README File for Python Code

Assessing Melbourne's Employment and Housing Correlations

Patrick Tjahjadi

The Python code is written from the Jupyter Notebook environment. There is one .ipynb file named 890003 associated with the code, but is separated into 17 blocks to simplify the code and make the data processing less condensed.

Summary of Python libraries used:

pandas – Used for reading the datasets and converting them into DataFrames.

numpy – Used to implement the VAT algorithm (More on 17th block) and to plot the line of best fit for the scatter plot to measure correlation.

matplotlib – Used to plot visualisations such as a scatter plot, bar graph, parallel coordinates and the VAT visualisation.

seaborn – Used to plot the VAT visualisation.

re – Used to remove the thousand and dollar separators from property prices so that they may be used for calculation, since it was impossible to calculate them in string format.

decimal – Used to convert the property prices into decimal form (which will be converted to floating point format for better use later).

Scipy – Used to measure the Pearson correlation coefficient and p-value of two variables. Also used to implement the VAT algorithm (More on 17th block).

1st Block

Import all libraries required for the study. The libraries used are **pandas, matplotlib, numpy, scipy, re, decimal** and **seaborn.** The purpose of these libraries is mentioned in the code as comments.

Next, define the years (2002 – 2016) and employment sectors (Business, Public Services, IT, Housing, Operations and Others) that are covered into lists. Read the csv files used for this study (2 datasets). Moreover, define the names of colours to be plotted in the visualisation later on.

2nd Block

Gather total number of people employed in each employment industry for each year. This means finding the value of each employment industry for the City of Melbourne (total) for each year (2002 – 2016). Add all the 20 employment industries and also the total number of employment combined.

This is method 1 as mentioned from the Integration phase of the report.

3rd Block

List the prices for house/townhouse and residential apartment for each year from 2002 to 2016. Put the values into lists starting from 2002 to 2016.

4th Block

Divide the employment industries into 6 sectors: Business, Public Services, IT, Housing, Operations and Others. Put the values into lists and gather the total number of people employed for each sector per year.

Next, list the number of people employed for each sector per year.

5th Block

Plots the parallel coordinates for the number of employment per year for each sector starting from 2002 to 2016. This is done using the **matploblib** library.

6th Block

Gather all instances of the number of people employed in each employment industries and separate them into 6 sectors.

This is method 2 as mentioned from the Integration phase of the report.

7th Block

Find the average price of house and apartment for each year. Since the dataset stores the price in string format using the dollar sign '$', it is impossible to calculate the average without converting them into floating point format. This is done using **re** and **decimal**, to remove the dollar sign and thousand separators.

Note: there is a **try except TypeError** code to prevent a TypeError should the code be run more than once (because the dollar sign and thousand separators are already removed).

8th Block

Calculate the Pearson correlation coefficient and its p-value for apartment price, house price and average housing price against the total number of people employed per year. This is done using **scipy.**

9th Block

Plot a boxplot of the instances of people employed in each sector. This is meant to detect outliers and growth of the number of people employed over the years using **matplotlib.**

10th Block

Provide a bar chart of the price of houses in Melbourne from 2002 to 2016 using **matplotlib.**

11th Block

Provide a bar chart of the price of apartments in Melbourne from 2002 to 2016 using **matplotlib.**

12th Block

Provide a scatter plot of the number of people employed vs apartment prices for each year. The corresponding Pearson coefficient and p-value is displayed.

13th Block

Provide a scatter plot of the number of people employed vs house prices for each year. The corresponding Pearson coefficient and p-value is displayed.

14th Block

Provide a scatter plot of the number of people employed vs average property prices for each year. The corresponding Pearson coefficient and p-value is displayed. Moreover, plot a regression line to detect correlation for this scatter plot. Libraries used were **matploblib** and **numpy**.

15th Block

Find the percentage increase in employment and property prices from 2002 to 2016. The value is kept in percentage with 2 decimal places.

16th Block

Normalise the values of all employment sectors. This is required for the VAT visualisation and clustering. The normalisation formula is given by:

$$Normalised\ Value = \frac{Old\ Value - Minimum\ Value}{Maximum\ Value - Minimum\ Value}$$

17th Block

Apply the VAT algorithm adapted from http://www.ece.mtu.edu/~thavens/code/VAT.m. Plot the VAT visualisation to detect clusters. This is done using **numpy, scipy** and **seaborn**.