Datasheet - TTC Bus Delays

The dataset in question is `bus-clean.RDS`, found in the `inputs` directory. This dataset was used extensively in this paper and formed the basis of further manipulation in the paper.

Motivation

Purpose of creation

To have a consistent, single source of data that all analysis could be based off. The dataset had to be accurate (e.g. no false / erroneous data) and manageable (e.g. in a format that could be easily manipulated in further analysis).

Authors

The original data was provided by the Toronto Transit Commission (TTC), accessible through the City of Toronto's Open Data Portal. This and similar data is made available as part of various public improvement initiatives, empowering individuals and independent organisations to solve civic issues in a data-driven manner. The specific dataset in this paper was created by the author to aid the analysis of bus delays specifically.

Composition

There are **488,764 instances** in the dataset.

Each instance of the data represents a delay event, associated with the following features:

- Date
- Route number
- Time of day
- Day of week
- Location of incident
- Incident type (e.g. route diversion, mechanical failure)
- Delay duration in minutes
- Gap in minutes between the current and following bus
- Direction of travel
- Vehicle ID

The dataset only contains events from Jan 1st, 2014 to Dec 31st, 2021. This is a subset of the original data, which is updated monthly and was last refreshed on April 6th, 2022.

The `delay duration` field was used as the target when creating models in this paper. The data is also self-contained and does not reveal personal information regarding any subpopulations.

Preprocessing

Significant preprocessing and cleaning was done to create this dataset from the original source. Approximately 30,000 instances were removed from the original data, largely due to missing or erroneous feature data. Several equivalent but differently-named features were also coalesced together. Finally, certain levels of categorical variables were aggregated. The full list of cleaning steps is as follows:

1. Renaming columns for consistency e.g. “Min Delay” to “min_delay”
2. Recasting the time field from a string to a time type
3. Dropping columns with limited data e.g. `incident_id` only had values for 5k/517k total rows
4. Coalescing equivalent but differently-named columns e.g. `delay` and `min_delay`
5. Dropping remaining rows with N/A values - 17.5k/517k total rows
6. Removing rows with delay durations beyond the 99th percentile (235 minutes); those beyond were deemed outliers that had little interpretable business benefit

All cleaning was done within the `R` environment. These cleaning functions can be carried out by running the `data-cleaning.Rmd` script within the `scripts` directory.

Uses

The dataset was used extensively to fit various linear models with delay duration as the dependent variable. Care should be taken to adjust for the data’s original right-skewed distribution, such as through log transformations.

Note that location-based data is currently not codified consistently, but stores location information as text fields that are syntactically inconsistent. Analysis regarding this data will likely be valuable, but depends on additional preprocessing.

Distribution and maintenance

Although publicly available, this dataset is not intended to be distributed in isolation. All licenses from the original data still hold, and similar versions can therefore be created by other organisations. As such, the dataset will not refresh automatically to encompass newly available data.

If others wish to refresh or augment this dataset, a larger dataset can be obtained through modification of the `data-download.Rmd` script. Data integrity is dependent on the Open Data Portal source, but effort has been made to ensure that datasets are backwards compatible and will be available to the public even if deprecated.