better air quality. | - Speed increases reduce emissions by **15–30%**. |
|  | Target high-emission zones for intervention. | - Forecast trends and high-emission zones allow targeted sustainability programs. |
| **Uber Company** | Optimize operational efficiency. | - Speed improvements reduce fuel costs and travel times. |
|  | Identify better traffic routes. | - Alternate route suggestions for high-emission zones like Brooklyn and Queens improve service efficiency. |

A graph showing a line

Description automatically generated

A graph with a line

Description automatically generated

A black and white text

Description automatically generated

A graph showing a graph of green and red stripes

Description automatically generated with medium confidence

**DISCUSSIONS**

By focusing on the 7–9 AM window, the project successfully identified actionable insights for emissions reduction during peak congestion periods. While this narrow focus achieved targeted results, expanding the scope could provide a more holistic solution. decision was made to better understand traffic patterns during the most congested times, which are critical for effective policymaking and transportation planning.

By running the model specifically during peak hours, stakeholders can gain more focused insights. This allows transportation authorities to design policies that directly address congestion and improve commuting efficiency.

Focusing on peak hours helps policymakers:

1. Identify areas with heavy traffic congestion and plan targeted solutions.
2. Create optimized routes to reduce travel time and improve road safety.

The results enable policymakers to prioritize high-emission zones like airports, environmental agencies to plan sustainability measures during peak emissions hours, and Uber to improve operational efficiency through route optimization and speed management.

**LIMITATIONS**

While the project provides valuable insights, there are some limitations to note:

1. **Limited Time Window:**  
   We only analyzed data from peak hours (7 AM to 9 AM). This gave us targeted insights but did not provide a full picture of traffic trends throughout the day. Analyzing the entire dataset could have offered more comprehensive findings.
2. **Optimization Methods:**  
   We did not explore alternative clustering methods for route optimization. Trying other methods could have led to more efficient solutions.
3. **Model Performance:**  
   We did not perform hyperparameter tuning on our models. Fine-tuning these settings could have improved the accuracy and overall performance of the results.
4. **Emissions Data:**  
   Real emissions data was not available, so we used proxy variables. This approach may have introduced some variability, reducing the precision of our results.
5. **Route Analysis:**  
   We did not analyze alternative routes in detail, which limited the depth of our recommendations. Additionally, we did not consider cost factors, even though cost is important for real-world decision-making.

**FUTURE WORK**

In future work, we can make several improvements to expand the project and enhance its effectiveness:

1. **Incorporate Cost Factors:**  
   Including cost considerations in route optimization will make our solutions more practical and applicable for real-world decision-making.
2. **Explore Advanced Techniques:**  
   By testing alternative algorithms and clustering methods, we can find more efficient and reliable solutions for managing traffic.
3. **Analyze the Full Dataset:**  
   Expanding the analysis to include the entire dataset will provide a complete understanding of traffic patterns throughout the day. This broader perspective can help with better strategic planning and decision-making.

**References:**

**Union of Concerned Scientists. (2020).** *Ride-hailing's climate risks: Steering a growing industry toward a clean transportation future.* Union of Concerned Scientists. Retrieved from https://www.ucsusa.org/resources/ride-hailing-climate-risks

**Huang, Y., et al. (2024).** *A holistic approach for equity-aware carbon reduction of ridesharing platforms.* arXiv. Retrieved from <https://arxiv.org/abs/2402.01644>

**OpenAI. (2024).** *Language model output used for report modification.* Retrieved from <https://chat.openai.com/>