COMP20008 Project Phase 3: Report

1 Community chronic disease risk and health risk: correlation analysis and community feature investigations

Author: XuLin Yang, 904904

2 Domain: communities, health

3 Questions:

(1) Identify the correlation between community citizens' chronic disease risks and citizens' health risks.
(2) What features in communities causes citizens to have these health risks and increase disease risks?

4 Datasets used:

All listed below has .csv data file and .json metadata file. And can be downloaded from aurin. All of these data are statistically counted in 2011.

1) LGA11[i] Chronic Disease Modeled Estimate
2) LGA11 Health Risk Factors Modeled Estimate
3) LGA11 Psychological Distress Modeled Estimate

These three data sets have the similar structure. All of them are labelled by LGA area code. As its name described, they describe features in each cases. See detailed metadata table at the end. (1) contains ten chronic diseases and data confidence level ('xxx - PRMSE'). (2) contains four health risks and data confidence level. (3) contains high psychological distress experienced and data confidence level. These data sets have noise data in about 0%-5% specified by confidence columns.

4) Local Government Area LGA profiles data 2011

This data set is a big data set. It is labelled by lga code. It illustrates facilitate service planning and policy development by enabling access to a broad range of data. It is used to analyse community features. See detailed metadata table at the end. There are something missing in the table. Although there are some outliers in the table, they can be used in later investigations.

	Data_set	#Rows	3 example columns	Column data type	Realiable data%
2	LGA11 Chronic Disease Modeled Estimate	80	"Arthritis"	Discrete	70%
3			"Arthritis_RRMSE"	Qualitative	
4			"Local_Government_Area_Code"	Categorical	
5	LGA11 Health Risk Factors Modeled Estimate	80	"Current_Smokers"	Discrete	84%
6			"Current_Smokers_RRMSE"	Qualitative	
7			"Local_Government_Area_Code"	Categorical	
8	LGA11 Psychological Distress Modeled Estimate	80	"People_felt_high_mental_pressure"	Discrete	97.50%
9			"People_felt_high_mental_pressure_RRMSE"	Qualitative	
10			"Local_Government_Area_Code"	Categorical	
11	Local Government Area LGA profiles data 2011	79	"Unemployment_rate"	Discrete	70%
12			"People_lack_physical_activity"	Discrete	
13			"Local_Government_Area_Code"	Categorical	

5 Pre-processing:

In the data pre-processing stage, firstly, we load data into excel to check whether each data column is matched its schema. Because sometimes, there can be a column named 'geometry_field' (which is not visible on aurin but is included automatically) but geometry data take 3 columns. In order to solve this problem, just delete this noisy data column. Then csv files are loaded by pd.csv_reader().

Secondly, after checking and loading data, columns with meaningful metadata is replaced. A columns renaming customized function rename_columns() us used. By doing this, real meaning for attributes can be used and all dataframes can have a common key 'Local_Government_Area_Code' when merging tables.

Data in data sets (1)-(4) are going to be split in three pd.DataFrame() for disease_risks, health_risks and lga_profile.

Errors in column name like in 'LGA11 Health Risk Factors Modeled Estimate.csv' {'obese_m_me _3_rrmse_3_11_7_13' : Obese Females 18 Years and Over - RRMSE } is wrong and corrected to 'Obese Males 18 Years and Over - RRMSE'. By doing this, wrong understanding of investigated feature will not be encountered.

Some useless rows (e.g. row.lga_name = 'Unincorporated Vic') are dropped. In consequence, higher efficiency and more accuracy can be achieved for analysis. Similarly, useless columns are not selected from raw data.

When splitting into dataframes studied later, several of this columns(e.g. 'GP_attendence' for 1000 people normalized to average citizen's 'GP_attendence') not normalized are normalized into the same range in the stage. These features are divided by population in that lga. As a result, the purpose to explore relations between features is achievable and feature going to be discussed are normalized in range $\in [0, 1]$.

There are some missing data in tables (e.g. lga_profile.LGA_land_use(rural/residential/industrial)) has some missing data as it has something like '<1%' in .csv and can't be read by pd.read_csv(). So using fillna(0) method can make data consistent with the original data set with losing minimum information.

Unreliable data will be dropped if 'xxx - PRMSE' == 1

By doing all the pre-processing steps described above, data are ready for analysis. Although some outliers might be in these data, most of them are dropped at unreliable row dropping steps. Others are investigated in further analysis.

6 Integration:

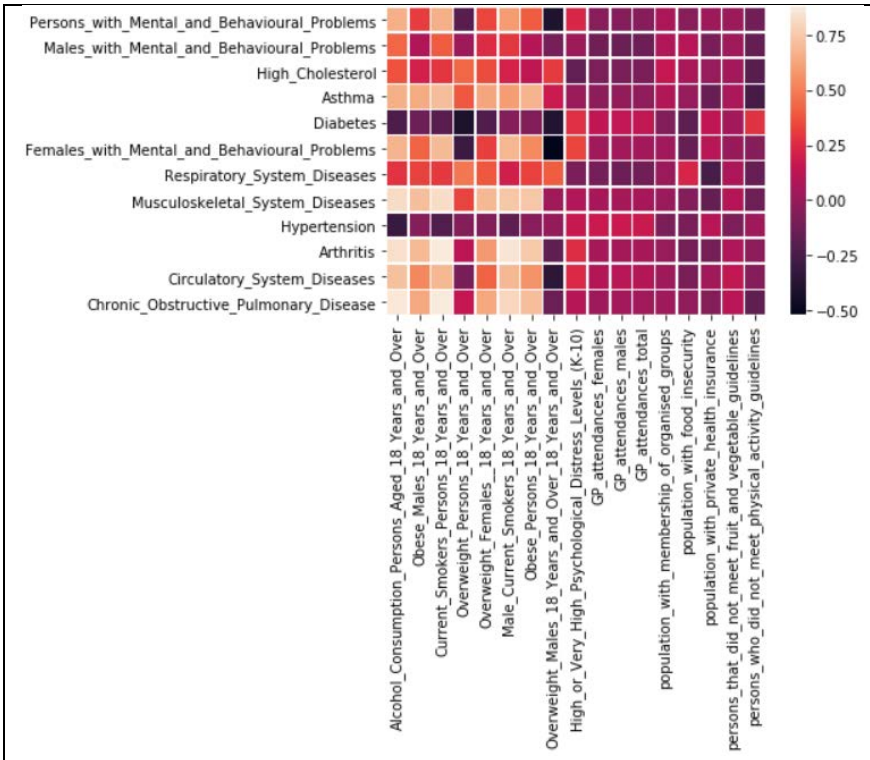
Some data needed to be derived from one columns. For example, LGA_name is a string with 'Cities (C), Rural Cities (RC), Boroughs (B) and Shires (S)' and lga name and type and be transferred. This group lga by its geographical location.

Some columns like 'Number_of_hospitals/health_services' are used to calculate the health service provided for each person by let 'Number_of_hospitals/health_services' divided by population in that lga area. The limitation of this method is: Some health service can be children cared, disable people and dental care service which is not related to the disease risk going to be discussed. As a result, some errors might be generated. Although error can be generated, this derived result can be used to reflect community level health infrastructure development level. And it might be a potential feature that affects health risk.

As these data sets are selected based on lga area code at year 2011, joining data sets can provide further information we needed for further investigation. The dataframe can be inner joined by its index based on lga_area_code as described in preprocessing stage. And we are able to explore deeper connection between community disease risk, community health risk and community feature attributes. The limitation of this method is some tables contain more data but lost at merging stage. That is ratio of using data is not 100%. And in the later analysis, all rows in different table can one to one matched which will not causes data incompleteness.

7 Results:

After investigating figure (1) heat map shown below, several health risk is highly or moderately correlated. Some of them are picked for further investigation.



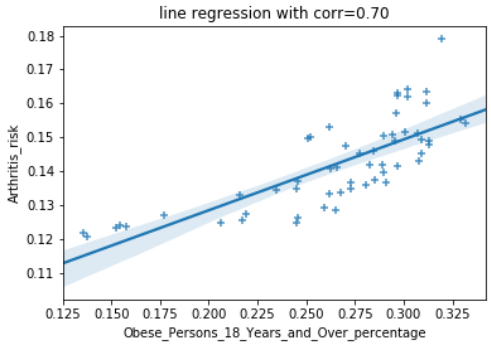
figure(1) heat map against disease risk and health risk

• The community with higher percentage of adults drinking alcohol has a higher risk of having the circulatory system disease($r^1 \approx 0.72$).

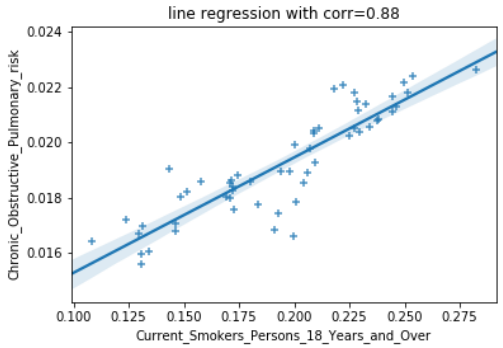
• The community with higher percentage of obese adults has a higher risk of having musculoskeletal system diseases($r \approx 0.72$) and arthritis($r \approx 0.70$).

• The community with higher percentage of adults who smoke results in a higher risk in asthma($r \approx 0.71$) and chronic obstructive pulmonary disease($r \approx 0.88$).

From the above investigation, linear regression model are applied to the two scatter plots ("obese adults v.s. arthritis risk" and "adult smoker percentage v.s. chronic obstructive pulmonary risk") Both figure(2) and (3) shows a significant increase in community's health risk will results an increase in community particular health risk. The relationship between health risk and disease risk is clearly illustrated by graph. Although there are some outliers in scatter plot. For example, some points at left bottom in figure(2) are outliers. But these outliers can also be used and provide potential information in the later discussion.



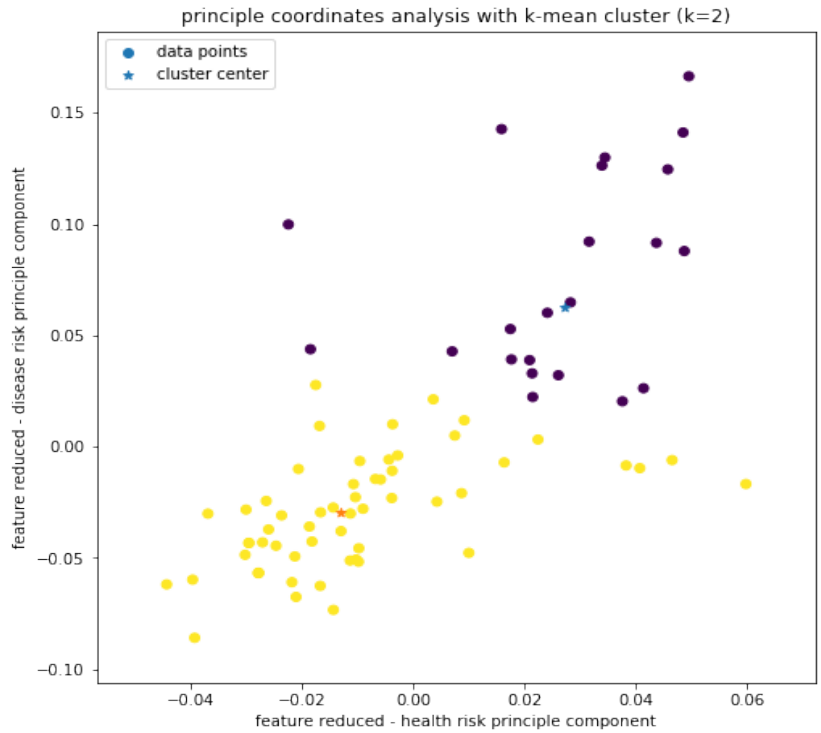
figure(2) linear regression model for obese adults v.s. arthritis risk



figure(3) linear regression model for adult smoker percentage v.s. chronic obstructive pulmonary risk

Then apply principle coordinates analysis in the discussion. As demonstrated in the correlation analysis above, five chronic disease (disease risk: circulatory system disease, musculoskeletal system diseases, arthritis, asthma and chronic obstructive pulmonary disease) are related to three health risk (smoker, obese and drinking alcohol). Both disease risks and health risks are feature reduced to two principle axes.

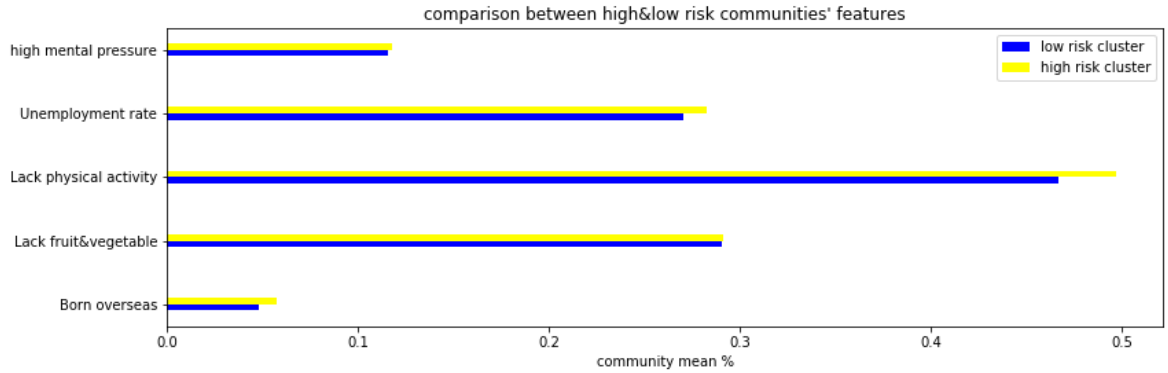
[1] Pearson correlation coefficient



figure(4) principle coordinated analysis

The two axis has variance [0.8, 0.9] indicates both health risk and disease risk can classify communities. K-mean cluster technique with k=2 is applied to the feature reduced scatter plot. The points are clustered into two groups (low risk cluster and high risk cluster)

Then we analyze two clustered data and studies community feature. As illustrated by figure(5), high risk community always have a higher percentage of investigated feature then low risk cluster. The difference between two types of communities can be summarized in these aspects: people felt high mental pressure, high unemployment rate, lack of exercises, and immigrants' population percentages. And the differences are about 3% to 5%. Lack of fruits and vegetables is not one of influences



figure(5) community feature for clustered data

In conclusion, there is clearly plotted evidence for chronic disease risk is correlated to citizens' health risk. Further investigation explores the aspects that cause community to have high risk. And the suggestion for high risk community can be summarized at this stage. The limitation of this analysis is it can be explored in a more detailed geographical area for smaller groups of people. Other potential factors like ages, races and incomes are not fully explored in this analysis.

8 Values:

Firstly, this analysis on the two questions is based the data selected from four data sets covering the three features going to be discussed. Not all columns in data files are needed for analysis. Secondly, some

wrong and missing data in the table needed to be corrected. Thirdly, different data columns have different ranges of data which make it impossible to compare and do visual analysis. Techniques of normalization are applied to multiple data columns. Furthermore, new feature can be derived from the existing data. And the newly derived can be used as features for communities for future analysis. Moreover, visual plotting techniques like heat map helps to give an intuitive interpretation of correlation of disease risk and health risk and unreliable data are dropped. To sum it up, with steps mentioned above, data is ready for analysis. If not doing all these above, raw data can't give appropriate information to solve proposed questions.

9 Challenges and Reflections:

The most time consuming part in this project is finding the correct data set to investigate with. Due to the data set selected, more potential health risk might influence chronic disease risk. Selecting columns also consumes a lot of time as these are big data sets and finding relevant columns is not easy. Another challenging part is doing clustering. Initially, clustering is applied to related health risks and disease risks respectively which results a lot of repeating steps and visualizations. After got hint from tutors, using pca analysis to do feature reduction respectively and then applying K-mean clustering can analysis multidimensional variables. After doing this, the efficiency of analysis is improved.

10 Question Resolutions:

From the results found above, the question (1) is answered in the correlation analysis section. The question (2) is answered in the second section of analysis. With the help of graph, how one feature influences the other is demonstrated clearly. The community with high risk (e.g.: Moreland, Kingston, etc) can decrease unemployment rate, provide health care for new arrivers and encourage people to do more physical activities.

Australian Government of Human and Health Service might be interested in the strong correlation between chronic disease risk and citizens' health risk. Local government might be interested in people felt high mental pressure, high unemployment rate, lack of exercises and immigrants population percentages and how these community features can influence health risk and disease risk. With the answers to these two questions, government department can know the health level of Victorians lga wide. For each local government, they can have a clear idea about what the kind of policies or community activities they can make or organize to help citizens to have a healthier life and lower the health risk in lga.

11 Code:

The python codes are about 400 lines. No other publically available code and no non-python codes are used.

- Primary pyhton library like pandas and numpy are used.
- re library is used to make it easier to process columns and reduce redundant works.
- json is imported to rename columns and read metadata from json files.
- seaborn library is used to do line regression modelling.
- sklearn library is imported as it is required for principle coordinates analysis and K-mean clustering.
- matplotlib library is required for graph plotting.

12 Bibliography

[i] Australian Statistical Geography Standard (ASGS): Volume 3, July 2011 http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/by%20Subject/1270.0.55.003~July%202016~Main%20Features~Local%20Government%20Areas%20(LGA)~7