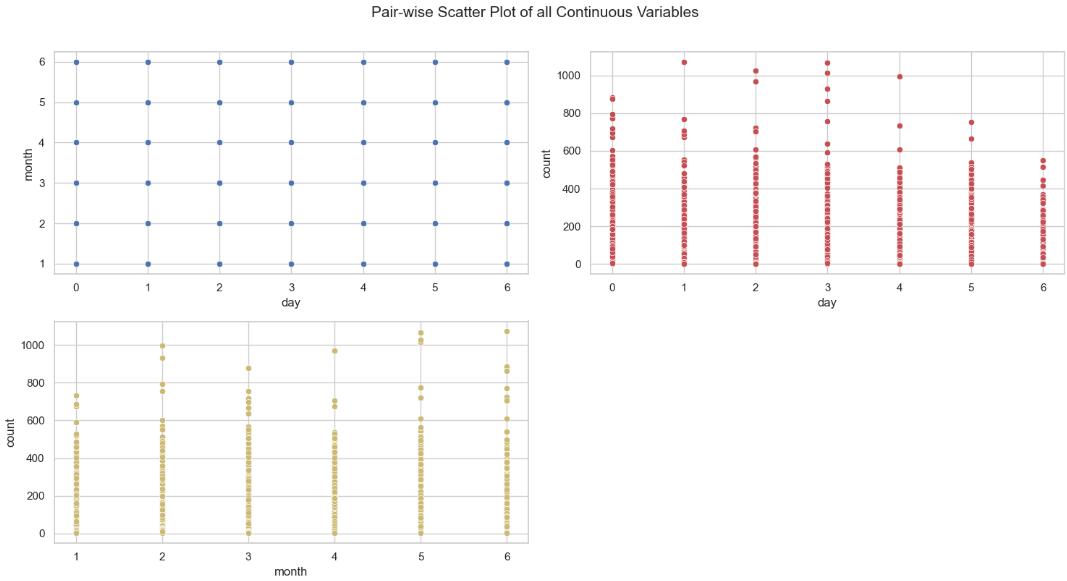
1. **Unclean Dataset :**

****

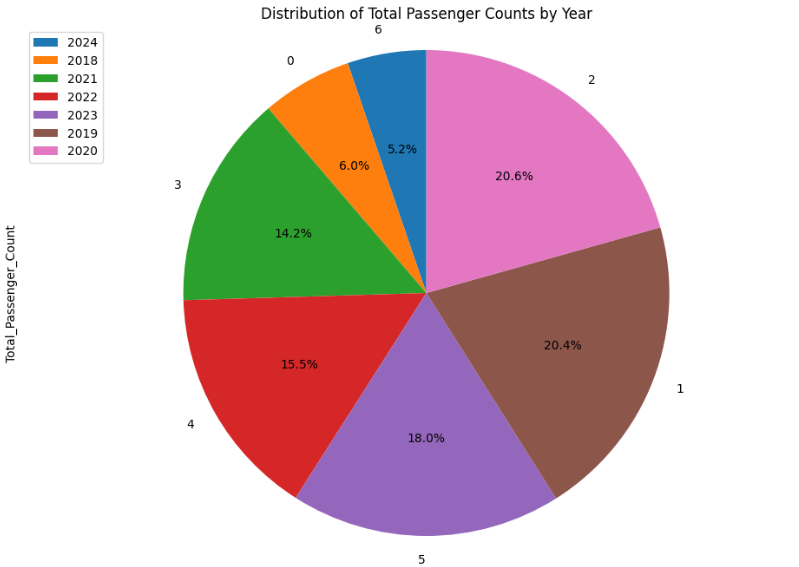
The analysis of the "thane\_ticket\_6months\_data1.csv" dataset involved performing Exploratory Data Analysis (EDA) using AutoViz, which randomly sampled 150,000 rows due to the large dataset size (2.4 million rows). The dataset consists of various columns with different data types, including integer-categorical, string-categorical, and numeric-boolean columns. Key insights from the dataset's quality assessment:

1. The "Unnamed: 0" column is likely an ID column and should be removed.
2. Columns like "d\_name," "s\_name," and "date" have high cardinality, suggesting the need for dimensionality reduction techniques (hash encoding or text embedding).
3. The "count" column has significant outliers (over 16,000 values) beyond the defined bounds, which should be capped or removed to prevent skewing the analysis.

**A screenshot of a computer

Description automatically generated**

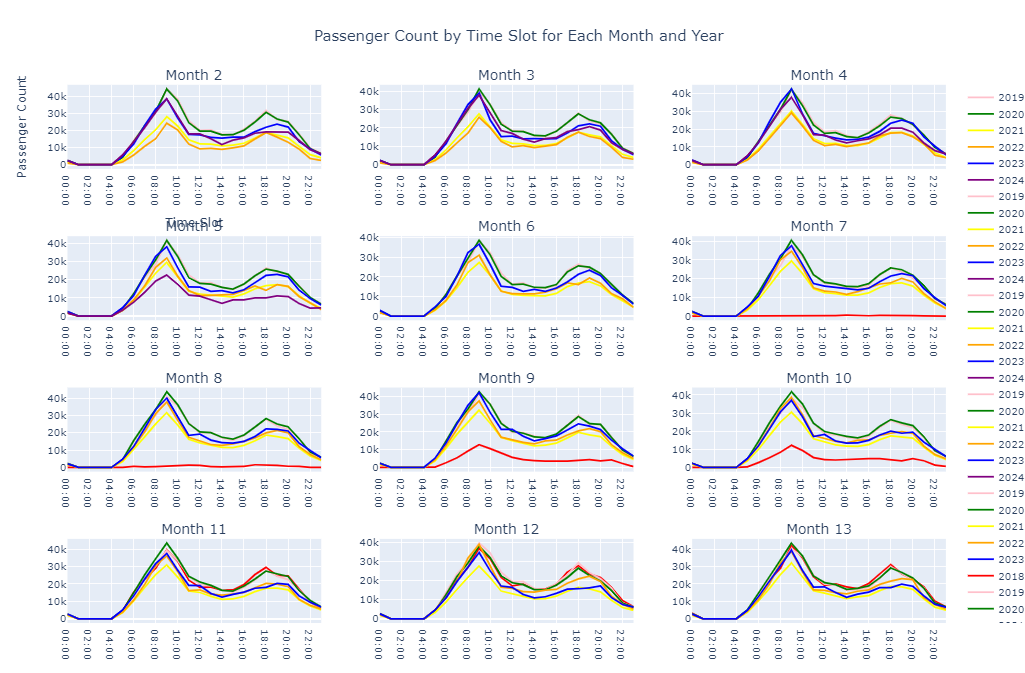
1. **Understanding the dataset**

****

The analysis calculates the total passenger count across all years and determines the percentage contribution of each year to the overall total. The yearcounts DataFrame was restructured to include the columns Year, Total\_Passenger\_Count, and Percentage, and sorted in ascending order by percentage. A pie chart was generated to visualize the distribution, with each year's contribution represented as a percentage slice. This visualization highlights the relative importance of each year in terms of passenger traffic, providing insights into trends or anomalies. The chart maintains clarity with equal axis proportions, labeled percentages, and a legend for ease of interpretation.

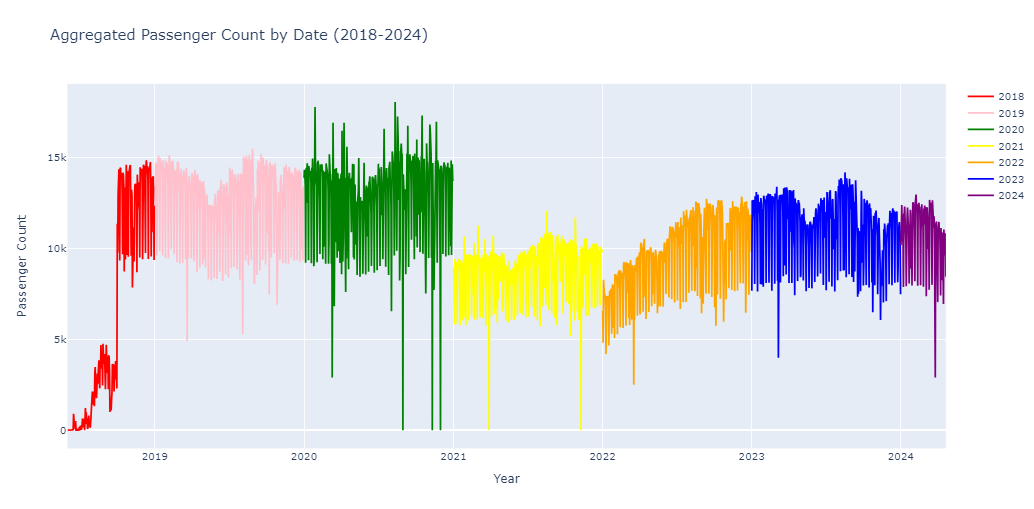
**Analysis of Passenger Count Trends Across Time Slots by Month and Year**

The graph displayed passenger count trends across different time slots for each month, categorized by year from 2018 to 2024. This visualization was plotted to identify temporal patterns and variations in passenger traffic over time, providing insights into peak travel periods and year-to-year changes. Each month was assigned a separate subplot, and passenger counts were aggregated and plotted as line charts for all available years, with distinct colors for clear differentiation. The purpose of this analysis was to understand how passenger behavior varied by time of day and season, aiding in resource planning and forecasting efforts in transportation systems.



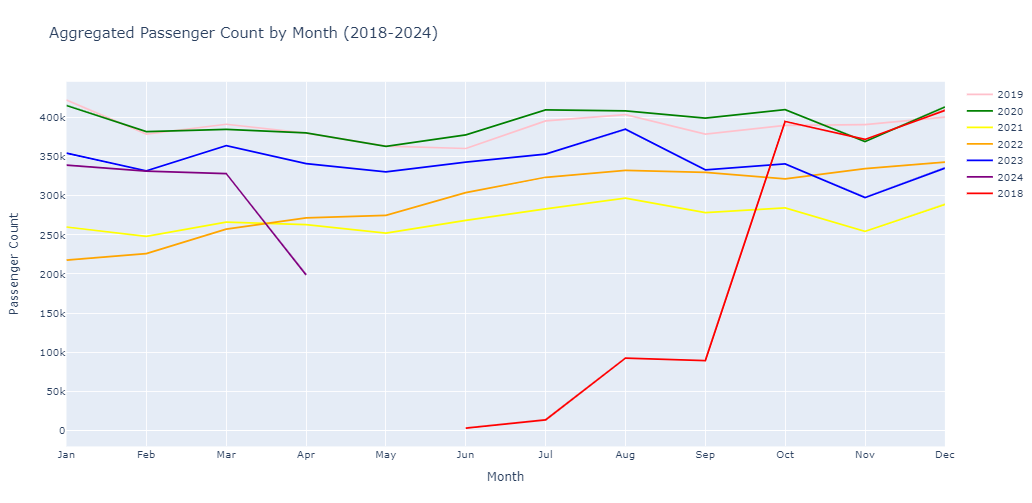
**Visualization of Aggregated Passenger Count Trends by Date (2018–2024)**

The graph illustrated aggregated passenger counts by date for each year from 2018 to 2024. This plot was created to analyze long-term trends in daily passenger volumes, providing a comprehensive view of year-to-year variations. Passenger counts were summed daily and grouped by year, with each year's data displayed as a distinct line in the chart using specific colors for differentiation. The visualization aimed to highlight seasonal peaks, annual fluctuations, and anomalies in passenger traffic, supporting the research in identifying patterns and assisting in strategic planning for transportation management.

****

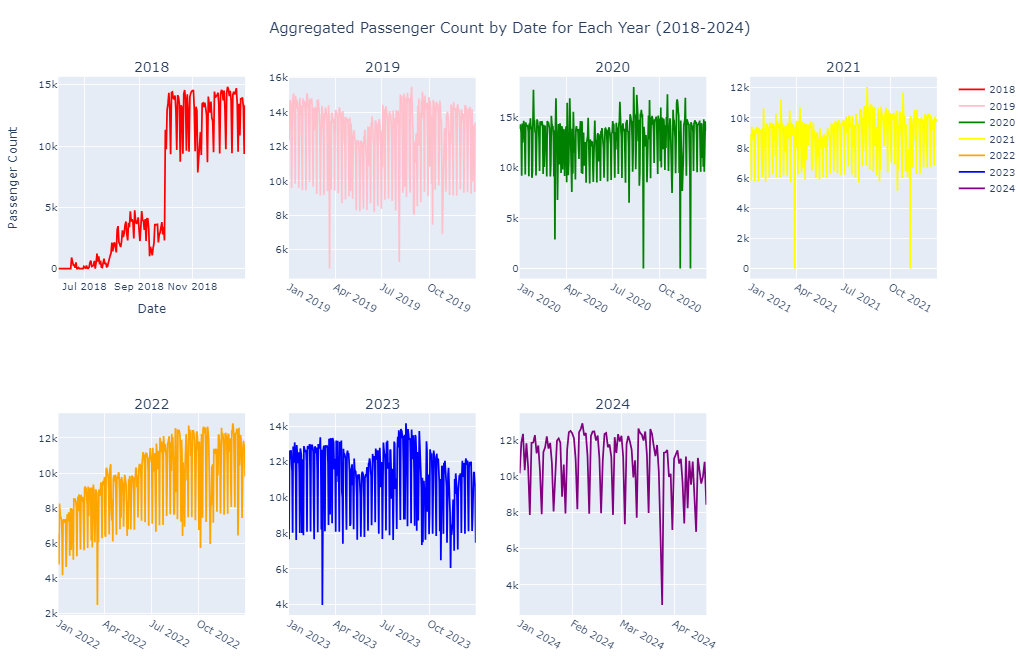
**Visualization of Aggregated Passenger Count by Month (2018–2024)**

The graph depicted aggregated passenger counts by month for each year from 2018 to 2024, highlighting monthly travel trends over several years. Passenger counts were summed across months and plotted for each year, with unique colors assigned to differentiate between years. The x-axis represented months, labeled from January to December, while the y-axis showed the corresponding passenger counts. This visualization aimed to uncover recurring monthly patterns, seasonal variations, and year-specific anomalies, providing critical insights for transportation planning and demand forecasting.



**Visualization of Aggregated Passenger Count by Date for Each Year (2018–2024)**

This graph presented the aggregated passenger counts by date for each year from 2018 to 2024, divided into individual subplots for detailed year-wise analysis. Passenger counts were summed on a daily basis, with each subplot corresponding to a specific year and uniquely colored for clarity. The x-axis in each subplot represented dates throughout the year, while the y-axis indicated the aggregated passenger counts. This arrangement facilitated a comparative analysis of daily trends, identifying seasonal patterns, yearly variations, and significant anomalies in passenger traffic over the observed period.



**1 . Data Preprocessing : DATA\_PRE\_PROCESSING\_station\_290224**

In the preprocessing phase, the dataset was carefully filtered and transformed to ensure relevance and consistency for analysis. Columns containing unnecessary or redundant information, such as Unnamed: 0, RouteNumber, Hour, SourceId, and DestinationId, were removed to streamline the data and focus on essential attributes. The dataset was filtered to include records specific to a particular station pair (SourceId = 17 and DestinationId = 11), narrowing the scope to a targeted route. Additionally, the Date\_Time column was converted into a datetime format to facilitate temporal analysis, and the data was further filtered to include only records from the years 2022 and 2023, ensuring the analysis focused on recent trends.

A dictionary was also created to map holiday types to their respective dates, categorizing specific days as public holidays, festivals, or weekends to capture the impact of these events on the data. This step allowed the integration of holiday-specific patterns into the dataset, enriching the analysis. For example, significant holidays and festivals were tagged, enabling differentiation between regular and special days. Finally, the index of the DataFrame was reset to maintain consistency after filtering, providing a clean and structured dataset for subsequent analysis and modeling. This preprocessing ensured the dataset's integrity while optimizing it for meaningful insights.

**2. Data Preprocessing and Translation : EDA - Holiday Col**

In the data preprocessing phase, the primary focus was on standardizing and translating the SourceName and DestinationName columns from Marathi to English for consistency and accessibility. Initially, the dataset contained route information with source and destination names in Marathi, alongside attributes such as RouteNumber, TripDirection, SourceId, DestinationID, issueDate, and datetime. A dictionary mapping Marathi names to their English equivalents was created to facilitate this translation. The columns SourceName and DestinationName were stripped of leading and trailing whitespaces and mapped to their respective English names using this dictionary. This transformation resulted in a clean, uniform dataset with translated names, such as "वागळे आगार" becoming "Wagle Depot" and "ठाणे स्टेशन" becoming "Thane Station(Satis)." This preprocessing step improved the dataset's usability for downstream analysis and modeling, ensuring a standardized format for station names across all records.

3. **Hourly vs Half Hourly Plots pt1**

This analysis presents the aggregated passenger counts for the entire year 2023, comparing data at two different time granularities: hourly and half-hourly. To facilitate this comparison, both the hourly and half-hourly datasets were grouped by date to calculate the total passenger count for each day. The resulting plots display the trends in passenger traffic, with the first subplot representing the hourly aggregated data (in blue) and the second subplot showing the half-hourly aggregated data (in red).

The reason for creating separate "hour" and "half-hour" columns is to explore the impact of different time granularities on the observed passenger flow patterns. By comparing these two time resolutions, we can gain deeper insights into the temporal variations in passenger traffic, which may vary depending on the level of detail used in data aggregation. This comparison helps in understanding how different time intervals might capture different aspects of passenger behavior.

A graph of a graph

Description automatically generated with medium confidence

4. **Hourly vs Half Hourly Plots pt2**

A graph of a graph

Description automatically generated with medium confidence

This plot provides a comparison of passenger counts across different time slots, using two datasets: half-hourly and hourly data.

The first subplot (on the left) represents passenger counts aggregated by half-hour time slots. The data is grouped into 48 time slots, corresponding to half-hour intervals throughout the day. Each bar in this plot shows the total passenger count for a specific time slot, with the x-axis representing the time slot number and the y-axis showing the passenger count.

The second subplot (on the right) shows the passenger counts aggregated by hourly time slots. This data is grouped into 24 time slots, representing each hour of the day. Similar to the first plot, the x-axis represents the hour of the day, and the y-axis displays the passenger count for each hour.

By comparing these two subplots, we can observe how the granularity of time (half-hourly vs. hourly) affects the distribution of passenger traffic across the day. The half-hourly data provides a finer resolution, capturing more detailed fluctuations in passenger flow, while the hourly data offers a more generalized view. This comparison helps in understanding how different time intervals might reveal varying patterns in passenger behavior.

A graph of a graph

Description automatically generated with medium confidence

This plot consists of two subplots, each showing a boxplot comparing passenger counts across different time slots for two different datasets: hourly and half-hourly, with separate visualizations for 2018 data.

The first subplot visualizes the passenger count distribution for hourly data (hour2018). Each boxplot represents the variation in passenger counts across different time slots within the day, with the x-axis representing the time slots and the y-axis representing the passenger count. The boxplots help highlight the spread, median, and any outliers in the passenger count for each time slot.

The second subplot provides a similar analysis but for half-hourly data (halfhour2018). Like the first plot, it visualizes passenger counts across 48 half-hour time slots, allowing for a more granular view of passenger flow throughout the day.

The comparison of these two boxplots helps to understand how the distribution of passenger counts varies across different time granularities (hourly vs. half-hourly), offering insights into the patterns and fluctuations in passenger traffic at different levels of temporal resolution.

A line graph of a graph

Description automatically generated with medium confidence

This analysis involves seasonal decomposition of passenger count data to identify underlying patterns in time series data. The first day's data from each month is extracted, and seasonal decomposition is performed using the additive model. The decomposition separates the data into four components: the original data, the trend, the seasonal component, and the residuals.

The **original data** shows the raw passenger count, while the **trend** component reveals long-term patterns such as overall growth or decline. The **seasonal** component identifies recurring patterns, like weekly or monthly cycles, highlighting peaks during holidays or weekends. The **residuals** show unexplained variations or noise, which could indicate anomalies or unexpected events.

This decomposition is important as it enables the identification of key factors driving passenger fluctuations. By understanding the trend and seasonal components, better forecasting and capacity planning can be achieved, and operational improvements can be made based on predictable patterns in passenger behavior.

5. **Thane-Station-Wagle-Depot 2022 EDA**

This graph presents a comparative analysis of passenger counts in 2023, using both half-hour and hourly data across 12 months. For each month, the graph displays two distinct time series: one representing passenger counts aggregated by half-hour intervals (depicted in blue), and the other representing counts aggregated by hourly intervals (depicted in red). By visualizing the data in this manner, the graph allows for a detailed comparison of passenger flow patterns throughout the year, highlighting variations in temporal resolution and enabling a deeper understanding of how passenger behavior might differ when observed in different time granularities. The subplots for each month provide a clear, month-by-month breakdown, allowing for an insightful examination of seasonal trends and variations in passenger traffic.

A graph of a passenger count

Description automatically generated with medium confidence

2.

A graph showing a graph

Description automatically generated with medium confidence

This specific plot focuses on visualizing the passenger flow for the station pair **Thane Station to Wagle Depot** for a specific date range from **January 1, 2023, to February 1, 2023**. The data is filtered from the hourly dataset to display only the relevant records for this station pair and date range.

The plot shows how passenger counts fluctuate over time during this period, with the x-axis representing time and the y-axis representing the number of passengers. By focusing on this particular station pair and date range, the analysis aims to understand patterns in passenger flow between these two locations for the specified timeframe.

**Why was this specific plot done:**

The plot was created to analyze and observe the passenger count flow for a specific station pair over a defined period. It helps to identify trends, peaks, or fluctuations in passenger traffic during this one-month period, providing insights into how passenger volume varies at a detailed, hourly level. This analysis could be valuable for operational planning, such as resource allocation or scheduling, based on the demand during the given timeframe.

3.

This analysis visualizes passenger count distributions for **Thane Station (Satis)-Wagle Depot** using box plots for both **hourly** and **half-hourly** data. The plots focus on identifying outliers in passenger counts over different time slots. The **hourly** data uses the **OneHourSlot** variable, while the **half-hourly** data uses the **HalfHourSlot** variable. Both plots reveal a minimal number of outliers, suggesting that passenger counts at this station pair follow a consistent distribution with few extreme values. This analysis helps identify typical passenger flow patterns and highlight any unusual data points requiring further investigation.

A graph of a graph of a graph

Description automatically generated with medium confidence

6. Misc Findings :

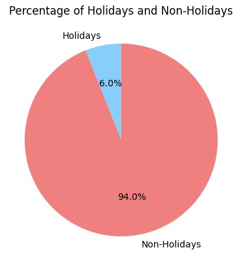
6.1 During which hour is the count of tickets high ?

From the analysis of the plot, it is evident that the highest count of tickets occurs during the time slot from **9 to 10 AM**. This peak suggests that this particular hour is a significant period for ticket sales, indicating a potential trend in customer behavior or demand during morning hours. Conversely, other time slots may show lower ticket counts, highlighting the importance of analyzing hourly data to optimize sales strategies and resource allocation.

A graph showing a number of tickets

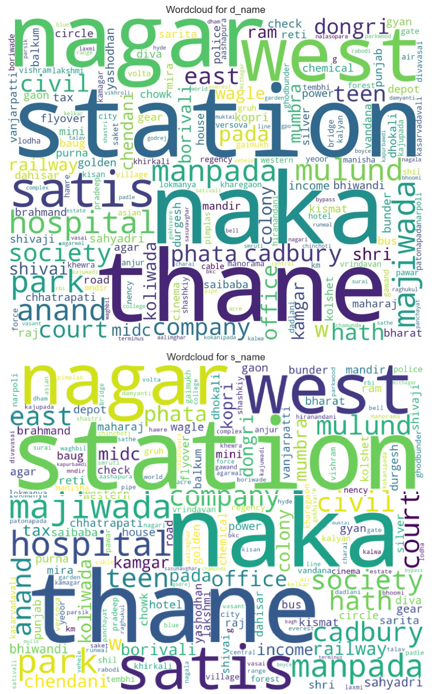
Description automatically generated

6.2 Number of Holiday Counts



The analysis of ticket sales reveals important insights into holiday and non-holiday ticket counts, as well as daily ticket fluctuations. The percentage of holidays is calculated to be **6.0%**, while the **non-holidays** account for **94.0%** of the total days. This distribution highlights that the majority of ticket sales occur outside holiday periods.In terms of ticket counts, the lowest recorded count in a day is **1**, while the highest is **1103**. This significant disparity indicates variability in ticket sales, suggesting that certain days attract far more customers than others.To visualize this data, a pie chart was created to represent the percentage of holidays versus non-holidays, effectively illustrating the predominance of non-holiday ticket sales. Overall, these findings emphasize the need for targeted marketing strategies during peak times and an understanding of customer behavior across different timeframes.

6.3 Word Cloud



1. Data Visualizations :