INSTITUTE OF HEALTH INFORMATICS

Graduate Programme in Health Data Science

Assessed Coursework Submission

Student:	LIU Jianyu
Module:	CHME0013: Data Methods for Health Research
Date due:	Monday, 21st January 2019, 12:00 midday
Word count: (excluding references, diagrams and appendices)	2498
Disability or other medical condition for which UCL has granted special examination arrangements:	
My learning development:	On this assignment, I have been particularly focusing on....
	data analysis and visualization for health research.
	In addition to general feedback, please give me feedback on .....

Assignment A

The packages used in this assignment are NumPy, Pandas, Matplotlib and SciPy. NumPy is used for scientific computing including the mean and standard error. Pandas is used most frequently in this assignment. Dataframe is used for calculation, organize data and display data through the research. It could not only help us read data from CSV files but also help us to solve the problems in multidimensional structured data sets. The pie chart is also base on the function which is used to support Dataframe. The pyplot in Matplotlib is used for data virtualization. The histogram, scatterplot and trend line are based on this function. SciPy is used to show the Gaussian curve based on the histogram and used for normal distribution test.

The data sources used in Assignment A are NHS Digital GP Practice Prescribing and NHS Digital GP Practice Demographics in April 2018. In NHS Digital GP Practice Prescribing, they provide information in practice level prescribing data. The variables including the practice code, the BNF code, the BNF name, the total actual cost, practice's address, practice's city, practice's postcode. In NHS Digital GP Practice Demographics, they provide information about patients registered at GP. The variables including the practice code and the number of patients. The practice code is existing in both data sources, and it could be identified as a potential identifier for data linkage.

To identify all GP practices located in London, we need to formulate the rule to support us. In the file 'codes_names_address.CSV', we could find practice address, city and postcode. When we check the information, we could find some noisy data result in the variable 'city' is empty and the information about the city is provided in the address. We could also find some round outside London is named as 'London'. After considering all the potential noisy data, the rule could be explained as identify the row which the variable 'Address2' or 'City' equal

'LONDON'. Based on this rule, we could identify all GP practices located in London, and the total number of it is 929.

The next step is calculating the total number of patients registered. We need to use the file ‘gp-reg-pat-prac-all.CSV' which including the number of patients in each practice. The list of practices in London can help us identify the targets. We need to summarize the number of patients and the result is 5,841,956.

In order to identify the total number of prescriptions, the file ‘Practice_prescribing.CSV' can be used and we also need to identify the practice in London by the list. The variable ‘QUANTITY' is mean the total quantity of each prescription in each practice code during April. So we can summarize the variable ‘QUANTITY', and the result ‘510136987' is mean the total number of prescriptions.

Similarly, the total actual cost of these prescriptions can be calculated by summarizing the variable ‘ACT_COST'. The total actual cost is 43322000.54.

The top 10 most frequent drugs prescribed is descending sorted by the summarized result of ‘QUANTITY'. In the table we could find the most common drugs is ‘Ensure Plus_Milkshake Style’ and the frequency of it is much higher than the second one (It was used for 17023232 times in April). It is a ready-to-drink, milkshake style oral nutritional supplement for people with, or at risk of developing disease-related malnutrition. The third highest frequency drug is ‘Fortisip Compact_Liq’ which is similar to ‘Ensure Plus_Milkshake Style’. We can infer that people with disease-related malnutrition account for a large proportion of the population. The second one is ‘Metformin HCl_Tab 500mg’ which the first-line medication for the treatment

of type 2 diabetes, particularly in people who are overweight. Diabetes is also a high-frequency disease in London.

The bottom 10 less common drugs are variety because there are many drugs which are only used for one time in April. The number of medications used one time in April is 297, and there also have a drug exist in the list but ‘QUANTITY' is 0, and it is a noisy data.

In Cambridge, the total number of patients registered is 311579, and there have 36 GP practices. The total number of prescriptions is 25232152, and the total actual cost is 2434403.94. In the top 10 most frequent drugs, nutrition is the most significant proportion of them, and this kind of drugs ranked 1, 2 and 4 in the list. The demand for is ‘Metformin HCl_Tab 500mg’ only ranked 9 and ‘Paracet_Tab 500mg’ for fever and ‘Dermol 500_Lot’ for skin condition ranked 3 and 5.

Comparing London and Cambridge, we could find the standard error between these two cities is different, which mean that the dispersion of patients in Cambridge is more uneven than London. We could also see the average patients per practices in Cambridge is 10051 and in London is 7821. Each practice in Cambridge needs to service more patients than in London. The active GP is mean the GP with at least one patient record. We could find that a large number of practices in London do not have any record in April.

To identify the prescriptions related to cardiovascular disease and antidepressants, we need to check the British National Formulary. We find the code of these two diseases are begin with ‘02' and ‘0403', so we need to extract all the samples which BNF code start with this two number. The total number of prescriptions across all practices for drugs related to

cardiovascular disease and antidepressants are 933262147 and 214223401. The total actual cost across all practices for medications related to cardiovascular disease and antidepressants are 90193834.02 and 16853470.86.

Actual cost and number of patients of each practice can be extracted from data source so we can calculate the relative costs per patient through ‘actual cost' divided by ‘number of patients'. The average cost per patients in April is around 11.61, and the average actual cost per practices is 87842.77. We can show a scatterplot of total spending across all practices in April and the number of patients in each practice in April to visualize the total monthly expenditure per registered patients. But the first step is data cleaning because some noisy data may influent the fit of the trend line. The method we used in the data cleaning process deletes the largest 0.1% data for both actual cost and number of patients. With the help of this process, we can show a reasonable scatterplot and a trend line. From the diagram1, we can find that as the number of patients grows, total spending across all practices has a growing trend.

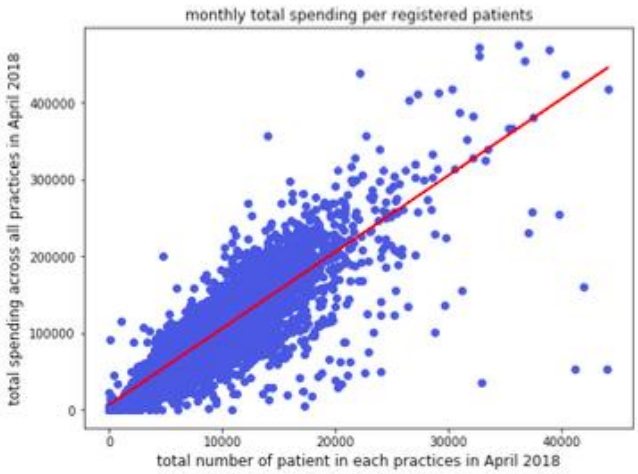

diagram1. monthly total spending per registered patients

But we could also find in the diagram2 of the number of patients per practices, the distribution of the histogram is not normally distributed which is mean that the distribution of

patients is uneven. This situation could result in some practices are very busy, and at the same time, another one is free.

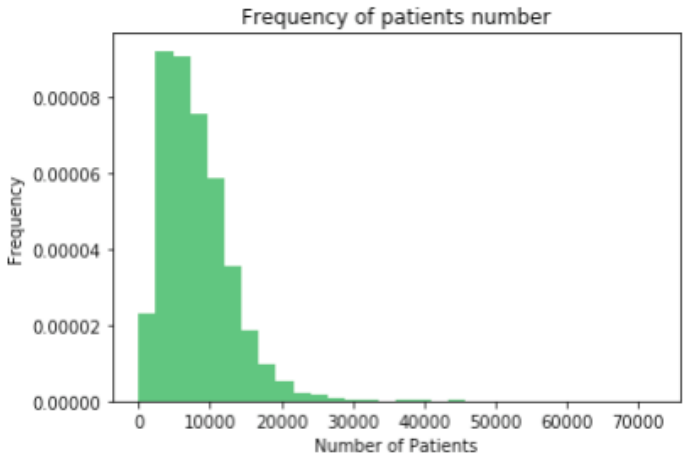

diagram2. Frequency of patients' number

The histogram for relative spending for all practices also needs data cleaning. For this histogram, we use the variable 'Relative_Costs_Per_Patient' and delete the bottom 1% and the top 1% of the value to avoid the influence of outliers. We also fit a Gaussian curve to support us to judge the distribution. From the diagram3, we can conclude that the distribution of relative costs per patient is approximating a normal distribution.

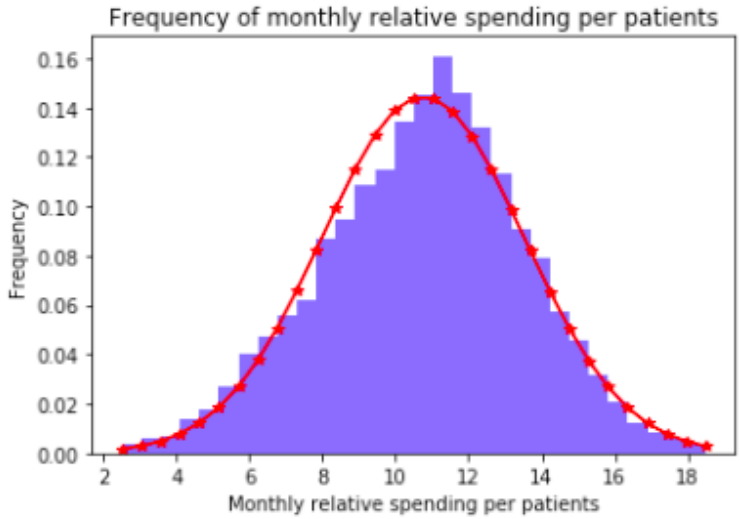

diagram3. Frequency of monthly relative spending per patients

Further test, we use the function ‘scipy.stats.normaltest’ to test whether a sample differs from a normal distribution. The null hypothesis of this test is the sample comes from a normal distribution. We receive the return P-value is less than 0.01, so we can reject the null hypothesis, and this is mean that the distribution of spending per practices is not a normal distribution. The diagram seems like negative skewness which means the average is less than the median, and the median is less than the mode.

Assignment B

The data sources used in Assignment B is the WHO Mortality (ICD-10 version) and Population datasets. It is a database of registered deaths compiled by WHO from data given by national authorities around the world. The cause of each death is classified by the circumstances that led to death. We will use the Mortality data, country codes and Population and live births datasets. The ‘Mortality data' including the variables country code, year, sex, death cause code (ICD-10) and death age. Another file which called ‘Documentation_1 Dec2018’ including the explanation of ICD-10 code so we could identify the target cause based on the variable ‘Cause’.

Through the file of country code, we could identify the country code of Iceland, Italy and New Zealand are 4160, 4180 and 5150, We need to focus on mortality data at the year of 2010, and we could get the total number of deaths of these three countries. The deaths number in Iceland in 2010 is 4038, in Italy is 1169230 and in New Zealand is 57298.

Similarly, we could extract the data of population from ‘Population_and_live_births.CSV'. Based on country code and year, we could identify the population of those three countries. The population in Iceland, Italy and New Zealand in 2010 is 318041, 60483386 and 4367360.

The population and mortality in these three countries differ significantly. We could calculate the mortality rate by using death number in 2010 divided by the population in 2010. The mortality rate in Italy is considerably higher than the other two countries, and the rate is 19.3 in a population of 1000. This rate is much higher than the result provided by The World Factbook (10.4 in a population of 1000) which is supported by the Central Intelligence Agency. This problem may be due to the noisy data and different data collection rules, but we could not identify the reason based on the available data.

By summarize the death number of the different age group in Italy, we could visualize the distribution of deaths. From the diagram4, we could find the most proportion of the people deaths at age 80 to 84 and 85 to 90. The average life expectancy in Italy provided by World Bank is 82.54 in 2016, and the peak of deaths frequency is just over this expectancy. We could also find the deaths frequency at 0 is significantly higher than several groups after it. This result is mean that children age under one year old is also an essential stage of people's healthy and people ought to pay more attention to the health of new-born.

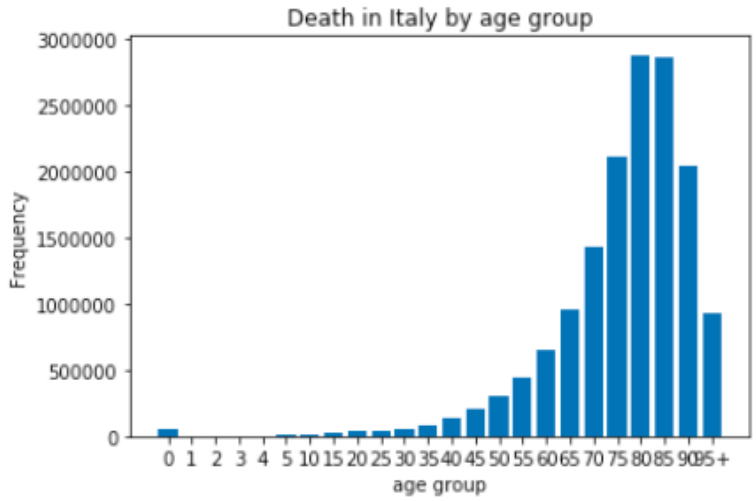

diagram4. Death in Italy by age group

For the variable ‘Cause', we could find that there are three different kinds of data type. Some of them are recorded in one letter followed by three numbers, some of them are recorded in one letter followed by two numbers, and the rest is recorded by four digits number. Through ‘Documentation 1 Dec2018', we could find the type of 4 digits number is a potential noisy data because it could be ICD-9 code or code in ICD 10 Mortality Tabulation List 1. For example, 1026 could mean 102.6, and it is an ICD-9 code which means ‘Bone and joint lesions due to yaws’. But it could also be a code in ICD 10 Mortality Tabulation List 1 which replace C00-D48. This data type makes people feel confusing so we could not judge this type of data. For another two types of data, they are all ICD-10 code in different levels of detail.

In our data source, we could find 620 kind of codes between C00 and D48 which is relative with a neoplasm (from C00.0 to D48.9). Based on this code, we could extract all the samples relative with neoplasm, and we could calculate the proportion of each deaths cause. The top5 cause is shown in the table, and we could check in the WHO website to identify the meaning of each code.

The majority type of neoplasm result in death in Italy is C34.9 which means ‘Malignant neoplasm of unspecified part of bronchus or lung', and the proportion of this type is 19.0% which is much higher than the second highest reason (only 6.9%). This phenomenon indicates that neoplasm of bronchus or lung is a majority reason for people's health problems. Furthermore, the danger of smoking should further attract people's attention.

In the diagram5, we combined the reasons other than the top five to highlight these five reasons. The top 5 causes of death relative with neoplasm are malignant neoplasm of unspecified part of bronchus or lung, malignant neoplasm of breast of unspecified site,

malignant neoplasm of colon unspecified, Malignant neoplasm of stomach unspecified and malignant neoplasm of pancreas unspecified. We could also find all of these five reasons are unspecified and we need to pay attention to the difficulty of diagnosis neoplasm type.

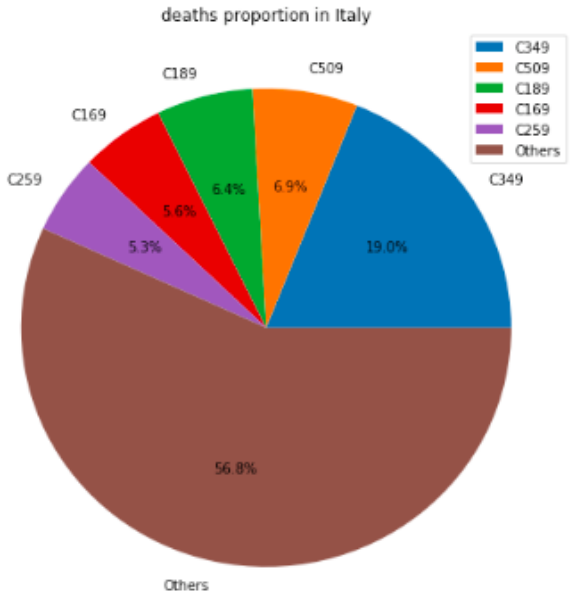

diagram5. deaths proportion in Italy

The top 5 age group in Australia dying with neoplasms cause of death could be identified by summarizing the death number in Australia of each age group. We could find except age group from 85 to 89, the number of deaths with neoplasms increases with age.

But when we focus on the proportion of dying with neoplasms cause of death in all the deaths people in each age group, we have a different finding. The top5 percentage is people from 60 to 64, and the proportion is 26%. This result is mean that 26% of people die in there 60 to 64 is because of neoplasms. And the top5 group is from 50 to 75, and this group could be identified as a high incidence of neoplasms in Australia. So it is clear that there have differences by age group for deaths from Neoplasms in Australia for 2010.

To compare and contrast the frequency of deaths by Neoplasms in Italy and Australia in 2010, we could summarize the deaths and population information in 2 different logic.

First, we can combine information from different type of Neoplasms. Based on this logic, we could compare the majority type of Neoplasms in these two countries. From 'Neo/Death%', we could calculate the proportion of deaths from each type of Neoplasms to the total deaths of Neoplasms. We could find the most frequency Neoplasms in both countries is C349, which means 'Malignant neoplasm of unspecified part of bronchus or lung'. But the frequency of C61, Malignant neoplasm of prostate, in Australia is much higher than Italy. From 'Neo/pop%', we could calculate the proportion of deaths from Neoplasms to the total population. The order of this variable is similar to 'Neo/Death%'.

Second, we can combine information from different age group. Based on this logic, we could compare the mortality from Neoplasms in each age group.

From the variable 'Neo/Death%', we could find Neoplasms is a majority cause of death at 5 to 9 and 60 to 64 years old in these two countries. Compare with Italy, Australia has better control of Neoplasms before 75 years old. The result is shown in diagram6.

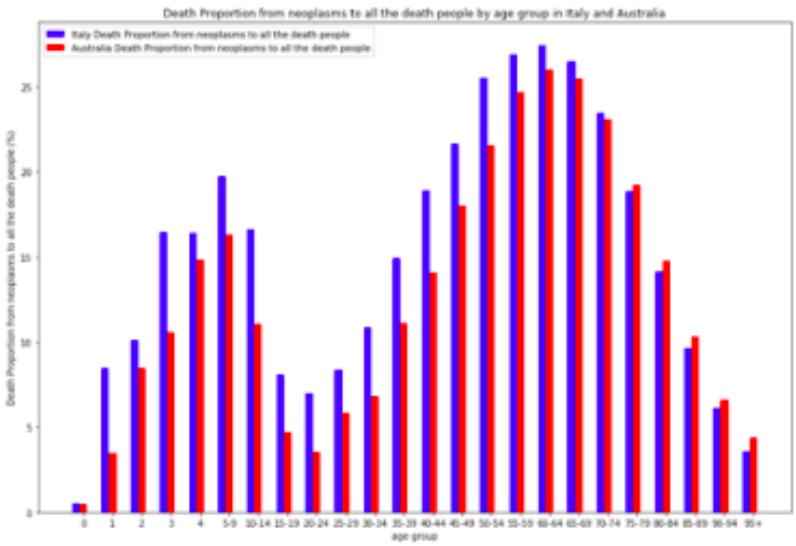

diagram6. Death Proportion from neoplasms to all the death people by age group in Italy and Australia

From the variable 'Neo/pop%', we could find the proportion of people deaths from Neoplasms to each age group's population continuous increase. The older the person, the more risk of Neoplasms should be prevented. The result is shown in diagram7.

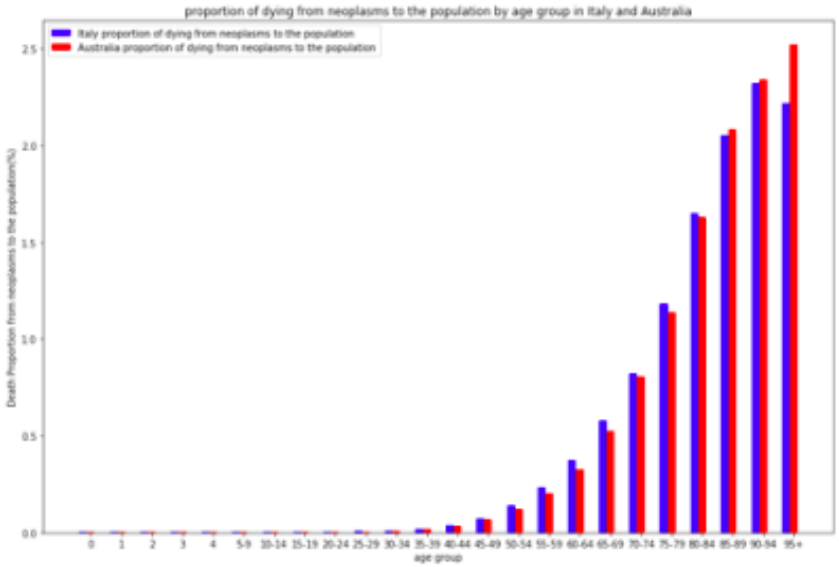

diagram7. proportion of dying from neoplasms to the population by age group in Italy and Australia

REFERENCES

Is Metformin the World's Next Wonder Drug? Can It Block Cancer Cell Growth? (2017, June 8). Retrieved January 20, 2019. Available at: http://trendintech.com/2017/06/07/is-metformin-the-worlds-next-wonder-drug/ [Accessed 21, January 2019].

Wingo, P.A. et al., 2003. Long-term trends in cancer mortality in the United States, 1930–1998. Cancer, 97(S12), pp.3133–3275.