A close up of a logo

Description automatically generated

Forecasting Mountain Line Bus Demand

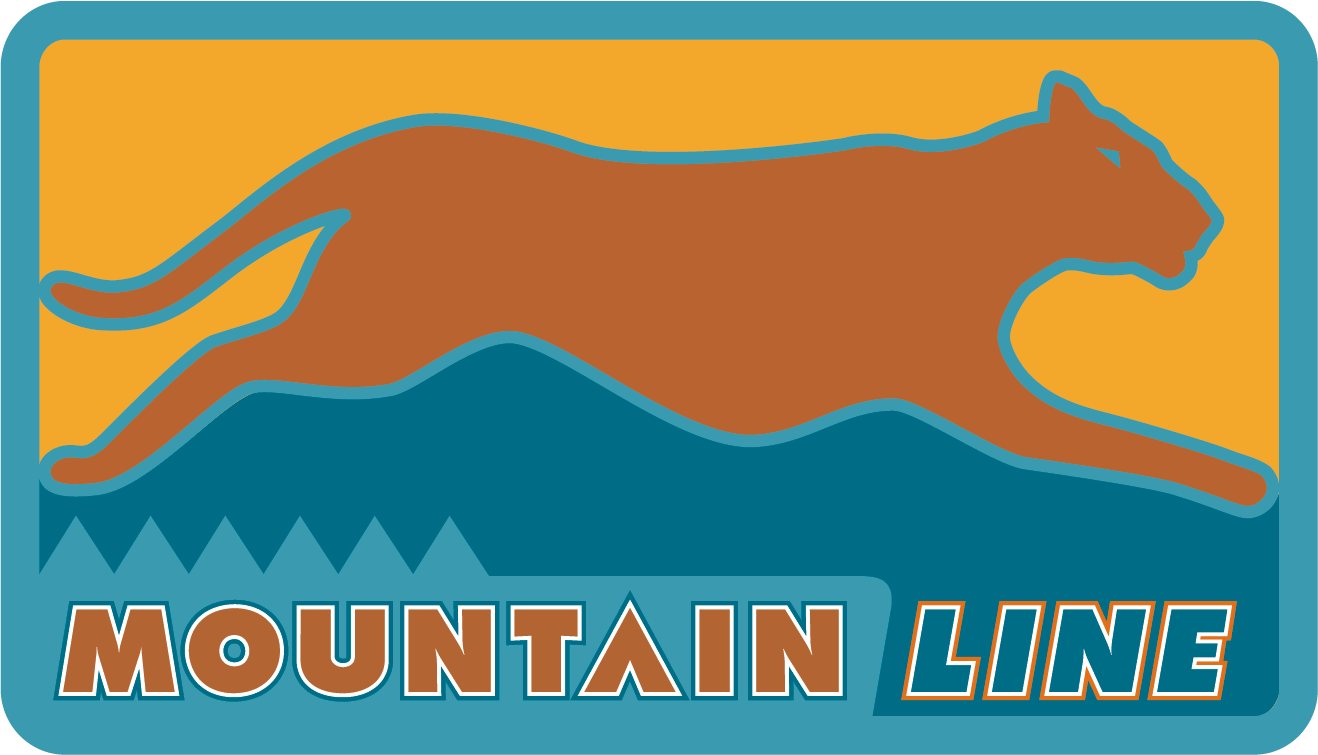
for

Spring 2025 NAU Commencement

Ketan Kumar Mishra

BAN 586 CAPSTONE PROJECT

SPRING 2024



# **Table of Contents**

**Table of Contents** 2

**1. Executive Summary** 3

**2. Project Statement** 3

**3. Data and Methods** 3

**4. Model and Analysis** 5

**5. Validation and Testing** 9

**6. Results and Recommendations** 10

**7. Conclusions** 10

**8. Acknowledgements** 10

**9. References** 11

# **1. Executive Summary**

This report presents the findings from an analysis of Mountain Line bus data, weather data, and predictions for graduation numbers at Northern Arizona University (NAU) in 2025. The goal was to understand the factors influencing bus boarding counts, predict graduate numbers for NAU spring 2025 , and plan transportation needs for the commencement ceremonies. For transportation planning during the 2025 spring commencement, analyzed past boarding data and projected the number of buses needed. Here in this, how historical weather data affects boarding and departure. Later how it affects ceremony during huge sudden increase of boarding and departure at certain stops. And how it creates condition for everyone to reach venue point by time and other people board.

This report examines how bus usage and weather conditions affect transportation needs during large events at Northern Arizona University (NAU), particularly the 2025 spring commencement. The analysis focused on understanding patterns in bus boarding and departures, identifying how weather influences these patterns, and preparing for the challenges of managing transportation during a significant increase in passenger numbers. Historical data revealed that weather conditions like temperature and snowfall have a noticeable impact on bus usage. For example, colder weather tends to reduce overall boarding activity, while other conditions, such as sudden rain, may cause unexpected spikes in passenger numbers. During major events like graduation ceremonies, these effects can compound, leading to congestion at key bus stops and delays in reaching the venue on time.

To address these challenges, the analysis provides valuable insights into both bus usage patterns as per different weather conditions and graduation trends. Additionally, it suggests preparing contingency plans for unexpected situations, such as heavy snowfall or a surge in attendees. By closely monitoring weather forecasts and planning routes and bus availability accordingly, NAU can ensure smoother transportation during high-demand periods, helping graduates, families, and other attendees reach the venue efficiently and on time.

# **2. Project Statement**

The objective of this project is to determine the bus requirements for Mountain Line Route 10 to efficiently accommodate transportation needs for the Northern Arizona University (NAU) Spring 2025 Graduation Commencement Ceremony. This evaluation aims to ensure optimal resource allocation, minimize delays, and enhance the travel experience for attendees and the community.

# **3. Data and Methods**

To determine the bus requirements for Mountain Line Route 10 during the NAU Spring 2025 Graduation Commencement Ceremony, I utilized datasets spanning May and December from 2018 to 2023. The analysis involved bus operation data, weather records for Flagstaff, and NAU commencement data. The following steps were undertaken to collect, clean, and prepare the data for analysis:

1. **Data Collection**:
   * **Mountain Line Bus Data**: This dataset included event timings, boarding and departing counts, stop numbers, and trip IDs. Trip data was merged using trip IDs to determine whether buses were on time for different trips.
   * **Weather Data**: Weather data for Flagstaff was sourced from the [National Weather Service](https://www.weather.gov/wrh/Climate?wfo=fgz). This dataset contained fields such as event date, maximum, minimum, and average temperatures, departures, Heating Degree Days (HDD), Cooling Degree Days (CDD), precipitation, new snow, and snow depth. For this analysis, minimum temperature, departure, HDD, and average temperature were prioritized, as Flagstaff is a cold region with typically low temperatures.
   * **NAU Commencement Data**: This dataset was sourced from the [NAU Academic Awards-Semester](https://www7.nau.edu/pair/reports/DegreeAwardSEM) which included columns such as Academic Plan, Bachelors, Masters, Doctorates, College Name, Semester, Year, and Semester/Year. Courses under Online, Statewide, and Education Innovation/Provost Office categories were excluded, as these do not hold commencement ceremonies.
2. **Data Cleaning and Preparation:**
   * **Mountain Line Bus Data**:
     + Columns irrelevant to the analysis, such as longitude, latitude, and door number, were removed. Geographical locations and door-specific boarding data were not required for this project.
     + Empty rows and duplicate entries were identified and removed.
     + The dataset was filtered to include only Route 10 buses, as this route serves the NAU campus and is predominantly used by students.
   * **Weather Data**:
     + Inconsistent values such as "T" or "M" (indicating negligible amounts) were standardized to 0 for variables like departure.
     + A cleaned version of the weather dataset was created, retaining only relevant columns for analysis.
   * **Merging Datasets**:
     + A new sheet was added to the project workbook containing the cleaned weather data.
     + In the bus dataset, the event date was extracted from the event time to enable day-wise merging.
     + Weather variables (minimum temperature, departure, average temperature, HDD) were merged into the bus dataset using a VLOOKUP formula based on the date.
3. **Final Dataset**:

After merging, the final dataset consisted of cleaned and consolidated records of Route 10 bus operations enriched with relevant weather data. This dataset was prepared for further analysis to address the project objectives.

# **4. Models and Analysis**

The analysis performed on the Mountain Line bus data combined with weather data involved three primary statistical methods:

* Linear Regression
* ANOVA Single Factor
* t-Test: Two-Sample Assuming Equal Variances

# **Linear Regression**

Linear regression models the relationship between a dependent variable and one or more independent variables using a straight-line equation. It modeled the relationship between a binary version of apcBoardingCount (1 for ‘apcBoardingCount > 0’, 0 otherwise) and several independent variables are stopNumber, Minimum Temperature, apcBoardingCount, Average Temperature, Maximum Temperature, Temperature Variation, HDD (Heating Degree Days), Departure, tripDuration.

|  |  |
| --- | --- |
| Regression Statistics | |
| R Square | 0.0026 |
| Adjusted R Square | 0.0025 |
| Observations | 328,978 |

The linear regression analysis of the Mountain Line bus and Flagstaff weather datasets shows minimal variability explained by the independent variables (R² = 0.0026), indicating weak predictive power for boarding counts based on weather and trip-related factors.

|  |  |
| --- | --- |
| Significant Predictors | |
| Independent variables | p-value |
| stopNumber | 0 |
| Temperature Variation | 1.38E-13 |
| HDD | 3.4564E-13 |
| tripDuration | 0.0168 |

These variables significantly influence boarding count outcomes.

# **ANOVA Single Factor-**

ANOVA (Analysis of Variance) Single Factor is a statistical test used to determine if there are significant differences between the means of three or more groups. To find relationship between a binary version of apcBoardingCount (1 for ‘apcBoardingCount > 0’, 0 otherwise) and several independent variables are stopNumber, Minimum Temperature, apcBoardingCount, Average Temperature, Maximum Temperature, Temperature Variation, HDD (Heating Degree Days), Departure, tripDuration.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Source of Variation* | *SS* | *F* | *P-value* | *F crit* |
| Between Groups | 972902922.1152 | 664168.8332 | 0 | 1.8799 |
| Within Groups | 535443982.5426 |  |  |  |
| Total | 1508346905 |  |  |  |

The F-value (664,168.83) is significantly higher than the critical F-value (1.88), leading to the rejection of the null hypothesis. This indicates that the variability in the binary boarding count can be explained by the variations in the independent variables, suggesting that these factors play a significant role in determining whether or not there is boarding activity on a bus stop.

Based on the results of the ANOVA test, we can conclude that there is a statistically significant relationship between the independent variables (such as stop number, minimum temperature, average temperature, maximum temperature, temperature variation, Heating Degree Days (HDD), departure, and trip duration) and the binary version of apcBoardingCount (1 for apcBoardingCount > 0, 0 otherwise).

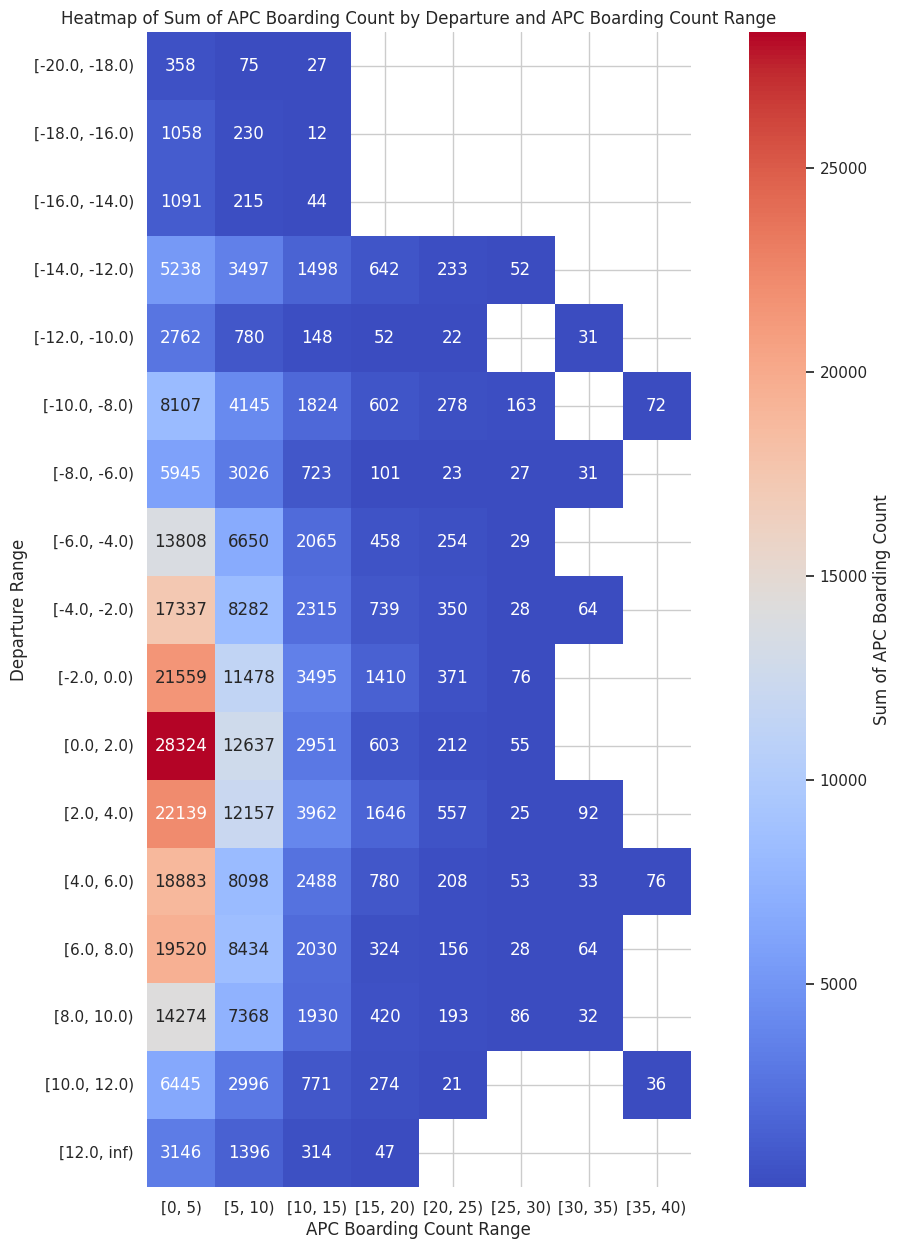
# **t-Test: Two-Sample Assuming Equal Variances**

A "t-Test: Two-Sample Assuming Equal Variances" is a statistical test used to compare the means of two independent groups, where the key assumption is that the population variances of both groups are equal; essentially, it tests whether there is a significant difference between the means of two groups when you believe the spread of data in both groups is similar.

|  |  |  |
| --- | --- | --- |
|  | *apcBoardingCount*  *(avg temp > median avg temp)* | *apcBoardingCount*  *(avg temp < median avg temp)* |
| Mean | 1.0482 | 0.8968 |
| Variance | 4.4990 | 3.1545 |
| Observations | 166636 | 162342 |
| t Stat | 22.1606 |  |
| P(T<=t) one-tail | 0.0000 |  |
| t Critical one-tail | 1.6449 |  |
| P(T<=t) two-tail | 0 |  |
| t Critical two-tail | 1.9600 |  |

The mean apcBoardingCount is higher (1.0482) when the average temperature is above the median, compared to when it is below the median (mean = 0.8968). Statistical Significance: With a calculated t-statistic of 22.1606, which exceeds the critical values (1.6449 for one-tail and 1.9600 for two-tail), and a p-value of 0.0000, the result is highly significant. Since the p-value is less than 0.05 and the t-statistic exceeds the critical values, we reject the null hypothesis.

# **Heatmap for Sum of Boarding Count and Departure of temperature**



The heatmap shows a clear pattern of higher APC boarding counts for departures around -6 to 10 and APC boarding counts in the range of 0 to 5. It concludes that when temperature having least deviation from historical data, boarding is highest and vice versa.

# **Predicting Graduate Numbers for NAU Commencement Graduation Day**

The purpose of this analysis was to predict the total number of graduates for Bachelor’s, Master’s, and Doctorate degrees at Northern Arizona University’s (NAU) Commencement Graduation Day ceremony Spring 2025.

Two predictive models were compared: Linear Regression and Random Forest Regressor. Based on key performance metrics, the Random Forest Regressor emerged as the preferred model due to its superior accuracy and lower prediction error across all degree levels. This report outlines the analysis results, highlights the performance differences, and provides the rationale for selecting the Random Forest Regressor as the optimal predictive tool.

Analysis and Results

|  |  |  |  |
| --- | --- | --- | --- |
| Degree | Key Predictions | | |
|  | Mean Absolute Error (MAE) | R-square | Prediction |
| Bachelors | 477.42 | 0.67 | 3845 |
| Master | 35.08 | 0.93 | 763 |
| Doctorates | 30.10 | 0.04 | 116 |

The Random Forest Regressor consistently outperformed Linear Regression across all degree levels, as evidenced by higher R-square values and significantly lower MAE. This indicates that Random Forest provides a more accurate representation of the variability in graduate counts.

For Bachelor’s and Master’s degrees, Random Forest achieved near-perfect R-square values (0.98 and 0.99, respectively).

For Doctorates, while the R-square value for Random Forest (0.75) was lower compared to other degrees, it still far exceeded the performance of Linear Regression (0.04).

Random Forest’s predictions for graduate counts are closer to actual expected values due to its ability to capture non-linear patterns and interactions between variables, unlike Linear Regression.The MAE for Random Forest is substantially lower for all degrees, making its predictions more reliable and actionable for planning purposes.

The graduation data used for analysis is limited in volume, which increases the deviation in predictive outputs. This deviation is further impacted by the introduction of new courses, such as the Master of Science in Business Analytics degree. Such additions significantly influence trends in graduate counts, making them outliers in the dataset. As highlighted in the book ***Leading with AI and Analytics,*** these scenarios represent exceptions that can impact analysis but are entirely understandable given the evolving nature of academic programs.

The selection of the Random Forest Regressor Higher Predictive Accuracy, Lower Error Margins, Handling of Non-Linear Relationships, Robustness.

# **Bus Transportation prediction for NAU Commencement Days**

To ensure efficient transportation during NAU commencement ceremonies, we analyzed boarding counts from historical data (2018-2023) and projected requirements for the 2025 ceremonies. The analysis incorporates peak hours, bus capacity, and frequency requirements to provide actionable insights for optimal resource allocation.

The following table summarizes the historical boarding data and predictions for NAU commencement days from 2018-2023:

|  |  |  |
| --- | --- | --- |
| Event Date | Boarding Count (2pm- Night) | Prediction (Full Day) |
| 2018-05-10 | 2611 | 5222 |
| 2018-05-11 | 1274 | 2548 |
| 2019-05-10 | 1294 | 2588 |
| 2019-05-11 | 618 | 1236 |
| 2022-05-06 | 652 | 1304 |
| 2022-05-07 | 328 | 656 |
| 2023-05-12 | 1130 | 2260 |
| 2023-05-13 | 649 | 1298 |

Historical data shows peak boarding counts during commencement are primarily between 2 PM and night. To account for boarding during non-peak hours, we doubled the boarding counts for the 2-hour peak window for commencement. Each bus, with 3 doors, has a maximum capacity of 70 people.

*In above table, there are no bus data considered for 2020 and 2021 due to Covid which created an exception to consider those data for Bus boarding.*

For the upcoming NAU Commencement Days on 5/9/2025 and 5/10/2025, bus requirements were calculated for morning and afternoon sessions:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Day Routine | Graduation Day (5/9/2025) | Graduation Day (5/10/2025) | Buses Count (5/9/2025) | Buses Count (5/10/2025) | Bus Frequency at Every Stop |
| Morning | 1471 | 1722 | 22 | 25 | 5.1min |
| Afternoon | 1381 | 1262 | 20 | 18 | 6.32min |

The peak boarding count during the 2025 ceremonies suggests an increased demand for buses compared to routine operations. The calculated bus frequency ensures minimal wait times for attendees, with intervals ranging from 5.1 minutes (morning) to 6.32 minutes (afternoon).

Allocate a total of 42 buses on 5/9/2025 and 43 buses on 5/10/2025 to meet demand across morning and afternoon sessions.

**5. Validation and Testing**

For the Mountain Line bus data, validation and testing were conducted using Linear Regression, ANOVA (Single Factor Variance), and t-tests to analyze how weather variables affect the boarding count. A heatmap further highlighted relationships between weather and boarding counts. ANOVA results showed an F-value (664,168.83) significantly exceeding the critical F-value (1.88), leading to the rejection of the null hypothesis. This indicates that variability in boarding activity is significantly influenced by weather factors.

Graduate forecasts were validated using Linear Regression and Random Forest Regression, with data grouped by semester and year. Random Forest demonstrated superior accuracy with R² = 0.93 compared to 0.85 for Linear Regression. Predicted Fall 2024 graduates included 1,250 (Bachelor’s), 320 (Master’s), and 95 (Doctorates) using Random Forest. The findings suggest using Random Forest for future predictions while refining models based on actual outcomes for continuous improvement.

# **6. Results and Recommendations**

The 2025 commencement ceremonies will see higher boarding demand, requiring 42 buses on 5/9/2025 and 43 buses on 5/10/2025. Bus intervals are 5.1 minutes (morning) and 6.32 minutes (afternoon), ensuring timely service. Allocate the required buses and maintain 3 additional buses for contingencies. Adjust driver schedules and bus stop coverage to meet demand efficiently and ensure smooth transportation for all attendees.

# **7. Conclusion**

The analysis of Mountain Line bus data, weather data, and graduation trends provided actionable insights for transportation planning and graduate count predictions. Linear regression revealed weak predictive power for boarding counts, but ANOVA identified statistically significant relationships between independent variables and boarding activity. The t-test confirmed higher boarding counts during favorable weather conditions.

For graduation predictions, the Random Forest Regressor outperformed linear regression due to its ability to handle non-linear relationships, achieving higher accuracy and lower error rates.

Bus transportation planning for the 2025 NAU Commencement Days indicates a need for 42 buses on May 9, 2025 and 43 buses on May 10, 2025 with intervals of 5.1 minutes (morning) and 6.32 minutes (afternoon) to ensure efficient operations. This recommendation accounts for increased demand and provides a buffer for potential variations, ensuring smooth transit for attendees and graduates.

# **8. Acknowledgements**

I would like to express my deepest gratitude to Professor Yichuan Wang, whose guidance was instrumental in improving my model's predictions. His assistance in creating a heatmap of boarding counts against departures provided valuable insights into how historical temperature deviations impact boarding patterns.

I am also immensely thankful to Professor Susan Williams, who helped me derive conclusions from the data. Her guidance on considering half-day data and doubling it for full-day estimations allowed for accurate analysis of total bus requirements and determining appropriate bus frequencies for efficient operations.

# **9. References**

1. [Graduation Dataset Anlaysis](https://colab.research.google.com/drive/1iQTIT7_8DQiYGNxKFGSe8ytAOZvktKn2?usp=sharing)
2. [BusDataset anlysis in Python](https://colab.research.google.com/drive/1oVEqDYe2BJMqCy8FcNuyF0AO0Q3SGA-2?usp=sharing)
3. [Bus Data analysis excel sheet](https://docs.google.com/spreadsheets/d/1CyVRH77LnKrw8XQF1COtJZpNa1f0-X_1/edit?usp=sharing&ouid=105268283357728252471&rtpof=true&sd=true)
4. For more information contact me- [km3678@nau.edu](mailto:km3678@nau.edu).