ETL Project - Road Fatalities in Western Australia

Contributors: Damini Patel and Kylie Burns

SUMMARY

This Data Base is created for Mainroads Western Australia to assess the number of crash fatalities experienced across WA. Using our data in conjunction with further research should give insight into the infrastructure and safety of our roads.

This can be done by analysing:

* Weather
* Car Safety Features
* Repairs of infrastructure
* Improved infrastructure
* Improved/Increased road safety measures
* Impact of new infrastructure

DATA SOURCES

Bureau of Infrastructure and Transport Research Economics

CSV: [https://www.bitre.gov.au](https://www.bitre.gov.au/)

Main Roads Western Australia - Open Data, Maps & Apps

CSV: [https://portal-mainroads.opendata.arcgis.com](https://portal-mainroads.opendata.arcgis.com/)

DATA CLEANUP & ANALYSIS

End User Questions

1. How many road fatalities does Western Australia record every year?

2. What vehicles are involved in most road fatalities?

3. What speed limit is associated with the highest road fatalities?

4. What age group is associated with the highest road fatalities?

5. What type of road user is most at risk of a road fatality?

Extract

2 of our 3 CSV files were obtained on the bitre.gov.au website. The 3rd CSV file was obtained from the Mainroads Western Australia website. The files from bitre.gov.au were not initially CSV files as the workbook contained more than 1 worksheet informing us that the null value for this data was -9. All worksheets were subsequently deleted, and we retained the data worksheet and converted the file into a CSV.

The dictionaries for all CSV files used have been saved into Resources folder.

Transformations

Each of our CSV files were imported and cleaned using Jupyter Notebook and Visual Studio Code and the resultant dataframes were saved as a CSV file.

We first imported the required dependencies required for the cleaning process. Following the import of the required dependencies, the data files were read in from the aforementioned sources, using the pd.read\_csv from pandas.

The CSV files can be found in the folder Resources.

After reading in the data files, the process of converting the data into data frames commenced.

Now that the data was in data frames, the data cleaning process was undertaken.

Column names were formatted by copying the original names from the imported data and renaming them to match column names created in our SQL database.

Two of our datasets had null values upon download and one did not (Crash\_Information\_(Last\_5\_Years)).

The data we were focussing on was for Western Australia only, so the ‘State’ column was filtered to only display data from WA.

With the data for WA only, the process of specifying the columns that were required could be undertaken and we filtered out columns of information that was not required.

The rows that contained null values (-9) were not removed as this would cause a mismatch between the 2 bitre files. Instead, rows that contained (-9) values were replaced with a string value of ‘unknown’ for the columns of string types and with the integer 0 for numeric columns. A check was done to ensure that no 0 values were present in these numeric columns originally.

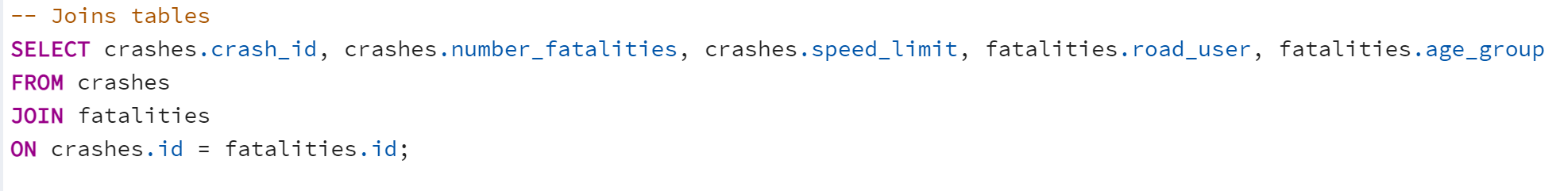
Once the cleaning of the data was finalised, the dataframe was saved to a CSV file.

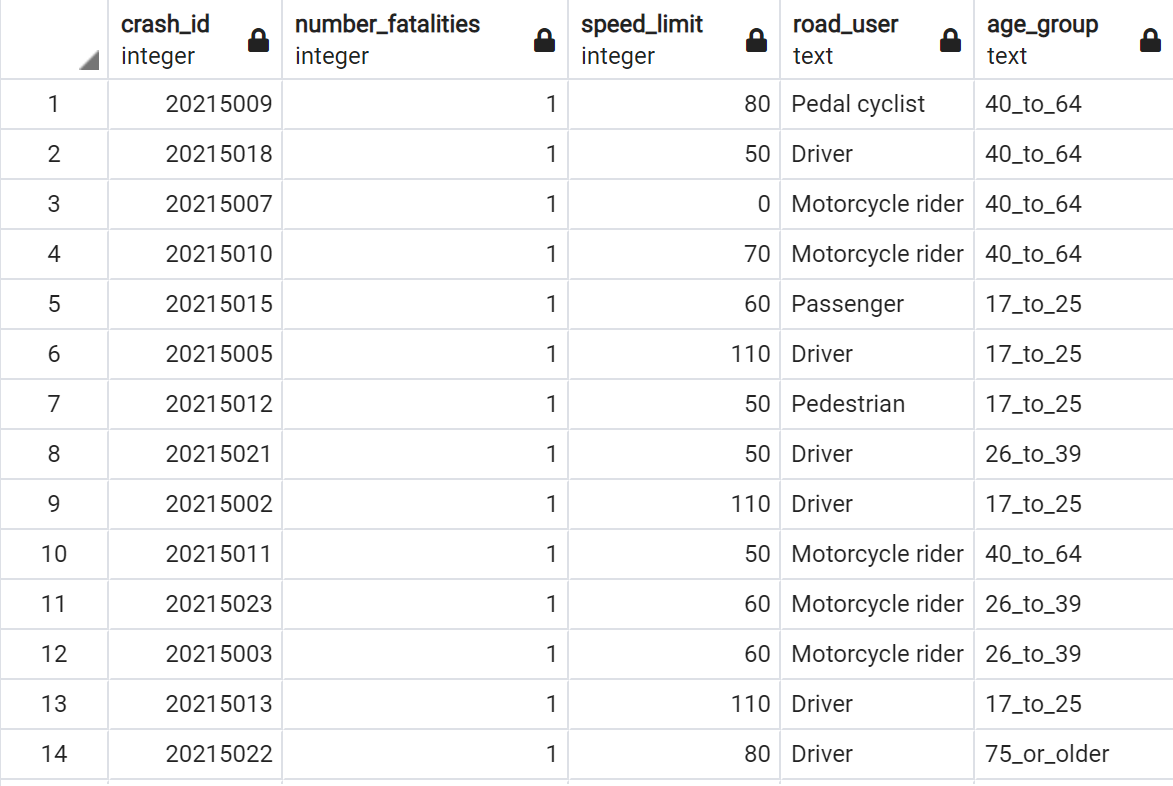
Load

The queries to create the tables can be found in sql file: fatal\_crashes.sql

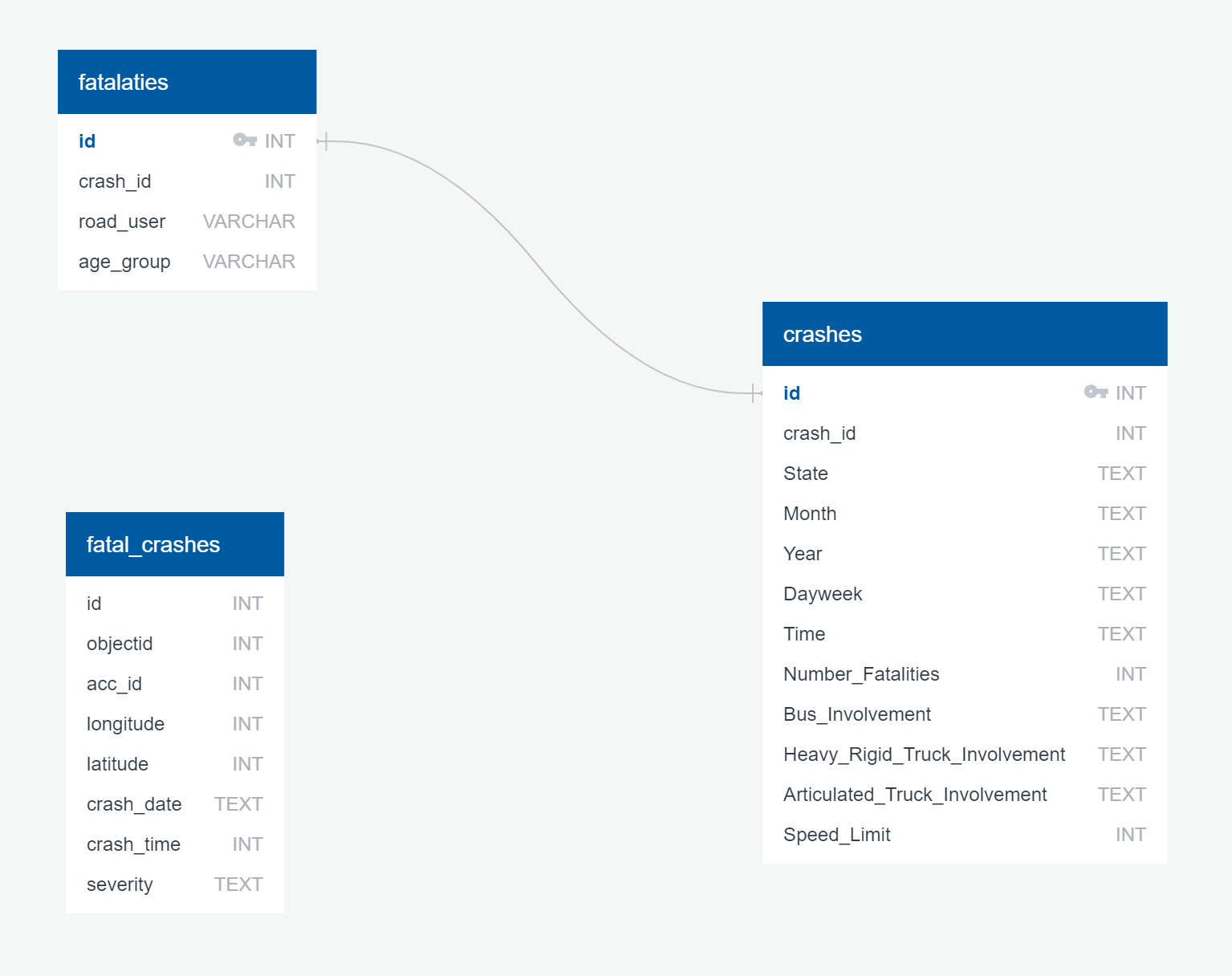
The following process was used to connect, and check the connection to the database:

1. Read in the three cleaned dataframes that were saved as CSV files.
2. create\_engine to create a connection to the Postgres database Road\_Fatalities
3. engine.table\_names to check connection and table names
4. df.to\_sql(name= ‘table name’, con=engine, if\_exists='append', index=False)
5. SELECT \* command in pgAdmin as well as in sqlAlchemy in the note book to check the data was successfully upload.

In pgAdmin the following join of tables was completed to demonstrate the connection and relationships between the tables.:



ERD below reflects the tables within the Road\_Fatalities Database



Further investigations/data

For further investigation an additional table (fatal\_crashes) was included that does not have a direct relationship with the other two tables. This fatal\_crashes table will allow the end user to conduct further investigation and gain additional information regarding the location by utilisation of the longitude and latitude.

The relationship between the three tables comes from the date and time of the crash.

This data will allow the end user to answer the questions posed earlier and will allow further investigation and deeper analysis of Road Fatalities in Western Australia.

An example of this could be finding the actual causes of the fatalities e.g. Driving under the influence, unroadworthy vehicle or driver fatigue.