CAUSES OF DEATH IN AUSTRALIA

FIT5147 Data Exploration Project

Abstract

This report using a range of visualization tools to analyze the fact and the trend of the causes of death in Australia across different states, genders by time

Jiahao Liu

jliu0044@student.monash.edu

Table of Contents

1. Introduction

1.1 Description

Australia, as a developed country with the forth and third highest life expectancy of males (79) and females (84) in the world, (Welfare, 2011) is also regarded as the tenth healthiest countries in 2016, just 4 points behind the first place Iceland. (Aubusson, 2016).

However, as the explosive growth of the population and the explosive growth in medical services needed, the growth rate of Australia's hospitals and hospital staff like doctors and nurses has completely lagged behind the growth rate of demand. (Duckett, 2016)And such situation will cause the following three problems:

a. The working intensity of the hospital will greatly increase. (HealthWorkforce, 2012)
b. The quality of medical services will fall as a result.
c. There will not enough medical services for remote area.

To find a way to prevent it is of great significance for Australia.

1.2 Motivation

In order to provide better healthcare services, the analysis of the causes of death not only allows the medical school to develop students in a targeted manner, but also allows the doctors to have a promising and clear research direction, which leads to a greatly improvement of the efficiency of medical services. It has far-reaching benefits not only for the government but also for citizens.

1.3 Question

a. What are the most common causes of death in the past decade in Australia across states, gender?
b. From 2006 to 2016, total deaths was gradually increased, why?
c. In the past decade, which causes of death have been controlled and which are more and more dominant, why?

2. Data Wrangling

2.1 Data Sources

There are two sets of data been used in these analysis:

1. 'Causes of death by year of occurrence(Australia)' by Australian Bureau of Statistics Retrieved from Website
This excel file contains 10 sheets, different regions stored in different sheets (extra 2 sheets are cover and Totality), All the causes of death encoded by ICD-10, which is a

medical classification by the World Health Organization(WHO). (Wikipeadia, 2014)
The number of people died in specific cause classified by year and gender.

2. 'Estimated Resident Population, States and Territories(Number)' by Australian Bureau of Statistics Retrieved from: Website
This excel file contains the population of Australia across different regions from 1981 to 2017, updated quarterly.

2.2 Steps of Data Wrangling

I have used python and tableau to wrangle the data:

Python：

Environment: Python 3.6.3. final. 0 and 4.5.0
Libraries used:

- Pandas (for data frame, included in Anaconda Python 3.6)
- Re (for regular expression, included in Anaconda Python 3.6)

Data Set 1 :

Step 1. Import data into data frame:
As different regions data stored in different sheets, I parsed the data sheet by sheet and then concatenate them together.

```
In [2]: excel_data = pd.ExcelFile('3303_13 causes of death by year of occurrence (australia).xls')
        excel_data.sheet_names

Out[2]: ['Contents',
         'Table 13.1',
         'Table 13.2',
         'Table 13.3',
         'Table 13.4',
         'Table 13.5',
         'Table 13.6',
         'Table 13.7',
         'Table 13.8',
         'Table 13.9']
```

Step 2. Drop useless columns and rows:
There were several useless data in this file:

- Unformatted header: For the excel file is for human read, there were lots of Nans and errors when we directly parse it into data frame.
- Derived columns: Such as the sum of male and female, or the sum among the year.
- Derived rows: There were lots of derived rows such as Chapter 1(A00 to B99), etc. (Used regular expression to filter them)

They were all useless, the data we need were the number of exact cause, gender and year of death.

```
In [3]: df = excel_data.parse('Table 13.1')

        # drop the columns which are all NaNs
        df = df.dropna(axis=1, how = 'all')

        # drop the rows which are all NaNs
        df = df.dropna(axis=0, how = 'all')

        df.columns = list(range(len(df.columns)))

        df = df.dropna(subset = [0])

        df = df[df[0].str.match('.+\([A-Z][0-9][0-9]\)')]
```

Step 3. Parse the useful data to a new data frame

As the dataset has been cleaned, I selected the useful part of it to a new data frame, and added the suffix to identify each region(prepared for afterwards concatenation and merge)

```
dfNSW = pd.DataFrame({
    '2006 male' : df[1],
    '2006 female' : df[2],
    '2007 male' : df[4],
    '2007 female' : df[5],
    '2008 male' : df[7],
    '2008 female' : df[8],
    '2009 male' : df[10],
    '2009 female' : df[11],
    '2010 male' : df[13],
    '2010 female' : df[14],
    '2011 male' : df[16],
    '2011 female' : df[17],
    '2012 male' : df[19],
    '2012 female' : df[20],
    '2013 male' : df[22],
    '2013 female' : df[23],
    '2014 male' : df[25],
    '2014 female' : df[26],
    '2015 male' : df[28],
    '2015 female' : df[29],
    '2016 male' : df[31],
    '2016 female' : df[32]
})
dfNSW = dfNSW.add_suffix(' NSW')
dfNSW['Causes of death'] = df[0]
```

Step 4: Concatenate all subset of data together

Did the same thing to each data from 8 regions, and then concatenate them into one data frame, based on the ‘Causes of death’(the function will automatically match the same column，they are concatenated by ‘Causes of death’)

```
In [14]: dfAll = pd.concat([dfNSW, dfVIC, dfQLD, dfSA, dfWA, dfTAS, dfNT, dfACT], axis = 1)
```

:

2007 male NSW	2008 female NSW	2008 male NSW	2009 female NSW	2009 male NSW	2010 female NSW	2010 male NSW	...	2012 male ACT	2013 female ACT	2013 male ACT	2014 female ACT	2014 male ACT	2015 female ACT	2015 male ACT	2016 female ACT	2016 male ACT	Causes of death
0	4	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	Cholera (A00)
0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	Typhoid and paratyphoid fevers (A01)
2	0	0	1	0	2	4	...	0	0	0	0	1	0	0	0	0	Other salmonella infections (A02)
0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	Shigellosis (A03)
5	9	3	2	3	10	14	...	4	4	0	0	0	3	0	1	0	Other bacterial intestinal infections (A04)

There are 1682 rows and 177 columns after wrangling, and ready for Tableau

Data Set 2:

Step 1: Import data into data frame
This set was much easier, all regions are in one sheet

```
In [18]: excel_data = pd.ExcelFile('310104.xls')
         df = excel_data.parse('Data1')
         df.head()
```

Out[18]:

	Date	Estimated Resident Population ; Male ; New South Wales ;	Estimated Resident Population ; Male ; Victoria ;	Estimated Resident Population ; Male ; Queensland ;	Estimated Resident Population ; Male ; South Australia ;	Estimated Resident Population ; Male ; Western Australia ;	Estimated Resident Population ; Male ; Tasmania ;	Estimated Resident Population ; Male ; Northern Territory ;	Estimated Resident Population ; Male ; Australian Capital Territory ;	Estimated Resident Population ; Male ; Australia ;	...	Estim Resi Popul ; Fem Austr
0	2006-03-01	3336005	2495951	1989764	765272	1030021	241193	108137	165351	10132982	...	1026
1	2006-06-01	3339035	2502687	1999858	766432	1034567	241227	108502	165814	10159424	...	1029
2	2006-09-01	3351254	2514544	2011909	768884	1040869	241864	109220	166549	10206440	...	1033
3	2006-12-01	3361292	2524859	2023921	770847	1047061	242549	109586	167405	10248923	...	1037
4	2007-03-01	3376796	2540885	2038366	773855	1055527	243232	110277	168533	10308891	...	1043

5 rows × 28 columns

Step2: Drop useless columns and rows

- Derived columns: Such as the sum of male and female, (Persons)or the sum among all region (Australia)
- Outrange year of data, the file contained population data from 1976 to 2017, what I used was 2006 to 2016

Step3: Reformat the data

- Data was recorded quarterly, what we need is the data for each year, so I got

the average of each season

- The average population is not integer, so I rounded the data to integer
- Rename the header in a unified format, e.g. (female VIC) ready for split and merge afterwards

```
In [21]: df = df.filter(regex = ('[A-Z]{2,3}$'))
         df
```

Date	male NSW	male VIC	male QLD	male SA	male WA	male TAS	male NT	male ACT	female NSW	female VIC	female QLD	female SA	female WA	female TAS	female NT	fema ACT
2006	3346896	2509510	2006363	767859	1038130	241708	108861	166280	3410731	2564748	2014404	787465	1020353	248370	100818	1698
2007	3394678	2557694	2057622	776575	1065559	244176	111497	169646	3453598	2608710	2065002	796138	1048330	249897	102983	1730
2008	3453669	2612300	2114359	786150	1100030	247428	115032	173019	3505718	2660002	2119272	805020	1081910	252061	105412	1758
2009	3509365	2669210	2167133	797108	1132248	250563	118413	176680	3557349	2715222	2170858	814445	1114170	254366	107834	1788
2010	3552786	2709534	2203070	805986	1157718	253152	120548	180155	3601030	2760640	2209655	822531	1140760	256030	109236	1822
2011	3591056	2748862	2239330	812866	1189454	254737	121542	183543	3640573	2805385	2248870	828910	1172154	256872	110160	1855
2012	3634007	2805044	2283240	821200	1230079	254698	123880	187739	3687890	2863962	2296857	837026	1205939	257264	112933	1897
2013	3682170	2865474	2321448	828653	1260796	254677	126921	191182	3741875	2927059	2341506	844820	1235404	258089	115608	1932
2014	3734151	2927865	2352379	836198	1272908	254767	126724	194066	3798953	2991890	2379346	853010	1252997	259262	116789	1963
2015	3787065	2992291	2377939	842624	1280544	255075	126443	197500	3855811	3058104	2413856	860586	1265574	260569	118014	2000

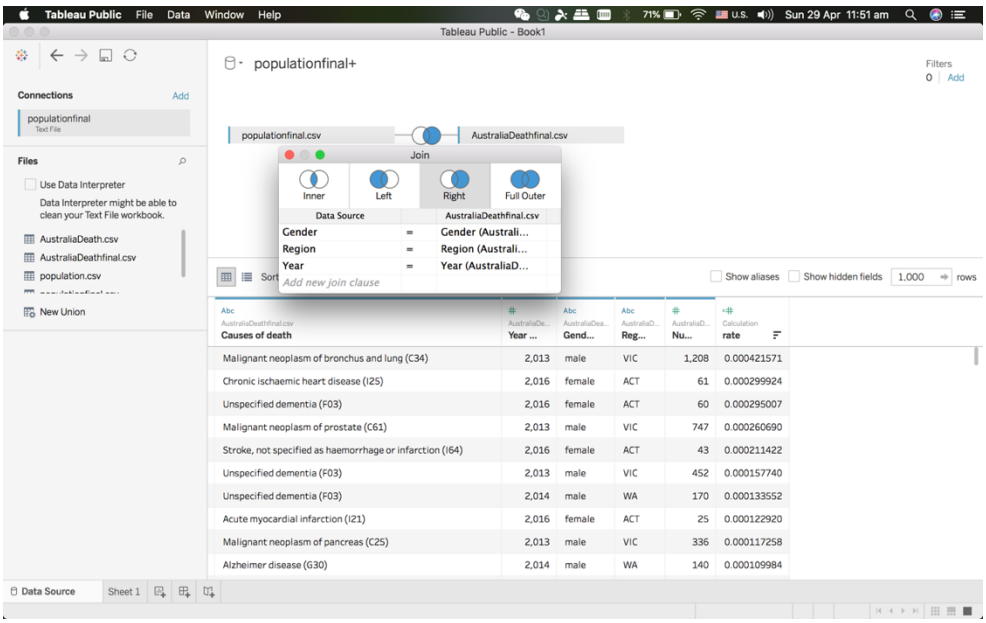
```
In [22]: df.to_csv('./population.csv')
```

There are 11 rows and 16 columns ready for Tableau

Tableau：

Tableau Desktop 10.5.3

1. Import both data set into Tableau pivot the data on each column and split the columns into gender, year and region. Export two
2. And then join two data set on gender, year and region. Calculate a death rate by using the population. The final data set is 2072231 rows and 5 rows, which is ready for data visualization.



3. Data Checking

For the data sets were retrieved from ABS, I thought it would be accurate, however, I still found some tricky errors in the data set of 'Causes of Death'

1. When I was wrangling the data, I have checked the data as well:

```
# check if the number of people is correct (male + female = total)
for i in range(1,32,3):
    print(df[i+2].equals(df[i]+df[i+1]))
```

```
True
True
True
True
True
True
True
True
True
True
True
```

The data contained the sum of males and females as Persons, for the national sheet, the total amount are all correct, shown as above.

However, in the region sheet, for example, NSW:

```
dfNSW = dfNSW.add_suffix(' NSW')
dfNSW['Causes of death'] = df[0]
# check if the number of people is correct (male + female = total)
for i in range(1,32,3):
    print(df[i+2].equals(df[i]+df[i+1]))
```

```
False
False
False
False
False
False
False
False
False
False
False
```

And I had a look back to the raw data:

	2006		
Cause of death and ICD-10 code	**Males**	**Females**	**Persons**
Total deaths	**23,514**	**22,603**	**46,117**
Causes of death			
CHAPTER I Certain infectious and parasitic diseases (A00-B99)	**433**	**383**	**816**
Intestinal infectious diseases (A00-A09)	10	9	19
Cholera (A00)	0	0	0
Typhoid and paratyphoid fevers (A01)	0	0	0
Other salmonella infections (A02)	3	0	3
Shigellosis (A03)	0	0	0
Other bacterial intestinal infections (A04)	1	2	6
Other bacterial foodborne intoxications, not elsewhere classified (A05)	0	0	0
Amoebiasis (A06)	0	0	0
Other protozoal intestinal diseases (A07)	0	0	0
Viral and other specified intestinal infections (A08)	6	2	10
Other gastroenteritis and colitis of infectious and unspecified origin (A09)	1	1	2

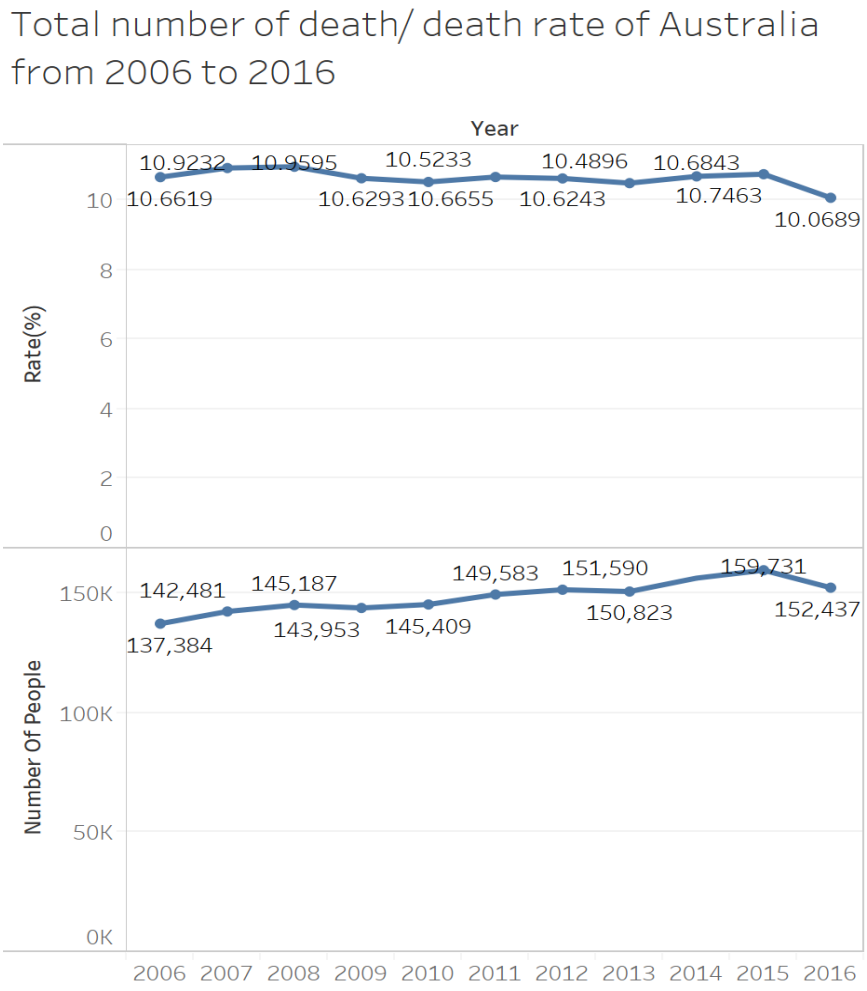
some errors happened in raw data (the number of persons was not equal to the sum of females and males). However, the total deaths were correct.

What I did was drop all derived data and got the derived data by calculating.

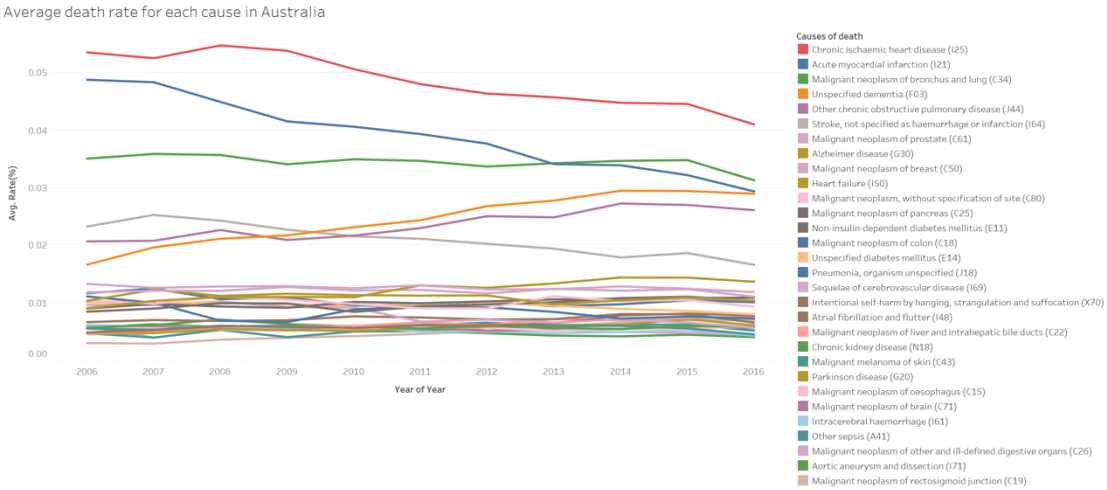
2. In the process of visualization, I found some data value actually missing but recorded as zeros, it was hard to repair and restore the raw data. Therefore, I used ‘average rate’ to do the analyze to deduct the influence by null data.

4. Data Exploration

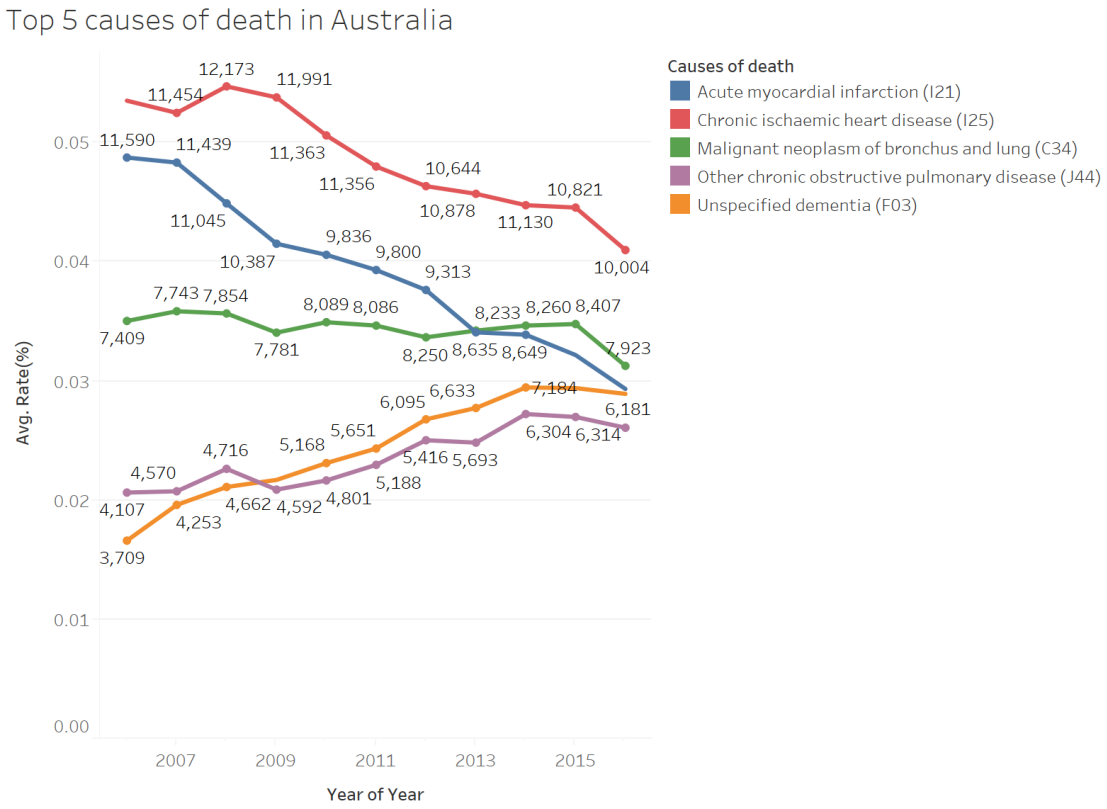
4.1 Explore overall data



Looked directly to the data, we might think that from 2006 to 2016, the death of Australian was increasing and decreased in 2016. However, as we take the population into account. Actually, the death rate of Australia was stable at around 10.7 and decreased to 10.0 in 2016



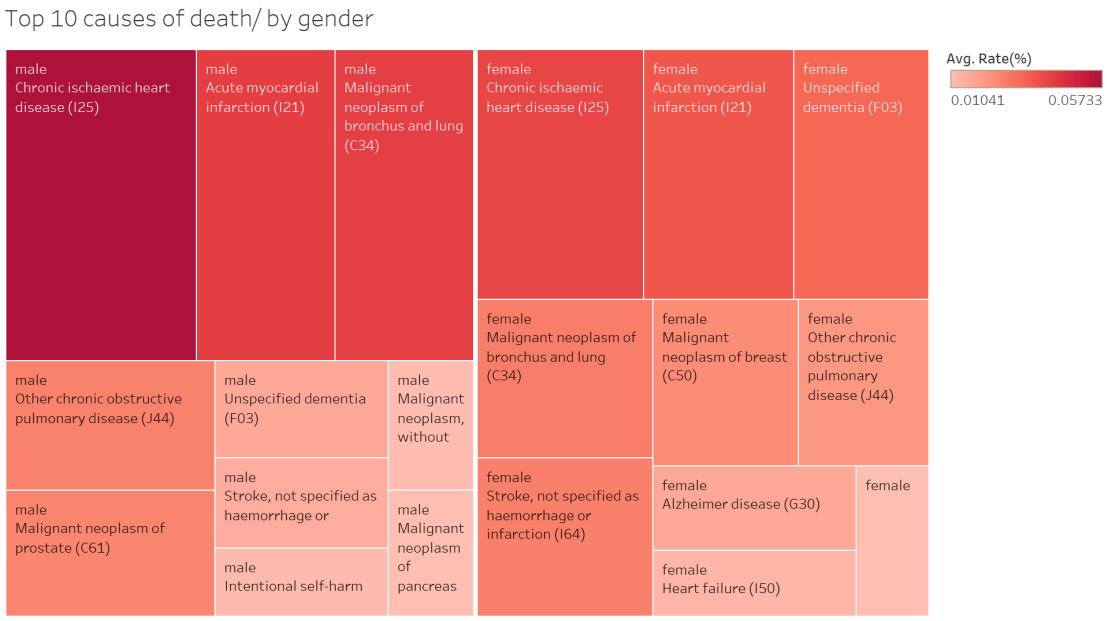
Too many types of causes in the category, in order to have a look for the rough trend of the cause, I plotted top 30 highest rate among 1794 causes, I found that there were actually nearly 10 types of causes account for over 0.1%.



The trend of average of death rate(%) for Year. Color shows details about Causes of death. The marks are labeled by sum of Number of People. The view is filtered on Causes of death, shows the top 5 highest death rate causes, which keeps Acute myocardial infarction (I21), Chronic ischaemic heart disease (I25), Malignant neoplasm of bronchus and lung (C34), Other chronic obstructive pulmonary disease (J44) and Unspecified dementia (F03).

They were the 5 dominant diseases in the past 10 years, and three of them were decreased especially Acute myocardial infarction(I21). However, we should be carful for the Other chronic obstructive pulmonary disease(J44) and Unspecified dementia(F03), which are likely becoming more and more dominant disease for Australian.

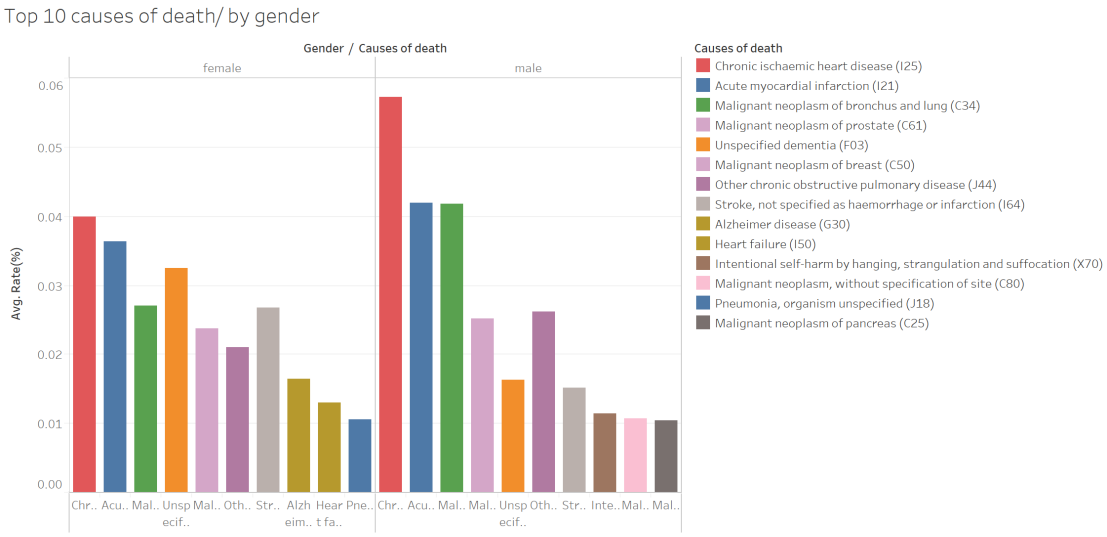
4.2 Explore data by gender



As the diagram shown:

- Chronic ischaemic heart is the most dominant disease, it was at the first place of death rate.
- followed by were Acute myocardial infarction(I21), Malignant neoplasm of bronchus and lung(C34), Other chronic obstructive pulmonary disease (J44) and Unspecified dementia (F03).

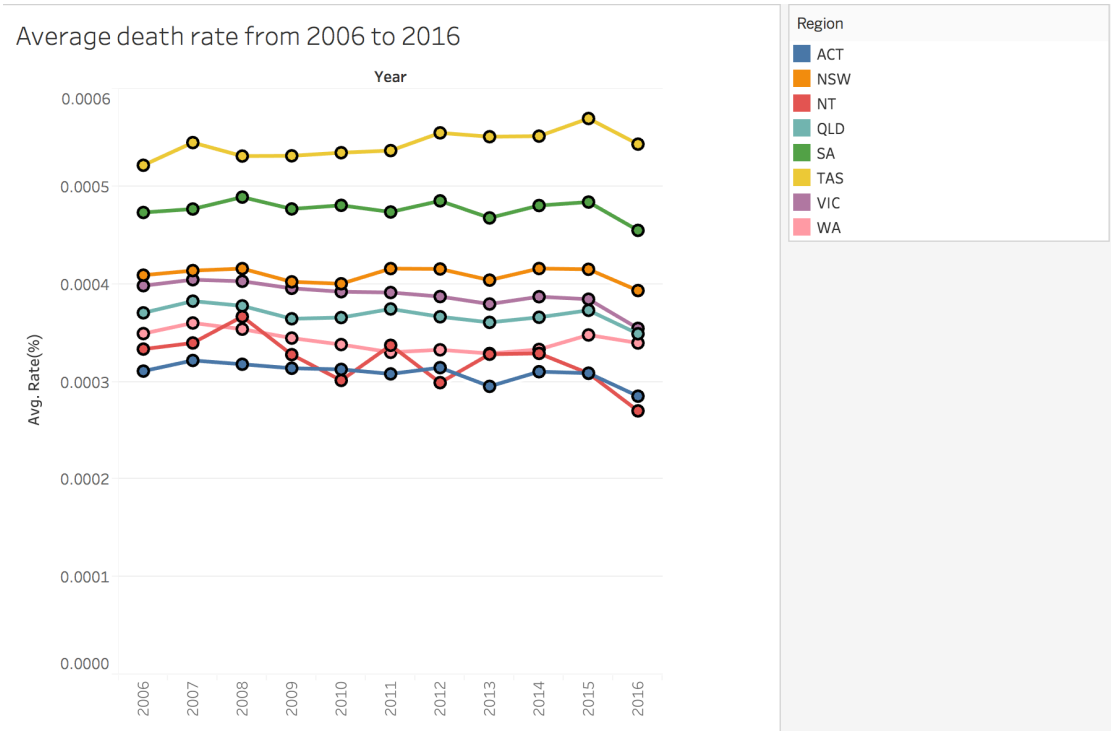
Then, I had a comparison between top 10 causes of death by gender:



As the diagram shown:

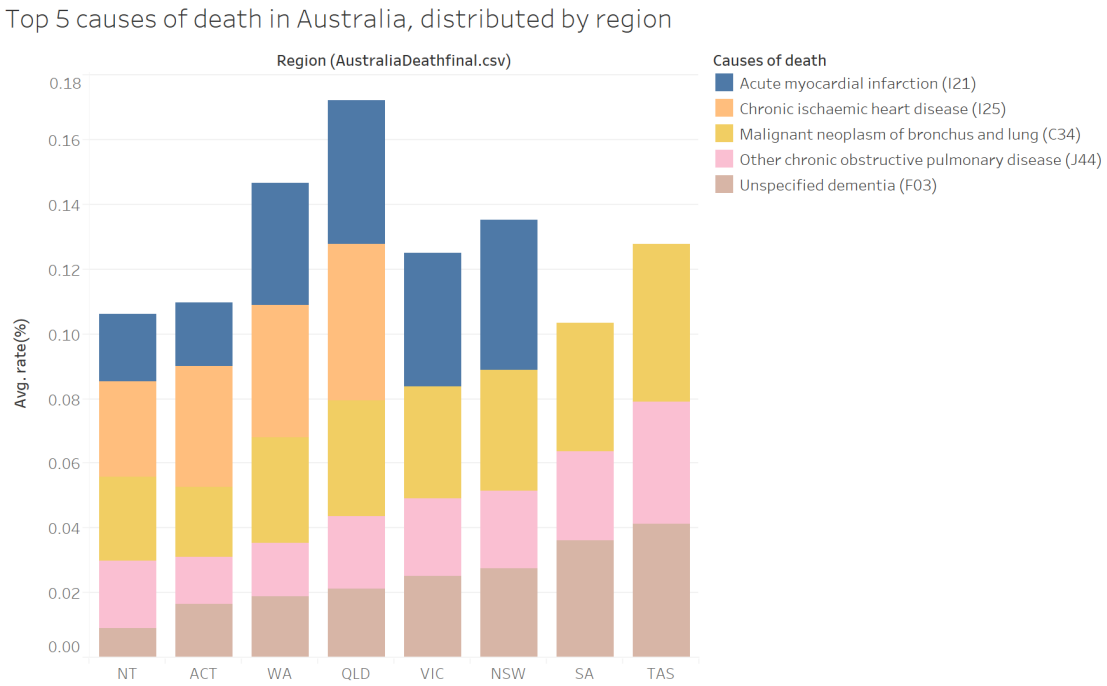
- In the top 10 list men have higher death rate at most of the diseases, except Unspecified dementia(F03) and Stroke(I64) .
- For men, Malignant neoplasm of prostate(C61) and for women Malignant neoplasm of breast (50) were in 4th and 5th place of top 10, which can be daily detected by simple way, which has lots of work to do.

4.3 Explore data by region



- Tasmania always had the highest death rate among Australia, nearly twice as much as ACT.
- The richest states VIC and NSW were in the middle.

- All states' death rate were stable between 2006 to 2016 except Northern Territory
- The whole nation's death rate decreased in 2016



Compared top 5 diseased find before, I tried to compare the data across region. Surprisingly, I found that in the data, there was no Chronic ischarmic heart disease(I25) record in VIC, NSW, SA, TAS. Also, no Acute myocardial infarction(I21) in SA and TAS. I have checked the raw data and they were all zeros. However, here was some data from Australian Heart Foundation (Foundation, 2014), which indicated that in 2014 there were over 1000 people died from IHD in Victoria, and the data from ABS didn't record it.
Therefore, compared to other regions, VIC, NSW, SA and TAS should have higher death rate than the diagram shown.

5. Conclusion

In a conclusion, the average death rate was decreased from 2015 to 2016, which was pretty optimistic, however, during 2006 to 2016, also there were diseases became more and more dominant, e.g. Other chronic obstructive pulmonary disease (J44) and Unspecified dementia (F03), which should be alerted.

Many diseases have an extremely high cure rate if they are effectively treated at the beginning of the disease, e.g. Malignant neoplasm of prostate(C61) and Malignant neoplasm of breast (50). The hospitals can greatly reduce the mortality of these diseases if they can promote knowledge of self - exam of these diseases.

After exploring the data, I can answer the questions at the beginning:
a. What are the most common causes of death in the past decade in Australia across states, gender?

The most dominant causes of death in the past decade were Acute myocardial infarction (I21), Chronic ischaemic heart disease (I25), they were the the most common causes in both gender and all regions.

b. From 2006 to 2016, total death was gradually increased, why?
Actually, the death rate was stable from 2006 to 2016 among Australia, the total number of deaths was increased was because of the growth of population.

c. In the past decade, which causes of death have been controlled and which are more and more dominant?
In the top 5 highest death rate disease. Other chronic obstructive pulmonary disease (J44) and Unspecified dementia (F03) were become more and more dominant and tend to be the highest disease in the future.

6. Reflection

I have learnt many aspects of data science from this project, from searching for raw data, wrangling data to formatted form, and then importing the data to tableau to do visualization.

I also met some difficulties. The first one was when I wrangled the raw data, there were 9 sheets needed to join, merge or concatenate. I spent a lot of time on data preparation, learned a lot of knowledge dealing with data on python. The other one was reformatting the data to tableau the data has 1.2 million records after reformatting and the software was not stable, I had to do everything carefully and tried different tools of visualization.

7. Bibliography

Aubusson, K. (2016, September 23). *Australia among top 10 healthiest countries in the world, Global Burden of Disease Study shows*. Retrieved from Sydney Morning Herald: https://www.smh.com.au/healthcare/australia-among-top-10-healthiest-countries-in-the-world-global-burden-of-disease-study-shows-20160923-grmyz5.html

Duckett, S. (2016, March 15). *The problems with Australia's hospitals – and how they can be fixed*. Retrieved from the conversation: https://theconversation.com/the-problems-with-australias-hospitals-and-how-they-can-be-fixed-54248

Foundation, A. H. (2014). Retrieved from https://www.heartfoundation.org.au/images/uploads/main/Your_heart/RES-113_Aust_heart_disease_statistics_2014_WEB.PDF

HealthWorkforce. (2012, March). *Wayback Machine*. Retrieved from https://web.archive.org/web/20140521003131/http://www.hwa.gov.au/sites/uploads/health-workforce-2025-volume-1.pdf

Welfare, A. I. (2011, March 12). Retrieved from https://web.archive.org/web/20110312160804/http://www.aihw.gov.au/life-expectancy-how-australia-compares/

Wikipeadia. (2014, April). Retrieved from https://en.wikipedia.org/wiki/ICD-10