Dixin Zhang

Rising Health Issues - Identifying Underlying Daily Factors of Chronic Disease

Domain: This project focuses on two domains: Health and Communities.

Question

The question this project is seeking to answer is **whether and how people's everyday life potentially contributes to the incidence rates of chronic diseases**. The term 'everyday life' is a broad concept, to specify, the question will be examined from two different aspects with respect to modelled estimates of chronic diseases. The two aspects are: people's **daily behaviours** - namely daily habit and underlying lifestyle choices; people's standard of living - namely weekly **personal income**. The project aims to find specific causes of certain types of chronic disease in order to gain a better insight of the initial question regarding 'how'. Chronic disease is Australia's biggest health challenge [1], so identifying causes is vital. It has long been difficult to cure chronic disease, if this study could analyse how people's daily behaviour affect the probability of getting chronic disease, and if government could target groups that are more likely to get chronic disease (for example, by income), there would be a high chance that chronic disease could be detected and treated in early stage, and would be beneficial to all Victorian residents.

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.

Pre-processing and Integration

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; cropped data with less space and length are handier for further cleaning. Also, all cleansed data are in percentage except for LGA area codes.

Meanwhile, raw datasets are CSV files; pre-processing is done by 'pandas' library of python since DataFrame is a commonly used pandas object that can be used for data representation.

In overall, raw number data, data concerning gender, LGA name, and data outside of Victoria are cropped by AURIN unless specified otherwise. Non-incorporated data are cropped using python. Missing values are replaced by mean values of given data by using the missing value method. Gender-related data and age-related data are cropped by Python 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 in 'chronic_df' and 'income_df' as they are from a reliable source (ABS) and this is done to prevent missing interesting results.

- Modelled Estimate data (1 dataset: 'chronic_df'): 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. This cleaned dataframe contains 8 columns, 78 rows; columns are grouped by different types of chronic disease and rows are grouped by LGA's.
- Survey data (6 datasets merged together: 'survey_df'): 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 by Python using boxplot by outlier detection method; however, suspected outliers are kept to ensure versatility of data. Outliers are replaced by the average data of Victoria State as provided in the raw dataset. All survey data are joined together into one condensed dataframe. The cleaned and merged dataframe contains 6 columns, 78 rows; columns are grouped by daily behaviours of people synthesised from survey data and rows are grouped by LGA's.
- Income data (1 dataset: 'income_df'): 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. For a more concise data representation, income data is grouped categorically to organise and reduce redundancy of data, ('<0', '1-299', '300-599', '600-999', '1000-1499', '1500-1999', '2000+' Australian dollars per week). The dataframe contains 7 columns, 78 rows; columns are grouped by income levels and rows are grouped by LGA's.

Table: Aggregated Clean Data

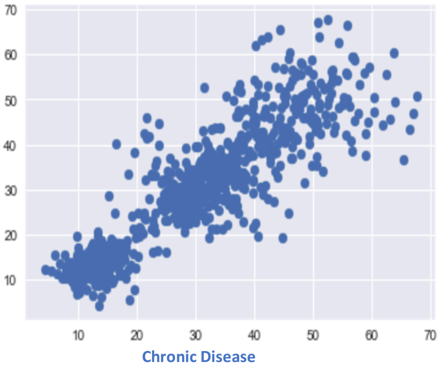
	'chronic_df'	'survey_df'	'income_df'
'chr_surv_df'	🟥	🟥	
'chr_inc_df'	🟩		🟩

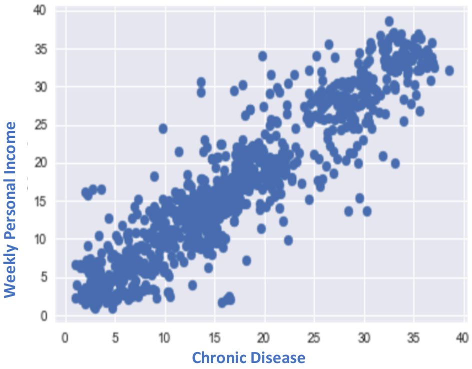
This table demonstrates how datasets were merged for the purpose of visualisation.

Moreover, columns are reordered and merged where it is necessary for visualisation in the later stage. As the table shows, two new DataFrames are created by inner-joining datasets to produce visualisations of the two aspects as listed in the investigation question. Ultimately, survey data is joined with chronic disease data to produce 'chr_surv_df', and income data is joined with chronic disease data ('chr_inc_df'). Methods are illustrated as below.

Datasets are joined by common LGA attributes. After data cleansing, all datasets match perfectly as they are all data of each Victoria LGA (in total 79 LGAs), which means that they all contain one column that is identical. Hence, the datasets are highly integrated, which helps 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).

In addition, column heading names are changed to human-readable name styles during this stage.



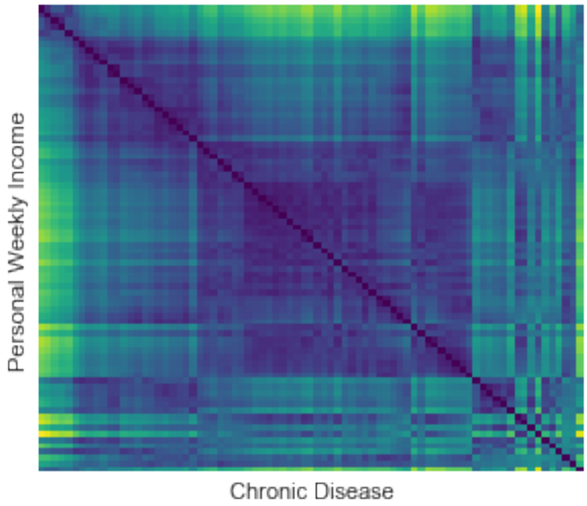


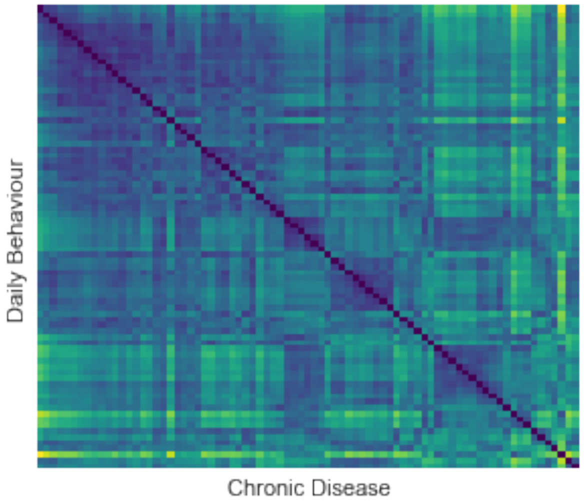
Dixin Zhang

Results

With these two condensed DataFrames ready for analysis, it is impractical and impossible to directly determine whether they contain correlation useful for analysis. Hence, lag plot method in pandas is used to determine randomness of data and to see if it is worthwhile for further investigation. 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, which proves high feasibility and value of the project is highly feasible and encourages further investigations.

For further investigations, clustering method is chosen for analysis; dissimilarity matrix is visualised in order to determine the clustering structure, the matrix is reordered with the aid of the VAT (visualisation for clustering tendency) algorithm.





As the dissimilarity matrix shows, there are four strong clusters along the diagonal in the heatmap of chronic disease and personal income, while there are four moderate clusters along the diagonal in the heatmap of chronic disease and daily behaviour (survey data).

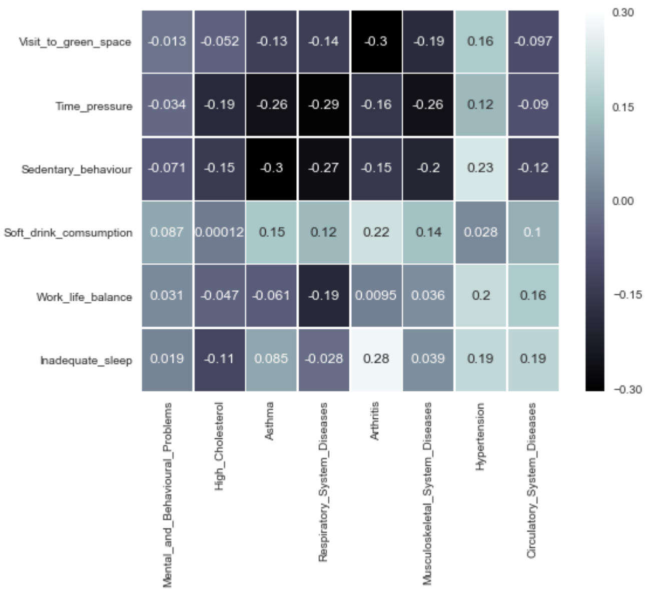
Thus, it was decided to calculate Pearson's correlation coefficient value r for each pair of variables from each dataframe in order to further visualise correlation tendency in a similarity matrix based on r value. (Pearson's correlation only works for linear relationship; it is chosen since lag plots show linear relation).

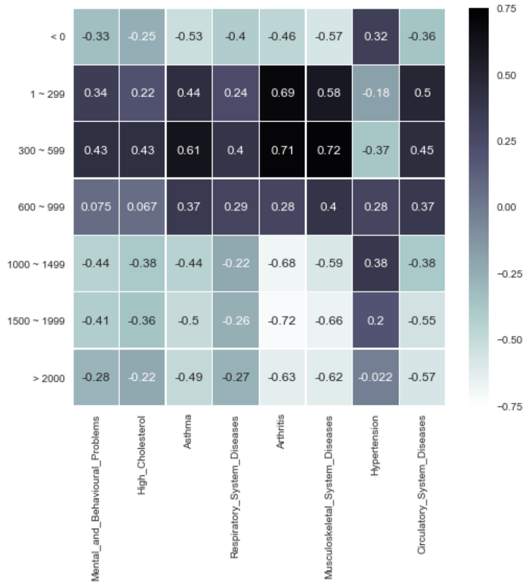
Next, r values are plot into heat maps using seaborn library of python. Heat maps show some strong correlation at the darkest and lightest shades, note that although some shows strongly negatively correlated, they are still useful results as we are not looking for positive relationship but linear relationship in general.

As for heatmap of chronic disease and survey data (daily behaviour), correlation appears to be moderate (consistent with the previous matrix), r value ranging from -0.3 to 0.3. Although some daily behaviours do not appear to have linear relationship with types of chronic disease, there are some specific highlights of our findings.

- Less sedentary behaviour leads to asthma and respiratory diseases
- low frequency of visits to green space, higher soft drink consumption and inadequate sleep leads to arthritis

- poor work-life balance, sedentary behaviour and inadequate sleep relates to hypertension;

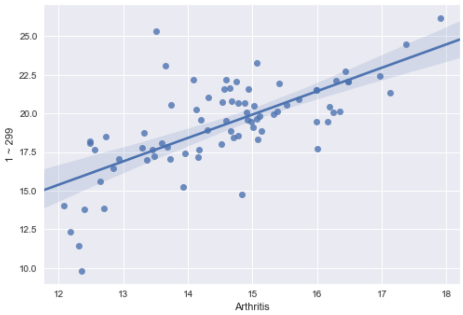


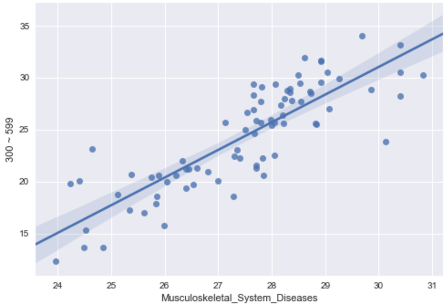


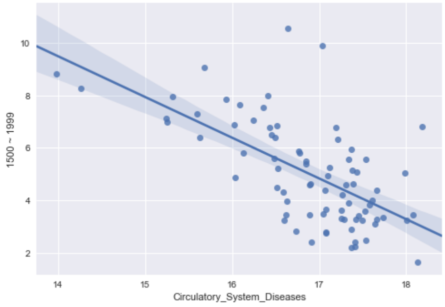
Strong and interesting correlative behaviour is shown by the heatmap of r values between chronic disease and various levels of personal income.

- The matrix shows darker colour for people with income lower than AU$1000, and lighter colour for people whose weekly income exceeds AU$2000.
- Except for people with negative income and people whose weekly income exceeds AU$2000, in overall, there is a clear trend that shows people with higher income tend not to have chronic disease (except for hypertension).
- This finding exhibits most significantly for circulatory system diseases and asthma, where the correlation r value strictly decreases as personal income increases. For example, the r values of asthma are 0.5, 0.45, 0.37, -0.38, -0.55, -0.57 as income goes up.
- As for people whose weekly income is greater than $2000, the above pattern mostly applies, but exhibits a significant drop on r for mental and behavioural problems and high cholesterol.
- As for people with negative income, actually, this category may largely contain full time students rather than purely unemployed people, so the trend does not apply to this group.
- Hypertension highlights a special case, where people with higher weekly income are more likely to have hypertension (again, except for people with negative income or people whose income is greater than $2000).

The scatterplots show some representative relationship between personal income and different types of chronic disease.







Value

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, or 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.

Raw data that have not been cleansed and integrated show little useful insights of the relationship between chronic disease and people's everyday life.

Challenges and Reflections

In reality, there is no known exact causes of chronic disease, and people's daily behaviours of course cannot be grouped into only six categories, there surly are far more factors of incidence rates of chronic disease than that stated in the result section. For example, gender, race and generic heritage are not considered; different race may lead to different eating habits, which could be a large factor for some chronic disease. Moreover, the daily behaviours specified in this study might be more correlated with other types of disease, but it remains unknown until other investigations are done.

- It was unknown if useful correlation exists until the stage of data visualisation, before that, pre-processing and integration of multi-dimensional data require heavy workload.
- One limitation is that the chronic disease data is a modelled estimate done by ABS, but not an actual statistics gathered from all hospitals in Victoria.

Question Resolution

Thus, the initial question is answered. In Victoria, people's weekly income strongly correlates to the chance of developing chronic disease, and people's daily behaviours do have relationship with getting chronic disease. Regarding the question of "how", certain types of daily behaviour contribute to certain types of chronic disease as illustrated in the result section, and there is a clear and significant trend of higher income and lower chronic disease rate despite some exceptions as illustrated in the result section. By filtering through complex datasets – authorities can have a better understanding of how people's daily life relate to higher incidence rate of chronic diseases. Such, motivating authorities to make positive changes and focus efforts on educating the general public and providing necessary public service as well as establishing related health welfare or services. Without necessary data wrangling, it would be difficult for authorities to decide what approach to take on lowering rate of developing chronic disease.

Code

400-500 lines of python code were written from scratch, primarily focusing on data cleansing, integration and visualisation of dataframes. Due to time constraint, code may not be efficient enough, and duplicated code are deleted and are written as function. Jupyter notebook is used for a better and organised representation. Python libraries used are pyplot, numpy, matplotlib, pandas, and seaborn. Code of the VAT algorithm is sourced from workshop 6 of this class, while other code is written from scratch. Specific steps of python code is detailed in the integration section.

Some methods implemented in Python code are: data transformation method, missing value method, outlier detection method, visualisation method, and clustering method. Scatterplots and heatmaps are visualised using seaborn library, lag plot is visualised using pandas.

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=6012954

[2] AURIN, DOI: https://aurin.org.au

[3] J.Bailey, University of Melbourne "lecture6-2017-final", In: *COMP20008 Elements of Data Processing*