ETL Project

Team 12: Victoria Traffic Analysis

A bridge over a river with a city in the background

Description automatically generated with medium confidence

An Nguyen , Seng Heang (Darren) Kim, Anjana Bandara

Table of Contents

[Overview 1](#_Toc71372186)

[Data Extraction 1](#_Toc71372187)

[Data Transformation 2](#_Toc71372188)

[Traffic Volume Data set 2](#_Toc71372189)

[The Hourly Traffic Volume Data set 3](#_Toc71372190)

[The Homogeneous Traffic Flow Data set 4](#_Toc71372191)

[Loading Data 5](#_Toc71372192)

[Analysis 6](#_Toc71372193)

[Traffic Volume Analysis 6](#_Toc71372194)

[The Hourly Traffic Volume 6](#_Toc71372195)

[The Homogeneous Traffic Flow 7](#_Toc71372196)

[Appendix 8](#_Toc71372197)

## Overview

The main focus of this project is to extract data from available sources and perform an ETL route on the extracted data. For this project we chose to analyse the behaviour of vehicle traffic in the state of Victoria. The three main focus areas for this project are,

* The traffic volume as a whole for Victoria
* The hourly traffic volume for Victoria
* Homogenous Traffic Flow

What we want to achieve through this analysis is, to outline the freeways and other linked roads to them with highest traffic volume. Furthermore, we will be outlining the highest traffic volume as per the time of the day. With the analysis of homogenous traffic flow, we will be creating a heat map with traffic flowing in and out of Victoria.

## Data Extraction

We extracted the **Traffic Volume** data through three sources in two formats.

Traffic volume data was extracted through an API from Victoria Open Data Hub. The data was extracted in json format. We had to develop a query URL to extract the data from the API.

Graphical user interface, text, application

Description automatically generated

**Hourly traffic volume** data was downloaded in a csv format through Kaggle.

Data source: <https://www.kaggle.com/daralm/hourly-traffic-volume-in-victoria>

**Homogenous Traffic Flow** was extracted from GeoJSON file, which is an open standard format designed for representing simple geographical features, along with their non-spatial attributes. It is based on the JSON format.

In order to read a GeoJSON file into GeoPandas DataFrame, we need to **import** geopandas as gdp

Then:



Data Source: https://discover.data.vic.gov.au/dataset/homogenous-traffic-flow

## Data Transformation

### Traffic Volume Data set

* Traffic volume data was extracted through the API and stored inside separate list. We extracted the data only needed for the analysis. For the data cleaning we used pandas library from python 3.

Text

Description automatically generated

* After storing the data in separate list, pandas were used to create a dataframe out of those lists.

Text

Description automatically generated

* We used pandas .isnull ().sum () function to find out all the missing values in the dataset. We found out there were eight rows with missing values. We dropped this rows using .dropna () function in pandas.
* Some of the columns were renamed after this step because we already had created the database tables and column names for each dataset.



* All the duplicate values were dropped after the above step.
* All column headers were changed into lower case cause the database table headers are in lower case.



* Finally, the cleaned dataframe was stored in a csv.

### The Hourly Traffic Volume Data set

* The csv was read using the pandas read\_csv function. The columns required for the analysis then extracted from the initial dataframe.
* The new dataframe headers are then renamed as per the table name headers in the database.

A picture containing text

Description automatically generated

* After this step, a function is created to transform day the day of the week column into the name of the day. And the function is applied for the whole day of the week column.

Graphical user interface, text, application, email

Description automatically generated

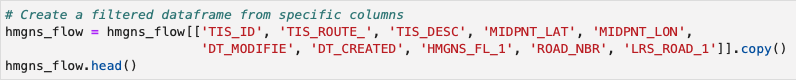
* Then the Dataframe is checked for duplicate values and the remaining duplicate values are dropped after that.

Graphical user interface, text, application, email

Description automatically generated

### The Homogeneous Traffic Flow Data set

This type of data after extracted into the DataFrame is quite simpler than the rest of the data set that we have. We have selected relevant data for our analysis as shown below:



## Loading Data

* PostgreSql was used to load the extracted and transformed data. We chose PostgreSql as our main database due to the compatibility it has with different data resources. And relational database seemed to be the best fit for our structured data.
* Data was loaded into the database using sqlalchemy create engine function.
* The dataset was merged together using the hmgns\_lnk\_id.

A picture containing text

Description automatically generated

**ERD for the SQL**

Graphical user interface, table

Description automatically generated with medium confidence

## Analysis

### Traffic Volume Analysis

Chart, bar chart

Description automatically generated

This graph shows the top ten link\_id’s with highest traffic volume in Victoria. Link id 297 has the highest traffic volume and link id 84 has the lowest volume. On average 1250 vehicles are passed on each link throughout Victoria.

### The Hourly Traffic Volume

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

These graphs show the traffic volume during school term and school holidays. As per the graphs, there’s a decrease in traffic volume during the school holidays. Weekends are always low with traffic but during seasonal functions traffic volume tend to increase.

### The Homogeneous Traffic Flow

Map

Description automatically generatedA map of a city

Description automatically generated with low confidence

Map

Description automatically generated

The homogenous traffic flow data is the extended dataset of traffic volume information which consists of latitude, longitude, and the traffic flow along a link that is representative of all travel along the whole link. Therefore, with the powerful tools in pandas, we are able to see the differences of the traffic flow in Victoria specifically on the map under various conditions.

## Appendix

**Extraction**

Table

Description automatically generated

**Transformation & Analysis – Filtering data**

**Graphical user interface, text, application

Description automatically generated**

**Loading to SQL: Database Connection and Data Frame Loading**

**Graphical user interface, text, application

Description automatically generated**

**Loading to SQL – Table Creation**

**Graphical user interface, text, application, email

Description automatically generated**