A Framework for Analysing Delays within Public Transport Networks

University of Queensland Master of Data Science
Group 10 Project Report - DATA7001

Johannes Volk[1], Snehin Raj Singh Kukreja[1],
Lasya Sahadeva Reddy[1], Boonyapat Sukosit[1],Zixuan Deng[1], Donghan Yang[1]
[1] The University of Queensland

Abstract

With this project we propose a framework to monitor and record the live state of public transport networks. This includes detailed route and vehicle descriptions, the current delays and other attributes defined in the open GTFS-RT specification. We simultaneously record the live precipitation and cloud coverage from freely accessible radar and satellite images and geotrack the complete public transport fleet exemplary for the Translink network in South-East Queensland, Australia. We conduct statistical analysis on the distribution of delays by route, area and time by recording this real-time data over multiple weeks. We compare the precipitation levels and delays in order to make conclusions about the robustness of the Translink infrastructure against precipitation by locations as well as other grouping criteria such as time of day and respective mode of transport or route number.

We give consent for this report or the videos of our presentations to be used as a teaching resource

Contents

1. Problem Solving

1.1. Introduction

This report details our findings for the group project for DATA7001 - Introduction to Data Science within the Master of Data Science at The University of Queensland, Australia. This project aims to solve an existing problem for stakeholders such as the general public, policymakers (e.g. local government), and operators of public transport systems. We use the Translink public transport network within South-East Queensland as an example to demonstrate the capabilities of using real-time data as a data source to build a historical collection of public transport states that can be used to report the past performance in terms of timeliness (section 2). We curate a dataset by preprocessing and transforming raw data streams into a form usable for downstream tasks (section 3). By the established collection of past observations we compute average delays by different criteria such as specific routes, suburbs, times of day, weekdays or overall. As such, we perform exploratory data analysis and further confirmatory analysis in order to quantify our assumptions to where, when and why delays might happen in the network and model our observed data using Regression as well as Classification methods. (section 4). We report our conclusions in section 6. We justify the usage of real-time interfaces to collect data by detailing possible use cases for computing live statistics in transport networks. We give instructions how to scale the proposed framework to a global scale (subsection 6.1).

2. Getting the data we need

The following section describes the used dataset more thoroughly and gives a detailed description of the data collection process and storage. As this project tries to investigate rather specific research questions it was necessary to curate our own dataset and make use of real-time APIs that we frequently polled and recorded. To this end we set up a remote server that captured data at different times in the weeks from March 2024 to May 2024. This includes periods of low to high precipitation as well as all days of the week, days with major events (e.g. ANZAC Day, Rugby/AFL games). With the established framework it is possible to extend our setup to any other city in the world that provides the "General Transit Feed Specification - Realtime" (GTFS-RT) interface to the public. Our observation is restricted to the southern area of the Translink network. This area fully captures the Brisbane and Gold Coast area. This area comprises urban but also more rural areas, distinct bus lanes but also shared roads and highways. The observation window can be scaled arbitrarily, if one merges map tiles, that are polled from the precipitation API. Figure 1 illustrates the whole data collection pipeline.

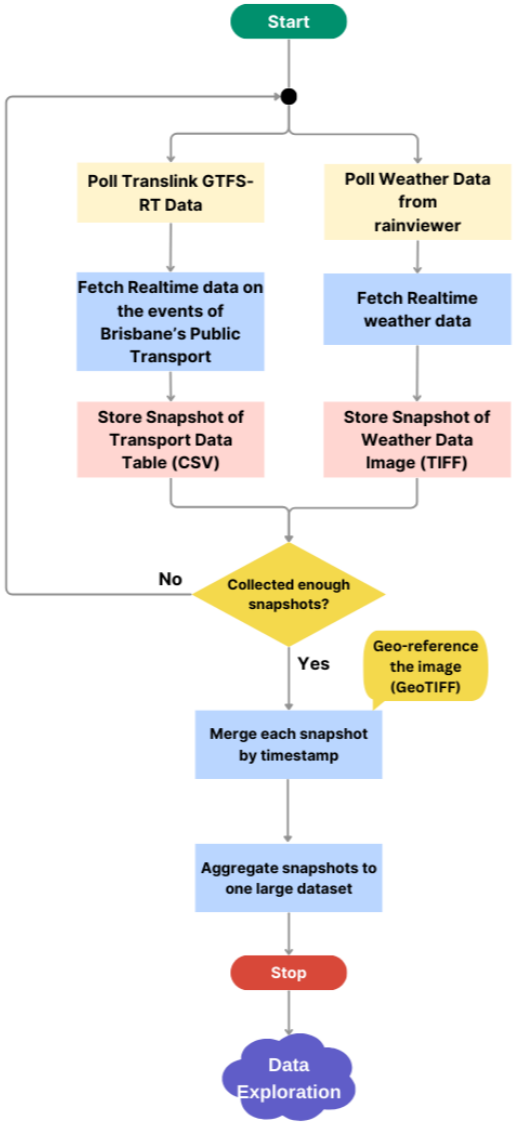


Figure 1. Schematic Flowchart illustration of the data collection procedure yielding a dataset ready for downstream tasks

2.1. Translink GTFS-RT

The GTFS-RT Translink API [8] provides us with the appropriate information about all Translink vehicles that operate in South-East Queensland. The interface is structured according to the "General Transit Feed Specification - Real Time" [1] format that is globally used by local public transport providers to share information about their operating fleet. This includes the location of stops, routes and all vehicles on these routes between stops. Most importantly, it provides details about the current delay of all vehicles in the network in real-time as well as expected delays for all upcoming stations. In this project we only record the current

delay until the next stop and don't evaluate the accuracy of the predictions. Any other indicative vehicle features provided by the API describe aforementioned attributes in more detail (e.g. full route names, route/vehicle IDs, etc.). Furthermore, we not only observe over 1000 busses operating in the network but all trains, ferries and trams. Figure 2 plots the current locations of Translink vehicles on a map [4, 9] to illustrate the vehicle density in the inner city of Brisbane. One can identify areas of higher congestion for different modes of transport (e.g. Buses in Brisbane CBD and Eastern Busway, trains at Roma train station). This project tries to identify more assumptions and translate them into quantifiable, statistically sound and provable facts.

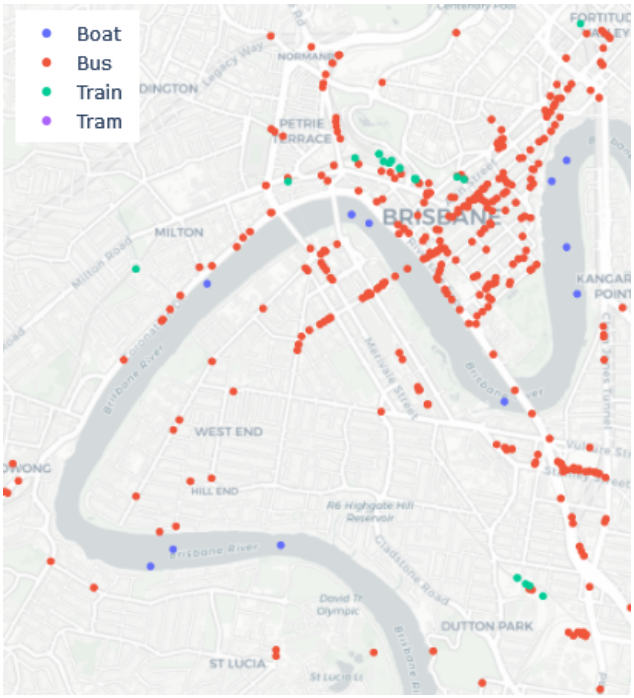
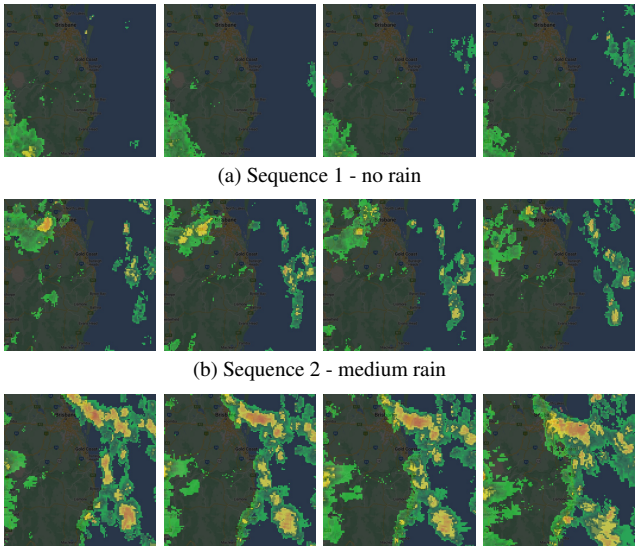


Figure 2. Map with Translink vehicles plotted at their respective location by type during one moment of the data collection process

2.2. Rain Radar Data

While collecting the live data from the Translink API, we also capture the current rain over the South-East Queensland region via the RainViewer API [10] in the same event loop. In Figure 3 one can review exemplary sequences of this data captured during different weather situations.

In order to connect these images as seen above with the individual data points of vehicles and their location at the time of the recording we georeference the radar images and store them in GeoTIFF format [11]. GeoTIFF is an open-source file format that extends the Tagged Image File Format (TIFF) data format in order to georeference satellite or



(a) Sequence 1 - no rain

(b) Sequence 2 - medium rain

(c) Sequence 3 - heavy rain

Figure 3. Three Sequences of four radar images each showing different levels of precipitation in South-East Queensland, Australia over 80 minute intervals

other geospatial images by adding additional information to the pixel values in a different layer. This allows us to query EPSG:4326 coordinates (i.e. longitude and latitude pairs) in any GeoTIFF radar image to obtain the respective rain levels at any location within the observation window at the moment the radar image was captured at. The Geospatial Data Abstraction Library (GDAL) [6] provides the necessary functionalities to first realize said georeferencing and consecutive querying of the radar images.

We also capture satellite images of the observation window as the RainViewer API provides it over an identical interface. However, evaluating a correlation between cloud coverage and delays are not part of this project and could be added to future extensions. For completeness, some examples for such satellite images are displayed in Figure 4.

2.3. Data Collection

In order to build a dataset needed to reliably record observations over six weeks from the 26th of March 2024 to the 10th of May 2024.

We leveraged the UQ Zones cloud computing resources for this course to record and store snapshots of live data over this timeframe. This data would later be downloaded to the personal computers of the group members for further exploration.

The rationale for using cloud computing resources was:

- The instance would be reliable with a high up-time rate
- The instance could be accessed from anywhere at any time allowing for better collaboration between users

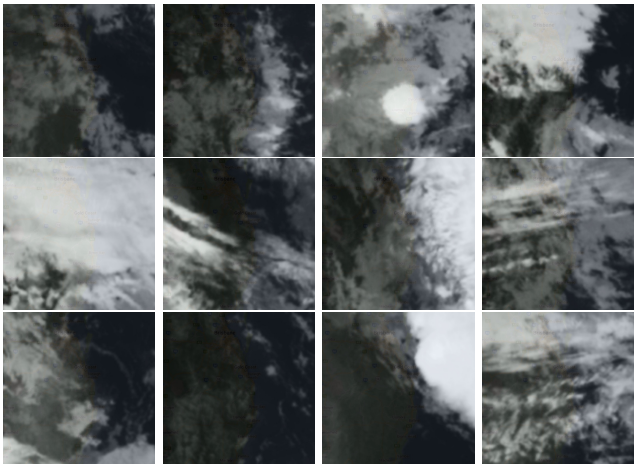


Figure 4. Twelve exemplery satellite images of the observed area in South-East Queensland, Australia at different times and weather situations

- The data would be stored centrally and be easy to access for all users

Resource	Specification
CPU	64-Core AMD EPYC 9334
RAM	4 GiB
OS	Ubuntu 20.04.5 LTS
Storage Size	80 GiB

Table 1. UQ Zones Resource Allocation Specification

Table 1 shows the specifications for the UQ Zones instance used. Although the memory might seem low for modern standards, since each snapshot of TransLink data (roughly 360 kiB) and each TIFF image (roughly 5 kiB) was written to storage and further aggregation was done on each group member's personal computer. Therefore there were no memory constraints in the collection of the data.

Once we started recording snapshots, we ran into an issue with the UQ Zones infrastructure as our instance would repeatedly crash after around 100 snapshots were recorded. The cause of the crashes is unknown, the implications of the crashes on the dataset are further discussed in 3.2.1 Missing Data

2.4. Further usage of data

We extend the dataset by annotating each location with the respective suburb it resides in. For this we integrate shapefiles [5, 7] and convert them into the GeoJSON [3] format to enable easy visualizations in Choropleth plots later on (see section 4). Figure 5 shows the geospatial tiling into suburbs. Herein, the area of the data we obtain from RainViewer API [10] is marked by the red square. This area covers among others the whole Brisbane and Gold Coast city area and provides varied road and traffic conditions of a city center as well as the outside suburbs over to rural areas.

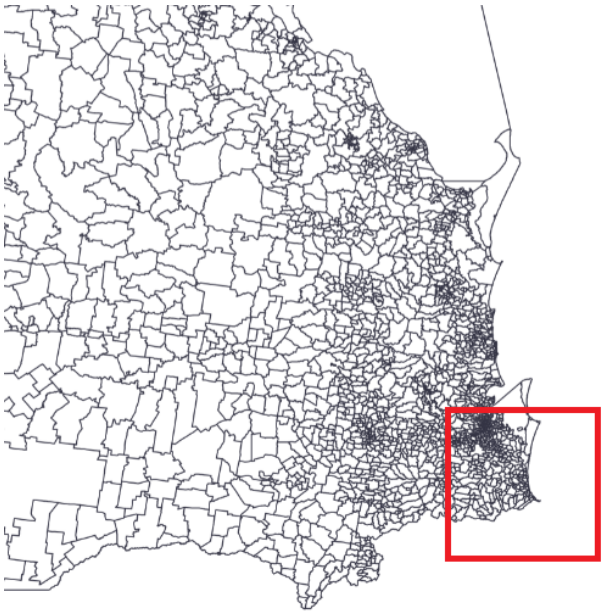


Figure 5. Cropped geospatial tiling of Queensland, Australia into suburbs [7]. The observation area in which we record radar data is marked with the red square

2.5. Data visualization

In the following we are going to examine our data closer in order to get a feeling for how the Translink network is structured. To this end, we aggregate all the data points from all frames like the one displayed Figure 2 and filter according to route types and other criteria. Among other things, this allows us to extract the observed vehicle routes. We present them in Figure 7 for trains, Figure 8a for ferries, Figure 10a for trams in the Gold Coast and for busses at the example for routes going to and from the University of Queensland St. Lucia Campus in Figure 8a.

As displayed in Figure 6 there seem to be irregularities with the actual routes a bus may take. For example bus route 139 (dark blue) displays abnormal location readings outside the expected route. This may be due to local road closures and consecutive route adaptations or other operational necessities such as vehicle rerouting to depots. However, it is not possible to exclude errors and it is necessary to explore this further during data curation (section 3).

As visualized in Figure 7, the Translink operators distinguish train lines going in the exact opposite direction. As such we will consider these pairs of train lines as two different routes as well.

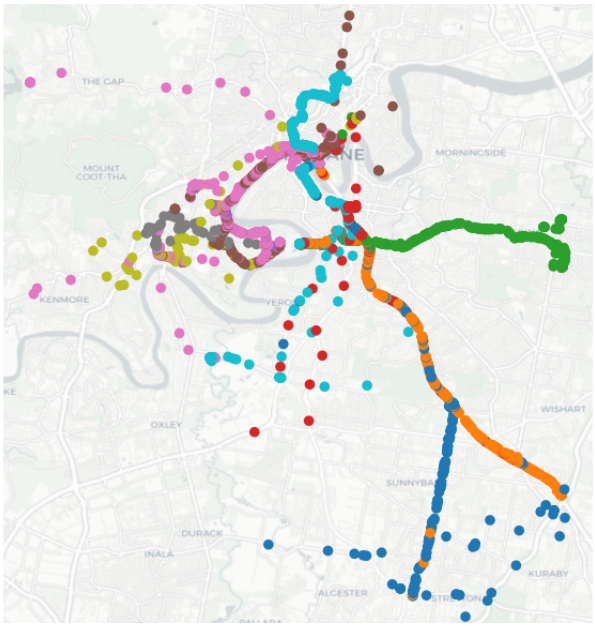


Figure 6. Clustering per bus routes servicing the University of Queensland St. Lucia campus as reconstructed with our data

Among other smaller local routes, the Translink network operates two large ferry systems the paid CityCat (”UQSL”, ”NHAM”) and the free Cityhopper (”SYDS”, ”NTHQ”), which are again split according to the direction of the route. One can see that the Cityhopper only covers a smaller fraction of the Brisbane River and the area serviced by the Citycat. Two other routes are not depicted as they only act as a bridge over in the northern part of Brisbane. While outlier delays are occurring over the whole ferry network (see Figure 8b), it is to mention that a higher density is observed upwards the stream.

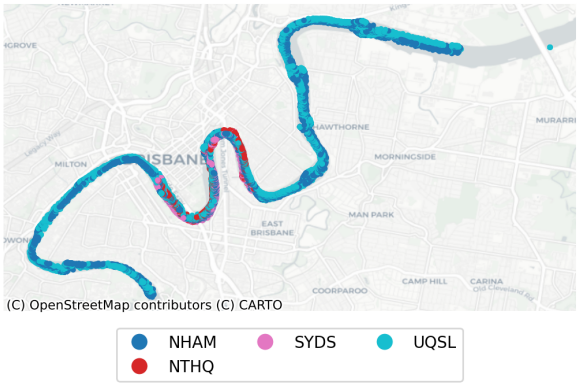
For the Trams (or G-Link) network, once again the API distinguishes between the north-bound (GLKN) and south-bound (GLKS) routes both located in the Gold Coast area. Apparently, the trams display more outlier delays in the coastal region. This might be explained through increased traffic in more urban areas compared to the less crowded suburban areas in the north west of the tram routes. However, compared to other modes of transport the delays are relatively low.

We can also visualize the delays using choropleth plots. An example can be seen in Figure 9, which shows the mean delay per suburb as recorded over the whole dataset.

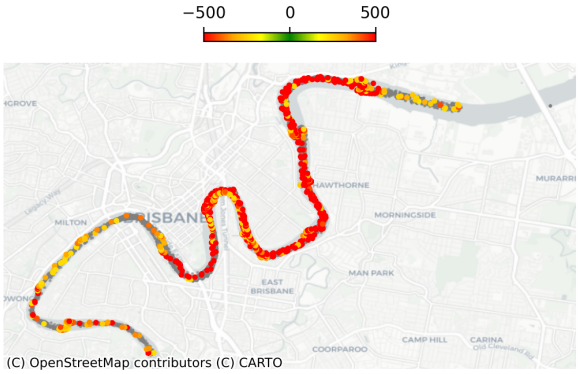
Further, we provide a web-based tool that allows for interactive visualization and exploration from the browser. This allows us to browse the dataset for the different observations as retained from the APIs at a specific point in time. We are able to filter by route name and can see the historical locations of all Translink vehicles at any moment within our observation window. For more details refer to our code base[1].

Figure 7. Clustering per train route within the Translink network as reconstructed by our captured data

[1] github.com/JohannesVolk/data7001-project



(a) Overview of the main ferry routes within the Translink network as reconstructed by our captured data



(b) Map scatter plot for outlier delays below the 2.5th quantile and beyond the 97.5 quantile on CityCat and Cityhopper ferries

Figure 8. Visualization of the four major ferry routes in the Translink network

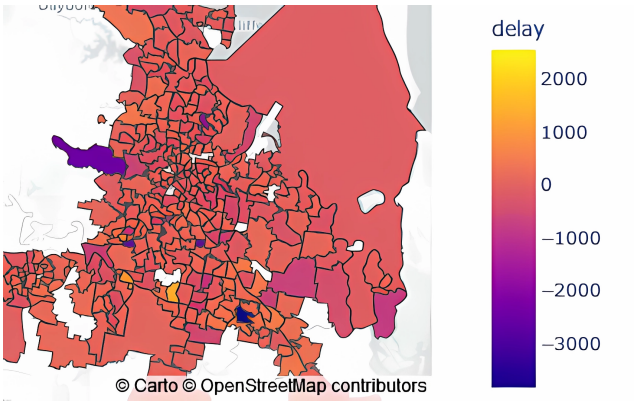
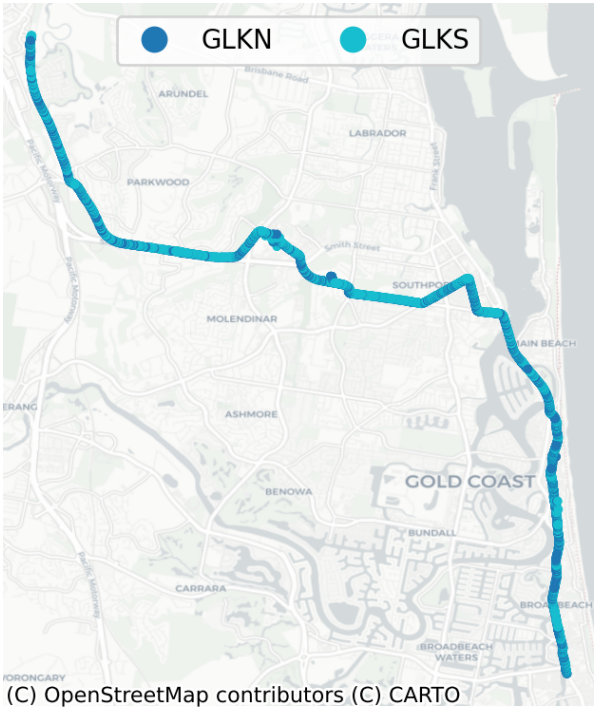
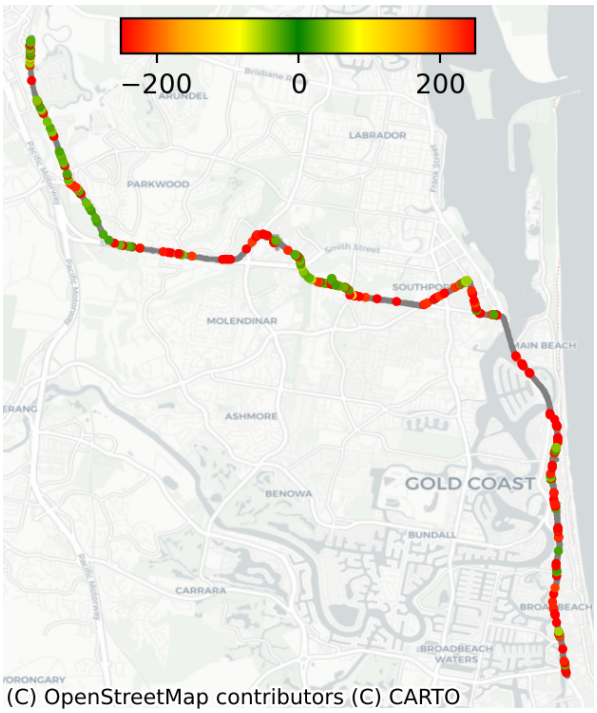


Figure 9. Choropleth of the mean delay per suburb



(a) Overview of the tram routes within the Translink network as reconstructed by our captured data



(b) Map scatter plot for outlier delays below the 2.5th quantile and beyond the 97.5 quantile on the G-Link trams in the Gold Coast

Figure 10. Visualization of the Tram network within South-East Queensland

3. Making the Data Fit for Use

3.1. Data Characteristics

Attribute	Value
File Size (GiB)	1.7
Format	CSV
Structure	Tabular
Number of Columns	29
Number of Rows	1,048,575

Table 2. Translink Data Characteristics

Our final datasets for further analysis and processing consists of the aggregated Translink dataset and the recorded weather data from the Rainviewer API.

The Translink dataset is structured, in a tabular form, with 29 features and around 1 million data points. We captured about 1.7 Gigabytes over six weeks (31 March 2024 to 10 May 2024). Some of the main features we explore are the:

- `Timestamp_Radar` (Unix Time - Integers)
- `Route Number` (String)
- `Stop Name` (String)
- `Latitude` and `Longitude` (Coordinates - Floats) - Provides us with the position of the vehicle
- `Delay` (seconds) - Calculated by taking the difference between the actual arrival time and the scheduled arrival time.

The Rainviewer dataset is a collection of TIFF images as mentioned in Section 2.2. Each image takes about 1 Megabyte in volume, we have collected around 1300 TIFF images with a total size of 1.3 Gigabytes. The rain is measured in decibels (dBz) and the brightness of pixels shows which regions have rainfall and their intensities as shown in Figure 3. Table 3 goes into more detail on the type of data captured in the TIFF image.

Attribute	Value
File Size (MiB)	1.0
Format	TIFF
Structure	Raster
Number of Columns	512
Number of Rows	512

Table 3. TIFF File Information

3.1.1 Missing Data

While recording snapshots from the Translink API and the Rainviewer API we encountered an issue with our Zones instance as it would periodically restart our environment after around one to two hours of recording. As a result, the data capture would have to be manually triggered daily over, this would result in missing data for periods between when the instance crashed and when the data capture was manually triggered again.

The missingness of the data is systematically related to the unobserved data. The time of day serves as a predictor of the likelihood of whether or not data was captured. Since it was more likely for the capture process to be triggered during the day (8 a.m. to 8 p.m.) as opposed to later at night (8 p.m. to 8 a.m.). Therefore we can classify the missing observations as Missing Not At Random (MNAR).

As a result, further explorations and results will have a bias towards how the transport system performs during the day and the weather patterns during the day.

3.1.2 Data Cleaning

During the data cleaning process, several columns were identified with missing values. To address this, the implemented code effectively skipped these blank values, maintaining the integrity of the dataset. Additionally, column assurance was ensured by verifying the presence of all required columns and filling any missing columns with `None` or appropriate default values. The `upcoming_stops` column, containing JSON data, required conversion and parsing to extract key information such as `arrival_delay` for detailed analysis. Furthermore, timestamps were standardized by converting them from UTC to the Australia/Brisbane timezone, ensuring consistency and accuracy in the time data. For instance, a Unix timestamp like `1714529308` was converted to the readable format `2024-04-04 20:42:34+10:00`. These steps ensured that the dataset was clean, consistent, and reliable for subsequent analysis.

3.2. Exploratory Data Analysis

3.2.1 Duplicate Localities

In Queensland, there is a possibility that two localities may have the same name. One example is there is an inner-city suburb 'Albion' with the ZIP Code 4010 in Brisbane and another remote locality "Albion" exists with the ZIP code 4822.

To remove the duplicates, our methodology was to first identify all known duplicate locality names in our dataset and manually check if each locality was within our observation area. Thankfully there was no case where two localities had identical names and were both located in the observa-

tion area. Hence the solution was to remove the duplicate locality record that was not within the observation window. More information can be found in our code base [2].

3.2.2 Analysis of Factors Contributing to Delays

The the delay boxplot from our dashboard was used to analyse the outliers in our dataset. Below are the observations made:

1. Vehicle type:

The box plots indicates that the major delay was caused by the buses followed by trains. There was no significant contribution from trams and ferries towards delay. It was also observed from Figure 11 that, there were hardly any outliers beyond a threshold of 10000 seconds, these outliers were removed from the dataset.

2. Time of the Day:

We performed data exploration on the variation of delays during different times of the day. We have considered three durations: Morning (7AM to 9AM), Afternoon (11 AM to 1 PM) and Evening (4 PM to 6 PM). It is obsereved that most of the delays were caused during peak hours of the day as it is during these hours, when the people need to commute to their work and school. It was also observed that the maximum delay caused was around 90 min - the possible reasons could be traffic congestion, public events or bad weather.

The classification model was built to analyze the correctness of the delay predictions for times of the day and it was found that, the predictions were more accurate in predicting the delays caused during afternoons.

	Morning	Afternoon	Evening
Accuracy	0.52273	0.57725	0.53446

Table 4. Accuracy scores of the classification report

The classification model demonstrated varying performance across different time periods. The midday period showed the best overall accuracy (0.58) with more balanced precision and recall for both classes. The morning and late afternoon periods had lower overall accuracy (0.52 and 0.53, respectively), with the model performing well in identifying delayed instances but struggling to correctly identify non-delayed instances. (compare Table 4)

These results highlight the importance of considering time-of-day variations in transit delay predictions, as the model's performance fluctuates based on the specific period. The precision and recall metrics indicate the model's strength in predicting delays, which is crucial for transit planning and management.

[2] github.com/JohannesVolk/data7001-project

3.3. Analysis of delays based on the zones

There are a total of 8 zones in southern Queensland translink. The objective of this analysis is to predict delays in a transportation system based on the zone_id using a Random Forest Regression model.

We cleaned the 'zone_id' column by converting non-integer values to NaN and dropping rows with NaN values and then we converted 'zone_id' back to integer type for further analysis. We selected 'zone_id' as the feature (X) and 'delay' as the target variable (y) for the regression model. We then split the data into training and testing sets (80% training, 20% testing) using train_test_split from sklearn.We have utilized GridSearchCV for hyperparameter tuning using 5-fold cross-validation and neg_mean_squared_error scoring. Finally we identified the best-performing Random Forest model based on mean squared error (MSE) on the training data.

The results from the Figure 17 shows that, the zone 6 has recorded the highest number of delays followed by zones 8 and 4 respectively.

Based on the above figure, the suburbs that fall under zone 6 are: Nambour, Maroochydore, Sippy Downs, Palmwoods, Mooloolaba and Varsity Lakes. There can be several reasons for high number of delays in this area, such as: traffic congestion, road work and construction, public events and festivals.

Based on the visualizations of delay predicted by the regression model, it is observed that zone 6 is predicted to encounter highest number of delays in the future. But the model achieved an MSE of approximately 2,048,429 on the test data. Hence, the model's performance in predicting delays across all zones can be further improved, as indicated by the relatively high MSE.

3.4. Outliers

From our data capture process there were two types of outliers identified in the data. We categorised the outliers into two categories: Type I - Outliers in terms of delays of buses (in seconds) and Type II - the geospatial position of buses in (longitude and latitude coordinates).

In this project, we handled Type I outliers by arbitrarily filtering out and discarding observed data points where the delay of the transport vehicle was more than 600 seconds (ten minutes). We do not explore mitigating Type II outliers shown in Figure 6.

Figures 11 to 14 show the presence of outliers in the arrival times of public transport vehicles with delays ranging up to 5350 seconds (about 1 hour and 30 minutes). These outliers could be due to errors with the reporting in the Translink API, resulting in bad data collected.

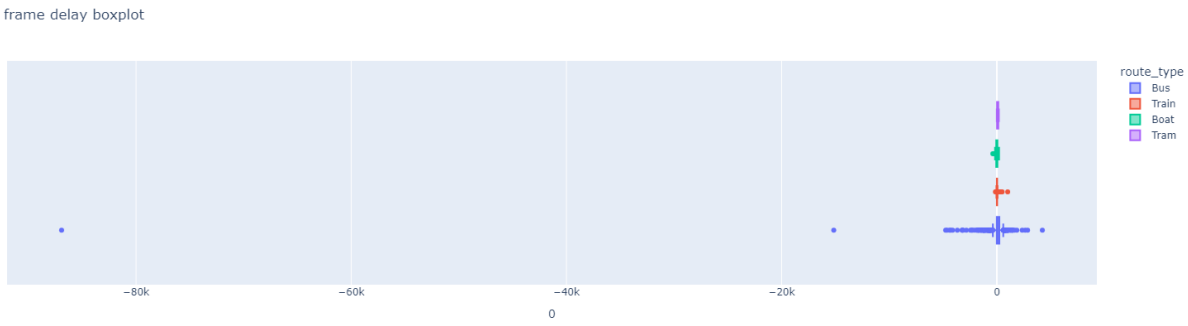


Figure 11. Delay Boxplot

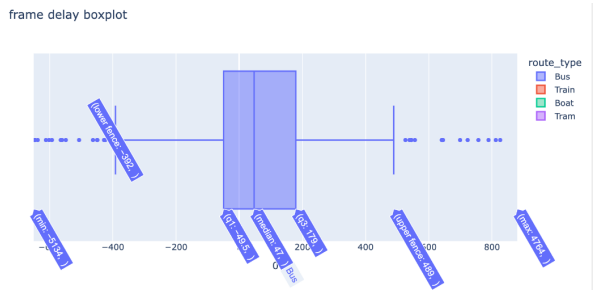


Figure 12. Delay Boxplot for 7AM - 9AM

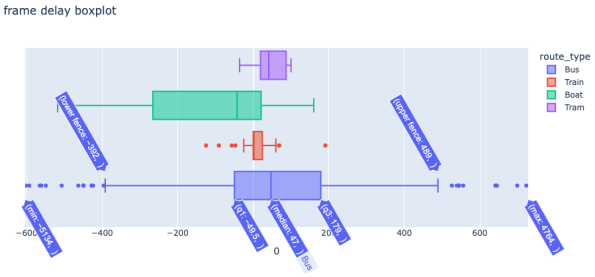


Figure 13. Delay Boxplot for 11AM - 1PM

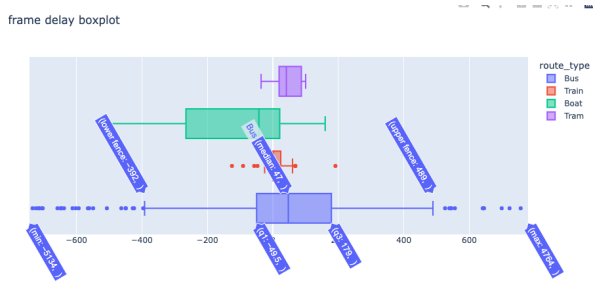


Figure 14. Delay Boxplot for 4PM - 6PM

Category	Outlier Description
Type I	Outliers in terms of the vehicle delay measured in seconds (Positive values indicates the vehicle is behind schedule and negative values indicates the vehicle is ahead of schedule)
Type II	Outliers in terms of the geospatial position of the vehicle where the vehicle is well outside the route it should be on. Measured in longitude and latitude coordinates.

Table 5. Outlier categories identified in our research

4. Making Data Confess

In this section, we embark on the critical task of exploring and transforming the dataset to extract meaningful patterns, uncover hidden insights, and prepare the data for advanced analytics and modeling. The overarching goal is to make the data ”confess” its underlying structure and relationships, enabling us to derive actionable insights and make informed decisions.

4.1. Classification

In this subsection, we delve into the application of classification techniques, specifically Random Forest and Support Vector Machine (SVM), to predict delays based on two key factors: rainfall and geographical zones. The primary objective is to develop robust predictive models that can accurately classify and anticipate delays in a given context, leveraging the power of machine learning algorithms.

The motivation behind using classification models stems from the need to proactively identify and mitigate potential delays, which can significantly impact various domains such as transportation, logistics, and service delivery. By harnessing the predictive capabilities of Random Forest and

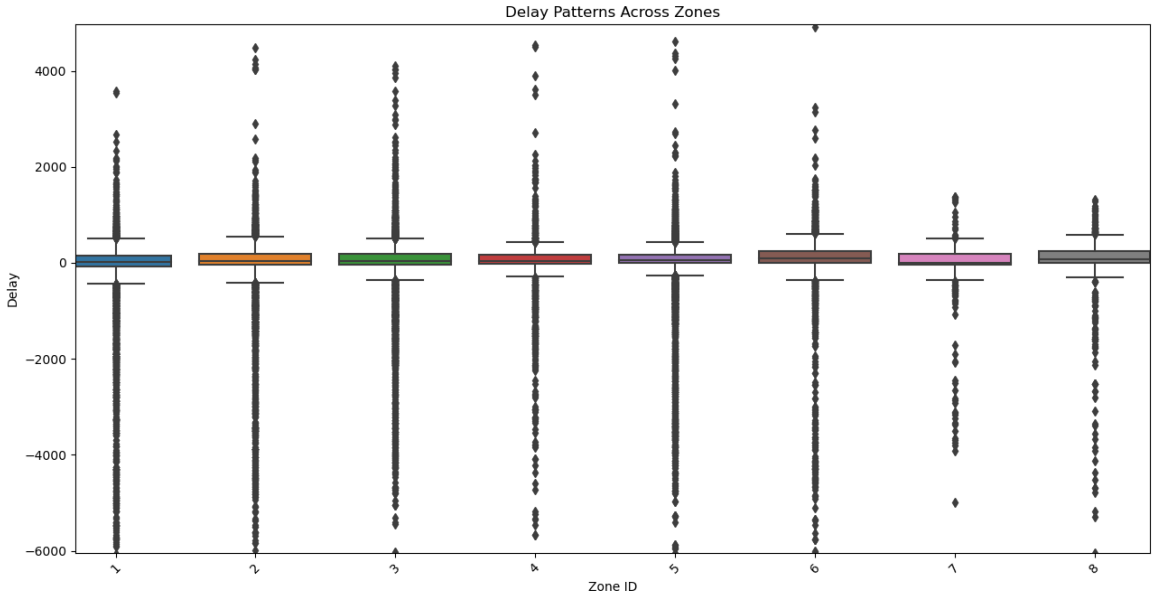


Figure 15. Delay patterns across the Translink zones

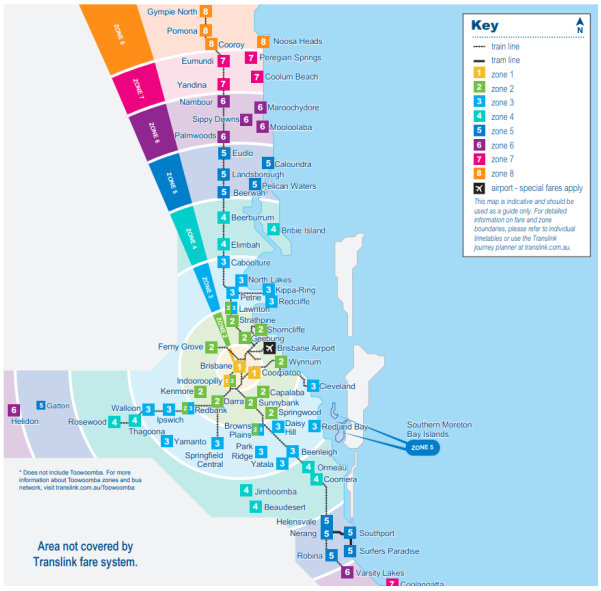


Figure 16. Zones partitioning South-East Queensland [12]

SVM classifiers, we aim to enhance decision-making processes and optimize resource allocation in scenarios where delays pose substantial challenges.

The choice of Random Forest and SVM as our classification algorithms is driven by their strengths in handling complex, high-dimensional data, capturing nonlinear relationships, and offering robust performance in classification tasks. Through rigorous model training, evaluation, and validation, we seek to identify the most effective approach for predicting delays and informing proactive decision-making strategies.

4.1.1 Analysis of Impact of Rain Intensity on Delay Severity

The objective of this analysis is to investigate the impact of rain intensity on delays in our transport dataset. The dataset consists of various features including rain intensity (rain_dbz), delay information (delay). The severity column was created for the training dataset using pandas by defining the ranges for delays categorized as 'no impact', 'minor delay', 'moderate delay', and 'major delay'. We employed a Random Forest Classifier to model the relationship between rain intensity and delay severity, aiming to predict delay severity based on rain intensity.

We loaded the dataset containing relevant information about rain intensity, delay, and delay severity. Missing values in the rain intensity column (rain_dbz) were filled using the mean of the available data. A scatter plot was generated to visualize the relationship between rain intensity (rain_dbz) and delay severity (severity). We split the dataset into training and testing sets using 80:20 ratio. A Random Forest Classifier with 100 estimators was trained on the training data to predict delay severity based on rain intensity. The model achieved impressive results with an accuracy of 100% on the test set.

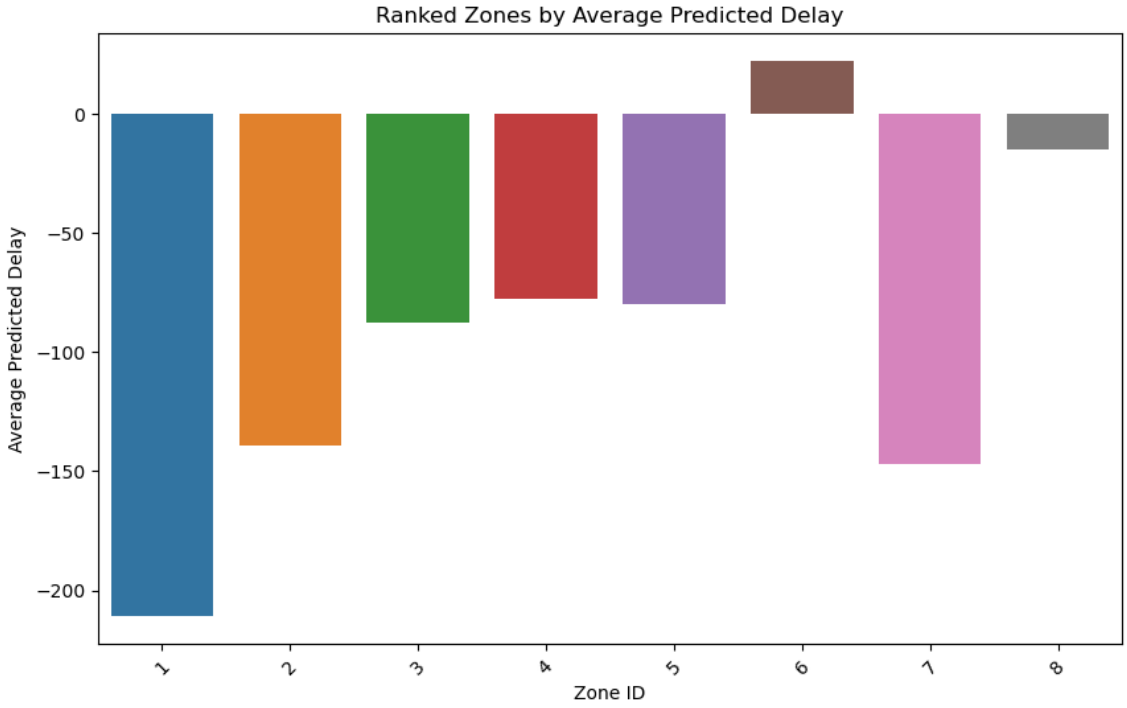


Figure 17. Zone ranks based on the predicted delay

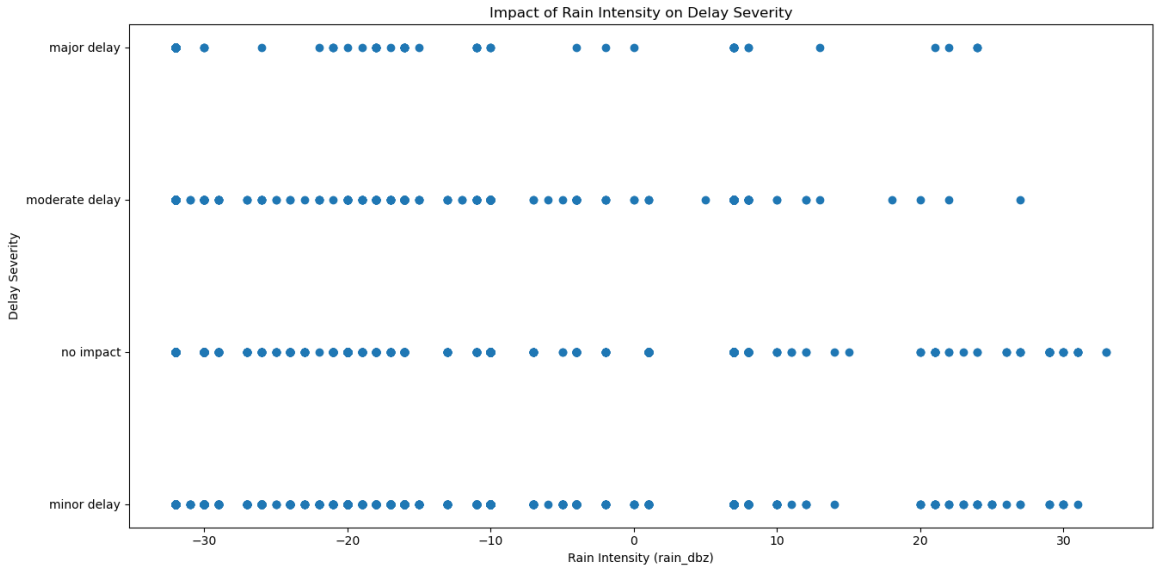


Figure 18. Impact of rain intensity on delay severity

Based on the above scatter plot in Figure 18 it is observed that, there are delays on the days when it did not rain and the number of delays on the non-rainy days are more when compared to the rainy days. Hence. Our findings lead

us to conclude that rain does impact the delay, although its significance is not very pronounced.

```
[[  125    0    1    0]
 [    0 3216    0    0]
 [    0    0  928    0]
 [    0    0    0 3172]]
```

	precision	recall	f1-score	support
major delay	1.00	0.99	1.00	126
minor delay	1.00	1.00	1.00	3216
moderate delay	1.00	1.00	1.00	928
no impact	1.00	1.00	1.00	3172
accuracy			1.00	7442
macro avg	1.00	1.00	1.00	7442
weighted avg	1.00	1.00	1.00	7442

Figure 19. Random Forest results

The analysis demonstrates a strong relationship between rain intensity and delay severity. The results obtained from Random Forest Classifier confusion matrix and classification report in Figure 19 show excellent performance metrics for the model's predictions across different delay categories. Confusion Matrix Analysis shows that for major delay, the model correctly predicted 125 instances out of 126 total instances with a precision of 1.00 and a recall of 0.99. The weighted average precision, recall, and F1-score are also 1.00, suggesting that the model performs consistently well across different classes, considering the class distribution.

In conclusion, the model shows good performance with perfect or near-perfect precision, recall, and F1-scores across all delay categories, resulting in an overall accuracy of 100%. These results indicate that the model is highly effective in accurately predicting delay categories based on the input data.

We also performed an analysis using random forest classifier to study the actual delay versus the predicted delay. From Figure 20 we can observe that, there are many overlaps between the actual and predicted delays which means that the model has performed well in predicting the delays correctly.

To validate the results obtained from the Random Forest model, we also employed Support Vector Machines (SVM) and included its findings in the report. The objective of this analysis was same as Random Forest model to explore the relationship between rain intensity and delay severity in a transportation dataset using Support Vector Machine (SVM) classification. The dataset was split into training and testing sets using a 80:20 ratio. An SVM classifier with a linear kernel was trained on the training data to predict delay severity based on rain intensity and delay features. The model achieved perfect accuracy (100%) on the test set, as evidenced by the confusion matrix and classification report in Figure 21

A scatter plot in Figure 20 was generated to visualize the actual delay severities ('y_test') versus the predicted delay severities ('y_pred') using SVM classification. Blue dots represent the actual delay severities, while red dots represent the predicted delay severities based on rain intensity and delay features. The plot confirms the accuracy of the SVM model in predicting delay severities.

The model's perfect accuracy across all severity levels indicates its effectiveness in categorizing delays, which can be valuable for transportation planning and management. Despite achieving 100% accuracy, further evaluation and validation are recommended to investigate potential biases in the dataset or model that could lead to overfitting or inflated accuracy.

To summarize, while the Random Forest and SVM model's 100% accuracy is impressive, it is imperative to conduct a thorough evaluation encompassing data quality, model complexity, generalizability, and additional metrics to ensure its reliability and applicability in real-world scenarios. Further validation and scrutiny will enhance our confidence in the model's predictive capabilities and decision-making usefulness.

4.2. Regression

4.2.1 Data Sampling

Due to the substantial size of the dataset, exceeding two million entries, we opted to employ a simple random sampling technique to manage the data volume efficiently.

To gain insights, we randomly selected data for a single day and created scatter plots depicting delay times for each type of transportation at ten-minute intervals within each hour. The provided Figure 22 illustrates the variation observed between 10:00 AM and 11:00 AM.

Upon examining the scatter plot, it becomes apparent that the distribution of data points exhibits a high degree of similarity, suggesting minimal variation in delay times for the same mode of transport over brief time intervals. Consequently, we determined that sampling data from each transport on an hourly basis would be appropriate for subsequent analysis.

4.2.2 Linear Regression

Among all the captured data, we specifically chose 'route_type', 'rain_dbz', 'lat', 'lon', and 'hour_12' as variables for studying the delay times. And these variables were selected based on several considerations:

- Route Type ('route_type'): Different transportation chosen may experience varying operational conditions and traffic situations, thus impacting delays differently.
- Rainfall Intensity ('rain_dbz'): Rainfall can be a significant factor influencing transportation delays due to

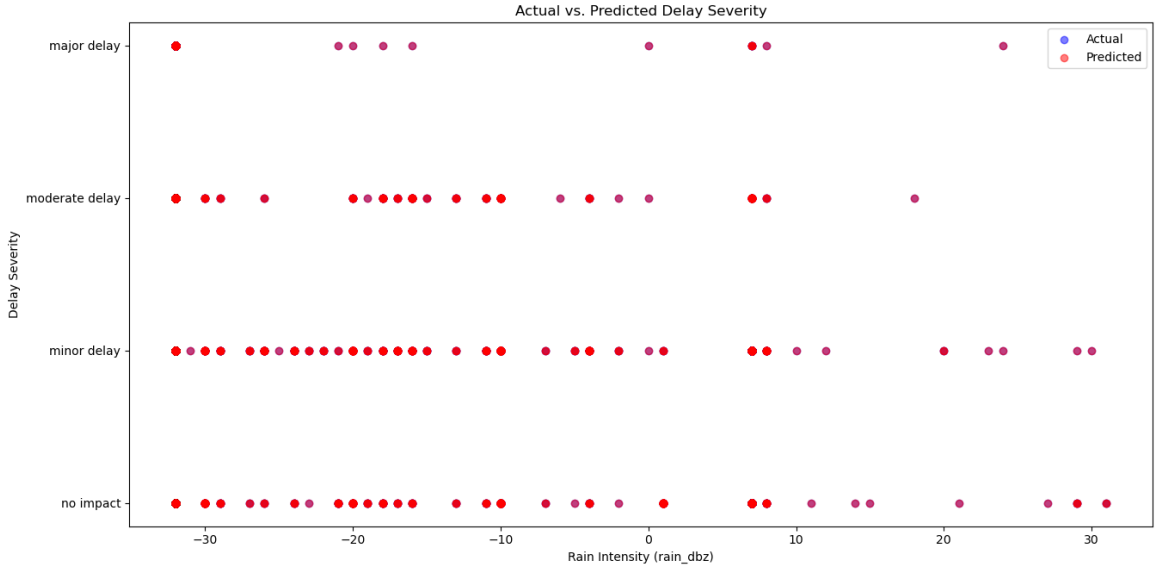


Figure 20. Scatter plot to compare actual and predicted delay severity

```
[[  126     0     0     0]
 [    0  3216     0     0]
 [    0     0   928     0]
 [    0     0     0  3172]]
```

	precision	recall	f1-score	support
major delay	1.00	1.00	1.00	126
minor delay	1.00	1.00	1.00	3216
moderate delay	1.00	1.00	1.00	928
no impact	1.00	1.00	1.00	3172
accuracy			1.00	7442
macro avg	1.00	1.00	1.00	7442
weighted avg	1.00	1.00	1.00	7442

Figure 21. Results of SVM model

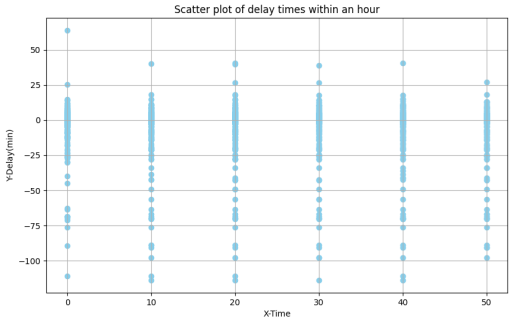


Figure 22. Scatter plot of delay times within an hour

adverse weather conditions leading to traffic congestion and operational instability.

- Geographical Location ('lat' and 'lon'): The geographical location of transportation may be associated with delays, as specific areas might face distinct traffic conditions or road situations.

- Time of Day ('hour_12'): Time is another important factor influencing delays of transportation vehicles, as delays are often more pronounced during peak commuting hours. Therefore, the model also considers the specific times when transportation vehicles are in operation.

By integrating these variables, our objective is to gain comprehensive insights into the factors contributing to transportation delays, facilitating valuable predictions and management strategies for delay mitigation.

Figure 23 presents the fitted results of our linear regression model. The R-squared value of 0.004 indicates a very limited degree of explanatory power, suggesting that the current set of variables may be insufficient to account for the variability in delay times, possibly due to issues such as multicollinearity or other factors. Moving forward, it will be essential to incorporate additional influential factors to enhance the comprehensiveness of our model.

The positive correlation observed between delay time and longitude and latitude suggests that delays are more prevalent in the southern coastal regions of Queensland.

OLS Regression Results

Dep. Variable:	0	R-squared:	0.004
Model:	OLS	Adj. R-squared:	0.004
Method:	Least Squares	F-statistic:	346.5
Date:	Thu, 02 May 2024	Prob (F-statistic):	0.00
Time:	09:55:35	Log-Likelihood:	-1.8226e+06
No. Observations:	419785	AIC:	3.645e+06
Df Residuals:	419779	BIC:	3.645e+06
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-852.1246	30.134	-28.278	0.000	-911.186	-793.063
route_type	-1.1068	0.047	-23.707	0.000	-1.198	-1.015
rain_dbz	-0.0023	0.002	-1.372	0.170	-0.006	0.001
lat	0.6033	0.098	6.132	0.000	0.410	0.796
lon	5.6839	0.206	27.528	0.000	5.279	6.089
hour_12	-0.1421	0.008	-18.329	0.000	-0.157	-0.127

Omnibus:	845809.577	Durbin-Watson:	1.491
Prob(Omnibus):	0.000	Jarque-Bera (JB):	14977516806.581
Skew:	-15.887	Prob(JB):	0.00
Kurtosis:	927.817	Cond. No.	1.65e+05

Figure 23. The summary result of linear regression

Conversely, other variables demonstrate an inverse relationship with delay time, indicating that delays are most frequently observed in buses, while trains experience minimal delays. Moreover, delays appear to be less frequent around noon and midnight.

Notably, the p-value for the rainfall variable exceeds 0.05, indicating that rainfall does not have a statistically significant impact on delay time within the context of this model. This unexpected outcome underscores the need for further investigation to explore whether unexplained variables or interactions can provide better insights into delay variations and inform more effective reporting and management strategies.

4.2.3 KNN Regression

Due to the insignificant fit of the linear regression model, we opted to employ KNN (K-Nearest Neighbors) regression analysis for improved data prediction. As shown in Figure 24, the model's outcome did not exhibit a perfect diagonal line. However, Figure 26 illustrates that the data segregated into four distinct sections based on the type of transportation, each showcasing its unique delay time patterns, which is logically sound. It is noteworthy that trams and trains experience the shortest delay times, while buses consistently show delays of 200 seconds.

The results of the KNN regression analysis demonstrate significant differences in delay patterns across each type of transportation, facilitating more precise predictions of delay situations for different modes of transport. Further analysis using KNN regression can explore nonlinear relationships within the data and interactions between different modes of transportation, providing deeper insights and predictive capabilities. Through this approach, we can effectively devise travel schedules and optimize transportation strategies to reduce delays and enhance efficiency.

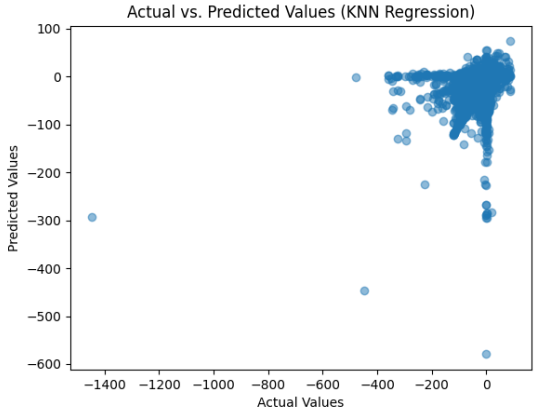


Figure 24. The model of KNN regression

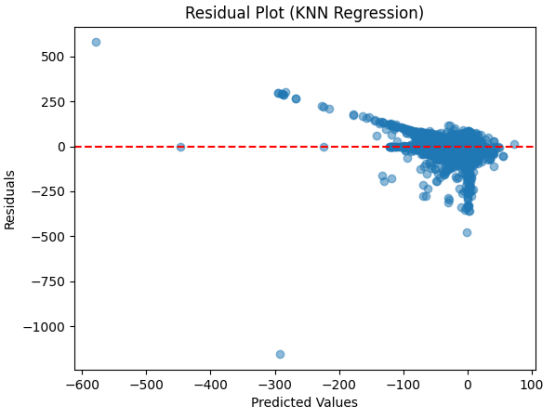


Figure 25. The residual plot of KNN regression

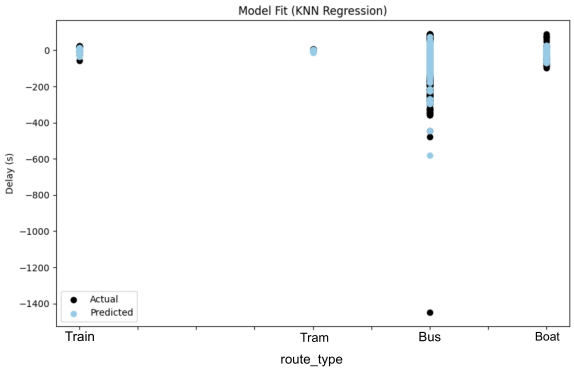


Figure 26. The prediction of data classified by KNN regression

5. Storytelling

For the storytelling session, we analyzed the delay in each type of public transport to see which type of public transport is most likely to be affected by other factor variables and can cause more delay if compared with others.

So we use a pie chart to divide public transport into four charts. Consists of boat delays, Bus delays, Tram delays, and Train delays to see the percentage of public transport that has delays less than 60 seconds, more than 60 seconds, more than 180 seconds, and more than 300 seconds as Figure 27.

As the result from the pie chart Figure 27a. The train is the public transport that has the least delay with 1% of the tram that delays more than 300 seconds and only 3% have delays of more than 180 seconds and less than 300 seconds.

On the other hand bus is the public transport that has the highest delay rate if compared with others, with 13% of the bus that have delays of more than 300 seconds and 15% delays of more than 180 and less than 300 seconds.

The other two public transport which are boat and train have similar rates of delay. So what can we say from the data that we get is bus is the public transport that most likely to be affected by other factor or variable to cause the delay.

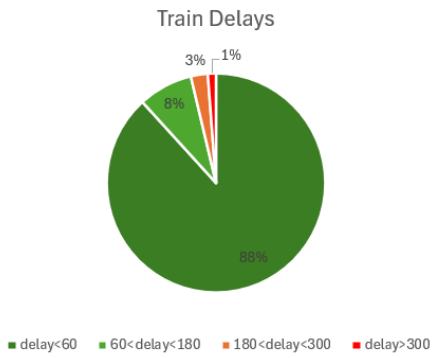
6. Conclusion

In this project, we introduced a methodology to observe and analyze the correlation between precipitation and delays within the Translink network operating in South-East Queensland, Australia. From the research, we also found the delays are mostly caused by buses during peak hours of the day and Non-delay prediction was found to be performing well during morning and evening hours, whereas the delay prediction was performing well during midday hours. Also, No significant delays were caused by ferries and trams and as you can see Zone 6 is predicted to have the highest number of delays the last thing is Rain did contribute to the delays, but the impact was not significant.

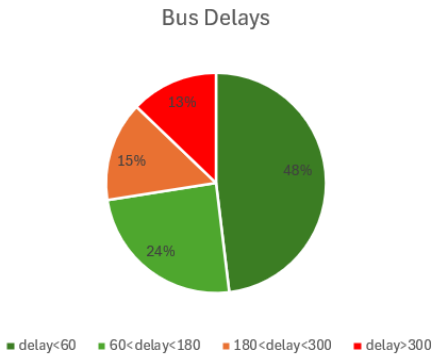
6.1. Future work

Our framework can be extended to any other city worldwide, that provides the used GTFS-RT interface [1]. Future work might compare different regions or countries by their abilities to keep public transport operating in a timely manner. This might also include statistical analysis about the robustness of whole public transport systems against weather and other types of precipitation as the RainViewer API also enables one to capture the current level of snowfall or cloud coverage.

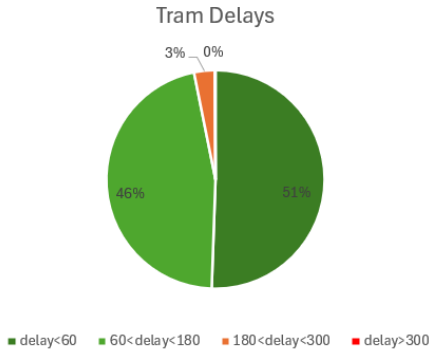
Going forward, any data capture should be done on a more reliable and robust cloud platform such as Amazon's EC2 [2] platform that would allow for the 24/7 monitoring



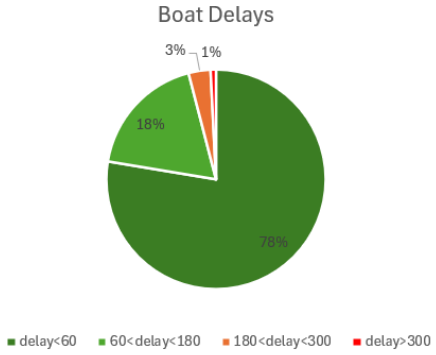
(a) Train Delays



(b) Bus Delays



(c) Tram



(d) Boat Delays

Figure 27. Pie Charts

of a transport network and more robust data analytics performed on high-performance clusters.

6.2. Responding to Peer Review Feedback

Regarding responding to peer review feedback to improve our project, we have given more explanation based on the research questions for example when the delay is happening and whether weather plays a role. Also in the problem statements we have changed from 'Whether a special event causes additional delays?' to 'Does weather play a role' based on the proposal directly on the weather.

Another thing is in the storytelling part, we provide more visualization for the audience to easier to see and understand in terms of a pie chart showing the percentage of delay rate in each transport and you can see which type of transport has the best or worst delay rate. We also added some visual graphs that show the delay prediction in each zone into our report to show the prediction result.

References

[1] Gtfs: Making public transit data universally accessible. https://gtfs.org/.

[2] Amazon Web Services. Machine learning on AWS, 2024. Accessed: 2024-05-08.

[3] H. Butler, M. Daly, A. Doyle, Sean Gillies, T. Schaub, and Stefan Hagen. The GeoJSON Format. RFC 7946, Aug. 2016.

[4] CARTO . Basemaps . https://carto.com/basemaps, 2024.

[5] Inc. Environmental Systems Research Institute. *ESRI Shapefile Technical Description*, 1998.

[6] GDAL/OGR contributors. *GDAL/OGR Geospatial Data Abstraction software Library*. Open Source Geospatial Foundation, 2020.

[7] Queensland Government. Feb24 - qld - localities - esri shapefiles - gda2020. https://data.gov.au/dataset/ds-dga-6bedcb55-1b1f-457b-b092-58e88952e9f0/distribution/dist-dga-5453ebd9-58f9-462c-a086-1f4cc883baf9/details?q=.

[8] Queensland Government. Translink gtfs real-time feed. https://translink.com.au/about-translink/open-data/gtfs-rt.

[9] OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org . https://www.openstreetmap.org, 2017.

[10] MeteoLab Inc. RainViewer. Weather maps api. https://www.rainviewer.com/api/weather-maps-api.html.

[11] N. Ritter and M. Ruth. The geotiff data interchange standard for raster geographic images. *International Journal of Remote Sensing*, 18(7):1637–1647, 1997.

[12] Translink . Translink Zone Map. https://translink.widen.net/s/hvpvz8n9dn/230109-seq-fare-zone.

7. Acknowledgements

In the following we detail our individual contributions to this project. More insights are given by our Github repository[3] and commit history:

[3] github.com/JohannesVolk/data7001-project

Type	Description	Name
Report	Abstract	Johannes
Report	Problem Solving	Johannes
Report	Getting data we need (sections 2., 2.1, 2.2, 2.4,2.5)	Johannes
Report	Section 2.3 Data Collection	Snehin
Report	Future work	Johannes & Snehin
Report	Flowchart Figure 1	Snehin
Report	Delay Choropleths, Histograms & Boxplots	Johannes
Report	Section 3.1 - Data Characteristics	Snehin
Report	Section 3.2.2 - Analysis of Factors Contributing to Delay	Lasya
Report	Section 3.3 - Analysis of delay based on the zones	Lasya
Report	Section 3.4 - Outliers	Snehin
Report	Making Data Confess Introduction	Lasya
Report	Data Sampling	Zixhuan
Report	Linear Regression	Zixhuan
Report	KNN Regression	Zixhuan
Report	Section 4.1 - Classification	Lasya
Report	Storytelling	Boonyapat
Report	Responding to Peer Review Feedback	Boonyapat
Report	Data Cleaning	Erik
Code	Data collection (includes georeferencing, API polling etc.)	Johannes
Code	Data collection Server	Snehin
Code	Data visualization (plots + webserver)	Johannes
Code	Classification	Lasya
Code	Regression	Zixhuan
Code	Classification & Data Cleaning	Erik
Project Pitch	Slides	Snehin (full version) & Johannes (draft)
Project Pitch	Video	Snehin
Trial Presentation	Slides 1-6 & advising on the rest	Snehin
Trial Presentation	Slides 7-10 & advising on the rest	Johannes
Trial Presentation	Slides 18-20	Zixhuan
Trial Presentation	Slides 21-22	Boonyapat
Trial Presentation	Slides 23 & 25	Lasya
Final Presentation	Slides 7-12	Johannes
Final Presentation	Slides 18-21	Lasya
Trial Presentation	Slides 23-25	Zixhuan
Final Presentation	Slides 22 & 27	Boonyapat
Final Presentation	Slides 13 - 16 & 25	Erik
Final Presentation	Slide 27	Lasya
Project idea	Idea & Concept	Johannes
Project coordination	Meetings and distributing work	Johannes & Snehin

Table 6. List of Contributions