EGD103 Assignment 2 Brief

Teaching Period 1, 2025

Introduction

In this assignment you will be analysing traffic data for QLD. This data is collected with traffic cameras by the Department of Transport and Main Roads. You will analyse this data to answer questions about what factors influence traffic counts.

Datasets

The main dataset for this assignment 'qld_traffic_2023.csv' is publicly available data collected from the Queensland Government's Open Data Portal. It includes traffic count data for 2023 based on factors such as location, direction of travel, time and day of week. Later in the assignment you will access some traffic data from earlier years. The raw data was collected from `https://www.data.qld.gov.au/dataset/queensland-traffic-data-averaged-by-hour-of-day-and-day-of-week`.

A data dictionary named 'field-descriptions.csv' is available. This gives a description of each variable in the dataset. Please refer to it if you are unsure about what information is contained in each column.

Assignment Instructions

Assignment questions are given on the following pages in parts A and B. You will be required to create a single JupyterLab Python notebook that answers all the assignment questions. Your notebook should be neatly structured and set out. Markdown cells should be used to provide headings, describe the assignment questions, and provide worded responses to questions. Code cells should be used to perform the data processing needed to answer the questions. No template is provided, so you will be starting your notebook file from scratch. For assistance with Markdown, you can visit `https://www.markdownguide.org/basic-syntax/`, or inspect the Markdown cells used in lecture and tutorial templates for the unit.

You are permitted to use any imports you would like for this assignment. Please include all imports at the top of your notebook before completing the assignment questions.

Assignment Questions

Part A: Preparing and Wrangling Data (1 mark each)

1. Import the 'qld_traffic_2023.csv' data using pandas. Display the first 5 and last 5 rows of the dataframe to confirm it has been imported successfully.

2. Call the info method to see a basic summary of the dataset. Based on the output answer the following questions:

   - How many observations (rows) are in the dataset?
   - What data types appear in the dataset?

3. Display a summary that shows how many missing values are in each column of the dataframe.

4. To ensure we are working with clean data for our analysis, filter out all rows of the dataframe that contain at least one missing value.

5. Provide code that displays the number of unique sites and the number of unique road names that are contained in the dataframe.

6. Provide code that evaluates the minimum value in each of the count columns (MON - WEEKEND_AVERAGE) to ensure there are no negative values.

7. Add a computed column named `TOTAL` that contains the sum of the counts from all 7 days.

8. Currently time is recorded as an object in the `HOURS` column. It would be much better to have a column that used a numeric representation for time. Create a new column called `START_TIME` that contains the start time from the `HOURS` column interval represented as an integer.

Part B: Summarising Data (2 marks each)

Use data aggregation and visualisation techniques to answer the following questions about the data:

1. **How do traffic counts vary by day of the week?** Aggregate to find the total value of each of the daily columns (ie. MON, TUE, ..., SUN). Visualise how the counts vary by day of week with a vertical bar graph. Comment on any trends that are observed.

2. **How do traffic counts vary by time of day?** Aggregate to find the average value in the TOTAL column for each start time. Visualise how the traffic counts vary by start time with a line plot. Comment on any trends that are observed.

3. **How does location effect traffic counts?** Compute the total number of traffic counts observed for each site in the data. Visualise how these counts are affected by location (ie. latitude and longitude) with a scatter plot. The x and y axes should be used for longitude and latitude, and colour should be used to indicate the total count for that location.

4. **What are the 10 busiest roads, and what is their average weekly traffic count?** Find the total weekly traffic count for each road name in the data. Sort to find the road names with the 10 highest weekly traffic counts. Visualise the results with a horizontal bar graph.

5. **How do the traffic counts compare between 2022 and 2023?** In this problem you will need to access the data in 'qld_traffic_2022.csv'. Import this data, and then perform the relevant wrangling so that the average weekly total for each site and time interval can be compared for 2022 and 2023 (note: this will require joining/merging the data on the side id and time interval). Then compute the correlation between the coutns and visualise the relationship with a scatter plot. The scatter plot should include a $y = x$ reference line to show where the 2022 and 2023 counts would be equal. Comment on how the years compare.

6. **Can we create a program that provides information for the closest site to a particular location?** Create a user-defined function that accepts 2 inputs: a latitude and a longitude. The function should find the site that is closest to the location described by the function inputs. The function should then filter the data for that site only and display a line plot that shows the traffic counts for weekdays and weekends at that site as a function of time (2 lines total). Call your function with an example latitude and longitude to demonstrate that it works.

7. **Your own research question:** Create your own research question about the data. Your research question should be clearly articulated, meaningful, and novel. You must select and correctly implement appropriate summarisation and visualisation techniques to answer that question.