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

<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).

Table

Description automatically generated with medium confidence

**Figure 3:** Merging the data frames

After this stage we also allowed for a CSV print of the merged Data frames

**Graphical user interface, text

Description automatically generated with medium confidence**

**Load**

The last step was to transfer our filtered tables into a database. We created a database and respective tables to match the columns from the final Panda’s Data Frames using QuickDBD exporting the table structure from an Entity Relationship Diagram (Figure 5)

Diagram

Description automatically generated with medium confidence

**Figure 5: ERD**

Graphical user interface, text, application

Description automatically generated

Some final cleaning also took place in the data base which also provided some testing of the integrity of the data transfer by running queries on the data (Figure 6)

**Table

Description automatically generated**

**Figure 6:** SQL data query

**Summary**

We used these datasets to create a data base that included data that is otherwise broken up and includes a lot of data that is otherwise redundant for the use by a layman analyst on this topic or in other words the data selected was recognizable, of interest and recognizable, but most importantly provided actionable insights into the crash data.

