Phase 4

* Introduction
  + Question
    - Good morning, my name is Rudy. Today, I am going to give a presentation in the domain of health about avoidable death and health risk factors. The question I am trying to answer is “Could we prevent avoidable death by reducing health risk factors?”.
  + Structure
    - This is the structure of my presentation today. I will go through them one by one
      * Motivation and outcome
      * Datasets
      * Data Wrangling Methodologies
      * Findings 1-4
      * Challenge
      * Summary
      * Question
* Purpose
  + In 2015, there are 4,800 people died from avoidable death which means they could be prevented. So, I want to find a way to prevent avoidable death. Avoidable death means the death can be avoided in early stage by changing habits, such as diet habits. Therefore, I want to figure out can we prevent avoidable death by reducing health risk factors, such as low-exercise and obese?
* Outcome
  + By analyzing appropriate datasets, my answer to the question is yes. We can prevent avoidable deaths by reducing health risk factors. Now, let me show you how I did it.
* Datasets
  + First of all, the two datasets I used are “LGA\_15\_Avoidable\_Death.csv” and “LGA\_15\_Health\_Risk\_Factors”. There are 2 main reason for I choosing them, first is that these two datasets have the information I want, then these two datasets are the data of same area at period of time which means it is easier to integrate.
* Data Wrangling Methodologies
  + Then, I am going to discuss the methodologies I used in data preprocessing, data visualization and data analyzing respectively.
  + Preprocessing (100)
    - In data preprocessing, in order to get appropriate data from raw datasets, I extract the features I want and transform them into consistent formats. Then, I integrate selected features from two datasets on their common primary key. After data integration, I solved missing values by deleting records with more than 3 features and imputation mean value for remaining missing values. Then, I used boxplot to detect outliers. Finally, I added 2 values, total avoidable death per 100,000 and total population has health risk factor per 100 in each area.
  + Visualization (100)
    - After data preprocessing, I have the ideal data to investigate. For further investigation, I used serval visualization methodologies which is scatter plot, heatmap, pie chart as well as the clustering map. In addition, the cluster map is which I plot use ARUIN by uploading processed data to AURIN.
* Data analyzing (with respect to graph)
  + After data wrangling, I found 4 things.
  + Findings and interests
    - Outlier
      * First of all, by detecting outlier. There is an important finding. Northern Grampians has extreme high number of avoidable death of 108.9 per 100, 000. It means that among every 100 people there are more than 1 people died from avoidable death in Northern Grampians. Therefore, I would suggest the government to pay more attention to this area.
    - Heatmap
      * Then, the heatmap shows the Pearson correlation between each kind of avoidable death and each kind of health risk factors. From the heatmap, we can see that the population take adequate fruit in each area has a negative correlation with each kind of avoidable death. Furthermore, the population of obese, low-exercise, risk waist measurement, psychological distress and smoking has a positive correlation with each kind of avoidable death respectively. Therefore, I guess we can prevent avoidable death by helping population has obese, low-exercise, risk-waist-measurement, psychological distress and smoking habits as well as encourage people to take adequate fruit (at least 2 slices per day).
    - Pie charts
      * The third finding comes from pie charts. The upper side pie chart is about different kinds of avoidable death. From this pie chart, it is clear that almost half of the avoidable death is caused by cancer. Thus, I would suggest the government to encourage citizen go to hospital for diagnosing cancer more regularly.
      * The pie chart in the bottom is about each kind of health risk factor. By inspecting this pie chart, it is clear that 68% population has obese, low-exercise and risk waist measurement. From this, I would suggest the government encourage people to exercise more as well as establishing more sports equipment.
    - Clustering map
      * Finally, this is the clustering map of total avoidable death per 100,000 in Victoria in 2015. There are three classes of total avoidable death per 100,000, which are 38.0 – 61.2 is class1, 61.2 – 78.9 is class and 78.9 -108.9 is class 3. From the map, intuitively we can see some areas are in class 3, I would suggest the government to pay more attention to this area.
* Challenges
  + Data processing is not easy, there are three main challenges I had during data processing
  + Finding datasets
    - The biggest challenge is finding datasets. It is extremely hard to find appropriate dataset at first, but, fortunately, after few days searching I found what I want finally.
  + Bad findings
    - Then, there are some bad findings. One of them are that the population overweight has a negative correlation with each kind of avoidable death. Which means we should encourage to be overweight in order to prevent avoidable death. It is very strange. After struggling, I dropped the feature of overweight.
  + Coding
    - Finally, I met a lot of coding problem during data wrangling, such as how to do visualization. I solved them by searching on the internet.
* summary
  + resolution to the question
    - In conclusion, we could prevent and reducing avoidable death by reducing risk health factor.
    - Furthermore, the government should take different action with respect to which class the area is in.