FIT5145 Introduction to Data Science Summer B 2018

Assignment 1 – 11:55pm 17th January 2018

This assignment aims to allow you to investigate and visualise data using various data science methods. It will test your ability to:

1. Read data files in Python (or R) and extract related data from those files;
2. Wrangle and process data into the required formats;
3. Use various graphical and non-graphical tools for performing exploratory data analysis and visualisation;
4. Use basic tools for managing and processing big data; and
5. Communicate your findings in your report.

You are free to use either Python or R to complete the tasks, although we recommend that you use Python because many of the required data wrangling and processing methods needed are included in the earlier Python tutorials.

You will need to submit a Jupyter notebook file in either Python or R containing the code you have written and the answers to each question. Marks will be assigned to reports based on their correctness, presentation and clarity. For example, higher marks will be given to reports containing graphs with appropriately labelled axes. As a general guide, a maximum of 2 marks are available for each plot and up to 1 mark for each of the associated questions.

Please include section headings and question numbers in your report.

Task A: Investigating Job Vacancy and Unemployment Rate Data (58 Marks total)

In the task, you are required to visualise the relationship between the number of job vacancies, the unemployment rate, Australian Graduate employment rates, the group of Eight Universities enrolments and the population of different Australian states. You will gain insights from how these relations and trends change over time. The data files used in this task were downloaded from the Commonwealth Department of Employment, Commonwealth Department of Education and Australian Bureau of Statistics. We have extracted the data from the original files and put into a simpler format. You can download the data from Moodle:

Resident Population Data *(EstimatedResidentPopulationByStateAndGender.csv):* This file contains quarterly data regarding the estimated resident population, grouping by state and gender, between 1/12/2005 and 1/6/2015

Student Enrolment Data *(StudentEnrolment.csv):* This file also has the number of students enrolled each year at eight Australian Universities:

- University of Adelaide (AU)
- University of Western Australia (UWA)
- University of NSW (UNSW)
- University of Sydney (USYD)
- University of Melbourne (Uni Melb)
- Australian National University (ANU)
- Monash University (Monash)
- University of Queensland (UQ)

Job Vacancies Data *(JobVacancies.csv):* This data file contains monthly data about the number of job vacancies (based on a count of online job advertisements newly lodged on SEEK, CareerOne and Australian JobSearch) across different Australian states, for the period between 1/1/2006 and 1/10/2016.

Unemployment Data *(EmploymentTimeSeries.xls):* This data file contains monthly data of employment and unemployment rate across different Australian states, for the period between 01/02/1978 and 01/10/2016.

Graduate Employment Data *(GraduateEmploymentTimeSeries.xls*): This data contains the graduate unemployment and employment rates per quarter for the Group of Eight universities since 1978.

Insurance Rates Data (InsuranceRates.csv.zip) contains data on health and dental plans offered to individuals and small businesses through the US Health Insurance Marketplace.

Do not change the names of these files.

A1. Investigating the Population Data (8 marks)

Have a look at the resident population data. We are only interested in the total values for each state (marked "Persons").

1. Create a data frame with these columns and rename the columns for each state.

2. In Python (or R) plot the population of Victoria, New South Wales and Queensland over time. You don't need to put the dates on the x-axis, just showing the index of each quarter is fine.

    a. Are the population values increasing or decreasing over time?
    b. Does the population data exhibit a trend and if so, what type?

3. Fit a linear regression to the Victorian population data and plot the linear fit. (HINT: In Python, you can use the "range(1,n)" function to generate a sequence of integer values: 1,2,...,n.)

    a. There should be a good linear fit to this data. What does this regression model predict the resident population in Victoria will be for the following dates: 1/9/15, 1/12/15, 1/12/16 and 1/12/17?

A2. Investigating Enrolment Data (10 marks)

Next have a look at the student enrolment data. The data is in the StudentEnrolment.csv file.

1. Create a data frame with each of the university enrolments and dates.

2. In Python (or R) plot the enrolments of eight universities over time. You don't need to put the dates on the x-axis, just showing the index of each quarter is fine.

    a. Which of the enrolment values are obviously increasing or decreasing over time? Comment on the rate of change.
    b. What state has the university with the highest enrolments? What university is this?
    c. Why are these universities more likely to have such high enrolments? (HINT: Re-visit your graph of the population data)

3. Fit a linear regression to the Victorian students enrolment data and plot the linear fit. (HINT: In Python, you can use the "range(1,n)" function to generate a sequence of integer values: 1,2,...,n.)

    a. Do these linear fits look good? Why does one linear fit looks better?
    b. What does the model predict Uni Melb and Monash's enrolments will be for the following dates: 1/12/15, 1/12/16 and 1/12/17.

A3. Investigating the Job Vacancies Data (10 marks)

Now have a look at the job vacancies data.

1. Create a data frame and include the dates.

2. Plot the job vacancies' of each state over time. You don't need to put the dates on the x-axis, just showing the index of each day is fine.

    a. What are the maximum and minimum values for job vacancies in Victoria over the period?

3. Fit a linear regression to the data and plot it.

    a. Does it look like a good fit to you? Would you believe the predictions of the linear model going forward?

4. Instead of fitting the linear regression to all of the data, try fitting it to just the most recent data points (say from the 90th data point onwards.)

    a. How is the fit? Which model would give better predictions of future vacancies do you think?

A4. Investigating the Unemployment Data (5 marks)

1. Create a data frame, including the dates.

2. Plot the unemployment data of each state over time. You don't need to put the dates on the x-axis, just showing the index of each day is fine.

3. It looks like the rate has been very high at times in the past..

    a. What was the maximum unemployment rate in Victoria recorded in the dataset?
    b. When did that occur?

A5. Investigating the Graduate Employment Data (5 marks)

1. Create a data frame. Include the dates.

2. Plot the unemployment data of each university over time. You don't need to put the dates on the x-axis, just showing the index of each day is fine.

    a. What do you notice about this graph in comparison to your other investigations?
    b. What conclusions can you draw from this?

A6. Visualising the Relationship between Unemployment and Job Vacancies (12 marks)

Now let's look at the relationship between unemployment levels and job vacancies.

1. Combine the data from the different files into a single table. The table should contain population values, job vacancy counts and unemployment rates for all of the different dates and different States/Territories.

   a. What is the first date and last date for the combined data?

2. Now that you have the data aggregated, we can see whether there is a relationship between unemployment and the number of job vacancies. Plot the values against each other.

   b. Can you see a relationship there? Is the relationship clear?

3. Select and plot only the data from Victoria.

   c. Can you see a relationship now? If so, describe this relationship based on the graph?

4. The different populations of the states will influence the number of job vacancies in each. Remove this effect by introducing a new column called 'Vacancy Rate' which contains the vacancy count divided by the population size, multiplied by 100 (2 marks).

5. Re-plot the new data.

   a. Is there a relationship between the unemployment rate and the job vacancy rate across all the data?

A7. Visualising the Relationship over Time (8 marks)

Now let's look at the relationship between unemployment levels and job vacancies over time.

1. Build a Motion Chart comparing the job vacancy rate, the unemployment rate, and the population of each state over time. The motion chart should show the job vacancy rate on the x-axis, the unemployment rate on the y-axis and the bubble size should depend on the population.

2. Run the visualisation from start to finish.

   a. Which state has the lowest job vacancy rate?
   b. Is the economy getting better or worse? I.e. was the Australian economy better in 2006/7 or 2014/5?
   c. Compared to the states, does the Northern Territory have higher or lower unemployment and higher or lower job vacancy rates?
   d. What might cause this? Would it make sense economically to move to NT?
   e. According to the graph, what happened at the end of 2008 and start of 2009? What might have caused this?
   f. Any other interesting things you notice in the data?

Task B: Exploratory Analysis on Big Data (32 Marks total)

In this part, you are required to do some exploratory analysis on the health insurance marketplace data. The file InsuranceRates.csv.zip contains data on health and dental plans offered to individuals and small businesses through the US Health Insurance Marketplace. This data was originally prepared and released by the Centers for Medicare & Medicaid Services (CMS), so please read the CMS Disclaimer-User Agreement before using this data. The data was then published on Kaggle. The file we provide is an extract from the data on Kaggle. Unzipped, the file is over 500MB and contains the following fields:

COLUMN	DESCRIPTION
BusinessYear	Year for which plan provides coverage to enrolees.
StateCode	Two-character state abbreviation indicating the state where the plan is offered
IssuerId	Five-digit numeric code that identifies the issuer organization in the Health Insurance Oversight System (HIOS)
PlanId	Fourteen-character alpha-numeric code that identifies an insurance plan within HIOS
Age	Categorical indicator of whether a subscriber's age is used to determine rate eligibility for the insurance plan.
IndividualRate	Dollar value for the insurance premium cost applicable to a non-tobacco user for the insurance plan in a rating area, or to a general subscriber if there is no tobacco preference.
IndividualTobaccoRate	Dollar value for the insurance premium cost applicable to a tobacco user for the insurance plan in a rating area

B1. Summarising the Data (9 marks)

Load the InsuranceRates.csv data in Python (or R).

1. Now answer the following questions:

    a. What are the possible values for 'Age'?
    b. How many states are there?
    c. How many rows and columns are there?
    d. How many years does the data cover? (Hint: pandas provides functionality to see unique' values.)
    e. How many unique ‘PlanId’s are there?
    f. What are the average, maximum and minimum values for the monthly insurance premium cost for an individual?
    g. Do those values seem reasonable to you?
    h. How many rows that have a non-empty ‘IndividualTobaccoRate’ are there?
    i. Among these rows, how much more on average do smokers pay for their insurance?

B2. Investigating Individual Insurance Costs (7 marks)

Now let's look more in detail at the individual insurance costs.

1. Show the distribution of ‘IndividualRate’ values using a histogram.

    a. Does the distribution make sense to? What might be going on?

2. Remove rows with insurance premiums of 0 (or less) and over 2000. (Use this data from now on.) Generate a new histogram with a larger number of bins (say 200).

    a. Does this data look more sensible?
    b. Describe the data. How many groups can you see?

B3. Variation in Costs over Time and with Age (4 marks)

1. Generate bar charts of insurance costs versus year and age.

    a. Are insurance policies becoming cheaper or more expensive over time? Is the median insurance cost increasing or decreasing?
    b. How does insurance costs vary with the age of the person being insured? (Hint: filter out the value 'Family Option' before plotting the data.)

B4. Variation in Costs across States (12 marks)

How do insurance costs vary across states?

1. Generate a graph containing boxplots summarising the distribution of values for each state.

    a. What is the difference between the lowest and the highest median insurance rates?
    b. How does the number of insurance issuers vary across states?

2. Create a bar chart of the number of insurance companies in each state to see. (Hint: you will need to aggregate the data by state to do this.)

    a. Which states have the lowest and highest number of insurance issuers respectively?
    b. Could competition explain the difference in insurance premiums across states?

3. Use a scatterplot to plot the number of insurance issuers against the median insurance cost for each state.

    a. Do you observe a relationship?
    b. Which state has the most expensive insurance issuer (in terms of median ‘IndividualRate’)?