**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, all of which have been explored and analysed in the Data Quality Report. This report was commissioned to address any issues with the datasets mentioned in the previous reports and outline a solution to the problem.

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

The following data quality issues were identified in this file. Please note that attempts at uploading this file to pandas were unsuccessful, as the dataset was over 10 Terabytes in size. A later Data analysis report will need to be commissioned to confirm these issues.

|  |  |  |
| --- | --- | --- |
| **Feature** | **Issue** | **Remedy** |
| DATASOURCE | Cardinality of 1 | Drop feature |
| DAYOFSERVICE | Includes time of day when not relevant | Remove Time |
| TRIPID | No issues | - |
| PROGRNUMBER | No issues | - |
| STOPPOINTID | No issues | - |
| PLANNEDTIME\_ARR | No issues | - |
| PLANNEDTIME\_DEP | No issues | - |
| ACTUALTIME\_ARR/ ACTUALTIME\_DEP | Less than 0.001% of dataset have logical integrity issues | Drop rows |
| VEHICLEID | No Relevant Issues | - |
| PASSENGERS | 100% of values missing | Drop feature |
| PASSENGERSIN | 100% of values missing | Drop feature |
| PASSENGERSOUT: | 100% of values missing | Drop feature |
| DISTANCE | 100% of values missing | Drop feature |
| SUPPRESSED/JUSTIFICATIONID: | 99.7% of values missing | Drop feature |
| LASTUPDATE | Administrative Activity | Drop feature |
| NOTE | 100% of values missing | Drop feature |

The main issue with this data as a whole is the files size and inability to be imported into pandas. The file needed to be separated into three data sets and each file was processed individually. Other than some logical integrity issues with the dataset, there were no substantial issues.

By logical integrity issues, this refers to entries where the actual departure time occurs before the actual arrival time. However, of nearly 117 million entries, less than 8 of entries had these constraints.

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

The following data quality issues were identified in this file:

|  |  |  |
| --- | --- | --- |
| **Feature** | **Issue** | **Remedy** |
| DATASOURCE | Cardinality of 1 | Drop feature |
| DAYOFSERVICE | Includes time of day when not relevant | Remove time |
| TRIPID | No Issues | - |
| LINEID/ROUTEID | Ambiguity as to which routes these relate to | Further analysis required |
| DIRECTION | Ambiguity as to if 1 means inbound/outbound | Further analysis required |
| PLANNEDTIME\_ARR | No Issues | - |
| PLANNEDTIME\_DEP | No Issues | - |
| ACTUALTIME\_ARR | 137,207 rows with missing feature value, integrity issues | Drop Rows (if unintentional) |
| ACTUALTIME\_DEP | 164,551 rows with missing feature value, integrity issues | Drop Rows (if unintentional) |
| BASIN | Cardinality of 1 | Drop Feature |
| TENDERLOT | No values (i.e. Cardinality of 1) | Drop Feature |
| SUPPRESSED | 4333 rows (0.00001985 of dataset) | Drop Feature |
| JUSTIFICATIONID | 4 justification codes missing where suppressed value exists. | Drop Feature |
| LASTUPDATE | Tracks administrative tasks associated with collecting/storing data | Drop Feature |
| NOTE | Repeats data featured in other columns | Drop Feature |

The DATASOURCE, TENDERLOT, BASIN and LASTUPDATE features were either used to track tasks associated with collection and storing the data, and not the phenomenon the data is trying to capture or had a cardinality of less than 2. For this reason, it is advised that these features be dropped from the dataset immediately, as they provide no information relevant to models for predicting bus delays.

Additionally, rows were SUPPRESSED and the JUSTIFICATIONID features are not NaN accounts for less than 0.1% of the entire data base, and it is unlikely this information can be used effectively in the models. For this reason, it is advised that the data be dropped.

Aside from these issues, there we no other issues associated with the following features: TRIPID, PLANNEDTIME\_ARR and PLANNEDTIME\_DEP. The DAYOFSERVICE includes 00:00:00 in each row when it is not necessary, but his can be easily removed.

The major issue with this table is the data quality issues for the ACTUALTIME\_DEP and ACTUALTIME\_ARR. Both these columns have rows with missing values, where there are 289,270 rows (13.25% of the total data set). Furthermore, there is a small number (103 rows or less than 0.00472% of the entire data set) where the departure times occur before the arrival times. In the

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

The following data quality issues were identified in this file:

|  |  |  |
| --- | --- | --- |
| **Feature** | **Issue** | **Remedy** |
| DATASOURCE | Cardinality of 1 | Drop Feature |
| DAYOFSERVICE | Includes time of day when not relevant | Remove Time |
| VEHICLEID | No Issue | - |
| DISTANCE | No Issue | - |
| MINUTES | Negative Values (0.004%) | Set to 0 or Drop |
| LASTUPDATE | Tracks administrative tasks associated with collecting/storing data | Drop Feature |
| NOTE | No data | Drop Feature |

Unlike previous datasets, there were relatively no issue with this data set. The only issues would be the presence of the negative values found in the minutes feature, but this occurred for only 11 entires (less than 0.01% of the entire dataset). Other than this, the only significant issues where the redundant features from the dataset.

**Conclusion**

These are the proposed changes to the dataset with the intention of cleaning the datasets. However, this report does not address what features to create or modify for the purposes of building models to fit the dataset. Additional features such as a difference in planned/actual arrival time or potential the journey time for one stop to another may be possible features to use as an outcome feature.