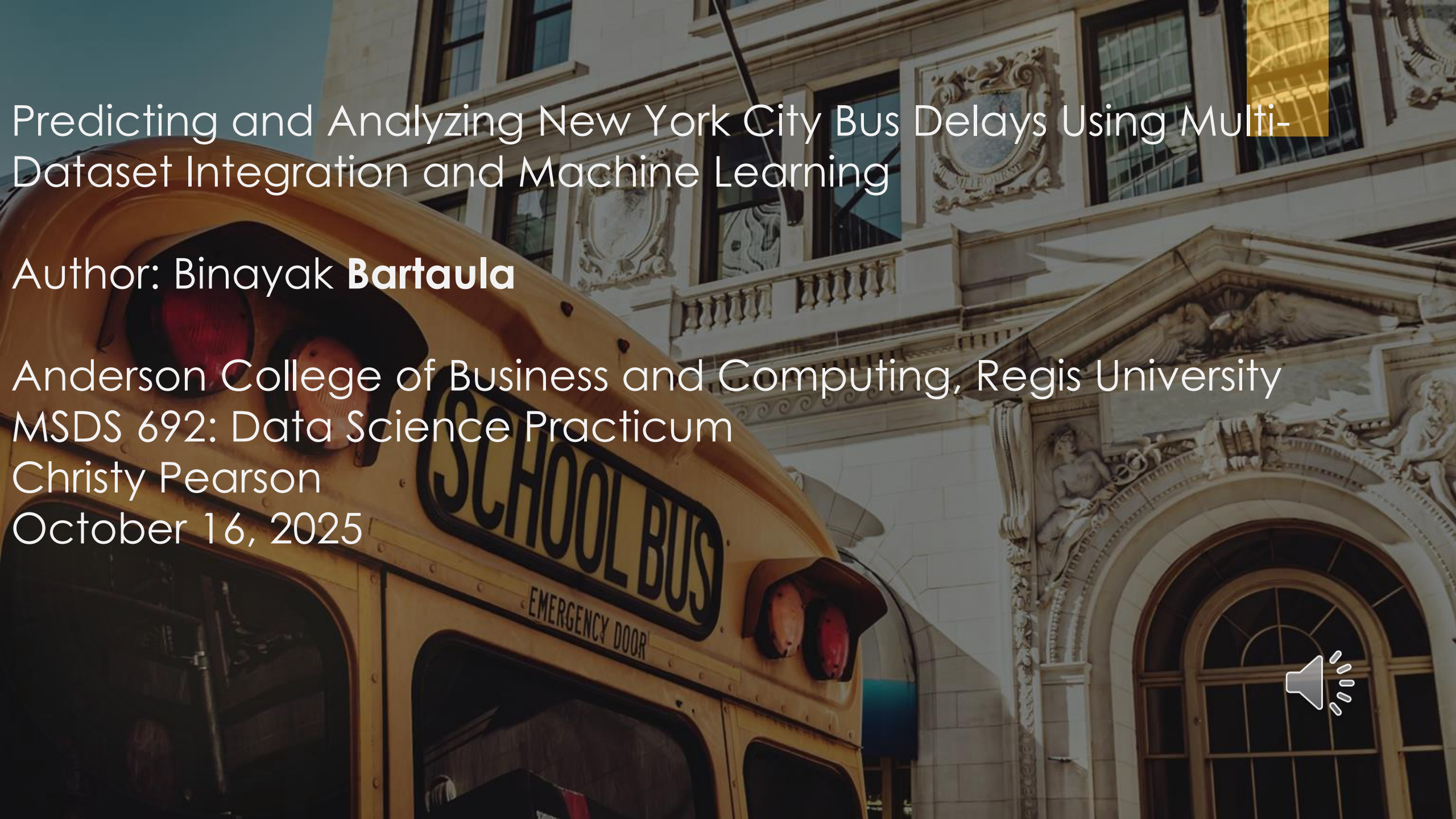
Predicting and Analyzing New York City Bus Delays Using Multi-Dataset Integration and Machine Learning

Author: Binayak **Bartaula**

Anderson College of Business and Computing, Regis University
MSDS 692: Data Science Practicum
Christy Pearson
October 16, 2025



The Daily Challenge of NYC Transit

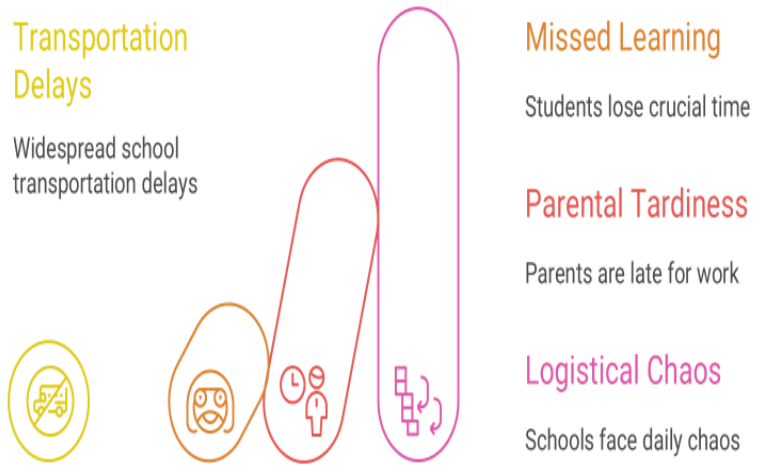
Why This Matters

•NYC's massive school transportation system causes frequent delays.

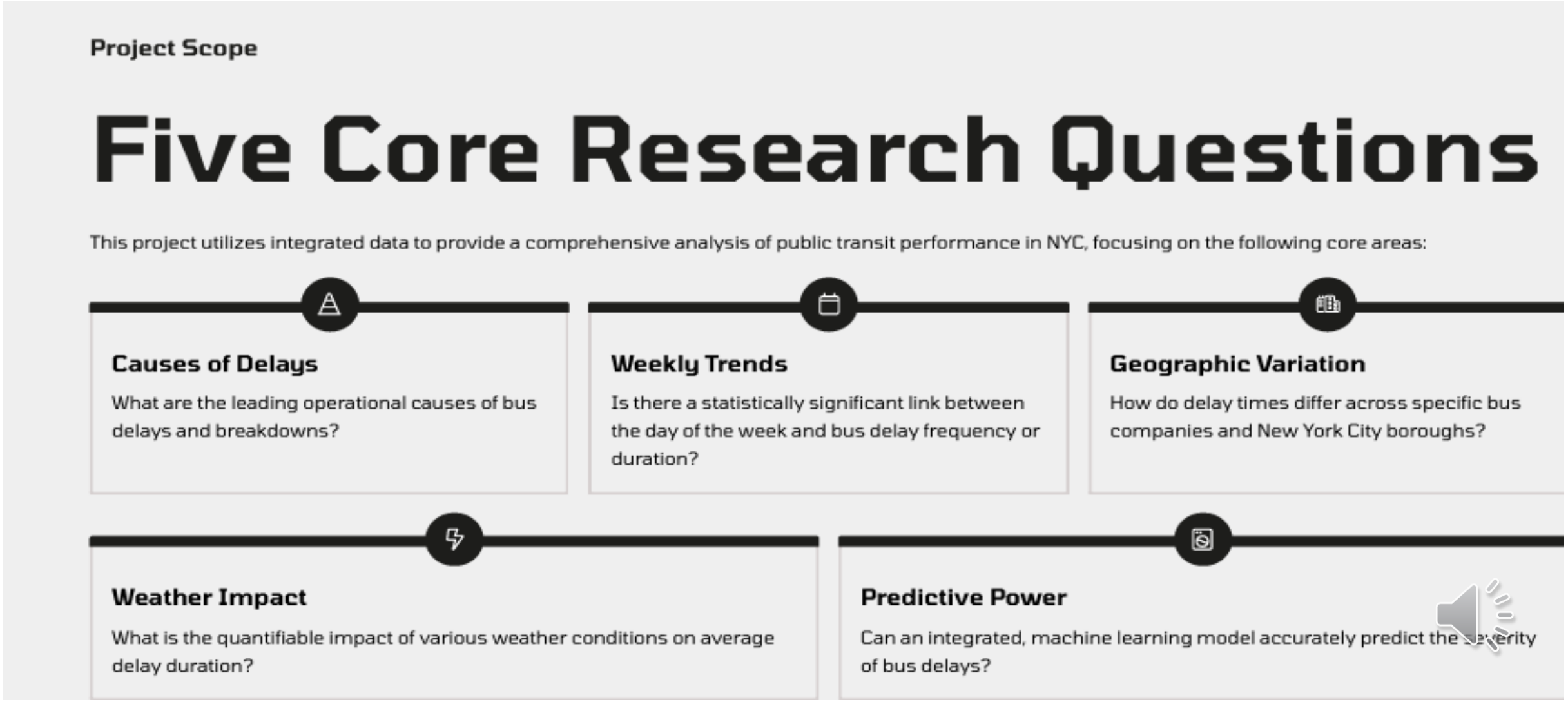
These delays impact everyone:

- Students miss crucial learning time.
- Parents are often late for work.
- Schools face daily logistical chaos.

Our Goal: Use data to find patterns that improve communication and resource management.



Five Core Research Questions



Data Sources

1. NYC Bus Operations Data 🚌

- A decade of operational history from NYC Open Data.
- Contains over 700,000 individual bus incidents

2. Daily Weather Intelligence 🌦️

- Comprehensive daily weather metrics from NOAA.
- Monitoring everything from snowfall and fog to temperature and wind speed.

Our Final Analytical Dataset

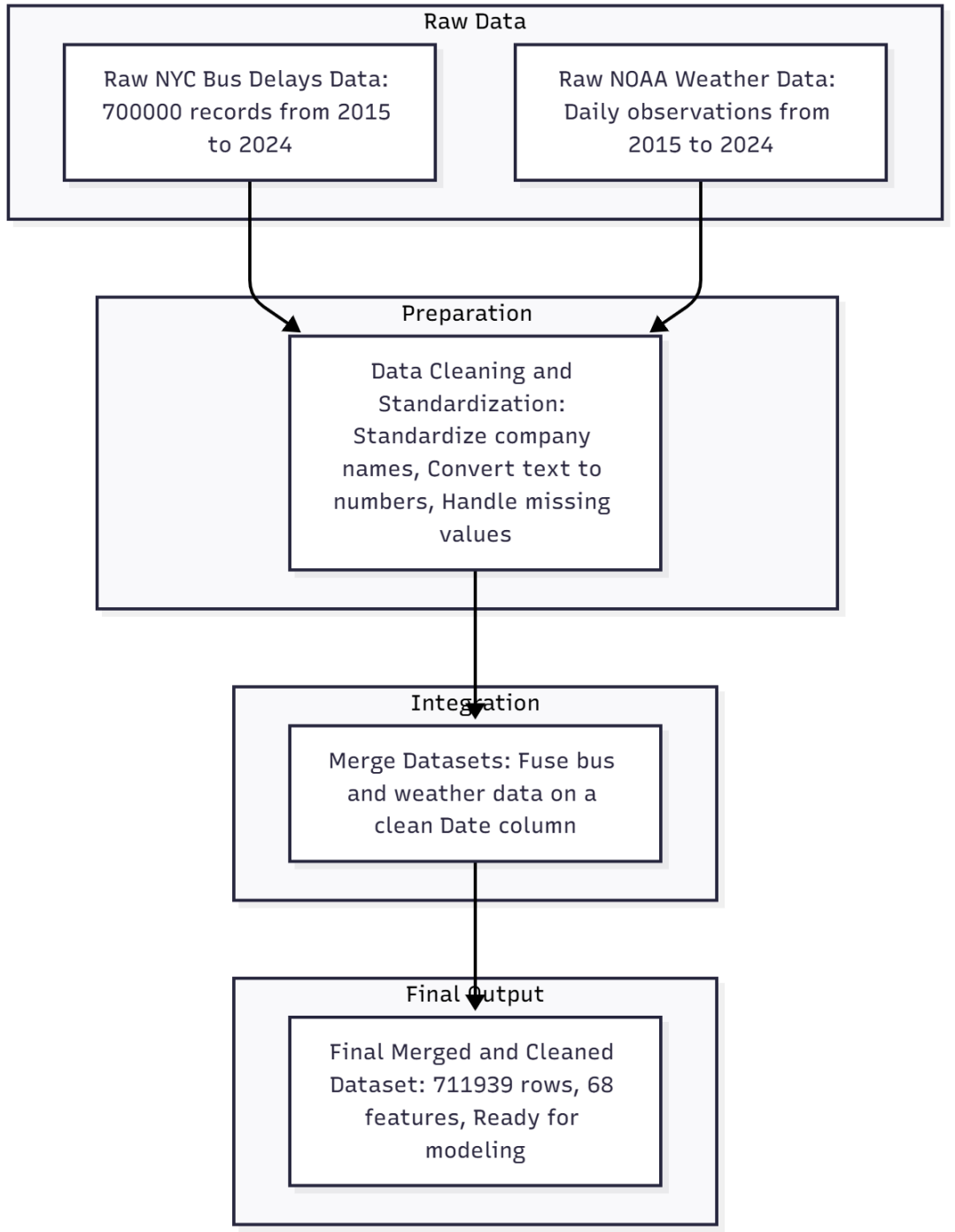
By fusing these two sources, I have created a robust, model-ready dataset with:

- 711,939 Enriched Rows
- 68+ Predictive Features

Data Cleaning & Preparation

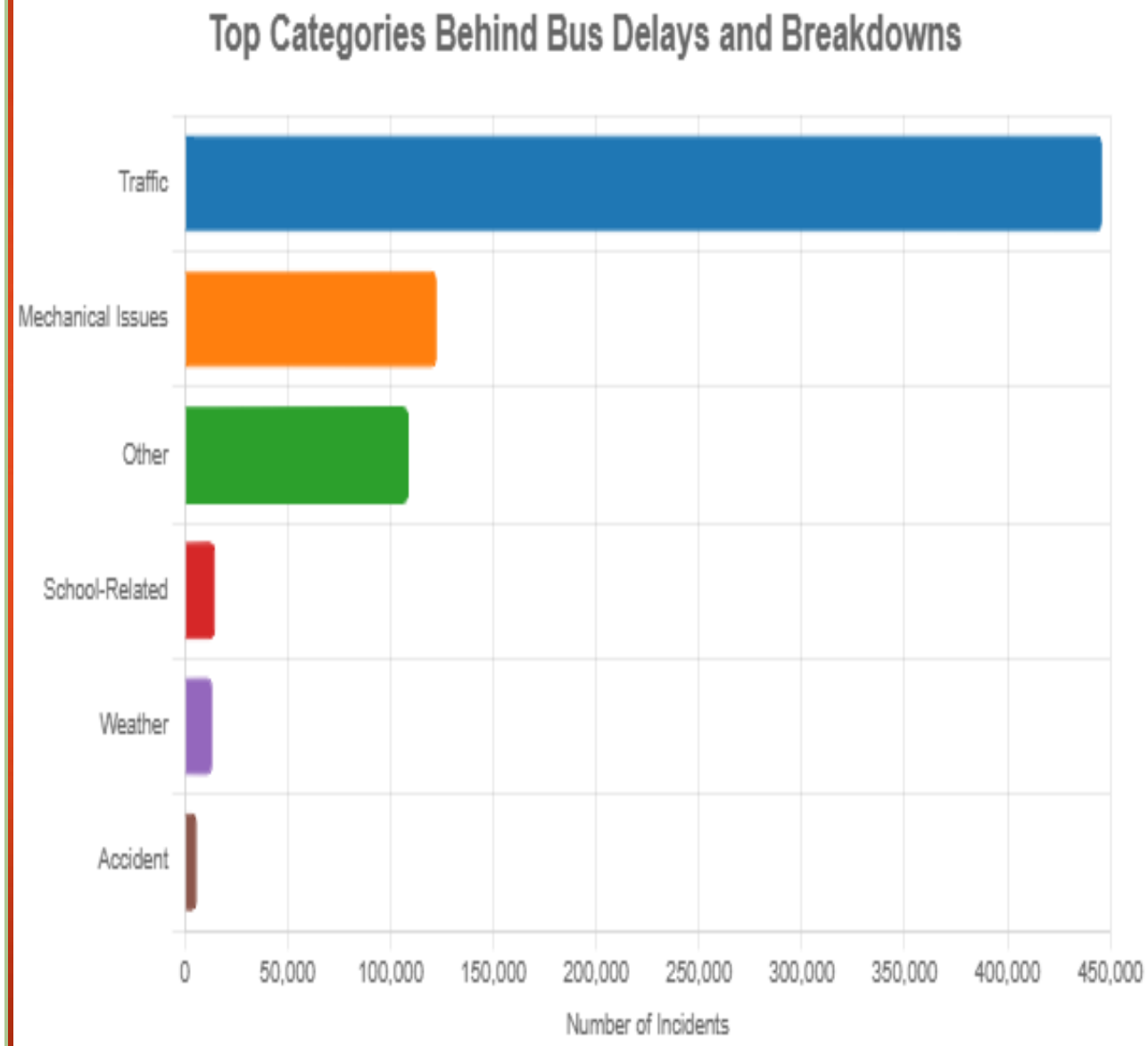
I executed a rigorous data overhaul to forge a pristine, model-ready dataset.

- Merged and cleaned 700,000+ records from NYC Bus and NOAA Weather datasets.
- Standardized formats, removed duplicates/outliers, and filled missing values.
- Streamlined the final dataset by eliminating all irrelevant columns and metadata.
- Achieved a final, flawless dataset of over 711,000 rows with zero missing values.



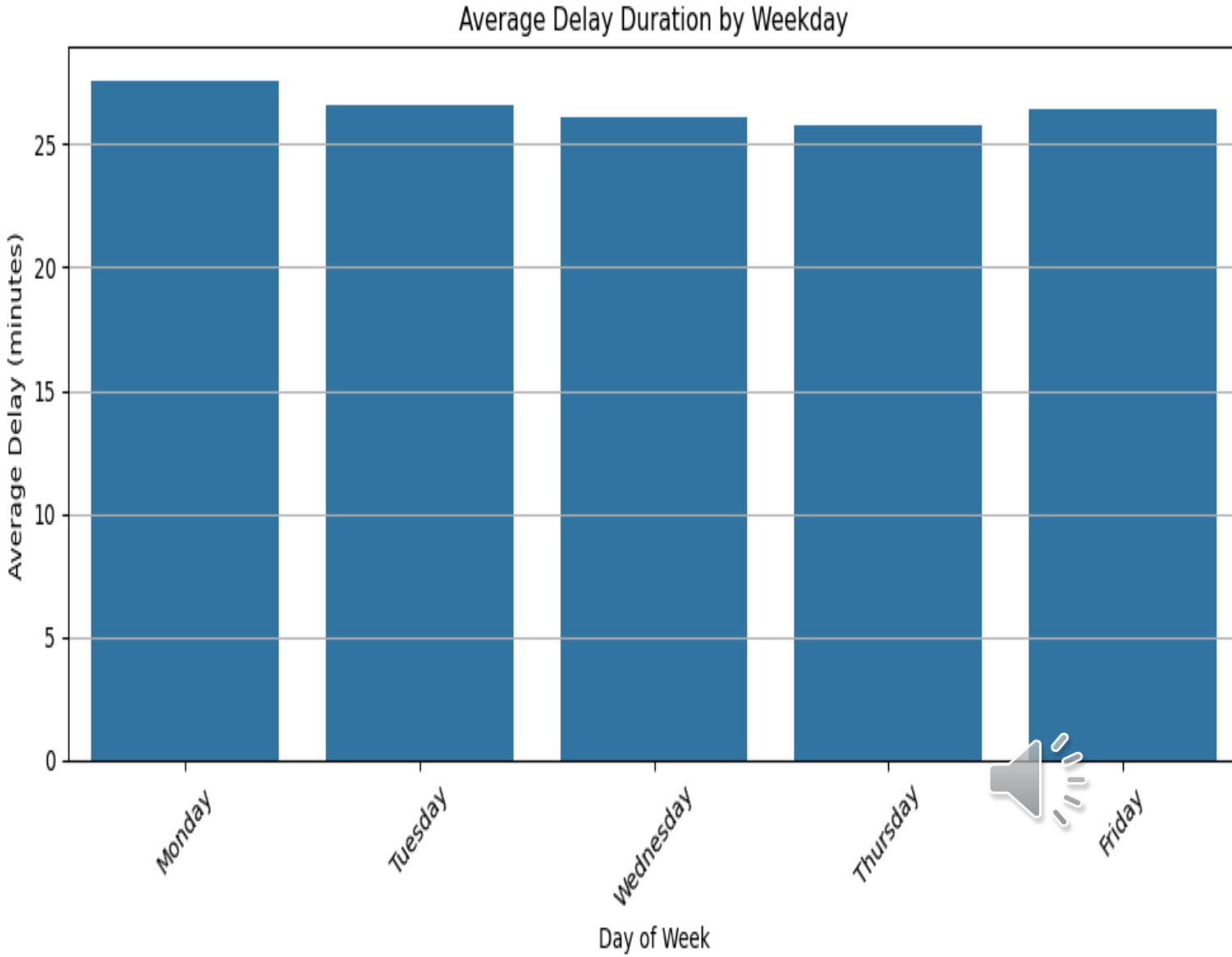
Research Question 1: What are the leading causes of delays and breakdowns?

- Traffic delays: 62% of incidents (446K+ cases), the primary cause.
- Mechanical issues: 17% (122K cases), the second major factor.
- Other causes: 15% (109K cases), needs further exploration.
- School & Weather : ~4% combined (14.8K & 13.6K cases).
- Accidents: <1% (5.9K cases), low impact.



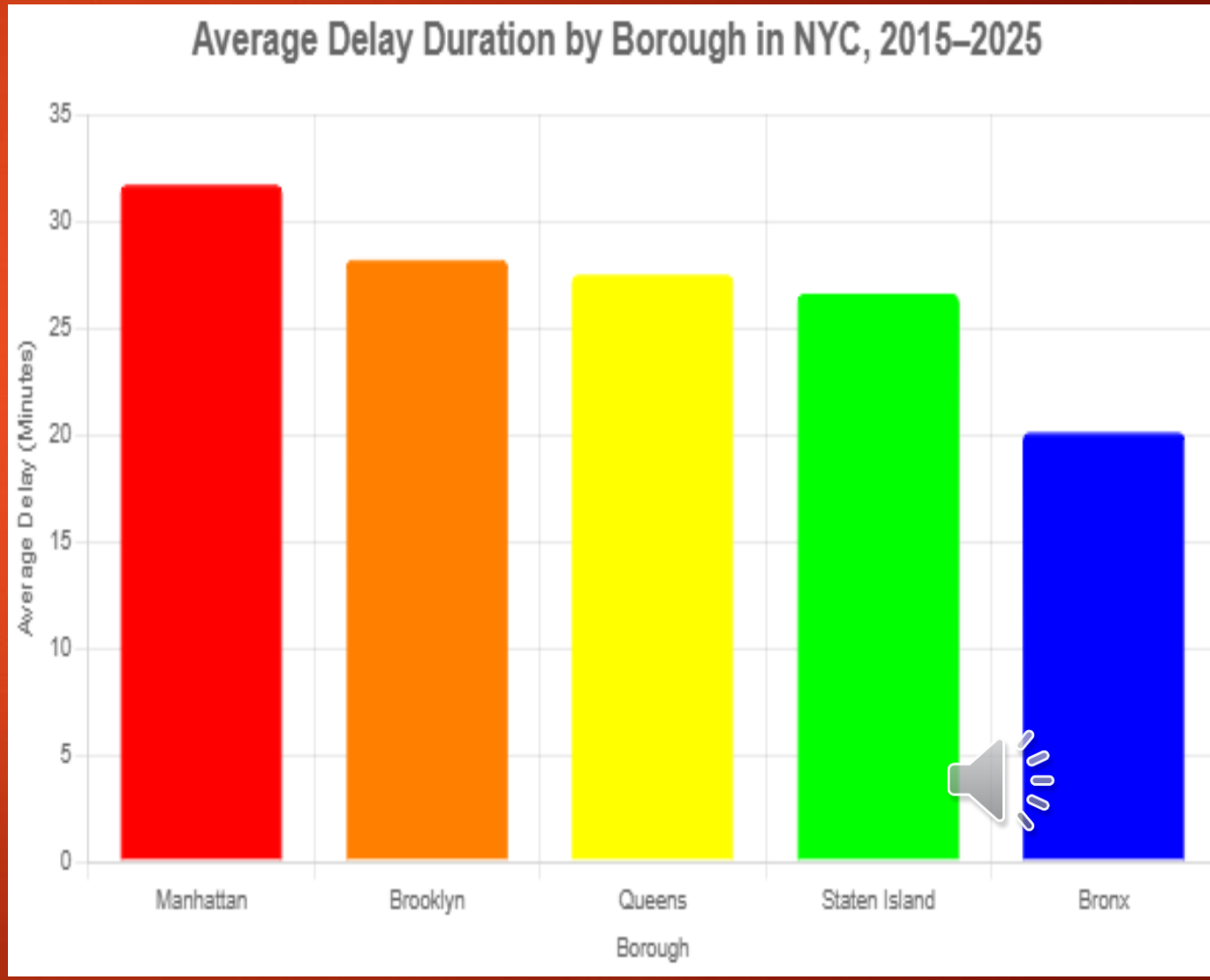
Research Question 2: Does the Day of the Week Impact Delay Durations?

- Monday delays peak: Highest average, likely due to weekend resets or traffic.
- Midweek (Tue–Thu) stable: Slightly lower, more consistent delays.
- Friday uptick: Minor increase, possibly from congestion or fatigue.
- This pattern suggests that day-of-week scheduling may affect operational efficiency.



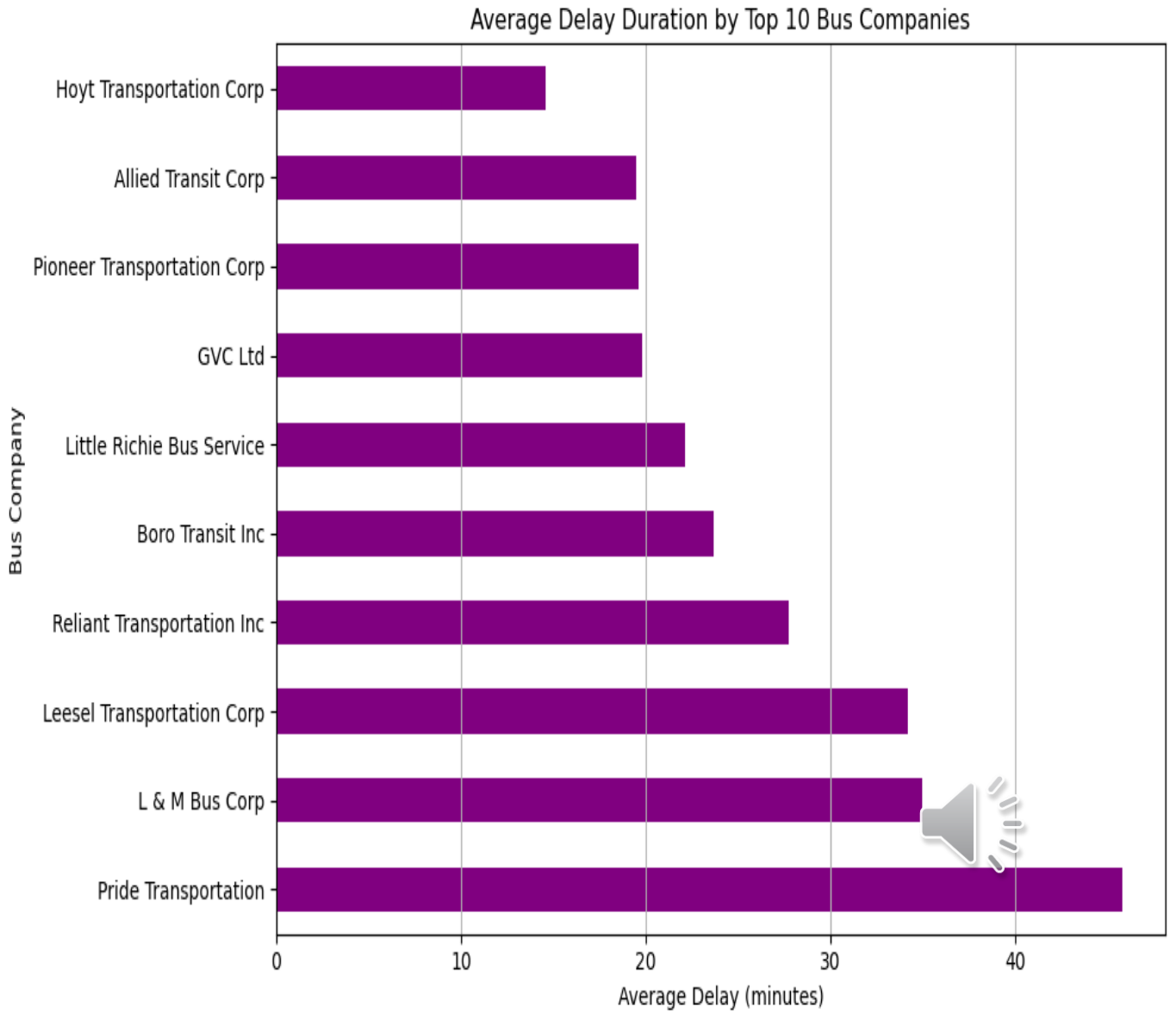
Research Question 3: Do Delays Differ by Borough?

- Manhattan: ~31.7 minutes delay.
- Brooklyn: ~28.2 minutes
- Queens: ~27.5 minutes
- Staten Island: ~26.6 minutes
- Bronx: ~20.1 minutes
- The 11.6-minute gap between Manhattan and Bronx suggests targeting Manhattan traffic interventions.



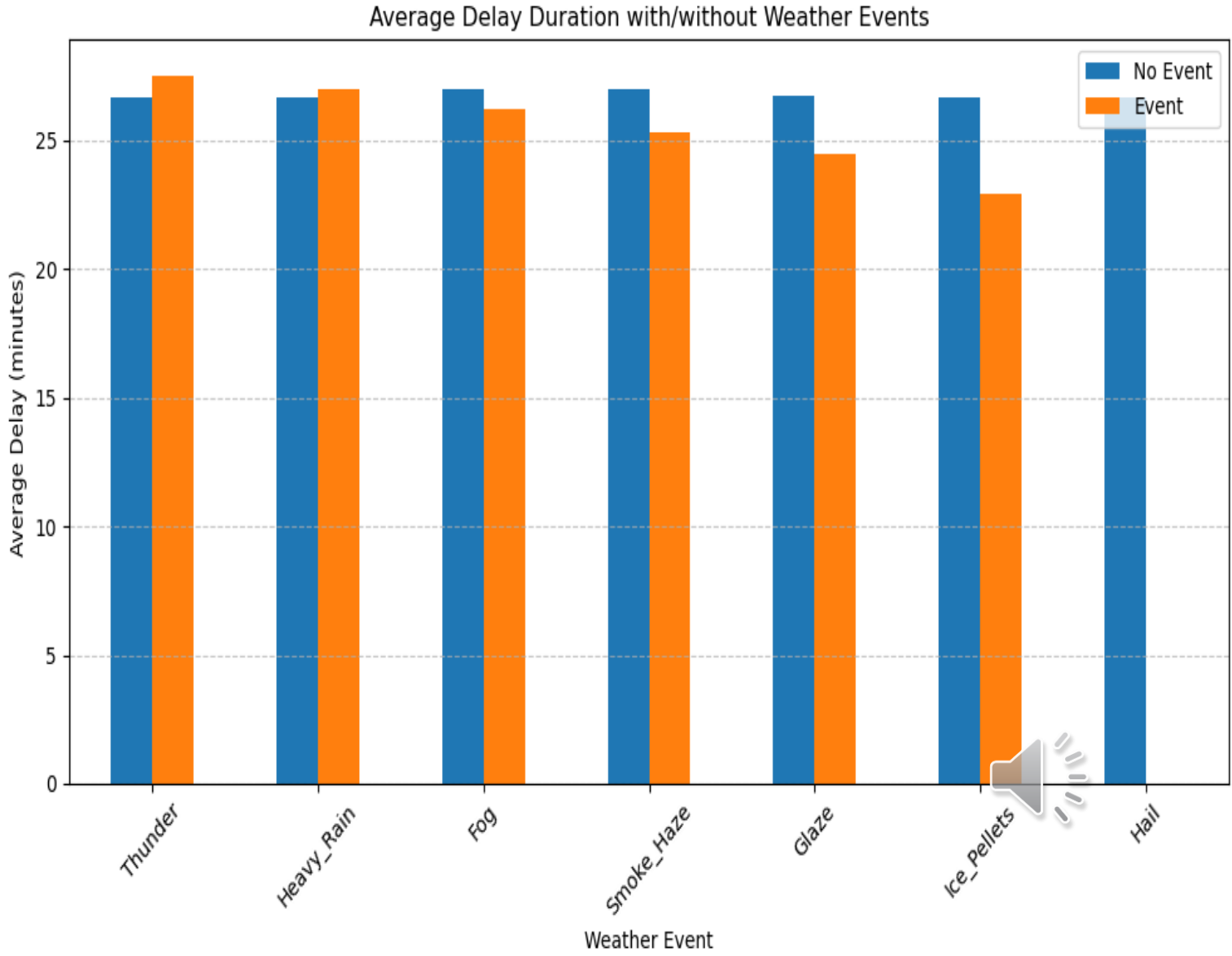
Research Question 3: Do Delays Differ by Bus Company?

- Pride Transportation: Highest delay, over 45 minutes.
- Hoyt Transportation: Lowest delay, ~14 minutes
- Wide performance gap: Over 30-minute spread, hinting at inefficiencies.
- Use data for contract oversight and targeted improvements.



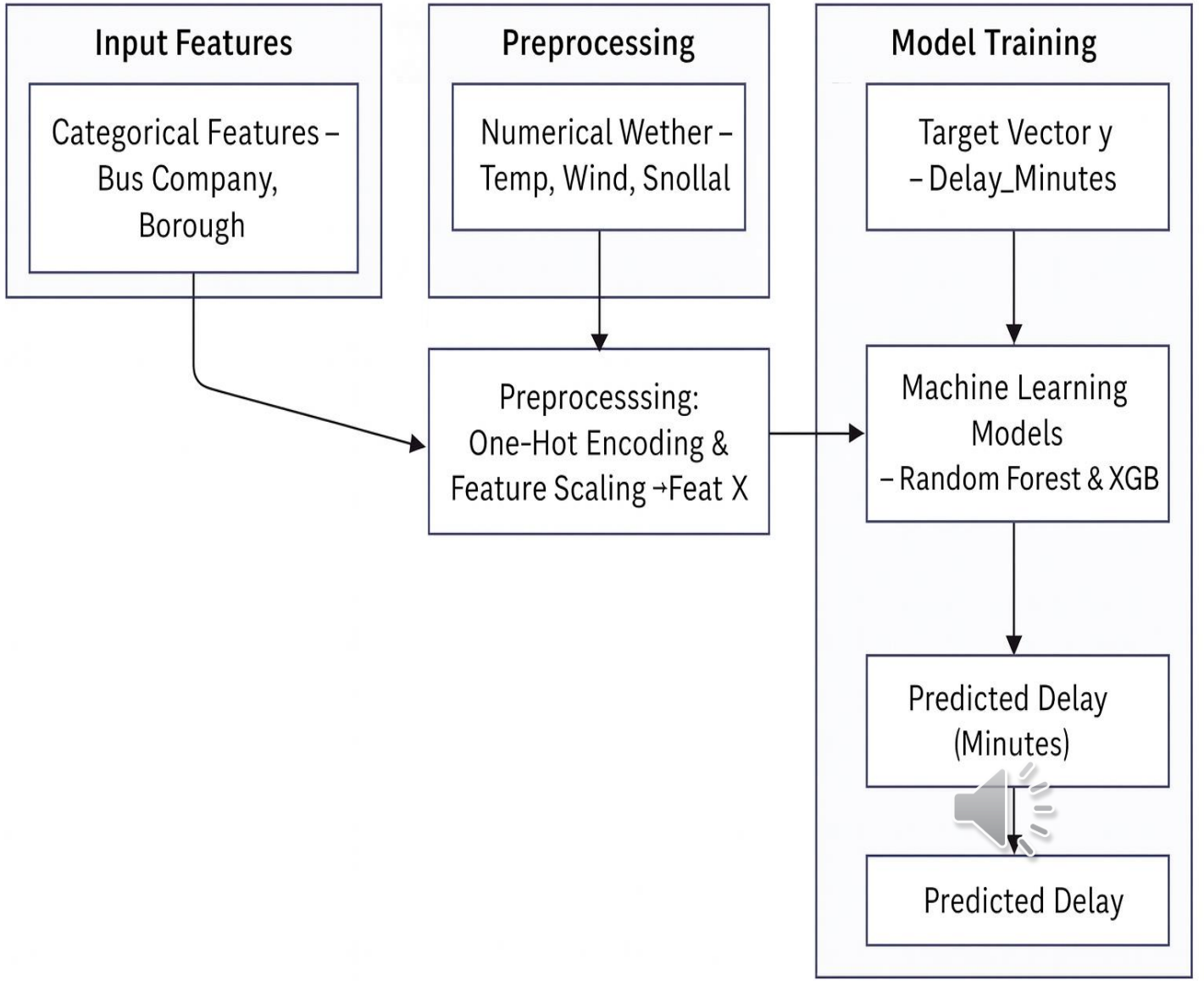
Research Question 4: Impact of Weather Events on Bus Delays

- Thunderstorms: Highest delay, dangerous driving conditions.
- Heavy Rain & Fog: Moderate delays, reduced visibility, and traffic.
- Ice Pellets, Glaze, Smoke/Haze, low or no delay, rerouting, or closures.
- Hail: no impact, rare in NYC.



Research Question 5: Can Delays or Breakdowns Be Forecasted Using Predictive Models?

- Data Processing: Cleaned 700,000+ records from transit and weather datasets for accuracy.
- Class Removal: Excluded delay classes with <2 entries (<1% of data) for robust training.
- Data Split: Used stratified sampling for an 80/20 train-test split
- Feature Prep: Applied one-hot encoding to categorical data and normalized numerical variables.
- Model Building: Developed ML models to predict delays, achieving RMSE ~17.6 minutes.



Model Selection and Comparison – Random Forest vs. XGBoost

- Tree-based models capture complex patterns & handle missing values well
- Built-in feature importance supports interpretability
- Ideal for large, real-world datasets
- Chosen for performance and interpretability over linear or SVM models

Model	RMSE	R2
Random Forest	17.62 mins	0.26
XGBoost	17.58 mins	0.26

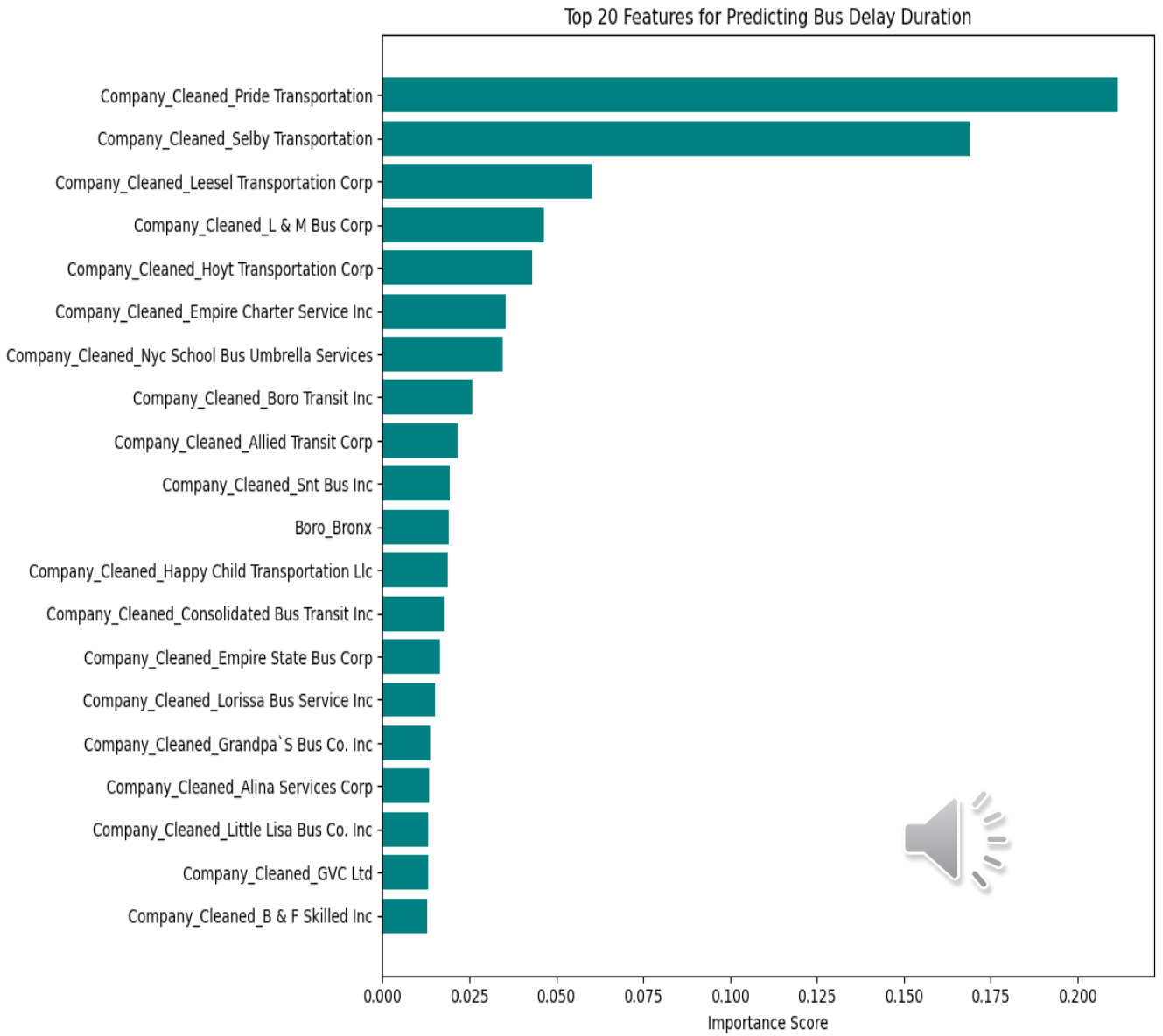
Interpretation:

- Both models perform similarly, suggesting diminishing returns with added model complexity.
- RMSE of ~17.6 minutes means the average prediction error is about 17 minutes.
- R^2 of 0.26 indicates modest predictive power; external factors likely influence delays.

What Drives Bus Delays? Top Predictive Features

Key Takeaways from Feature Importance Analysis(based on XGBoost model)

- Company identity key: Pride and Selby show unique delay patterns.
- Bus companies dominate: 8 of the top 10 predictors, highlighting internal impact.
- Bronx challenges: Top 20 ranking reveals location issues.
- Model success: Confirms learning of operational and regional patterns.
- Target underperforming companies for improvements.



Recommendation: Actionable Insights & Recommendations

- Target high-delay bus companies, e.g., Pride & Selby Transportation, to prioritize audits, retraining, or contract review.
- Address borough-specific issues. Manhattan exhibits longer delays when exploring traffic coordination or rerouting options.
- Focus on Monday operations. Delay durations peak on Mondays, which suggests a need for improved weekend-to-weekday transition planning.
- Weather has a modest influence. Thunder slightly increases delays; most events are not affected significantly by weather, so weather-resilient planning may not require a significant overhaul.
- Leverage ML predictions for daily operations. Use delay risk scoring by company, day, and weather to proactively manage dispatching and communication.

Conclusion & Key Takeaways

- Analyzed over 700,000 NYC school bus incidents. Cleaned, merged, and enriched with weather data to uncover delay patterns.
- Identified major causes of delays: "Heavy Traffic" is by far the most common, accounting for over 60% of all reports.
- Delay duration varies by weekday, borough, and bus company. Manhattan and Mondays experienced the most extended average delays.
- Built and compared ML models (Random Forest & XGBoost). XGBoost performed slightly better with an RMSE of 17.58 minutes and R^2 of 0.26.
- Top predictive features revealed company-level trends. Some bus companies are consistently linked to longer delays, offering actionable focus areas.
- Project demonstrates how data science can drive more innovative transportation planning, from auditing vendors to improving scheduling and communication with parents.

Tools and Libraries used

- Numpy: Python fundamental package for scientific computing
- Pandas : Python data wrangling library
- Scikit-Learn : For Machine Learning and Vectorization
- Seaborn : Statistical visualization too built on top of matplotlib.
- Matplotlib : Foundational library for visualizations.
- Jupyter Notebook : Interactive Python Programming server-client interface.

Thank You & Questions

Thank You! I appreciate your time and attention.

Questions? I am happy to answer any questions or discuss insights further.