**What datasets were used and how where they linked?**

For this analysis we wanted to look at 4 sources of data.

First, we need a measure of average wages in Melbourne. This came from: “Total hourly rates of pay excluding bonuses: all sectors by state, original (financial year index numbers for year ended June quarter)” taken from the Australian Bureau of Statistics website. This is timeseries data containing an index of average rate of pay (this was for states not cities so there is likely some error in that extrapolation) that had yearly numbers for all states from 1998 to 2020. This is an index of the relative value of labour, so the numbers do not correspond to dollar values.

We aimed to compare these wages to the consumer price index. This is an index of relative prices for different types of household expenditure. We got our data from: “CPI: All Groups, Index Numbers and Percentage Changes” taken from the Australian Bureau of Statistics website. This is timeseries data containing value for all capital cities with biannual numbers from 1948 to 2021.

We then wanted to compare wages to housing prices. Once again this was sourced from the Australian Bureau of Statistics website, specifically from: “Residential property price index, index numbers and percentage changes” which contains timeseries index values of the average residential property price. This contains values for all capital cities with biannual numbers from 2003 to 2020.

These were all we needed for the main analysis, but we also wanted to look at power consumption in Melbourne. Since studies have shown a strong relationship between earnings and power usage, specifically, this is independent of unreported income (black markets etc.) so could theoretically be substituted for wages in our analysis. This was taken from: “Annual electricity consumption – NEM” from the Australian Energy Regulator website. It contains timeseries data of each states power consumption in terawatt hours annually from 1999-2000 to 2020-2021.

These data sources were linked based on year (power based on second year in time period). So we could compare wages to corresponding house prices etc.

**Wrangling/Analysis Techniques:**

The same process was done for each capital city and also for the average of the cities.

1. CSV files were created containing the year and relevant index, for house prices and CPI. (Power consumption was also done for Melbourne). This was done manually because the format that the data was provided in did not allow for reading from csv or even from excel using python.
2. These CSV files were read into python as a pandas DataFrame.
3. The Date column for these DataFrames were altered to only contain the last two digits of the year. This was done to remove inconsistency in the date format.
4. The data was the grouped by year, with the average of a year taken when multiple years were available.
5. These separate DataFrames were then combined into a single DataFrame for the city.
6. Sklearn’s LinearRegression model was then trained on wages to House prices.
7. Two datapoints were then predicted using this model, and the gradient calculated.
8. The R^2 value of the was regression calculated using the score method.
9. The residuals for each of the wage values was calculated by subtracting the predicted values from the actual values.
10. Steps 6 to 9 were then repeated for wages to CPI.
11. Scatterplots were then made comparing wages to both CPI and house prices.
12. The residuals were plotted.

We then also plotted wages to power usage to test the theory that it was an effective measure. But the very low correlation ruled that out.



Figure Hourly Pay to CPI and House Prices

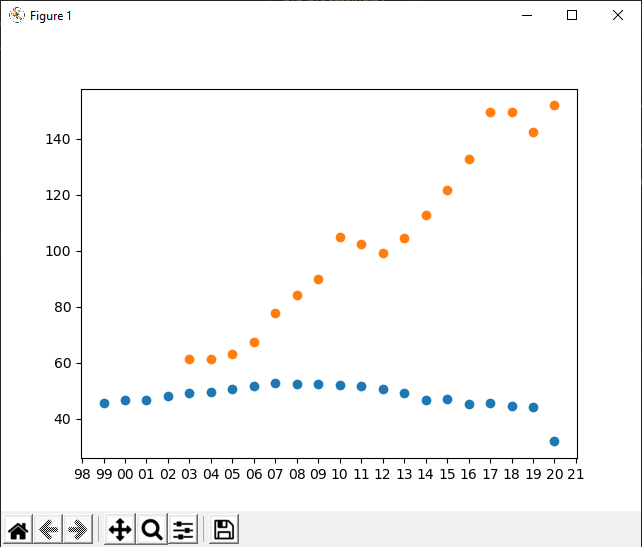


Figure Power (x) to Hourly Pay (y)