Appendix A: **Data Quality Report**

A.1 Overview

The dataset used is Dublin Bus travel data from 2018 in the form of three plain text documents: RT_Leavetimes, RT_Trips and RT_Vehicles. Each file contains data accumulated from 1st January 2018 until 31st December 2018. Below is an independent breakdown of each file and its features, giving an overview of the data, while outlining data quality issues found and how they will be addressed.

By using a combination of summary statistics and visual data representation such as bar charts, I will analyse the data and the relationship between features.

A.2 Summary

Below is an independent breakdown of each file and its features. In order to better estimate bus travel times, weather information for 2018 was joined with travel information. However this data appeared to have no effect on arrival time predictions and was dropped.

A.2.1 RT_Leavetimes

This dataset represents stop to stop information for individual trips throughout the year. It consists of 117 million rows with 18 features, each row corresponding to an individual stop along a specific bus route. The features are as follows:

- DataSource - Bus Operator Code
- DayOfService - Date of service
- TripID - Unique trip code
- ProgrNumber - Sequential position of stop in the trip
- StopPointID - Unique stop point code
- PlannedTime_Arr - Planned arrival time from stop, in seconds
- PlannedTime_Dep - Planned departure time from stop, in seconds
- ActualTime_Arr - Actual arrival time from stop, in seconds
- ActualTime_Dep - Actual departure time from stop, in seconds
- VehicleID - Unique vehicle code

- Passengers - Number of passengers on board
- PassengersIn - Number of boarded passengers
- PassengersOut - Number of descended passengers
- Distance - Distance measured from beginning of trip
- Suppressed - Whether trip is suppressed or not
- JustificationID - Fault code
- LastUpdate - Time of last recorded update
- Note - Free note

A.2.2 RT_Trips

This dataset contains information for full journeys made throughout the year. There are 2 million rows with 16 features, each row representing data for a single, end-to-end bus journey. The features are as follows:

- DataSource - Bus Operator Code
- DayOfService - Date of service
- TripID - Unique trip code
- LineID - Unique line code
- RouteID - Unique route code
- Direction - Route direction
- PlannedTime_Dep - Planned departure time of trip, in seconds
- PlannedTime_Arr - Planned arrival time of trip, in seconds
- Basin - Basin code
- TenderLot - Tender Lot
- ActualTime_Dep - Actual departure time of trip, in seconds
- ActualTime_Arr - Actual arrival time of trip, in seconds
- Suppressed - Whether the trip has been suppressed
- JustificationID - Fault code
- LastUpdate - Time of last recorded update
- Note - Free note

A.2.3 RT_Vehicles

This data represents information regarding the individual buses operating each day during the year. It contains 272,000 rows and 7 features, with each row corresponding to a single bus. The features are as follows:

- DataSource - Bus Operator Code
- DayOfService - Date of service
- VehicleID - Unique vehicle code
- Distance - Distance travelled by the vehicle on this day
- Minutes - Amount of time worked by vehicle on this day
- LastUpdate - Time of last recorded update
- Note - Free note

A.3 Data Cleaning

Each dataset was investigated and preprocessed individually before being joined to make one large dataset. After an initial logic check that all dates in the DAYOFSERVICE column were within the year 2018, the following actions were taken in order to clean the data:

1. Removing Null and Constant Columns
   RT_Leavetimes
   - Constant column 'DATASOURCE' was dropped.
   - Columns 'PASSENGERS', 'PASSENGERSIN', 'PASSENGERSOUT', 'DISTANCE' and 'NOTE' were all dropped as they were null.

   RT_Trips
   - Constant columns 'DATASOURCE' and 'BASIN' were dropped.
   - Empty column 'TENDERLOT' was dropped.

   RT_Vehicles
   - Constant column 'DATASOURCE' was dropped.
   - Empty column 'NOTE' was dropped.
2. Eliminating Uninformative Features
   RT_Leavetimes
   - Constant column 'DATASOURCE' was dropped.
   - Columns 'PASSENGERS', 'PASSENGERSIN', 'PASSENGERSOUT', 'DISTANCE' and 'NOTE' were all dropped as they were null.

   RT_Trips
   - 'SUPPRESSED' and 'JUSTIFICATIONID' were dropped as they had a high number of empty rows.

- 'LASTUPDATE' was dropped as it appeared to have been arbitrarily contrived and contained no useful information.
- 'NOTE' was dropped due to high cardinality and it was unclear what the data represented.

RT_Vehicles

- 'LASTUPDATE' was dropped as it appeared to have been arbitrarily contrived and contained no useful information.

3. Logic Check
   RT_Leavetimes
   - Three rows contained an 'ACTUALTIME_DEP' value which was less than the 'ACTUALTIME_ARR' value, which is impossible. The difference in these values was at most two seconds and so it would suggest a logging error. The 'ACTUALTIME_DEP' values for these rows were changed to the 'ACTUALTIME_ARR' values.

   RT_Trips
   - Rows with empty values for 'ACTUALTIME_ARR' and 'ACTUALTIME_DEP' were dropped.

   RT_Vehicles
   - Rows with a negative value for 'MINUTES' were dropped.

Once the individual datasets were cleaned, further processing was undertaken in order to join each dataset on specific columns. Columns in RT_Leavetimes were renamed in order to make a clean join and reduce unnecessary extra features. The following feature names were changed:

- 'PLANNEDTIME_ARR' to 'PLANNEDSTOPTIME_ARR/DEP'
- 'ACTUALTIME_ARR' to 'ACTUALSTOPTIME_ARR'
- 'ACTUALTIME_DEP' to 'ACTUALSTOPTIME_DEP'

A.4 Joining Datasets

The following actions were undertaken in order to join all three datasets. This created a single dataset with the intention of being used for modelling.

- RT_Leavetimes and RT_Trips were joined on 'DAYOFSERVICE' and 'TRIPID' which resulted in a new dataset with 13.5 million less rows and 18 features in total.
- A logic check was undertaken to make sure stop departure times don't occur before the trip starts and that trips don't finish before the last stop arrival. This resulted in a loss of a further 20,000 rows.
- 'ACTUALSTOPTIME_ARR' was dropped.
- This dataset was then joined with RT_Vehicles on 'DAYOFSERVICE' and 'VEHICLEID', resulting in a new dataset with 103 million rows and 20 features.

After the three datasets were joined, the following actions were taken to further clean the data and create new features appropriate for modelling.

1. Deriving New Features
    - 'TIMEPASSEDSINCE_DEP' was derived by taking the difference between 'ACTUALSTOPTIME_ARR' and 'ACTUALTIME_DEP'.
    - 'WEEKDAYOFSERVICE' and 'MONTHOFSERVICE' were derived from 'DAYOFSERVICE'.
2. Dropping Features
    - 'DAYOFSERVICE', 'TRIPID', 'NOTE', 'DISTANCE' and 'MINUTES' were all dropped due to high cardinality.
    - 'VEHICLEID' was dropped since it is unavailable from real time data.
    - 'PLANNEDSTOPTIME_ARR/DEP', 'ROUTEID', 'PLANNEDTIME_DEP', 'ACTUALSTOPTIME_DEP', 'PLANNEDTIME_ARR' and 'ACTUALTIME_ARR' were all dropped due to showing signs of high collinearity. The correlation matrix can be seen below:

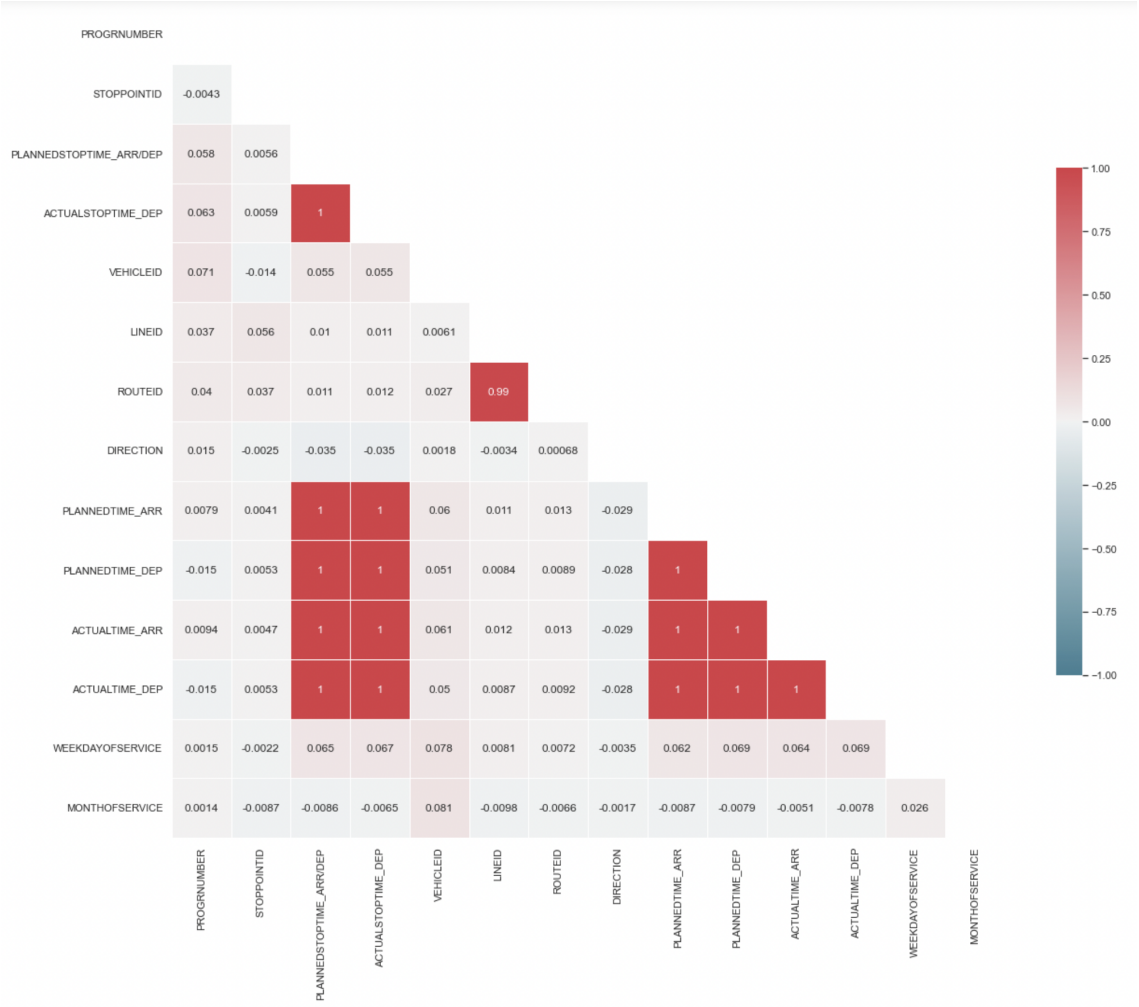


Figure A.1: Correlation Matrix of features

After cleaning the data, the following features were used for modelling: 'PROGRNUMBER', 'STOPPOINTID', 'LINEID', 'DIRECTION', 'ACTUALTIME_DEP', 'WEEKDAYOFSERVICE' and 'MONTHOFSERVICE', while the target feature was "TIMEPASSEDSINCE_DEP'. An example of the cleaned data is shown below:

	PROGRNUMBER	STOPPOINTID	LINEID	DIRECTION	ACTUALTIME_DEP	WEEKDAYOFSERVICE	MONTHOFSERVICE	TIMEPASSEDSINCE_DEP
0	12	119	1	1	47427	0	1	585
1	13	44	1	1	47427	0	1	631
2	14	7603	1	1	47427	0	1	644
3	15	45	1	1	47427	0	1	662
4	16	46	1	1	47427	0	1	710
5	17	47	1	1	47427	0	1	744
6	18	48	1	1	47427	0	1	762
7	19	49	1	1	47427	0	1	849
8	21	52	1	1	47427	0	1	931
9	22	265	1	1	47427	0	1	1036

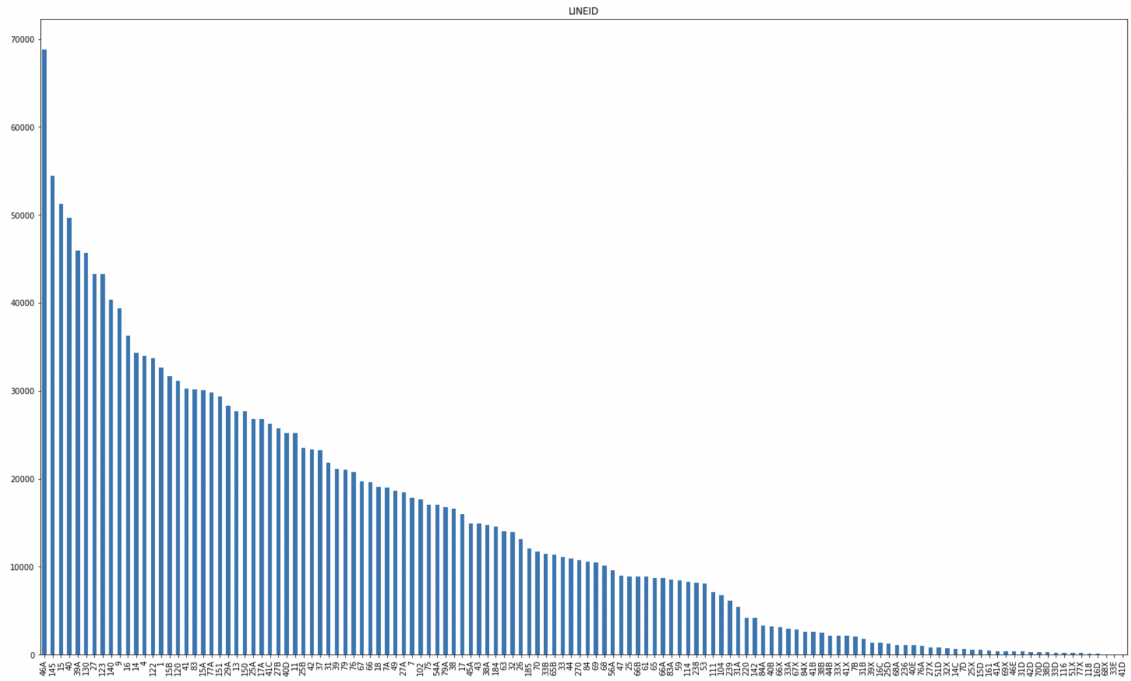
3. Dummy Features

   - Both 'WEEKDAYOFSERVICE' and 'MONTHOFSERVICE' were changed to 'category' types and dummy features were derived from each.

A.5 Review Categorical Features

Many of the features in each of the three datasets have a high cardinality making them difficult to use for analysis. For this reason I will focus on three features in RT_Trips that I believe give a useful insight into the travel data as a whole, LINEID, WEEKDAYOFSERVICE and MONTHOFSERVICE.

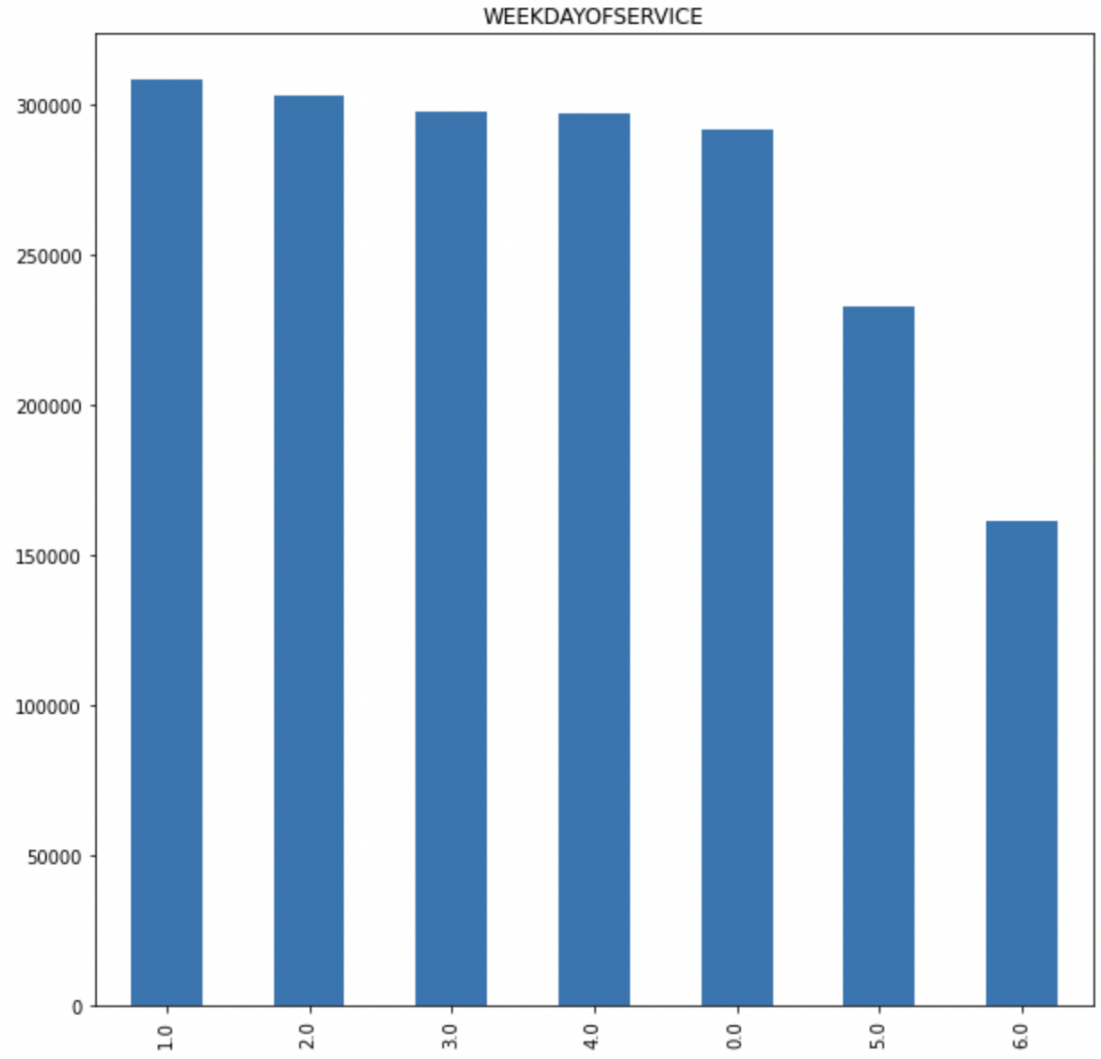
A.5.1 LINEID



The bar chart above shows the number of trips for each route taken in 2018. From this chart we can deduce that 46A is the most data rich bus route, compared to 41D which had the least

amount of trips in 2018. This information may be useful in determining the effect of sample size on machine learning models.

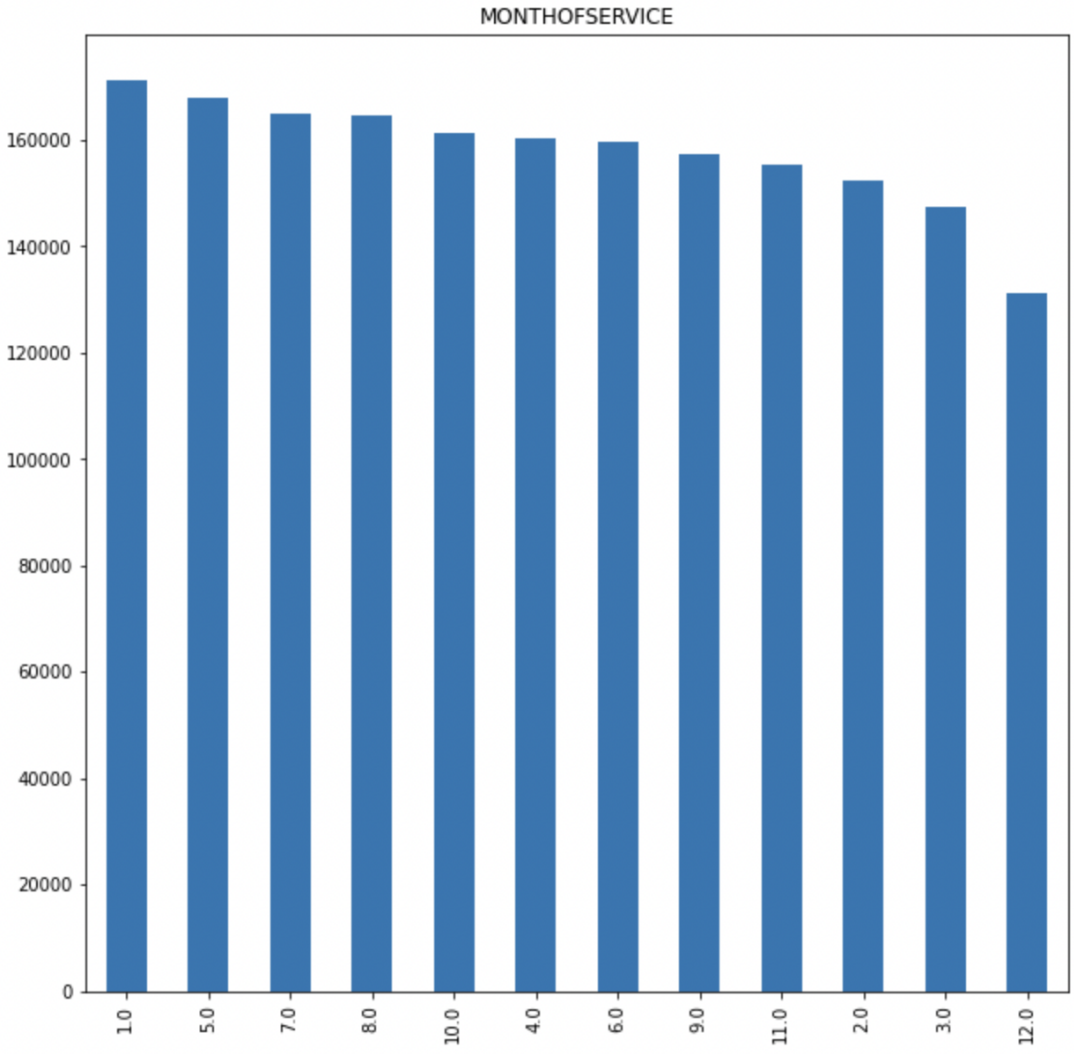
A.5.2 WEEKDAYOFSERVICE



The bar chart above indicates that the majority of bus journeys took place on Monday, while Saturday had the least amount of trips. We can see from the chart that the number of bus journeys decreases for the weekend as Friday, Saturday and Sunday have the least amount of service.

A.5.3 MONTHOFSERVICE

The bar chart shows the amount of bus journeys that took place for each month of 2018. Most trips took place in January and the least amount of trips took place in December, indicating that there does not appear to be any seasonal trend regarding the demand for bus journeys.


MONTHOFSERVICE
160000
140000
120000
100000
80000
60000
40000
20000
0
1.0
5.0
7.0
8.0
10.0
4.0
6.0
9.0
11.0
2.0
3.0
12.0