PHASE 2

Name: Renjie Meng; Student ID: 877396

2018-04-23

1. Title: Could we prevent avoidable death by reducing potential health risk factors？

Identify approaches to reduce avoidable death in Victoria

1. Domain: health
2. Version2:

The question of this project is trying to answer is “Could we prevent avoidable death by reducing potential health risk factors?”

Many unnecessary deaths like cancer are due to potential health risk, such as obesity and unhealth diet, which may be prevented. Local government and people would be interested in this topic because if we can effectively prevent avoidable death, we will make life safer and happier.

People would be benefit if there is relationship between avoidable death and potential health risk factors, they could reduce the chance of dying from avoidable death by avoiding and eliminate the related potential health risk factors. Furthermore, the local government could post the effect of potential health risk factors could lead to death to the public in order to encourage people to avoid and eliminate potential health risk factor.

What are the ways to reduce avoidable death in Victoria? We do not want death unnatural, the doctor and citizens will be interested in this topic. They can gain information about how to prevent avoidable death happen, the doctor could predict more precisely with this information to help people to live longer.

What who how

1. Version2

If the raw gives us a good positive relationship b/w one specific potential health risk factor and avoidable death. The innovative information can be advice about the taking action of avoiding and eliminating the specific potential health risk factor in order to prevent avoidable death is reasonable.

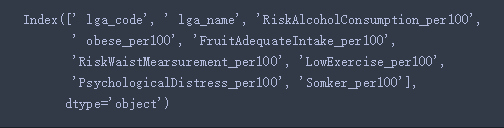
I would start to find approach from the relationship between different health risk factor and different kinds of avoidable death. In order to figure out, appropriate approaches to avoid preventable death. If there is a good positive relation between number of people who has specific health risk factor and number of avoidable death, or there is a negative related between number of people fruit intake adequate and avoidable death, then we could give advice on preventing avoidable death by increasing fruit take and reducing health risk factors when it is not too late.

Lack of knowledge about how new knowledge can be created

No reference to previous work

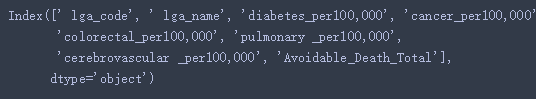
1. The two datasets displayed as bellow:

* Dataset1 is Estimated number of males/females/people aged 18 years and over whose health was at risk in one of the following categories: psychologically distressed, blood pressure, overweight, obesity, waist measurement, smoking, alcohol consumption, fruit consumption and no or low exercise, 2014-15And, I choose risk waist measurement, high blood pressure, fruit intake, obese, low exercise, psychological distress and smoker per 100 population to analysis.



Link: <https://portal.aurin.org.au/>

* Dataset2 is ‘LGA15 Avoidable Mortality-By Selected Cause - 2010-2014’ listed different kinds of Avoidable Mortality Causes at age 0-74 from 2010-2014 in Victoria. The specified causes of death are: cancers, colorectal cancer, breast cancer, circulatory system diseases, ischaemic heart disease, cerebrovascular disease, respiratory system diseases, chronic obstructive pulmonary disease, deaths from select external causes of mortality, suicide and self-inflicted injuries, other external causes of mortality, transport accidents. And, I select diabetes, cancer, colorectal, pulmonary and cerebrovascular to analysis.



Link: <https://portal.aurin.org.au/>

1. Lack of analysis about the raw data

Added about the analysis of the raw data.

* The raw dataset contains too many features so I reduce them by deleting unnecessary features with excel. The raw dataset of avoidable death contains total number of avoidable death of specific cause. But at the beginning I want to find relationship between each health risk factor and total number of all kinds of avoidable death in general. To determine which potential health risk factor is related with avoidable death.

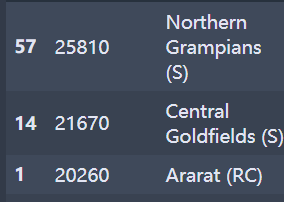
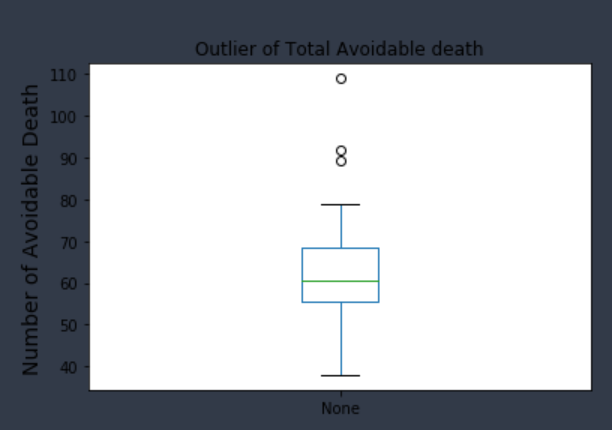
For the data preprocessing, I delete all the columns except the percentage column and index by excel for both dataset. For the rest columns, I renamed the in the format with no space between each char. Then I load the tow data sets, and using pandas data frame to add a new value called ‘Avoidable\_Death\_Total’ indicates total number of avoidable death per 100,000 population. As well as, delete all the row has null value with the .dropna() methods. For data integration, I integrate two data sets on ‘lga\_id’.

For data analysis and visualization, I would plot scatter plot between avoidable death and each health risk factor in order to find correlation. If there is a good or strong which mean greater than 0.3 for Pearson correlation, I would continue to use these two data sets to answer my question. And I would drop the useless health factor where it has weak correlation with ‘Avoidable\_Death\_Total’

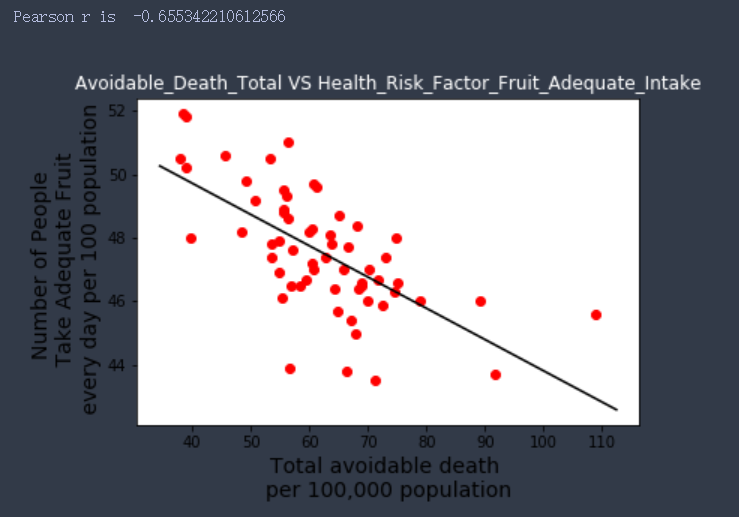
1. Lack of outlier and performance about missing value as well as labels for figures

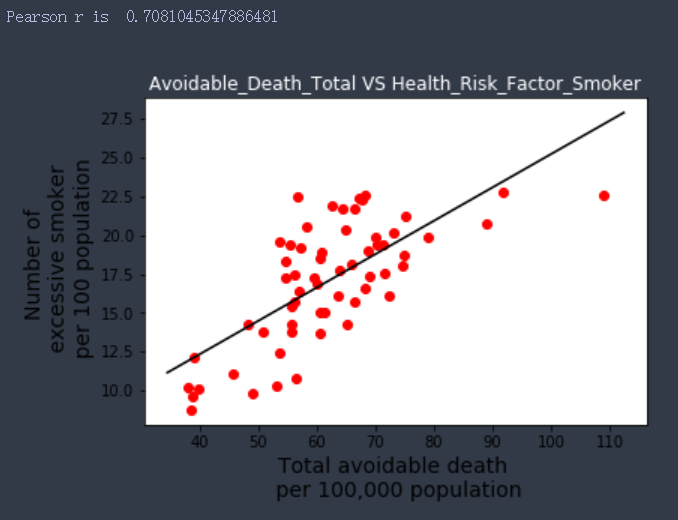
Labels has already been added on.

Missing values is solved by deleting every row containing missing values.

Outlier  There are three outliers in Avoidable\_Death\_Total and there are the three-area illustrated above. It is ok to preserve these outliers since it is real data, but we must pay more effort to these three area for reducing avoidable death.

After finishing the first step of data analysis, I move all the noise data away and I add a new value ‘Avoidable\_Death\_total’ indicates total number of avoidable death aged 0-75 in Victoria from 2011-2014. This is because I want to figure out the overall relation between avoidable death and each health risk factor. In order to determine whether my project is feasible or not.





For data analysis and visualization, I plot nine scatter plot between avoidable death and each health risk factor rate in total, but I dropped 3 weak correlation plots. Then I learn the learn linear regression as well as Pearson correlation for each scatter plot. Then, I choose two representative plots among whole plots which is Figure 1 and Figure 2 above.

Figure1 is scatter plot between ‘Avoidable Death Total Per 100,000 population’ and ‘Number Of People who take adequate fruit each day per 100 population ‘. The Pearson correlation of the two attributes are -0.655 which is nearly a strong negative relation. And the linear regression in plot also show a negative relation between two attributes. This means people who take enough fruit everyday would have a relatively low possibility of get avoidable death. This indicates that people can prevent avoidable death by taking enough fruit every day. Also, this support my project of identifying methods to prevent avoidable death.

Figure2 shows the scatter plot between ‘Total Avoidable Death per 100,000 population’ and ‘Smoker per 100 population’. There is a Pearson correlation of 0.708 between two attributes, which means a nicely strong positive relation. Furthermore, the linear regression also illustrates a positive relation between two attributes visually and intuitively. All the information above indicates that there is a trend that the avoidable death has a high possibility to happen on those smokers. Thus, the approaches of preventing avoidable death can be pay more attention to smokers and increase the frequency of going to hospital as well as helping them quit smoking. Again, this support my project of identifying approaches to prevent avoidable death.

Therefore, there is a good correlation between two datasets which means I can continue my plan.

1. The future plan is to discuss how to prevent each kind of avoidable death depend on health risk factors. As well as, how to help people prevent avoidable death for each area.

Since there is a nice correlation between total avoidable death and each health risk factors, I could continue the first program. Then, there are ‘lga\_code’ and ‘lga\_name’ in both datasets, I could group by region in order to figure the relation between area and the approaches of preventing avoidable death in order to take action with respect to each region effectively.

5. pre-processing

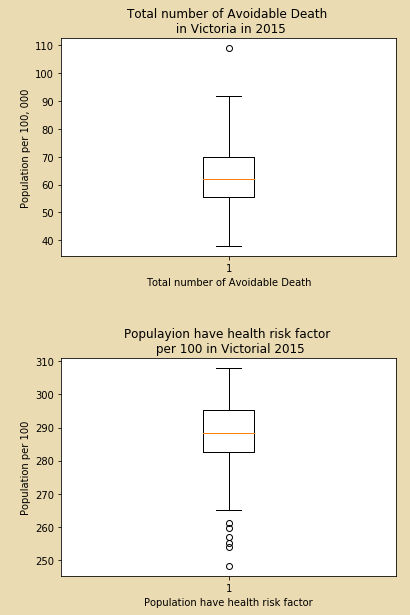
* Data format
  + There are lots of noisy data in both original datasets which are the features that I do not need. Then I delete the features that I do not want by excel. Then the feature name for both datasets are consisting of abbreviations, I change them into full spelling with no space between each word. Furthermore, move the ‘lga\_code’ to the first column in both dataset for further data integrating. Sort both datasets on ‘lga\_code’ in an ascending order by pandas for future integration as well.
* Missing value

I would do data integration first, then deal with missing values since it is easier to delete records that contains too little information from one table than from two tables.

In dataset “health risk factor”, after deleting the records above there is no missing values.

For dataset “Avoidable Death Causes”, there are 18 rows have missing value, it is not suitable to deleting all of them. And I do not have that expert knowledge to fill these missing values. But I can do imputation. Thus, my strategy is to choose missing values to do imputation with mean value. And I will choose missing value by the number of null value it contains. There are five main causes of avoidable death, I decide to remain records with at most 3 null values, which is about deleting 4 records. This is because these records contain less information and I think I can throw them away.

* Adding values

The raw datasets do not have representative data feature for comparing, then I created new features from raw data. Two new features added, one is the ‘Avoidable\_Death\_Total’ which sum all kinds of avoidable death per 100, 000 for each area. Another is ‘Health\_Risk\_Factor\_Total’ which sum all people have health risk factor per 100 excludes the Adequate Fruit Intake feature for each region. Since Take Adequate Fruit is health.

**Figure 1 Boxplots**

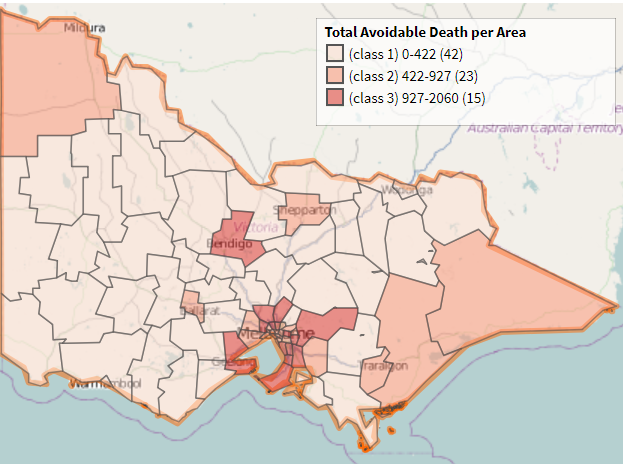
- Outliers

By analysising the boxplot of total number of avoidable death, I see that there is an area have unnormal high avoidable death and the area is Northern Grampians (S). I would recommend those area’s people with more power, but it is still ok to preserve it. Then the second boxplot of population have health risk factor only gives extreme small outliers which means those area’s people are health, it is ok to preserve them.

6. Integration

- integrate on common primary key feature ‘lga\_code’. First there are exact same lga code in both dataset, so it is easier to integrate them on this feature. Then, I sort both dataset on ‘lga\_code’ at ascending order with excel.

7. Results



8. Value