COMP20008 Assignment 2 Written Report

Is there a Correlation between Air Pollution Emissions (of SO2) and the Personal Income of Individuals in Victoria?

Group 103

Park Chang Whan

******* *******

******* *******

******* *******

What is the research question and how is it related to the theme of understanding the livability, inclusiveness, health and sustainability of communities in Victoria?
Sulfur dioxide (SO_2) is a toxic chemical compound released by volcanic activity, copper extraction and burning of fossil fuels. According to studies done by the Smithsonian Institution, these three destructive aspects have been on the rise as a result of climate change, technological advancements and other suitable factors, in turn polluting the environment with increasing levels of SO2. Therefore, pollution directly impacts our environment and standards of living by releasing toxic chemicals such as sulfur dioxide, and has the potential to serve repercussions on our health. Additionally, Studies by the Australian Bureau of Statistics have outlined that personal incomes in Australia have risen by 3-5% (among various categories). By having access to (and demanding) more services, goods and technology as a result of higher incomes, it is hypothesised that there will be a reduction in pollution in Victoria. To solidify (or counteract) the hypothesis, this data science study aims to highlight the role personal incomes have in influencing pollution, in particular: Is there a Correlation between Air Pollution Emissions (of SO2) and the Personal Income of Individuals in Victoria? This report could be used as a benchmark in order to evaluate the sustainability and liveability of communities affected by pollution, on condition that pollution increases in the future.

What are the datasets you've used and how have you linked them together?
For pollution of chemical compounds, the dataset utilised for this data science study is produced by the National Pollutant Inventory. It is composed of 786818 rows with a combination of float (for categories such as amount of emissions), integer (for categories like report date) and string formats (for its various headings). Although it provides data on relatively unimportant materials for this report, such as the business that conducted the study, longitude/latitude, and jurisdiction facility ID, this dataset still proves to be vital due to its valuable information on reporting year(s), polluting chemical substances, amount of emissions per kilogram and locations (state-wise). The data spans the period 1998/1999 to 2019/2020.

The dataset providing information on the personal income of greater Australia was provided by the Australian Bureau of Statistics and consists of multiple sheets. Each new sheet has a row representing a smaller statistical area of Australia, with the last having each row representing local government areas. Each statistical or local government area in every row is represented as an area integer (as well as a string for that area's name). There are five column groups with each group having five columns representing five financial years of data. These are Earners (persons) as well as median age of workers, total sum earned in each area, mean sum earned and median sum earned, each represented as an integer.

Ultimately, the various datasets will be linked by respective years in order to obtain accurate data and identify trends year-by-year. Additionally, when outputting any graphs/figures, having both datasets sorted by respective years will prove to be useful and beneficial. With the relationship of income and pollution being outlined before, we can link the datasets together by comparing the per capita GDP along with the cost for technology in improving the quality of the environment. For example, higher income would mean high affordability of such technology, leading to a reduction in pollution rate of SO2.

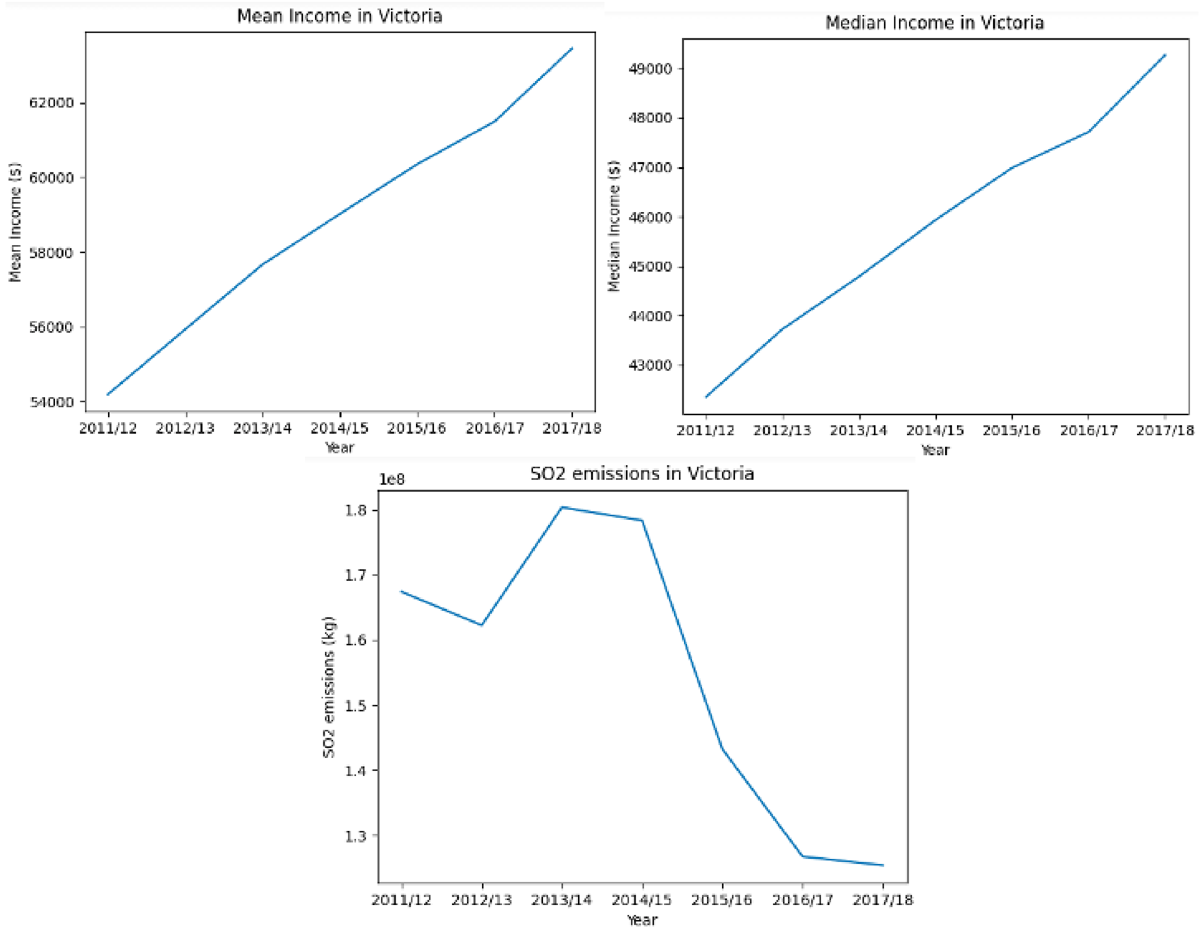
What wrangling and analysis methods have you applied? Why have you chosen these methods over alternatives?
Due to the Person Income Dataset being significantly large with information that wasn't required, we decided to use only one of the sheets (Total income by GCCSA - TIBGCCSA) and saved it as a csv file to use as we only needed values for Victoria (which were identical throughout the sheets).

We mainly used pandas to start our process of data mining. The data from the TIBGCCSA dataset was pulled into a dataframe that contained only the income from Victoria and also only the Mean and Median income columns by using the .loc() method. The data from the emissions .csv file was similarly entered into a dataframe using the .loc() method to only take the data related to sulfur dioxide emissions in Victoria. The .groupby() method was used to group the rows by years, which we then calculated the sum of emissions for that year using aggregate method.

Starting our diagnostic analysis, the data frames are plotted together on two different scatter plots, one with mean income for each year in Victoria (x-axis) and one with median income (x-axis), and both have the sulfur dioxide emissions per year on the y-axis. This enables a visual inspection to see if there is a relation between the income of individuals in Victoria and the amount of sulfur dioxide emissions. The values in the two data frames are also able to be used to calculate the Pearson *R* Correlation coefficient using a function from the python statistics library to indeed see if there is a relationship between the two variables and how strong that relationship is. The data was plotted on line graphs so the general trend in each variable and between income and emissions can be seen as the years pass. A linear regression model (which we used for predictive analysis) for both the median income and mean income is also done to determine, if any, what effect changing the x variable (income of victoria in that year) would have on the y variable (sulfur dioxide emissions of that year). In order to check our assumptions for the linear regression as well (linearity and independence of residuals), we plotted the residual plot to visualize and verify them.

The scatterplot was chosen to represent the data from the two datasets because they are only two numeric values being compared over the same points in time. This way we are able to use it to compare with the results we get when doing the regression analysis. Therefore the scatter plot is the clearest form of visualisation that can be used to represent the two data values over the same periods of time. Having a scatterplot and only two different variables changing also makes calculating the Pearson *R* Correlation coefficient simple and also very informative as if there is a strong or weak relationship between the two sets of data in the dataframe and if the relationship is positive or negative. Line plots were also used with the two variables changing by year on the same set of axes to show the basic trends of both income and emissions per year and then superimposed so the trends could be compared to see if they are both strong or positive or inversely related. Due to there only being a comparison between two numeric variables other visualisation options such as a bar chart or histogram would be unsuitable due to there being no frequencies of something occurring being compared. Similarly other options such as a heat map are unsuitable again, because only two numeric values are being compared. Therefore leaving scatter plots as the best option for both visualisation and for understanding if there is a correlation between the two numeric values.

What are the key results your research has obtained?
After obtaining relevant datasets and applying the data wrangling techniques identified, many graphs were outputted for analyses and discussion of results.



Figures 1,2,3: Yearly Mean Income, Median Income, SO2 Emissions in Victoria, respectively

As seen above on *Figures 1* and *2* above, mean and median income in Victoria had a strong upward trend from 2011 to 2018. During the same time period, it can be seen that SO2 emissions (*Figure 3*) have significantly decreased. The spike from 2012-2014 can be attributed to Victoria further industrialising their economy and services, perhaps ensuring locations receive more than sufficient energy by burning extra fossil fuels. Overall, the ultimate trend observable is: as personal income increases, SO2 emissions decrease.

Many studies have been conducted on the relationship between emissions of toxic pollutants (mainly CO2) and rising levels of income. These studies suggest that emissions are largely driven by income and energy consumption, specifically, pollutant levels increase with income, which counteracts the results of our study. However, Australia is the 10th richest country in the world, with, approximately, $52,000 (USD) GDP per capita. Assuming this wealth is distributed evenly across states, it can be said that Victoria is affluent, hence being technologically advanced and having the ability to execute structural changes to battle pollution levels of SO2. Therefore, as income in Victoria increases, per capita GDP increases and technology to contest SO2 pollution becomes more affordable and widely used, ultimately reducing the toxic pollution levels.

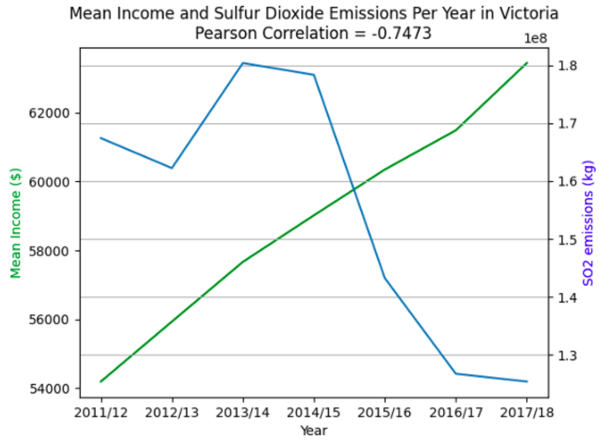


Figure 4: Mean Income vs SO2 Emissions, with a Pearson Correlation of -0.7473

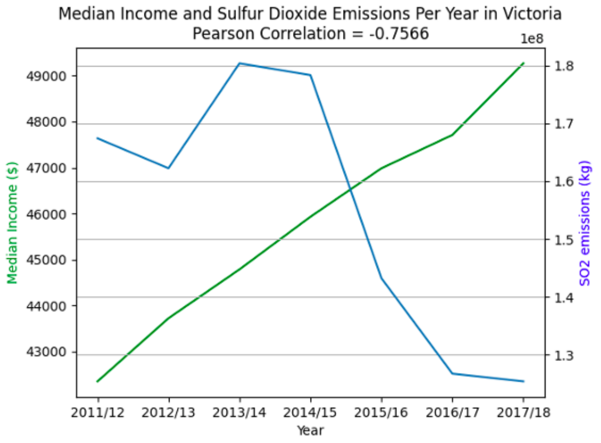


Figure 5: Median Income vs SO2 Emissions, with a Pearson Correlation of -0.7566

In order to further evaluate the relationship between these two variables, a multivariate analysis was done. Pearson's *R* Correlation, commonly used in linear regression, was used to determine the strength of the relationship between income and emissions. Both the median and mean incomes have similar *R* coefficients of -0.7473 and -0.7566, respectively, indicating a strong, clear and inverse correlation with SO2 emissions.

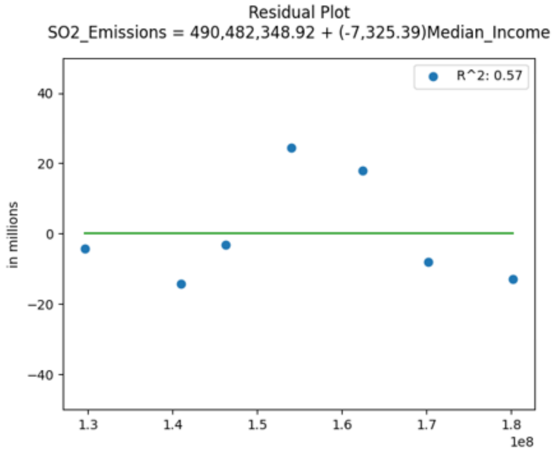


Figure 6: Yearly Median income vs Yearly Sulfur Dioxide Emissions Residual plot

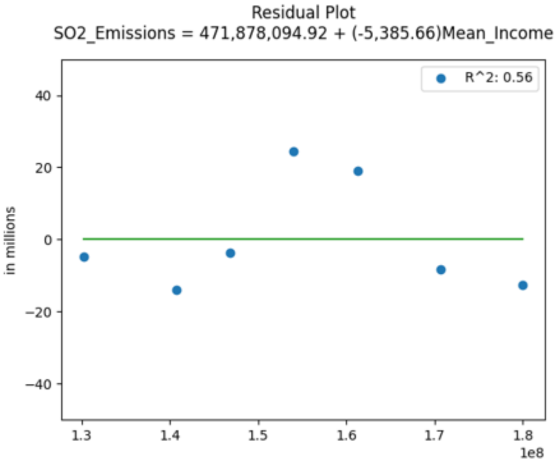


Figure 7: Yearly Mean income vs Yearly Sulfur Dioxide Emissions Residual plot

Although the Peason correlation shows a strong negative relationship, we know that correlation does not mean causation. So in order for us to go further and explain the impact of changes in income on the emissions, we decided to use linear regression. We have found the equation(*Figure 6,7*) with their respective coefficient of determination (R^2). With the coefficient of determination, we found that 57% (56%) of variation in emissions is explained by variation in the median (mean) income. We believe that this is a significant percentage as the numbers we are dealing with for emissions are in the tens of millions. In addition, we made sure to check our assumption of linear regression for linearity and independence of residuals and were able to verify our assumptions (*Figure 6,7*). So, by looking at the equation and the scatterplots (*Figure 8,9*), we can visualise and verify the relationship between income and emissions.

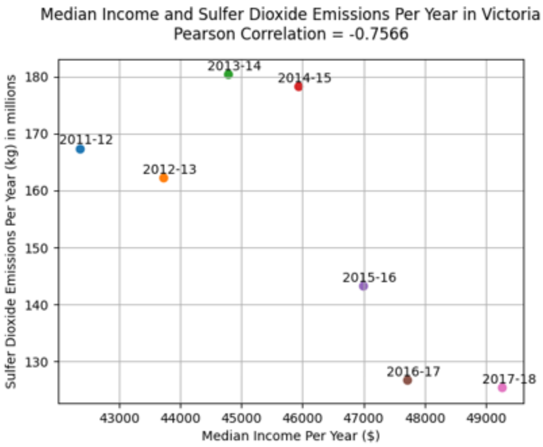


Figure 8: Yearly Median income vs Yearly Sulfur Dioxide Emissions scatterplot

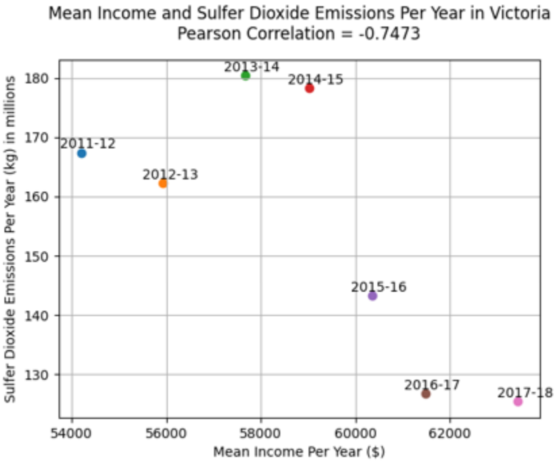


Figure 9: Yearly Mean income vs Yearly Sulfur Dioxide Emissions scatterplot

Why are your results significant and valuable?

A strong rising income trend has been illustrated in this study, showing that Victoria is consistently obtaining wealth, and this appears to continue in the future. This is significant as, through other various analyses, the trend could be extrapolated to determine approximately when in the future SO2 pollution levels approach to zero, hence allowing for healthier, more sustainable communities in Victoria. Additionally, the data will prove valuable to policymakers and lawmakers to set a standard for industries, households and factories to further manage the issue of pollution. This in turn has other benefits, such as impeding climate change and stopping depletion of the ozone layer. Finally, we believe that the information can be used to encourage or focus more funds into technology to reduce the rate of pollution, increasing liveability in Victoria.

What are the limitations of your results and how can the project be improved for the future?

From the information observed from the graphs, we can see the trends of the data. However, it is hard to foresee the predicted mean income for future or even the SO2 emissions due to the different factors involved. For example, like with the recent COVID outbreak, this information may not be able to be used to predict future results. As for the wrangling process, there were some complications faced in data mining and manipulating the data to be entered into a data frame correctly due to the empty cells present in the original table. Therefore, the process of wrangling was a bit complicated but is still a minor issue. The lack of points also contributed to the issue of the data presentation. Further limitation to our results is that despite there being a correlation between increased income and a decrease in sulfur dioxide emissions, this correlation does not necessarily mean causation and there could be other factors other than income that result in a downward trend in sulfur dioxide emissions. Hence, the analyses discussed, although relatively correct, may have some degree of uncertainty and lose slight credibility.

In our opinion, this project can be further improved with better formatting of data to allow for a more simple wrangling process without as much manipulation required. Perhaps even evaluating on a monthly basis can provide us with more accurate and precise results. As for the uncertainty in the future, increased knowledge of the economy might be helpful to some extent but this would mean more data to be considered. Therefore, the research will be more complicated and lengthy in process but can return more accurate and precise information. Additionally, it would be useful to identify the income-pollution relationship with other chemicals, in order to broaden scope and determine which toxic gases pose the highest threat to the liveability, sustainability and health of Victoria.