PHASE 3

Name: Renjie Meng; Student ID: 877396; Date: 2018-05-08

1. Tittle: Prevent avoidable death in Victoria
2. Domain: Health
3. Question: "Could we prevent avoidable death by reducing health risk factor in Victoria?" In 2015, around 4800 people died from avoidable death. Many avoidable deaths like cancer may due to potential health risk factor such as smoking and obesity. Since this topic covers citizen's health, the local government and citizen would be interested in because they want a safer and longer life. If there is negative relationship between health risk factor and avoidable death, people and the local government may be benefit. People could prevent avoidable death by avoiding and changing health risk factors, as well as the local government could arise people's health awareness by publish the relation between health risk factors and death.

```
1  print("Total Avoidable Death in Vic in 2015 is %.0f people." %(BigDF['Avoidable_Death_Total'].sum()))
```

executed in 20ms, finished 08:54:31 2018-05-08

```
Total Avoidable Death in Vic in 2015 is 4795 people.
```

4. Datasets:

   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.

```
Index(['FruitAdequateIntake_per100', 'LowExercise_per100',
       'RiskWaistMeasurement_per100', 'Somker_per100',
       'PsychologicalDistress_per100', 'obese_per100',
       'Health_Risk_Factor_Total'],
      dtype='object')
```

  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, ischemic 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.

```
Index(['diabetes_per100,000', 'cerebrovascular_per100,000',
       'colorectal_per100,000', 'cancer_per100,000', 'pulmonary _per100,000',
       'Avoidable_Death_Total', 'lga_code', 'lga_name'],
      dtype='object')
```

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

5. Preprocessing

- Data Format: By inspecting the raw data, there are many noise data. I delete useless columns and then rename remain columns with the format that no space in names. Then, I swap the column position for primary keys of both datasets for future integration. Then, I sort both data set on primary key in an ascending order.
- Missing Value: By inspecting the raw data, there are 18 out of 80 rows have missing values. It is not good to delete them all or input with 0, since about 25% of the dataset will be curious and the datasets will be not representative. Then, it is also not good to input all missing values with mean, because there are 5 of 18 missing values miss more than 3 of 5 field attributes, if I fill them, it would change the variance of datasets obviously, which is not what I want. Therefore, I decide to move records with more than 3 missing values per row as well as fill those remain records with mean values. My methods could change the variance of datasets and loss some data. However, I achieved balance between losing data and losing variance.
- Outlier Detection: I did outlier detection for three features.

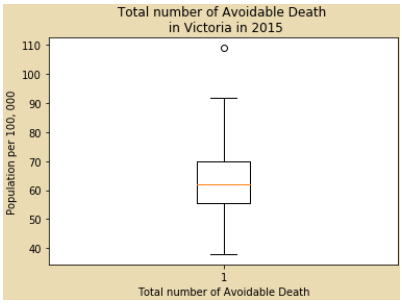

Figure 1

i. First one is the total number of avoidable death in each area. And the result boxplot (Figure 1) shows that there is an area has very large number of avoidable death. By checking datasets, I found the area is "Northern Grampians (S)". The value is 108 which does not affect further processing, so it is ok to maintain it. However, in the final report, I would suggest the government to pay more attention to Northern Grampians.

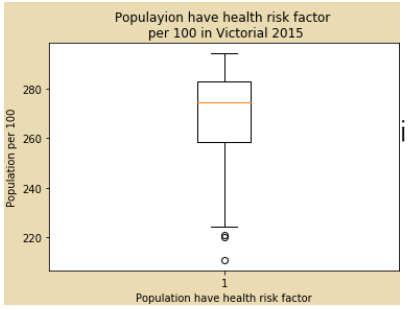

Figure 2

ii. Second one is population having health risk factor for each area. The boxplot (Figure 2) shows some very small outliers which mean small population have health risk factor in those area. These outliers is true values will not affect further processing, so it is ok to have them.

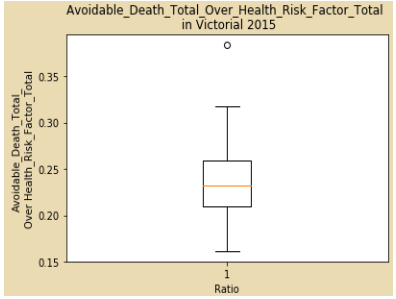

Figure 3

iii. The last one is the ratio of total avoidable death over population having health risk factor in each area which try to illustrate the influence of health risk factors. The boxplot(Figure3) shows one area has extremely high ratio. After checking, the area is "Northern Grampians (S)" again. I must suggest the government to focus on this area for preventing avoidable death.

6. Integration:

I integrate both datasets on 'lga_code' and 'lga_name'. By viewing both whole datasets, they have exactly same 'lga_code' and 'lga_name', it is easy and good to integrate. In the data processing, I have already sort both datasets on 'lga_code' and 'lga_name' in an ascending order. Then, I integrate two datasets. Fortunately, my method does not occur issues, since two datasets has exactly same primary key.

7. Results

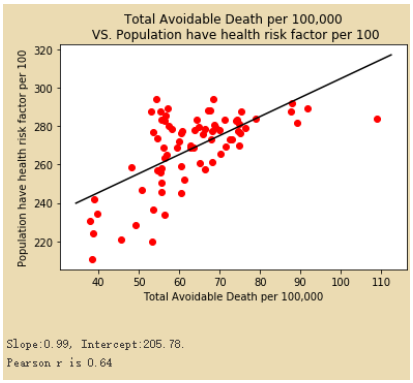

Figure 4

■ Figure 4, illustrates that there is a nearly-string correlation between total avoidable death and population have health risk factors. Which may indicate that less population having potential health risk factor, there would be less avoidable death in each area. This support further investigation on this topic about specific relation between each kind of avoidable death and each kind of health risk factor.

■ Figure 5 is a heat map of Pearson correlation between each avoidable death causes and each health risk factors. This figure shows that the population takes adequate fruit has a negative relation with each kinds of avoidable death causes. And population with obese, low-exercise, smoking, risk waist measurement and psychological distress have a positive relation with each kind of avoidable death causes.

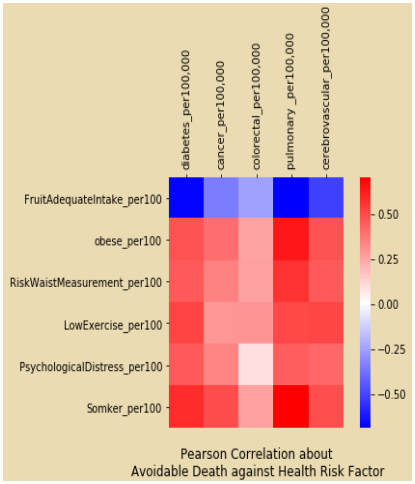

Figure 5

■ Figure 6 shows the proportion of each kind of avoidable death causes. By inspecting figure 5, I would suggest that the local government encourage people to do cancer diagnose regularly, because near 48% avoidable is caused by cancer. Of course, the local government would also encourage people to diagnose the other diseases as well.

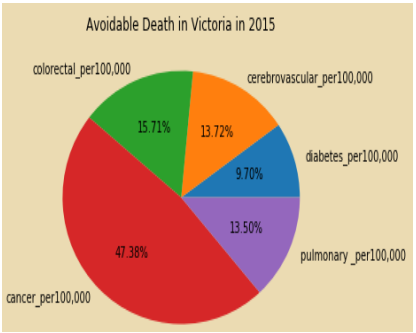

Figure 6

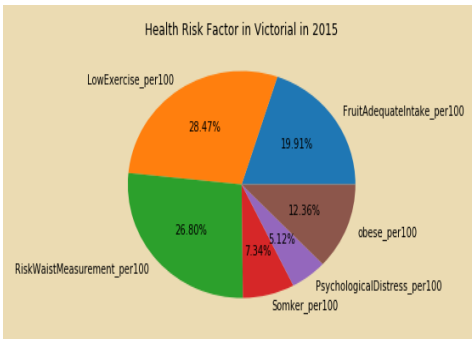

Figure 7

■ Figure 7 illustrates the percentage of each kind of health risk factor. Low-exercise, obese as well as risk waist measurement occupy almost 68% of all health risk factors. From this, I would suggest people and the local government to be more focus on these three health risk factors as well as take effective actions, such as building more public sports equipment or build more gyms.

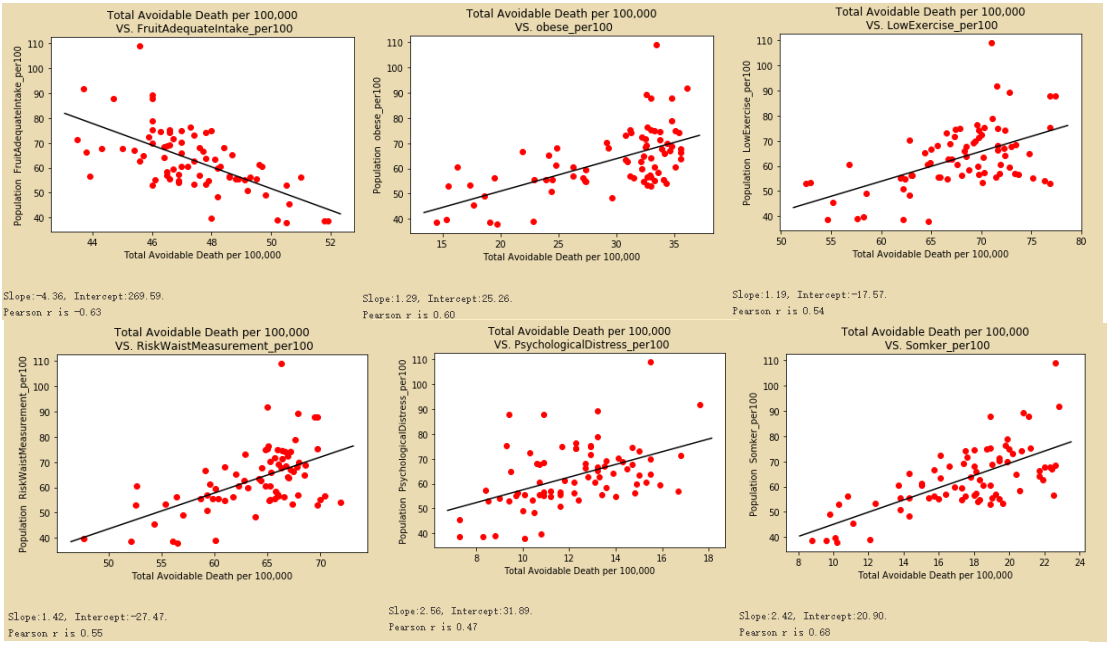

Figure 8

- Figure 8 shows scatter plots of adequate fruit intake, obese, low-exercise, risk waist measurement, psychological distress and smoker per 100 against total avoidable death respectively. The goal is to investigate the relationship between each kind of health risk factor and total avoidable death. In order to reduce avoidable death from these health risk factors. The scatter plots as well as the Pearson correlation both show that the population take adequate fruit, the less total avoidable death will be. As well as, the more population of obese, low-exercise, risk waist measurement, psychological distress and smoker per 100, the more total avoidable death will be. Therefore, I would summarize that in order to prevent avoidable, the local government should encourage people to take enough fruit as well as avoiding obese, low-exercise, risk waist measurement, psychological distress and smoking.

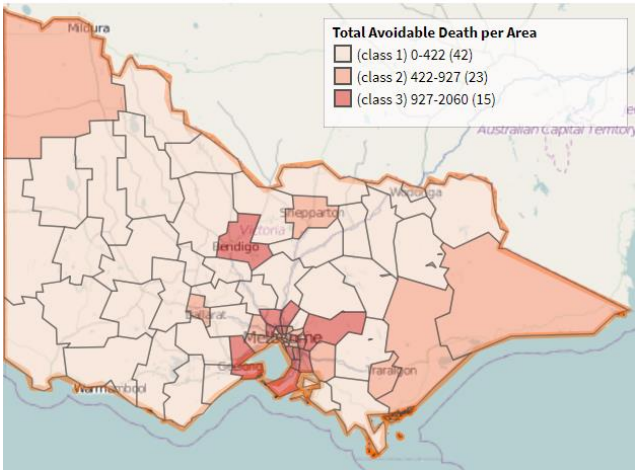

Figure 9

- Figure 9 shows 3 clusters of area by total avoidable death for each area on the Victoria map. By inspecting this map, I would suggest the government to take different action depends which cluster the area belongs to. In addition, I would suggest the local government to take more powerful action for area in class 2 and 3.

8. Value
    - Compare to the raw data, I generate 3 added values which are 'Avoidable_Death_Total', 'Health_Risk_Factor_Total' and 'Ratio'. 'Avoidable_Death_Total' is the sum of each kind of avoidable death for each area. 'Health_Risk_Factor_Total' is the sum of each kind of health risk factor for each area. 'Ratio' is 'Avoidable_Death_Total' against 'Health_Risk_Factor_Total' for each area. Because the raw data is too specific, I want to find overall correlation then I added the first two data. The last added value indicates the influence of health risk factor to avoidable death.
9. Challenge and Reflections
    - The biggest challenge is finding appropriate dataset. During processing, I am struggling with missing values since there are too many. The integration need to sort the datasets, I struggled at first since I do not know how to sort .In visualization, after I draw scatter plot and compute correlation, I found that population of overweight has a negative relationship with total avoidable death. Which means overweight is good. It is not reasonable, so I delete the feature. For the same reason, I drop 'RiskAlcoholConsumption_per100', 'HighBloodPressure_per100' as well.
10. Question Resolution
    - After data processing, I would say we can preventing avoidable death by reducing obese, low-exercise, risk waist measurement, psychological distress and smoking since there are positive correlations as well as encourage adequate fruit intake because of negative correlation. Furthermore, I would suggest the local government and people pay more attention to Northern Grampians (S) since there are many avoidable deaths as well as a high ratio. Also, I would suggest the local government and people to take action such as build more sports equipment or gym as well as increase the consumption of fruit depends the class in Figure 9 (previous page) in order to be effective.
11. Code
    - I write about 350 lines of code for the whole data processing. 130 lines for preprocessing and integrating. And about 220 lines for data visualization and analysis.
    - The major libraries I use are pandas, scipy as well as matplotlib.
    - I used the public code about linear regression in order to find the linear regression line. And some code from this subject's workshop.
    - I use the clustering on map on the website AURIN in order to plot clustering on the map of Victoria.
12. Bibliography
    - The AURIN; Link: https://portal.aurin.org.au/