ETL Project

Road Fatalities in Western Australia

ETL DATA ANALYSIS – PROJECT 2 – 13th March 2021

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 SOURCE

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

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

DATA CLEANUP & ANALYSIS

The following outlines the steps of importing and cleaning data in Jupyter Notebook and Visual Studio Code and subsequently exporting the cleaned data frames to a SQL server.

Each of our CSV files were imported, cleaned and the resultant dataframes were saved as a CSV file.

Beginning by importing the required dependencies required for the coding. 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 which were CSV, the process of converting the data into data frames commenced.

Now that the data was in data frames, the data cleaning process was undertaken. Two of our datasets had null values upon download and one didn’t (Crash\_Information\_cleaned).

The data we were focussing on was for Western Australia only. The column ‘State’ was filter 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 filter out columns of information that was not required.

Formatting column names and headers, the original header row from the imported dataset was renamed. The renaming was to get the columns to match column names created in our SQL database.

The rows that contained null values were not removed, however rows that contained null values were replaced with a string value of ‘unknown’ value.

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

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: