Assessing Melbourne's Employment and Housing Correlations

Patrick Tjahjadi

Domain

The domain of the study covers employment and real estate.

Question

The aim of this report is to assess the correlation between the number of people employed and the price of housing within a particular time period. This report aims to answer the following question: "**Do the increase in employment rates contribute to the rise of house prices in Melbourne?**" With the steady increase of house pricing in Melbourne over the past decade, this report studies the trend from 2002 to 2016 and assesses employment rate as a factor in changing house prices. Although this report only considers the City of Melbourne area, this information can be used as a reference for other suburbs all across Victoria and Australia as geographical area will be abstracted away.

Datasets

This report uses the following 2 datasets:

1. 'Median House Prices – By Type and Sale Year': Contains data about the median house and residential apartment prices from 2000 to 2016 for the City of Melbourne. This dataset provides not only the median price but also the number of transactions for each type of residential housing each year. This dataset is extracted from the City of Melbourne website. Link: https://data.melbourne.vic.gov.au/Property-Planning/Median-House-Prices-By-Type-and-Sale-Year/i8px-csib
2. 'Employment by Block by Industry': Contains data about the number of people employed within a particular block, classified by industry (ANZSIC1) and yea**r**. The 20 sectors provided are considered for this report to assess which employment sector affects house pricing the most. It covers the year 2002 to 2016 for the City of Melbourne area. This data is also extracted from the City of Melbourne website. Link: https://data.melbourne.vic.gov.au/Economy/Employment-by-block-by-industry/b36j-kiy4

These datasets were chosen because they are both provided by the City of Melbourne website, which is a reliable, credible and accurate source. Moreover, the datasets were first assessed so as to be deemed suitable for this study (More on **Pre-processing**). Both datasets are in CSV format.

Pre-processing

Prior to processing the data, several pre-processing methods were done to ensure consistency. Firstly, both CSV files have to be opened in Excel and screened for any missing data, outliers and formatting used for highlighting or beautification (different text colour, font size, etc.). Given the huge sets of data provided by the two sources, it is crucial for the datasets to be cleaned so that only the necessary data for this report is used. Hence, pre-processing data was all done through Microsoft Excel (For modifying data entries in the CSV) and Python (For overall data processing). Several methods were done for this pre-processing phase and its limitations:

1. Unnecessary data was removed. This included the data that were recorded prior to 2002. Since the number of employment dataset only recorded the year starting from 2002, the median house prices for 2000 and 2001 in the dataset, which are the first four rows, are not included in this report. The number of house transactions are also not required.

2. For missing data in ‘Employment_by_block_by_industry.csv’, it is assumed that no people work for that particular industry within that time period and suburb. Hence, blank entries are given the number zero in Microsoft Excel.
3. Outliers detected in the datasets are best left unchanged because they may provide interesting information within a particular time period and suburb. Moreover, these outliers are assumed to be accurately recorded due to the City of Melbourne being a reputable source. Several interesting inferences could then be made due to those outliers and is hence better left unchanged. Outliers are detected by a boxplot visualisation (**Figure 5**).
4. For prices in dollars, the dollar signs and thousand separators are removed in Python using the libraries **re** and **decimal**. This will allow arithmetical operations for the ‘Median_House_Prices_-_By_Type_and_Sale_Year.csv’.

In detail, before plotting the graphs used for this report, the data was processed through Python by means of a **for** loop. The data wrangling process can be separated into 2 different methods: one to instantly append all the number of employment of each sector per year into lists for the “City of Melbourne (total)” to gather total number of employment per year compared to house prices, and another to iterate through the entire dataset by block except the “City of Melbourne (total)”, to gather instances of number of employment visible through a boxplot (**Figure 5**).

Integration

To reduce the number of columns required for visualisation, the 20 employment industries were classified into different categories, namely: Business Sector for ‘Business Services’, ‘Finance and Insurance’, ‘Retail and Trade’ and ‘Wholesale Trade’; Public Services Sector for ‘Education and Training’, ‘Electricity, Gas, Water and Waste’, ‘Health Care’ and ‘Public Administration and Safety’; IT Sector for ‘Admin and Support’ and ‘Information Media and Telecommunications’; Housing Sector for ‘Accommodation’ and ‘Real Estate Services’; Operations Sector for ‘Construction’, ‘Manufacturing’, ‘Rental and Hiring’ and ‘Transport, Postal and Storage’; and the Others Sector for ‘Arts’, ‘Agriculture’, ‘Food and Beverage’ and ‘Other’. This will allow simpler visualisation and analysis compared to having 20 comparisons.

Before the datasets are processed, the two CSV files were combined together and read via Python’s **pandas** library. The pre-processing and integration process hence adds value for future processing. However, several limitations through the integration process can be found:

1. Subjective splitting of sectors – The method on how to classify the 20 employment industries into 6 different sectors is categorical. There is no one clear definition on which employment industry belongs to which sector.
2. Redundant code for graphing – Since in order to plot the visualisation for different years, the study involves manually creating 15 lists for data, corresponding to each year.
3. Missing data issues – There are various missing data that are present for the number of employment. Even by replacing them with zero, there are too many industries and blocks with zero number of employment. Due to this, the VAT visualisation for the number of people employed in the employment sectors produces 6 notable clusters in a peculiar behaviour (**Figure 1**). Moreover, the boxplot visualisation is vague in determining outliers (**Figure 4**).

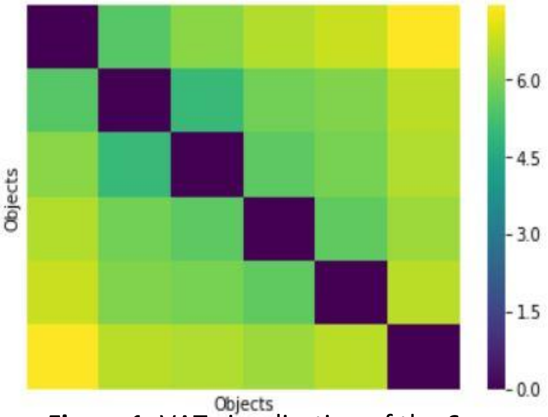

Figure 1: VAT visualisation of the 6 employment sectors

Results

Firstly, this report must assess the change in house pricing from 2002 to 2016 in Melbourne by means of a bar graph. The price of housing has notably increased since 2002.

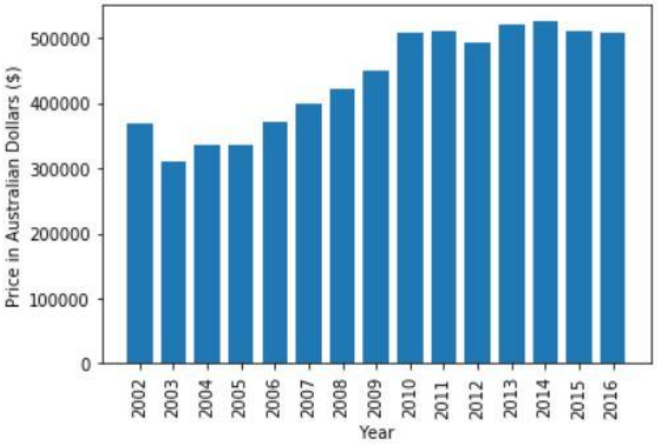

Figure 2: Median Apartment Price in Melbourne

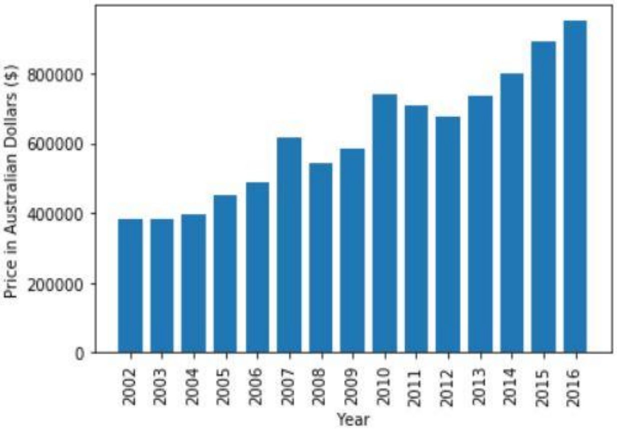

Figure 3: Median House Price in Melbourne

Needless to say, housing prices have increased steadily over the past decade (**Figure 2, Figure 3**). Interestingly, however, there are some years in-between that have higher or lower median house prices. Through speculation, the increase in the country's economic growth, along with the increase in people's purchasing power has led to more demand than supply and enabled people to purchase housing. Hence, this has led house prices to increase to reduce the shortage of housing supply. This is further backed by the fact that there is indeed a rise in the number of employment in Melbourne since 2002 (**Figure 4**). With more people being employed, the purchasing power of people increases, forcing property to increase its prices to maintain competitiveness.

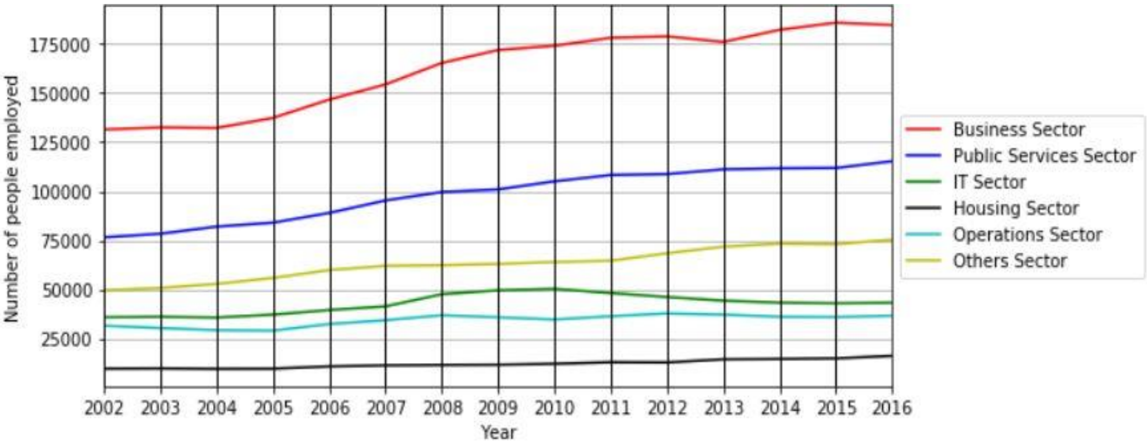

Figure 4: Number of employment per year for each sector (City of Melbourne)

Next, this report assesses the clusters found by normalising the six employment sectors. The normalisation process was done by applying the following formula through all instances:

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

After normalisation, the clusters are detected by means of a VAT visualisation using a VAT algorithm. This was initially done to assess the clusters of each sector. However, the result has a peculiar behaviour of having 6 notable clusters (**Figure 1**). Through this peculiar behaviour, it is difficult to assess the predictability of the clusters and to analyse the clusters any further.

Interestingly, the number of people employed in the business sector has experienced the most growth. This can also be proven by means of a boxplot by noticing that the business sector has indeed the biggest spread of data points. The purpose of the boxplot is to measure all the instances of the number of people employed in a particular sector for each block and year. However, since there are many zero entries in the dataset, the depiction of outliers is vague but can still be detected. Nevertheless, the visualisation gives an interesting insight about the number of people employed in each sector with its outliers (**Figure 5**).

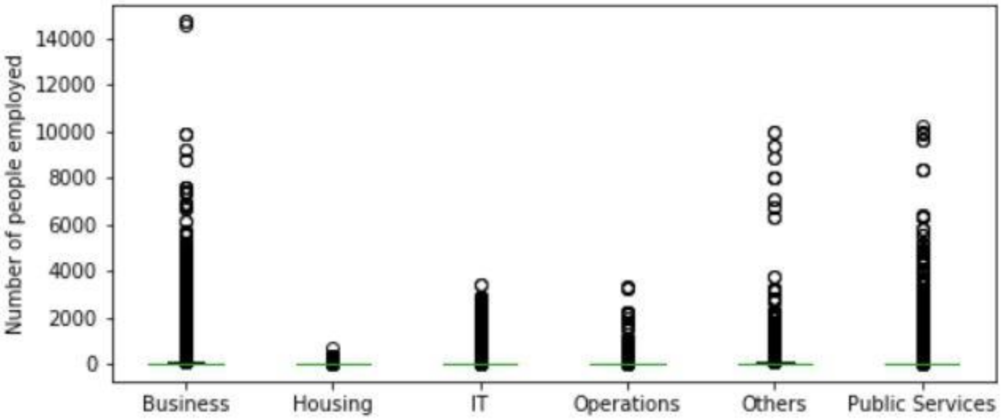

Figure 5: Instances of people employed for each sector in the City of Melbourne

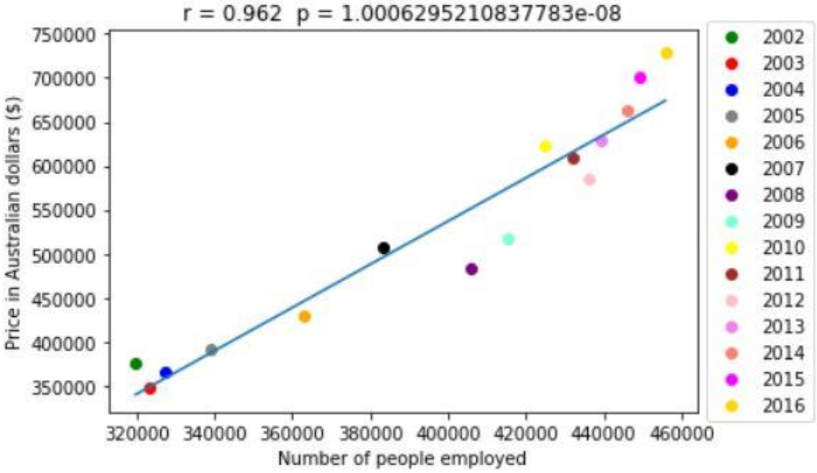

Figure 6: Median house and apartment price averaged vs number of people employed (City of Melbourne)

From what the analysis have shown so far, there is indeed an increase in both the number of employment and the average price of housing in Melbourne from 2002 to 2016. Hence, a correlation analysis must now be conducted to prove that there is indeed a positive correlation between the two factors. Hence, a scatter plot was drawn to examine correlation. Moreover, the Pearson correlation coefficient and its corresponding p-value was provided to indicate the significance of the correlation (**Figure 6**). Note that it is possible that a more accurate result can be obtained by removing the outliers of each employment sector and replacing the missing values from the dataset with a realistic value instead of zero, but the presence of outliers and missing values were finally decided to be kept to conclude a more pragmatic analysis.

Through Python, it is found that there is a percentage increase of housing price from 2002 to 2016 by 194.14% while the number of people employed increased by 142.54%. Backed by the Pearson correlation coefficient of 0.936 and the shape of the regression line, this concludes that the number of people employed is a factor towards changing house prices in Melbourne (**Figure 6**).

Value

The raw data provided by both datasets are untidy and contain several outliers and missing values. Therefore, it would be hard to justify any correlation between the two datasets without processing them. Hence, processing the raw data adds value towards inferences.

By processing and visualising the data, people can easily see the correlation between the two factors by giving them human-readable visual access towards the big, messy data. Moreover, processing and visualising the data gives the Victorian government and the general public more context so they can make better inferences about the correlation between the number of employment and house prices without the necessary costs required for a census or survey. This is invaluable for parties that are involved in maintaining employment and real estate issues.

Challenges and Reflections

It is important to note that even though the study found that there is a positive correlation between the number of people employed and the increase in house prices with a high Pearson correlation coefficient (**Figure 6**), this study did not take other factors into account such as: Percentage of people in poverty, Melbourne's population, Number of foreigners purchasing property, etc. These factors could have various effects on the correlation in the change of house prices in Melbourne over the past decade. Several issues were encountered during this study:

1. There exists a notable number of missing values within the dataset that makes it difficult to detect outliers within the number of people employed in each block and year.
2. This study has attempted to visualise a heatmap of the number of people employed in each sector. However, since there are a lot of values involved, the processing time takes too long for Python to process. Ultimately, it is decided to not use the heatmap for this study.
3. More datasets would be required to thoroughly assess the factors that affect property prices. The initial question was "**What are the variables that affect changing property prices?"** but found that employment rate is one of the factors that affects prices the most.

Question Resolution

As shown from the results, this study has sufficiently answered the posed question and successfully abstracted away geographical bias. Hence, this report can be applied in all areas besides Melbourne. **The increase in employment rates indeed contribute to rising house prices.** However, further research would identify all other variables mentioned previously and would yield a more accurate result, possibly including geographical areas outside the City of Melbourne.

Since there is a correlation between employment rate and house pricing, the Victorian Government could impose appropriate income and real estate taxes to assess overall affordability and people's purchasing power. Furthermore, people can use this report for further reference in real estate purchase, lease, investment, etc., possibly boosting the region's economy and welfare.

Code

Every code during this study is done using Python. The Python libraries used were **pandas, matplotlib, numpy, scipy, re, decimal** and **seaborn.**

Methods done include: Reading data from csv files and storing them in **panda**'s DataFrame, removing dollar signs and thousand separators for arithmetic operation using **re** and **decimal**, generating VAT visualisation[1] and clustering using **numpy, scipy** and **seaborn,** graphing plots using **matplotlib**, correlation analysis by calculating Pearson coefficient and p-value using **scipy,** feature transformation (normalisation) and integration of employment industries into sectors.

Bibliography

1. Michigan Technological University (VAT algorithm link: http://www.ece.mtu.edu/~thavens/code/VAT.m). Adapted from MATLAB version.