**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 | - |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

The main issue with this table as a whole is the files size and inability to be imported into pandas.

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 | Drop Rows (if unintentional) |
| ACTUALTIME\_DEP | 164,551 rows with missing feature value | 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 ess 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 direction feature did not have any missing values.

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

The following data quality issues were identified in this file:

|  |  |  |
| --- | --- | --- |
| **Feature** | **Issue** | **Remedy** |
| DATASOURCE | Cardinality of 1 | Drop |
| DAYOFSERVICE | Includes time of day when not relevant | Remove Time |
| VEHICLEID | No Issue | - |
| DISTANCE |  |  |
| MINUTES |  |  |
| LASTUPDATE |  |  |
| NOTE |  |  |