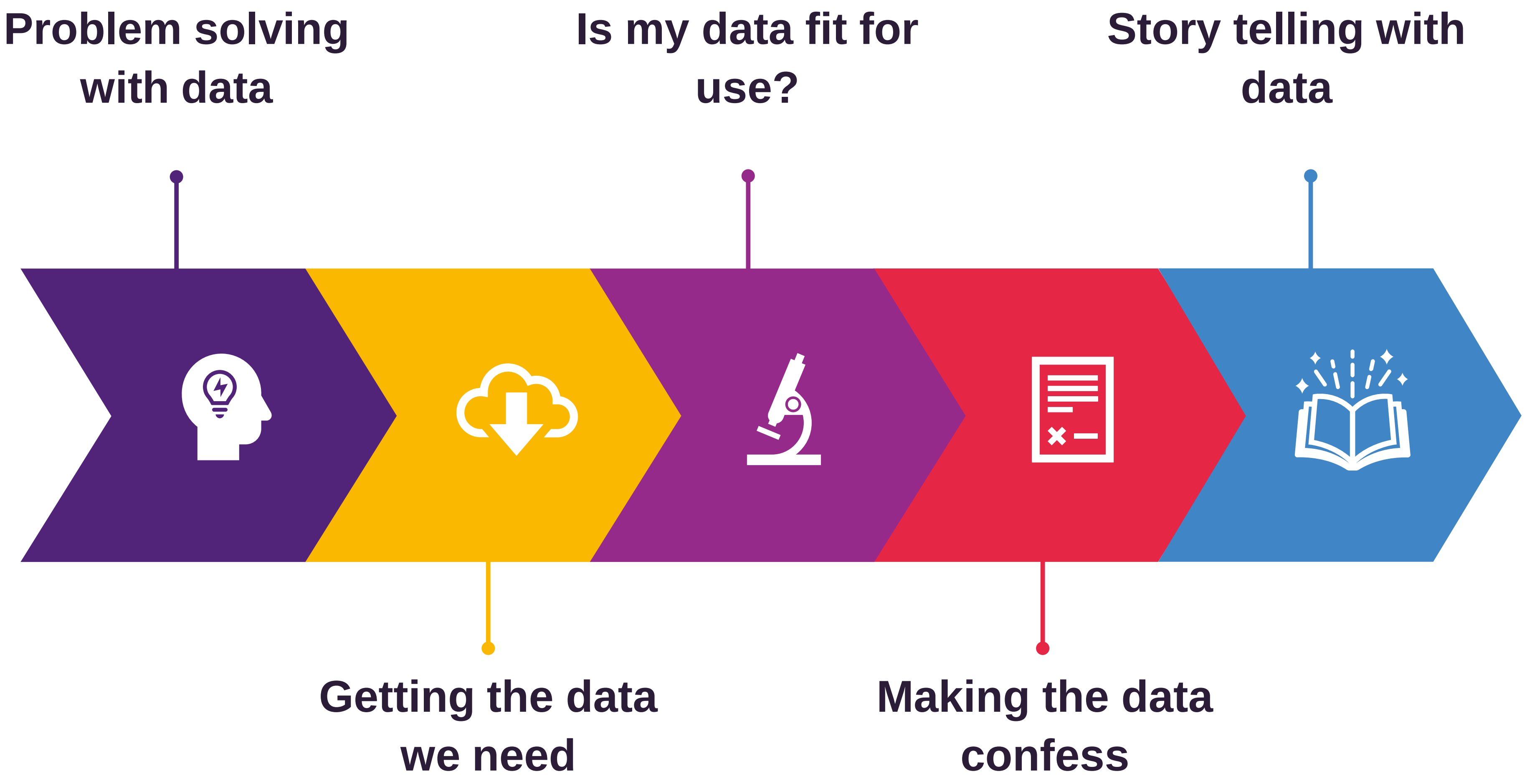
Beyond the Bus Stop

Insights into Brisbane's Public Transport

- Group 18 – Greg, Linh, Haoqing, Miya, Chaitanya, Wilson

The University of Queensland
Australia

https://www.abc.net.au/listen/programs/brisbane-drive/electric-bus-comes-to-redland-city/13208392


Problem solving with data
Getting the data we need
Is my data fit for use?
Making the data confess
Story telling with data

Problem Solving with Data

Motivation

As UQ students reliant on public transport, we aim to use data to improve efficiency, match services to demand, support planning, and study weather impacts on commuters.

- **How do weather conditions impact on the trip volume?**
- **How do people from different suburbs use public transport?**
- **Can we forecast future trip volumes?**

Getting the Data I Need

Dataset Dimensions

		Files	Rows	Columns
Queensland Government	• Origin Destination Trips 2022-2025	38	300-500K each	9
	• Origin Destination Stop Locations Jan 2022-Feb 2025	1	16932	5
	• Translink Time Table (current)	3	2.6M	11
Australian Government Bureau of Meteorology	• Weather Observations Brisbane City (Queensland) from April 2024 to March 2025	12	28-31 each	21
Australian Bureau of Statistics	• 2021 Census Data pack - General Community Profile (Postal Areas)	1	433	109
Brisbane City Council	• Brisbane Suburb Boundaries (Arcgis)	1	195 Suburbs	

Is my Data Fit for Use?

Missing Data

Missing Not At Random

- *Fare Evasion*

  No ticket = No trip recorded.

- *Smart Ticketing*

  Smart Ticketing Trips are not recorded (system limitation).

  Affects:

  - Rail - Mid-2022+
  - Ferry - April 2024+
  - Bus - March 2025+

Missing at Random

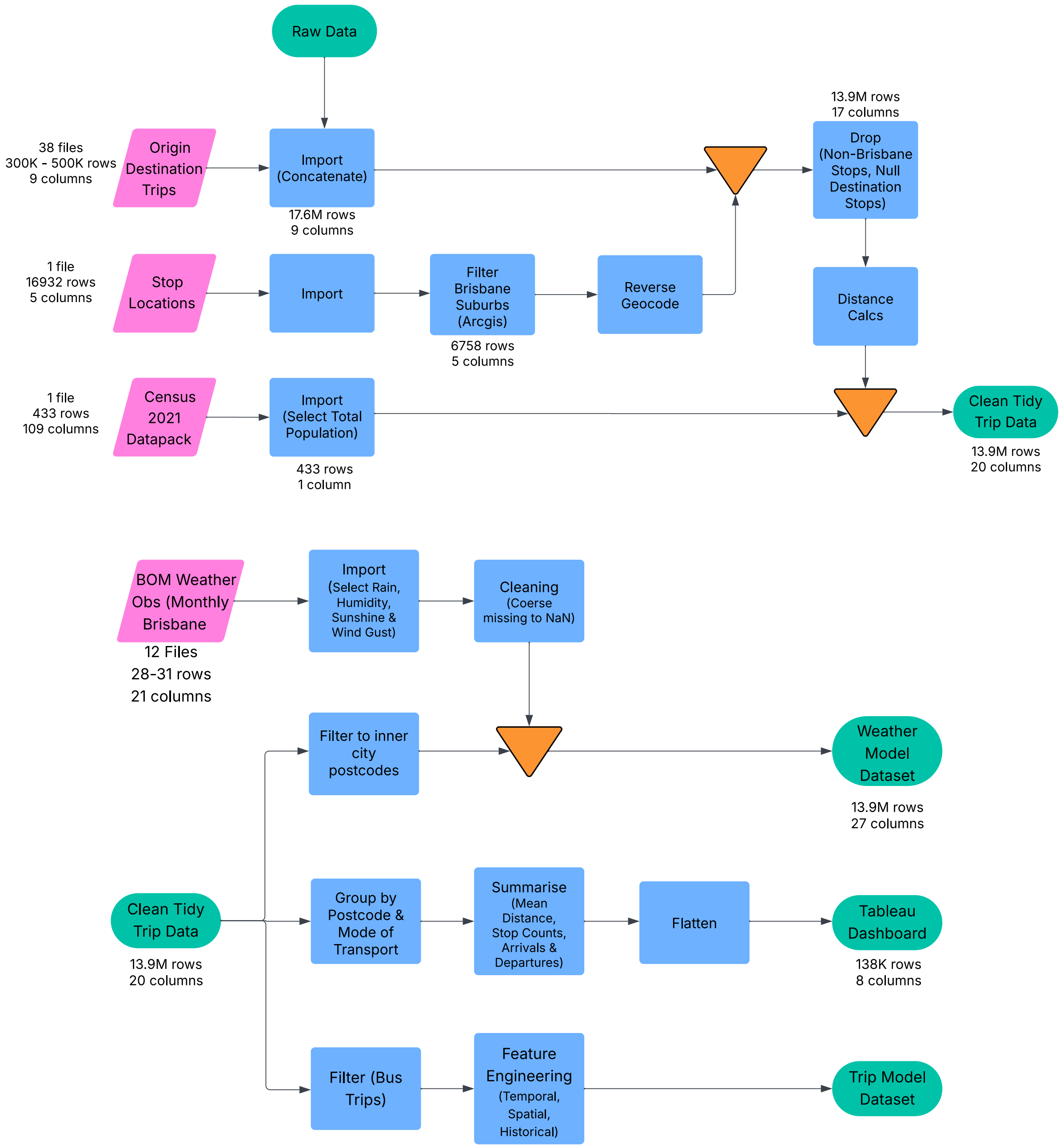
- *Paper Tickets*
  - Destination not recorded. (Records dropped)

Missing Completely at Random

- Missing Weather Observations.
  - Variables dropped if too many missing (eg Wind Gusts).

Is my Data Fit for Use?

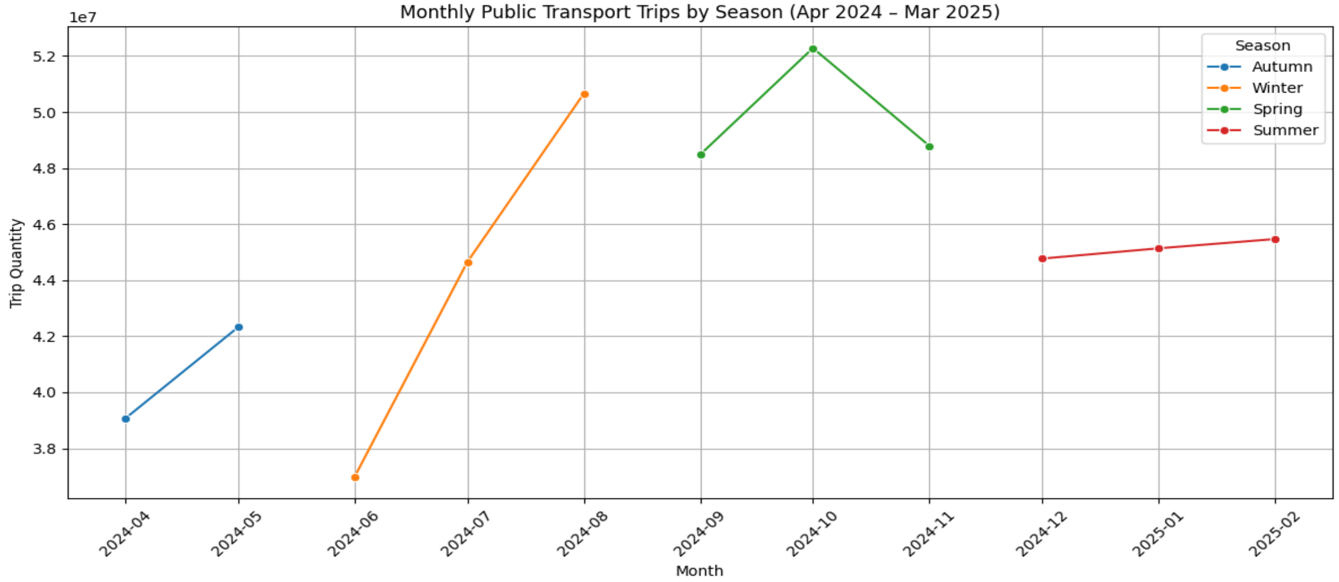
Data Processing



Research Area 1: **Does weather conditions have an impact on the trip volume?**

EXPLORATORY DATA ANALYSIS

- Let us analyze a seasonal trend of ridership over a year (Apr '24 to Mar '25)



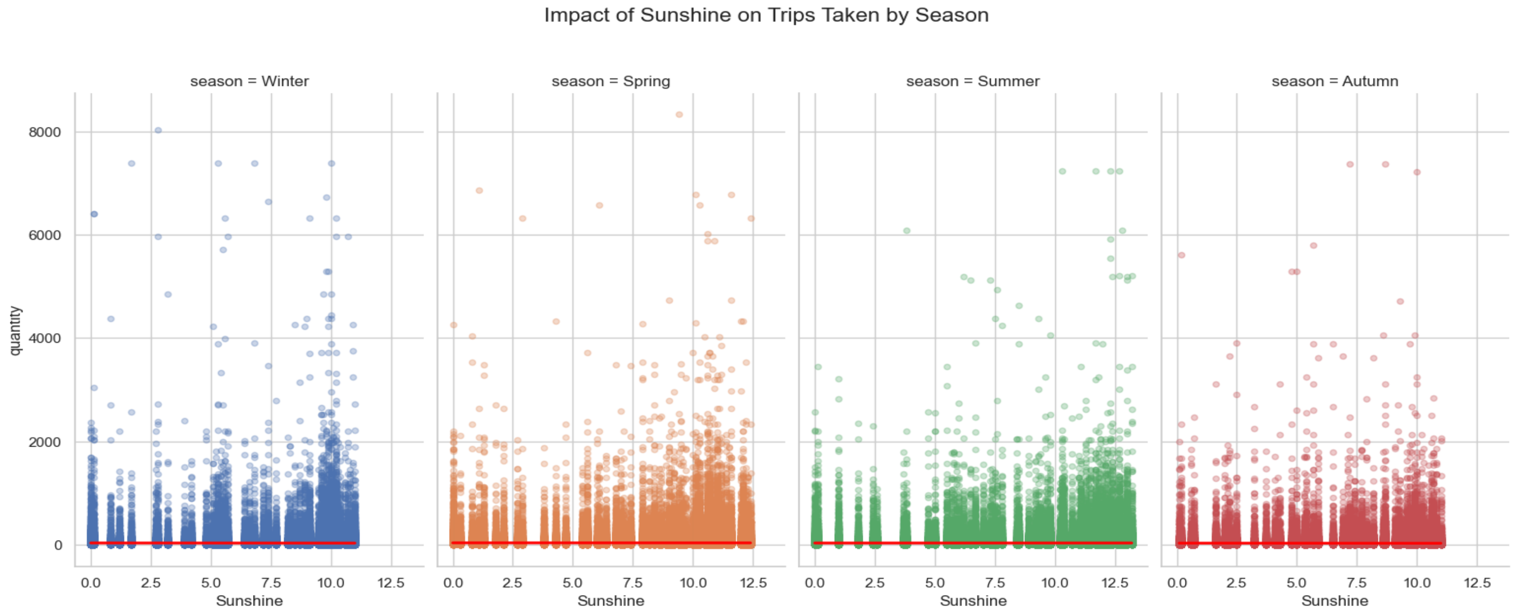
Key Observations:

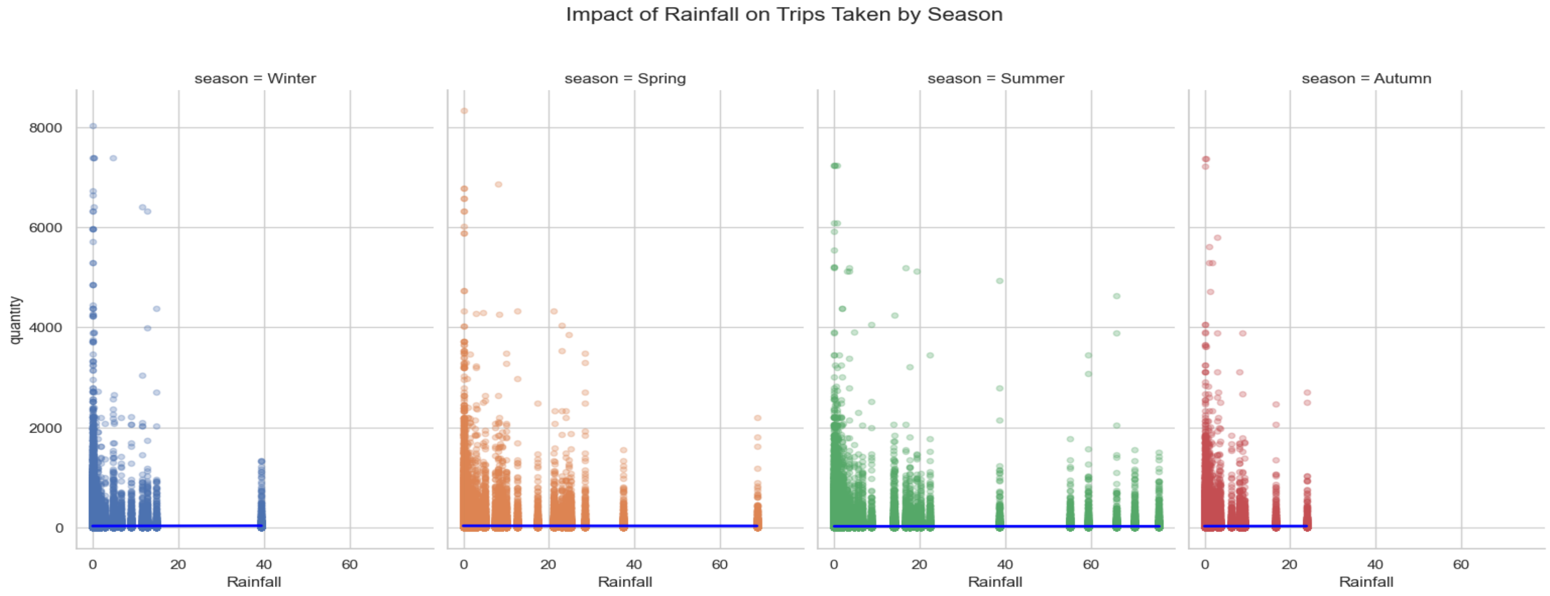
- **Autumn:** Lowest trip volumes
- **Winter:** Highest trip volumes, peaking in August
- **Spring:** Maintains consistently high usage with moderate fluctuation
- **Summer:** Shows moderate but steadily increasing trip volumes. Dips in December.

Research Area 1: Does weather conditions have an impact on the trip volume?

EXPLORATORY DATA ANALYSIS

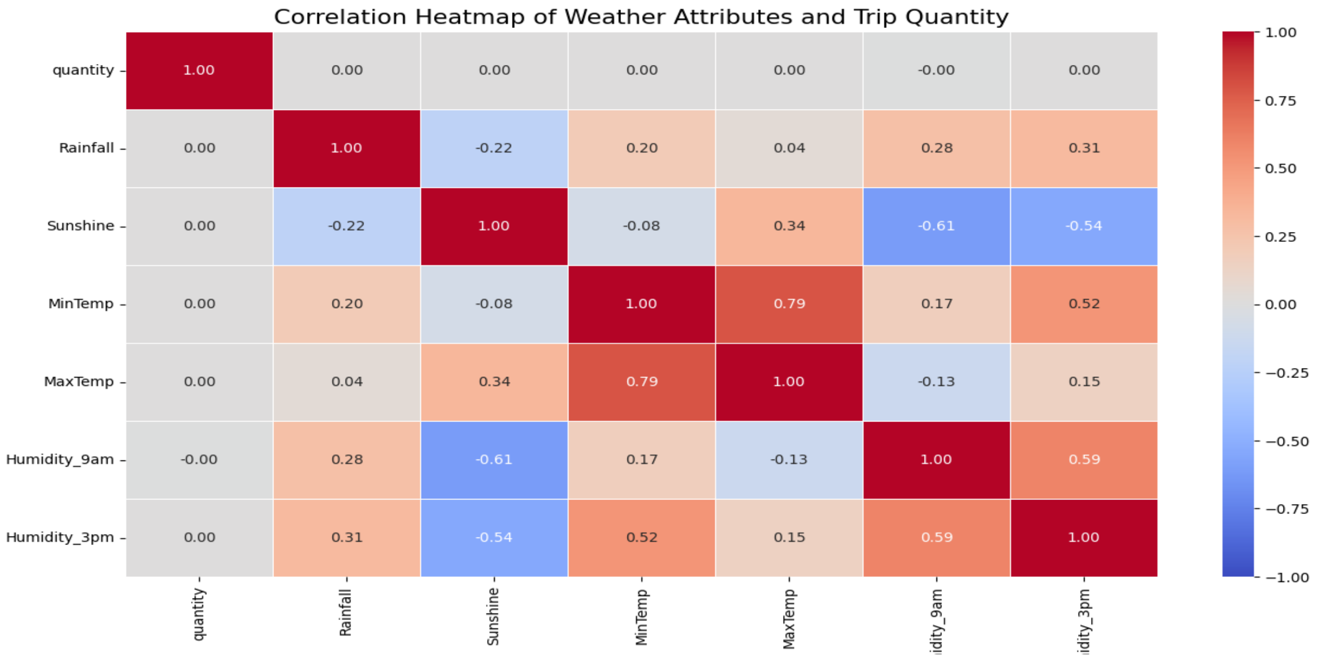
- Let us analyze the influence of sunshine and rainfall on trip volume (per season)





Research Area 1: Does weather conditions have an impact on the trip volume?

- Let us analyze the correlation between weather factors and public transport trips



Variables	Coefficient	Std Error	t-value	p-value	95% CI
Intercept	35.3003	0.100	351.422	0.000	[35.10, 35.50]
Sunshine	0.0748	0.011	6.897	0.000	[0.054, 0.096]
Rainfall	0.0150	0.004	4.143	0.000	[0.008, 0.022]

R-squared =0.000 (indicates weather does not explain variation in trip count)

F-statistic =27.46 (significant, but very tiny effect)

Number of observations: 13.86 million

<u>Research Area 2:</u>

How people from different suburbs use public transport?

EXPLORATORY DATA ANALYSIS

Purpose : Understand how people are commuting from different suburbs using public transport

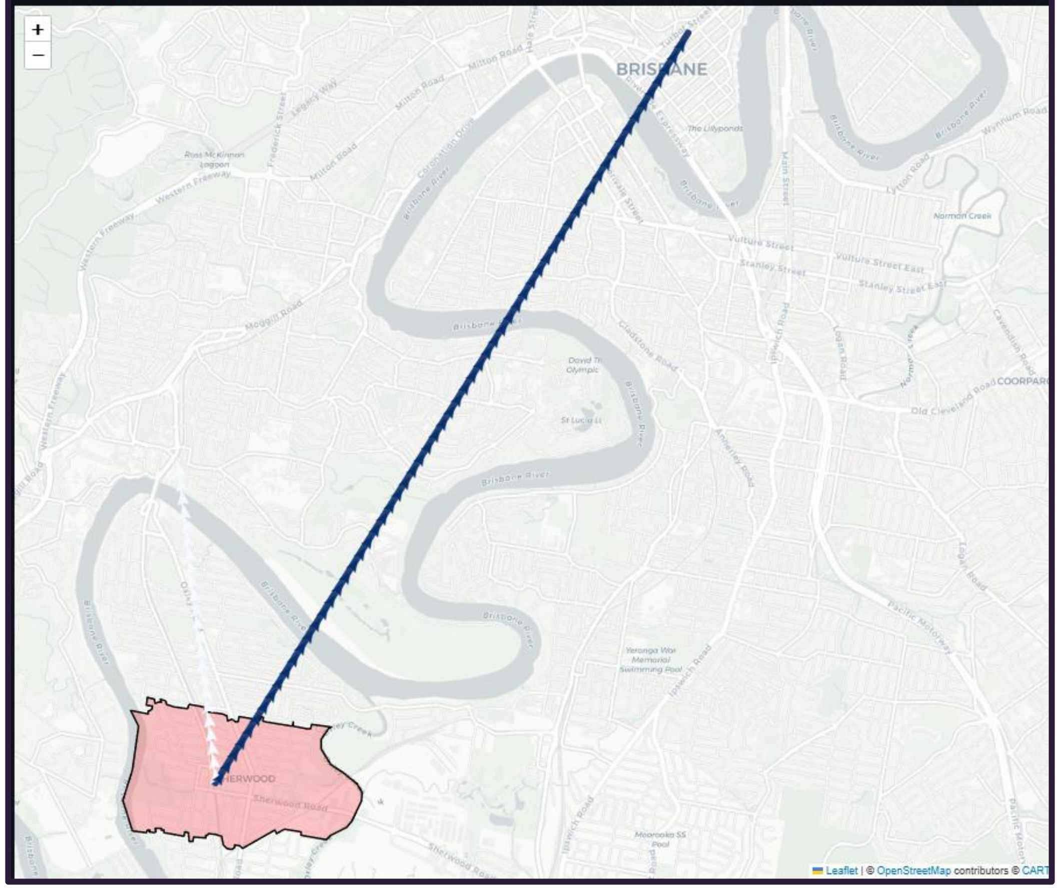
- Transport Type? *E.g. Bus/Train/Ferry*
- Transport Route? *E.g. Bus 412*
- How many transfers made?
- Popular routes/stops?

E.g. Top 1% of Trips for Brisbane City

	origin_stop_name	destination_stop_name	route	quantity
0	Central station, platform 1	Fortitude Valley station, platform 1	Rail	30664
1	Central station, platform 1	Northgate station, platform 1	Rail	25896
2	Central station, platform 1	Ferny Grove station, platform 1	Rail	20719
3	Central station, platform 1	Eagle Junction station, platform 1	Rail	20682
4	Roma Street busway, platform 1	QUT Kelvin Grove station, platform 2	330	20106
5	Central station, platform 1	Indooroopilly station, platform 1	Rail	20106
6	Central station, platform 1	Toowong station, platform 1	Rail	19503
7	Central station, platform 1	Darra station, platform 1	Rail	18060
8	Central station, platform 1	Nundah station, platform 1	Rail	17937
9	Central station, platform 1	Albion station, platform 1	Rail	17791

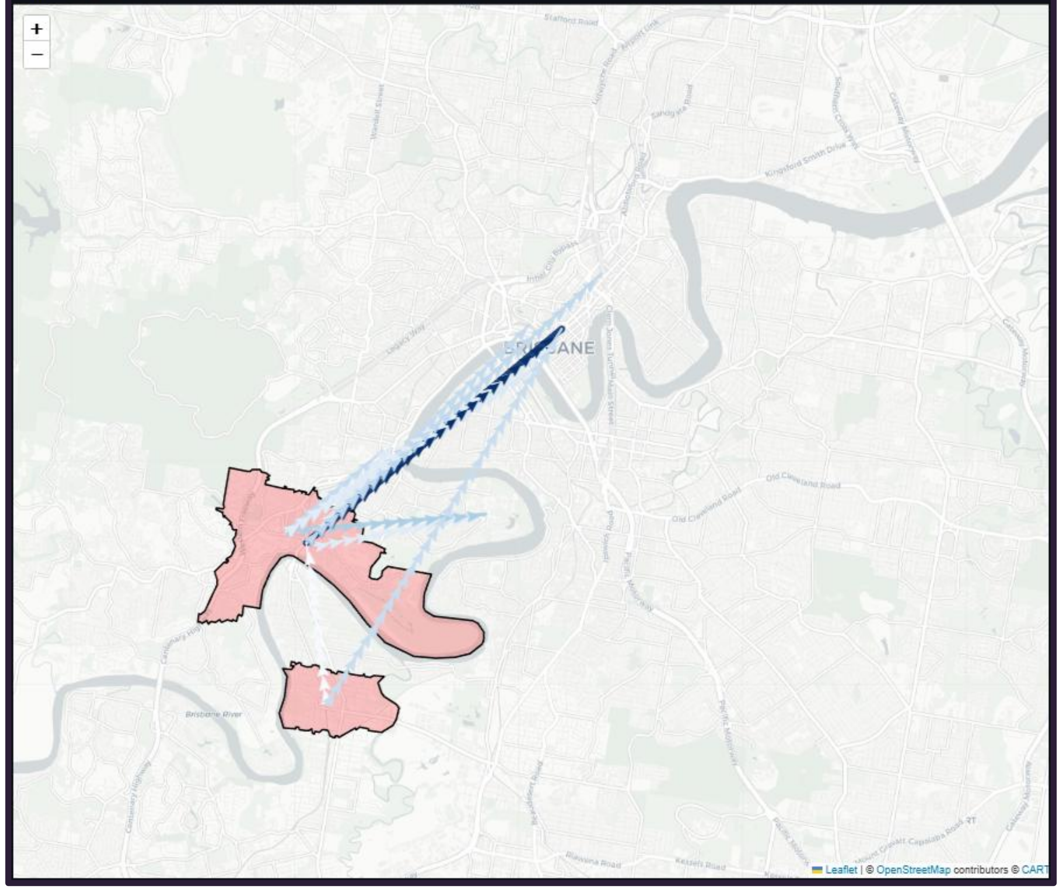
E.g. Suburb Sherwood

Top 1% trips from: SHERWOOD



origin_stop_name	destination_stop_name	route	quantity
Sherwood station, platform 1	Central station, platform 1	Rail	6578
Sherwood station, platform 1	Indooroopilly station, platform 1	Rail	2433

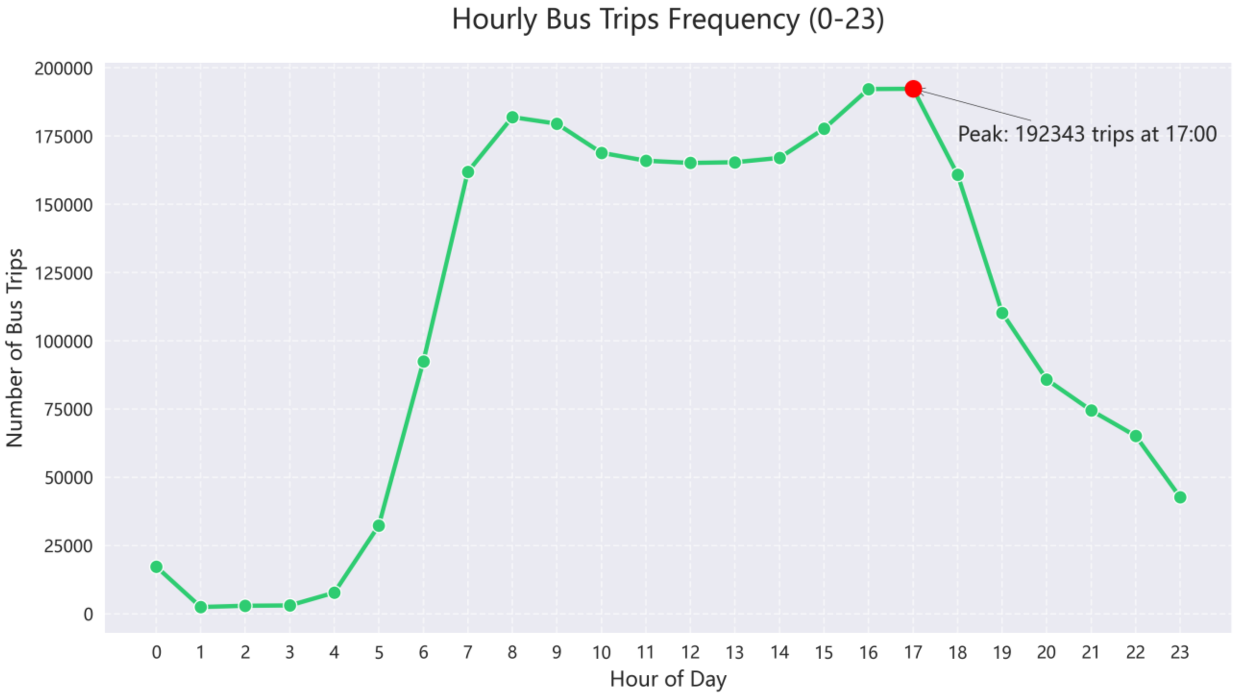
Top 1% trips from: INDOOROOPILLY, SHERWOOD

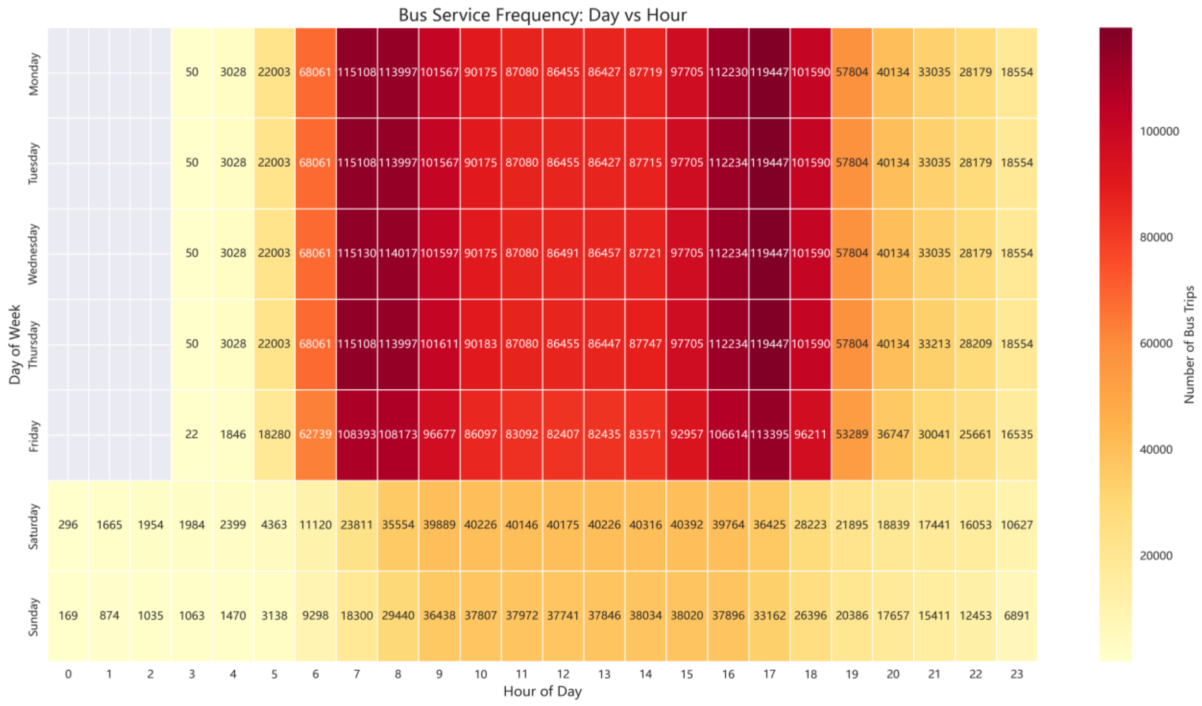


	origin_stop_name	destination_stop_name	route	quantity
0	Indooroopilly station, platform 1	Central station, platform 1	Rail	20594
1	Indooroopilly Shopping Centre station, stop C	University of Queensland	427	8460
2	Indooroopilly station, platform 1	Fortitude Valley station, platform 1	Rail	7240
3	Indooroopilly station, platform 1	Roma Street station, platform 10	Rail	6962
4	Indooroopilly Shopping Centre station, stop B	High St at Toowong, stop 14 (temp relocation)	415	6848
5	Sherwood station, platform 1	Central station, platform 1	Rail	6578
6	Indooroopilly Shopping Centre station, stop B	Cultural Centre station, platform 1	425	5776
7	Indooroopilly Shopping Centre station, stop B	Queen Street station	425	4757
8	Indooroopilly station, platform 1	Toowong station, platform 1	Rail	4585
9	Lambert Rd near Central Ave, stop 35	University of Queensland	427	4537

Research Area 2:
How people from different suburbs use public transport?

EXPLORATORY DATA ANALYSIS





Research Area 3:

Passenger Demand Forecasting

MAKING THE DATA CONFESS

Training Setup

All models were trained on a preprocessed and feature-engineered dataset that included:

- Temporal features (day type, hour bucket, month)
- Spatial identifiers (route, direction)
- Historical trends (lagged passenger counts, rolling means, percentage changes)

Train-Test Split

To simulate real-world forecasting, the dataset was split chronologically:

- **Training set:** All data from 2022 to 2024 (5820423 rows)
- **Test set:** All data from 2025 (432883 rows)

	route	direction	day_type	hour_bucket	origin_stop	destination_stop	distance_km	lag_1	lag_2	lag_3	rolling_mean_3	pct_change	year	month_num
0	0	6	0	0	1026	3043	15.896244	22.0	18.0	16.0	18.666667	0.454545	2022	4
1	0	6	0	0	1026	3043	15.896244	32.0	22.0	18.0	24.000000	0.031250	2022	5
2	0	6	0	0	1026	3043	15.896244	33.0	32.0	22.0	29.000000	-0.181818	2022	6
3	0	6	0	0	1026	3043	15.896244	27.0	33.0	32.0	30.666667	-0.074074	2022	7
4	0	6	0	0	1026	3043	15.896244	25.0	27.0	33.0	28.333333	-0.280000	2022	8
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5820418	892	2	0	2	3276	3246	2.744571	2.0	1.0	1.0	1.333333	-0.500000	2024	10
5820419	892	2	0	2	3276	3246	2.744571	1.0	2.0	1.0	1.333333	0.000000	2024	11
5820420	892	2	0	2	3276	3246	2.744571	1.0	1.0	2.0	1.333333	1.000000	2024	12
5820421	892	12	0	2	3246	3276	2.744571	3.0	1.0	1.0	1.666667	1.666667	2024	11
5820422	892	12	0	2	3246	3276	2.744571	8.0	3.0	1.0	4.000000	0.250000	2024	12

Model Selection Rationale

Four models were chosen to represent varying levels of complexity:

- **Linear Regression, Decision Tree, XGBoost, LightGBM**

These models help compare simple vs. advanced approaches on the same dataset, balancing accuracy and complexity.

Research Area 3: Passenger Demand Forecasting

MAKING THE DATA CONFESS

Results & Insights

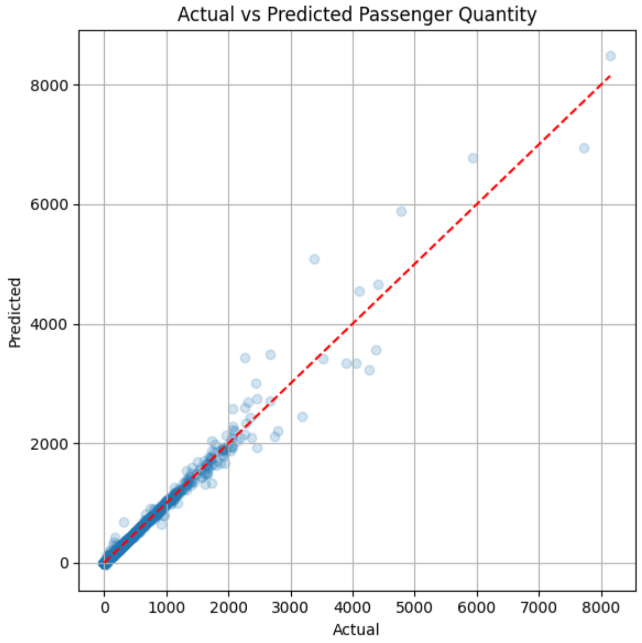
Models Comparision

- LightGBM outperformed all models, accurately capturing temporal, spatial, and trend-based patterns
- Linear Regression served as a baseline but lacked capacity for non-linear patterns.
- Decision Tree and XGBoost offered moderate gains but were less consistents

Model	MAE	RMSE
LightGBM	**0.29**	**5.91**
Decision Tree	1.32	7.35
XGBoost	0.59	11.12
Linear Regression	3.51	14.51

Prediction Accuracy – LightGBM

- Predictions closely align with actual values along line, indicating strong overall model performance.
- Accuracy is **high for typical demand levels**, with minimal deviation.



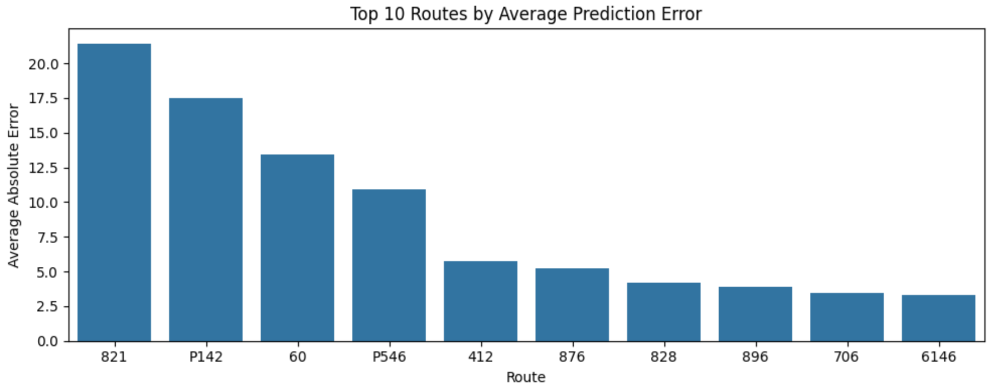
Research Area 3:

Passenger Demand Forecasting

MAKING THE DATA CONFESS

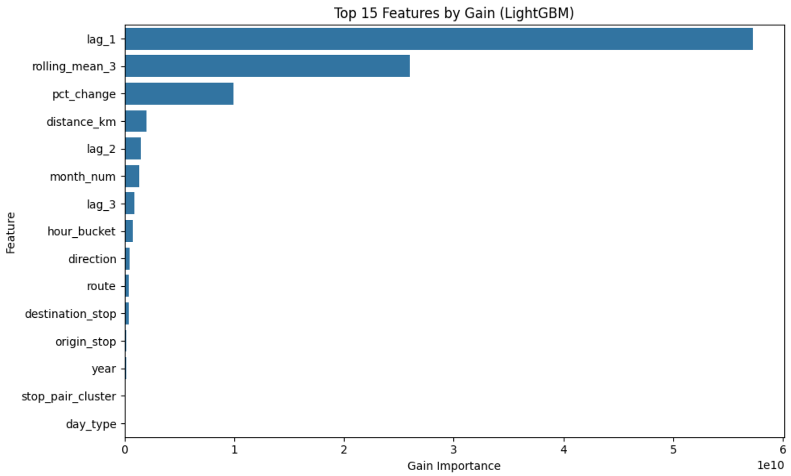
Route-Level Error Analysis

- **Unstable demand patterns** contribute to high prediction errors on certain routes.
- **Limited historical data** reduces model accuracy for less frequently used or newer routes.
- **External factors** changes should be considered for high-error routes.



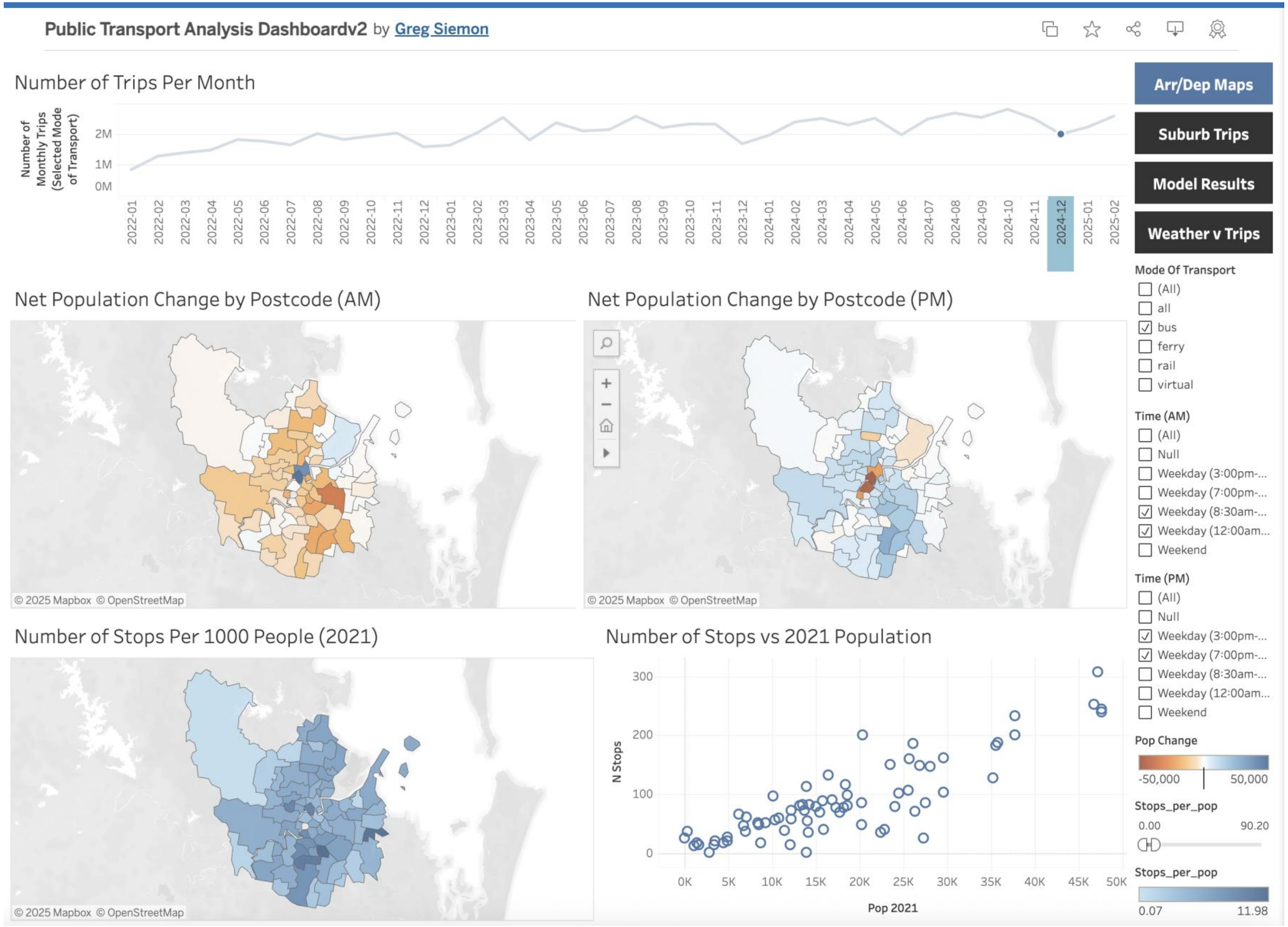
Feature Importance (Gain-Based)

The model relied most on temporal trend features, highlighting the importance of recent demand and short-term patterns. Spatial features and day_type contributed little, suggesting that time-based variables already captured their effects.



Storytelling with Data

Tableau Dashboard (Demo)



[Link]

Storytelling with Data

Recommendations and Further Study

Recommendations

- Align service levels with routine patterns (eg. school terms, work peaks)
- Consider increasing the number of stops in areas with <6 stops per 1000 population. Lack of access may be contributing to low number of trips. - Translink
- Connect Tableau to a Database (flat files too big for it to handle reliably)

Further Study

- Include more variables: day of week (weekend/weekday, holidays, local events, school calendars etc)
- Explore lag and extreme weather impacts
- Trip segmentation: leisure vs commute (Regular commuters travel regardless of the weather whereas leisure trips are likely to be affected by poor weather)
- Estimate Missing Data – eg fare evaders, smart ticking, etc
- Integrate timetable analysis with trip data.