ETL Project

Team

* Tealiie Le
* Kaayin Kee
* Stuart Karpin

**Extraction**

We used 4 datasets from the DataVic website which is Victoria’s open data platform

<https://discover.data.vic.gov.au/organization/vicroads>

DataVic is the place to discover and access Victorian Government open data.

DataVic is owned by the State Government of Victoria, and is administered by the Victorian Department of Premier and Cabinet.

The [DataVic Access Policy](https://www.data.vic.gov.au/datavic-access-policy) was endorsed by the Victorian Government in August 2012 for implementation throughout the Victorian public service.

In particular we used data sets from the ‘Crash Stats’ suite of data. In particular the following data sets

<https://discover.data.vic.gov.au/dataset/crash-stats-data-extract>

* ACCIDENT
* ACCIDENT\_LOCATION
* ROAD\_SURFACE\_CONDITION
* ATMOSPHERIC\_COND

**Transformation**

Our first steps in cleaning up the datasets involved importing the different CSV files into separate data frames

Graphical user interface, text, application, email

Description automatically generated

**Figure 1:** Importing CSV data into separate data frames

Following steps included renaming common columns of the data sets with the same name and format (Figure 2).

Graphical user interface, text, application

Description automatically generated

**Figure 2:** Filtering and renaming columns for ease of merging

In all 4 data sets the Accident No was the common reference making it the logical variable to merge the data on using an outer join to capture all uncommon data into one data frame(Figure 3).

A picture containing table

Description automatically generated

**Figure 3:** Merging the data frames

With all the datasets combined (Figure 5) into one universal table each accident event was reordered to a more logical format and contained a higher concentration of usable ‘at a glance’ data.

Graphical user interface, application, Word

Description automatically generated

**Figure 5:** Final Output after merging all the datasets

**Load**

The last step was to transfer our final output into a database. We created a database and respective tables to match the columns from the final Panda’s Data Frame using QuickDBD exporting the table structure from an Entity Relationship Diagram (Figure 6)

Graphical user interface, website

Description automatically generated

**Figure 6: ERD**

**Summary**

We used these datasets so we could identify

These indices can be used to identify