EGD103 Assignment 2 Brief

Teaching Period Summer, 2024

Introduction

In this assignment you will be analysing flight delay data for domestic flights in Australia. This data is collected by the Bureau of Infrastructure and Transport Research Economics (BITRE). You will analyse this data to answer questions about what factors affect flight delays.

Datasets

The main dataset for this assignment 'flight_delays.csv' is publicly available data collected from BITRE. It includes monthly aggregated flight delay data for each flight route and airline in Australia. Later in the assignment you will access some other flight data 'domestic_flights.csv' which is also available from BITRE.

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 'flights_delays.csv' data using pandas. The 'Month_Start' column should be imported in datetime format. Display the first 5 and last 5 rows of the dataframe.

2. Call the info method to see the data type of each variable. Use this to confirm that the 'Month_Start' column was successfully imported in datetime format.

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

4. Display all rows of the dataframe that contain missing values.

5. Provide code that displays the earliest and latest dates in the joined data to determine the date range.

6. Provide code that displays all the unique cities in the data and all the unique airlines.

7. Add a computed column to the dataframe that includes the year, and another computed column that includes the month name.

Part B: Summarising Data (2 marks each)

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

1. **How do flights and delays vary with time?** Find the total number of flights, departure delays and arrival delays for each month in the data. Then visualise how the monthly totals vary with time using a line plot. Comment on any patterns or trends observed.

2. **What proportion of domestic flights belonged to each airline in 2005?** Compute the total number of flights for each airline in 2005. Then create a pie plot that shows what ratio of 2005 flights belonged to each airline.

3. **How does airline affect the likelihood of your plane landing late?** Compute the total number of flights and arrival delays for each airline. Use this result to compute the arrival delay percentage for each airline. Sort these results by value and then visualise the results in a horizontal bar graph.

4. **How strongly do late arrivals correlate with late departures?** Compute the correlation between departures and arrivals. Construct a scatter plot that visualises the relationship between the two variables,

and also plot the function $y = x$ on the same axes to reference where late departures and arrivals are equal on the graph. From your results, comment on the relationship between late arrivals and departures.

5. **Can we create a program that gives an annual snapshot of arrival delays?** Create a user-defined function that accepts 4 inputs: a dataframe, a year, a city 1 string and a city 2 string. The function should filter the dataframe for the year and cities selected. It should then output a line plot that displays the arrival delays on each month of the year for each airline (ie. should have a separate line for each airline). The plot title should include the year and city names. Test your function by calling it on the data for an example date and region.

6. **Does the route distance have an effect of late arrivals?** In this problem you will need to access the data in 'domestic_flights.csv'. Import this data, and then perform the relevant wrangling on it so that you can join the route distances to the main dataset. You should then create a scatter plot that shows the relationship between route distance and the percentage of arrival delays.

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.