The Commute Chronicles

An Analysis of Travel Efficiency, Weather Impacts, and Crowding

Opeoluwa Daniel Oyedeji

December 5, 2025

Contents

1 Introduction

2 Methods and Results

2.1 Exploratory Data Analysis

2.1.1 Dataset Overview

```r
df <- read.csv("bus_trip_data.csv")
df$date <- as.Date(df$date) # Convert date column to R Date type
```

The data has 10 columns and 28 rows. To ensure the table fits within the document margins, the data set is displayed below in two parts: the "To Campus" details and the "To Home" details.

date	day_of_week	dep_to_campus	arr_to_campus	dur_to_campus
2025-09-29	Monday	7:00:00 AM	9:18:00 AM	138

dep_to_home	arr_to_home	dur_to_home	rain_to_home	crowded_unilink_to_home
3:08:00 PM	5:30:00 PM	142	FALSE	TRUE

2.1.2   Visual Inspection of Trends

Before running any complex statistical tests, we visualize the raw data to understand its shape. We are looking for two things: the **distribution** of the data (e.g. is it bell-shaped?) and any **trends** over time (e.g. is traffic getting worse?).
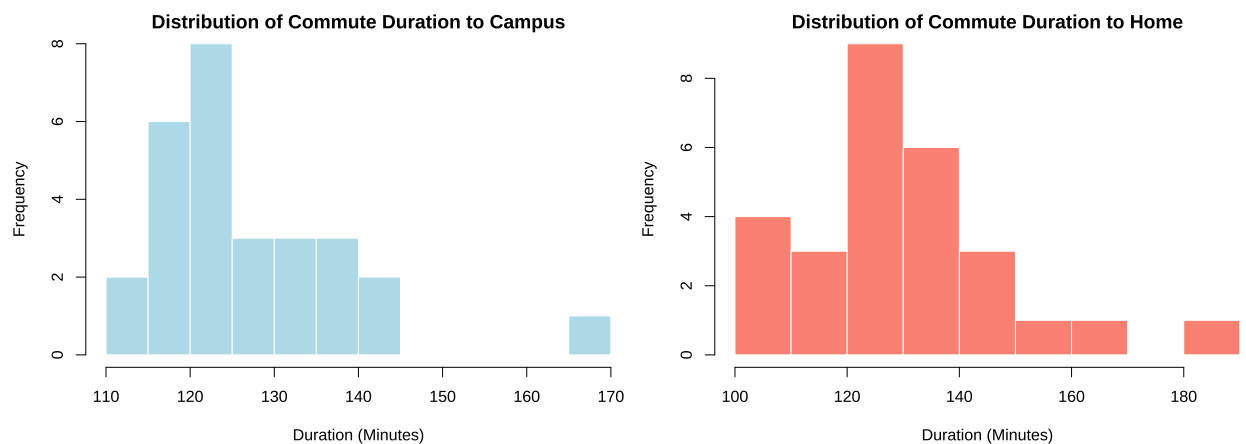

Figure 1: Histograms displaying the distribution of travel times for trips to campus (left) and trips to home (right) with 10-minute bin intervals.

An examination of the distributional shape (Figure 1) highlights the following patterns:

- Both histograms exhibit a *positive skew* and a detached bar at the far right, indicating a *potential outlier* that may need to be addressed in our statistical tests.

- Both histograms are *unimodal* with the mode being around the *120-130 minute interval*.
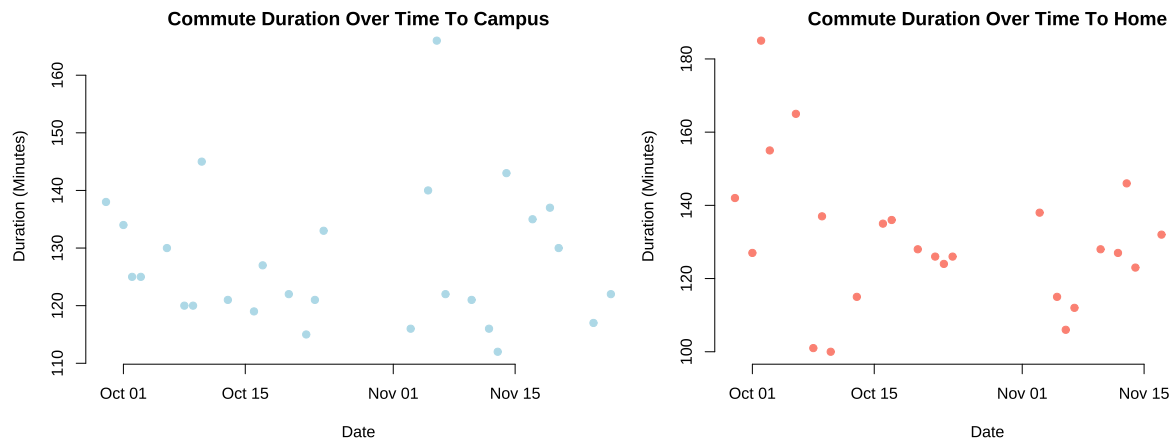
Figure 2: Scatter plots tracking daily commute durations for trips to campus (left) and return trips to home (right) over the observed period.

Plotting the commute durations over time (Figure 2) highlights a stark contrast in stability between the two journeys:

- Trips to campus appear stable over time, showing only random noise without a noticeable increase or decrease in duration.

- Trips to home exhibit high variance, with widely dispersed data points, suggesting unpredictable delays that warrant further statistical testing.
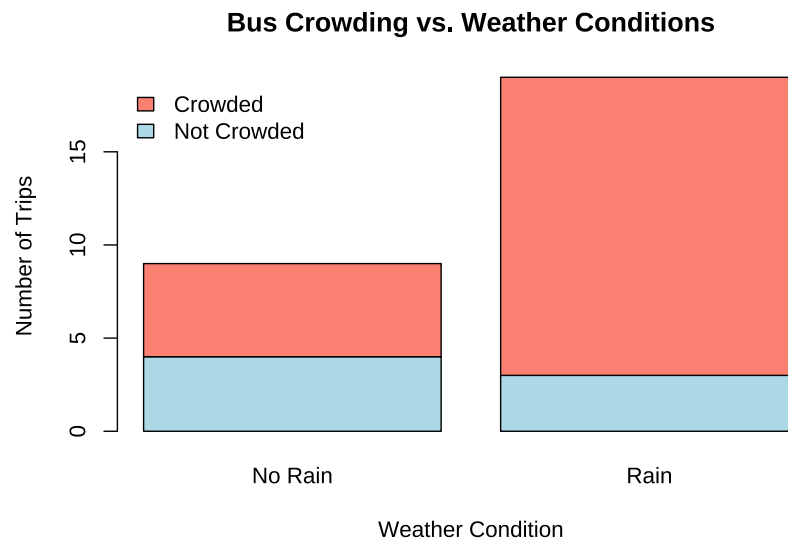

Figure 3: Stacked bar chart showing the proportion of crowded UniLink buses to home during clear weather versus rain.

An analysis of the interaction between weather and bus capacity (Figure 3) suggests a strong correlation:

3

- **On clear days**, the distribution is balanced, with the bus being crowded roughly half the time.

- **On rainy days**, the vast majority of trips experience overcrowding, likely due to an increase in students choosing public transport over walking or cycling.

2.2 Hypothesis Testing and Methodology

While the exploratory analysis highlighted potential trends, such as the volatility of the evening commute and the correlation between rain and crowding, visual inspection alone is insufficient to draw firm conclusions. In this section, we define the specific hypotheses derived from our observations and outline the statistical methods selected to test them.

2.2.1 Comparing Commute Directions (Paired t-test)

The scatter plot (Figure 2) suggested that the journey home is not only more volatile but potentially longer on average than the morning commute. To confirm this, we analyse the difference in means between the two trips.

Method Justification: Since each day (or record) consists of a trip to campus and a trip to home, a paired student t-test is the appropriate method to control for day-specific variance (e.g., a generally traffic-heavy day affecting both trips).

- H_0 (Null): There is no significant difference in mean duration between the trip to campus and the trip home ($\mu_{campus} = \mu_{home}$).

- H_a (Alternative): The trip home takes significantly longer than the trip to campus ($\mu_{home} > \mu_{campus}$).

2.2.2 Predictability of Arrival Times (Linear Regression)

We aim to understand the relationship between departure time and arrival time. Specifically, we want to assess if a linear model can accurately predict arrival times, or if traffic delays introduce non-linear variance.

Method Justification: We use *Simple Linear Regression* to quantify the relationship between the independent variable (Departure) and the dependent variable (Arrival). Rather than a hypothesis test, the goal here is to assess the goodness of fit (R^2) to see how much of the variance in arrival time can be explained purely by the time of departure.

2.2.3 The "Rain Effect" on Crowding (Fisher's Exact Test)

The bar chart (Figure 3) indicated a strong dependency between rainy weather and bus overcrowding. We aim to calculate the statistical significance of this association.

Method Justification: We are analyzing two categorical, binary variables (`rain_to_home` vs. `crowded_unilink_to_home`). This structure results in a 2×2 contingency table which Fisher's Exact Test is well suited for.

- H_0 (Null): Weather conditions and bus crowding are independent.

- H_a (Alternative): Rainy weather significantly increases the likelihood of the bus being crowded.

2.2.4 Weekday Crowding Variance (Pearson's Chi-Square)

We investigate the specific phenomenon of the *"Thursday Rush."* Personal experience suggests that Thursdays are disproportionately crowded compared to other weekdays. This is probably because on Thursdays, I depart campus around 18:00 (6 PM), coinciding with the general close of business for local offices, while on other days I leave earlier.

Method Justification: We examine the frequency of crowded buses across the distinct days of the week. Pearson's Chi-square Test is chosen to determine if the distribution of crowded trips differs significantly from what we would expect by chance. Unlike Fisher's Exact Test (which we used for a 2×2 comparison), Chi-square is ideal for this 2×4 matrix (Crowded/Not Crowded vs. Mon/Wed/Thu/Fri).

2.2.5 Conditional Probability of Events: Rain vs. Crowding (Bayes's Theorem)

While the previous tests determine if a relationship exists, we use Bayes's Theorem to quantify the dependency between the event **"It is Raining"** and the event **"The Bus is Crowded."** Specifically, we aim to measure how the presence of rain updates the probability of the bus being full.

Method Justification: This is not a hypothesis test for statistical significance, but a Probability Update. We will calculate the **Prior Probability** ($P(Crowded)$), the baseline chance of a crowded bus on any random day, and compare it to the **Posterior Probability** ($P(Crowded|Rain)$).

- If $P(Crowded|Rain) = P(Crowded)$, the events are **Independent** (Rain has no effect).

- If $P(Crowded|Rain) > P(Crowded)$, the events are **Dependent**, and rain acts as a positive risk factor.