**Analysis**

**of**

**distinctive cause of deaths**

Yu Lin Shih (CIN 305857869)

Chou I Cheong (CIN 305176968)

Department of Information System

California State University, Los Angeles

**Table of Contents**

[Abstract 1](#_Toc500269921)

[Introduction 1](#_Toc500269922)

[Methodology 2](#_Toc500269923)

[Dataset URL 2](#_Toc500269924)

[Data format and description 2](#_Toc500269925)

[Data Cleaning 3](#_Toc500269926)

[1.Substring 3](#_Toc500269927)

[2. Data Type Conversion 6](#_Toc500269928)

[3. Data Consistency and Completeness Enforcement 7](#_Toc500269929)

[Show/Apply Summary Statistic 10](#_Toc500269930)

[Statistical Averages of Data 10](#_Toc500269931)

[Count, Mean, Standard Deviation, Min & Max 10](#_Toc500269932)

[Leading cause of death summaries (Count, Unique, top and freq) 11](#_Toc500269933)

[Analysis & Visualizations 11](#_Toc500269934)

[Research problems 11](#_Toc500269935)

[1. What is the top five of leading cause of death in California? 12](#_Toc500269936)

[2. What is the interesting findings of top five of leading cause of death in California when associated with age and death number? 14](#_Toc500269937)

[3. What is the relationship of leading cause of death between nationwide and California and why is it an important issue? Based on our findings, what preventative strategies can be suggested to adapt on improving overall health care service? 17](#_Toc500269938)

[Reference 19](#_Toc500269939)

# **Abstract**

This study aims to help California state government to improve public healthcare service and raise public awareness of corresponding disease which caused death in California by using Python to investigate and analysis the relationship of top leading cause of death between nationwide and the leading causes of death in California, and how geographical environment and external factors may contribute to public health challenges across the state that suggest California state government allocate proper resource and adapt suitable measures in appropriate fields. Besides, we examine the efficiency of preventative measures and strategies in terms of health care spending of California and it’s mortality rate of leading cause of death. Eventually, the findings would provide the state government, doctors and researchers great insights to understand the performance of leading cause of death’s prevention strategies they applied, thus, adapting more effective preventive strategy and to improve state’s overall health care service.

# **Introduction**

In 2016, U.S. healthcare spending reaches new peak $10,345 per person and hit $3.35 trillion1, Higher health care expenditure input to health care service, however, the mortality has slightly increased. While a rare increase that was not driven by top 5 distinctive causes of death, but was driven by more people dying from drug overdoses, Alzheimer and suicide1. Top 5 leading cause of death is not an indicator of mortality; each state has their own distinctive cause of leading in terms of different environment. According to Centers for Disease Control and Prevention report2, indicated that the distinctive cause of death in Louisiana is Syphilis instead of top 5 of distinctive causes of death in U.S. Public only perceive top 5 of distinctive causes of death as the indicator of causing death, however, they do not recognize what they probably going to die in their state. Every state has different culture, different geographical environment and various health care challenges. In this project, our object of investigation is California. Through this study, our purpose is to help California state government to adapt appropriate strategies and plans in order to improve overall health service of state and allocate resource in a right place. In Louisiana case, it can be suggested that Louisiana state government adapt measures or allocate resources to raise public awareness of sex protection, prevention is more effective than remedial, because prevention cost is less expensive than remedial cost. The findings also can help doctors and physicians to understand how preventative plan can help people live longer and healthier live. While the public also have a visualized picture to help them understand how they may going to die by which distinctive cause of death and raising their awareness of public health.

# **Methodology**

## **Dataset URL**

1. <https://data.cdc.gov/dataset/Deaths-in-122-U-S-cities-1962-2016-122-Cities-Mort/mr8w-325u>Deaths in 122 U.S. cities - 1962-2016. 122 Cities Mortality Reporting System

2. <https://wonder.cdc.gov/controller/datarequest/D76> Centers for Disease Control and Prevention. “Underlying Cause of Deaths 1999-2015.

## **Data format and description**

Data in this report are based on information from government official record. The total size of datasets is approximately 488 KB and the format of file is CSV. This report contains several datasets from Centers for Disease Control and Prevention, National, underlying 5 leading causes of death 1999 - 2015 and California underlying cause of death 1999 – 2015 these two data sets are the primary data set for investigating the top five of leading causes of death in California, and also examine the relationship between the top five of leading causes of death, age and number of deaths. Last but not least, analyzing the correlation relationship between the top of leading cause of death of nationwide and California in 2015. According to Leveraging Your Financial Intelligence: At the intersection of Money Health and Happiness6 (Douglas Lennick, ‎Roy Geer, ‎Ryan Goulart, 2017) indicated that “nearly 75 percent of all deaths in the United States are attributed to 10 distinctive causes of death, with the top three of distinctive cause of death accounting for over 50 percent of all deaths. Understanding the relationship of leading cause of death in California and national average could be executing specific and preventative measures to improve public healthcare service of California.

Regrading to Los Angeles Times7 indicates that alcohol is responsible for about 88,000 deaths in the U.S. each year, while liver cancer, hypertension, stroke and accidents may be partially related to alcohol use. Then the distinctive cause of death in California is Alcoholic liver disease in 2014, thus, analyzing the distinctive cause of death in California in 2015 can know the issue of alcohol use is improving or getting worst. Therefore, underlying Cause of Deaths 1999-2015 dataset will be using to examine the hypotheses of the distinctive cause of death caused by Alcoholic liver disease in 2014 will remain the same result in 2015.

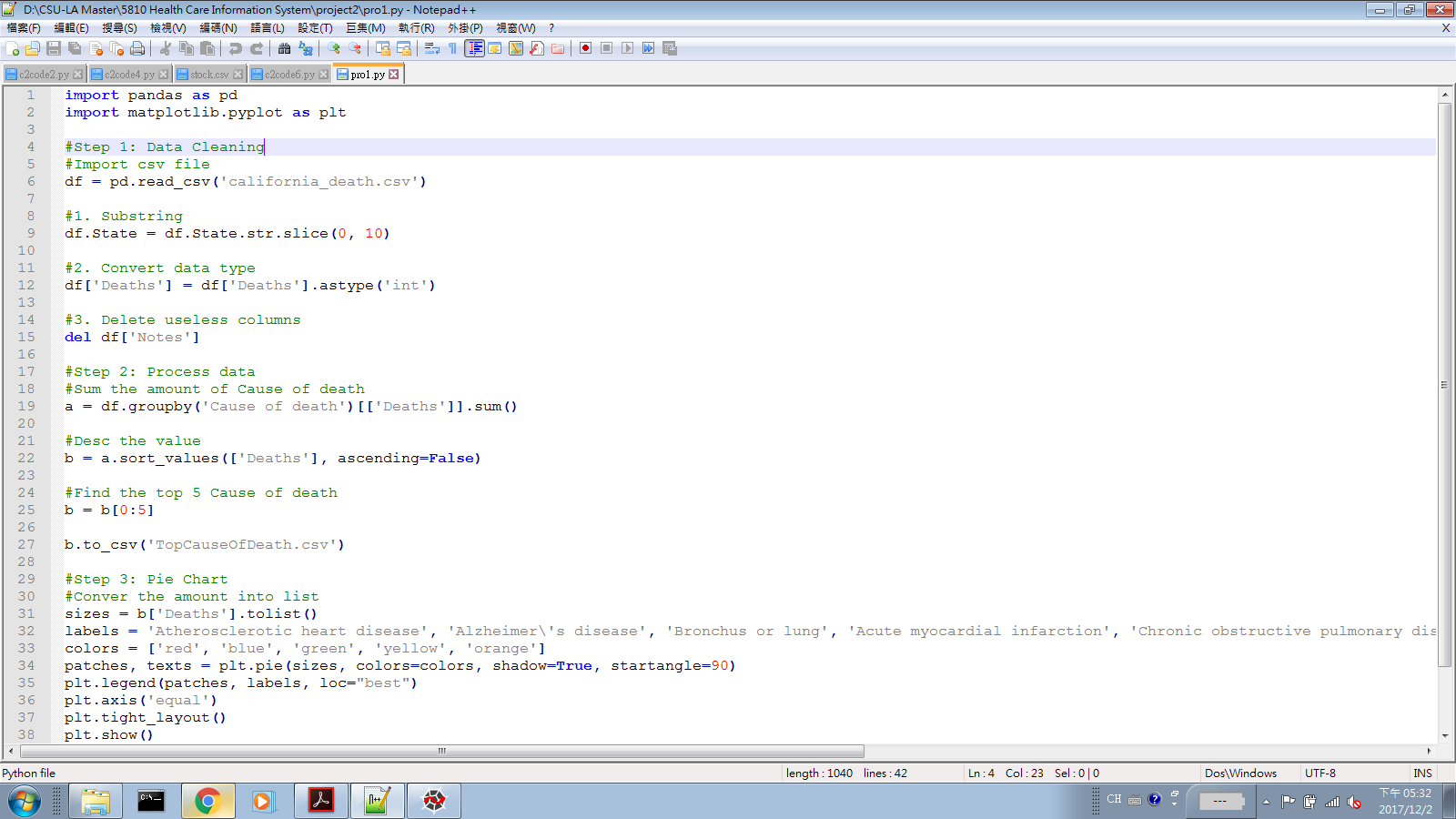
# **Data Cleaning**

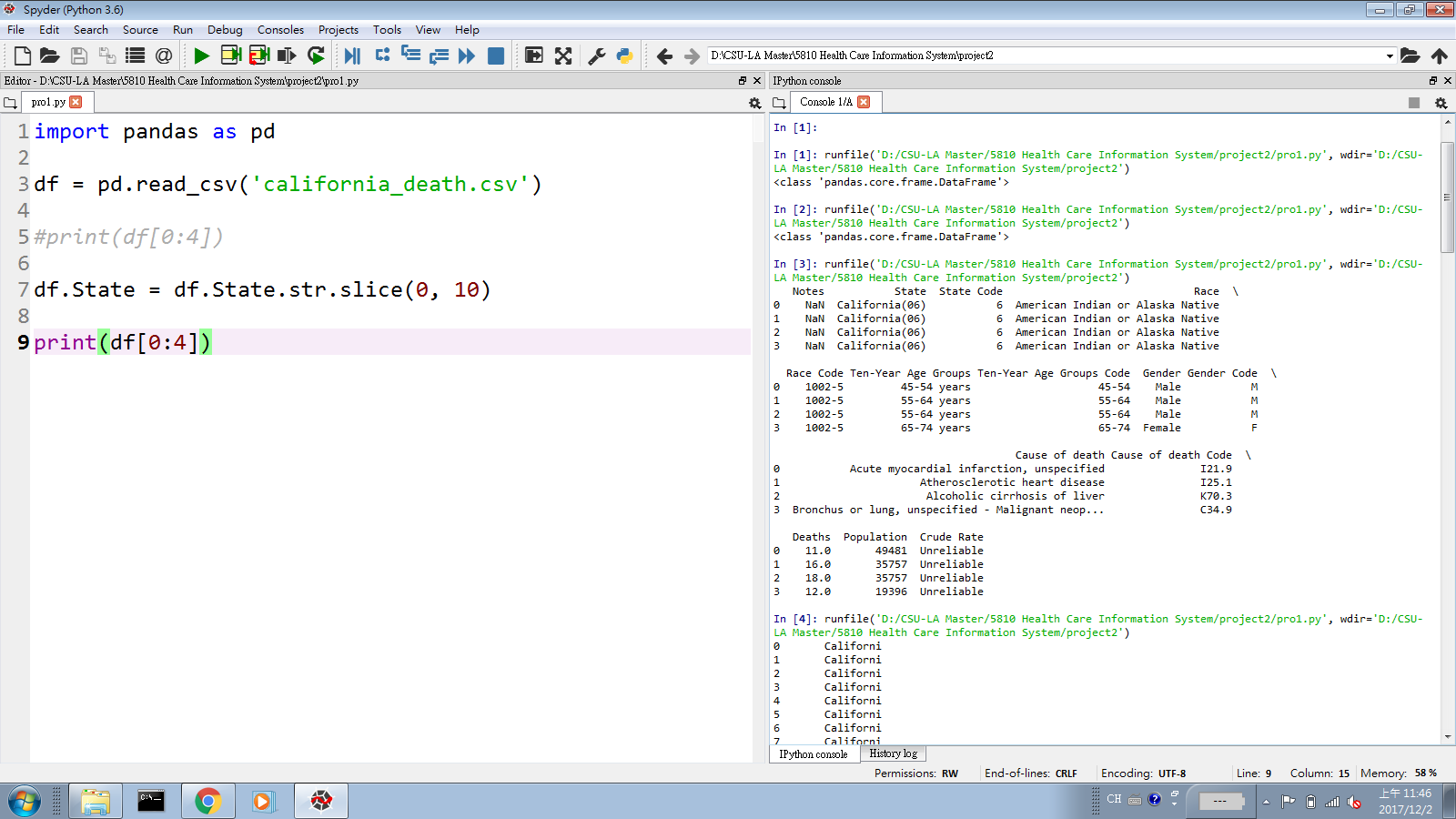
## **1.Substring**

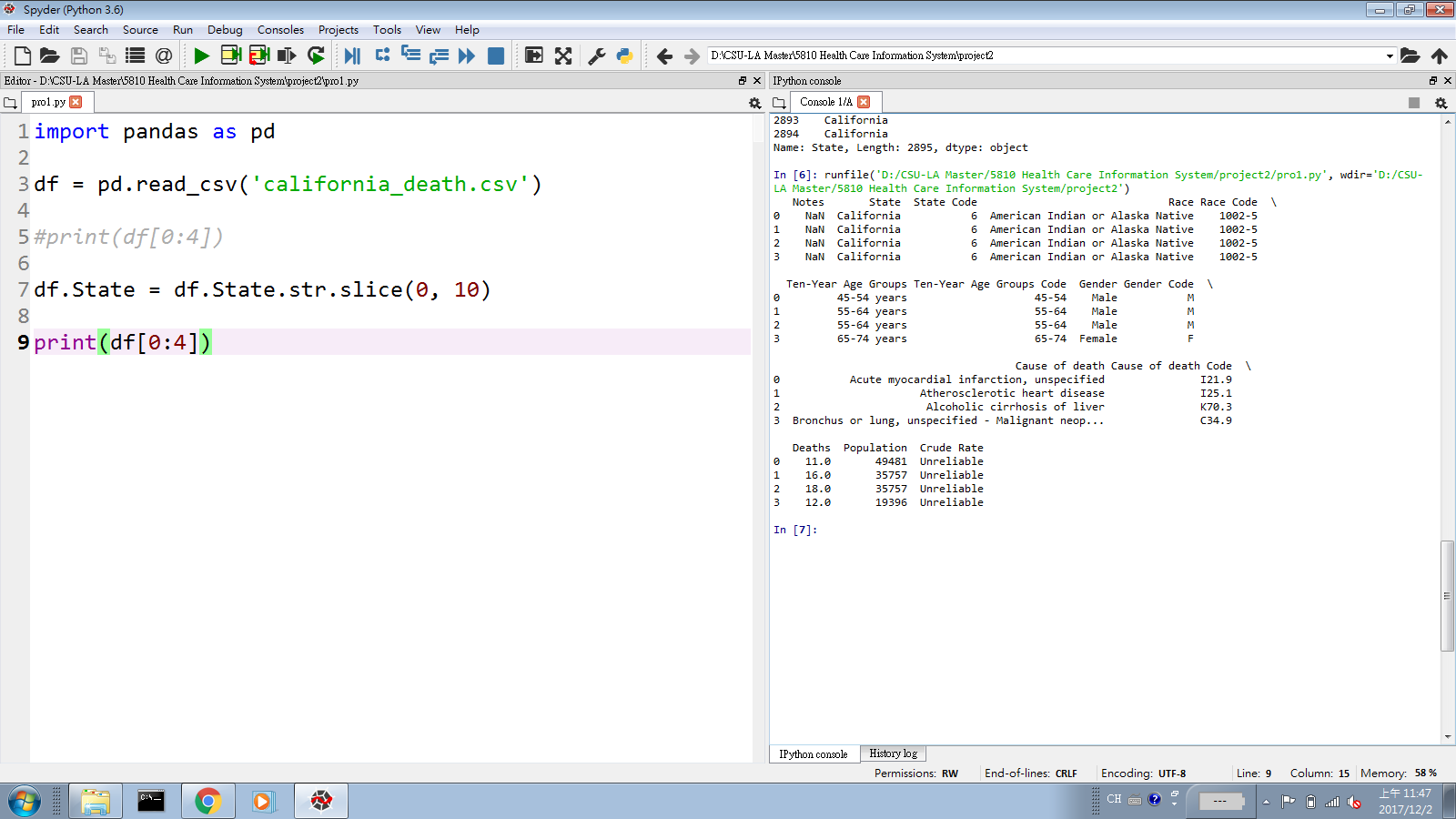
In California death dataset, data owner use California (06) to represent California state, here California and 06 has duplicate meaning, therefore, we use ‘California’ to replace California (06) to simplify data in order to a perfectly clear and unambiguous meaning, besides, to avoid confusing for people who do not know what 06 is stand for.

**Using categories: Pandas Data frame, strings**

Using Pandas Data frame stores CSV file and use strings to remove ‘(06)’ from California(06) to California.

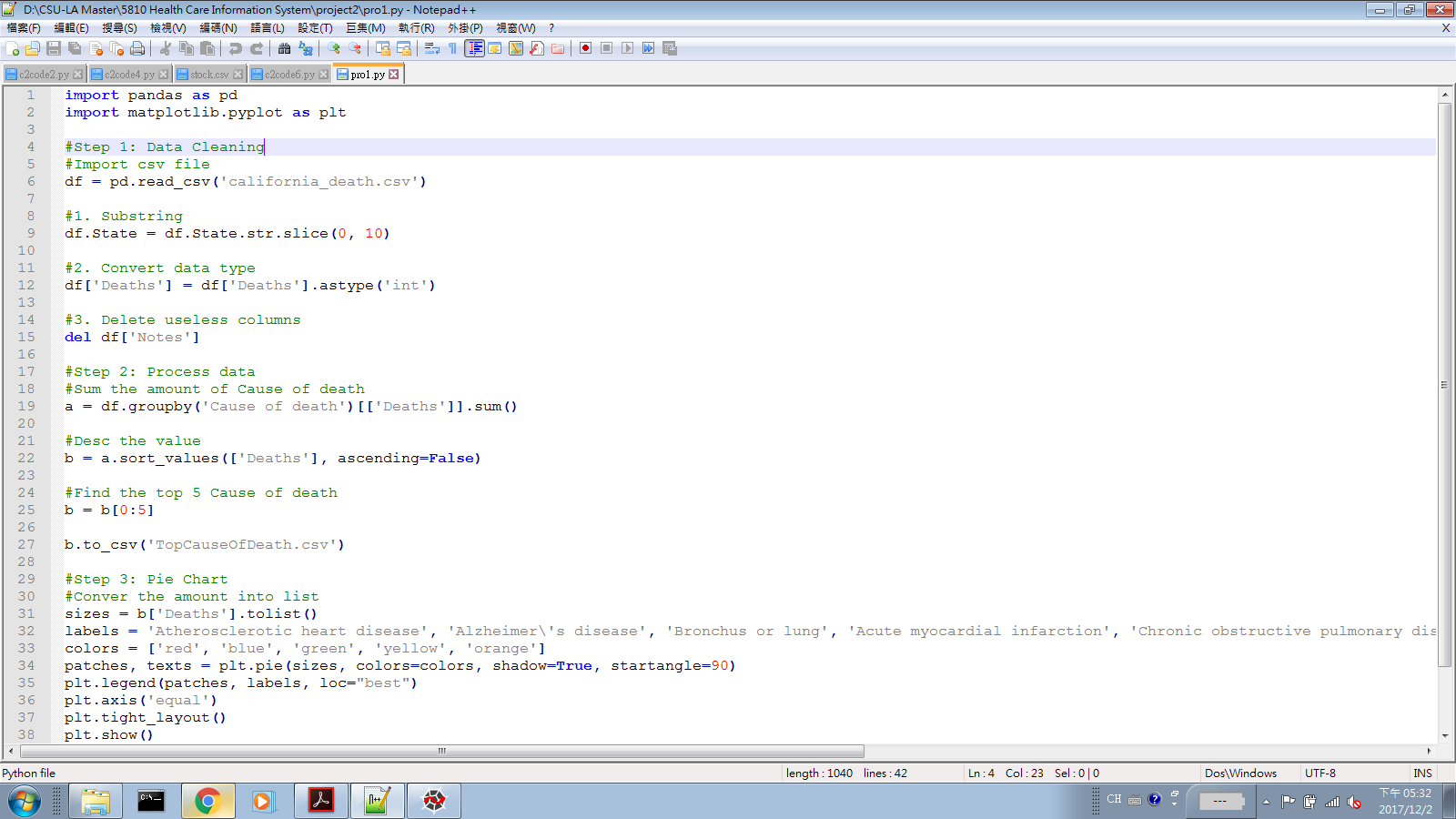
****

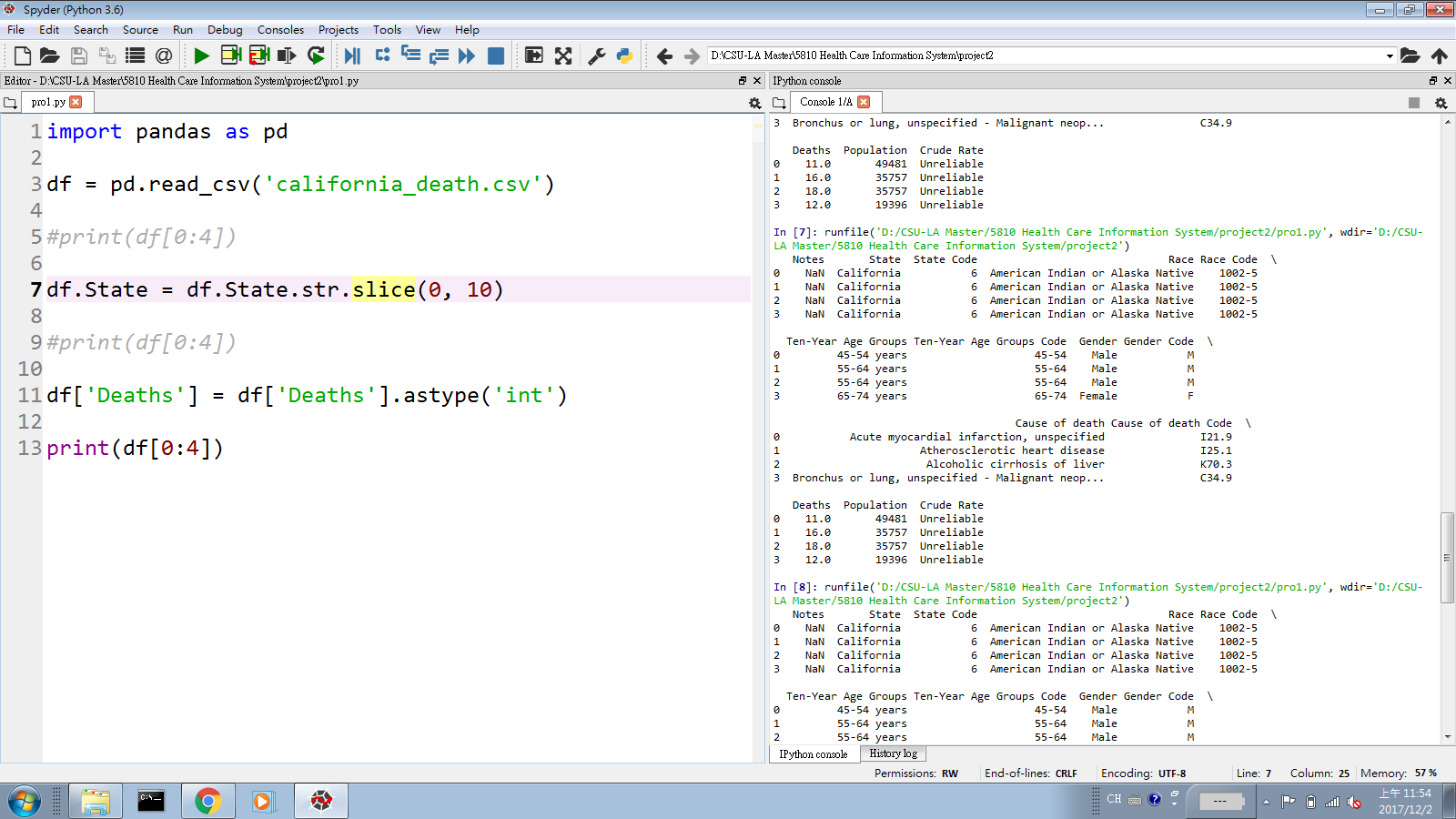
****

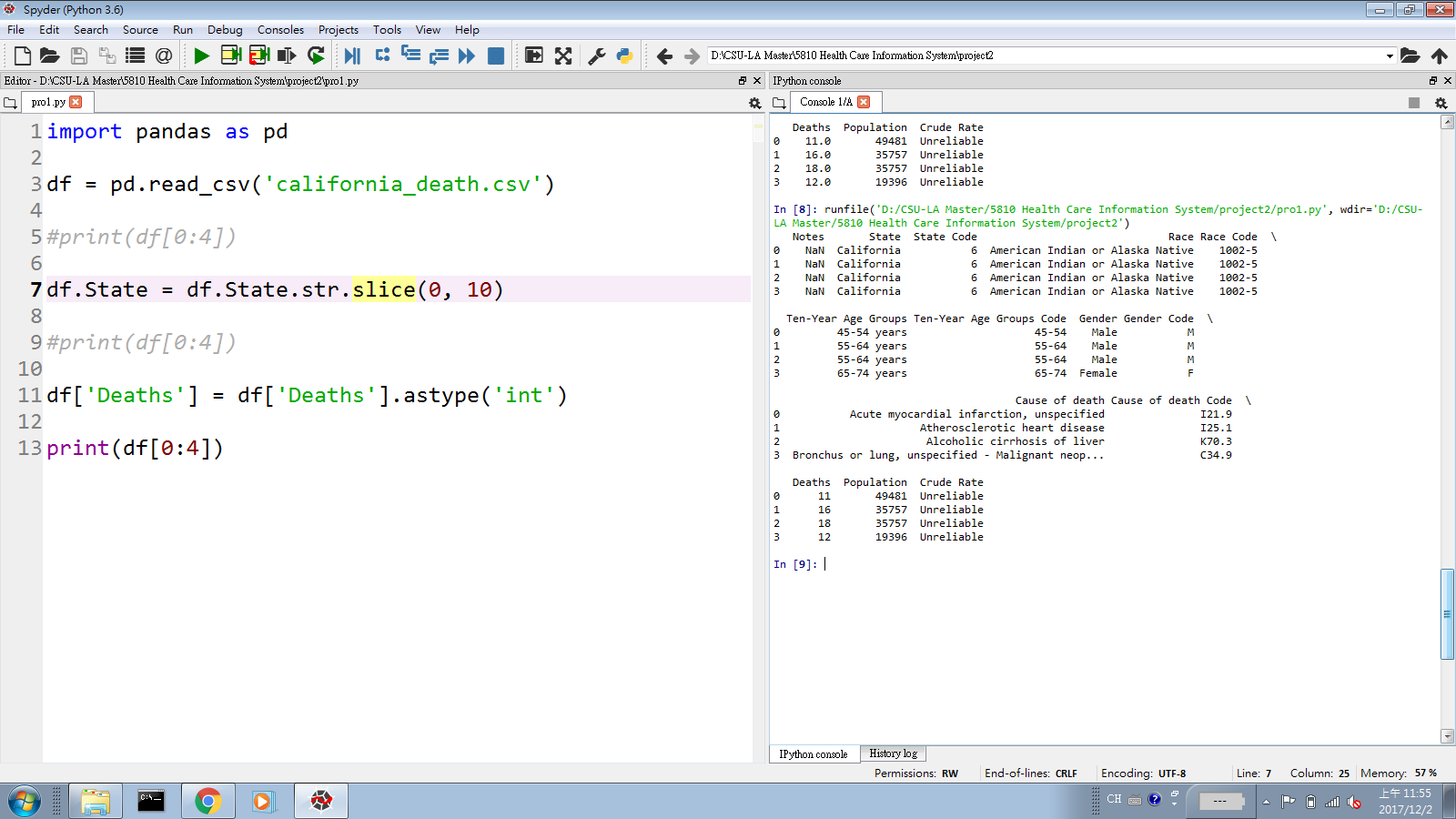
****

## **2. Data Type Conversion**

Since the datatype of death column is float, while death number cannot used float to represent the number of death, it must be shown as integer type, so that we change the datatype of death from float to integer in order to ensure the accuracy of result and avoid the occurrence of error.

****

****

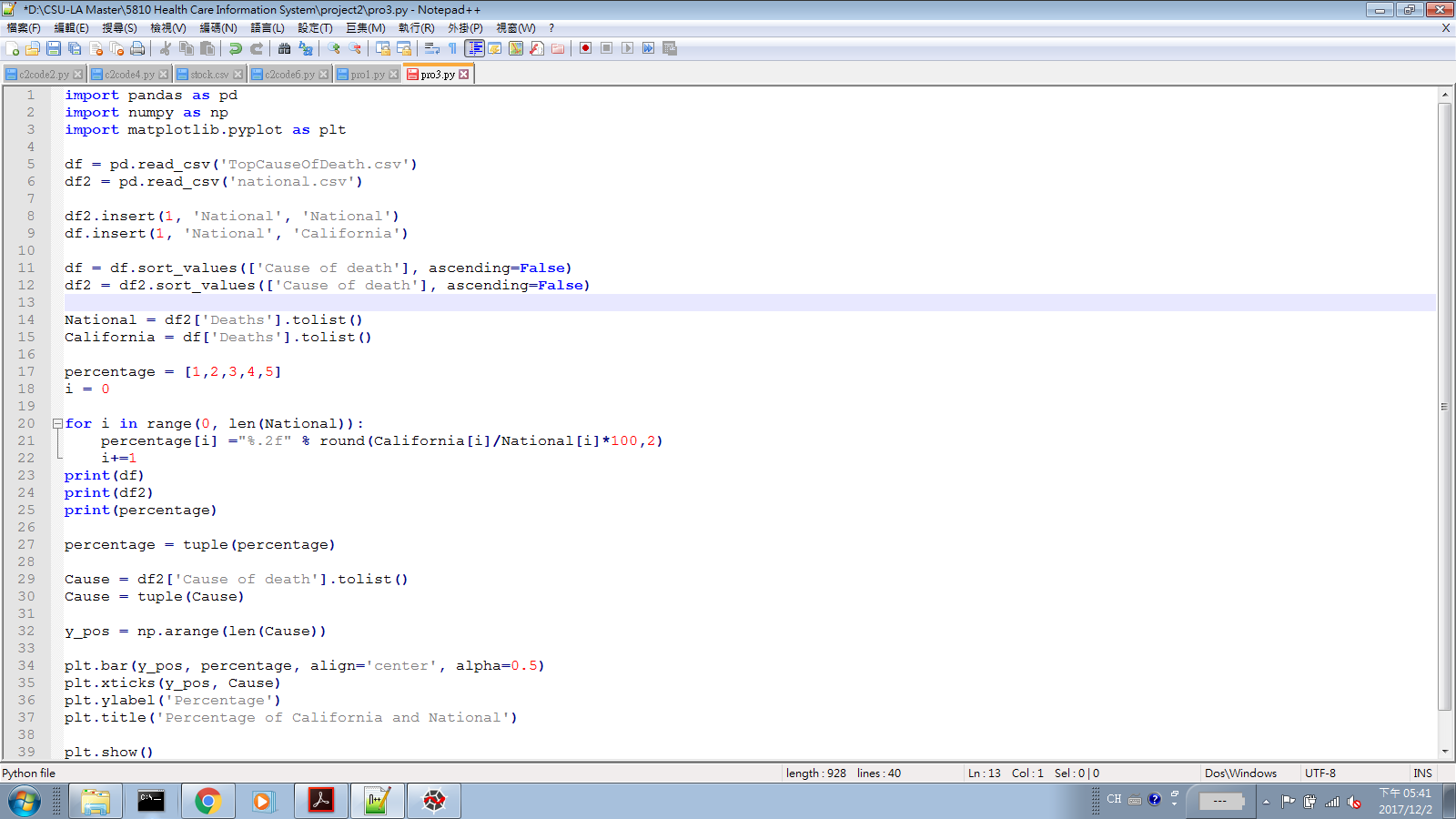
****

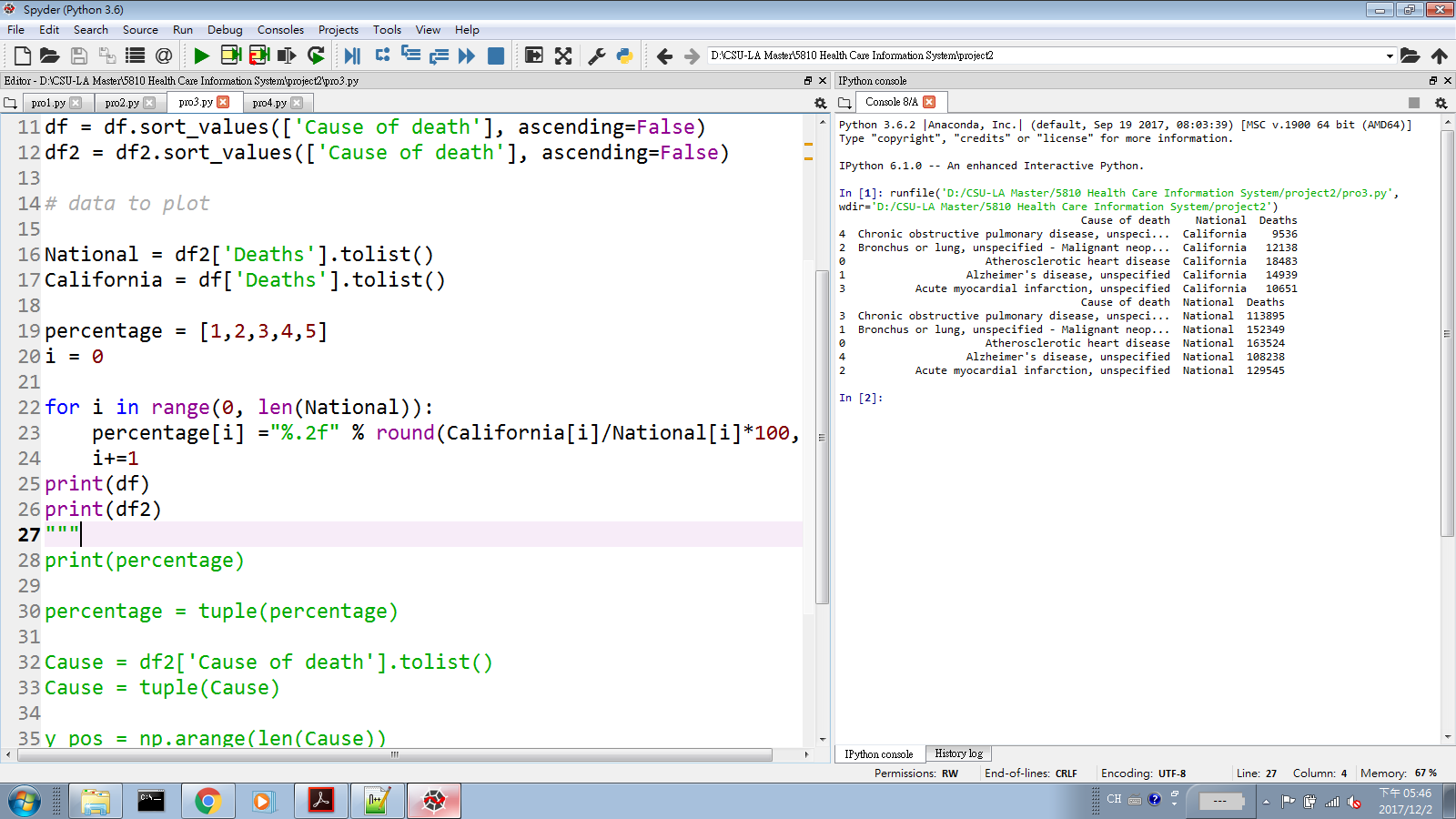
## **3. Data Consistency and Completeness Enforcement**

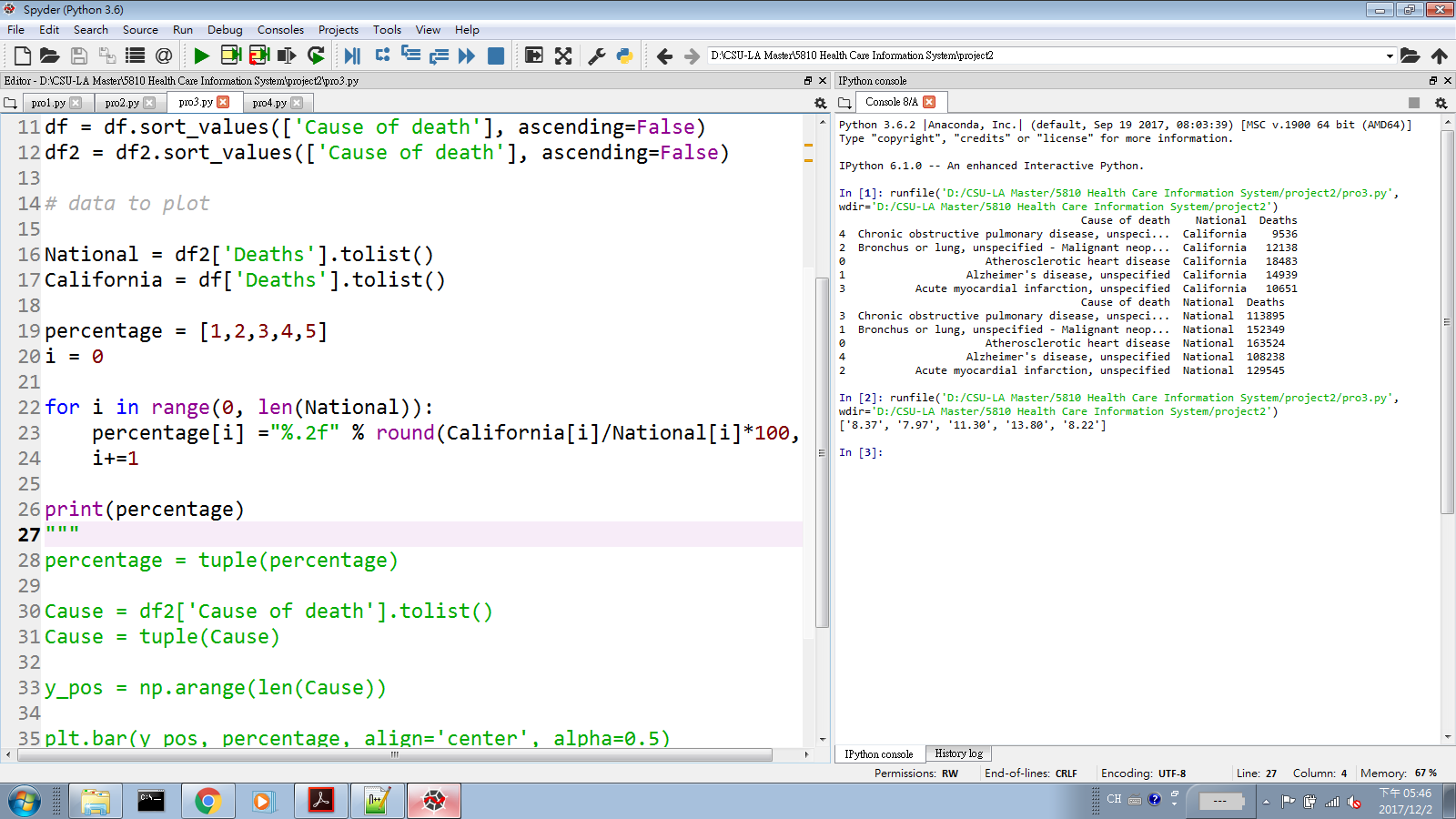
Due to one of our visualizations is to show the percentage of top 5 distinctive causes of death in California and compared with nationwide for the same leading causes of death. Therefore, we join two datasets which are California death and national death in data frame and calculate the percentage of top 5 leading cause of death in California accounted for how many percentage of same 5 leading cause of death of nationwide in order to maintain the consistency and completeness of data.

**Using Data Type: Lists**

Lists we using it here is for extracting the data.

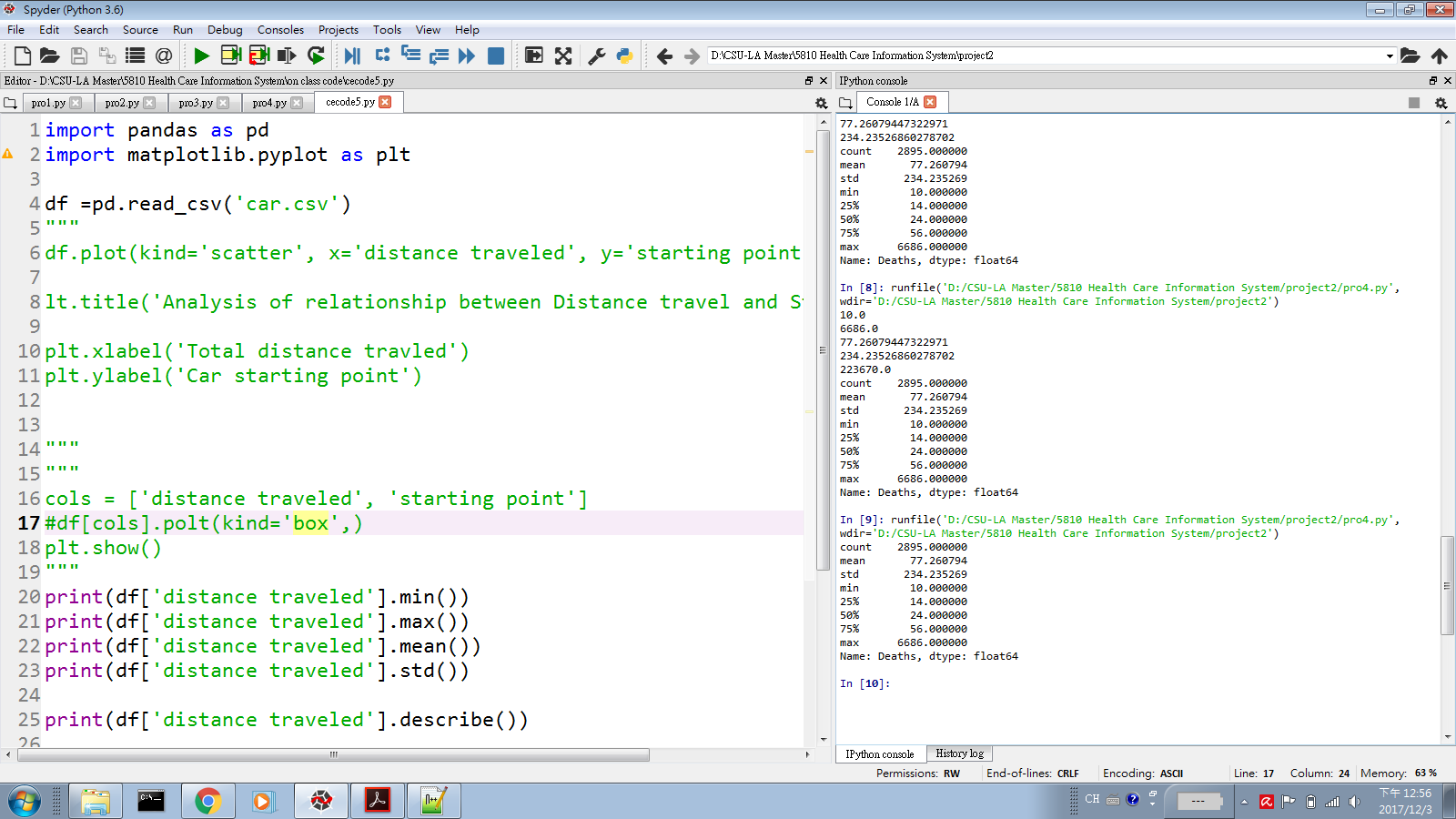
****

****

****

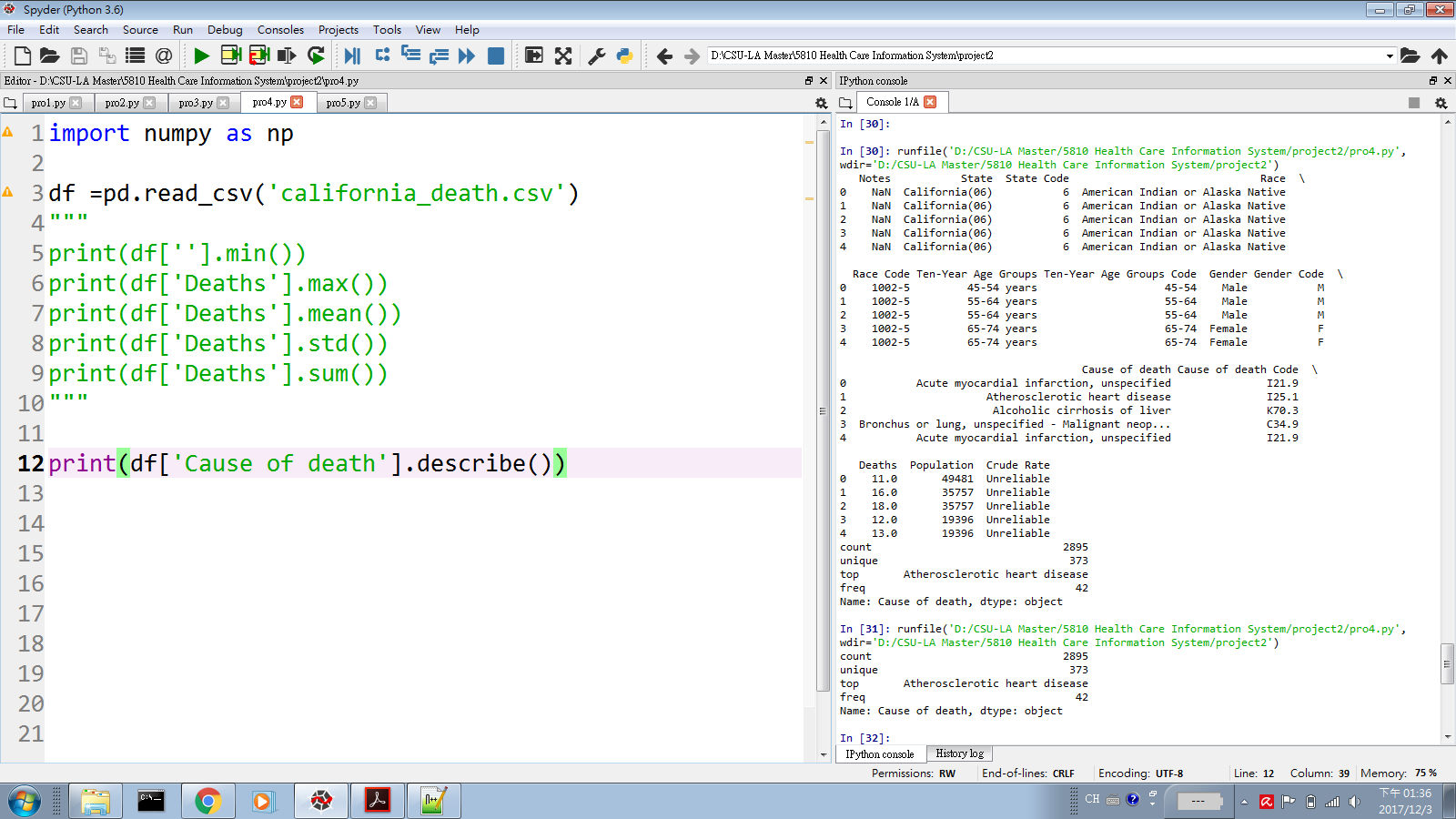
# **Show/Apply Summary Statistic**

## **Statistical Averages of Data**

****

### **Count, Mean, Standard Deviation, Min & Max**

The result from the above statistical averages table, we can obviously see that there are 2,895 records in California death dataset. For Mean, we use death number as our object, the total number of death in California 223,670 divided by 2895 records, the mean for 2,895 records of death amount column is 77.260794 with standard deviation 234.235,269, the deviation also illustrated the mean is 234.24 far from the normal. On the other hands, it depicted that the minimum number of death is 10, while the maximum number of death is 6,686.

****

### **Leading cause of death summaries (Count, Unique, top and freq)**

In this summary, our investigation object is cause of death. In **Count**, which displays that California death dataset has 2895 records, then **Unique** showed that there is 373 categories of leading cause of death makes people died in California. And the **top** also demonstrated Atherosclerotic heart disease is the top 1 leading cause of death in California. While **freq** counts the which is repeated most often in the dataset, and that Atherosclerotic heart disease has the highest number of occurrence, which is 42 times.

# **Analysis & Visualizations**

## **Research problems**

There are three facets of the research problem as below:

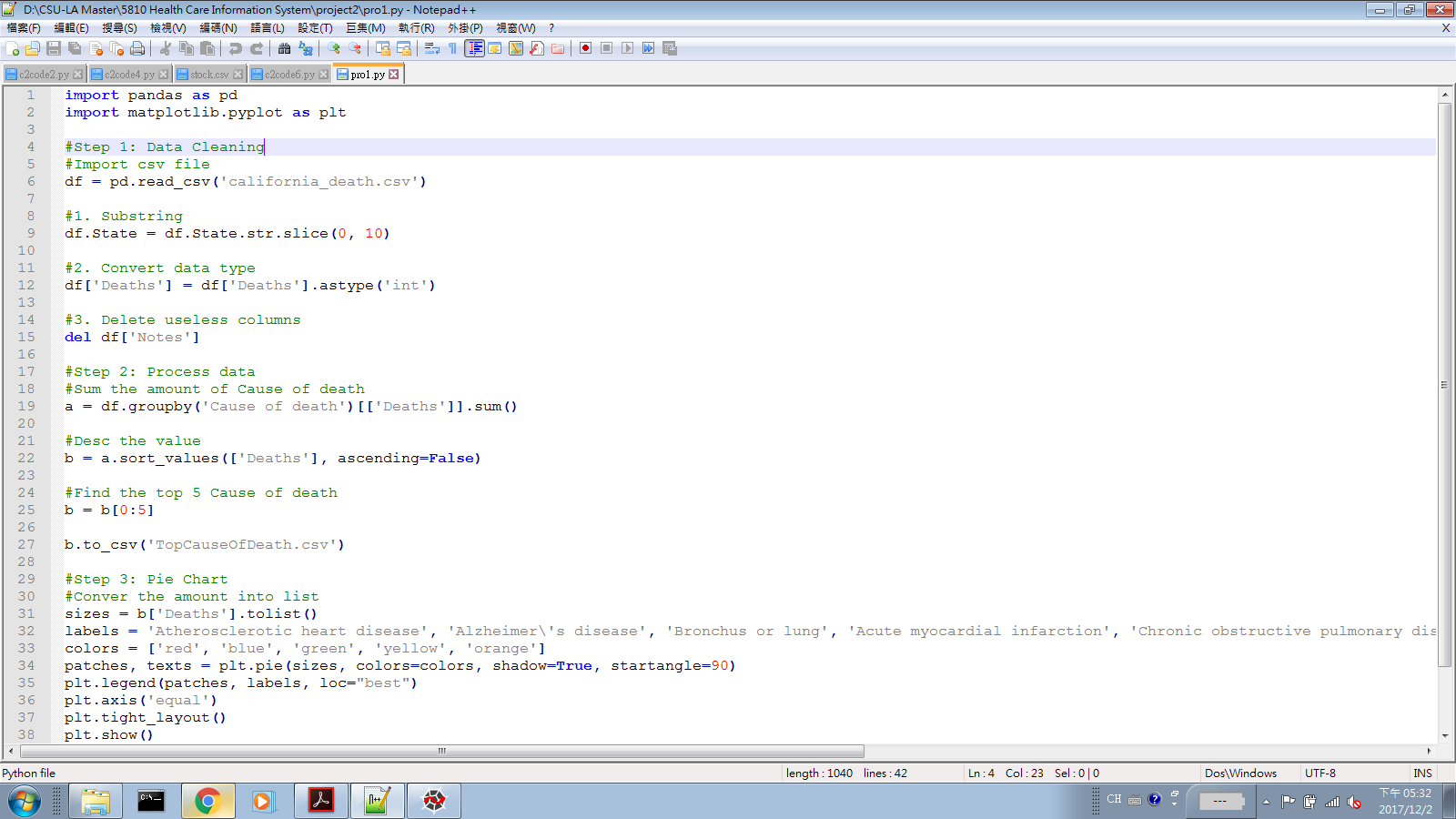
* What is the top five of leading cause of death in California?
* What is the interesting findings of top five of leading cause of death in California when associated with age and death number?
* What is the relationship of leading cause of death between nationwide and California and why is it an important issue? Based on our findings, what preventative strategies can be suggested to adapt on improving overall health care service?

### What is the top five of leading cause of death in California?

According to our finding, the result indicated that the top five of leading causes of death in California are Atherosclerotic heart disease, Bronchus or lung - Malignant neoplasms, Acute myocardial infarction, Alzheimer’s disease and Chronic obstructive pulmonary disease. While Atherosclerotic heart disease is top 1 leading of cause death, which caused 18,483 people died in 2015. The 2nd leading cause of death make 14,939 people died The result showed as the below table 1.

**Using Data Type: Files**

Using files to upload CSV file as data frame.

****

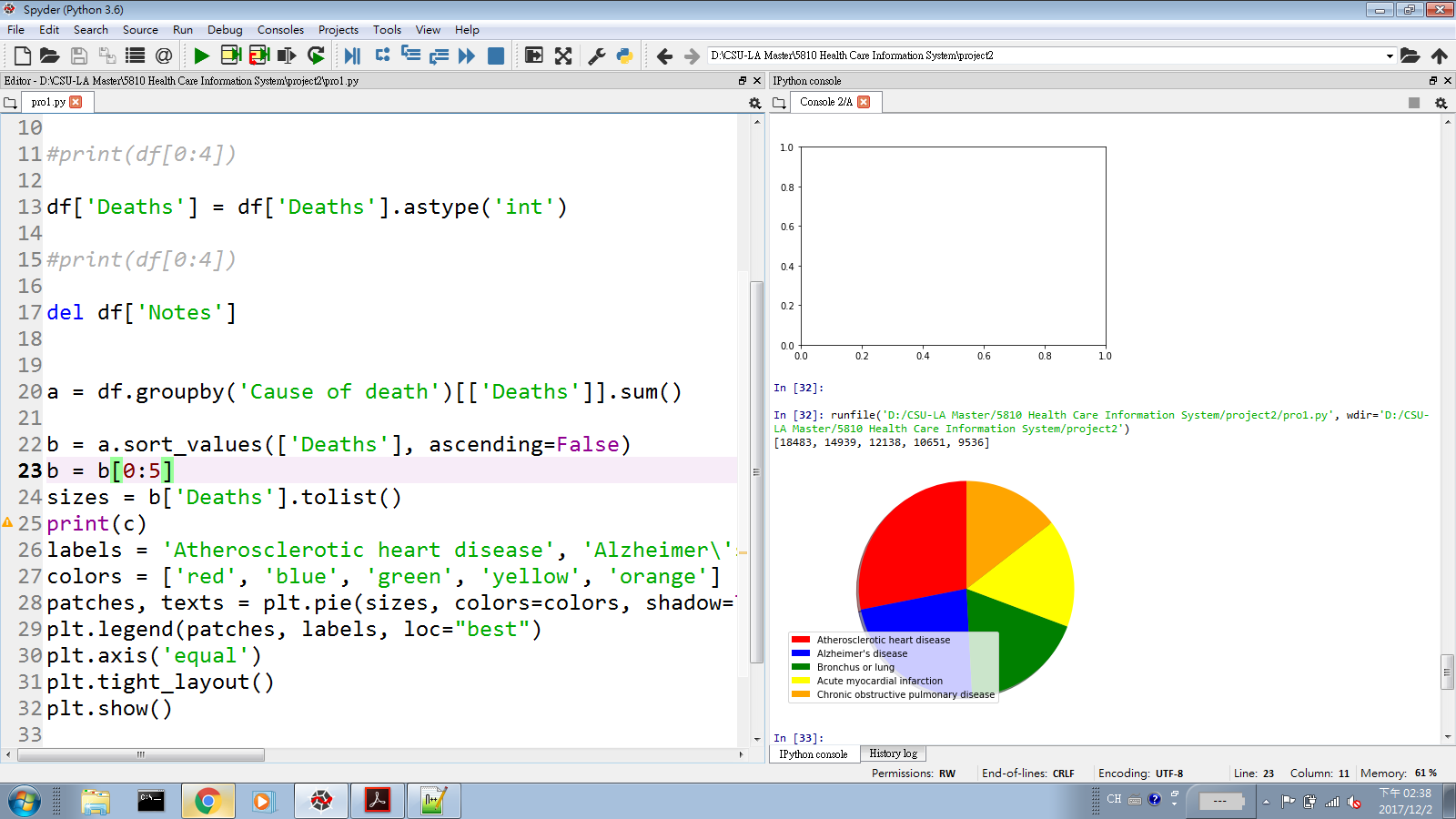
****

Table 1.

### What is the interesting findings of top five of leading cause of death in California when associated with age and death number?

From the below table 2, it can obviously see that the death of 5 leading causes of death remained fairly unchanged in 25 years to 54 years. But the death of Atherosclerotic heart disease has significantly increased from 65-74 years to 85+ years. While the death of Bronchus or lung - Malignant neoplasms and Chronic obstructive pulmonary disease has slightly rise between 65-74 years, 75 to 84 years and 85+ years. Besides, the age group between 75-84 years and 85+ years saw a dramatic growth in deaths of Alzheimer’s disease, moreover, Alzheimer’s disease has higher deaths in 85+ years than other leading causes of death among top 5 leading causes of death, it is approximately 10,778 deaths. The interesting finding is other leading causes of death experienced a different degree of increase from 65-74 years to 85+ years, whereas the deaths of Acute myocardial infarction has a slight decrease from 65-74 years to 85+ years.

**Using Data Type: Functions**

Functions help us to calculate the number of death in each age group.

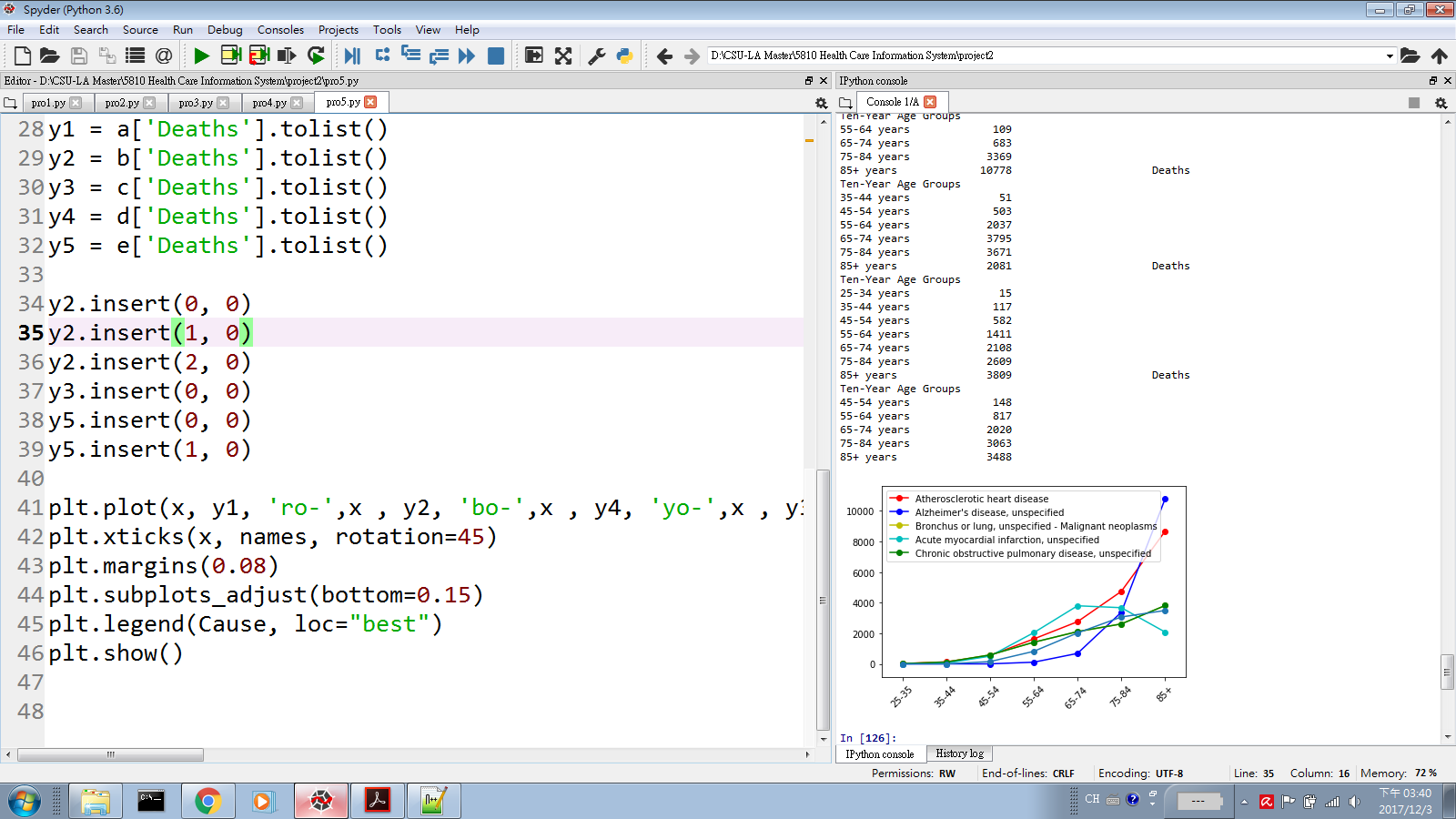
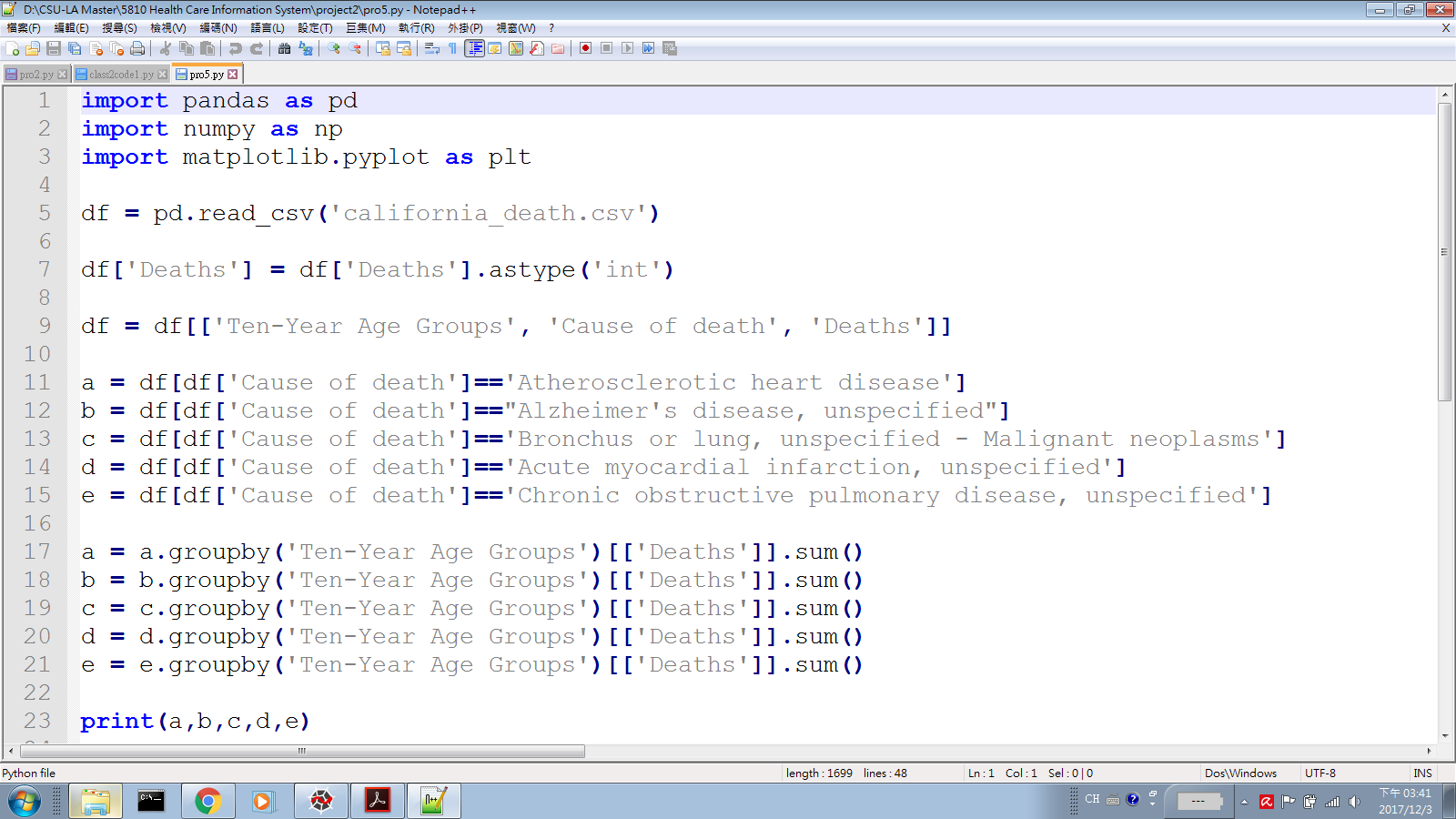
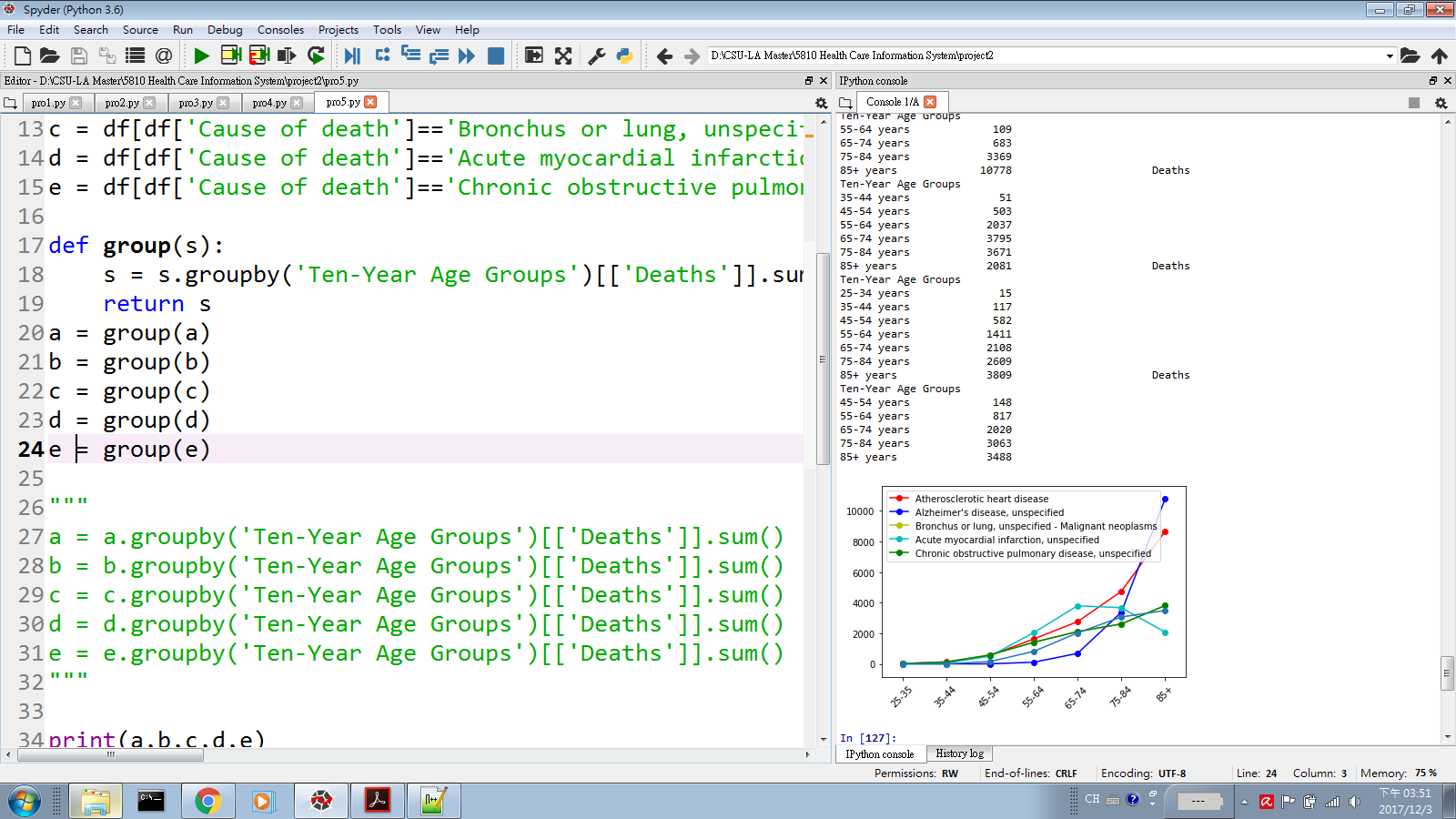
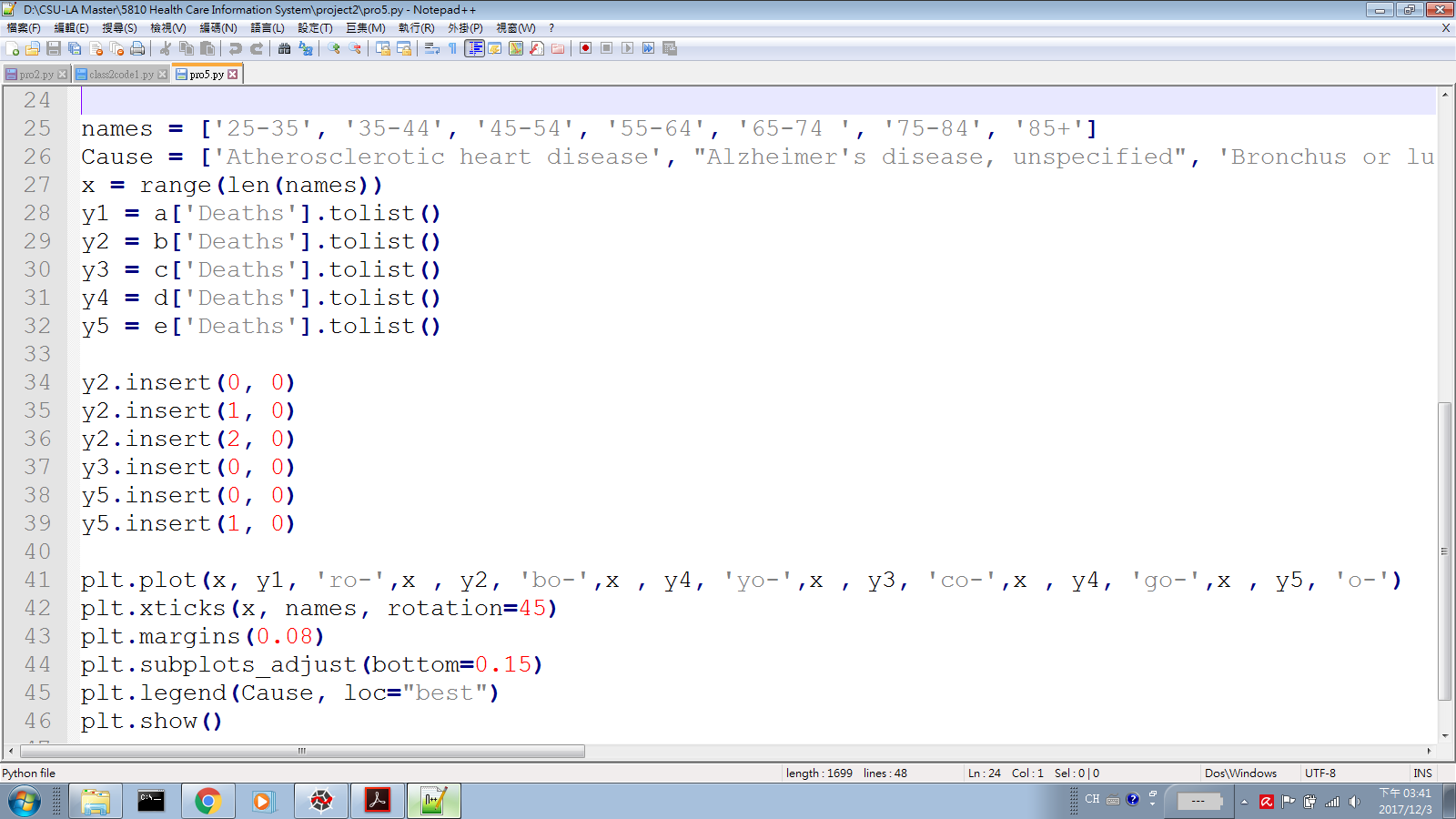
****

Table 2.

****

****

****

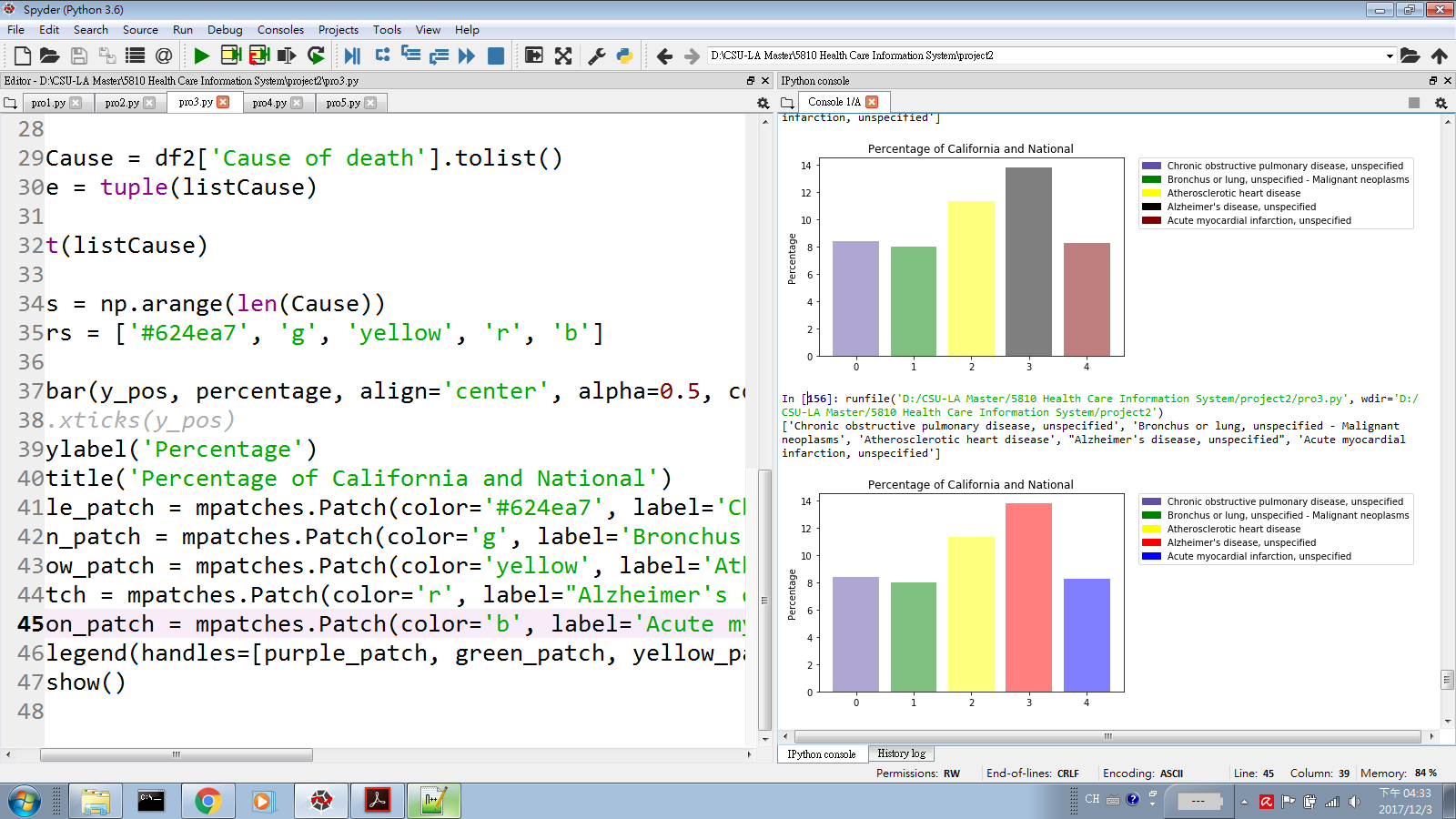
### What is the relationship of leading cause of death between nationwide and California and why is it an important issue? Based on our findings, what preventative strategies can be suggested to adapt on improving overall health care service?

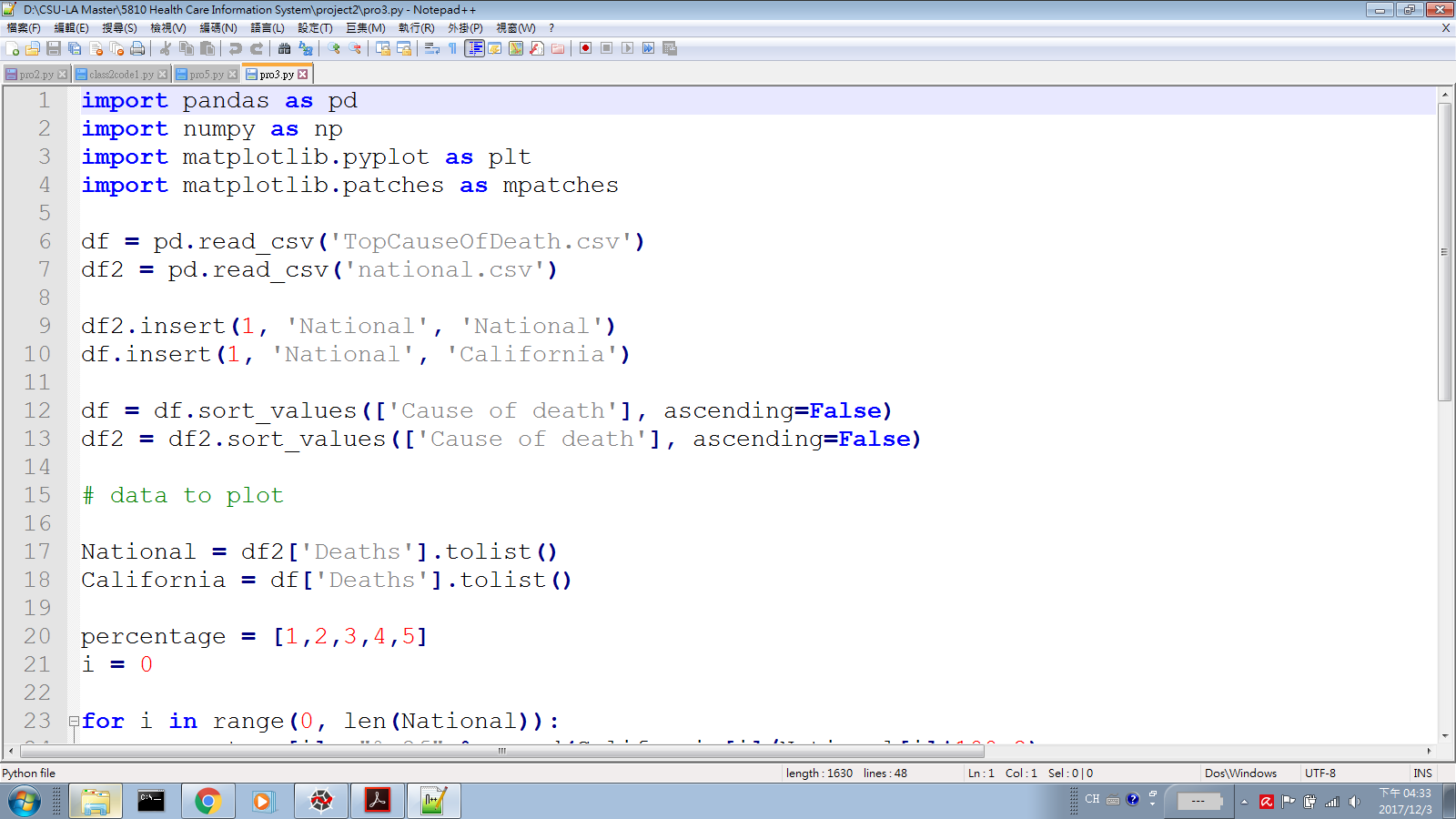
A look at percentage of California and compared with nationwide for the same leading cause of death will reveal some interesting observations. The result illustrates that the top 1 of leading cause of death (Atherosclerotic heart disease) in California accounted for 11.3% of the total death of nationwide Atherosclerotic heart disease. On the other hands, the Alzheimer’s disease which is top 2 of leading cause of death in California, accounted for 13.8 % of the total death of nationwide Alzheimer’s disease, but it contributes 13.8% death to the total death of Alzheimer’s disease in U.S which statistic is higher than Atherosclerotic heart disease.

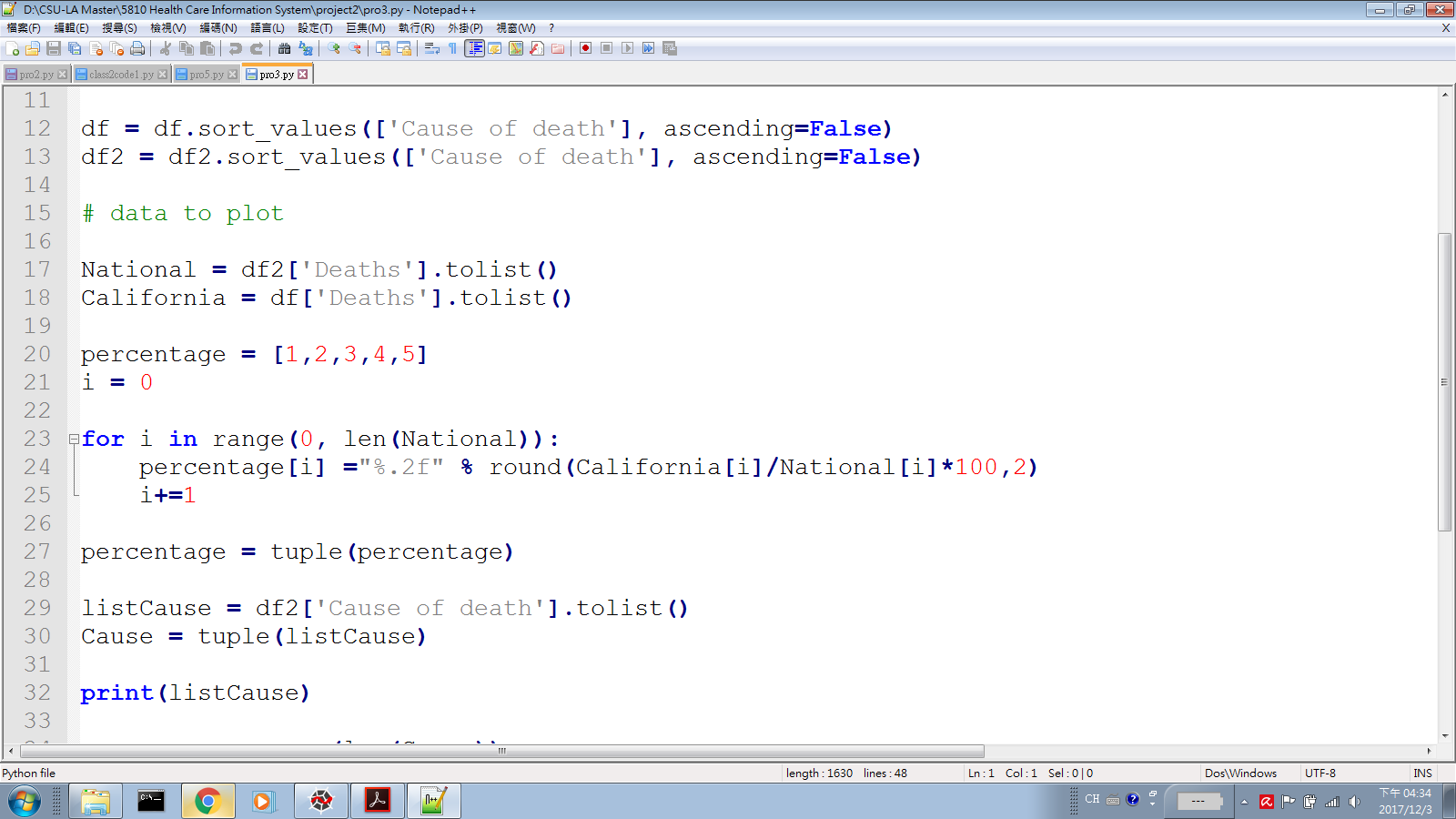
Heart Disease is the top leading cause of death in U.S and worldwide, as well as the top leading cause of death in California. And Alzheimer’s disease also is top 6 of leading causes of death in U.S, but California contributes approximately 14% death to the total death of Alzheimer. Therefore, California state government ought to put resources or hold events to help public understand how they may going to die by which distinctive cause of death and raising their awareness of public health, thus they can prevent or reduce the risk of these disease through change bad habits and lifestyles. Doctors and physicians also can suggest patient to eat a diet with lower salt, sugar and total fat, but higher in fruits and vegetables. Avoiding excessive intake of alcohol, quit smoking and workout regularly.

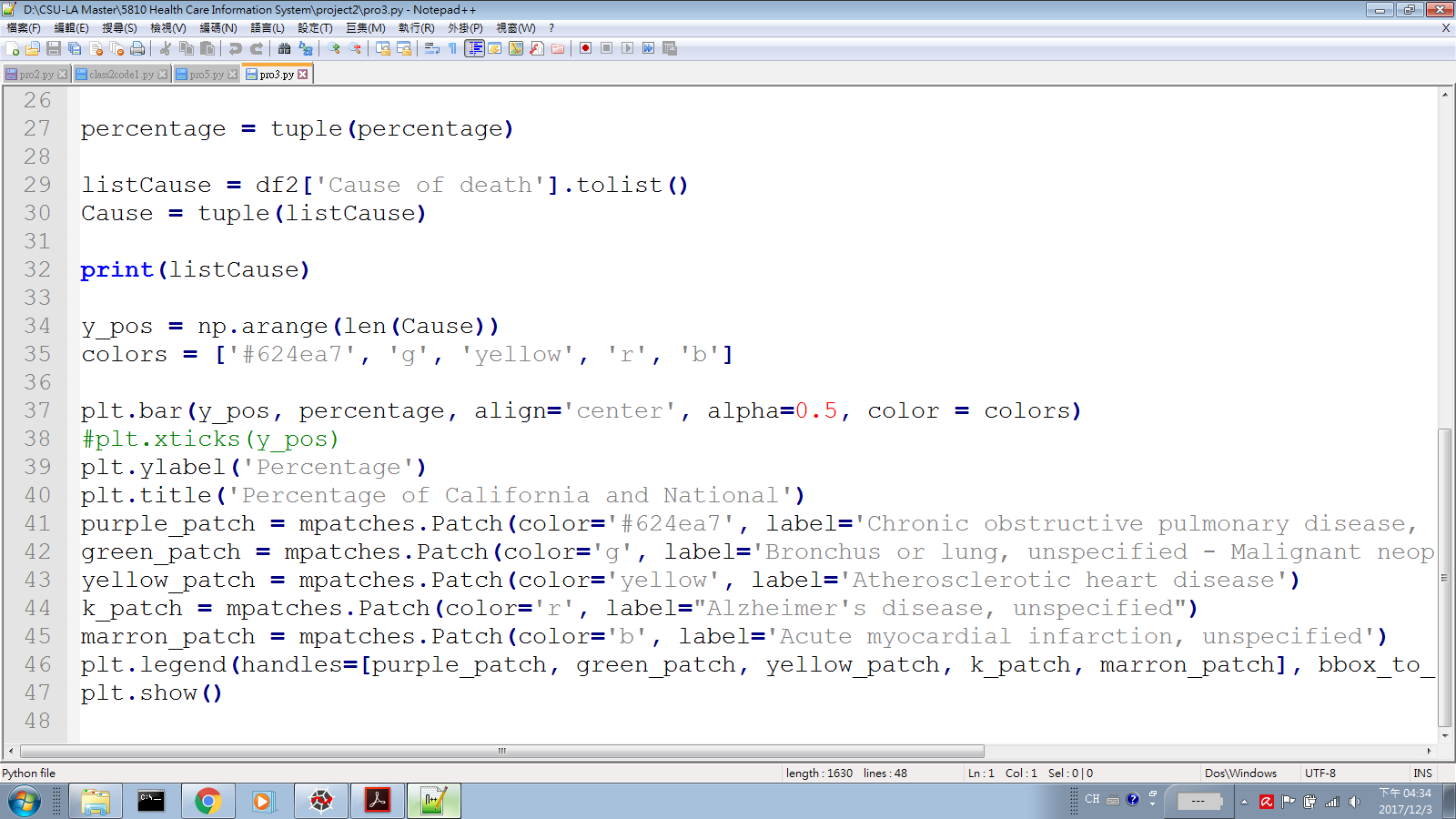
**Using Data Type: Tuple**

We convert list to tuple and create the chart with tuple, then, set tuple as y-axis of bar chart.

****

****

****

****

# **Reference**

1. Sabrina Tavernise. “First Rise in U.S. Death Rate in Years Surprises Experts”, 01 June, 2016

<https://www.nytimes.com/2016/06/01/health/american-death-rate-rises-for-first-time-in-a-decade.html> Accessed 09 November, 2017

2. Alexa Lyons. From Louisiana? Then You’re Probably Going To Die From Syphilis, Sorry, 21 May, 2015 <https://coed.com/2015/05/21/death-map-united-states-disease-details/> Accessed 09 November, 2017

3.Centers for Disease Control and Prevention. “Underlying Cause of Deaths 1999-2015. “<https://wonder.cdc.gov/controller/datarequest/D76> Accessed 09 November 2017

4. Centers for Disease Control and Prevention. “Deaths in 122 U.S. cities – 1962 -2016. 122 Cities Mortality Reporting System” <https://data.cdc.gov/dataset/Deaths-in-122-U-S-cities-1962-2016-122-Cities-Mort/mr8w-325u>Accessed 09 November, 2017

5. Centers for Medicare & Medicaid Services. “National Health Accounts State Health Accounts Residence” , 15 June, 2017 [https://www.cms.gov/Research-Statistics-Data-and Systems/Statistics-Trends-and Reports/NationalHealthExpendData/NationalHealthAccountsStateHealthAccountsResidence.html](https://www.cms.gov/Research-Statistics-Data-and%20%20Systems/Statistics-Trends-and%20Reports/NationalHealthExpendData/NationalHealthAccountsStateHealthAccountsResidence.html) Accessed 09 November 2017

6. Douglas Lennick, ‎Roy Geer, ‎Ryan Goulart. Leveraging Your Financial Intelligence: At the intersection of Money Health and Happiness, 2017

<https://books.google.com/books?id=4n45DwAAQBAJ&pg=PA91&lpg=PA91&dq=Nearly+75+percent+of+all+deaths+in+the+United+States+are+attributed+to+10+leading+causes+of+death&source=bl&ots=hJBHGpgysY&sig=mDkFf5abJ2COOc_43xNwFsRAAq0&hl=en&sa=X&ved=0ahUKEwiwpc2JgLnXAhVDJiYKHboACOIQ6AEIVjAH#v=onepage&q=Nearly%2075%20percent%20of%20all%20deaths%20in%20the%20United%20States%20are%20attributed%20to%2010%20leading%20causes%20of%20death&f=false>

7. Karen Kaplan. “About 88,000 U.S. deaths each year traced to alcohol use, study says”, 13 March, 2014 <http://www.latimes.com/science/sciencenow/la-sci-sn-alcohol-related-deaths-years-lost-sxsw-20140313-story.html> Accessed 09 November, 2017