DS8007 ADVANCED DATA VISUALIZATION

PROJECT REPORT

Analyzing the Impact of Delays on Toronto's TTC Buses Network

Group Members

Divyansh Agrawal - 501288889

Ahnaf Shahriyar Chowdhury - 501314107

Table of Contents

1 Introduction

Public transportation reliability is a necessary factor in the daily lives of people. In Toronto, the TTC bus network has thousands of customers daily but faces delays due to many reasons, including traffic, operational issues, emergencies, weather, etc. Because of these, a huge gap is created while providing services. As a result, passengers get frustrated, making the transit system less organized.

The main goals of this project include carrying out exploratory data analysis on TTC bus delay data, creating useful visualizations, and obtaining insights that can inform bus schedules for optimization as well as delay minimization. The scope of this project involves completing data preprocessing operations such as missing value handling, outlier removal of outliers, and data transformation. The project, following data preprocessing, will endeavour to offer in-depth insights on factors that lead to delays on various routes, incident types, as well as on diverse timeframes. Bar plots, heatmaps, and box plots will be utilized to find patterns in addition to trends in data. The analysis will concentrate on the top 10 most delayed routes, analyzing variables such as incident types, operational inefficiencies, and external factors influencing service delivery. This project will provide valuable insights that can help improve the performance of the TTC bus system to make operations smoother, minimize passenger wait times, and improve the overall public transit experience.

2 Data Description

Data utilized in this project is the 2024 TTC Bus Delay Data, which offers complete insights on delays on different bus routes in the Toronto Transit Commission system. The data records delays on buses in addition to factors such as the number, type of incident, delay duration, as well as the hour at which delays were experienced. The data offers a complete record that shows on which routes delays are most common and why.

Dataset Format:

The dataset is provided in Excel (.xlsx) format, which consists of the following key columns:

- **Date:** The date of the recorded delay.
- **Route:** The bus route number where the delay occurred.
- **Time:** The time of the delay incident.
- **Day:** The day of the week the delay occurred.
- **Location:** The specific location or area where the incident or delay occurred.
- **Incident:** The type of incident responsible for the delay, such as General Delay, Traffic Diversions, Emergency Services, etc.
- **Min Delay:** The total time, in minutes, that the bus was delayed due to the incident.

- **Min Gap:** The time gap between buses, which can be affected by delays.
- **Direction:** The direction in which the bus was traveling (e.g., N for North, S for South).
- **Vehicle:** The vehicle number of the bus involved in the incident.

Dataset Source:

The dataset being used is `ttc-bus-delay-data-2024` and can be found on the open Toronto site. The link is as follows:

`https://open.toronto.ca/dataset/ttc-bus-delay-data/`

Dataset Preprocessing Steps Overview:

Handling Missing Values and Duplicate Rows:

- The Columns `Route` and `Direction` were filled with ”Unknown” to cover missing data.
- Duplicate rows were handled by deleting them.

Outlier Detection and Removal:

- Outliers in `Min Delay` and `Min Gap` were found by employing the Interquartile Range (IQR) approach.
- Outliers that exceeded these limits were excluded from the dataset to prevent extreme values from affecting the analysis.

Feature Engineering:

- An additional column for `"Hour"` was added by using the `Time` column, making it possible to analyze by varying times throughout the day.
- Location data was utilized to geocode addresses to map them using latitude and longitude.

Data Transformation:

- The `Date` column was also converted to datetime in order to support analysis based on time, enabling weekly or monthly aggregations.
- A `Week` column was derived from the `Date` column to organize the data by week to perform temporal analysis.

Sorting and Ranking:

- Routes were categorized by total delay to concentrate on the most delayed routes.
- Incident types were grouped by total delays to highlight the most common delay-causing incidents.

3 Detailed Data Preprocessing

Data Preprocessing for Insight 1:

Aggregating the Data by Route:

We group the dataset by the `Route` column and calculate the total (`Min Delay`) for each route. This will give us a clear view of the total delays for each route.

Reason: Aggregating the data allows us to see the total delay in each route. This aggregation is necessary for identifying the routes that are critical in terms of delays. We can prioritize routes with the highest delays by summing up the `Min Delay`.

Sorting Routes by Total Delay:

Once the total delays have been aggregated, we need to sort the routes based on the total delay. This allows us to target the routes that are causing the most delays.

Reason: Sorting the routes by total delay helps us focus on the top 10 routes that are contributing the most to delays. These are the routes that need special attention, such as schedule optimization, addition of more buses, or improving traffic management on these routes.

Data Preprocessing for Insight 2:

Aggregating Data by Route:

We group the data by `Route` and aggregate the `Min Delay` for each route, which sums the delays for each route.

Reason: Aggregating delays by `Route` helps us understand how much delay each route is experiencing in total, which is critical for identifying the routes that need improvement.

Sorting Routes by Total Delay:

After we combine all the delays for every route, we rank all the data in descending order to get the most delayed routes to appear at the top. Subsequently, we choose the top 10 most delayed routes.

Reason: Sorting by total delay enables us to quickly see which routes are causing delays. This is crucial in selecting routes that need operational intervention. We select the top 10 routes because they have the most impact and should be fixed quickly to minimize delays.

Reshaping the Data:

In order to prepare data for creating the heatmap, we have to reshape it. The heatmap requires data in matrix form, thus we reshape it.

Reason: The heatmap function requires that data is in 2D form such that all routes become columns, and their respective delay values appear in the matrix. We make sure that data is in the right form for the heatmap by making `Route` as index and transposing (T).

Data Preprocessing for Insight 3:

Grouping by Incident Type:

We segment the dataset according to the `Incident` column to obtain the overall delays resulting from every type of incident. This enables us to see the delay distribution by incident type.

Reason: By segmenting by incident type, we can measure to what extent every incident type adds to the total delay. This process assists in determining which incident types to address from an operations standpoint so that we can target areas causing most delays.

Sorting Incident Types by Total Delay:

After we compute each incident type's total delay, we rank the incident types in descending order based on their total delay to determine which incident types contribute most to delays.

Reason: Sorting in descending order enables us to concentrate on those incident types that have an impact in terms of causing delays. This enables us to rank incident types in terms of priorities for investigation or operational adjustments to minimize delays.

Data Preprocessing for Insight 4:

Grouping by Route and Incident Type:

We group by `Route` and `Incident` to accumulate the total `Min Delay` for every combination of route and incident type.

Reason: By grouping the data by `Route` and `Incident Type`, we can observe how individual incident types contribute to the overall delay for each route.

Sorting the Routes by Total Delay:

To concentrate on the top 10 routes, we first calculate the overall delay for every route and rank the routes in descending order according to their overall delay.

Reason: Sorting routes by total delay allows us to recognize which routes were most delayed, which in turn we are most interested in analyzing. We narrow our focus to the top 10 routes for making visualization more practical and useful.

Data Preprocessing for Insight 5:

Calculating IQR for Min Delay and Min Gap:

Prior to removing outliers, we have to obtain the Interquartile Range (IQR) for `Min Delay` and `Min Gap`. IQR is a measure used to gauge the data's spread as well as to spot outliers.

Reason: The IQR is utilized to determine where most data points lie. Through calculating Q1 (25th percentile) and Q3 (75th percentile) for `Min Delay` as well as `Min Gap`, we have an accurate idea about data spread and can set outlier thresholds.

Defining Lower and Upper Bounds for Outlier Removal:

We use IQR to set the minimum and maximum bounds for both `Min Delay` and `Min Gap`. Any data points that lie outside these boundaries will qualify as outliers and thus will be excluded.

Reason: The IQR rule is used in calculating fences for outlier identification. Any data points that lie outside the lower and higher bounds are labeled as outliers and excluded from the dataset to make the analysis representative of the usual system behavior, not influenced by extreme values that might produce misleading results.

Removing the Outliers:

Now we remove rows from the dataset for which `Min Delay` or `Min Gap` is outside the specified limits.

Reason: By eliminating the outliers, we have ensured that the data that remains captures typical patterns in bus delays and gaps, which is essential for reliable analysis. Outliers have the ability to distort relationships between variables, thus by removing them we increase the dependability of our results.

Data Preprocessing for Insight 6:

Filtering Top Locations:

Filter the data to retain the 30 most frequent values in the `Location` column. The selection is done based on value frequency.

Reason: Geocoding text addresses is a time-consuming task and is typically limited by API rate limits. Transforming the top 30 locations will enable efficient processing while retaining the high-impact data points.

Aggregating Delay and Incident Data:

For each selected location, it aggregates the information using the `Location` column. It finds the average delay (`Min Delay`) and the most frequent incident type by utilizing a counter function.

Reason: This aggregation allows us to understand both the severity (average delay) and the reason (leading incident type) at each location. It makes it easier to figure out which locations are worst and why.

Geocoding Location Names:

Every address is converted to its corresponding geographical coordinates (latitude and longitude) through the Nominatim service of the geopy library with a rate limit applied.

Reason: Geocoding is required to transform text locations to coordinate format suitable for spatial mapping.

Dropping Geocoding Failures:

Addresses that are unable to yield valid coordinates are removed from the dataset.

Reason: Invalid or missing coordinates would result in errors in plotting or meaningless points on the map and are hence excluded to maintain the sanctity of the visualization.

Data Preprocessing for Insight 7:

Converting the Date Column:

The `Date` column is changed to datetime type from string so as to support time-based analysis.

Reason: Datetime conversion is needed to allow time grouping and resampling. It ensures the `Date` column is handled as part of a time series.

Adding a Weekly Time Period Column:

A new column `Week` is added by converting each `Date` to its corresponding week start date through a time period function.

Reason: This conversion helps in aggregating data by periods of weeks and thus easier identification of trends over time and smoother and more intelligible visualization.

Categorizing by Route and by Weekly Delay Count:

The data is categorized as `Week` and `Route` for getting per-week delay incidence count for a route.

Reason: Summing up counts of delay provides an organized data set wherein delay patterns can be charted as a function of time by route. It helps one view which routes are consistently problematic.

Top Routes Filtering:

Out of all the routes, top 10 routes with highest total delays are selected for visualization.

Reason: Filtering prevents the animation from getting cluttered and jumbled. It highlights the delay-susceptible and pertinent paths most, increasing clarity and effectiveness in visualization.

Pivoting Data for Animation:

The data aggregated is pivoted into wide format so that every path is a column, and week rows are shown. Zero is utilized for filling missing values. Cumulative delays' counts are added across weeks.

Reason: This is required for frame-by-frame animation. The broad format with running totals allows the chart to show how delays accumulate and how rankings of routes change over time.

Data Preprocessing for Insight 8:

Extracting the Hour from Time:

Time column is converted to datetime and hour is extracted to create a new column called `Hour`.

Reason: Extracting the hour enables analysis by time of day. This is necessary in order to determine if delays or gaps in service are more common during peak or off-peak hours.

Convert Direction to Categorical Type:

The `Direction` column is changed to a categorical variable to utilize it as an effective color differentiator in the scatter matrix.

Reason: Categorical encoding is necessary for grouping by color in visualizations. This will enable us to observe if direction has any relationship with delay or gap patterns.

Dropping Missing Values and Sampling:

Rows with missing values in columns `Min Delay`, `Min Gap`, `Hour`, or `Direction` are dropped. A random sample of 1000 rows is selected for plotting.

Reason: Removing null values ensures data integrity for plotting. Sampling improves performance and readability by avoiding cluttered plots.

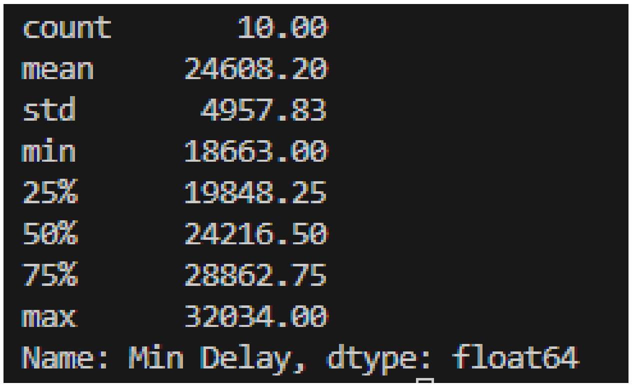
4 Exploratory Data Analysis (EDA)

INSIGHT 1:

Problem Statement:

The goal is to identify the routes that experience the most delays so that operational improvements can be implemented, such as adjusting schedules, adding more buses, or improving traffic management. This is important because it can reduce delays, eventually improving passenger satisfaction and overall system efficiency, and also ensuring that buses run on time and reducing passenger waiting times.

Data Statistics and Distribution:



```
count       10.00
mean     24608.20
std       4957.83
min      18663.00
25%      19848.25
50%      24216.50
75%      28862.75
max      32034.00
Name: Min Delay, dtype: float64
```

Figure 1: Data Statistics for TTC Bus Delays

Figure 2 illustrates the distribution of delay totals over routes, with an extremely heavy clumping of data points at zero. This implies that most routes have very short delays, but some have substantially higher delays. This high degree of skewness means that efforts to improve operations should concentrate on reducing the small number of high-delays routes while keeping most routes in line.

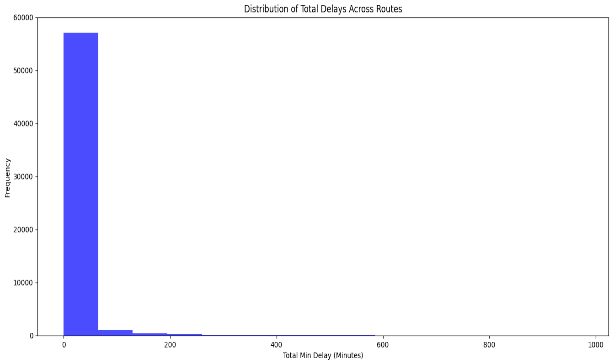


Figure 2: Data Distribution of Delays

Visualization:

A bar plot is best suited for representing total delay by route since it enables easy comparison among routes. The bar plot effectively highlights the most delayed routes, enabling us to know which routes have to be attended to with utmost urgency.

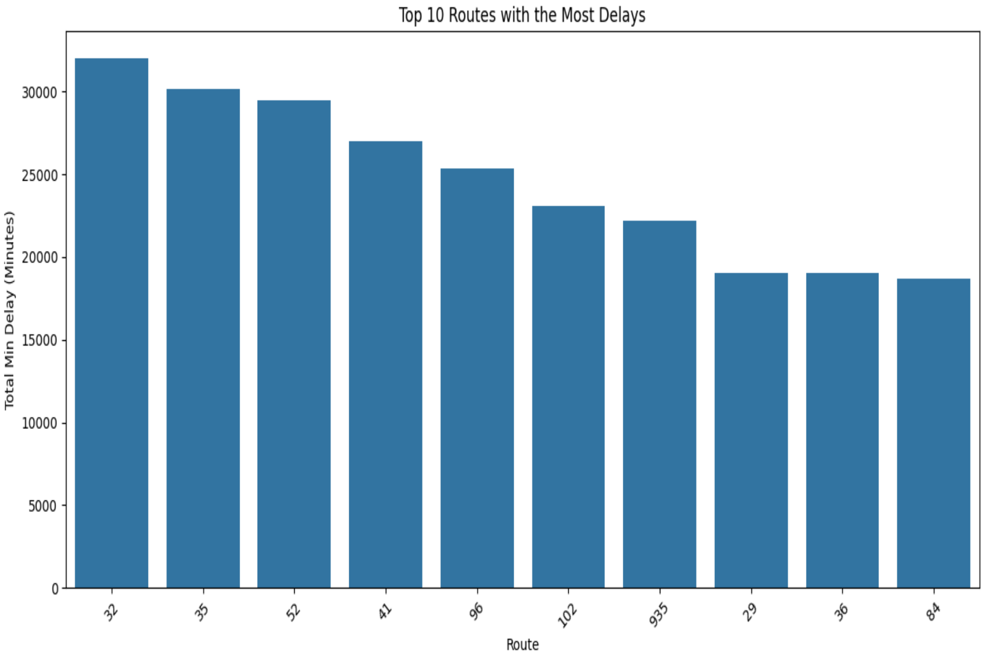


Figure 3: Total Delay by Route (Bar Plot)

Analysis from the Insight:

From the bar plot of the top 10 routes with the most delays, we can draw several insights:

- Route 32 stands out as the highest in terms of total delay. This indicates that operational changes are most needed on this route to reduce delays.
- Other top routes, like Route 35, Route 52, and Route 41, also experience significant delays, highlighting areas that need targeted improvements.
- This visualization enables us to quickly contrast delays across routes and determine the most troubled routes that need to be targeted for upgrades.

Top delaying routes, such as Route 32, Route 35, and Route 52 are primarily responsible for delays and ought to receive priority for operational enhancement.

INSIGHT 2:

Problem Statement:

The goal of this analysis is to determine which routes have the most delays. By seeing the total delay on every route on one plot, we can immediately spot routes that should have maximum priority in terms of enhancing scheduling or operational efficiency. Bus delays pose a serious problem for public transportation, affecting passenger experience as well as system efficiency. Delay reduction can contribute to enhancing overall service quality.

Visualization:

The heatmap displays a color-coded matrix in which color intensity is used to indicate the cumulative delay. The color range YlOrRd (yellow to red) is used to represent delay seriousness, where more serious delays will have more intense colors, allowing for easy visual verification of the most affected routes. The annot=True parameter superimposes delay values on the heatmap for increased readability.

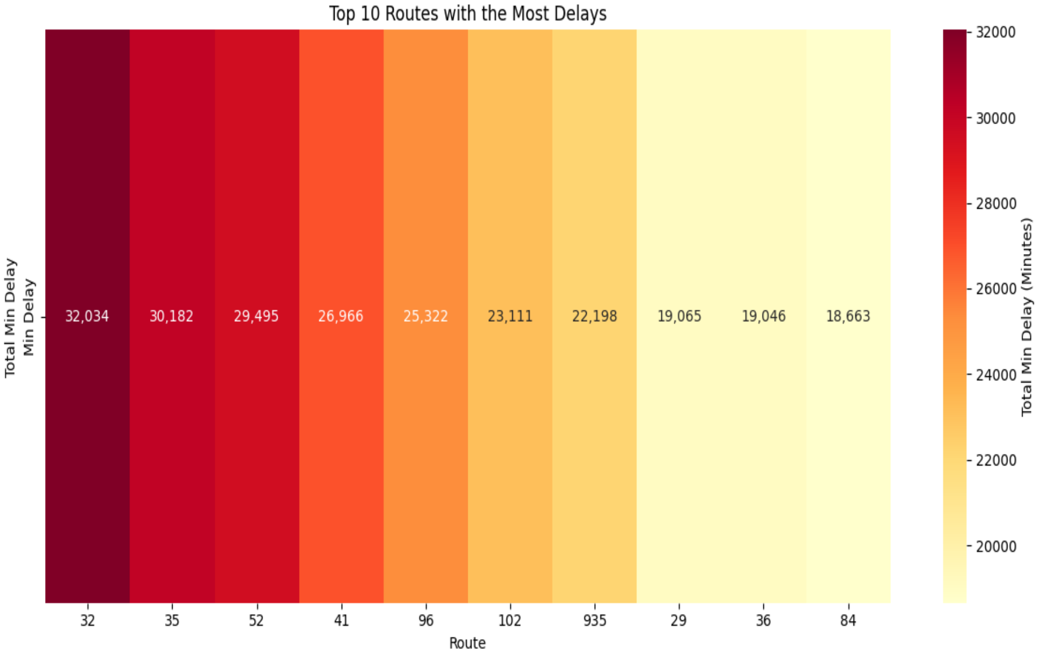


Figure 4: Total Delay by Route (Heatmap)

Analysis from the Insight:

The heatmap provides an immediate and intuitive visualization of the total delays by route. Each route is represented as a column, with the delay values shown inside the color-coded matrix.

- Route 32 has the highest delay (32,034 minutes), indicating that it is the most problematic route.

- Route 35 and Route 52 also have high delays, suggesting that these routes should be prioritized for operational improvements.

- The color intensity represents the severity of the delays, with darker shades indicating higher delays.

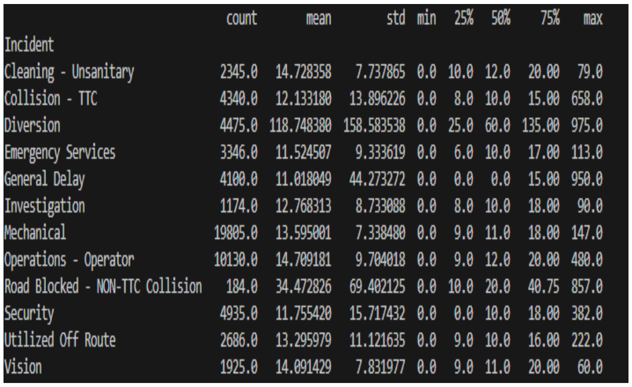
Routes 32, 35, and 52 have the largest delays, most likely because of traffic congestion, service disruptions, or insufficient manpower. The heatmap strongly reflects these routes so that operational teams can optimize timetables or add more resources to decrease delays. Route 32 is especially high priority, and it is necessary to analyze whether traffic management or bus frequency is needed to break delays. Route 32 has the greatest total delay, after Route 35 and Route 52, all of which need urgent analysis. The next step is to study why delays occur on these routes and to apply measures such as optimizing schedules, adding more vehicles, or optimizing traffic management in rush hours to decrease delays and increase overall service efficiency.

INSIGHT 3:

Problem Statement:

The goal of this analysis is to determine how individual incident types contribute to overall delays on bus routes. This analysis is important since it offers practical insight for operational performance enhancement for those incident types most responsible for delays. This is crucial for knowing which types of incidents most often delay service, enabling targeted operational adjustments. That is, if most delays result from Diversions or Mechanical Issues, traffic management can be streamlined, or bus maintenance can improve to minimize delays on such routes.

Data Statistics and Distribution:



```
                                 count        mean         std  min   25%   50%     75%    max
Incident
Cleaning - Unsanitary           2345.0   14.728358    7.737865  0.0  10.0  12.0   20.00   79.0
Collision - TTC                 4340.0   12.133180   13.896226  0.0   8.0  10.0   15.00  658.0
Diversion                       4475.0  118.748380  158.583538  0.0  25.0  60.0  135.00  975.0
Emergency Services              3346.0   11.524507    9.333619  0.0   6.0  10.0   17.00  113.0
General Delay                   4100.0   11.018049   44.273272  0.0   0.0   0.0   15.00  950.0
Investigation                   1174.0   12.768313    8.733088  0.0   8.0  10.0   18.00   90.0
Mechanical                     19805.0   13.595001    7.338480  0.0   9.0  11.0   18.00  147.0
Operations - Operator          10130.0   14.709181    9.704018  0.0   9.0  12.0   20.00  480.0
Road Blocked - NON-TTC Collision 184.0   34.472826   69.402125  0.0  10.0  20.0   40.75  857.0
Security                        4935.0   11.755420   15.717432  0.0   0.0  10.0   18.00  382.0
Utilized Off Route              2686.0   13.295979   11.121635  0.0   9.0  10.0   16.00  222.0
Vision                          1925.0   14.091429    7.831977  0.0   9.0  11.0   20.00   60.0
```

Figure 5: Data Statistics for Incident Types Delays

Figure 6 shows that delays by incident type have a very skewed distribution, with all but a very few delays clumped at one end. The shape suggests that although most incidents have relatively short delays, certain incident types tend to have very significant delays.

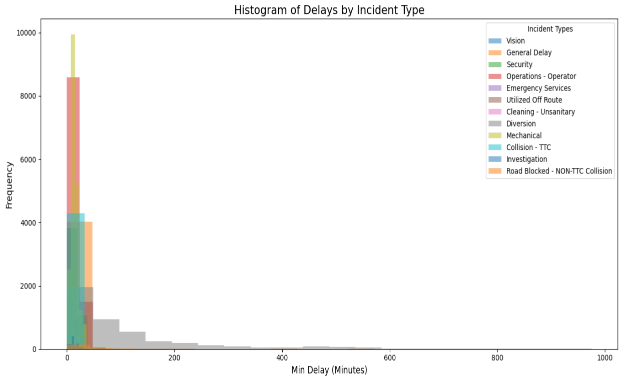


Figure 6: Data Distribution of Delays by Incident Type

Visualization:

A bar chart is suitable for illustrating the total delays resulting from all incident types because it is possible to visually compare the relative size of delays. This makes it easy for stakeholders to quickly recognize which incident types require attention.

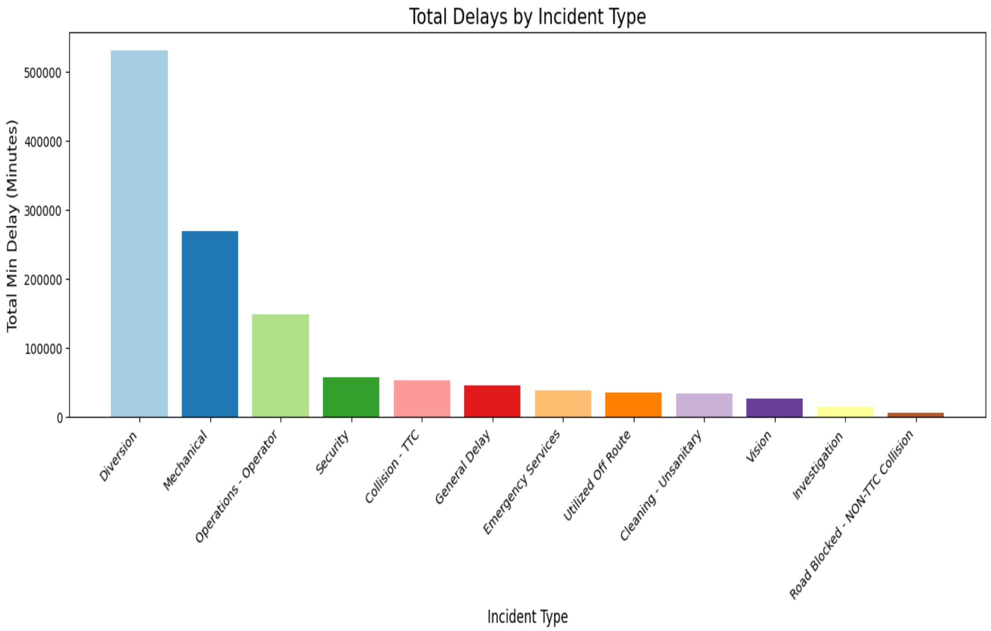


Figure 7: Total Delay by Incident Type (Bar Chart)

Analysis from the Insight:

From the bar chart of delays caused by different incident types, we can draw several insights:

- **Diversions:** The bar chart reveals that Diversion incidents have the greatest share in delays, followed by Mechanical failures and Operations-Operator incidents.

- **Mechanical Failures:** Mechanical faults also play an important role in causing delays, but these can most likely be mitigated by enhanced maintenance measures.
- **Operations and Security:** Incident types like Operations and Security have moderate delay impacts, suggesting that efforts at enhancing operational effectiveness as well as security need to increase.

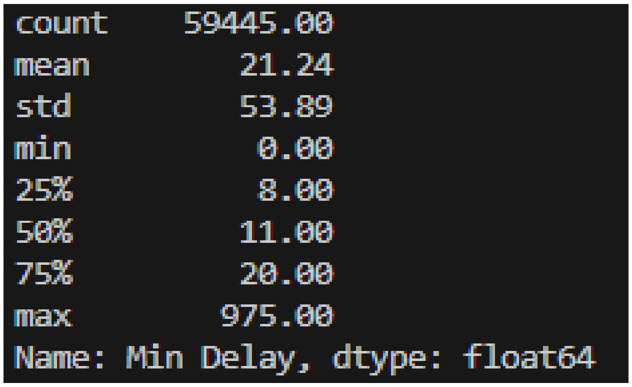
Based on the bar chart, General Delay ranks lower in terms of overall delay impact, meaning that General Delay is less impactful on delays compared to Diversions and Mechanical Issues. The operational focus should be on reducing Diversions and Mechanical Issues, followed by addressing General Delay. The analysis suggests prioritizing operational improvement efforts on Diversions and Mechanical Failures, followed by improving security incident management.

INSIGHT 4:

Problem Statement:

The aim of this analysis is to determine the contribution that individual incident types make to overall delays on the 10 most delayed routes. This is important to know in order to understand which incident types have the most impact on delays, and thereby to optimize operations on these routes. It is crucial because knowing the particular delays that most often occur enables targeted intervention. If particular types of incidents (e.g., security incidents or road diversions) account for most delays, efforts to improve in these areas can be prioritized.

Data Statistics:



```
count    59445.00
mean        21.24
std         53.89
min          0.00
25%          8.00
50%         11.00
75%         20.00
max        975.00
Name: Min Delay, dtype: float64
```

Figure 8: Data Statistics for Incident Types on Top 10 Delayed Routes

Visualization:

The stacked bar plot is suitable for representing the contribution to total delay by all incident types for every route. Every bar will represent one route, and segments in the bar will represent delays due to varying incident types.

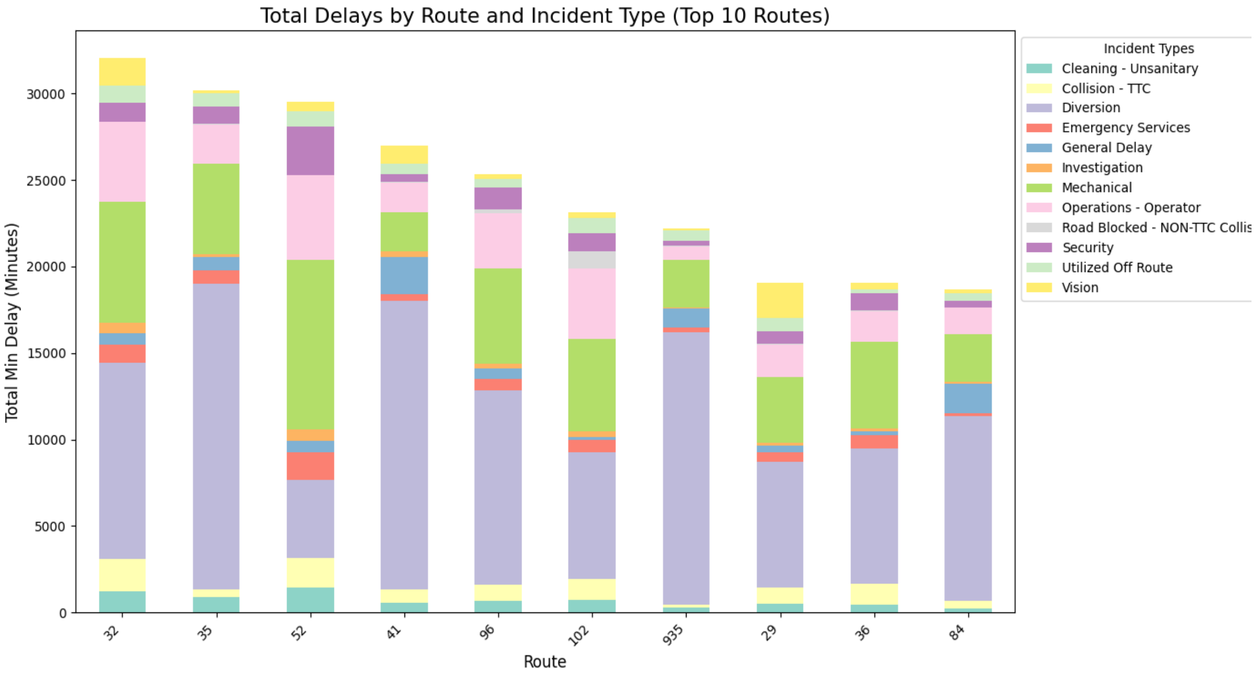


Figure 9: Total Delay by Route and Incident Type (Stacked Bar Plot)

Analysis from the Insight:

The stacked bar chart emphasizes that Mechanical breakdowns and Diversions account for most delays on the top 10 routes, followed by Operations - Operator and Security incidents. These incident types constitute the most serious challenges for the system and thus must first be addressed through operations improvement.

- **Diversions:** For Route 32 and Route 35, much of their delay is due to Diversions, indicating that enhancing traffic planning and management on alternate routes could decrease delays.
- **Mechanical Failures:** Route 52 shows notable delays due to Mechanical issues and Diversions, suggesting that more frequent maintenance and better handling of traffic diversions would help minimize delays on this route.
- **Operations and Security:** Operations and Security incidents contribute moderately to delays, suggesting that improving operational effectiveness and security measures should be a focus as well.

The data clearly illustrates that Mechanical issues and Diversions account for the most delays on the top 10 routes. These need to be some of the highest priorities to address in order to realize greater efficiency throughout the transit system. The large delays on some routes caused by Mechanical (e.g., Route 52) or Diversions (e.g., Route 32) indicate areas for intervention that need to be targeted by operational approaches.

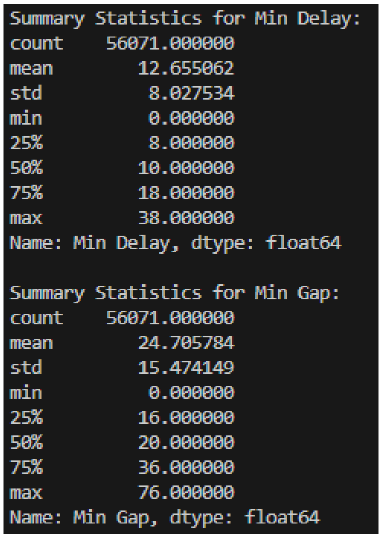
To minimize delays and maximize bus schedules, efforts ought to concentrate on solving most frequent incident types, such as Diversions, Mechanical breakdowns, and operational faults. Vehicle maintenance improvement, improved traffic flow management, and planning can potentially minimize delays at peak hours and in high-traffic areas. By resolving all these, the system can maximize operational effectiveness, minimize disruptions in service, and maximize passenger satisfaction.

INSIGHT 5:

Problem Statement:

The goal of this analysis is to examine Min Delay in relation to Min Gap. We're interested in determining if greater delays have bigger gaps between buses. This is very important if we're going to optimize bus timetables and increase efficiency in the TTC system. This is crucial because if there is a positive correlation, it means that as delays rise, service gaps between buses also rise, resulting in longer wait times for riders. Knowing this relationship can improve bus operations.

Data Statistics:



```
Summary Statistics for Min Delay:
count    56071.000000
mean        12.655062
std          8.027534
min          0.000000
25%          8.000000
50%         10.000000
75%         18.000000
max         38.000000
Name: Min Delay, dtype: float64

Summary Statistics for Min Gap:
count    56071.000000
mean        24.705784
std         15.474149
min          0.000000
25%         16.000000
50%         20.000000
75%         36.000000
max         76.000000
Name: Min Gap, dtype: float64
```

Figure 10: Data Statistics for Min Delay and Min Gap

Visualization:

The scatter plot illustrates the relationship between Min Delay and Min Gap. Every point is one bus delay with an accompanying gap. By eliminating the outliers, we can more clearly see the underlying relationship between these variables.

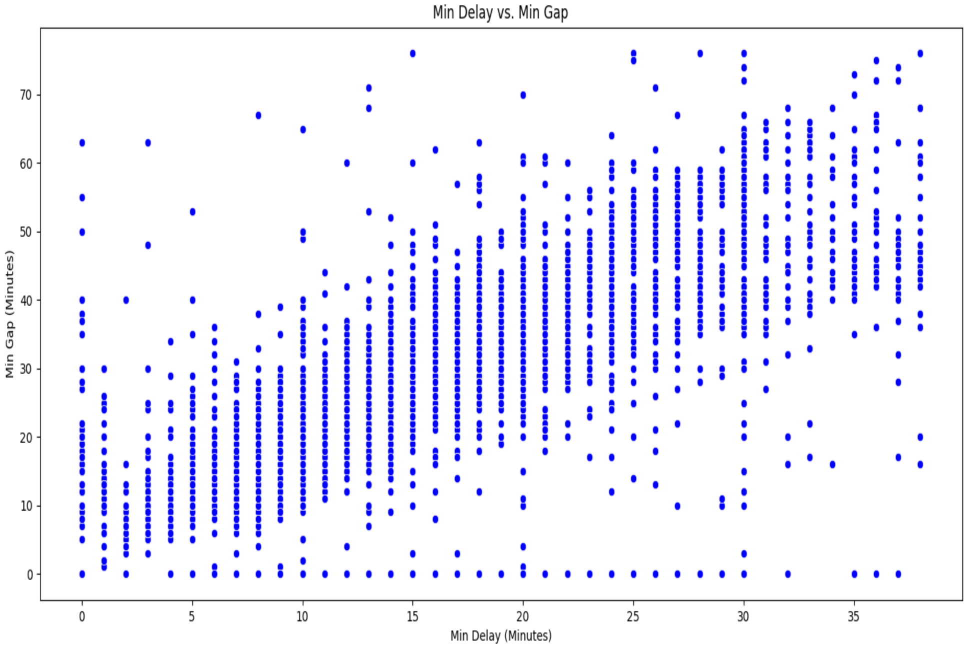


Figure 11: Min Delay vs. Min Gap (Scatter Plot)

Analysis from the Insight:

If points are distributed in a diagonal trend from bottom-left to top-right, it implies that longer delays have bigger gaps between buses. This is an important finding for optimizing bus schedules. However, if points are distributed randomly with no particular trend, it implies that delays and gaps are independent, and delay reduction will not necessarily have an impact on gaps

- The positive relationship that is evident in the scatter plot is between Min Delay and Min Gap, reflecting that higher delays have resulted in higher gaps between buses. This implies that for minimizing passenger waits, there is a need for better punctuality on buses. By reducing delays, service gaps can also be diminished, making for an efficient, passenger-supportive system.

- Eliminating the extremes offers us a more accurate insight into Min Delay and Min Gap's real relationship by removing the distortion brought about by outlier values. By targeting routes that have maximum delays, we can go about optimizing scheduling to avoid delays, causing wide gaps between buses, thus enhancing service efficiency as well as minimizing passenger wait times.

- The majority of data points lie towards the lower end of the graph, indicating that most delays and gaps are relatively small. However, some outliers show extremely large delays and gaps, which may need further investigation.

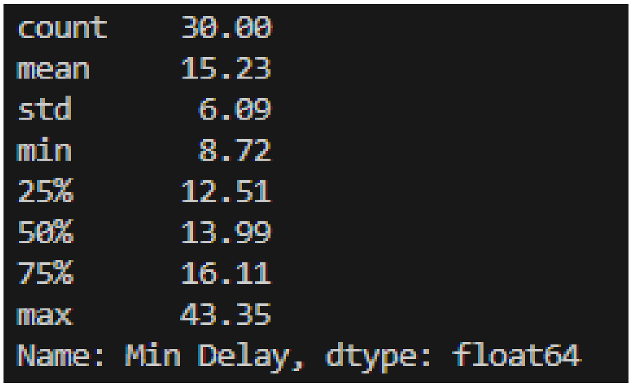
Analysis finds that Min Delay is positively correlated to Min Gap, indicating that more extended delays also produce greater service gaps. Removing outliers serves to emphasize the main relationship between these variables, making our analysis clearer. The operational implication is to decrease delays since decreasing delays directly reduces gaps between buses, resulting in efficient service and greater passenger satisfaction.

INSIGHT 6:

Problem Statement:

This analysis aims to identify the physical locations in Toronto where bus delays are most concentrated. It also shows what kinds of incidents are generating the most delay at each hotspot. The intention is to guide geographic and cause-specific plans for public transit.

Data Statistics and Distribution:



```
count    30.00
mean     15.23
std       6.09
min       8.72
25%      12.51
50%      13.99
75%      16.11
max      43.35
Name: Min Delay, dtype: float64
```

Figure 12: Data Statistics for Delay Locations

Figure 13 shows that most delays at all locations cluster in a 10 to 20-minute range, with some extending outside that range. The frequency drops off rapidly for delays in longer times, suggesting that very long delays are quite uncommon. This is a pattern that implies most delays are short, but that there exist from time to time some that greatly increase delay length.

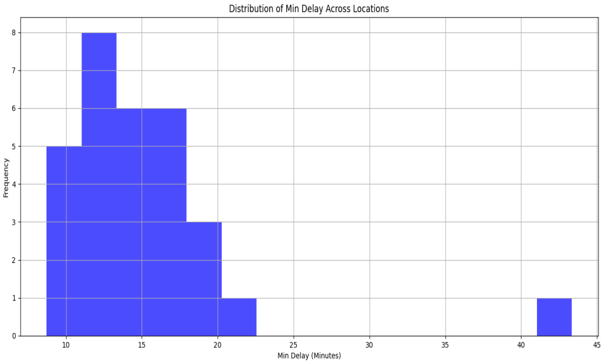


Figure 13: Data Distribution of Delays by Location

Visualization:

An interactive map is generated by making use of the folium library. Every location is plotted as a circular marker. Each bubble's size is proportional to the average delay at that location. The bubble color represents the most common incident type, with individual colors used to emphasize various categories. Popups also show location name, mean delay, and the most common delay reason.

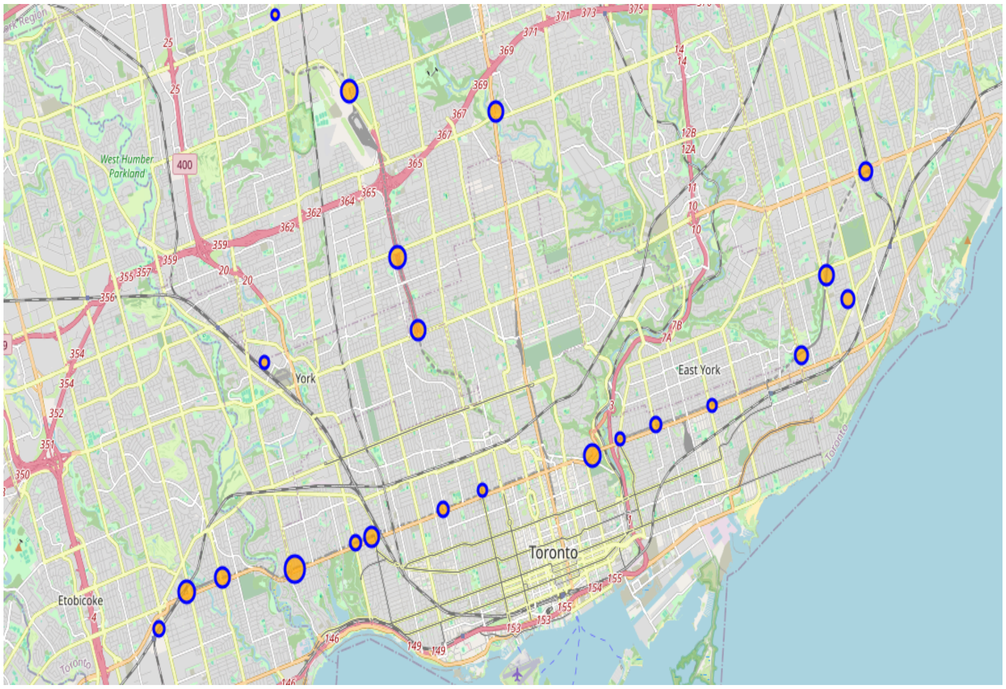


Figure 14: Geospatial Visualization of Delay Hotspots

Analysis from the Insight:

From the interactive map of delays, we can draw several insights:

- The analysis shows that delays are highly concentrated along principal transit corridors, particularly in downtown Toronto.
- Some of the intersections and transfer points both have high severity and frequency of delay.
- Larger bubbles on the map indicate locations with longer average delay durations.
- Individual hotspots are marked by the dominance of particular incident types. Some locations, for instance, experience high delays due primarily to Security-related incidents.
- This trend suggests that targeted efforts, such as enhanced traffic enforcement, signal timing modifications, or better security measures, can reduce delays at specific points.

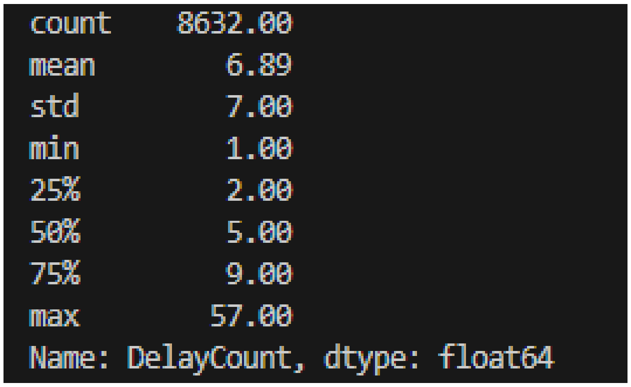
This geospatial intelligence provides an easy and insightful way of understanding where and why TTC bus delays are happening. It transforms numeric data into spatial information that is easy to interpret and act upon. This makes it highly valuable not only to data analysts but also to decision-makers and operational personnel involved in transit planning and public safety.

INSIGHT 7:

Problem Statement:

The purpose is to indicate how the most delay-susceptible TTC bus routes change over time. The goal is to determine the routes that are always highly delayed and how their ranking changes week by week.

Data Statistics:



```
count    8632.00
mean        6.89
std         7.00
min         1.00
25%         2.00
50%         5.00
75%         9.00
max        57.00
Name: DelayCount, dtype: float64
```

Figure 15: Data Statistics for Delay-Susceptible Routes Over Time

Visualization:

The animation is a horizontal bar chart, plotted with the matplotlib animation module. One frame per week. - Each bar is a bus route - The bar length is the total number of delay events - The animation moves in weeks to indicate how routes go up or down in ranking

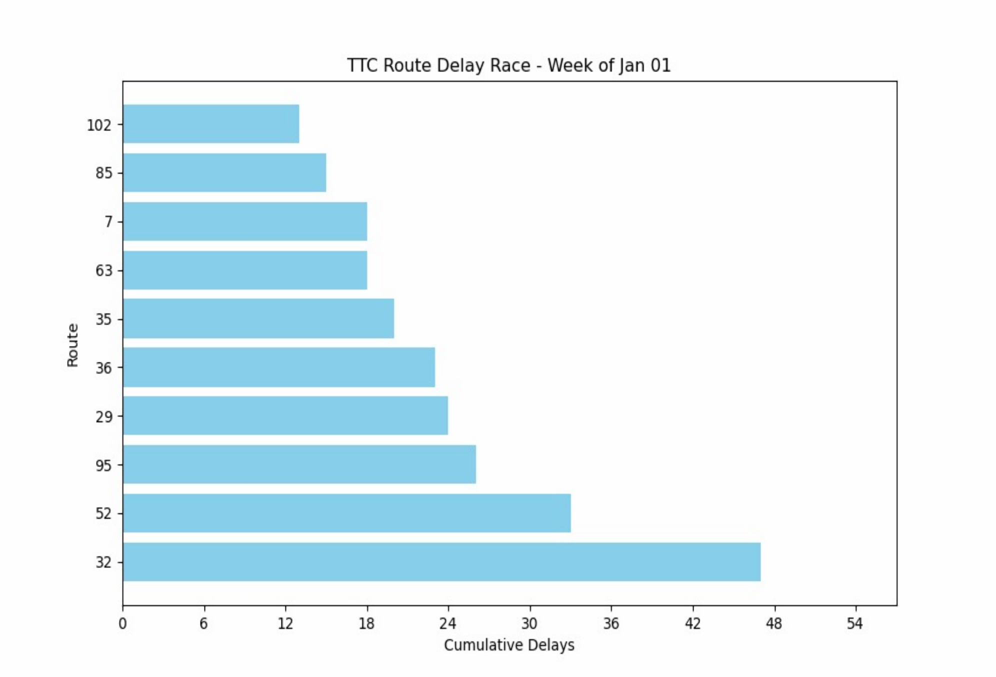


Figure 16: Animation of Delay-Susceptible Routes Over Time Snapshot

Analysis from the Insight:

From the animation of delays over time, we can draw the following insights:

- The animated visualization indicates that some routes have significant amounts of delays over many weeks, indicating systemic operation issues or persistent high demand.
- Other paths have fluctuating rankings, which may suggest transient forces such as construction, weather, or events.
- The animated film elucidates how volatile the issue of delay is and substantiates the utility of weekly checks over fixed analysis.

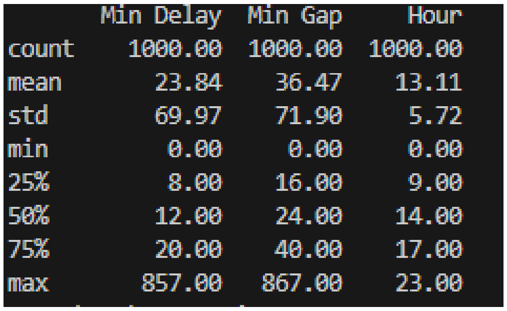
This bar chart race depicts delay patterns on a time-series by route. It is a good vehicle for communicating dynamic issues and visually engaging to technical and non-technical consumers alike. By strongly correlating what routes most contribute to unreliability of the service over time, the insight is useful towards making more effective scheduling and operation decisions.

INSIGHT 8:

Problem Statement:

The objective in this analysis is to examine the connection between delay duration, service gap, time of day, and travel direction. This information is useful in establishing any correlations or multivariate patterns that can influence scheduling and operational strategies

Data Statistics:



```
       Min Delay  Min Gap     Hour
count    1000.00  1000.00  1000.00
mean       23.84    36.47    13.11
std        69.97    71.90     5.72
min         0.00     0.00     0.00
25%         8.00    16.00     9.00
50%        12.00    24.00    14.00
75%        20.00    40.00    17.00
max       857.00   867.00    23.00
```

Figure 17: Data Statistics for Delay, Gap, Hour, and Direction

Visualization:

A scatter matrix (or pair plot) is generated with the use of the seaborn library. The variables included here include Min Delay, Min Gap, Hour, and Direction. - Off-diagonal plots provide pairwise scatter plots for all variables. - The diagonal displays histograms of all variables' distributions. - Different colors are used to depict varying directions, imagining multivariate patterns over categorical direction codes.

Analysis from the Insight:

From the scatter matrix, we can draw the following insights:

- The scatter plots indicate a high correlation between Min Delay and Min Gap, indicating that larger delays are associated with larger service gaps.
- Delay and Gap also appear to be strongly skewed with most values clustered at lower values.
- The Hour distribution has more activity later in the afternoon and early evening, which reflects rush hours for Toronto
- Direction does not show a noticeable visual trend along any of the axes, although some directions may see some clustering under specific delay or hour conditions, which can reflect route-based or directional inefficiencies.

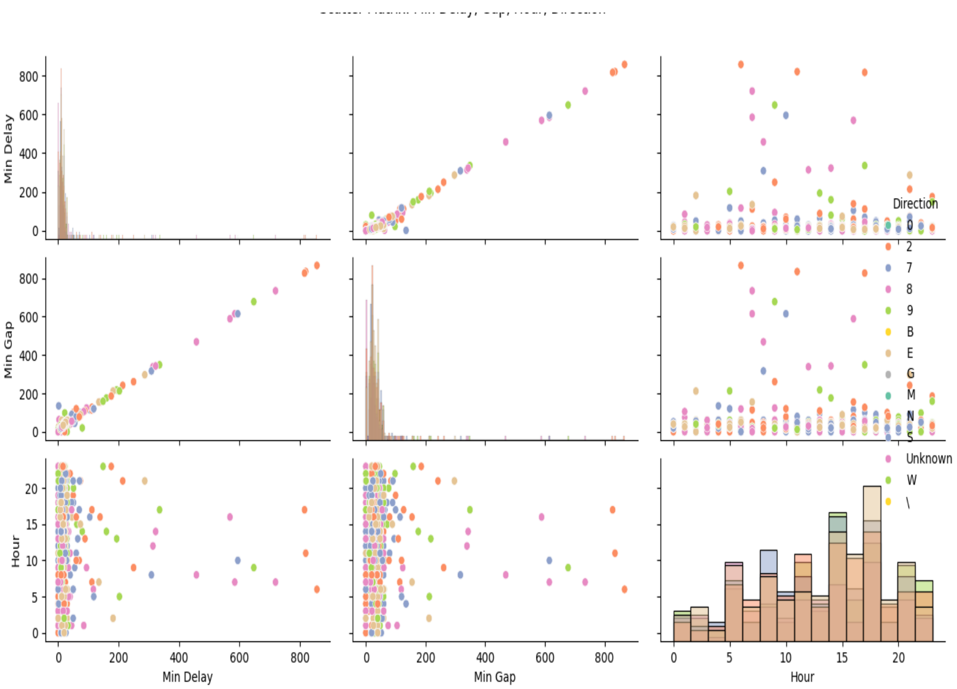


Figure 18: Scatter Matrix for Delay, Gap, Hour, and Direction

Each cell in this multivariate scatter plot summarizes one aspect regarding relations between delay, gap, time of day, and direction. This is most valuable in examining combined effects, for example, whether high delay in peak hours is only occurring in certain directions. Such results can inform more targeted operational measures, e.g., adjustment of headway or time- and direction-dependent rerouting.

5 Conclusion

Analysis of delays in the TTC buses has yielded useful insights on the most contributory factors to delays, most delay-prone routes, types of incidents, and their influence on service gaps as well as on operational efficiency. Route 32, Route 35, and Route 52 were consistently reported to be the most delay-prone, reflecting an urgent need for operational upgrades, including improved traffic management, mechanical upkeep, and addressing security-related incidents. Further analysis indicated that delays have a positive association with service gaps, reflecting that reducing delays will have an associated impact on decreasing passenger waiting times. Geospatial analysis showed that delays were most likely to cluster at high-traffic intersections, which can be corrected by targeted remedial measures such as improved security or optimized traffic flow. Temporal analysis using animated bar charts showed that certain routes have repeated delays, which necessitate long-term changes to scheduling.

In all, results from this study underscore the importance of data-based decision-making that can optimize bus schedules, operational efficiency, and passenger experience. Based on this study, future interventions should aim to study the root causes for delays, especially for the most problematic routes, and introduce real-time traffic management solutions as well as frequency adjustments at peak hours. Subsequent research could study external factors, such as special occasions or weather, influencing delays and use predictive models to manage these disruptions.