FIT1043: Assignment 1

Due: Sunday, 18 September 2016

The aim of this assignment is to investigate and visualize data. In particular, we will identify trends, gain insights from how those trends change over time and attempt to predict future values.

The data we will use contains Airport traffic data and comes from the Australian Bureau of Infrastructure, Transport and Regional Economics.

- The CSV file contains 30 years worth of monthly data listing international air traffic volumes in terms of passengers, freight and mail, to and from Australian airports.
- The file is available on Moodle and publicly available from data.gov.au at the url: https://www.data.gov.au/dataset/international-airlines-traffic-by-city-pairs

Hand-in Requirements

Please hand in a **PDF file** containing your answers to all the questions.

- You can use Word or other word processing software to format your submission. Just save the final copy to a PDF before submitting.
- Make sure to include screenshots/images of the graphs you generate in order to **justify your answers** to all the questions. (Note that Google Sheets doesn't currently allow users to save Motion Charts as images, so you will need to use screen-capture functionality to create appropriate images.)
- We would like to see the Google Sheets you use to generate the Motion Charts and the Python code you write to format the data. Please **include links to your sheets, and a copy your Python code** in your submission, (either submitting scripts / Jupiter notebooks, or copying the code into your report).

Python Availability

You will need to use Python to complete the assignment. You can do this by either:

1) running a Jupyter Notebook on a computer in the labs;
2) accessing a Jupyter Notebook sever at https://jupyterhub.erc.monash.edu; or
3) installing Python (we recommend Anaconda) on your own machine.

Assignment Tasks

Parts A to C of the assignment involve building a series visualisations using (GapMinder-style) Motion Charts in Google Sheets and then answering a series of questions about the data. Parts D and E involve graphing the data in Python and answering further questions. Part F involves investigating a different dataset.

Part A: Visualising Airport Traffic

The aim of the first part of the assignment is to build a Motion Chart showing passenger, freight and mail quantities over time for different airports in Australia. More specifically, you should:

- Generate a Motion chart with Passengers_Total on the x-axis, Freight_Total_(tonnes) on the y-axis and where the color the bubbles is given by the Mail_Total_(tonnes).
- Note that the CSV file is both too big and not in the right format to visualise directly using Google Sheets, so you will first need to aggregate the data **using Python** at both the 'Year' and 'AustralianPort' level. (Code to help you do this was discussed during the lectures.)
- You can then output the data as a CSV file, which you can upload to Google Sheets in order to create the appropriate visualisation.

Having graphed the data as a Motion Chart, answer the following questions:

1. Why are the values lower in 2016 than they were in 2015?
2. Which city has the largest number of international air passengers traveling through it in 2015?
3. In which year did Brisbane have almost the same number of passenger numbers as Melbourne?
4. Has the number of passengers travelling through Sydney airport ever decreased from one year to the next? If so, when did it happen? Any idea why that might have occurred?

Part B: Comparing Countries

Repeat the above steps to generate a new spreadsheet and visualization, this time showing the data aggregated by Country and Year, and then answer the following questions:

1. Which country had the largest total passenger numbers coming from or to Australia in 2015?
2. Which country had the largest freight volume in 2015? Why do you think that is?
3. Which country had the largest mail volume in 2015? Why might that be?
4. In which year did Singapore overtake New Zealand in terms of the freight volume?

Part C: Visualizing Foreign Ports

Repeat the above steps to generate a new spreadsheet and visualization, this time showing the data aggregated by ForeignPort and Year.

1. When did Emirates start flying to Australia? How do you know that? [HINT: You may want to change the scale on the X and Y axes to be logarithmic (Log) rather than linear (Lin).]
2. In 2015, we see a number of cities with zero passengers but significant amounts of freight. Which city sees the most freight (but no passengers). Why would so much freight go to that location? [HINT: you'll need to Web search to find the answer to the last question.]

Part D: Time of Year Analysis

We will now investigate the amount of traffic flowing through different ports at different times of the year. To do this you will need to aggregate the data at the month level. **Use Python to plot** the average quantity of postage both TO and FROM the US over the different months of the year. Don't forget to add axes to both plots.

1. Which month of the year are the most mail packages sent both to and from Australia? Why do you think that is?
2. Overall, do we see more mail volume entering Australia from the US or leaving Australia for the US?
3. Bonus Challenge Question 1: (**NB: This question is difficult and should not be attempted before completing all other parts of the assignment including Part F.**) The graph above doesn't take the yearly growth trend (investigated in Part E) into account when calculating the monthly variation and could be inaccurate as a result. To fix this problem, we need to remove the trend information from values before calculating the monthly average. We can do this by dividing the mail volume for each month and year by the average value across that year[1] BEFORE aggregating the data across years. The challenge is to generate that graph. In time series modeling, the values you are computing are referred to as multiplicative seasonal effects.

[1] Ideally the average value across months of the year should be centered around the particular year, (i.e. include the 6 months before and 5 months after the current month), although averaging over calendar years will likely produce very similar results.

Part E: Predicting Future Traffic Volumes

Graph **in Python** the total annual passengers at Sydney Airport against year.

1. Does the data show a clear trend? If you don't see one, try plotting the data after aggregating at the year level.
2. Is there a problem with one of the datapoints? If so, why is that?
3. Remove the problematic datapoint and run a simple linear regression in Python by modifying the code from the Python tutorial. Does the linear fit look to be a good fit to you?
4. How fast are international passenger numbers increasing each year? [Hint: What is the slope of the linear fit above?]
5. What does the linear model predict for passenger volume in 2020? [HINT: Get the slope (m) and intercept term (c) for the linear fit above and use the function Y=m*X + c, to calculate the prediction for X=2020.]
6. Try fitting the linear model only to the data from the year 2004 onwards. What happens to the prediction for 2020? Which prediction do you trust more? Why?
7. Bonus Challenge Question 2: (**Attempt only after completing Part F.**) Use the estimates of multiplicative seasonal effects from the previous challenge question to predict the monthly passenger volumes for 2020, by (i) dividing the predicted annual passenger numbers by 12 to calculate the monthly average, and (ii) multiplying by the appropriate seasonal effect value. Plot the resulting predictions.

Part F: Repeat the Analysis on Another Dataset

As discussed in the lectures, there is a huge amount of public data available online. For example, the Australian, US, UK, Singapore and Indian governments all provide websites with links to datasets:

- https://www.data.gov.au/
- https://www.data.gov/
- https://data.gov.uk/
- https://data.gov.sg/
- https://data.gov.in/

And Kaggle, a private company which runs data science competitions, also provide a list of their publicly available datasets:

- https://www.kaggle.com/datasets

Your task is to find some interesting data and do an analysis similar to Parts A through E above. Look for some data **with a temporal component** and:

1. Aggregate it as required using Python.
2. Interpret it using a Motion Chart.
3. Predict future values using linear regression.