ETL Project

Road Fatalities in Western Australia

ETL DATA ANALYSIS – PROJECT 2 – 13th March 2021

Participants: 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 frame to a SQL server.

Beginning by imporng environments for the code. This included the tradional Pandas

module, as well as SQLAlchemy and a few others. Next, we imported the module dotenv and

used the load\_dotenv to access our local environment and get the user and password to log

into the Postgres server.

Following module import, our team read in our data files from the aforemenoned sources,

using the os module as well as the read\_excel aribute from Pandas.

The Excel files can be found in the folder Resources.

Imporng the dataframes, we began the data cleaning process. Both of our datasets were

complete upon download. No rows were removed, and the datasets contain no null values or

rows. Most of the cleaning process involved the elecon data from this year. Mainly, the

process involved the renaming of columns, and reformang of the frame to simplify the

headers.

Formang column names and headers, the original header row from the imported dataset

was removed with df.drop , then reset the index of the dataframe to account for the

dropped row, and to ensure that the county indices will match our COVID-19 dataset index.

The next step was to join together our datasets. We merged the tables on the County

column, using the elecon dataframe index. This was an inner join. Aer the combinaon of

tables, we performed some final data cleaning and made some cosmec changes for

digesbility and clarity. This process included renaming COVID and votes columns to provide

clarity between the two values provided their close proximity within the dataframe.

Following our compleon of the data cleaning process, we shied our focus to uploading the

data to a SQL server. This process required some addional quick cleaning, which involved

renaming columns to match column names created in our SQL database. This was done to

allow for easier column calls within the database, considering using the original names would

degrade usability of the database in SQL.

The queries to create the tables can be found in sql file: etl\_project.sql.

The following process was used to connect, check the connecon and upload the data to the database: