**Assignment: Description**

The aim of this assignment is to investigate and visualize data using various data science tools. 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 to performing exploratory data analysis and visualization;
4. use basic tools for managing and processing big data; and
5. communicate your findings in your report.

You will need to submit two files:

1. A **report in PDF** containing your answers to all the questions. Note that 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. (Marks will be assigned to reports based on their correctness and clarity. -- For example, higher marks will be given to reports containing graphs with appropriately labeled axes.)
2. The **Python code** as a Jupyter notebook file (or R code as a script or notebook file) that you wrote to analyse and plot the data.

**Tasks:**

There are two tasks that you need to complete for this assignment, plus a third (optional) task for higher credit.

You are free to use either Python or R to complete the tasks, although we recommend that you use Python. You may also do some preprocessing of the data in Excel if you wish (see suggestions below).

**Task A: Investigating Job Vacancy and Unemployment Rate Data**

In the task, you are required to visualize the relationship between the number of job vacancies, the unemployment rate, and the population of different Australian states, and gain insights from how these relations and trends change over time. The data files used in this task were originally downloaded from [the Department of Employment](https://employment.gov.au/) and [Australian Bureau of Statistics](http://www.abs.gov.au/). We have extracted the data from the original files and put into a simpler format. You can download the data from Moodle:

* **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.
* **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.
* **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.

**A1. Investigating the Population Data**

Have a look at the resident population data. You will see many columns. We are interested only in the total values for each state (marked "Persons"), so you can drop the other columns and rename the columns for each state if you wish. (HINT: The file isn't very big so you can make the changes in Excel if you want.)

1. In Python plot the population of Victoria, New South Wales and Queensland over time. (HINT: You don't need to put the dates on the x-axis, just showing the index of each quarter is fine.)
   * Are the population values increasing or decreasing over time?
   * Does the population data exhibit a trend and if so, what type?
2. Fit a linear regression using Python 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.)
   * Does the linear fit look good?
   * Use the linear fit to predict the resident population in Victoria for the dates: 1/9/15, 1/12/15, 1/12/16, and 1/12/17.

**A2. Investigating the Job Vacancies Data**

Now have a look at the job vacancies data.

1. Use Python to plot the job vacancy counts for Victoria over time. (HINT: Pandas contains a "transpose()" method and Excel can also be used to transpose data.)
   * What are the maximum and minimum values for job vacancies in Victoria over the time period?
2. Fit a linear regression to the data and plot it.
   * Does it look like a good fit to you? Would you believe the predictions of the linear model going forward?
   * Instead of fitting the linear regression to all of the data, try fitting it to just the most recent datapoints (say from the 85th datapoint onwards). How is the fit? Which model would give better predictions of future vacancies do you think?

**A3. Investigating the Unemployment Data**

Now have a look at the unemployment data.

1. Use Python to plot the Unemployment Rate **for Victoria** over time.
   * It looks like the rate has been very high at times in the past. What was the maximum unemployment rate in Victoria recorded in the dataset and when did that occur?

**A4. Visualizing the Relationship between Unemployment and Job Vacancies**

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

1. Use Python to combine the data from the different files into a single table. The table should contain population values, job vacancy counts and unemployment rates for the different dates and different States/Territories.
   * 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.
   * Can you see a relationship there?
3. Try selecting and plotting only the data from Victoria.
   * Can you see a relationship now? If so, what relationship is there?
4. The different populations across 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.
   * Is there a relationship between the unemployment rate and the job vacancy rate across all the data?

**A5. Visualizing the Relationship over Time**

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

1. Use Python to 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. (HINT: A Jupyter notebook containing a tutorial on building motion charts in Python is [available here](https://www.alexandriarepository.org/wp-content/uploads/20170628102639/MotionChart-Activity.zip).)
2. Run the visualization from start to finish. (Hint: In Python, to speed up the animation, set timer bar next to the play/pause button to the minimum value.) And then answer the following questions:
   * Which state generally has the lowest job vacancy rate?
   * Is the economy generally getting better or worse? I.e. was the Australian economy better in 2006/7 or 2014/5? Explain your answer.
   * Compared to the states, does the Northern Territory generally have higher or lower unemployment and higher or lower job vacancy rates? What might cause this? Would it make sense economically to move to NT?
   * According to the graph what happened at the end of 2008 and start of 2009? What might have caused this?
   * Any other interesting things you notice in the data?

**Task B: Exploratory Analysis on Big Data**

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)](https://www.cms.gov/cciio/resources/data-resources/marketplace-puf.html), so please read the [CMS Disclaimer-User Agreement](https://www.cms.gov/CCIIO/Resources/Data-Resources/Downloads/Data-Disclaimer-User-Agreement.pdf) before using this data. The data was then published on [Kaggle](https://www.kaggle.com/). 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 enrollees. |
| 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. Summarizing the Data**

Load the InsuranceRates.csv data in Python (or R) and answer the following questions:

1. How many rows and columns are there?
2. How many years does the data cover? (Hint: pandas provides functionality to see 'unique' values.)
3. What are the possible values for 'Age'?
4. How many states are there?
5. How many insurance providers are there?
6. What are the average, maximum and minimum values for the monthly insurance premium cost for an individual? Do those values seem reasonable to you?
7. How much more on average do plans for smokers cost?

**B2. Investigating Individual Insurance Costs**

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

1. Show the distribution of ‘IndividualRate’ values using a histogram.
   * 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).
   * Does this data look more sensible?
   * Describe the data. How many groups can you see?

**B3. Variation in Costs across States**

How do insurance costs vary across states?

1. Generate a graph containing boxplots summarizing the distribution of values for each state.
   * Which state has the lowest median insurance rates and which one has the highest? (Hint: you may need to rotate the state labels to be able to read the plot.)
   * Is there much variation in costs across states?
2. Does the number of insurance issuers vary greatly across states?
   * 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.)
3. Could competition explain the difference in insurance premiums across states?
   * Use a scatterplot to plot the number of insurance issuers against the median insurance cost for each state.
   * Do you observe a relationship?

**B4. Variation in Costs over Time and with Age**

Generate boxplots (or other plots) of insurance costs versus year and age to answer the following questions:

1. Are insurance policies becoming cheaper or more expensive over time?
   * Is the median insurance cost increasing or decreasing?
2. How do insurance costs vary with the age of the person being insured? (Hint: filter out the value 'Family Option' before plotting the data.)
   * Do older people pay more or less for insurance than younger people? How much more/less to they pay?

**Task C: Exploratory Analysis on Other Data**

Find some publicly available data and repeat some of the analysis performed in Tasks A and B above. Good sources of data are government websites, such as: data.gov.au, data.gov, data.gov.in, data.gov.uk, ...