Phase 2A

1.Title of Project: **Rising Health Issues and Communities - Identifying Neglected Causes of Chronic Disease**

2. Domain: **Health and Communities**

3 & 4.

The question I am seeking to answer is **whether and how people's everyday life potentially contributes to the chance of developing chronic diseases**. Chronic disease is Australia's biggest health challenge [1], so identifying causes is vital. The question will be examined from two different aspects with respect to modelled estimates of chronic diseases. The two aspects are: people's quality of life - namely daily behaviour, habit, the choice people make subtly, and underlying **lifestyle choices**; people's standard of living - namely **personal income**. In addition, this project aims to find specific leading causes of certain types of chronic disease in order to gain a better insight of the initial question regarding 'how'.

Local hospitals and health professionals can diagnose patients' medical condition by asking them questions related to the factors stated in the result. Schools and organisations can use those results to teach children and adolescents to be aware of their unhealthy habits and behaviours. The result could be used as an indicator of both official and self-assessments of people's health conditions. The result models could also be used for risk analysis across medical and insurance industries, or for educational purposes for governments to promote people to live healthier by doing small changes in their lives.

5. **Datasets**

These datasets are all available from AURIN. (https://aurin.org.au) Although these data are also available from other reputable websites (for example, ABS), AURIN provides a handy functionality on visualising and selecting data attributes and study area. So datasets are selected and downloaded from AURIN.

- **LGA11 Chronic Disease** - Modelled Estimate: Modelled estimates of various chronic diseases by LGA across Victoria along with the relative root mean square error of each estimate.
- **LGA Visit to Green Space** - Survey feedback of the percentage of survey respondents across Victoria who report have visited green space weekly or more frequent.
- **LGA Time pressure** - Survey feedback of the percentage of survey respondents across Victoria who report having time pressure or feeling rushed most of the time.
- **LGA Daily soft drink consumption** - Survey feedback of the percentage of survey respondents who report drinking soda every day over the last 7 days across Victoria.
- **LGA Sedentary behaviour** (sitting hours per day) - Survey feedback of the percentage of survey respondents who sit for greater than or equal to 7 hours a day.
- **LGA Inadequate sleep** (<7 hours per weekday) - Survey feedback of the percentage of survey respondents who report on average sleep less than 7 hours every weekday.
- **LGA Adequate Work-life balance -** Survey feedback of the percentage of employed survey respondents who disagree that their work and family life often interfere with each other.
- **LGA11-based B17B Total Personal Income** - From ABS Census 2011, it is a comprehensive dataset that provides information about weekly personal income by age and sex by LGA in Victoria.

6.

Raw data that have not been cleansed and integrated show little useful insights of the relationship between chronic disease and people's standard and quality of life. It would be difficult for governments or medical institutes to determine and identify the underlying factors of chronic disease with such limited information as separate datasets.

By joining datasets together – especially all survey data, authorities can have a better understanding of people's quality of life (lifestyle choices), and how the choice people make substantially relate to the risk of getting chronic diseases. As such, motivating authorities to make a change and focus efforts on educating the general public and providing necessary public service. By filtering through complex datasets – especially personal income data, authorities can have a better understanding of people's standard of life, and identify groups with higher incidence rates. As such, motivating authorities to make positive changes such as establish related health welfare or services. Without these necessary data wrangling, it would be difficult for authorities to decide what to prioritise on focusing regarding preventing residents from getting chronic diseases.

7 & 8. **Initial Investigation and Explanation**

The initial investigation is divided into several steps: data pre-processing, integration, as well as some initial results and visualisations.

Phase 2A

Data pre-processing

Initially, these large datasets need to be cleaned to fit the purpose of this investigation. The aim of data cleaning is to obtain useful and valid data of consistent formats. As mentioned, AURIN provides a great tool for choosing specific dataset attributes or schema prior to download, which has high efficiency in data cleansing – saves a huge amount of time on writing code to crop most of the useless data; cropped data with less space and length are handier for further cleaning. Another thing to note is that all data are in percentage except for LGA area codes.

Meanwhile, all datasets are CSV files, and pre-processing is done by 'pandas' library of python as the DataFrame object is a commonly used pandas object which is used to represent datasets. Jupyter notebook is used for a better and organised representation.

Data specified as below are cropped or manipulated after consideration.

- In overall, raw number data, data concerning gender, LGA name, and data outside of Victoria are cropped by AURIN unless specified otherwise. Check for missing data; non-incorporated data are cropped using python.

- Modelled Estimate data (1 dataset): High error data, (data with high relative root mean square error) are cropped for more reliable and accurate results. As a result, Chronic obstructive pulmonary disease and diabetes data are cropped to reduce errors of the result. Outliers are not cropped for more realistic results as high error data are already cropped, and the data are from a reputable source.

- Survey data (6 datasets): Only percentage data regarding specific LGA and LGA code data are kept. Data regarding confidence interval and average value are discarded as they are not useful for integration. As it is a survey data, outliers are cropped using boxplot in python; however, suspected outliers are outliers are kept to ensure versatility of data. Outliers are in turn replaced by the average data of Victoria State as provided in the raw dataset.

- Income data (2 datasets): Only the second dataset includes non-gender data, so the whole dataset of B17A is discarded. The B17B dataset contains hundreds of attributes, which is of less efficiency and impossible to pre-process in AURIN. So data regarding gender is discarded using Python code, but age data is kept for now, which may be useful in the later stage of the analysis. 'Persons' data are raw number data and are normalised into percentage data by divide by total population data at the end of each row. Gender-related data and age-related data cropped for two reasons – to reduce the size of this enormous datasets, and for consistency with the standard of modelled estimates data (modelled estimates data of chronic disease is for all age and gender groups.). Although there may be limitations, outliers are not cropped as they are from a reliable source (ABS) and this is done to prevent missing interesting results.

Integration

Two new DataFrames need to be created by inner-joining datasets to produce visualisations of the two aspects as listed in the investigation question. Before this, all survey data are joined together into one condensed dataframe ('survey_df'); weekly personal income data is grouped into several categories to organise and reduce redundancy of data; they are categorised into: '<0', '1-299', '300-599', '600-999', '1000-1499', '1500-1999', and '2000+' Australian dollars per week('income_df'). Chronic disease data is named as 'chronic_df'. Ultimately, survey data is joined with chronic disease data('chr_surv_df'), and income data is joined with chronic disease data ('chr_inc_df'). Methods are illustrated as below.

Datasets are inner joined by common LGA attributes. After data cleansing, all datasets match perfectly as they are all data of Victoria LGAs (as many as 79 LGAs in total). This helps to not losing any valuable data of every LGA since there are no missing data in each row of the datasets (if there are missing data, then the whole row need to be cropped). So the datasets are highly integrated, which will help the completeness and unbiasedness of the analysis results as a whole (actually, in this case, merge can be used instead of join, as the datasets are already matched up, but join method is used just in case and for consistency of style).

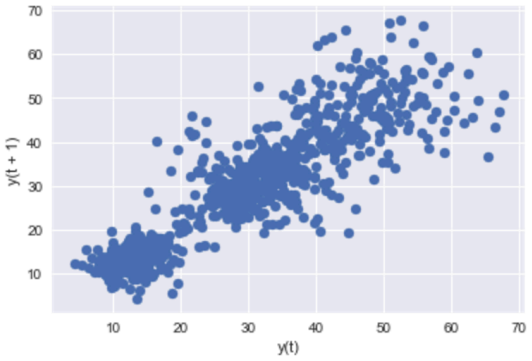
After this, column heading names are changed to human-readable name styles; columns are reordered and merged where it is necessary for visualisation in the later stage.

Phase 2A

Initial Result and Visualisation

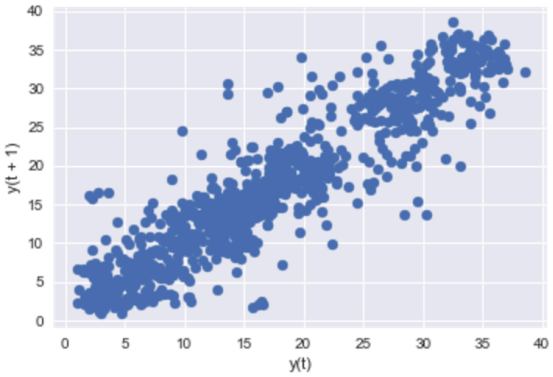
```
lag_plot(chr_surv_df)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x11c11c7b8>
```

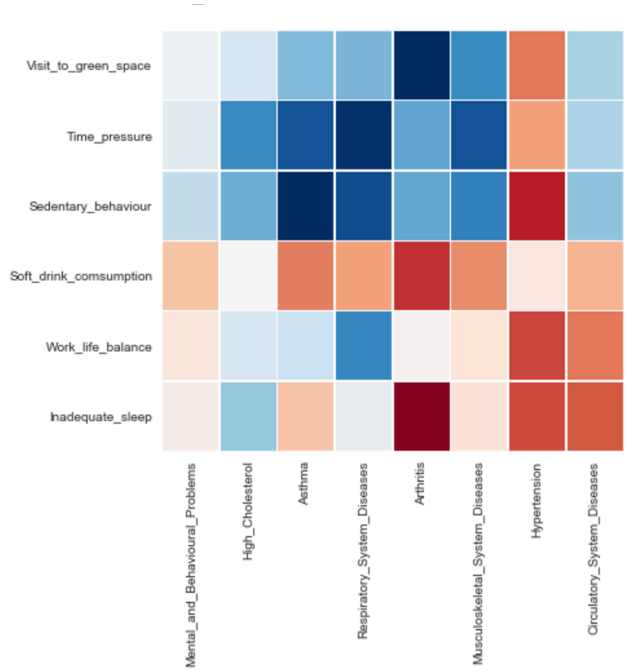


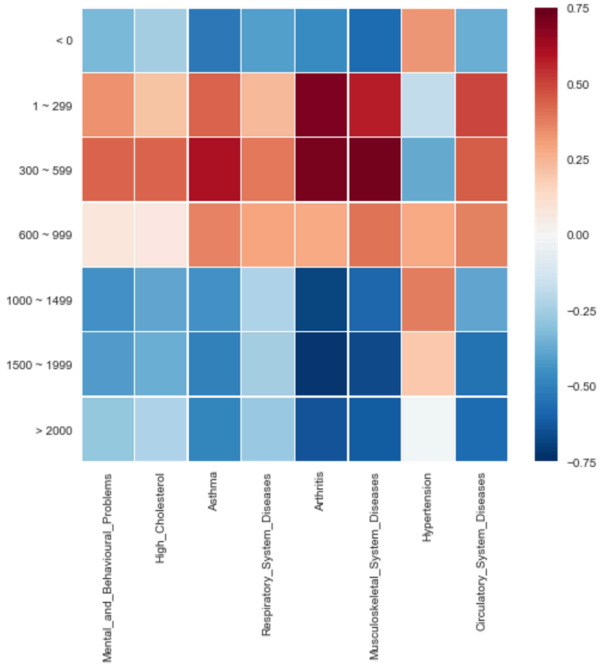

```
lag_plot(chr_inc_df)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x11c2913c8>
```



With these two condensed DataFrames ready for analysis, it is impractical and impossible to determine whether they contain correlation useful for analysis. Hence, lag plots are used to determine randomness of data and to see if it is worthwhile for further investigation. In general, random data do not exhibit any correlated behaviour in the lag plot, while non-random behaviour in the lag plot shows that the data are correlated in some ways and worth for further investigation. There are strong linear correlations as shown in the plots, so it can be stated confidently that the project is highly feasible and is very likely to yield interesting and useful results.





Furthermore, Pearson's correlation coefficient value r is calculated for each pair of columns from the dataframes. (Pearson's correlation only works for linear relationship; it is chosen since lag plots show linear relation). Then r values are plot into heat maps using seaborn library of python. Heat maps show some strong correlation at the darker shades, and in overall, there is a significant tendency of colour difference in different row – an overall red for weekly income under $1000; an overall blue for healthy lifestyles. However, sedentary behaviour data shows a different result to what was expected. All in all, this project is highly feasible in the later stage, and further in-depth analysis can be done to add more value to raw data in an innovative way. For example, use scatterplots for analysis on specific cases with significant tendency or correlation.

References

[1] Australia Institute of Health and Welfare "Australia's biggest health challenge". In: *Australia's health 2014* (20124), DOI: http://www.aihw.gov.au/publication-detail/?id=60129547205