Data Analytics of public transport

Contents

[ABSTRACT 3](#_Toc55166869)

[INTRODUCTION 4](#_Toc55166870)

[LITERATURE REVIEW 5](#_Toc55166871)

[LITERATURE REVIEW DISCUSSION 10](#_Toc55166873)

[METHODOLOGY 13](#_Toc55166874)

[DATA COLLECTION 13](#_Toc55166875)

[DATA PRE-PROCESSING 14](#_Toc55166876)

[DATA CLEANING 14](#_Toc55166877)

[TIME SERIES 16](#_Toc55166878)

[ARIMA MODEL 16](#_Toc55166879)

[OTHER MODELS 17](#_Toc55166880)

[ACCURACY TESTING 18](#_Toc55166881)

[EXPLORATORY DATA ANALYSIS 19](#_Toc55166882)

[RESULTS 22](#_Toc55166883)

[ARIMA 22](#_Toc55166884)

[Accuracy calculation results 27](#_Toc55166885)

[DISCUSSION AND CONCLUSION 28](#_Toc55166886)

[REFERENCES 29](#_Toc55166887)

[APPENDIX 32](#_Toc55166888)

[API REQUEST 32](#_Toc55166889)

[HEATMAPS 32](#_Toc55166890)

[DATA EXCERPT 35](#_Toc55166891)

# ABSTRACT

Despite the fact that punctuality is an advantage of rail transportation compared to other long-distance transports, there are frequent train delays. Railway networks in Sydney are facing challenges in avoiding train delays. This means that the travellers may be delayed for their specific destination.

The aim of this study is to analyse and predict the train delay using time series data and machine learning models. Our study goes through the following stages, Data collection, data cleaning, Data pre-processing, testing different models and then Evaluating.

A preliminary literature review was carried out and papers were analysed on the basis of pre-determined features; specifically, the machine learning techniques used, the structure of dataset targeted, and the methodologies used. We collected data from NSW open data hub and carried out ARIMA along with other models as benchmarks. The results obtained shows that ARIMA model has the lowest RMSE value 44.48377 for arrival dataset and 40.4671 for departure dataset. Hence, ARIMA performed better than benchmark models.

**Keywords**: Machine learning, ARIMA model, time-series, train delay, alerts, neural network, prediction, MAE, RMSE

# INTRODUCTION

There is so much of research going on in field of applications of artificial intelligence in transport fields, including predicting train delays. Many different prediction models have been used to predict train delay using datasets from TfNSW Open Data Hub. This report examines the work done so far in past in the field of AI in transport and identifies the limitations. It also presents a comparison between different prediction models.

In 2019 the punctuality rate has been very poor due to extreme weather conditions like bush fire. Many trains were diverted and cancelled at the end moment as there was no method of predicting the weather conditions and automatically alerting the railway authorities. Suddenly in February 2020, the punctuality rate drastically fell down due to the COVID-19 pandemic. Trains were cancelled without alerting the travellers. Restriction on number of people using the compartments reduced thus increasing waiting time for daily commuters. Government asked people to stay indoors so, many people stopped travelling. Frequency of trains came down as more than half of trains were empty during off-peak times. The entire architecture was replanned. Recently, situation is coming back to normal with punctuality rate of 97.9%. We need a systematic approach to predict the train delays and alerts the passengers.

While a lot of different models have been used in the articles in literature review, (Maini, 2020) is the only article we found that used ARIMA to forecast delays. We intend to further explore this and see if ARIMA performs better than benchmark models.

The rest of this report is organized in the following manner. Section 1 is the literature review. Section 2 is the methodology. Section 3 is exploratory data analysis. Section 4 is the results of ARIMA Model and benchmark models. Section 5 is discussion of results and conclusion.

# LITERATURE REVIEW

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Year of publication | Reference/Authors | Methodology/tools | Attributes Used/ conclusion | Research Gap/ Future work |
| 2012 | (Büker & Seybold 2012) | Monte carlo simulation | Europtirails /TIS the expected time-of-arrival (ETA) of trains passing the Brenner-corridor | The process for estimating the rescheduling decisions could be improved by applying the delay propagation mapping method |
| 2018 | (Gaurav & Srivastava 2018) | Random Forest Regressors (RFRs) and Ridge Regressors (RRs) as two types of individual regression models in N-OMLMPF | India train running status information (Train Data) over the period of March 2016 to February 2018. | Recurrent Neural Networks (RNN) have the tendency to learn past details and predict the next state. The dynamic nature of predictions would require an online learning algorithm to constantly learn the changing behaviour of railway network and delays. |
| 2019 | (Jiang, C et al. 2019) | **random** **forest** **regression** (RFR) model, **multiple** **linear** **regression** (MLR), support **vector** **machine** (SVM), and **artificial** **neural** **networks** (ANN) | Train operation records from Wuhan-Guangzhou (W-G) **high**-**speed** **railway**. | The primary and indirect delays needs to be combined into an cohesive recovery model to improve High Speed Rail(HSR) service operations planning |
| 2019 | (Jiang, S et al. 2019) | Semi-parametric model, linear models, Weibull distributions, Binomial logistic regression and Random Forest. | Sweden Railway data | The semi-parametric model, which integrates standard logistic regression and Random Forest algorithm, is not widely used in analysing railway delay in Sweden and is expected to be more developed |
| 2019 | (Jianqing et al.) | Matrix factorization, Multiview learning | Literature review | Further work on matrix factorization. |
| 2018 | (Oneto et al. 2018) | Big data analytics Extreme learning machines Shallow architecture Deep architecture | TM data provided by RFI | External information from outside sources such as weather information, traveller information from tourism databases, railway status or any data sources that may affect rail transit operations, and needs to be considered. |
| 2019 | (Wang & Zhang 2019) | boosted regression trees, DBSCAN algorithm, gradient-boosted regression trees model (GBRT) | For this study, a three-month dataset of weather, train delay and train schedule records were collected. Railway Customer Service Centre of China on 28 December 2017. The geographical information system (GIS) data for 2761 railway stations | Data limitations due to lack of recent data used in the modelling, and other factors contribute to the big difference between delay prediction and actual times. |
| 2020 | (Wen et al. 2020) | Random forest model, artificial neural network model, LSTM model | Running data of the Dutch railway Rotterdam Central to Dordrecht section | Several train delay influence factors overlooked. The proposed model does not reflect many infrastructure factors |
| 2019 | (Wu et al. 2019) | The novel ConvLSTM Encoder-Decoder model with CPS | Sydney Train GTFS dataset 5 days | Critical Point Search (CPS) will be applied in order to do classification on whole train network data. |
| 2013 | (Yaghini et al. 2013) | Artificial Neural Network, decision tree, multinomial logistic regression | monthly files from 2005 to end of 2009 are used from Iranian Railway | The accuracy of model can be improved by metaheuristic methods such as genetic algorithms or simulated annealing or hybrid algorithms. Training time can be improved through methods like particle swarm optimization or continuous ant colony optimization |
| 2010 | (Suwardo et al. 2010) | ARIMA model | Time series data of bus travel time were collected starting from January to December 2007 for two typical days a month, workday and weekend. | The results show that the travel average model, MA (2) and MA (1) are close to the observed observation values using the moving average model. They are indicated by the values of MARE and MAPPE. The expected moving average models have lower MARE and MAPPE values than other temporal ARMA models. |
| 2017 | (Gao et all. 2017) | autoregressive integrated moving average (ARIMA) and artificial neural network (ANN) | The half hourly data of historical prices are obtained from UK Reference Price Data from March 22nd to July 14th 2010 and the predictions are derived from a sliding training window with a length of 8 weeks. | The ARIMA and ANN models were compared for electricity price prediction. ARIMA model outperformed in this study. |
| 2020 | (Xu, et al. 2020) | Autoregressive Integrated Moving Average (ARIMA) model, Long Short-Term Memory (LSTM) model | The data set used in this experiment is from the dam monitoring data of Manwan Hydropower Station, and the typical positive vertical line data is selected as the dam deformation data set. A total of 6884 dam deformation data from June 12, 1999 to April 16, 2018 was selected. |  |
| 2019 | (Widiyaningtyas et al. 2019) | ARIMA | data from IRC Malang City which contains data on passengers from1 August 2014 to 31 October 2017 | Prediction accuracy can be improved with more data than 1188 data used in this study |
| 2018 | (Marc & Thomas 2018) | (AR(MA((X)–)GARCH-type model.  Naïve forecasts | Empirical forecasting literature reviews on electricity spot markets.  86 articles | This is the first article to apply a meta-analysis methodology to a literature review of the literature on empirical forecasts in the spot energy market. |

## Table.1. Literature Review

## LITERATURE REVIEW DISCUSSION

Machine learning is a field that fall categorically under artificial intelligence. Several approaches existing within machine learning rely on algorithms and models. In this review, we will look at some of these specific techniques. ARIMA models are popular model for time series data, and some type of ARIMA model was used in six out of ten studies. There is a summary of machine learning methods used in column 4. Only key articles have been tabulated.

Delay prediction in transport is a vast field. There are implications in quality control, customer satisfaction, profit generation and business intelligence. There have been investigative studies done on public and private transport systems including buses, trains, ships and taxi services. The most common is data on trains as they are widespread in populous metropolitan cities across the world. This data is used to make trip planning apps as well as decide routes and frequency.

Other studies were also found and included especially those dealing with time series data. An article predicted volume of passengers on trains with ARIMA model on time series data. A key meta-analysis done on 86 articles tabulated the performance of several models on time series data involving electricity demand forecast.

The meta-analysis yielded several important observations that we took into consideration while choosing our models. Firstly, according to study conducted by Marc and thomas, forecasts of a hybrid (AR(MA((X)–)GARCH-type models outdo their AR(MA)(X) equivalents. ARMA(X) models, on the other hand, sustain improved predictions than AR(X) models. Merging ARMA or GARCH models with mode switching methods and threshold models does not yield improved prediction results (Marc & Thomas, 2018).

Secondly, this literature review concluded that the prediction from Neural Network does not outperform that from time series models(Marc & Thomas 2018).

Thirdly, ARMA approaches that are refined and contain diversity in their attuning process, are more preferred than simple ARMA models with predicting improvements by over 5 per cent(Marc & Thomas 2018). Combining the models will help to achieve different modes of time series approach. These model types can combine the advantages of several individual models. If different predictions are easily found in one prediction, the efficiency increases by 7%. These predictions are based on a number of individual predictions, and the finishing prediction is made using a weighing or selection algorithm.

Hybrid models are another approach to integrate models together, in which predictions from one model serve as inputs to another. The hybrid model is 38% better than the time series model.(Marc & Thomas 2018)

Fourthly, according to (Marc & Thomas 2018) Naïve forecasts, simultaneously show worse performance than time series models with 30 percent forecast accuracy. This result is to be expected since naive forecasts are only used as examples for other models.

The most common data used was historical data as can be seen in the last column of table 1. Historical transport data was used to train models. There were different levels of details in different study and different features were selected to predict delay. Some studies combined historical data with other aspects such as traveller data, weather data, transport utilisation and congestion data to predict delay. Due to time constraints and other limitations we only used historical data in our study, converted to time series format.

Machine learning can complement the services of experts by flagging common seasonal variations in delay and highlighting problem areas for a smoother functioning transport system. Some variation of machine learning is used in many transport apps. However, the diversity and dimensionality of data significantly expands the scope of a machine learning project. In majority of the studies, one machine learning model is seen as performing better than the others. However, the studies also discussed the limitations such as time, memory and CPU power taken to analyse big data and categorical data. In his paper (Ghofrani et al. 2018) discusses some popular use of Big data analytics and machine learning in the field of railway transportation. In some studies the model showed unexpected deviations on test data(Wang & Zhang 2019). All machine learning models are data hungry models. Along with machine learning development in big data analytic techniques will cause better delay prediction.(Kitamura & Deible 2020).

# METHODOLOGY

In this stage, we have analysed and compared different models to find the model which gives most accurate prediction. As shown in Figure 1, we have used several stages to come to the conclusion.

Accuracy testing

Using Prediction model

Analysing different models

Data Pre-processing

Data collection

Figure 1: Different steps during study

## DATA COLLECTION

People living in Sydney prefer trains for their commute to work but train delay is causing them delay in reaching at their destination. Train network authorities are facing difficulties in removing this delay as there is no real time monitoring system. To minimize the delay in trains, firstly we need to predict the future delays which in turn should be then sent to Sydney train authority. First, we collected the Sydney train delay data from the TfNSW Open Data Hub and Developer Portal (https://opendata.transport.nsw.gov.au/). The Open Data Hub contains the transportation data for NSW . In the [Data Catalogue](https://opendata.transport.nsw.gov.au/dataset) we can access all of available APIs and downloadable datasets.

So, we collected two weeks data from the open data platform by accessing the Public Transport - Realtime Trip Updates API using CURL operation. The time period of data collection is from 5th august 2020 to 18th August 2020. The response data is Shown in the figure below.

As shown in figure 2, From response body, we get ample data like details of trip id, schedule relationship, route id, timestamp but the data which is of importance is station ID, destination delay and arrival delay. The station id is a unique identifier of each station in NSW network of stations. The delays are here measured in Unix time (1 Unix time=1 Second). The data collected was in the protobuf format, which was not easy to interpret so, we used python parser to convert the data into CSV format.

Now the data was in columns which was readable for us. Our columns were; station\_id, arrival\_delay and departure\_delay. The code to receive this data from TfNSW open data hub has been attached in Appendix.

Graphical user interface, text, application, email

Description automatically generated

Figure 2: Screenshot of the Curl request

## DATA PRE-PROCESSING

Data pre-processing is a crucial stage in Web usage mining process because the log data can be quite unstructured and heterogeneous in nature. The target of data cleaning is to handle the missing data through the identification of the noises and hence to remove outliers and resolve inconsistencies. This step is performed before the classification process. During this phase several processes are performed including data cleaning, converting coefficients to digital conversions, converting coefficients to date, and checking stationarity. The purpose of data cleansing is to remove missing values and change the data format from .xlsx to .csv (Widiyaningtyas et al., 2019).

## DATA CLEANING

In order to enhance data quality, the data cleaning step is used to delete obsolete entries from the log file that are not useful for research purposes. Data cleaning is typically a procedure unique to the site and entries are deleted depending on the form of analysis. To clean the two weeks data in the CSV format. The data in the CSV data, had some inconsistencies like unwanted strings, redundant data so we used MS Excel to remove them. Using R, we filtered non-zero values of delay for exploratory data analysis. (Chu et al, 2015). After studying the data and doing literature review, we decided to convert data to time series and use ARIMA Models. We would compare the performance of the models using evaluation measures like RMSE.

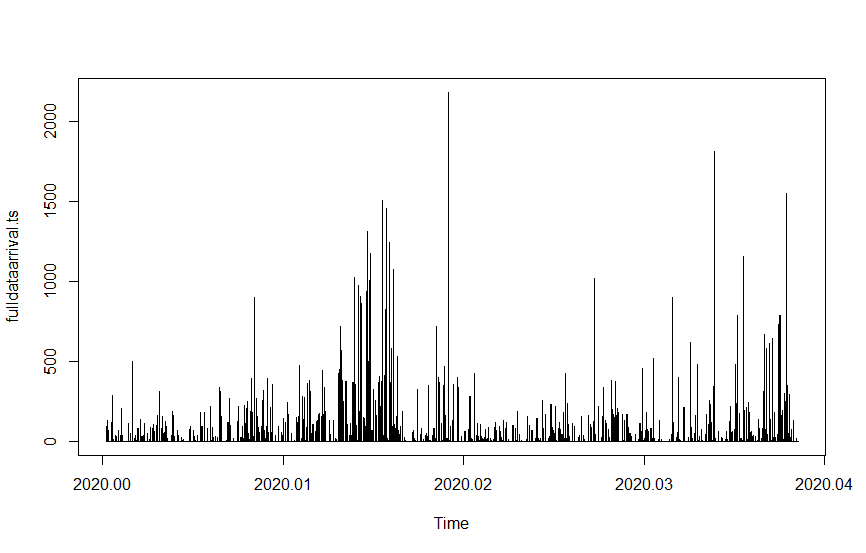
Stationary monitoring is used to insure whether there is a rising or declining data pattern. In this analysis, the stationary test is performed by observing the graph shape of the dataset. The data is not stationary if the graph has a diagonal path, but if the graph is in a straight line, then the data is stationary. As shown in Figure 3, the information shows no general pattern, so it can be applied to the ARIMA model.

Figure 3: Dataset graph

## 

## TIME SERIES

In order to identify the right models for the delay prediction of trains using our dataset, we researched different languages and data formats. When it comes to observing and predicting temporal performance, the time series dataset using R language was the best one to perform our predictive analysis. According to literature review time series models consider trend and seasonality factors and are simple to understand and execute for any time-related datasets.

**What is time series?**

Time-series data emerge naturally when tracking a process over a given period of time. The time-series analysis method has the advantage of taking into account the fact that the observed data over time may have an underlying structure. Therefore, the goal of the time series analysis is to understand this basic method and the driving forces of the observed data, as well as to predict future events. The following notation is used to characterize the previous N observations: y1,...,yN . In addition, we are interested in the importance of progress in the future of time series h. That is, it predicts yN + h. The parameter h is called the horizon because it indicates how much into the future you want to predict.

The observed time series data can be broken down into several potential components: random or trajectory component, trend unit (systematic linear or nonlinear trend in series) or seasonal component (repeating patterns at intervals).

Correlation is an important concept of time series and it has 2 types,

* **Auto-correlation,** whichreflects both *direct* and *indirect* properties
* **Partial Auto-correlation,** which reflects *only the direct properties*

## ARIMA MODEL

The ARIMA approach is capable of forecasting past data with the effect of data that is theoretically difficult to grasp and has a high degree of shortterm estimation and is capable of meeting periodic data variability. The application of ARIMA algorithm is to estimate time series data in multiple areas and deliver strong performance (Ariyo et al, 2014).

The principle of ARIMA model is to predict time series is that the predicted object will generate a data sequence over time. The data that we have used is of train delay in seconds over 2 weeks. We described the sequence by using parameters in the ARIMA model. At the point when the model and the sequence are effectively coordinated, the ARIMA can anticipate future succession by consolidating the past data (Xu et al, 2020)( Gao et al, 2017).

In the ARIMA model formula, p is the autoregressive order, et is white noise and μ is the variable. The mathematical formula is shown below

Xt = μ + θ1Xt-1 + θ2Xt-2 +⋯+ θpXt-p + et

ARIMA models is segregated into three divisions, firstly: autoregressive (AR) model, moving average (MA) model, autoregressive moving average (ARMA) model, which consists of properties of the first two models and integration.

After analysis of the literature the ARIMA model was the best fit for our data, and consistently predicted the time series data more accurately.

It should be checked on its assumptions to decide the ARIMA approach that fits the needs, since one great strategy for a scenario is not inherently appropriate for another.

## OTHER MODELS

In the evaluation we will compare different methods, from naive forecasters to the correct ARIMA model. These are naive forecasting approaches, for example using the last observed value of the current time series as a forecast, only as a benchmark to see if ARIMA performs better or worse in comparison.

**Average method:**



It takes average or mean of all the observed values to predict the forecast value. In the average method, time patterns are completely ignored. It predicts the projected value as the average of the observed values so far.

**Naïve method:**



It takes into account the last few observations to predict the forecast value. In naïve method, the last observation is then used as the next indicator and the horizon is not regarded.

**Seasonal naive forecast:**



It is similar to the naïve method, it looks at the last few values to predict the forecast value, within a certain time period or of the last season e.g. April 2018, April 2019 data. It has season-parameter m.

**Drift method:**



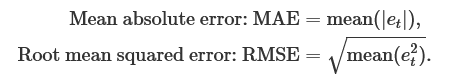
It takes first and last observed values in the dataset and draws a straight line between the two points. This type of model is more suitable for linear trends and not great for seasonal-affected data. The drift method makes it possible to increase or decrease predictions linearly over time, expressing the amount of change over time (drift) as the average change in historical data.

## ACCURACY TESTING

Accuracy testing is carried out to show the discrepancy between actual value and predicted value. There can be four possible outcomes with regards to the measurement results: accurate and precise, accurate but nor precise, inaccurate but precise, plus inaccurate and imprecise (Leeflang et. al 2008). We used two measures-

**Mean Absolute Error (MAE):** The MAE score is the value derived from the average absolute error expected by assigning equal weight to all data with the actual value (Xue et. al 2005). MAE values that are closer to 0 indicate more accuracy, while those close to 1 are less accurate. MAE is selected because the generation of an average error is more intuitive than the entire data, because it gives the entire data the same weight.

**RMSE (Root Mean Squared Error):** It is also an average of all absolute errors except the errors are weighted more heavily.



## EXPLORATORY DATA ANALYSIS

In this section we have analysed the delays by using R programming to make these visualizations. Refer appendix to refer the code used to generate these graphs.

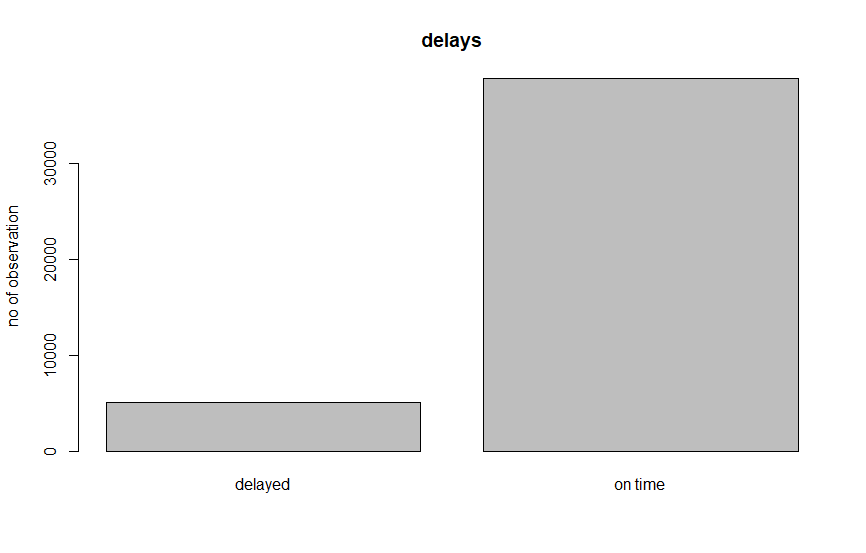


Figure 4: delays

This barplot shows the number of observations over delays. The majority of obervations have zero arrival and departure delay. If an observation had non zero reading in either arrival delay or departure delay or both it was counted in the delay bar. 38840 total readings were extracted over two weeks out of which 5055 readings had delays i.e. 13% of the readings had delays.

Chart, scatter chart

Description automatically generated

Figure 5: Study of arrival and deaprture delays

For this scatterplot, the nonzero delay values were filtered into a dataset. Then the points were plotted with arrival delay on x axis and departure delay on y axis. Red points are when arrival delay was equal to departure delay. Otherwise the points are blue. So this graph shows some blue points below red line and some blue points above red line. There is an important distinction between these. The blue points above the red line have departure delay which is greater than arrival delay. This train is getting progressively more late; secondary delays are being introduced. The blue points below the red line have departure delay which is less than arrival delay. This train is getting closer to being back on schedule with progressively less delay.

For generating the heatmaps the data was divided by days of the week into separate data frames. As shown in figure 6, the complete dataset was spanned across two weeks; thus there was two days of data for each weekday. Then the data was grouped by station-id and average delay calculated for each group for arrival delay and departure delay. This information was then converted to a matrix and the heat maps generated. These heatmaps give an overview of the pattern of delay for each day of the week.

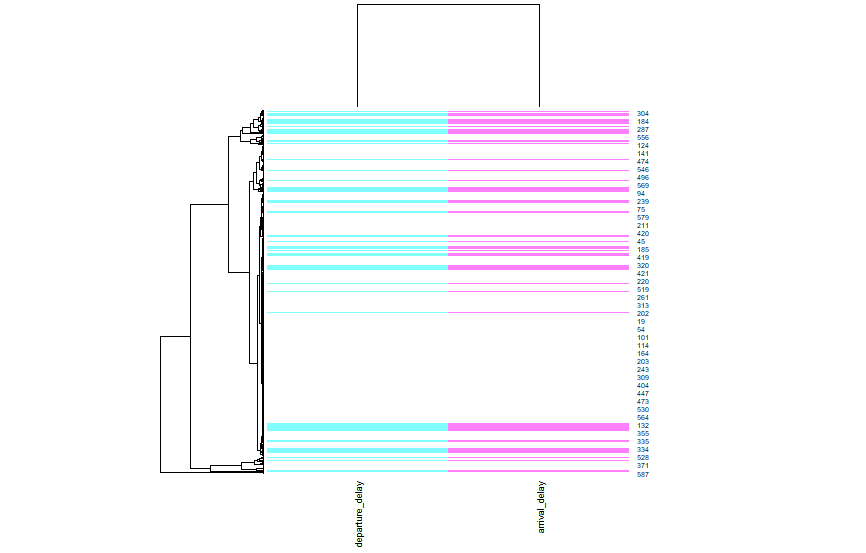


Figure 6: Heatmap of Mondays

Though we would assume that weekends would have less frequent delay than weekdays, data shows this is not true. Wednesday has the least delay affecting the least number of stations, followed by Sunday. Monday, Friday, and Saturday show frequent delays affecting many stations. The heatmaps of rest of days are shown in Appendix.

On Monday, greater than 50% of stations experience delays, more than any other day of the week. This may be due to the transition to work on Monday. On other days, the number of stations with delays is less than 50% of total stations, with Sunday having delays affecting least number of stations.

# RESULTS

## ARIMA

We used auto.arima method from the forecast package in R. This runs several models and outputs the best ARIMA model based on AIC (Akaike Information Criteria) and BIC (Bayesian Information Criterion) values. –

**Arrival delay data**

Data: fulldataarrival.ts (Time series)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model Airma** | MA1 | MA2 | MA3 | MA4 | MA5 |
| **S.E.(standard error)** | 0.0048 | 0.0049 | 0.0051 | 0.0052 | 0.0052 |

Information Criteria

Akaike Information Criteria, AIC= 457745.3

Akaike Information Criteria corrected, AICc= 457745.3

Baynesian Information Criteria, BIC= 457797.4

This shows that 5 ARIMA models were run with different p ,d, q values and the best results were with ARIMA(0,1,5). 0,1 and 5 are P, D and Q values respectively.

|  |  |  |
| --- | --- | --- |
| P | 0 | P is the number of autoregressive terms, |
| D | 1 | D is the number of nonseasonal differences needed for stationarity, and |
| Q | 5 | Q is the number of lagged forecast errors in the prediction equation. |

It is worth noting that these parameters for knowledge do not appear to be good guides for selecting the correct order of differentiation (d) of a model, but rather for selecting the p and q values. This is because the distinction alters the data from which the probability is measured, rendering the AIC values not comparable between models with different differential orders.

ACF Plot

A Correlogram (also called ACF Plot) is a graphical way to illustrate normal distribution in data that varies over time. Autocorrelation is when an error moves to a corresponding point in time at one particular moment in time.

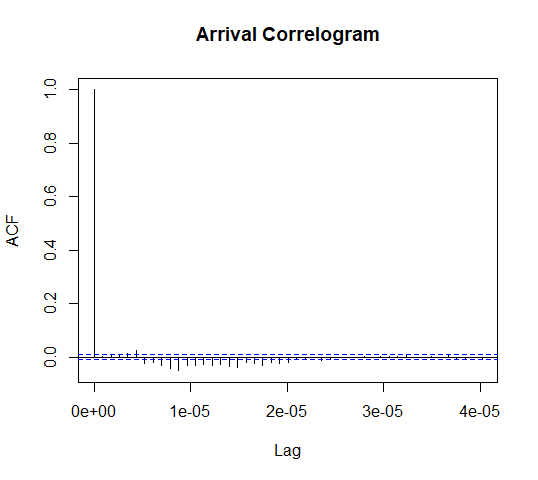


Figure 7: ACF plot

This plot was made with 20 LAG values. The two dotted lines are significant bounds. The auto correlations for in sample forecast errors do exceed significance bounds for the Lag values from 1e-05 to 2e-05. For the rest of the lag the auto correlations don’t.

PACF Plot

You can tentatively define the numbers of AR (p) and/or MA (q) terms expected by looking at the autocorrelation function (ACF) and partial autocorrelation (PACF) plots

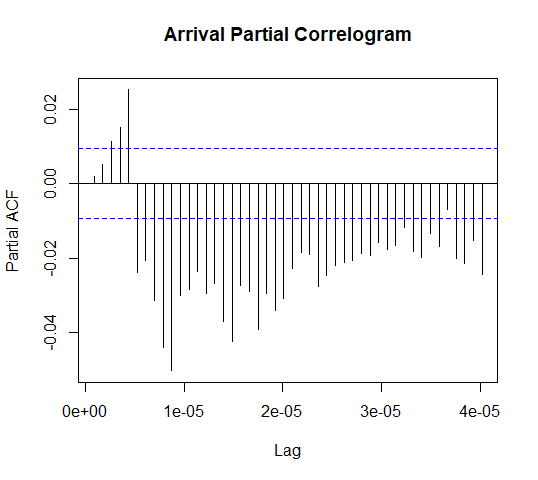


Figure 8: PACF plat

The two dotted lines are significant bounds. The auto correlations for in sample forecast errors do exceed significant bounds for all Lag values.

**Departure delay data**

Model Details

Series: fulldatadep.ts

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model Airma** | MA1 | MA2 | MA3 | MA4 | MA5 |
| **S.E.(standard error)** | 0.0048 | 0.0048 | 0.0052 | 0.0051 | 0.0052 |

Information Criteria

Akaike Information Criteria, AIC= 449437.5

Akaike Information Criteria Corrected, AICc= 449437.5

Baynesian Information Criteria, BIC= 449489.6

This shows that 5 ARIMA models were run with different p, d, q values and the best results were with ARIMA(0,1,5). 0,1 and 5 are p,d and q values respectively.

|  |  |  |
| --- | --- | --- |
| P | 0 | P is the number of autoregressive terms, |
| D | 1 | D is the number of nonseasonal differences needed for stationarity, and |
| Q | 5 | Q is the number of lagged forecast errors in the prediction equation. |

The ARIMA model parameters are the same for departure data as arrival data. This is because the two datasets are almost the same with only small differences within values. This is seen in exploratory data analysis scatter plot where a lot of the values are same for arrival and departure.

It is worth noting that these parameters for knowledge do not appear to be good guides for selecting the correct order of differentiation (d) of a model, but rather for selecting the p and q values. This is because the distinction alters the data from which the probability is measured, rendering the AIC values not comparable between models with different differential orders.

ACF Plot

A Correlogram (also called ACF Plot) is a graphical way to illustrate normal distribution in data that varies over time. Autocorrelation is when an error moves to a corresponding point in time at one particular moment in time.

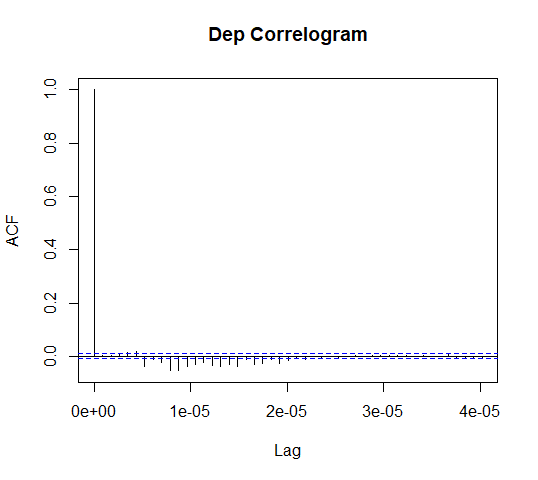


Figure 9: ACF plot

The two dotted lines are significance bounds. The auto correlations for in sample forecast errors do exceed significance bounds for Lag values 1e-05 to 2e-05. For the rest of the lag the auto correlations don’t.

PACF Plot

You can tentatively define the numbers of AR (p) and/or MA (q) terms expected by looking at the autocorrelation function (ACF) and partial autocorrelation (PACF) plots of the distinct sequence.

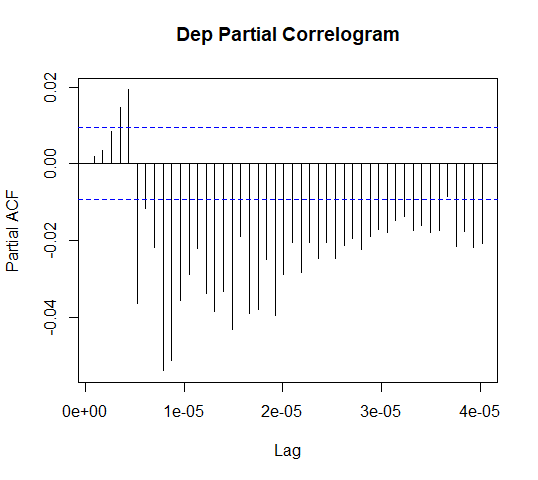


Figure 10: PACF plot

The two dotted lines are significance bounds. The auto correlations for in sample forecast errors do exceed significance bounds for all Lag values.

## Accuracy calculation results

To compare performance of all the four models we calculated evaluation measures for each i.e. RMSE (Root Mean Square Error) and MAE (Mean Absolute Error).

ARRIVAL DELAY

|  |  |  |
| --- | --- | --- |
| **ARR** | **RMSE** | **MAE** |
| Mean | 84.0503 | 28.38403 |
| Drift | 45.421 | 4.364033 |
| Naïve | 45.421 | 4.362031 |
| Arima | 44.48377 | 5.906509 |

DEPARTURE DELAY

|  |  |  |
| --- | --- | --- |
| **DEP** | **RMSE** | **MAE** |
| mean | 84.0503 | 28.38403 |
| drift | 45.421 | 4.364033 |
| naive | 45.421 | 4.362031 |
| ARIMA | 40.4671 | 4.959584 |

ARIMA model has the lowest RMSE value 44.48377 for arrival dataset and 40.4671 for departure dataset. Thus, it has the best forecasting ability out of all 4 models. Naïve method has the lowest MAE at the third decimal place, however drift, naïve and ARIMA all have MAE that are close together.

# DISCUSSION AND CONCLUSION

ARIMA model is used to predict a value changing over time. It is used to predict fluctuations in prices over time, number of passengers over time etc. However, ARIMA has not been widely applied to forecast delays. (Maini, 2020) is the only article we found that used ARIMA to forecast delays. In this article the author acquired flight data and conducted 2 Arima models on it. He then compared the two model’s performance using accuracy measures like RMSE.

We decided to expand on using ARIMA for forecasting delays. A model’s performance is compared against benchmarks like average method to see if the predictions are more accurate than simply taking the average of all observations. Another benchmark is to see if the model’s predictions are more accurate than simply taking the last value recorded as the forecasted values. We tested the performance of ARIMA against these. This is the contribution from our research.

ARIMA performed better than the other models. The evaluation measures reflected that with difference between Mean RMSE(84) and ARIMA RMSE(40) being over 40 for departure ARIMA and 40 for arrival ARIMA. The difference between drift and naïve and ARIMA was 5 points for departure ARIMA. The ARIMA model is good for forecasting data which does not have seasonal variations and has stationarity. These features are present in our data. This margin can be increased by fine- tuning the model.

In future we could try to filter data so that only delays greater than zero are in the dataset and see how the accuracy is affected. We could also try to take data over a longer period of time and see if it improves accuracy. All models are data hungry and show some improvement with large amount of data. We can also attempt to cluster similar data by weekdays and see if accuracy improves. There is a lot of potential in this field.

# REFERENCES

Rogge-Solti, A., Vana, L. and Mendling, J., 2015, December. Time series petri net models-enrichment and prediction. CEUR Workshop Proceedings.

Hyndman, R.J. and Athanasopoulos, G., 2014. OTexts. com. Forecasting: principles and practice/Rob J Hyndman and George Athanasopoulos. OTexts. com [Heathmont, Victoria], print edition. edition, 2014.

Gao, G., Lo, K. and Fan, F., 2017. Comparison of ARIMA and ANN models used in electricity price forecasting for power market. Energy and Power Engineering, 9(4B), pp.120-126.

Xu, G., Jing, Z., Mao, Y. and Su, X., 2020, August. A dam deformation prediction model based on ARIMA-LSTM. In 2020 IEEE Sixth International Conference on Big Data Computing Service and Applications (BigDataService) (pp. 205-211). IEEE.

Ariyo, A.A., Adewumi, A.O. and Ayo, C.K., 2014, March. Stock price prediction using the ARIMA model. In 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation (pp. 106-112). IEEE.

Chu, Y.C., Hsu, C.C., Lee, C.J. and Tsai, Y.T., 2015, October. Automatic data extraction of websites using data path matching and alignment. In 2015 Fifth International Conference on Digital Information Processing and Communications (ICDIPC) (pp. 60-64). IEEE.

Kumar, V. and Khosla, C., 2018, January. Data Cleaning-A thorough analysis and survey on unstructured data. In 2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 305-309). IEEE.

Büker, T & Seybold, B 2012, ‘Stochastic modelling of delay propagation in large networks’, Journal of Rail Transport Planning & Management, vol. 2, no. 1-2, pp. 34-50.

Gaurav, R & Srivastava, B 2018, ‘Estimating train delays in a large rail network using a zero shot markov model’, in 2018 21st International Conference on Intelligent Transportation Systems (ITSC), pp. 1221-6.

Ghofrani, F, He, Q, Goverde, RMP & Liu, X 2018, ‘Recent applications of big data analytics in railway transportation systems: A survey’, Transportation Research Part C, vol. 90, pp. 226-46.

Jiang, C, Huang, P, Lessan, J, Fu, L & Wen, C 2019, ‘Forecasting primary delay recovery of high-speed railway using multiple linear regression, supporting vector machine, artificial neural network, and random forest regression1’, Canadian Journal of Civil Engineering, vol. 46, no. 5, pp. 353-63.

Jiang, S, Persson, C & Akesson, J 2019, Punctuality prediction: combined probability approach and random forest modelling with railway delay statistics in Sweden\*, IEEE, 978-1-5386-7024-8978-1-5386-7023-1, Conference, <http://ezproxy.uow.edu.au/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=edseee&AN=edseee.8916892&site=eds-live>.

Jianqing, W, Luping, Z, Chen, C, Jun, S, Sim Kim, L & Jianming, Y ‘Data Fusion for MaaS: Opportunities and Challenges’, in 2018 IEEE 22nd International Conference on Computer Supported Cooperative Work in Design (CSCWD). Proceedings, Place of Publication: Piscataway, NJ, USA; Nanjing, China. Country of Publication: USA.

Kitamura, G & Deible, C 2020, ‘Retraining an open-source pneumothorax detecting machine learning algorithm for improved performance to medical images’, Clinical Imaging, vol. 61, pp. 15-9.

Marc, G & Thomas, P 2018, ‘Forecasting performance of time series models on electricity spot markets: a quasi-meta-analysis’, International Journal of Energy Sector Management, vol. 12, no. 1, pp. 103-29.

Oneto, L, Fumeo, E, Clerico, G, Canepa, R, Papa, F, Dambra, C, Mazzino, N & Anguita, D 2018, ‘Train Delay Prediction Systems: A Big Data Analytics Perspective’, Big Data Research, vol. 11, pp. 54-64.

Suwardo, W, Napiah, M & Kamaruddin, IJJotioeM 2010, ‘ARIMA models for bus travel time prediction’, pp. 49-58.

Wang, P & Zhang, Q-p 2019, ‘Train delay analysis and prediction based on big data fusion’, Transportation Safety and Environment, vol. 1, no. 1, pp. 79-88.

Wen, C, Mou, W, Huang, P & Li, Z 2020, ‘A predictive model of train delays on a railway line’, Journal of Forecasting, vol. 39, no. 3, pp. 470-88.

Widiyaningtyas, T, Muladi & Qonita, A 2019, Use of ARIMA Method To Predict The Number of Train Passenger In Malang City, IEEE, 978-1-5386-8448-1, Conference,http://ezproxy.uow.edu.au/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=edseee&AN=edseee.8834663.

Wu, J, Zhou, L, Cai, C, Dong, F, Shen, J & Sun, G 2019, Towards a General Prediction System for the Primary Delay in Urban Railways, IEEE, 978-1-5386-7024-8

978-1-5386-7023-1, Conference, <http://ezproxy.uow.edu.au/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=edseee&AN=edseee.8916868&site=eds-live>.

Yaghini, M, Khoshraftar, MM & Seyedabadi, M 2013, ‘Railway passenger train delay prediction via neural network model’, Journal of Advanced Transportation, vol. 47, no. 3, pp. 355-68.

Li, Chunshien, Chiang, Tai-Wei. "Complex Neurofuzzy ARIMA Forecasting—A New Approach Using Complex Fuzzy Sets", Journal IEEE Transactions on Fuzzy Systems, June 2013, Volume: 21, Issue: 3.

maini, r., 2020. Time Series Analysis- Forecasting Departure Delays. [online] Medium. Available at: <https://medium.com/analytics-vidhya/time-series-analysis-forecasting-departure-delays-776d22636179> [Accessed 27 October 2020].

# APPENDIX

## API REQUEST

import requests

headers = {

    'Accept': 'application/x-google-protobuf',

    'Authorization': 'apikey iT9QxFPpO4g5pKsmMjLYUogP95nyObkBMOxn',

}

params = (

    ('debug', 'true'),

)

response = requests.get('https://api.transport.nsw.gov.au/v1/gtfs/realtime/sydneytrains', headers=headers, params=params)

print(response)

print(response.status\_code)

## HEATMAPS

The heatmaps show the data which was divided by days of the week into separate data frames. The complete dataset spanned two weeks; thus there was two days of data for each weekday. Then the data was grouped by station-id and average delay calculated for each group for arrival delay and departure delay.

Chart, bar chart

Description automatically generated

Figure: Heatmap of Tuesdays

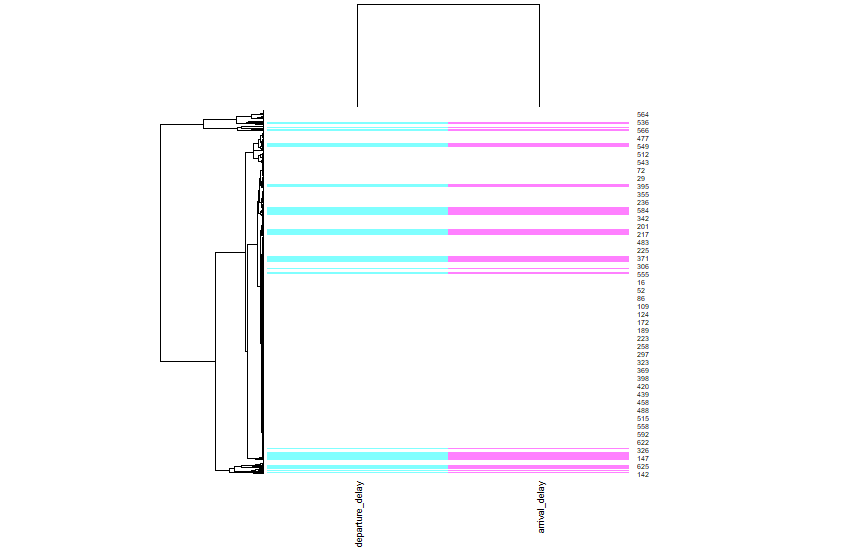


Figure: Heatmap of Wednesdays

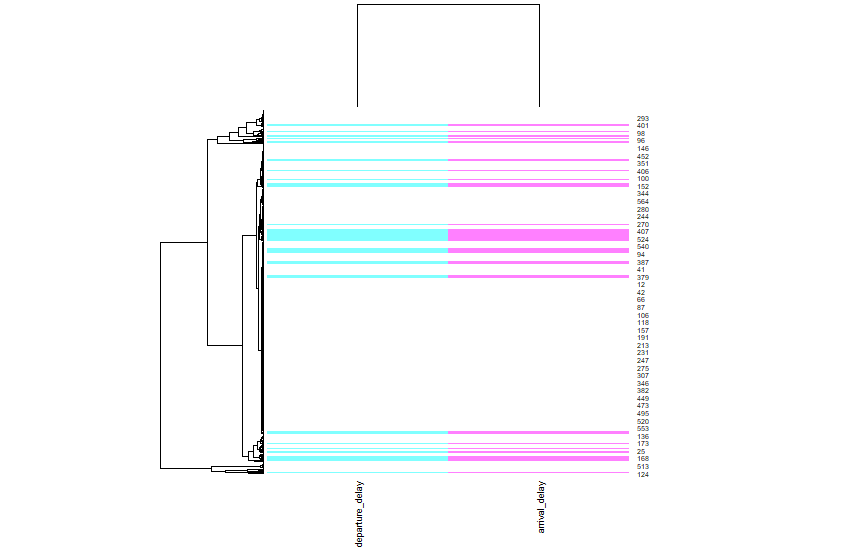


Figure: Heatmap of Thursdays

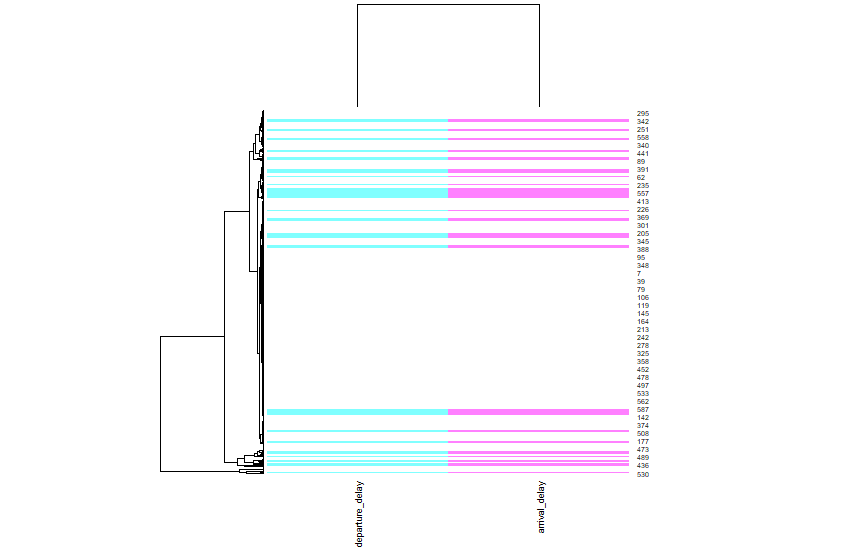


Figure: Heatmap of Fridays

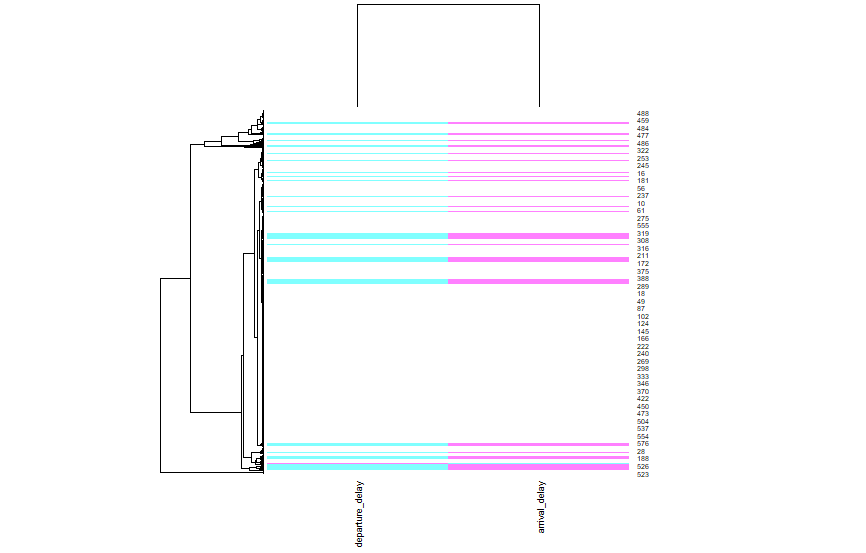


Figure: Heatmap of Saturdays

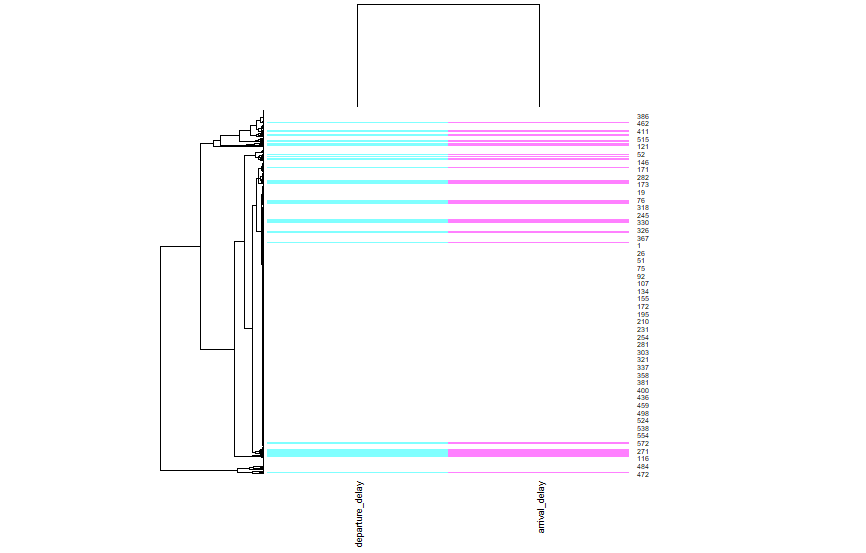


Figure: Heatmap of Sundays

## DATA EXCERPT

|  |  |  |
| --- | --- | --- |
| stop\_id | arrival\_delay | departure\_delay |
| 2508162 | 211 | 211 |
| 2515142 | 211 | 211 |
| 2500392 | 211 | 121 |
| 2500402 | 121 | 121 |
| 2500372 | 121 | 121 |
| 2526172 | 121 | 121 |
| 2530232 | 121 | 0 |
| 2527162 | 0 | 0 |
| 2529201 | 0 | 0 |
| 2529220 | 0 | 0 |