PHASE 2

Name: Renjie Meng; Student ID: 877396

2018-04-23

1. Title: Identify approaches to reduce avoidable death in Victoria

2. Domain: health

3. 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.

4. 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.

5. 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([' lga_code', ' lga_name', 'RiskAlcoholConsumption_per100',
       ' obese_per100', 'FruitAdequateIntake_per100',
       'RiskWaistMearsurement_per100', 'LowExercise_per100',
       'PsychologicalDistress_per100', 'Somker_per100'],
      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, 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.

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

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

6. 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'

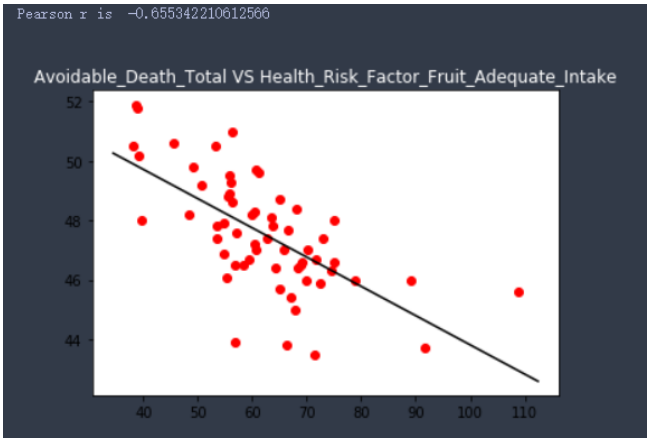

Figure 1

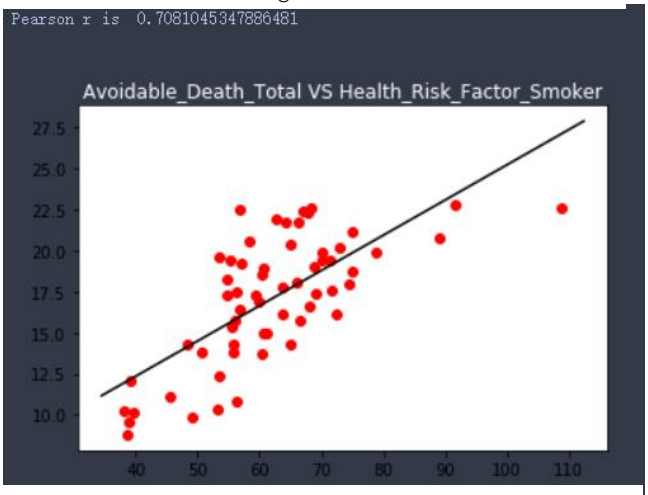

Figure 2

7. 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.

8. 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.