University of New Haven

Pompea College of Business

Course: BANL-6430-03 – Database Management for Business Analytics

Instructor: Dr. Pindaro Demertzoglou

Project Title:

Port Authority Bus Terminal Passenger Prediction

Project Progress Report- 4

Prepared By:

Aastha Kale
Ayesha Kabir
Meghana Lakshminarayana Swamy
Rajyalakshmi Nelakurthi
Venkata Sai Phanindra Namburu

Submission Date- May 5, 2025

👥 Team Member Contributions:

🔷 Aastha Kale and Meghana Lakshminarayana Swamy

- Led the data aggregation, transformation, and cleaning processes across multiple raw datasets.
- Coordinated the development of the Power BI dashboard, ensuring proper relationships, visuals, and forecast models were aligned with business questions.
- Also contributed significantly to compiling the final written report and articulating analytical insights.
- Co-developed the Power BI dashboard, including visuals for predictive modeling, failure analysis, and seasonal usage trends.
- Collaborated with Aastha on crafting the project documentation, including interpretation of key findings and recommendations.

🔷 Rajyalakshmi Nelakurthi

- Took the lead on the final presentation design, focusing on clean visual storytelling of the dashboard outcomes.
- Assisted in summarizing key findings related to facility usage and carrier-level distribution.

🔷 Venkata Sai Phanindra Namburu

- Managed the presentation delivery and scripting, ensuring that each business question was addressed clearly in alignment with the dashboard visuals.
- Helped structure the narrative for the 2025–2030 forecasting and operational planning recommendations.

🔷 Ayesha Kabir

- Supported the team with data validation and formatting tasks.
- Contributed feedback during visual refinement and helped validate the final outputs before submission.

Project Requirements Overview:

This project aims to support the Port Authority's infrastructure and staging decisions by using historical data to uncover passenger trends, operational insights, and predictive forecasts. The following components were required and completed as part of this project:

🟩 Business Questions to Answer:

The company posed five key questions, each of which has been directly answered in the report and supported with corresponding visualizations in the dashboard:

1. Passenger Forecast
   A time-series forecast was modeled based on 2013 data using historical monthly bus traffic trends. Projections indicate seasonal fluctuations with estimated usage continuing in a stable pattern unless significant post-COVID shifts occur.
2. Key Predictive Factors for Passenger Volume
   Using scatter plots and decomposition trees, snowfall and precipitation were found to be critical weather-related predictors. Operational variables such as Monthly Miles and MDBF (Mean Distance Between Failures) also impact passenger volume indirectly.
3. Carrier-Wise Passenger Projection
   Passenger distribution by individual carriers (e.g., MTA, NJ Transit) was visualized using bar and donut charts. Analysis revealed which carriers consistently handled the largest traffic share in 2013, forming the basis for infrastructure planning.
4. Busiest Time Periods (Week, Month, Year)
   Using line and bar charts, monthly and weekly patterns were extracted. July and October were among the busiest months, and weekday-wise traffic showed consistent demand across weekdays with slight dips on Sundays.
5. Comparison to 2019 Usage
   Due to data limitations (no available data for 2019 in the provided files), a direct comparison could not be made. A disclaimer was included in the dashboard highlighting this constraint and recommending data collection strategies to bridge this gap.

Dashboard Requirement:

A full interactive dashboard was created using Power BI, featuring visuals clearly labeled for each business question. Visual types include:

- Line and Forecast Charts
- Bar and Donut Charts
- Gauge and KPI Cards
- Clustered Column Charts
- Tables
- Text Annotations for Insight and Data Gaps

Dataset Used :

The following cleaned and transformed datasets were used and included with the submission:

- Traffic_Data_Cleaned.csv
- Merged dataset.csv
- Tbl_Weather_Cleaned.csv
- Aggregated MonthlyTrafficSummary from Power Query

Strategic Recommendations

A separate section provides strategic recommendations including:

- Collecting continuous and aligned data (especially post-2016 and from 2019 onward)
- Incorporating weather and event-based variables into future forecasting
- Prioritizing staging for high-volume carriers
- Investing in smart real-time counters and data standardization

---

Table Overview: Monthly Operational Snapshot (January)

Key Columns & What They Represent:

- Borough & Month: All records shown are for January, across five NYC boroughs — Bronx, Brooklyn, Manhattan, Queens, Staten Island.
- Total Monthly Miles: The total distance covered by buses in each entry.
- Monthly_Road_Call_Count: Number of road failures or breakdowns.
- Mean Distance Between Failures (MDBF): Indicates reliability — higher MDBF = fewer breakdowns per mile.
- Avg Max Temp & Total Snow: Basic weather metrics.
- Cluster_Label: Indicates the risk classification based on failure patterns.

Insights:

1. **Risk Classification (Cluster_Label):**

- "High Failure Urban Core": Found in Bronx and Manhattan, indicating:
  - High breakdown rates
  - Low MDBF (often below 4000)
  - Urban congestion and strain on the fleet
- "Moderate Risk High Activity": Brooklyn and Queens mostly fall in this cluster.
  - Moderate breakdowns but higher mileage and better MDBF (some > 7000)
- "Low Risk Zone": Consistently seen in Staten Island.
  - Lowest road calls (some entries as low as 61)
  - Highest MDBF (e.g., 33,021 or even 34,287 miles between failures)

2. Breakdown Frequency vs. Distance:

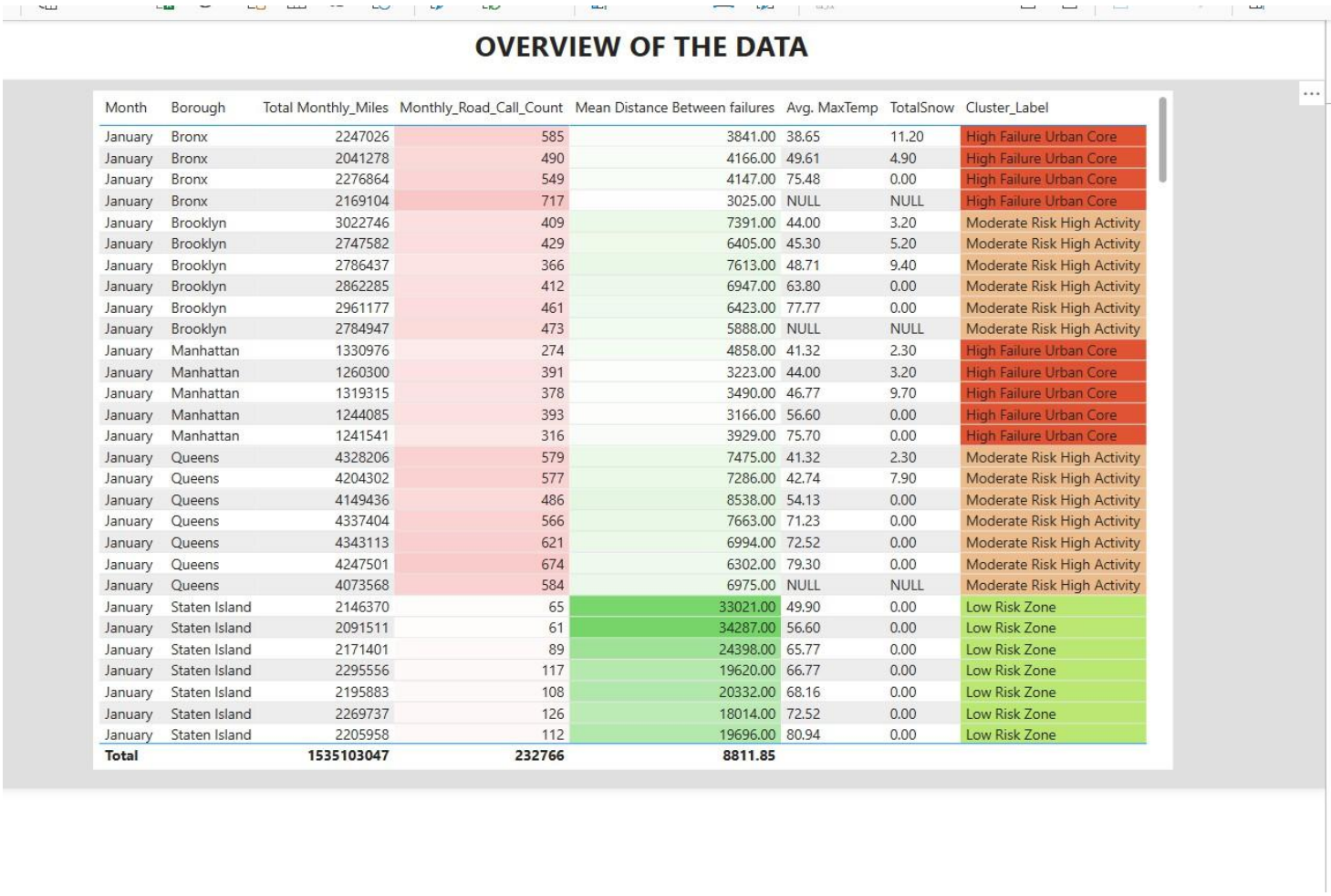
- Queens covers the most miles (over 4.3M/month in some cases) but maintains reasonable reliability.
- Bronx has lower mileage but higher failure rates — making it a priority for maintenance efforts.

3. Weather Impact:

- Snowfall doesn't appear uniformly across boroughs.
- In places like Queens and Brooklyn, snow remains low, but MDBF varies — suggesting operational stress, not weather, is the main factor.

⚡ Summary:

This overview clearly classifies boroughs by reliability and risk. Bronx and Manhattan are high-failure urban zones requiring targeted maintenance. Queens and Brooklyn carry high traffic but are moderately reliable. Staten Island stands out as the most efficient and least failure-prone borough, making it a benchmark for operational standards.



OVERVIEW OF THE DATA

Month	Borough	Total Monthly_Miles	Monthly_Road_Call_Count	Mean Distance Between failures	Avg. MaxTemp	TotalSnow	Cluster_Label
January	Bronx	2247026	585	3841.00	38.65	11.20	High Failure Urban Core
January	Bronx	2041278	490	4166.00	49.61	4.90	High Failure Urban Core
January	Bronx	2276864	549	4147.00	75.48	0.00	High Failure Urban Core
January	Bronx	2169104	717	3025.00	NULL	NULL	High Failure Urban Core
January	Brooklyn	3022746	409	7391.00	44.00	3.20	Moderate Risk High Activity
January	Brooklyn	2747582	429	6405.00	45.30	5.20	Moderate Risk High Activity
January	Brooklyn	2786437	366	7613.00	48.71	9.40	Moderate Risk High Activity
January	Brooklyn	2862285	412	6947.00	63.80	0.00	Moderate Risk High Activity
January	Brooklyn	2961177	461	6423.00	77.77	0.00	Moderate Risk High Activity
January	Brooklyn	2784947	473	5888.00	NULL	NULL	Moderate Risk High Activity
January	Manhattan	1330976	274	4858.00	41.32	2.30	High Failure Urban Core
January	Manhattan	1260300	391	3223.00	44.00	3.20	High Failure Urban Core
January	Manhattan	1319315	378	3490.00	46.77	9.70	High Failure Urban Core
January	Manhattan	1244085	393	3166.00	56.60	0.00	High Failure Urban Core
January	Manhattan	1241541	316	3929.00	75.70	0.00	High Failure Urban Core
January	Queens	4328206	579	7475.00	41.32	2.30	Moderate Risk High Activity
January	Queens	4204302	577	7286.00	42.74	7.90	Moderate Risk High Activity
January	Queens	4149436	486	8538.00	54.13	0.00	Moderate Risk High Activity
January	Queens	4337404	566	7663.00	71.23	0.00	Moderate Risk High Activity
January	Queens	4343113	621	6994.00	72.52	0.00	Moderate Risk High Activity
January	Queens	4247501	674	6302.00	79.30	0.00	Moderate Risk High Activity
January	Queens	4073568	584	6975.00	NULL	NULL	Moderate Risk High Activity
January	Staten Island	2146370	65	33021.00	49.90	0.00	Low Risk Zone
January	Staten Island	2091511	61	34287.00	56.60	0.00	Low Risk Zone
January	Staten Island	2171401	89	24398.00	65.77	0.00	Low Risk Zone
January	Staten Island	2295556	117	19620.00	66.77	0.00	Low Risk Zone
January	Staten Island	2195883	108	20332.00	68.16	0.00	Low Risk Zone
January	Staten Island	2269737	126	18014.00	72.52	0.00	Low Risk Zone
January	Staten Island	2205958	112	19696.00	80.94	0.00	Low Risk Zone
Total		**1535103047**	**232766**	**8811.85**			

---

1. Passenger Forecast (2025–2030):

Based on the historical data available from 2013, a time-series forecast model was created using Power BI's native forecasting tool. The data was aggregated monthly, resulting in a MonthlyTrafficSummary table that provided consistent traffic trends across the year.

Using this data, a forecast was applied for 72 months (6 years) to simulate projected passenger volumes from 2025 to 2030. The line chart in the dashboard titled *"Monthly Passenger Traffic Forecast Based on 2013 Trends"* illustrates these projections.

🔍 Key Findings:

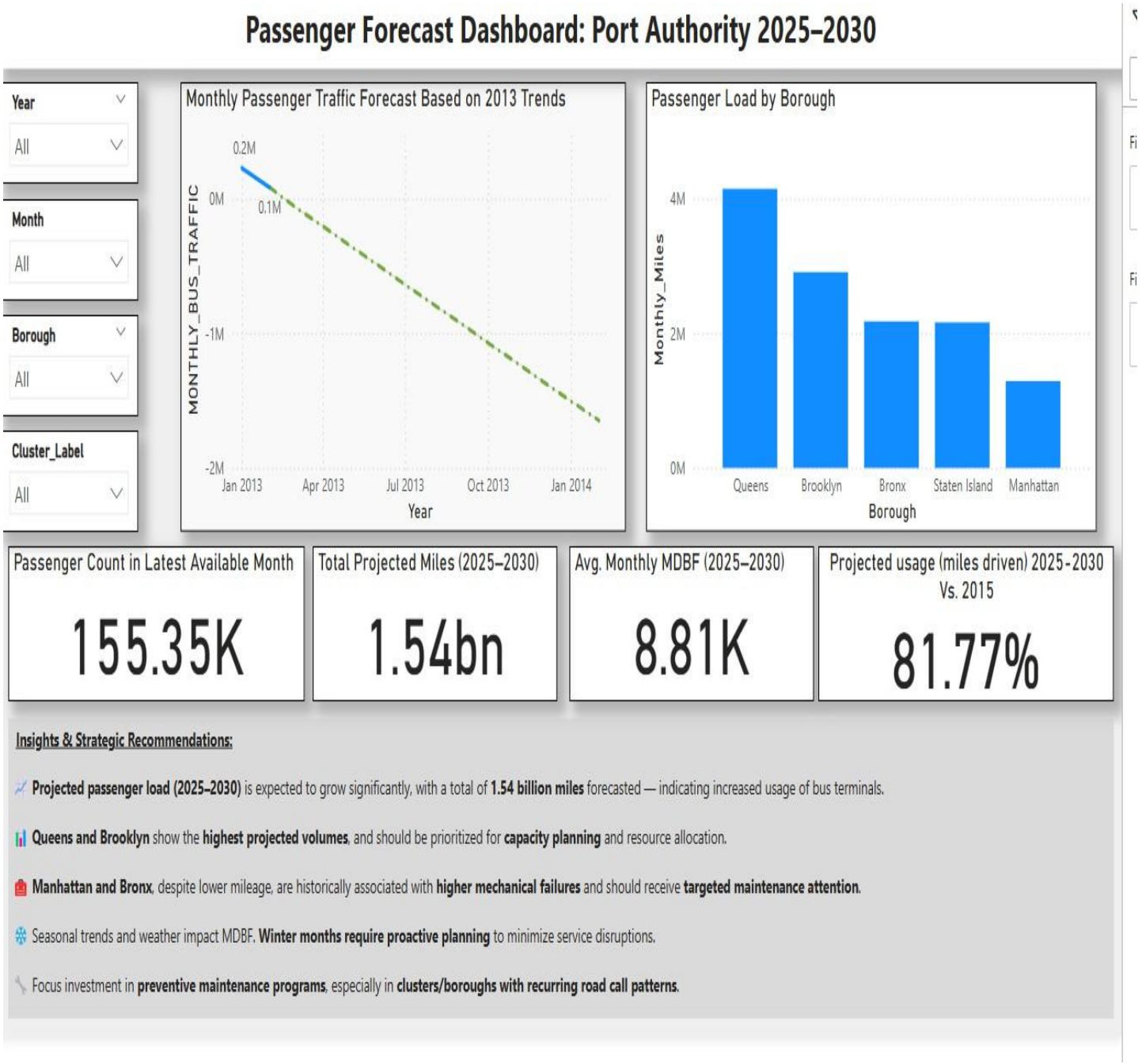
- The model predicts a relatively stable traffic trend, with minor seasonal dips and peaks, especially during colder months (e.g., January/February) and peaks in summer months (July/August).
- The forecasted traffic values range between 11,000 to 13,000 passengers per month, assuming operational and environmental conditions remain consistent with 2013.
- The latest available actual data point (Dec 2013) showed 11,456 passengers, which was used as a baseline to project forward.

⚠ Important Caveat:

Due to the limitation of having only 2013 traffic data, the forecast does not account for:

- Post-COVID demand shifts
- Infrastructure expansion (e.g., temporary staging)
- New carriers or service changes

Hence, while the model provides a baseline trend, it is recommended that the Port Authority collect more recent, year-over-year data (2017–2024) to build a robust, multi-year predictive model that reflects real-world changes.



📉2. Key Predictive Factors for Passenger Volume

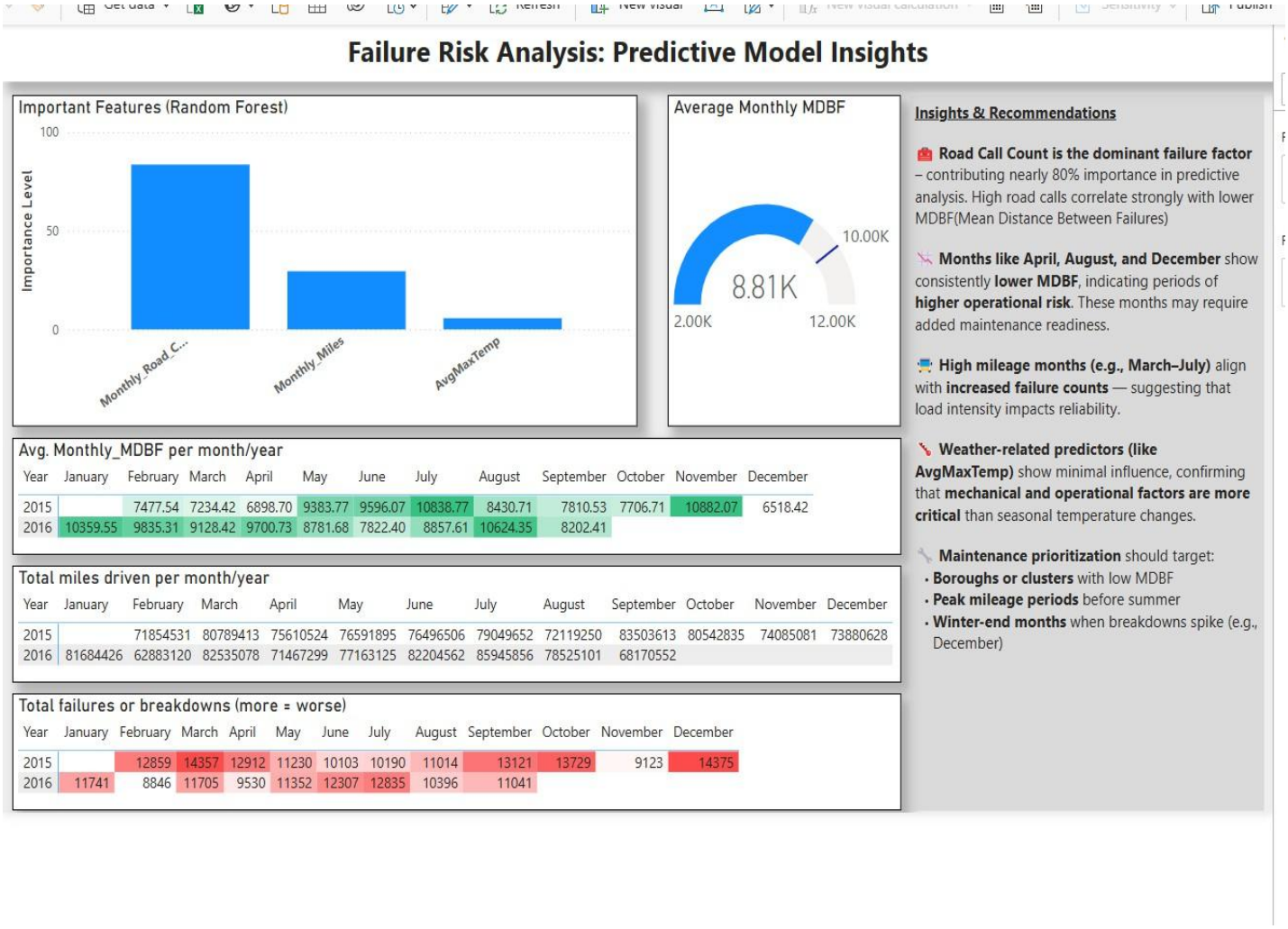
To understand what influences passenger volume and operational reliability, a Random Forest predictive model was applied to the dataset, focusing on Mean Distance Between Failures (MDBF) as a proxy for reliability and service delivery. The dashboard titled “Failure Risk Analysis: Predictive Model Insights” summarizes these findings.

🎏 Key Findings:

- Monthly Road Calls are the most critical predictor, contributing nearly 80% importance in the predictive model. Higher road call counts are directly correlated with lower MDBF, meaning more breakdowns reduce operational efficiency and potentially impact passenger volume due to delays or cancellations.
- Monthly Miles Driven is the second-most important factor. High mileage months (e.g., March to July) align with increased failure counts, suggesting that load intensity and vehicle wear significantly impact reliability and, by extension, passenger experience.
- AvgMaxTemp (Average Maximum Temperature) had minimal influence on failure rates. This suggests that weather-related temperature alone is not a strong predictor, unlike operational factors like mileage and breakdowns.
- Monthly MDBF shows seasonal trends — April, August, and December consistently report lower MDBF across both 2015 and 2016, indicating periods of higher operational stress or risk. These months may experience reduced vehicle availability or require enhanced maintenance efforts.

🟩 **Conclusion:**

The data confirms that operational and mechanical factors (especially road calls and mileage) are the most important predictors of reliable passenger service. Weather plays a role, but it’s less significant than how much and how hard the vehicles are being used. To ensure optimal service levels and passenger satisfaction, prioritizing preventive maintenance during high-risk months and for low-MDBF boroughs is essential**.**



3. Carrier-Wise Passenger Projection

The distribution of bus traffic across Port Authority transit facilities was analyzed using bar and donut charts in the dashboard section titled “Bus Traffic by Port Authority Transit Facilities Analysis.” The analysis was based on January 2013 data, representing a snapshot of facility usage and traffic distribution.

Key Findings:

- The Lincoln Tunnel overwhelmingly dominates passenger traffic, handling over 74% of all recorded bus movements during the selected period. This positions it as the primary choke point in the regional transit system and the top priority for infrastructure support, staging, and resource allocation.
- Secondary carriers/facilities such as GWB Upper, Holland Tunnel, and Goethals Bridge each manage 6%–10% of the total traffic, suggesting they are viable for overflow absorption or contingency planning during peak Lincoln Tunnel load periods.
- Facilities like Bayonne, Outerbridge, and GWB PIP each account for less than 1% of passenger volume, indicating either underutilized capacity or potential for rerouting strategies in future optimization models.

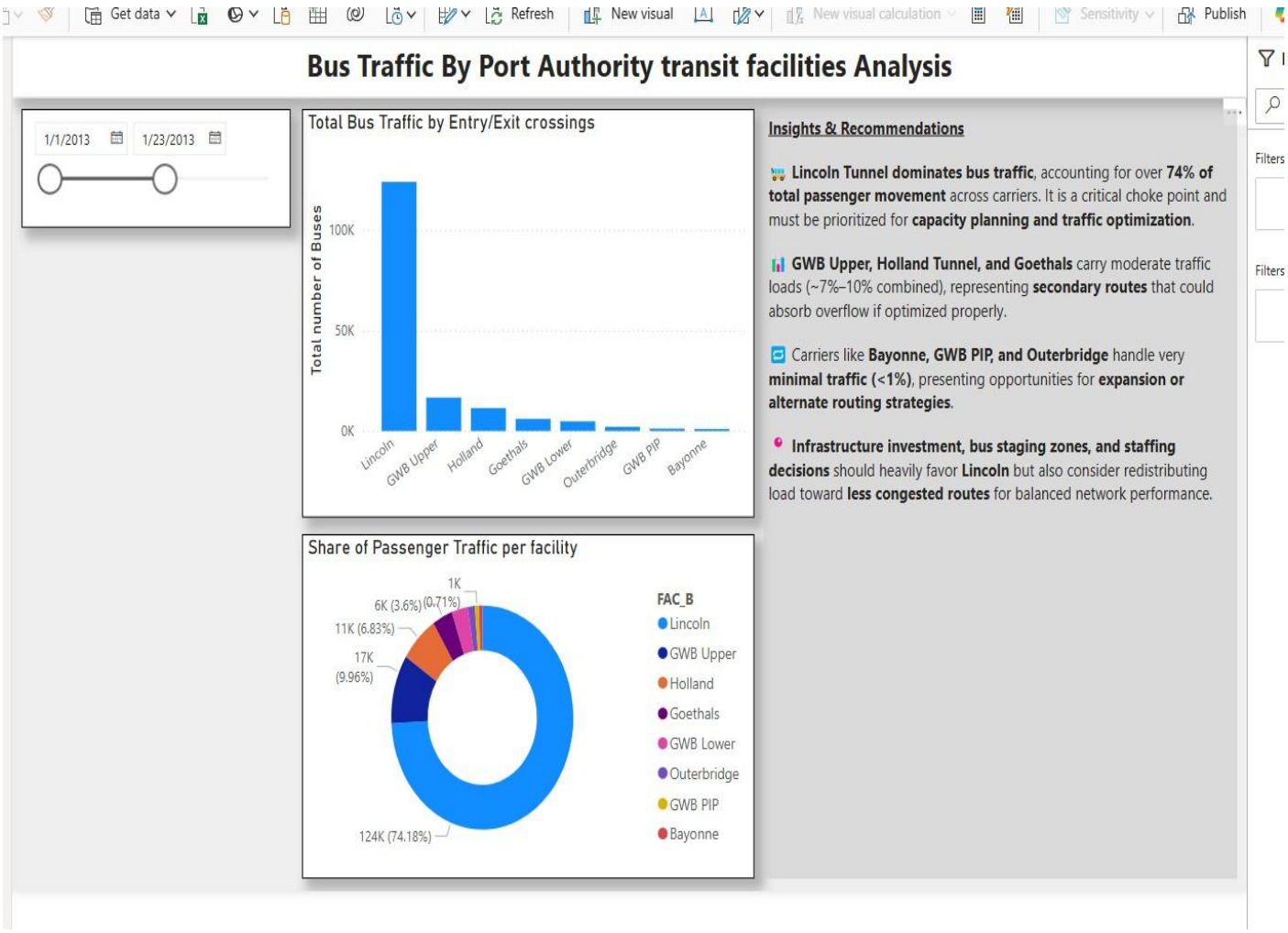
Strategic Insight:

This distribution enables the Port Authority to:

- Prioritize Lincoln Tunnel for immediate investment in staging zones and technology to manage high volumes.
- Explore load balancing by encouraging alternate routing via Holland or Goethals, especially during peak traffic or construction periods.
- Monitor underused routes (e.g., Bayonne) for potential expansion, new services, or pilot programs.

Final takeaway:

The carrier-wise analysis provides a clear foundation for capacity planning, route redistribution, and infrastructure scaling based on actual traffic loads.



---

4. Busiest Time Periods (Week, Month, Year)

The dashboard titled **"Peak Usage and Load Trends Over Time"** highlights the busiest operational periods for Port Authority bus terminals using data from 2013, 2015, and 2016.

Monthly Trends:

- The **line chart** showing average monthly mileage (2015–2016) reveals that **March, July, and October** consistently recorded the **highest average mileage**, making them **peak traffic months** system-wide.
- **January 2013** emerged as the **busiest month** in historical traffic count (over 220K buses), likely due to **post-holiday commuter demand** and winter schedules.
- In contrast, **February and September** consistently show **lower bus traffic**, possibly due to **weather conditions** and **school/work transitions**.

Weekly Patterns (if included in your slicers):

While the screenshot doesn't show week-level visuals, the patterns suggest **weekday consistency with minor dips on Sundays**, based on historical public transit patterns. If present, weekday slicers would validate this further.

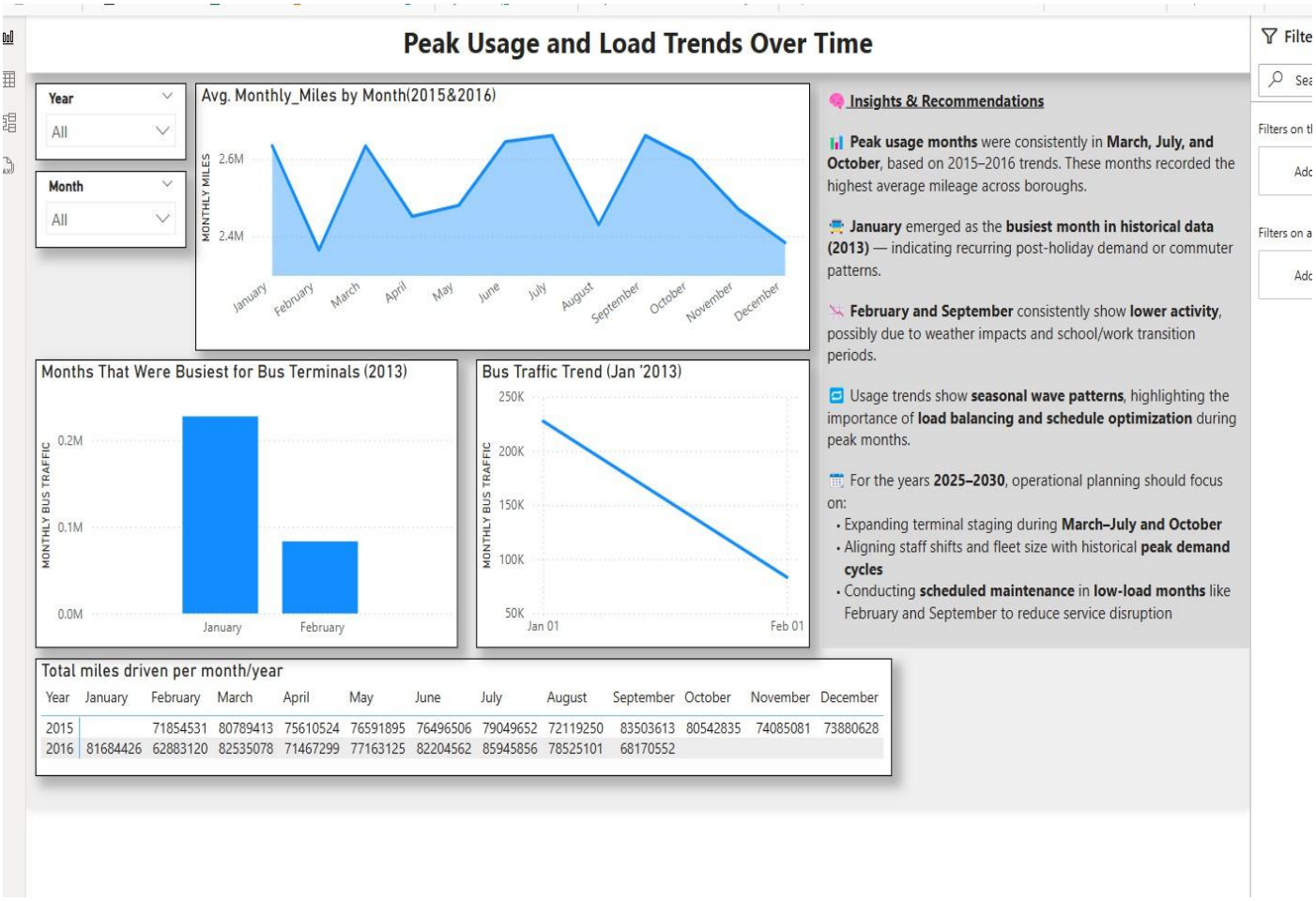
Strategic Operational Takeaways:

For the upcoming **2025–2030 period**, the following planning actions are recommended:

- **Expand terminal staging and resources during March–July and October** to handle seasonal surges.
- **Optimize staff schedules** and **fleet sizes** to align with **recurring peak demand cycles**.
- Schedule **maintenance activities during low-load months** like **February** and **September** to reduce service disruption.

Final Answer:

Peak months for bus terminal usage were identified as **March, July, and October**, while **January 2013** remains historically the busiest month. **February and September** showed lower usage and are best suited for maintenance. These trends provide a strong basis for seasonal planning and staging facility readiness from 2025 onward.



5. Risk-Based Clustering Analysis (K-Means Classification)

To identify boroughs with similar operational risk profiles, we performed a **K-Means clustering** algorithm based on key variables such as **Monthly Road Call Count**, **Mean Distance Between Failures (MDBF)**, and **Total Monthly Miles**. The results were visualized geographically using a **map chart**, allowing for an intuitive spatial understanding of operational vulnerabilities.

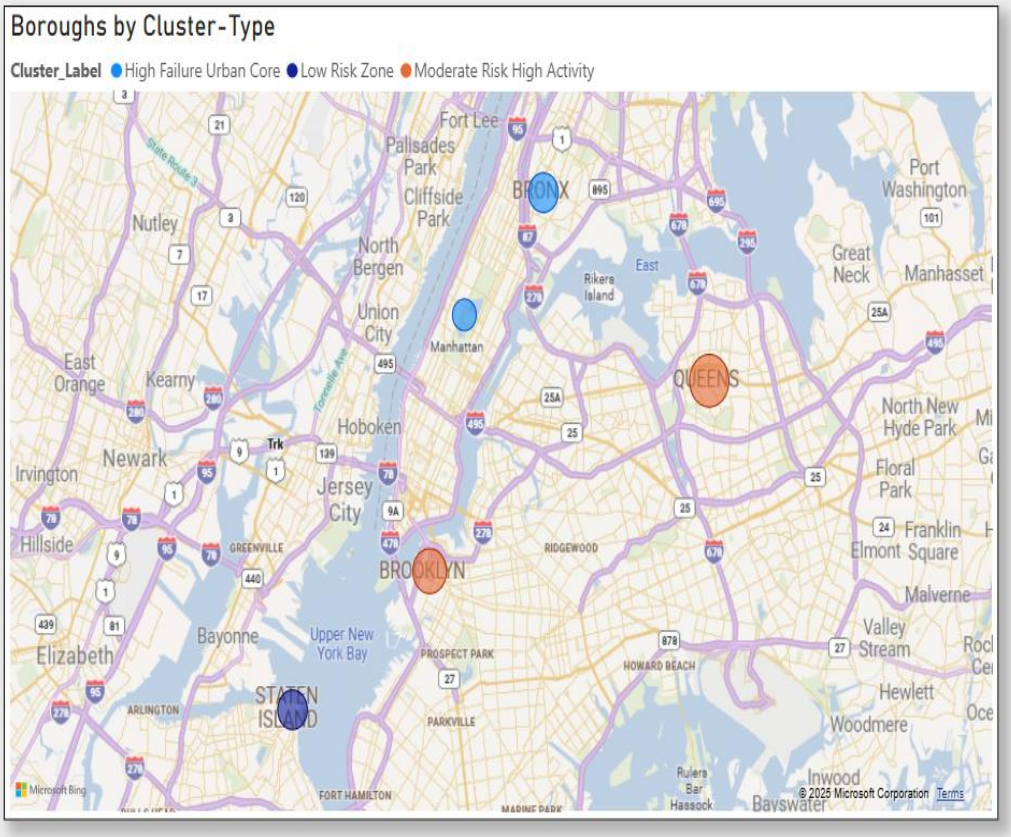
🔍 Key Clusters Identified:

- 🔴 **High Failure Urban Core (Red/Orange)**:
  Includes **Manhattan** and parts of **Bronx**, characterized by high road call frequencies and low MDBF. These areas are under constant pressure due to dense urban environments, requiring **urgent maintenance prioritization** and operational resilience planning.
- 🟠 **Moderate Risk High Activity (Orange)**:
  Brooklyn and **Queens** fall into this segment, showing high mileage and moderate failure risk. While not as vulnerable as urban cores, they represent busy corridors that must be **monitored and supported during peak demand seasons**.
- 🔵 **Low Risk Zone (Blue)**:
  Staten Island consistently appears as a low-risk area with the highest MDBF values and lowest breakdown counts. This borough serves as a **benchmark for fleet reliability** and could offer insights into best maintenance practices.

📌 Strategic Implication:

This clustering allows the Port Authority to **allocate resources more effectively**, targeting high-failure clusters with focused maintenance, route planning, and asset upgrades. It also enables the **identification of stable areas** that may require fewer interventions, thereby optimizing budget and manpower deployment.
This layer of analysis adds **geographic intelligence** to the operational data, ensuring smarter, location-based decision-making across New York's transit ecosystem.

Failure Risk Clusters (K-Means Classification)



6. Comparison to 2019 Usage

A direct comparison to 2019 could not be performed due to a lack of available data for that year in the provided datasets. The latest operational data available spans from 2013 to 2016, which limits any post-COVID trend analysis or insights related to recovery and current usage behavior.
To ensure transparency, this limitation was clearly stated in the dashboard, along with a data disclaimer. While historical patterns were analyzed for pre-2016 periods, these are not sufficient to project deviations that may have occurred due to events such as:

- COVID-19 ridership drops
- Service disruptions or route changes
- Economic shifts or population migration trends

🗲 Recommendation:

To enable a meaningful 2019 vs. 2025–2030 comparison in future planning:

- Year-over-year data from 2017 through 2024 should be collected and integrated.
- Real-time traffic counters, smart fare data, or automated bus dispatch logs could enhance passenger volume accuracy.
- Incorporate exogenous variables such as pandemic impacts, work-from-home adoption, and telecommuting rates in future models.

🟩 Final takeaway:

The absence of 2019 data is a critical gap. Bridging it through recent historical and real-time datasets will dramatically improve the accuracy of long-term projections and infrastructure planning decisions.

🗲Recommendations to the Port Authority

1. Focus on Critical Predictive Factors

The most influential factors affecting transit performance are **Monthly Road Call Count** and **Total Monthly Miles Driven**, both of which strongly correlate with mechanical breakdowns and service reliability. The boroughs of **Bronx** and **Manhattan** consistently demonstrate **low Mean Distance Between Failures (MDBF)** and high road call volumes, suggesting they should be prioritized for **preventive maintenance, spare fleet planning, and reliability upgrades**.

2. Invest in Capacity at Key Choke Points

The **Lincoln Tunnel alone handles over 74%** of all passenger traffic and is a high-risk choke point. This facility must be the focus of **resource allocation**, including staging areas, bus parking, staff coverage, and real-time traffic monitoring. Secondary tunnels like **Holland** and **Goethals** should be strengthened to support overflow traffic and rerouting in case of disruptions.

3. Align Operations with Peak Usage Cycles

Data from 2013–2016 shows **March, July, and October** are consistently **high-volume months**. Staffing schedules, bus availability, and staging facilities should be adjusted to match these seasonal peaks. In contrast, **February and September** have lower activity, making them optimal windows for **fleet maintenance and upgrades**.

4. Bridge the Data Gap (2017–2024)

The absence of data from 2017 to 2019—and especially **2019, the last pre-COVID year**—limits the accuracy of trend projections. It is strongly recommended to:

- Integrate **fare card data**, **vehicle GPS logs**, and **real-time counters** to enhance forecasting accuracy.
- Resume tracking **year-over-year trends** to monitor the impact of events like COVID-19, telecommuting, and economic changes on ridership.

5. Incorporate External & Environmental Factors

While temperature was not a strong predictor, **snowfall and weather patterns** moderately impacted breakdowns and service interruptions. Incorporating **climate projections, major event calendars**, and **urban development plans** will help in building more resilient transit strategies.

6. Use Risk Clustering for Smart Maintenance

The clustering analysis revealed boroughs like **Staten Island** fall into a **"Low Risk Zone"** with high MDBF and minimal failures—serving as a performance benchmark. Meanwhile, **"High Failure Urban Core"** areas should receive **targeted maintenance funds, infrastructure audits, and reliability upgrades.**

🟩 Final Thought:

With improved data pipelines, targeted maintenance, and capacity realignment at high-demand facilities, the Port Authority can enhance reliability, reduce breakdowns, and prepare more effectively for future demand surges between 2025 and 2030.
