EODP Assignment 2 Report

Research question and how it affects sustainability of the community of Victoria

Needless to say, the crime rate of a certain location strongly affects the safety and quality of life of its citizens. It also has a huge influence on the well-being and development of the location, in terms of economy, happiness etc. Therefore, a deep analysis of how crime rates are affected is of huge importance. We believe that one of the main factors that motivate people to commit acts of crime like burglary, theft, tax fraud, etc is their level of income. With money being of vital importance to everyone, we infer that the lower the level of income, the higher is one's tendency to commit crimes.

In this project, our aim is to deduce a correlation between the level of income and the likeliness to commit criminal activities. By extension, we would be able to obtain a more comprehensive view of the liveability in Victoria. We strongly believe that a detailed analysis of this combination of datasets would reveal some very beneficial results and conclusions that would be useful to help lower crime rates.

Datasets and how they are linked together

A total of 8 datasets were obtained, with 7 of them being downloaded from Australia's public data website (https://data.gov.au/dataset/ds-sa-860126f7-eeb5-4fbc-be44-069aa0467d11/details?q=crime), namely:

- "2011-12-data_sa_crime.csv"
- "2012-13-data_sa_crime.csv"
- "2013-14-data_sa_crime.csv"
- "2014-15-data_sa_crime.csv"
- "2015-16-data_sa_crime.csv"
- "2016-17-data_sa_crime.csv"
- "2017-18-data_sa_crime.csv"
-

concatenated into one dataframe with the name of "combine.csv" which forms an extensive Pandas dataframe containing the attributes "Reported Date", "Offence counts", "Offence Level 2 description" etc from years ranging from 2012 to 2017.

On the other hand, a dataset named "Standard_Income_Statement_2014-15.csv" was obtained from Australia's public data website (https://data.gov.au/dataset/ds-melbourne-https%3A%2F%2Fdata.melbourne.vic.gov.au%2Fapi%2Fviews%2F2ph9-es73/details?q=income). It is a comprehensive collection of the data regarding the yearly income of citizens of Victoria, and the data we extracted was specifically from the years ranging from 2012 to 2017 as well. We analyzed the attributes "year", "actual/budget/planned income", "income statement description" and "value".

Wrangling, analysis methods and key results and why they are significant

To begin with the wrangling and analysis that we have applied, we would be using a more general view in terms of offence types. To visualise the data in order for us to have a more comprehensive view, we plotted a few diagrams.

In Diagram 1, a bar chart is plotted, with “Offence count” on the y-axis and “year” on the x-axis. From the file “combine.csv”, we first grouped the dataset by “year”, then we aggregated the “Offence count” values according to their respective years, resulting in a dataset containing 6 rows and the same number of columns as before. Then, we plotted the bar chart to obtain a detailed visualisation of the distribution of crime cases from 2012 to 2017, which was observed to be relatively even without any consistent trends of increase or decrease.

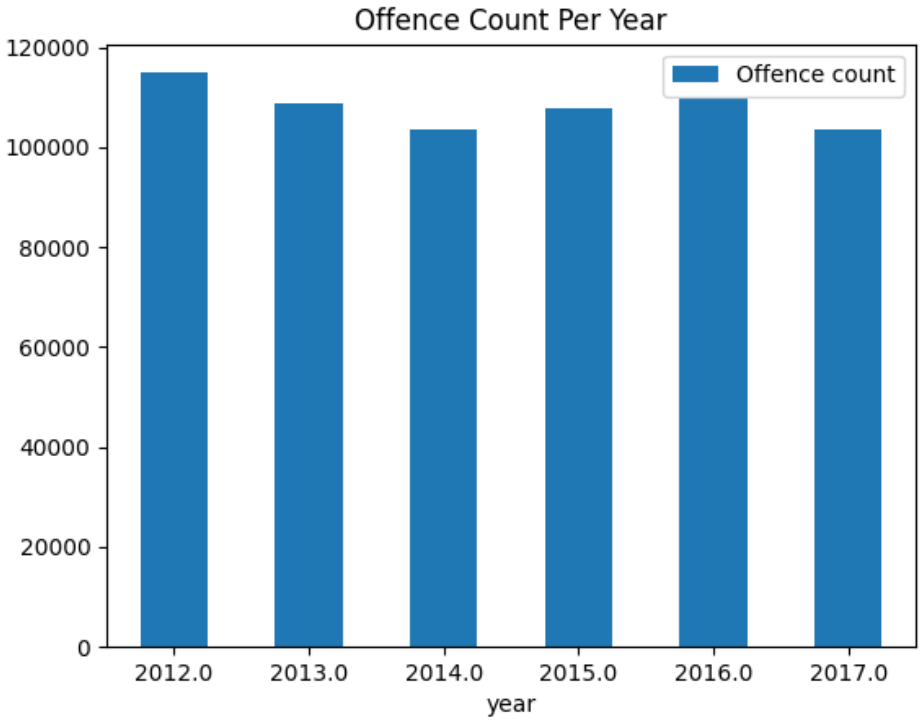

Diagram 1

At the same time, Diagram 2 was plotted as well, it being a bar chart with “Income” as the y-axis and “Year” as the x-axis. Using similar preprocessing steps, we grouped the dataset by “year”, and we aggregated the total income earned in the span of their respective years, ranging from year 2012 - 2017. In this case, a steady increase in income per year can be observed, with 2017 having the highest amount (>800000) and 2012 having the least (about 700000).

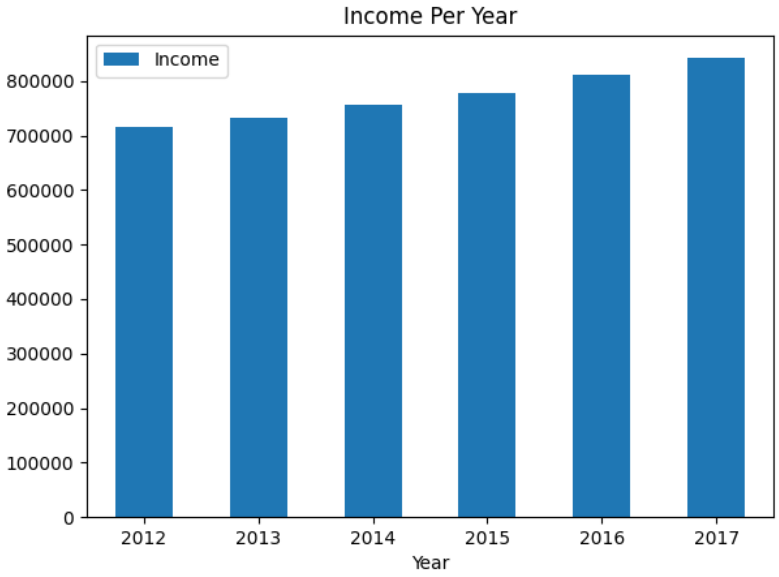

Diagram 2

Following that, a scatter plot was plotted, as can be viewed on Diagram 3. With “Income” as the x-axis and “Offence count” as the y-axis, it was to our surprise that there was not a clear or strong relationship observed as the number of offence counts fluctuated year by year and no trend was apparent. As it was deduced that the relationship between offence count and income is linear, we utilized Pearson’s correlation to calculate the **Pearson coefficient which resulted in a value of -0.569**, indicating a weak negative relationship between the two variables. Hence, after reviewing the combine.csv file, we decided to come up with a more specific preprocessing method to combat the lack of clarity.

year	Offence Count	Income(k)
2012	114917	714914
2013	108699	731206
2014	103574	755176.6
2015	107639	777805.9
2016	109863	812509.8
2017	103477	841948.7

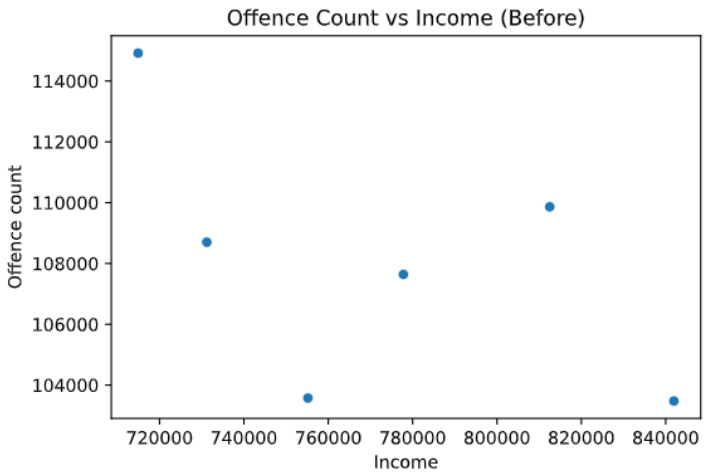

Diagram 3

At the second stage of our data wrangling process, we first plotted Diagram 4, a bar chart showing the distributions of the number of offences across the variety of crime types. We grouped the dataset by “Offense Level 2 Description” and plotted a bar chart of crime count vs crime type. It was evident that some crime types such as 'ACTS INTENDED TO CAUSE INJURY', 'PROPERTY DAMAGE AND ENVIRONMENTAL' and "SERIOUS CRIMINAL TRESPASS' constituted significant proportions to the total number of offence counts, but the “THEFT AND RELATED OFFENCES” contributed to too large of a proportion, reducing the ability of the data to cover all crime types. Hence, we removed the “THEFT AND RELATED OFFENCES” Offence counts.

Therefore, “combine.csv” was preprocessed again (“ex_combine.csv”), but this type of non-monetary related crimes were dropped as again, they are not related to the income. We used this data to plot another scatter plot (Diagram 5), and drew a line of best fit. It was very clear that there was a linear, negative relationship, and the **Pearson coefficient calculated is -0.830**, translating into a strong negative correlation.

year	Offence Count	Income(k)
2012	71974	714914
2013	67709	731206
2014	64152	755176.6
2015	66091	777805.9
2016	65501	812509.8
2017	61657	841948.7

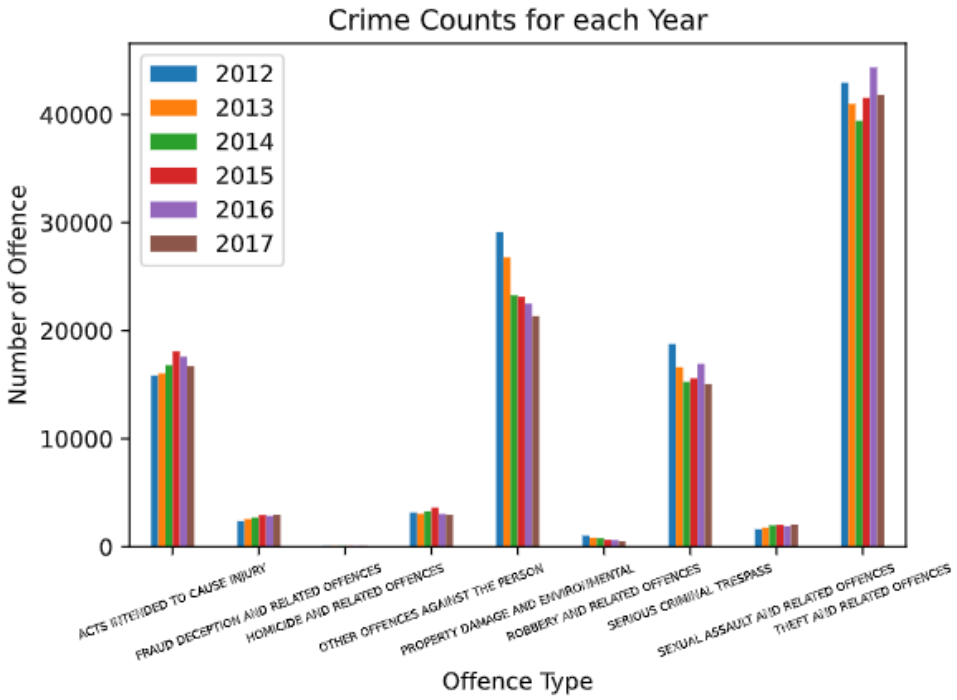

Diagram 4 (Before remove "Theft and related Offences")

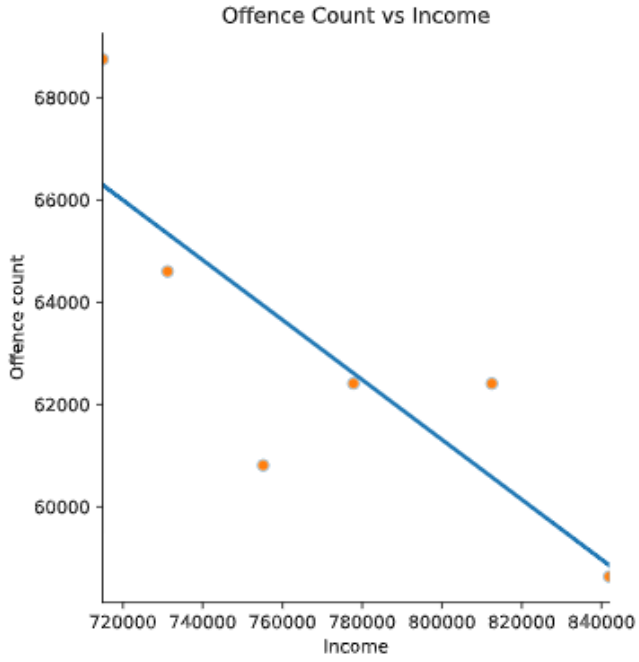

Diagram 5 (Scatterplot after remove "Theft and related Offences")

Why this method is chosen instead of other alternatives

One of the reasons that Pearson's correlation was selected to calculate the relationship between offence count and income is that the Pearson correlation is useful to assess the strength of the linear relationship between crime rate and standard income, and it can be determined by seeing how close the scatter plot is to a straight line. Both our variables are discrete and quantitative, so Pearson's correlation is suitable.

In addition, using Pearson's correlation, hints of potential causal relationships may be obtained, although it is not necessarily implied. The possible value of the Pearson coefficient is within the range [-1<x<1], and the closer its absolute value is to 1, the stronger the relationship, and the negative sign indicates a negative relationship. The Pearson coefficient we obtained in this case is -0.819, and since it is negative and its absolute value is really close to 1. It can be deduced that as the income increases, the crime rate decreases.

There is another alternative we could have utilised to determine the correlation between the two data, but as there is no non-linear correlation between crime rate and income, Mutual Information would be inappropriate.

Limitations

There are a number of limitations of our results.

1. Firstly, not all types of crimes are related to monetary purposes. From the first stage of our data analysis process, it was concluded that a number of crime types were not related to the correlation of data at all, and this adds a lot of impurities to the analysis. Criminal activities due to personal disputes, conflicts of interest in business or work, or even for personal entertainment or fulfilment, etc do not contribute to income whatsoever.
2. Secondly, the total value of the income cannot be determined solely by the total amount of income, as the fluctuations of the economy may cause the value per dollar to change. For example, there may have been a high inflation rate in 2016, causing people to have less purchasing power per dollar.
3. Another limitation could also be the possible lack of data which might not give accurate calculations of the data. Also, not only does income affect the crime rate, it is possible for it to happen vice versa as well.
4. Besides, one's level of income does not provide an accurate estimate of one's level of wealth. For Example, one may have different avenues of passive income that increase their total wealth.
5. Another possible scenario is that even if one's income is of a decent to a high amount, there may be factors that contribute to a decrease in wealth. For example, losses in investments,pre-existing debt, or even expensive medical costs for a loved one.
6. Lastly, as we aggregated the total income, we failed to consider income inequality as well, and that may play a huge part in affecting the crime rates.

Improvements

1. Firstly, research on economic changes in Victoria may help to determine the percentage of alternation of income on other economical factors. Deduct those alternation percentages to obtain a more accurate result based on crime only.
2. Other than using the datasets on income only, the datasets containing the wealth of citizens of Victoria should be taken into account, as it provides a true estimation of the monetary possession per citizen.
3. Lastly, research on any income inequality in Victoria reported to determine the rate of crime of that particular area.