**Data Quality Report:**

The Dublin Bus dataset is divided into 3 separate files: rt\_leavetimes\_DB\_2018.txt, rt\_trips\_DB\_2018.txt, and rt\_vehicles\_DB\_2018.txt. The purpose of this report is to gain an understanding of the data (i.e.: what is the phenomenon that the data is trying to represent, what does each feature represent in the dataset, and are there any data quality issues associated with the data). This information can be used to determine any issues and how to correct them before models are trained and prepared.

1. **rt\_leavetimes\_DB\_2018.txt**

This dataset modes the times at which different buses passed by a particular stop. Each line denotes one particular vehicle at a particular stop. Note that these entries do not require that a bus must stop at the station with the intention that passengers be allowed to enter/exit the bus stop, as there are several entries where the expected time of arrival is identical to the expected time of departure. The columns consist of the following features:

|  |  |
| --- | --- |
| **Feature** | **Description** |
| **DATASOURCE** | How the data is stored. Upon inspection, the data appears to have a cardinality of 1, where the only value in each column is “DB”. This means to columns provides little insight into the phenomenon itself and is merely reflecting administrative tasks associated with the collection of data. |
| **DAYOFSERVICE** | This returns the day corresponding to when the data was taken. |
| **TRIPID** | A unique identifier for a bus trip itself. For example. The 15b going into Dublin city at 9 am would have a different TRIPID to the 15b going into Dublin city at 3 pm. In the original database, this served as the primary key for the table. |
| **PROGRNUMBER** | The programme the bus stop relates to (eg: 12, 13, 14, 15, 16) |
| **STOPPOINTID** | The unique identifier for the bus stop that the arrival and departure relate to. |
| **PLANNEDTIME\_ARR** | This denotes the time a bus is expected to arrive at the station. Note that the bus does not need to stop at the station to be included in the dataset. A bus might pass through one station on there way to another station, and this will be included into the dataset. These data points can be distinguished from how the Planned Time of Arrival is equal to the Planned Time of Departure (PLANNEDTIME\_DEP). This feature is measured (in seconds).  Although some would argue that this does not give us a significant amount of data to use, as many of the data points are businesses stops that have nothing to do with their main route, it does provide us with information as to if a bus is late to its next destination. If we find information relating to the weather and traffic conditions of that location, it could give us some insight as to why time bus arrived at the time it did and could potentially explain any delays with the bus. |
| **PLANNEDTIME\_DEP** | Records the scheduled time of departure (in seconds from 00:00:00 that day), not taking into account delays |
| **ACTUALTIME\_ARR** | Records the actual time the bus was expected to arrive at the station (in seconds from 00:00:00 that day) |
| **ACTUALTIME\_DEP** | Records the actual time the bus was expected to depart from the station (in seconds from 00:00:00 that day) |
| **STOPPOINTID** | A unique identifier to distinguish one bus stop from another. |
| **PROGRNUMBER** | Although this is not confirmed, this may relate to a certain bus route that the bus is traveling through. For example, in the first 100 rows of the data set, the document makes reference to the following program numbers: 12, 13, 14, 15, 64, 65, 66, Which are all bus routes in Dublin city. |
| **VEHICLEID** | A unique identified used to distinguish one vehicle from another. |
| **PASSENGERS, PASSENGERSIN, PASSENGERSOUT:** | These features capture the number of passengers passing through the station each day. However, upon initial inspection, each of these columns are blank, and contain no relevant information. Thus, it is difficult to discern what this information means. |
| **DISTANCE** | This denotes how far the stop point is from the next location (in meters). Similar to the features relating to the passenger data It is difficult to discern, as the column is left blank. |
| **SUPPRESSED/JUSTIFICATIONID:** | Was the bus suppressed (i.e. did something prevent the bus from completing its journey) and if so, what was the error code relating to the incident. |
| **LASTUPDATE** | The last time this entry was updated. This information is more for administrative purposes rather than depicting the phenomenon taking place. |
| **NOTE** | This column denotes any additional notes that the author deemed necessary to include. However, this column is largely left blank. |

1. **rt\_trips\_DB\_2018.txt**

This table reflects the trips themselves, and provides more information on the trip, such as what bus route the trip corresponds to, and how long did the bus take to reach the end of the route. The contents of the file are made up of 2.18 million points and includes the following 16 features:

|  |  |
| --- | --- |
| **Feature** | **Description** |
| **DATASOURCE** | Identical to the previous table. The feature only has one value (DB) and does not yield any relevant data |
| **DAYOFSERVICE** | The day that the bus trip took place. |
| **TRIPID** | A unique identifier to distinguish one trip from another. It appears to be connected to the previous dataset’s trip column, as they share identical values. This is also shared with the rt\_leavetimes table and can be used to connect the two together in an SQL style database. |
| **LINEID/ROUTEID** | The Line ID denotes what bus route the datapoint corresponds to. The ROUTEID denotes a particular version of the route. For example, the first entry in the table has a LINEID of 68 and 68\_80. |
| **DIRECTION** | Denotes if the bus is inbound (1) or outbound (2) (i.e., are travelling into or out of Dublin city). |
| **PLANNEDTIME\_ARR** | Records the scheduled time of arrival (in seconds from 00:00:00 that day), not taking into account delays |
| **PLANNEDTIME\_DEP** | Records the scheduled time of departure from the previous destination (in seconds from 00:00:00 that day), not taking into account delays |
| **ACTUALTIME\_ARR** | Records the actual time the bus was expected to arrive at the station (in seconds from 00:00:00 that day) |
| **ACTUALTIME\_DEP** | Records the actual time the bus was expected to depart from the previous station (in seconds from 00:00:00 that day) |
| **BASIN** | It is left uncertain what the “Basin” corresponds to in this situation. Additionally, this column is missing significant amounts of information, and it is advised that this column is dropped for this reason. |
| **TENDERLOT** | What lot does the bus reside in. Upon initial inspection, this column appears to be missing significant amounts of information and will likely need to be dropped. |
| **SUPPRESSED/JUSTIFICATIONID** | This reflects if the bus was suppressed (i.e., did something prevent the bus from completing its journey) and if so, what was the error code relating to the incident. |
| **LASTUPDATE** | The last time the information in this column was updated. |
| **NOTE** | This column denotes any additional notes that the author deemed necessary to include. However, this column is largely left blank. |

1. **rt\_vehicles\_DB\_2018.txt**

This table reflects the vehicles themselves. It includes the following features:

|  |  |
| --- | --- |
| **Feature** | **Description** |
| **DATASOURCE** | Identical to the previous table. |
| **DAYOFSERVICE** | The day the bus was in transit. |
| **VEHICLEID** | The unique identifier corresponding to a particular vehicle. This appears to have served as the primary key for this table in the original database. |
| **DISTANCE** | The total distance (in meters) the bus travels over that day. |
| **MINUTES** | The time (in minutes) that the bus spent in transit that day |
| **LASTUPDATE** | The last time the information in this column was updated. |
| **NOTE** | This column denotes any additional notes that the author deemed necessary to include. However, this column is largely left blank |

Overall, this represents a much smaller dataset compared to the other tables in this list. The data relates to the businesses themselves and does not directly provided any information as to if the bus was late to its destination. Does, the file is the least relevant to the data set.

**Duplicates**

There were no duplicate points in any of these tables, although this has yet to be confirmed for the leavetimes table as there were difficulties in loading the data sets (see File Size section). This is ideal, as duplicate data points may induce bias into the datasets (Naumann and Herschel, 2010, pp.1-3).

**Missing Values**

A summary of the missing values in each feature in the three tables can be found in appendix 1. The only major issue was the missing values in the ACTUALTIME\_ARR and ACTUALTIME\_DEP, as these features will play a critical factor for model building alongside the PLANNEDTIME\_ARR and PLANNEDTIME\_**DEP** times.

**File Sizes**

The files had the following sizes:

|  |  |  |
| --- | --- | --- |
| **File Name** | **File Size (KB)** | **File Size (GB)** |
| rt\_leavetimes\_DB\_2018.txt: | 11301458278 | 10,777.9105 |
| rt\_trips\_DB\_2018.txt | 231974052 | 221.2277 |
| rt\_vehicles\_DB\_2018.txt | 17395353 | 16.5895 |

As demonstrated above, the rt\_leavetimes is 10.53 Terabytes large. This has made analysing the data in this file extremely difficult, as neither the Linux Virtual Machine nor the Data Processing Server provided by UCD is capable of running the Pandas scripts associated with the file. Based on error message generate from attempts to run the data, there is at least 116 million rows in the data set. A solution for accessing the data will need to be addressed in the data quality plan.

**Data Integrity**

The rt\_trips\_DB file, where there are 103 rows where the departure time occurs before the arrival time, which is impossible to occur. For this reason, it is advised that these rows be dropped. Although data loss is undesirable, this only results in a data loss of 103 data points of a data set with 2,182,637 points, meaning that this represents less than a 0.005% data loss. There were also 3 entries in the leave ties file that suffered this issue as well, as well as 11 entries in the vehicles file where the time a bus spent in transit was negative.

**Summary Statistics:**

Summary statistics of each feature can be found in appendix 1.

**Conclusion**

Overall, the datasets are relatively clean. The leave times dataset is too large to be analysed in pandas, and a secondary data quality report may be prepared addressing any issues with that data set. The vehicles dataset has not serious issues, but the data is less relevant to the project, as it only informs users of the vehicles use rather than if the vehicle was late to its destination. The leave times file contains valuable data, but due to the technical difficulties associated with accessing the data, another data quality report will be required to analyse this data set.

**Appendix 1:**

**Leave Time Analysis:**

Note: The table contain the Leave Time values was over 10 TB in size. For this reason, the dataset needed to be divided into three smaller files, the first two (segments aa and ab) containing 40 million rows, and the final table (segment ac) contain 36 million rows. Due to looming deadlines and time constraints, the tables contain these data points were not listed in the appendix, but are available on the Data processing server.

**Trip Analysis:**

Continuous Data

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **details** | **count** | **mean** | **std** | **min** | **0.25** | **0.5** | **0.75** | **max** |
| PLANNEDTIME\_ARR | 2182637.0 | 55078.7477 | 17424.8223 | 18203.0 | 39683 | 55276 | 69006 | 91680.0 |
| PLANNEDTIME\_DEP | 2182637.0 | 51533.2928 | 17571.6373 | 16200.0 | 36000 | 51600 | 65160 | 87600.0 |
| ACTUALTIME\_ARR | 2045430.0 | 55062.3883 | 17338.6992 | 17546.0 | 39668 | 55189 | 69019 | 94378.0 |
| ACTUALTIME\_DEP | 2018086.0 | 51706.1166 | 17526.9482 | 15974.0 | 36518 | 51618 | 65398 | 97177.0 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **details** | **count** | **unique** | **top** | **freq** | **first** | **last** |
| DAYOFSERVICE | 2182637 | 360 | 12/02/2018 | 7122 | 01/01/2018 | 31/12/2018 |

Categorical Data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **details** | **count** | **unique** | **top** | **freq** |
| DATASOURCE | 2182637 | 1 | DB | 2182637 |
| TRIPID | 2182637 | 658964 | 7324685 | 19 |
| LINEID | 2182637 | 130 | 46A | 76728 |
| ROUTEID | 2182637 | 588 | 46A\_74 | 37182 |
| DIRECTION | 2182637 | 2 | 2 | 1100273 |
| BASIN | 2182637 | 1 | BasDef | 2182637 |
| TENDERLOT | 0 | 0 |  |  |
| SUPPRESSED | 4333 | 1 | 0.0 | 4333 |
| LASTUPDATE | 4333 | 1 | 0.0 | 4333 |

Missing Data

|  |  |
| --- | --- |
| **details** | **missing (%)** |
| DATASOURCE | 0 |
| DAYOFSERVICE | 0 |
| TRIPID | 0 |
| LINEID | 0 |
| ROUTEID | 0 |
| DIRECTION | 0 |
| PLANNEDTIME\_ARR | 0 |
| PLANNEDTIME\_DEP | 0 |
| ACTUALTIME\_ARR | 6.286294973 |
| ACTUALTIME\_DEP | 7.539091475 |
| BASIN | 0 |
| TENDERLOT | 100 |
| SUPPRESSED | 99.80147867 |
| JUSTIFICATIONID | 99.80161612 |
| LASTUPDATE | 99.80147867 |
| NOTE | 0 |

**Vehicle Analysis:**

Continuous Data

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **details** | **count** | **mean** | **std** | **min** | **0.25** | **0.5** | **0.75** | **max** |
| DISTANCE | 272622 | 177486.3285 | 71334.43238 | 0 | 126765 | 177420 | 227126.25 | 473864 |
| MINUTES | 272622 | 45597.64322 | 14881.38987 | -70145 | 39380 | 46994.5 | 57911 | 76651 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **details** | **count** | **unique** | **top** | **freq** | **first** | **last** |
| DAYOFSERVICE | 272622 | 360 | 04/05/2018 | 911 | 01/01/2018 | 31/12/2018 |

Categorical Data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **details** | **count** | **unique** | **top** | **freq** |
| DATASOURCE | 272622 | 1 | DB | 272622 |
| VEHICLEID | 272622 | 1152 | 2693284 | 342 |
| LASTUPDATE | 272622 | 360 | 25-JUN-18 19:01:32 | 911 |
| NOTE | 0 | 0 | 0 | 0 |

Missing Data

|  |  |
| --- | --- |
| **details** | **missing (%)** |
| DATASOURCE | 0 |
| DAYOFSERVICE | 0 |
| VEHICLEID | 0 |
| DISTANCE | 0 |
| MINUTES | 0 |
| LASTUPDATE | 0 |
| NOTE | 100 |

**Bibliography**

Naumann, F., & Herschel, M. (2010). An introduction to duplicate detection. Morgan & Claypool.