Predicting and Analyzing Traffic Patterns

(COMP3125 Individual Project)

\*Note: Do not used sub-title

Kevin Son   
[DATA SCIENCE FUNDAMENTALS](https://wentworth.brightspace.com/d2l/home/48677)

# Introduction (*Heading 1*)

# Traffic congestion in urban areas represents one of the most significant challenges facing modern cities, particularly in the Greater Boston area where commuters spend an average of 96 hours per year in traffic delays [1]. This research focuses on analyzing traffic patterns and their determinants using data science techniques. According to the 2023 INRIX Global Traffic Scorecard, Boston ranks as one of the most congested cities in North America, significantly impacting daily commuter patterns [1]. While current research has primarily focused on broad traffic patterns, this study aims to bridge knowledge gaps by investigating specific factors such as peak traffic timing, weather correlations, and the impact of local events using machine learning approaches including Linear Regression and Random Forest algorithms. The findings from this research will not only help individual commuters make informed decisions about their daily commutes but also provide valuable insights for urban planning and traffic management systems, especially considering the recent shifts in work patterns due to remote work policies [2].

# Datasets

## Source of dataset

The datasets used in this analysis were obtained from reputable statistical sources:

1. Congestion data: "statistic\_id235786\_most-congested-city-centers-in-the-north-america-2023.xlsx" from Statista.
2. Cost data: "statistic\_id1305426\_cost-per-driver-of-traffic-congestions-in-the-us-by-urban-area-2019.xlsx" from Statista.
3. Hours lost data: "statistic\_id1305446\_most-congested-urban-area-in-the-us-2019.xlsx" from Statista.

These datasets were generated in 2023 and 2019 respectively, using data collected and analyzed by the Statista research organization, a reputable provider of market and consumer data.

## Character of the datasets

The datasets are provided in Excel (.xlsx) format and contain the following key information:

|  |  |  |
| --- | --- | --- |
| Dataset | Parameters | Units |
| Congestion data | City, Congestion\_Level | Percentage |
| Cost data | City, Cost\_Per\_Driver | US Dollars |
| Hours Lost data | City, Hours\_Lost | Hours |

The data was cleaned by removing any rows with missing values for the key parameters. Additionally, the city names were standardized by removing any parenthetical information (e.g., "Los Angeles (CA)").

To combine the datasets, a merge() function was used to join the datasets on the "City" column, creating a comprehensive dataset with all the relevant metrics for each city.

Furthermore, the analysis included creating new derived features such as "Cost Per Hour", "Congestion Impact", and "Cost Efficiency" to enable deeper insights into the relationships between the variables.

# Methodology

## Linear Regression Model

The primary method used in this analysis is a Linear Regression model to predict the hours lost in traffic based on congestion levels and cost per driver.

1. Data Preparation The relevant datasets were loaded and merged, with city names cleaned to ensure consistency. Feature engineering was then performed to create additional metrics, such as cost per hour, congestion impact, and cost efficiency.
2. Model Training and Evaluation The data was split into training and testing sets, and the features were scaled using StandardScaler from scikit-learn. The Linear Regression model was trained using the fit() method, and its performance was evaluated using the R-squared score and root mean squared error (RMSE). Feature importance was also analyzed to understand the relative contribution of each variable.

## Correlation Analysis

In addition to the predictive modeling, a correlation analysis was performed to investigate the relationships between the traffic metrics. A heatmap was generated to visualize the correlation coefficients between congestion level, cost per driver, and hours lost.

Identify applicable funding agency here. If none, delete this text box.

## Efficiency and Impact Ranking

To address the research questions regarding cost-efficient and high-impact cities, the dataset was further analyzed to rank the cities based on the derived features. Specifically, the cities were ranked by their cost efficiency (cost per driver divided by congestion level) and congestion impact (congestion level multiplied by hours lost).

# Results

Example: After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

## Model Performance

The Linear Regression model was trained to predict the "Hours Lost" based on the "Congestion Level", "Cost Per Driver", and "Cost Per Hour" features. The model's performance is summarized as follows:

Root Mean Squared Error (RMSE): 3.3959

The RMSE of 3.3959 suggests that, on average, the model's predictions are off by about 3.4 hours compared to the actual values.

Feature Importance:

* Congestion Level: 0.0115
* Cost Per Driver: 24.4815
* Cost Per Hour: -0.3959

The feature importance results show that the "Cost Per Driver" has the highest positive impact on the "Hours Lost", while the "Cost Per Hour" has a negative impact. The "Congestion Level" has a relatively small but positive influence on the target variable.

## Correlation Analysis

The heatmap shows a strong positive correlation (1.00) between "Cost Per Driver" and "Hours Lost", as well as between "Congestion Level" and "Hours Lost". There is a moderate negative correlation (-0.39) between "Cost Per Driver" and "Congestion Level".

These correlation patterns suggest that as congestion levels and cost per driver increase, the hours lost in traffic also tend to rise. Additionally, cities with higher congestion levels tend to have lower cost per driver, indicating a trade-off between these two metrics.

## Efficiency and Impact Ranking

Based on the features created through data engineering, the cities were ranked by their cost efficiency and congestion impact.

Most Cost-Efficient Cities:

1. Los Angeles – Cost Efficiency: 28.754717
2. San Francisco – Cost Efficiency: 37.789474
3. Boston – Cost Efficiency: 54.878049

Highest Impact Cities:

1. Boston – Congestion Impact: 61.09
2. Chicago – Congestion Impact: 55.10
3. Los Angeles – Congestion Impact: 54.59

The results show that Los Angeles is the most cost-efficient city, with the lowest cost per driver relative to its congestion level. In contrast, Boston has the highest congestion impact, with the highest combined congestion level and hours lost.

These findings suggest that some cities, like Los Angeles, is able to manage congestion more effectively, while others, like Boston, experience disproportionately high costs and hours lost due to traffic.

# Discussion

While the analysis provided valuable insights, there are a few areas for improvement. The inability to calculate a reliable R-squared score is a limitation, as this metric is crucial for evaluating the model's predictive performance. Investigating the root cause and exploring alternative modeling techniques could enhance the results.

Additionally, incorporating more diverse data sources and features may better capture the complex factors influencing traffic congestion and costs. Conducting case studies on top-performing and underperforming cities could also provide valuable insights into effective strategies.

# Conclusion

The key findings from this analysis include:

* The Linear Regression model was able to provide insights into the relationship between traffic congestion, cost, and hours lost, though the R-squared score could not be reliably calculated.
* Correlation analysis revealed strong positive correlations between cost per driver, congestion level, and hours lost, as well as a moderate negative correlation between cost per driver and congestion level.
* Los Angeles was identified as the most cost-efficient city, while Boston had the highest congestion impact.

These insights can inform policymakers and urban planners in developing strategies to address traffic congestion and its economic burden. By understanding the drivers of high costs and lost productivity, cities can work to optimize their transportation systems and infrastructure to improve efficiency and reduce the impact on commuters and businesses.

##### References

1. [1] INRIX. (2023). "INRIX 2023 Global Traffic Scorecard." Retrieved from <https://inrix.com/scorecard/>
2. [2] Bick, A., Blandin, A., & Mertens, K. (2023). "Work from Home after the COVID-19 Outbreak." Federal Reserve Bank of Dallas, Working Paper No. 2023-3.