Sedentary Behaviour and Chronic Diseases: Identifying Correlations and Potential Risks

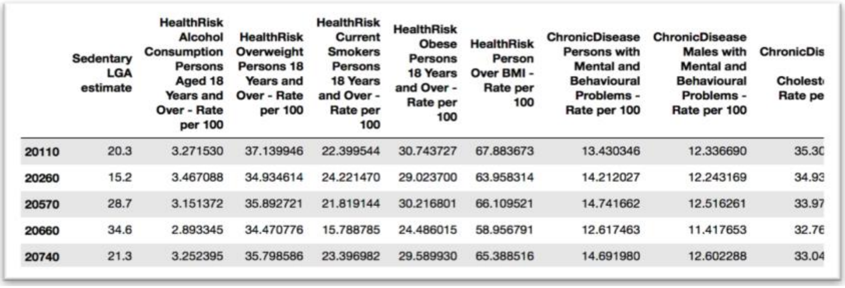
1. **Domain & Question**:
   The selected **domains** of this report are Human behaviour and Health.
   The **question** to be investigated is:
   - Are there significant correlations between sedentary behaviour (sitting 7 hours or more per day) and incidence of chronic diseases in Victoria?
   - If any, what is the nature of the correlations? And what are the causes to this result?

2. **Data Sources[1]**:
   - LGA11[2] Health Risk Factors – Modelled Estimate (named 'HealthRisk.csv')
     Modelled estimates of health risk factors: *overweight and obesity* by LGA 2011, for 2011 to 2013. All the attributes are given by rate. (80 rows × 32 columns)

   - LGA11 Chronic Disease – Modelled Estimate (named 'ChronicDisease.csv')
     This dataset contains information on modelled estimates of chronic diseases including: circulatory system diseases, type 2 diabetes, high cholesterol, and hypertension disease for 2011-13 by LGA 2011. (80 rows × 32 columns)

   - LGA11 Sedentary Behaviour (Sitting Hours per Day) (named 'Sedentary.csv')
     The proportion of people who sit for 7 hours or more per day. Modelled estimates by LGA 2011. (80 rows × 32 columns)

These datasets were chosen as they come from reliable sources and have the common LGA area_code attribute, which simplifies successful integration.



	Sedentary LGA estimate	HealthRisk Alcohol Consumption Persons Aged 18 Years and Over - Rate per 100	HealthRisk Overweight Persons 18 Years and Over - Rate per 100	HealthRisk Current Smokers Persons 18 Years and Over - Rate per 100	HealthRisk Obese Persons 18 Years and Over - Rate per 100	HealthRisk Person Over BMI - Rate per 100	ChronicDisease Persons with Mental and Behavioural Problems - Rate per 100	ChronicDisease Males with Mental and Behavioural Problems - Rate per 100	ChronicDis Cholest Rate pe
20110	20.3	3.271530	37.139946	22.399544	30.743727	67.883673	13.430346	12.336690	35.30
20260	15.2	3.467088	34.934614	24.221470	29.023700	63.958314	14.212027	12.243169	34.93
20570	28.7	3.151372	35.892721	21.819144	30.216801	66.109521	14.741662	12.516261	33.97
20660	34.6	2.893345	34.470776	15.788785	24.486015	58.956791	12.617463	11.417653	32.76
20740	21.3	3.252395	35.798586	23.396982	29.589930	65.388516	14.691980	12.602288	33.04

Figure 1: Sample health risk data

   - LGA11 Age Distribution – Person (named 'AgeDistribution.csv')
     This dataset contains information about the number of persons and their proportion of the total population by age groups to 85 years for the year 2013, and over by LGA 2011. (80 rows × 39 columns)

   - LGA11 Labour Force (named 'LabourForce.csv')
     This dataset initially by Torrens University Australia - Public Health Information Development Unit. Unemployment rates and labour force participation rates (2012) and female labour force participation (Census 2011) by LGA 2011. (80 rows × 11 columns)

3. **Pre-processing and Python**:
   In pre-processing, the datasets were manipulated by using *Excel*, *pandas* library for Python:
   - The *.csv* files read in *Jupyter Notebook* are already pre-processed in *Excel* including change any formatting used on them to get better viewing (headers, number formatting).
   - Since it also parses metadata stored in a JSON file while initializing, the name of each column is too long to index, renaming and replacing column name are required. Editing attributes' name to make sure consistency of content. Using underscore instead of space in column name to avoid Key Error

[1] All datasets can be downloaded from **AURIN** (Australian Urban Research Infrastructure Network): **https://aurin.org.au**

[2] **LGA**: Local Government Area, in this report it refers to the municipality (or suburb) of Victoria from 2011.

in later process.

- Sorting by area code of Victoria to ensure multiple data sets' area code are matching each other.
- For clearer visualisation and more convenient analysis of the data, removing unrelated column is required. To avoid error, 'rate per 100' was the preferred unit instead of 'count' in the attributes, where there are no missing values. Although there were some outliers in the data, they are included in the result because the data source is credible and can give interesting results.
- To reach accurate and clear results, all the data was normalized into range [0,1]. This was done by using the normalized formula provided from lecture.

Limitation:

Using Excel to rename and delete row data: time consuming, hard to correlate each attributes' name with row data. It may be solved by using Regular Expression to search specific column name, then return the matched column. It may also can delete columns with the provided words. Since it is possible to delete columns shared the same words by mistake, in pre-processing, manipulating data by *Excel* to keep accuracy.

4. **Integration and Python:**

To connect the multiple datasets, using inner join on the common attribute 'area_code', this method can also remove the areas which were not present in all three datasets. Finally, using Concat in pandas to make the data condense into a contiguous DataFrame for later analysis.

To analyse the data and detect patterns, visualisations were applied. All Graphs were plotted using a combination of Python functions in plotting library *MatPlotLib* and *Seaborn* (*seaborn.regplot* is used to plot the regression line with the scatter plot. *seaborn.heatmap* and *seaborn.clustermap* are used to clearly display the correlation matrix.). For geographic distribution images, *AURIN* provides a tool to generated. Using *Jupyter Notebook* to implement visualization and commenting. Markdown language is used to make more clear comment. All code is written in Python 3.

5. **Results**

Part A: *First investigation at Chronic Diseases:*

In this report, six chronic diseases are selected. The datasets of these diseases against the sedentary LGA estimates were plotted on scatter plots where each point represents the statistics of a Victorian suburb obtained from AURIN, and a regression line is drawn to, if possible, identify a trend. The results can be discussed in two parts.

a. Expect results:

In both *Figure 2* and *Figure 3,* a positive correlation between the sedentary LGA estimate and the selected chronic disease can be observed through the linear regression of the data plots, with positive PCC[3]s of 0.32 and 0.30 respectively. This positive relationship coincides with speculation, based on previous influence of the mass media. However, the absolute values of these two PCCs are relatively small compared to other datasets to be discussed in the next part. These small absolute values might fail to suggest a strong correlation between the two variables in each case and the presence of obvious outliers might be an important contributor to this outcome.

[3] **PCC**: Pearson's Correlation Coefficient

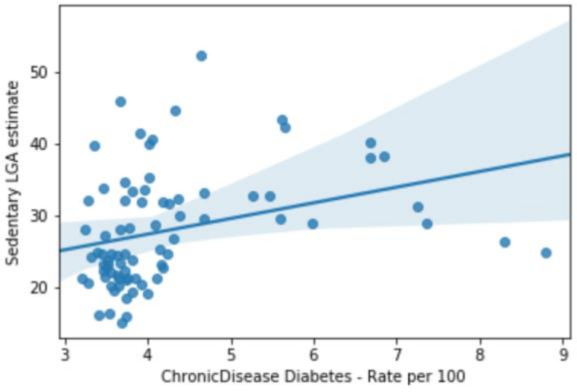


Figure 2: Sedentary Behaviour vs. Diabetes (%)

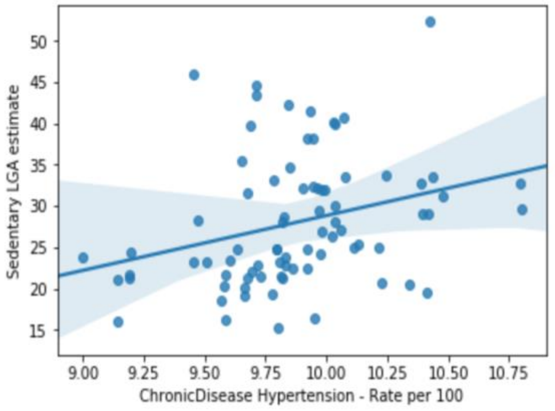


Figure 3: Sedentary Behaviour vs. Hypertension (%)

b. Counterintuitive results:

Figure 4-7 show that the diseases selected are negatively correlated to sedentary behaviour with a positive PCC in each case, which is counterintuitive, especially evident in *Figure 6* and *7*, where overweight and obesity each has a clear inverse proportional relationship with sedentary behaviour. **Though against speculation**, the results are very convincing especially in *Figure 4*, *5* and *7*, with relatively high PCC absolute values of 0.68, 0.73, 0.72 respectively.

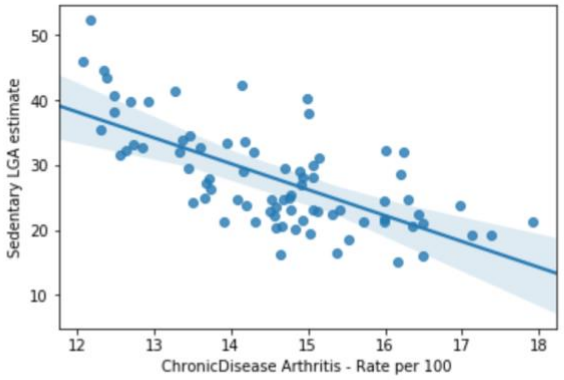


Figure 4: Sedentary Behaviour vs. Arthritis (%)

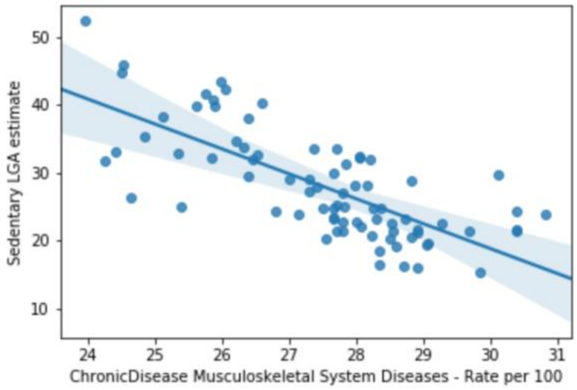


Figure 5: Sedentary vs. Musculoskeletal (%)

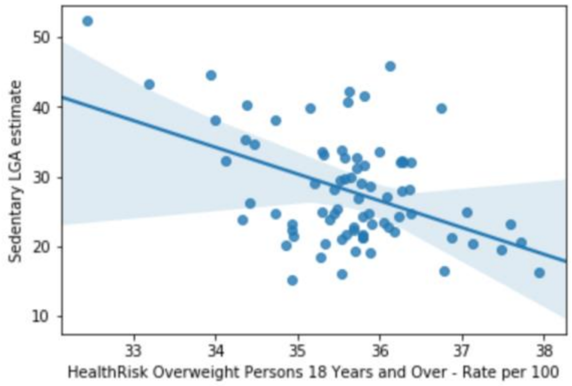


Figure 6: Sedentary behaviour vs. Overweight (%)

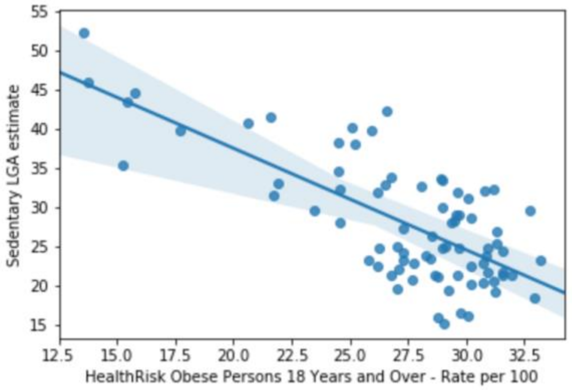


Figure 7: Sedentary behaviour vs. Obesity (%)

These counterintuitive results might be caused by different complicated factors, like population, age and labour force distribution, which will be discussed in the following parts.

Part B: *Heat Map of Chronic Diseases:*

As expounded in the above investigation, an overview is given by how the chosen six possible relatives diseases are correlated with sedentary behaviour. However, further analysis requires more detailed investigation. Therefore, further investigation is given by displaying correlations with a heat map and a clustered heat map.

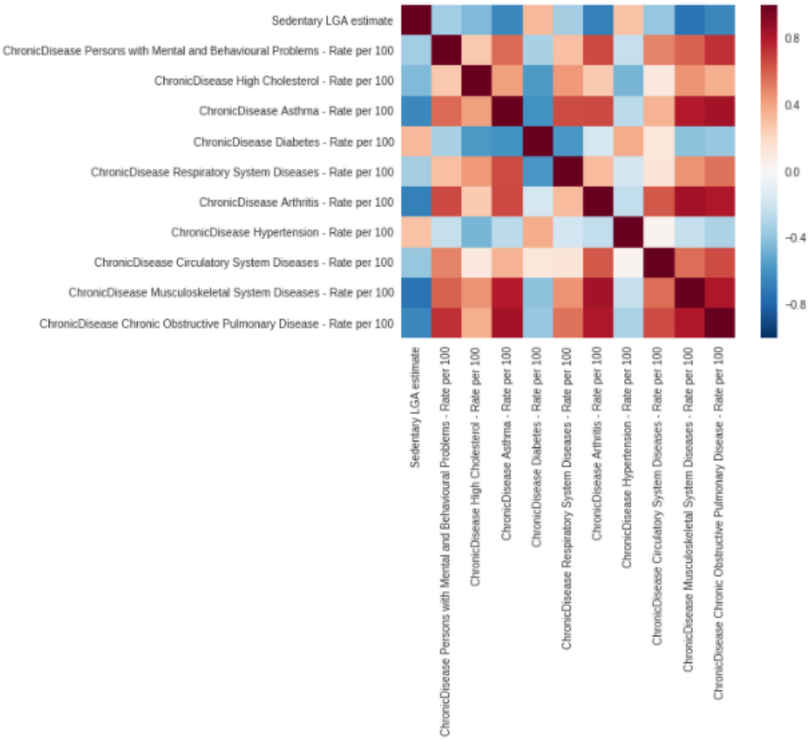


Figure 8: Heat Map of Chronic Diseases

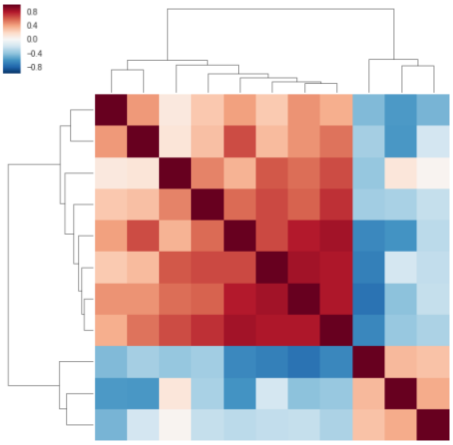


Figure 9: Cluster Heat Map of Chronic Diseases (remove label)

As we can see form the heat map (left), sedentary status is uncorrelated or negatively correlated (PCC of r < 0.4) to most of the chronic diseases, which shares the same result with the previous finding. In the same map, we can observe that diseases have either positive or negative correlations with each other. For example, Asthma and Chronic Obstructive Pulmonary disease have a PCC of more than 0.8, and Asthma and Musculoskeletal System disease have a PCC of more than 0.9. It shows that some diseases may have concurrent symptoms.

The right plot is the clustered heat map, since the labels are the same with that of heat map, they are not displayed. Sedentary behaviour lies in the last third column/row. As we can see, almost all diseases have correlations with each other, which matches the above assumption. The unpredicted results in the previous investigation is caused from this.

Part C: *More Aggregate data:*

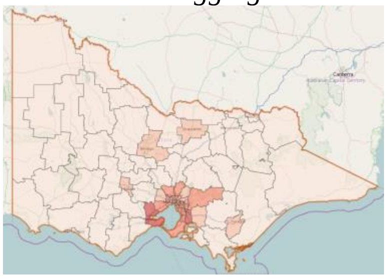


Figure 10:Population Distribution Map

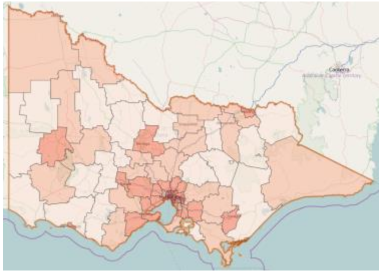


Figure 11:Sedentary Behaviour Distribution

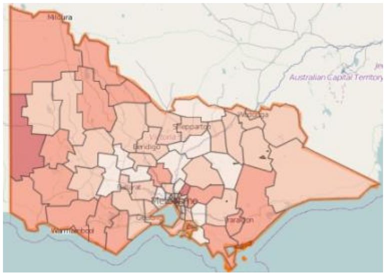
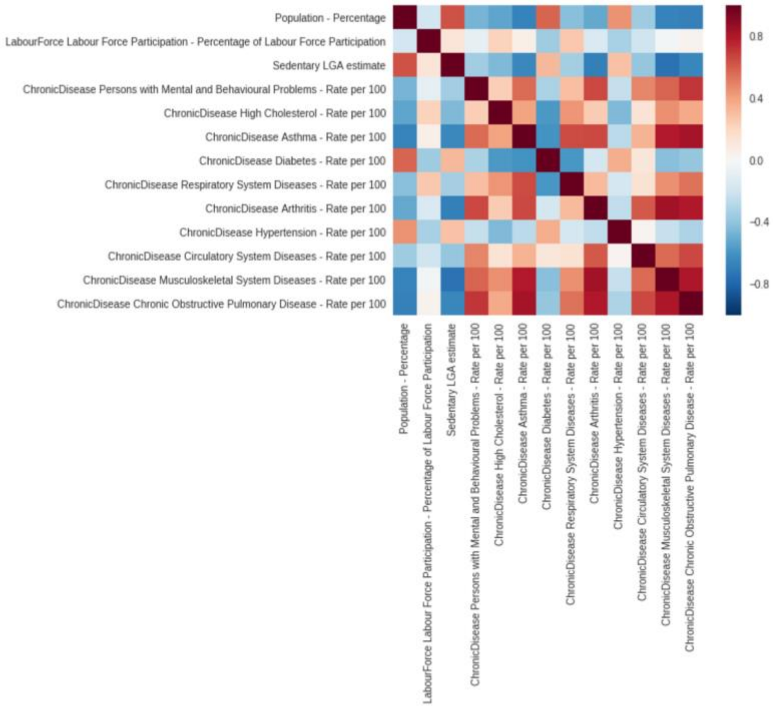
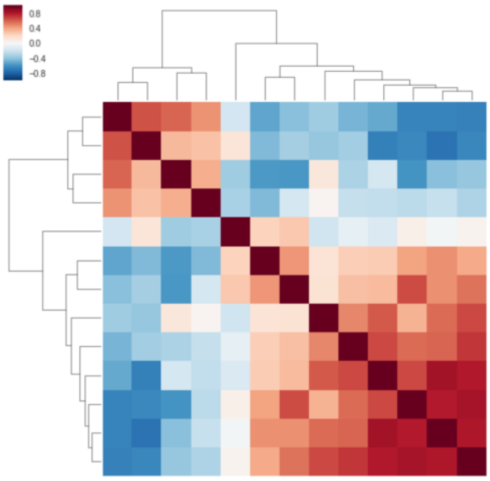


Figure 12: Labour Force Distribution Map

In this section, three distribution maps of population, sedentary status and labour force are generated using AURIN, then further aggregation is given according to the datasets. As the three maps illustrate, the distributions of population (*Figure 10*) and sedentary status (*Figure 11*) share a similar pattern, but those of labour force (*Figure 12*) and sedentary status (*Figure 11*) essentially differ.

The Chronic Disease dataset is set according to the data selected from real total population and percentage of labour force participation. Python calculates the percentage out of the total population, since normalizing the data is a necessary. A heat map and a clustered heat map are generated accordingly.





As shown in the heat map, sedentary status and population has a PCC of more than 0.6, which indicates that they have a relatively high positive correlation. Meanwhile, they have negative correlations with most chronic diseases, while positive correlation with tow disease.

6. **Value:**
   - The essential step of this project is the pre-processing several datasets, where much of time was spent on. This manipulation added the most value to the raw datasets.
   - Comparing to raw data, using visualization and PCC also give a clearer observation about the results, like using regression to examine the correlation between sedentary behaviour and chronic diseases. This manipulation increases the reliability of the results.

7. **Challenges and Reflections:**
   - To find the relationships between sedentary behaviour and chronic diseases, datasets consistency in year and suburb are needed. It takes a lot of time to finding suitable datasets for this research.
   - The datasets I got contain several chronic diseases. Picking appropriate visualizations in Part A is quietly difficult. Heap Map was applied to figue out the correlation between required attributes easily.
   - The second research question is quite challenging, since several kinds of cardiovascular diseases' datasets are not available on the Internet, along with the death rate of chronic diseases. The investigation has a lot limitation.

8. **Conclusion (Question Resolution):**
   - Through comparing the sedentary status with chronic disease incidences in 80 Victorian suburbs, both positive and negative correlations have been proposed.
     The negative correlations against my speculation were analysed in **depth** because they are counterintuitive and they contradict with the study published by Department of Health (2017), which states that sedentary status is a critical factor for overweight, obesity and other chronic diseases.
   - Unfortunately, the second research question cannot be resolved form the investigation. In Part B, it was found that each disease is correlated with the other diseases, which complicates the study because the influence from other diseases cannot be kept as constants when comparing to sedentary status. Further multifaceted datasets are required to obtain a more reliable result.

9. **Bibliography**:

Department of Health (2017) – Sedentary Behaviour: http://healthywa.wa.gov.au/Articles/S_T/Sedentary-behaviour

Note: The margins of this report set in *word* is Top Bottom Left Right 2 cm.