The Commute Chronicles

An Analysis of Travel Efficiency, Weather Impacts, and Crowding

Opeoluwa Daniel Oyedeji

December 5, 2025

Contents

1 Introduction

For any university student, time is a limited and valuable resource. However, my geographical situation presents a significant challenge: I spend about four hours each day commuting. This adds up to twenty hours a week in transit (equivalent to a part-time job) and creates a considerable gap in my efficiency. Due to the discomfort of the journey, I cannot effectively use this time for studying, making my commute a dead zone in terms of productivity.

The motivation for this project is to reclaim agency over my travel time. Due to overcrowding, particularly on Thursdays, buses often bypass stops, leaving passengers stranded. I aim to shift from passive waiting to active management by analysing my travel logs to identify patterns, assess the impact of weather on crowding, and deduce actionable insights to optimise my journeys.

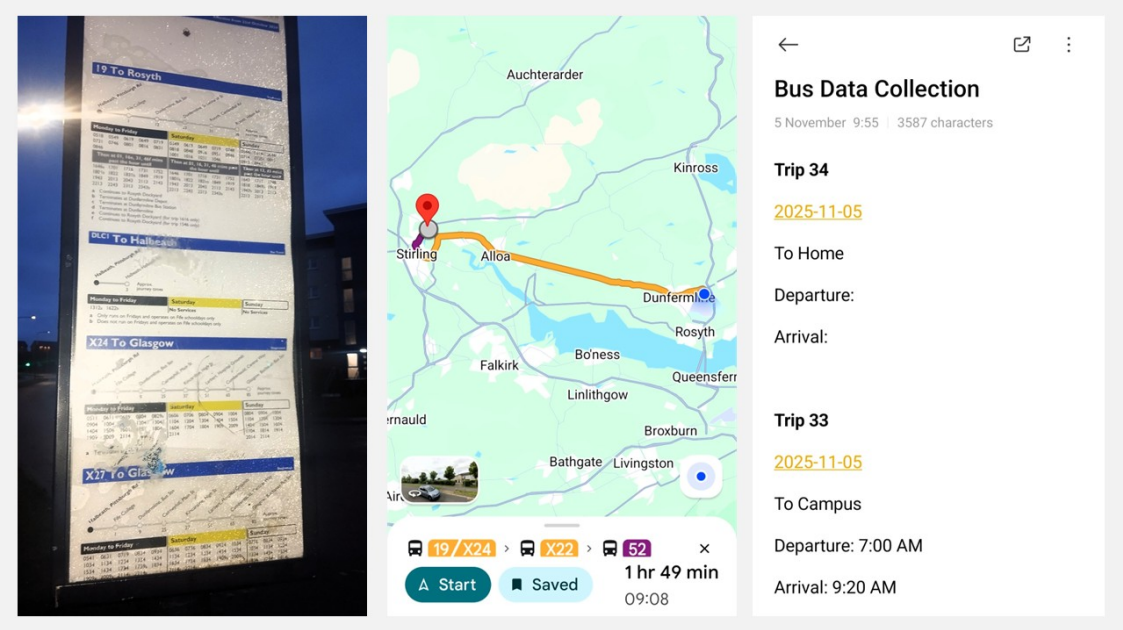

Figure 1: Data Collection Source. The collage shows the physical bus stop (left), the Google Maps route of the journey (centre), and the notepad app used to manually log daily trip details (right).

1.1 Dataset Description

The dataset is a self-collected log of my daily commute to and from campus. It records departure and arrival times, trip duration, and simple 'True/False' indicators for rain and bus crowding.

```
bus_data <- read.csv("bus_data.txt")
```

The data has 10 columns and 30 rows. A sample of the dataset 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
2025-10-01	Wednesday	7:00:00 AM	9:14:00 AM	134
2025-10-02	Thursday	11:40:00 AM	1:45:00 PM	125
2025-10-03	Friday	6:00:00 AM	8:05:00 AM	125
2025-10-06	Monday	7:05:00 AM	9:15:00 AM	130

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	TRUE	TRUE
12:00:00 PM	2:07:00 PM	127	TRUE	FALSE
6:18:00 PM	9:23:00 PM	185	TRUE	TRUE
11:15:00 AM	1:50:00 PM	155	FALSE	FALSE
3:15:00 PM	6:00:00 PM	165	FALSE	FALSE

2 Methods and Results

2.1 Exploratory Data Analysis

Before running any statistical tests, we visualize the data to understand its shape or spot any patterns.

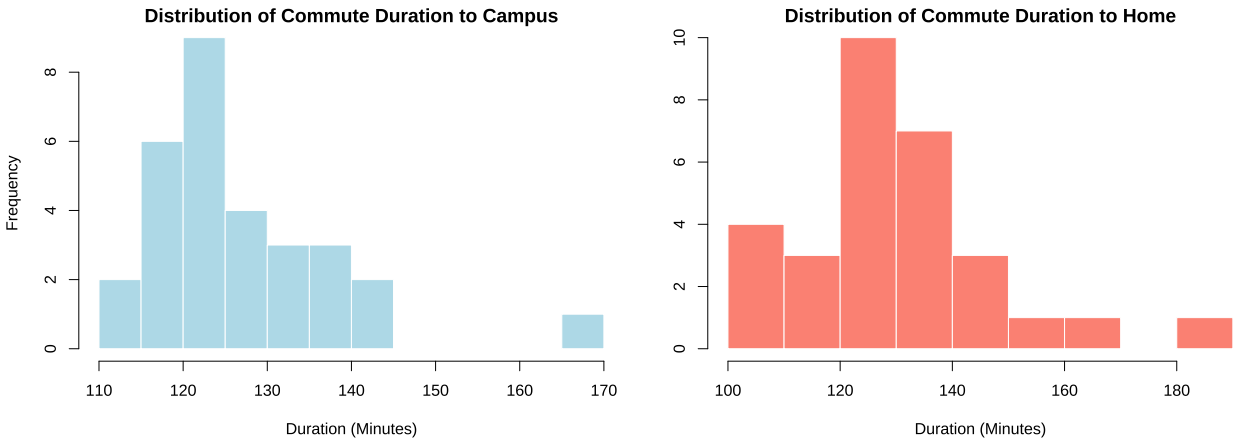

Figure 2: Histograms displaying the distribution of travel times for trips to campus (left) and trips to home (right) with 10-minute bin intervals.

An examination of the distributional shape (Figure 2) 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*.

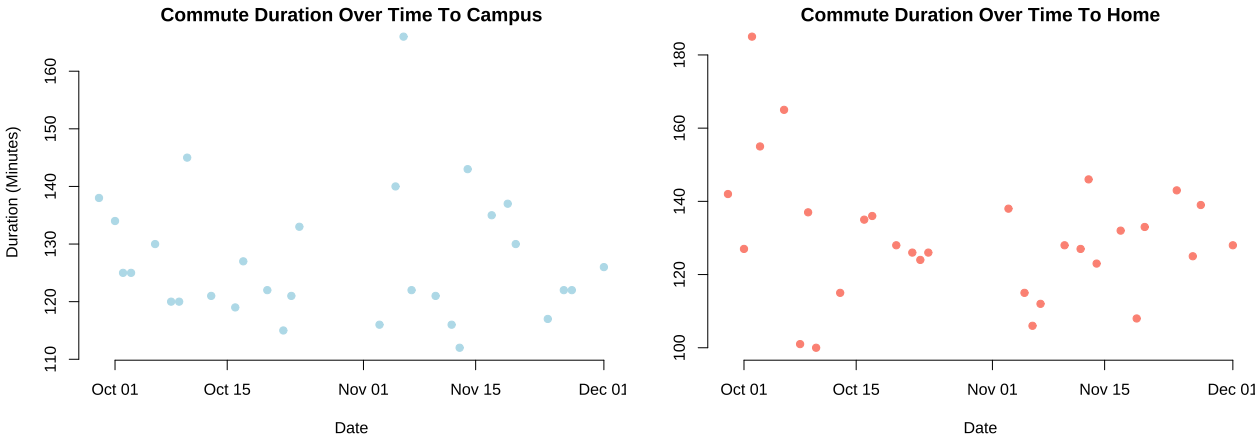

Figure 3: Scatter plots tracking daily commute durations for trips to campus (left) and return trips to home (right) over the observed period.

Plotting the commute durations over time (Figure 3) highlights a difference in *volatility* 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 a higher variance, with widely dispersed data points, suggesting unpredictable delays that call for further statistical testing.

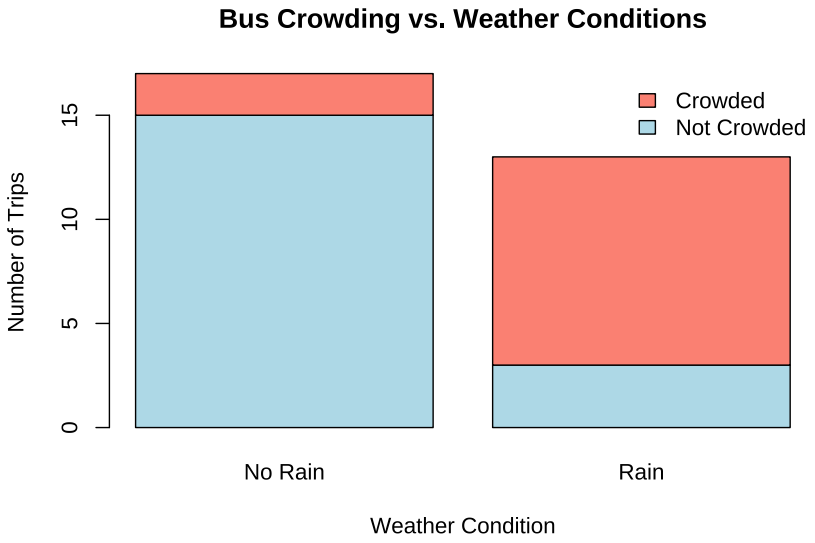

Figure 4: Stacked bar chart showing the proportion of crowded UniLink buses to home during clear weather versus rain.

An analysis of the interaction between weather and bus capacity (Figure 4) suggests a strong correlation:

- **On clear days**, most of the trips are not crowded.
- **On rainy days**, majority of the trips experience overcrowding, likely due to an increase in students choosing public transport over walking or cycling.

2.2 Statistical Analysis Plan

While the exploratory analysis highlighted potential trends, such as the volatility of the trips back home 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 3) suggested that the trip back home is not only more volatile but potentially longer on average than the trip to campus. 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). We use a *one-sided* test because the journey home is expected to take longer, as it typically occurs later in the day when more people are outside.

- H_0 (Null): There is no difference in mean duration between the trips to campus and the trips home ($\mu_{campus} = \mu_{home}$).
- H_a (Alternative): The trips home take longer than the trips 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 4) indicated a strong dependency between rainy weather and bus overcrowding. We aim to calculate the statistical significance of this association. We use a one-sided test because we anticipate that rain will increase the likelihood of overcrowding.

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 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 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.

2.3 Results and Interpretation

2.3.1 Difference in Commute Durations

To determine if the journey home is indeed longer than the journey to campus, we perform a **paired t-test**. This controls for day-specific traffic conditions by comparing the two specific trips made on each date.

$$H_0 : \mu_{home} = \mu_{campus}$$
$$H_a : \mu_{home} > \mu_{campus}$$

```
t_result <- t.test(bus_data$dur_to_home, bus_data$dur_to_campus,
                   paired = TRUE, alternative = "greater")
t_result
```

	T_Statistic	Degrees_of_Freedom	p_Value	Mean_Difference
t	0.64	29	0.264	2.8 min

- **Decision:** Since the **p-value** (0.264) is greater than the significance level ($\alpha = 0.05$), we **fail to reject the null hypothesis**.
- **Interpretation:** There is **insufficient** evidence to conclude that the trip back home takes longer than the trip to campus.

2.3.2 Predictability of Arrival Times

We use **Linear Regression** to assess how accurately the *departure time* predicts the *arrival time*.

$$H_0 : \text{Slope} = 0 \quad \text{(Departure time does not predict arrival time)}$$
$$H_a : \text{Slope} \neq 0 \quad \text{(There is a linear predicitive relationship)}$$

```
# Combine the departure times and arrival times for campus and home trips
all_dep <- c(bus_data$dep_to_campus, bus_data$dep_to_home)
all_arr <- c(bus_data$arr_to_campus, bus_data$arr_to_home)

# Convert Strings to POSIXct (Date-Time objects)
# We use a dummy date so R sees them as times on the same timeline
dep_time <- as.POSIXct(all_dep, format="%I:%M:%S %p")
arr_time <- as.POSIXct(all_arr, format="%I:%M:%S %p")

# Calculate the Linear Model
fit <- lm(as.numeric(arr_time) ~ as.numeric(dep_time))
fit
```

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept (Unix Epoch)	-2.354711e+07	1.429123e+07	-1.6477	0.1048
Departure Time	1.013300e+00	8.100000e-03	125.1397	0.0000

We modeled the relationship using the linear equation:

$$\text{Arrival} = A + B(\text{Departure}) + \epsilon$$

Validating this model against the data reveals:

- **Decision:** Since the slope is not zero, we **reject the null hypothesis**. There is evidence that arrival time is dependent on departure time.
- **Intercept (A):** Omitted since `POSIXct` data is stored as seconds since 1970. It holds no practical meaning for this daily commute analysis.
- **Slope (B):** The slope is 1.013, indicating an almost 1-to-1 relationship. For every minute you leave later, you arrive approximately one minute later.
- **Error Margin (ϵ):** The typical deviation from the predicted arrival time is 14.8 minutes (Residual Standard Error).
- **Predictive Power (R^2):** The model explains 99.6% of the variance in arrival times.

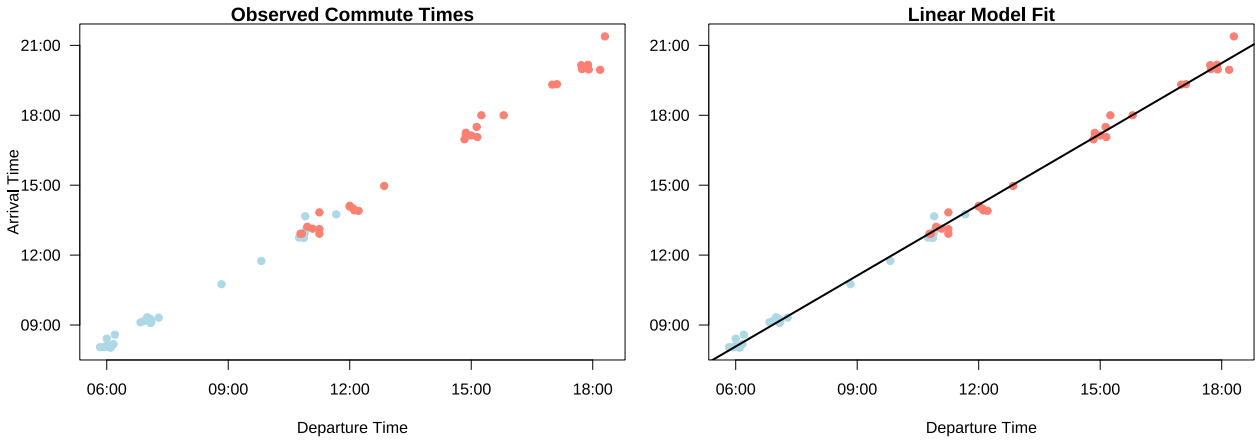

Figure 5: Regression analysis of commute times. The left panel displays all departure vs. all arrival times, colored by direction (Blue: To Campus, Pink: To Home). The right panel overlays the linear regression model (black line), indicating a strong predictive relationship.

2.3.3 Impact of Rain on Crowding

We apply **Fisher's Exact Test** to determine if rainy weather increases the odds of the UniLink bus being overcrowded. Fisher's test compares the Odds Ratio (OR). An odds ratio of 1 means the events are independent. Since we are testing if rain increases crowding, the alternative is "greater than 1."

$$H_0 : OR = 1 \quad \text{(Rain and Crowding are independent)}$$
$$H_a : OR > 1 \quad \text{(Rain increases the odds of Crowding)}$$

```
contingency_table <- table(bus_data$rain_to_home, bus_data$crowded_unilink_to_home)
rownames(contingency_table) <- c("Clear Weather", "Raining")
colnames(contingency_table) <- c("Not Crowded", "Crowded")
contingency_table
```

	Not Crowded	Crowded
Clear Weather	15	2
Raining	3	10

```
fisher_result <- fisher.test(contingency_table, alternative = "greater")
fisher_result
```

Odds_Ratio	p_Value
21.31	0.000465

- **Decision:** Since the p-value (0.000465) is less than 0.05, we **reject the null hypothesis**.
- **Interpretation:** The data indicates that rain impacts commute comfort. The odds ratio suggests that the odds of encountering a crowded bus are approximately **21 times** higher on a rainy day compared to a clear day.

2.3.4 Testing the "Thursday Rush"

To investigate the "Thursday Rush," we perform a **Pearson's Chi-square Test** to check if bus crowding is dependent on the specific day of the week.

$$H_0 : p_{Mon} = p_{Wed} = p_{Thu} = p_{Fri} \quad \text{(Crowding rates are equal across all days)}$$
$$H_a : \text{At least one day has a different crowding rate}$$

```
# Order days logically (Mon-Fri), not alphabetically
days <- factor(
  bus_data$day_of_week,
  levels = c("Monday", "Wednesday", "Thursday", "Friday") # No Tuesday trips
  )
contingency_table_chi <- table(bus_data$crowded_unilink_to_home, days)
rownames(contingency_table_chi) <- c("Not Crowded", "Crowded")
contingency_table_chi
```

	Monday	Wednesday	Thursday	Friday
Not Crowded	7	5	1	5
Crowded	2	2	7	1

```
chi_result <- chisq.test(contingency_table_chi, simulate.p.value = TRUE)
chi_result
```

Chi_Square	p_Value
10.45	0.008

- **Decision:** Since the p-value (0.008) is less than 0.05, we **reject the null hypothesis**.
- **Interpretation:** There is **sufficient** evidence to support the claim that specific days are inherently more crowded.

2.3.5 Bayesian Risk Update: The Impact of Rain

Finally, we apply **Bayes's Theorem** to quantify how a rainy weather forecast should update our risk assessment for bus overcrowding.

- Let H be the Hypothesis that **"The Bus is Crowded"**.
- Let D be the Data that **"It is Raining"**.

$$P(H|D) = P(H) \cdot \frac{P(D|H)}{P(H) \cdot P(D|H) + P(\bar{H}) \cdot P(D|\bar{H})}$$

Where:

- $P(H)$: The *prior* probability that the bus is crowded.
- $P(H|D)$: The *posterior* probability that the bus is crowded. The probability the bus is crowded given that it is raining.
- $P(D|H)$: The probability it is raining given that the bus is crowded.
- $P(D|\bar{H})$: The probability it is raining given that the bus is not crowded.

Probability Calculation:

```
# The columns are stored as TRUE (1) and FALSE (0)
# So the mean is the probability of TRUE (1)
p_H <- mean(bus_data$crowded_unilink_to_home)
p_not_H <- 1 - p_H

crowded_days <- subset(bus_data, crowded_unilink_to_home == TRUE)
not_crowded_days <- subset(bus_data, crowded_unilink_to_home == FALSE)

p_D_given_H <- mean(crowded_days$rain_to_home)
p_D_given_not_H <- mean(not_crowded_days$rain_to_home)

p_posterior <- (p_H * p_D_given_H) / ((p_H * p_D_given_H) + (p_not_H * p_D_given_not_H))

risk_factor <- p_posterior / p_H
```

- *Prior Probability* ($P(H)$): 40%.
- *Posterior Probability* ($P(H|D)$): 76.9%.
- Rain acts as a *Risk Multiplier* of 1.92×.

Visualizing the Belief Shift: To visualize this update, we model our belief as a **Beta Distribution**. The *Prior (Blue)* represents our uncertainty before checking the weather, while the *Posterior (Pink)* shows how the presence of rain shifts our expectation toward "Crowded."

```
# We need counts for alpha/beta parameters
total_trips <- nrow(bus_data)
crowded_trips <- sum(bus_data$crowded_unilink_to_home) # TRUE is 1, FALSE is 0
rain_trips <- subset(bus_data, rain_to_home == TRUE)
crowded_rain <- sum(rain_trips$crowded_unilink_to_home)

alpha_prior <- crowded_trips
beta_prior <- (total_trips - crowded_trips)

alpha_post <- crowded_rain
beta_post <- (nrow(rain_trips) - crowded_rain)

# Plotting
curve(dbeta(x, alpha_post, beta_post), from = 0, to = 1,
      col = "salmon", lwd = 3, lty = 1,
      main = "Bayesian Update: Impact of Rain on Crowding Probability",
      xlab = "Probability of Crowding", ylab = "Density",
      ylim = c(0, 5)) # To ensure both plots fit vertically

curve(dbeta(x, alpha_prior, beta_prior), add = TRUE,
      col = "lightblue", lwd = 3, lty = 2)

legend("topright", legend = c("Posterior (rain)", "Prior (general)"),
       col = c("salmon", "lightblue"), lwd = 3, lty = c(1, 2), bty = "n")
```

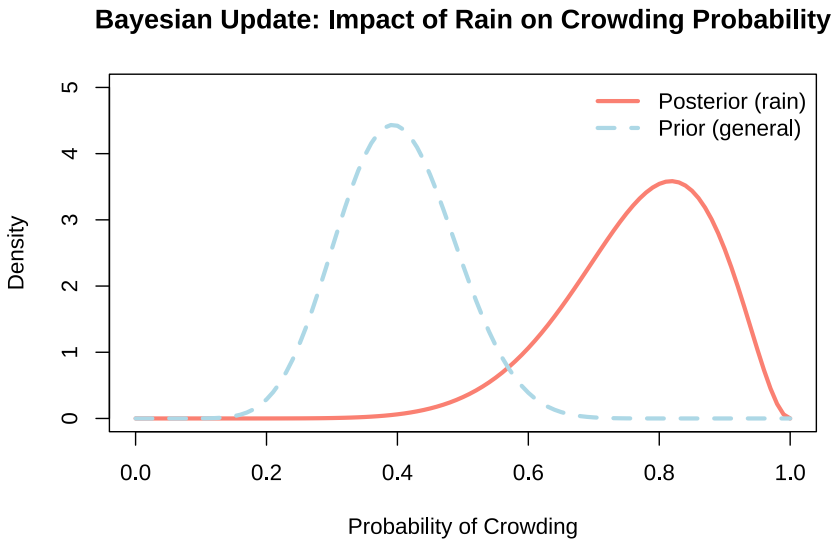

Figure 6: Beta distribution plots showing the shift in belief. The Prior (blue dashed) represents the baseline probability of overcrowding across all days. The Posterior (pink solid) shows the updated probability given rainy weather, illustrating an increase in the risk of crowding.

3 Conclusion

This project set out to analyse the reliability and comfort of my daily commute to and from campus using a data-driven approach. By applying statistical tests to personal bus logs, we have moved beyond anecdotal evidence to identify distinct patterns in travel efficiency and bus overcrowding.

3.1 Summary of Statistical Findings

The following table summarizes the core hypotheses tested and the conclusions drawn from the data:

Research Question	Method	Key Result
Is the journey home longer than the journey to campus?	Paired t-test	**No.** Both trips take similar times.
Can we accurately predict arrival time based on departure?	Linear Regression	**Yes.** A strong linear relationship exists ($R^2 \approx 0.996$), but random traffic noise creates a residual error of $\pm$ 14.8 minutes regardless of departure time.

Research Question	Method	Key Result
Does rain lead to overcrowding?	Fisher's Exact Test	**Yes.** Rain is a predictor of crowding. The odds of a full bus increase dramatically ($21\times$) during rain.
Is overcrowding random across the week?	Pearson's Chi-Square	**No.** Crowding is systematic and dependent on the day. Thursdays showed a disproportionate frequency of crowded trips compared to the rest of the week.
How much should I worry about rain?	Bayes's Theorem	**Double the Risk.** The presence of rain acts as a risk multiplier ($1.92\times$), shifting the probability of a crowded bus from a baseline (40%) to a near-certainty (76.9%).

3.2 Limitations and Areas for Concern

While the analysis reveals clear trends, several limitations should be considered when interpreting the results:

- **Sample Size:** The dataset consists of only 30 observations (days), which is sufficient for initial hypothesis testing. However, a larger sample size ($n > 100$) would provide more robust estimates, particularly for the Pearson's Chi-square test of individual days.
- **University Timetables:** The Chi-square analysis assumes *"Day of the Week"* is the primary factor for crowding. However, the university's lecture schedule (e.g., a large class finishing at 6 PM on Thursdays) likely acts as a hidden variable driving the *"Thursday Rush"*.
- **Seasonality:** Data was collected over a single semester. Seasonal changes (winter darkness, exam periods) may alter these patterns, meaning the regression model may need re-calibration for future months.
- **Binary Definition of "Crowded":** The variable `crowded_unilink_to_home` is subjective (`TRUE`/`FALSE`). A more detailed metric (e.g., "Seats Available") would allow for deeper analysis.

3.3 Final Thoughts

The Commute Chronicles started as a simple record of daily travel but grew into a story about reliability versus chaos. The data shows that the morning commute follows fixed routines and is mostly

predictable, while the evening commute needs flexible planning.

For the remainder of the semester:

- Go home earlier on Thursdays to avoid the systematic rush.
- Treat a rainy forecast as a signal to prepare for a crowded journey.

This project demonstrates that even small, personal datasets can provide actionable insights when we analyze them carefully with statistics.

Project repository and commit history available at: github.com/Daniel-Ope06/the-commute-chronicles