



**Foundations of Analytics Final Report**

**Instructor: Dihui Lai, PhD**

**Group C1:**

**Qianyue Ma, Yiwen Chen, Garry Ge,**

**Qingyuan Wang, Qinxin Chen**

**Washington University in St. Louis**



Table of Content

[**Background**](#_bemstsbju6b7) **3**

[**Approach**](#_tj977g3n024a) **4**

[**Result**](#_2rzbn761zwk8) **6**

[**Interesting Findings**](#_22x3zkylkrfh) **16**

[**Conclusion**](#_g5bj7iqmgxjg) **17**

[**Appendix**](#_onfpdqc7xcdr) **17**

# 

# Background

Our group was hired as data scientists by a prestigious insurance company. The vice president of our department would like to launch a new life insurance product and asked us to do data analytics on the previous mortalities and come up with some critical trends, features, and finally, business-purpose recommendations. After thoroughly discussing, our group decided to work based on the morality data (from 2005 - 2015) from CDC. Every year, Centers for Disease Control and Prevention (CDC) provides detailed statistics of deaths and their underlying causes in the United States. Mortality data is a fundamental source of demographic, geographic, and cause-of-death information. This is one of the few sources of health-related data that are comparable for small geographic areas and are available for a long time period in the United States. This data is reliable and trustworthy as it has been used for numerous research and it is used by the government.

There are eight features included in this report: (1) State and county of residence, (2) year of death (rather than the full date of death), (3) race (1968-98: recoded to white, black, other; 1999-present: recoded to white, black, American Indian or Alaska Native, Asian or Pacific Islander), (4) sex, (5) for 1999-present: Hispanic origin (not Hispanic or Latino, Hispanic or Latino) (6) age group at death (specific age recoded to 16 age groups), (7) underlying cause of death (4-digit ICD code), and (8) 69 ICD-8, 72 ICD-9, 113 ICD-10 cause-of-death recode (depending on the data year).

Our tasks were around the problems below:

1. Major causes of death in the US
2. Death distribution against age
3. Top 3 causes of death for each age band
4. Causes against year. Trends in causes
5. NLP method for assigning ICD description to potential 5000 patients

# Approach

Our method is to build a logistics regression model. But before modelling, data cleaning was a crucial step in order to speed up the time of concatenation and get rid of distractors.

We first filtered relevant variables for our research by dropping missing value (missing over 90%) and kept meaningful variables only, which could help us create a relatively clean dataset and improve our following progress. Then by studying the potential relationship between death caught by diseases and other parameters, we designed our model format, that is, logistics regression. An important step we did in our study is that we tuned our parameter C with cross-validation, and then we fit the entire dataset with tuned parameters. Finally we got our model and formula and tried to make some interpretations with the outcomes and analysis of charts. Last but not least, we made some predictions based on reasonable threshold.

Our model formula is shown below.

**Y (probability of getting one ICD-Code) = ß0+ß 1 \* age+ß 2 \* sex\_F+ß 3 \* race\_1+ ß4 \* race\_2+...+ßk \* resident\_3**

In this formula, Y means the probability of death getting one ICD-Code, X are six features we chose before modeling, they are age, gender, race, marital state, month and resident state.

In this formula, value of β is related to our dependent variable Y. If β increases, the possibility of people who died due to diseases will increase. If β decreases, the possibility will decrease. If β is positive, it means the corresponding X is positively related to Y. If β is negative, X is negatively related to Y.

Using Hypothesis Testing, we set a significance level of 0.05 to measure whether it is significant to Y. For Y, we decide to set a threshold value. After trying different numerical number, we finally determine the value of 0.5 to maximize our performances. For those value more than 0.5, we define them as 1; for those value less than 0.5, we define them as 0.

To point out a basic condition, all of our variables are dummy variables, of which the value is either a 1 or a 0.

Then we started to use our model to solve every problem listed in the project. Based on data from the most recent two years, we hope to resolve the first two problems. By roughly studying data of 15 years, we find that there is not very much difference within the time period. Therefore, we think these two years’ data is sufficient for us to give comparatively accurate answers to the major causes of death and death distribution against age.

The first question is to determine the major cause of death. With the help of data visualization tools we determined the two major cause of death. From the manner of death, we picked the natural manner. From the ICD code, we picked the top three disease, that is, Atherosclerotic heart disease, Malignant neoplasm of bronchus or lung, unspecified, and unspecified dementia. We employed a basic type of statistics method to give an initial processing of the whole dataset. We downloaded all the dataset to create a sequence in periods and found a similar trend within 15 years. Therefore we focus on the recent two years’ data and verify our assumption successfully.

The second question is to get a death distribution against age with 5-year band. We have sorted all the data together with a rough analysis and we found a similar death distribution against age. Therefore we did a death distribution chart with parameters of ICD codes and manners.

To solve the third question of determining the top 3 causes for each age band, we divided all the ages into 5-year bands, and get 27 groups of different age features. All the 36 causes of death were fitted into the model and then we pick up the top 3 causes for each age band. After sorting all the causes, we found for all top 3 causes there are 4 causes have the most appearance.

To study the possible trends of causes of death, we chose 4 top causes and made a chart against years.

For the last problem, we need to determine if any patients have medical conditions that are associated with ICD codes of major causes of death. We first converted transcriptions into vectors, then calculate the similarity measure between ICD description and medical transcripts. After that, we use max similarity for each row to make ICD types corresponding to max similarity. Then we decode ICD types. Finally by setting a threshold, we assigned ICD description to a medical transcripts only if the similarity score meets the threshold.

# Result

Question 1

**What are the major causes of death in the US?**

To understand the major causes of death we had analysed the mortality dataset for 2014 and 2015. There are thousands of causes of death and our group decided to find the top 10 causes of death amount 2014 and 2015 period. Based on the ICD code the top3 cause of death is Atherosclerotic heart disease, Malignant neoplasm of bronchus or lung, and Unspecified dementia.

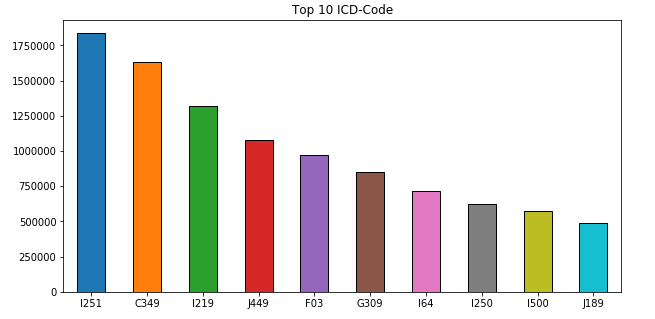


Figure 1.1: Top 10 Causes of Death in the US

|  |  |  |  |
| --- | --- | --- | --- |
| I25.1 | Atherosclerotic heart disease | | |
| C34.9 | Malignant neoplasm of bronchus or lung unspecified | | |
| F03 | Unspecified dementia | | |
| I21.9 | Acute myocardial infarction unspecified | | |
| J44.9 | Chronic obstructive pulmonary disease unspecified | | |
| G30.9 | Alzheimer's disease unspecified | | |
| I64 | Stroke not specified as hemorrhage or infarction | | |
| I50.0 | Congestive heart failure | | |
| I25.0 | Atherosclerotic cardiovascular disease so described | | |
| J18.9 | Pneumonia unspecified | | |

Question 2

**For different causes of death, how does the death distribution look like against age (For example,histogram of 5 year age band)?**

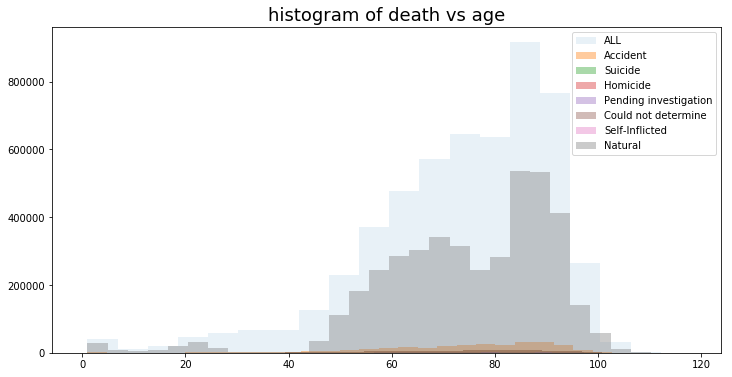


Figure 2.1 Histogram of Death Against Age (Manner of Death Based)

From Figure 2.1, the group found that from the age 0 to 10 the number of deaths have a decrease tend. The number of deaths continue to increase as the age increased. Most people dead between 50 to 94. At the age band 85-90 the number of deaths had reached the peak point. The number of deaths for age over 90 had a huge decrease due to the natural age limitations of human beings.

Question 3

**For each age band (5-year band), what are the top 5 causes of death? Do they differ?**

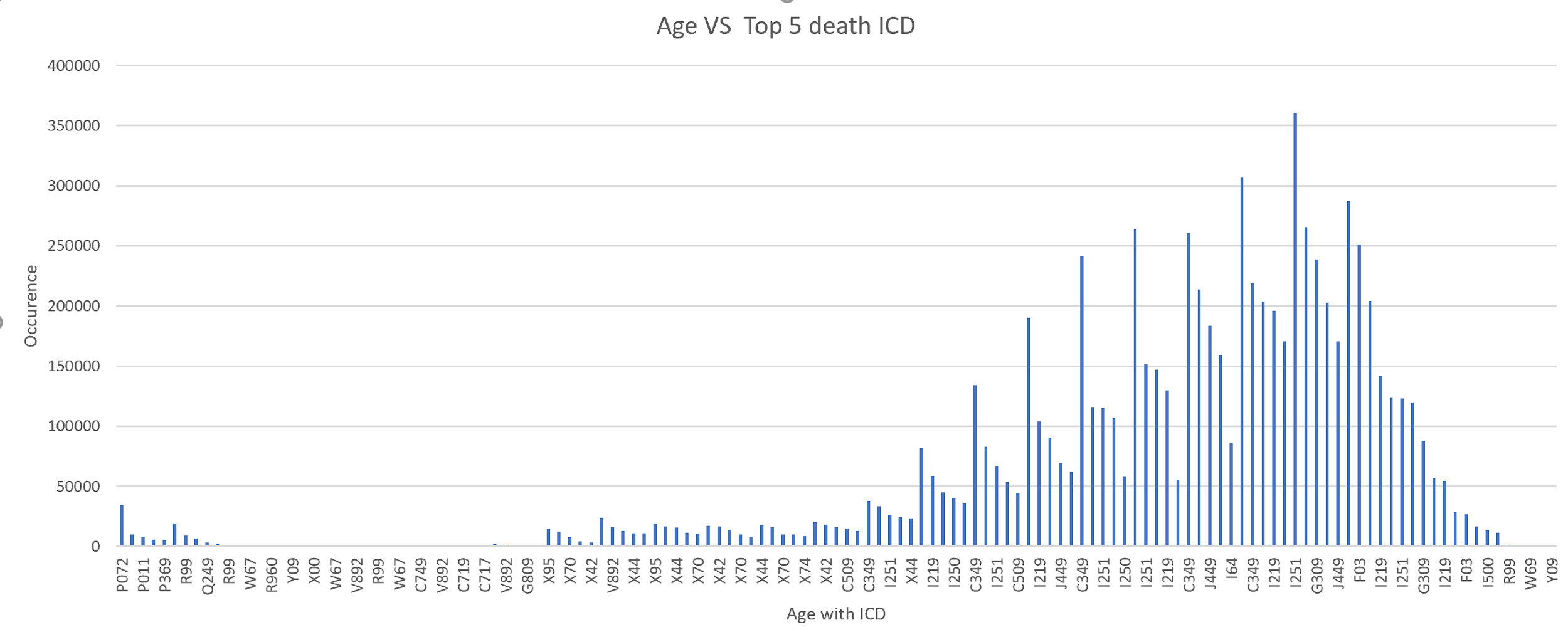


Figure 3.1 Top 5 Causes of Death (5-year band based)

The top dead ICD code various between the age. Some diseases only appeared during early age and some only appeared in the group that over 100. Extreme immaturity of newborn and infant death syndrome are the top death disease in group 0 to 4 and 5 to 10 respectively. In age group over 100, accidental drowning and submersion while in natural water and exposure to excessive natural heat, initial encounter are causing most dead.

On the other hand, there are four diseases appeared frequently in age group 90 to 120. Those diseases are Malignant neoplasm of unspecified part of bronchus or lung, Atherosclerotic heart disease of native coronary artery, Unspecified dementia, and Alzheimer's disease. Those diseases cause most dead and appeared most frequently. The group had investigate those diseases in the following questions.

Question 4

**Are the causes of death changing over time? Are there any significant increasing or decreasing trending some causes?**

In this section, we attempted to explain the association between cause of death and year using machine learning methods. Since it is impossible to consider more than 3000 types of diseases in one model, we collected top 3 death causes across 11 years and found out the main causes are concentrated in 4 ICD codes, which were the input of our later models.

In order to demonstrate relationships between these four diseases and death, we established 4 logistic regression models with death binary indicator as target variable and other features like year, race and marrigment to determine the probability of being caught by such disease (setting 2005, married, male.. as base dummy combination). We took all related variables into consideration to reduce irreducible error and logistic regression formula is shown below:

*+ +...*

We will use Betas to interpret unit change effects on the estimated target variable and p-values to decide significance level of each beta. In this case, betas represent for difference of death between predicted year and 2005, that is, betas for 2006-2015 could partial express changing trend setting 2005 as a baseline comparison. We observed that for year dummies, p-values are all quite close to zero which indicated significant influence of year on death. In order to illustrate causality explicitly between year and death, we extracted year coefficients from estimation summary report, and the result is shown in the below graphs:

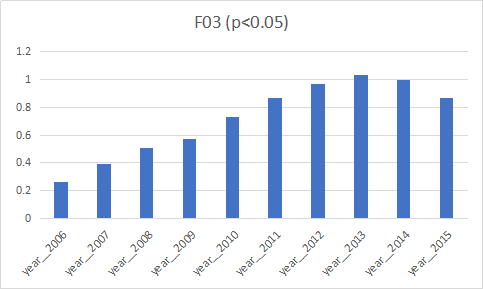


Figure 4.1: Year Coefficients across year for unspecified dementia

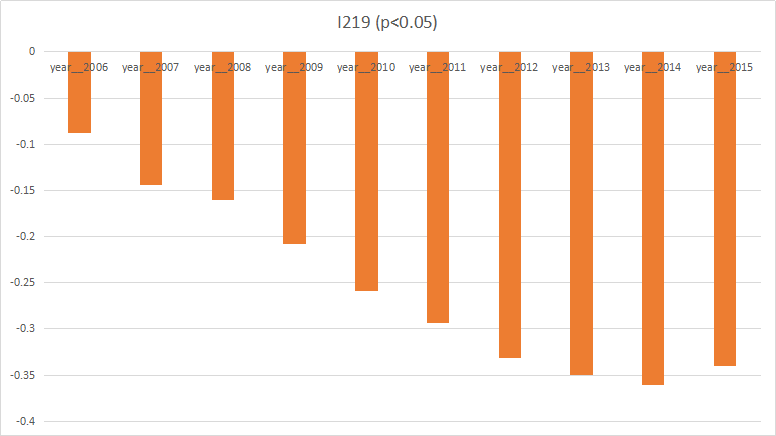


Figure.4.2: Year Coefficients across year for myocardial infarction

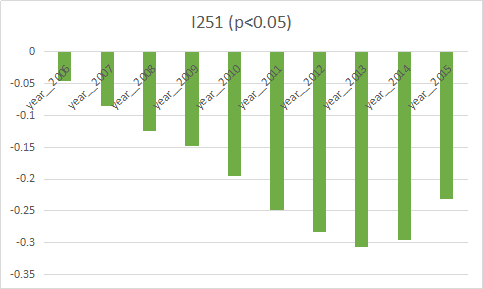


Figure.4.3: Year Coefficients across year for heart disease

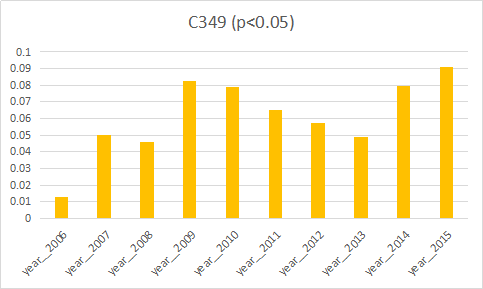


Figure.4.4: Year Coefficients across year for lung disease

Take heart disease for example, we could observe certain trend as year passed, possibility of being died due to heart disease increased from 2005 to 2013 and then slightly declined, which partially corresponded with real world data and proved the feasibility of our models.

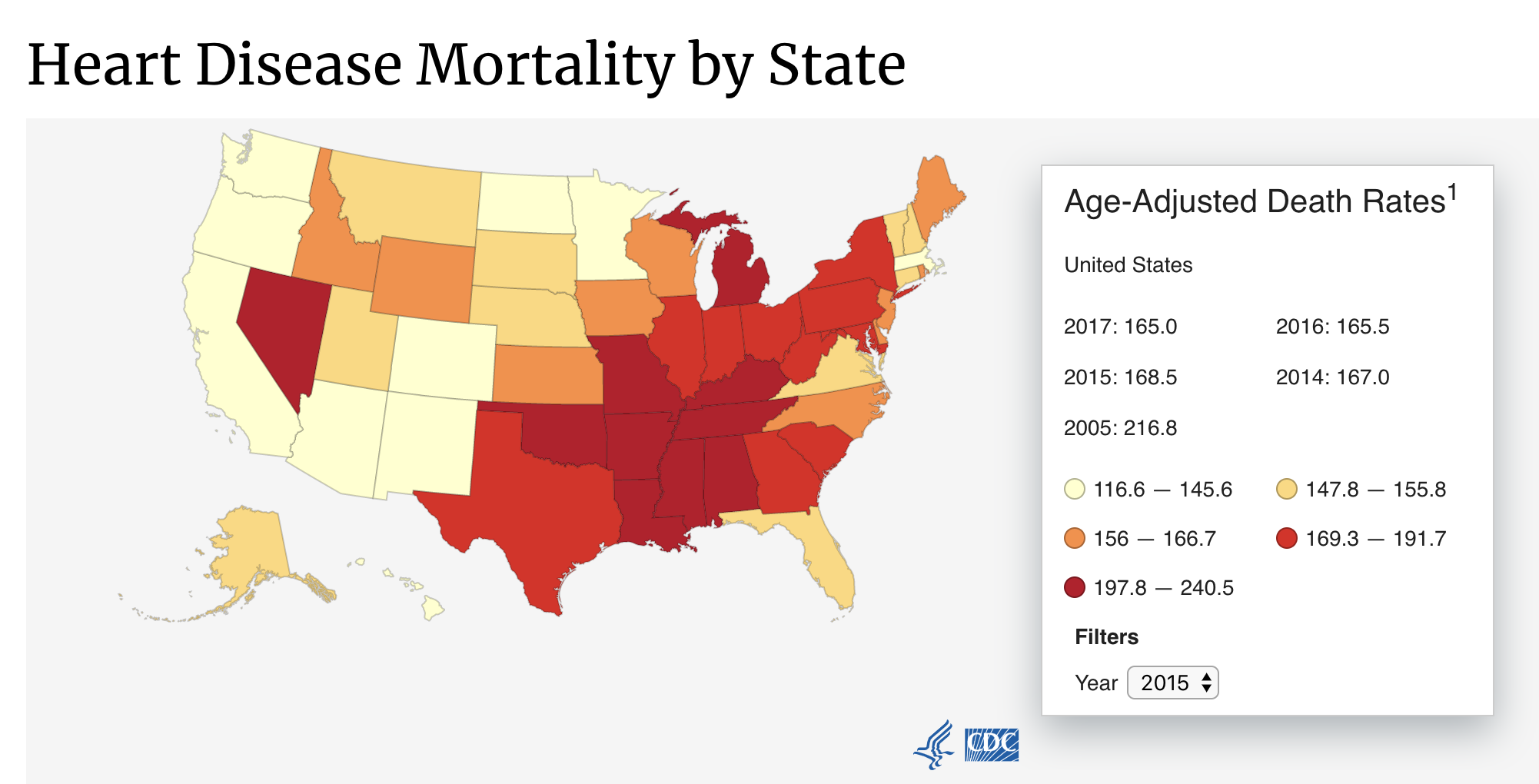


Figure.4.5: Nationwide heart disease death rate across year

Question 5

In this section, we got an Excel file containing medical transcriptions of 5,000 patients. We used natural language processing and cosine similarity to diagnose these 5,000 patients based on their descriptions. Two files were needed in order to solve question 5. The one is ‘medicaltranscription.csv’, which was already given. The other one was an Excel file which contains ICD type numbers with descriptions. We did have this file written manually since there were over 10 thousands of ICD codes, which took time to analyze all of them. Therefore, in order to simplify our work, we decided to divide all ICD codes into nearly 40 types based on the column “**39\_cause\_recode**” in the data of each year. Indexes were the types of ICD with “description” as the only column in this file. It looks like:

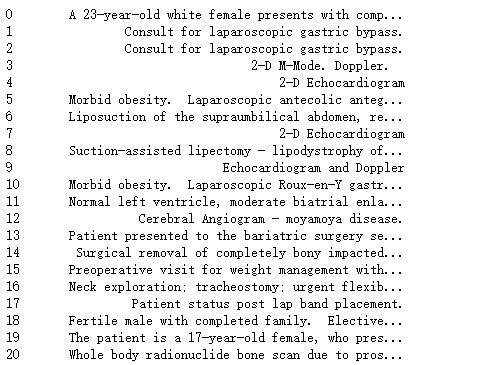
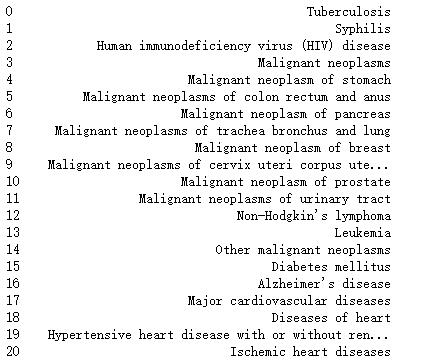


Figure 5.1 ICD Description Figure 5.2 Patient Description

There are four main steps in order to solve this problem. For step 1, we need to convert those transcriptions into vectors. Following some parts of the steps from Assignment 4 & 5, we used “word\_tokenize”function from nltk and convert the corpus into a list of words. Then instead of creating a vocabulary for all words, we applied POS-tag on each word and created a vocabulary of nouns only (i.e. NN, NNP,NNS, and NNPS). Next, we converted the “description’ of each case in “medicaltranscriptions.csv” and ICD description in “ICD-description.csv” to vectors using the pretrained word to vector dictionary “PubMed-and-PMC-w2v.bin”. The result looks like:

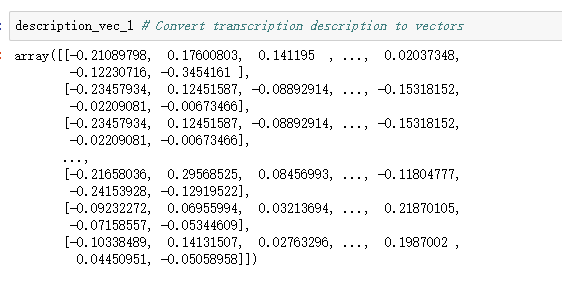
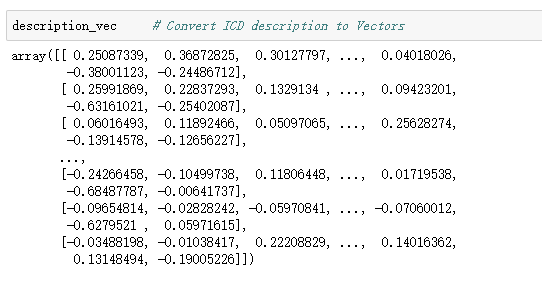


Figure 5.3 ICD Description to Vector Figure 5.4 Transcription Description to Vectors

For step 2, we calculated the similarity measure between ICD description and medical transcripts by importing “cosine\_similarity” from “sklearn.metrics.pairwise”:

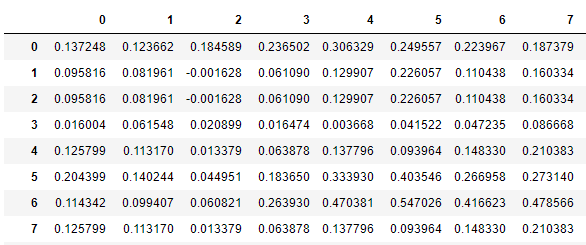


Figure 5.5 Similarity Measure Between ICD Description and Medical Transcripts

In figure 5.5, there are 5000 rows representing patients in the file and 42 columns representing ICD type numbers. We calculated the similarities of all ICD types for each patient. After getting this result, as step 3, we decided to get the maximum number of each row so that there should only be one ICD type for each patient, which furthest matches his/her physical description. Then we got the column names in which the maximum numbers of similarities locate and decoded them to the icd description:

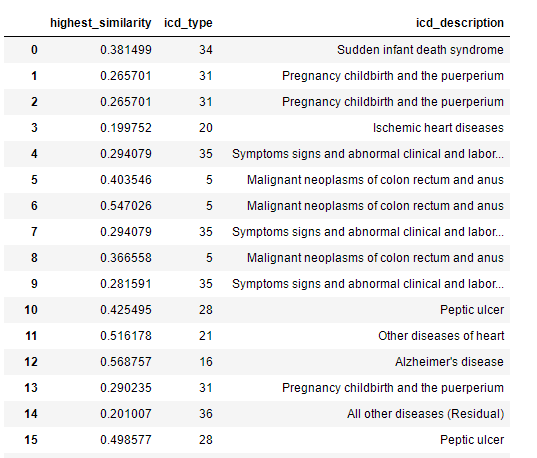


Figure 5.6 Max Similarity for Each Row with icd\_type Number and Description

Noticing that not all similarities are high enough, we considered that the accuracy could be guaranteed only when the similarity is above a certain threshold. Therefore, finally, we set our threshold to be 0.75 and then we assigned ICD descriptions to medical transcripts only if the similarity score is above 0.75:

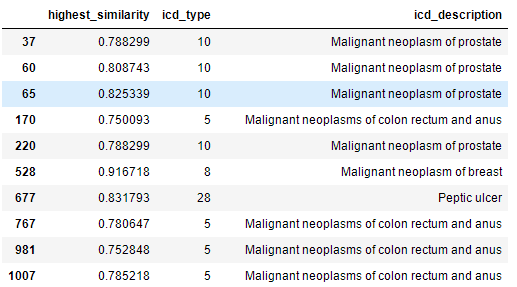


Figure 5.7 Result defined (highest\_similarity > 0.75)

As a result, of 5000 patients, there were only 56 of them whose similarities were above 0.75. To test the accuracy, we shuffled the result and pick the first 10 patients to manually check if the description matched ICD type description. The accuracy, surprisingly, is 100% (10/10).

# Interesting Findings

After we modeled the data using Logistic Regression and came up with the result, we found something interesting. We take model of “Unspecified Dementia” for example, first, noticing that the p-value of “Age” is lower than 0.05, we conclude that age is significant. Also, since the coefficient of age is 0.0888, there is a positive relationship between age and Unspecified Dementia, which is true because the elders are often at the highest risk of dementia.

Manner, however, turns out to be a bad predictor for this model because of the p-value. Only “Natural” is significant and we believe that it’s because nearly 78% of deaths are natural, which account for most of all manners of death.

For race and resident, the black people and intrastate residents have a positive relationship with dementia; asians/pacific islanders and foreign residents have negative relationship with dementia.

There’s another interesting fact on marital status. By looking at the coefficients, we were surprised to find that divorced people have the highest probability of getting “Unspecified Dementia” Oppositely, there’s a negative relationship between married people and “Unspecified Dementia” so we may conclude that it’s not that easy for married people to get dementia.

There’s no exaggeration to say that these interesting findings may help insurance companies reconsider their insurance plans for different groups of people with different medical conditions.

# Conclusion

According to the results, our group has several proposals towards this insurance plan.

1. First of all, we suggest that this insurance should not include infant or add extra charge to infant or children who are below the age of ten because children tend to have a higher death rate than the other age groups and there are various types of death that can cause that.
2. Talking about age groups, elders who are above the age of 100 should not be included in this insurance plan because they have the highest rate of natural death.
3. Since our insurance plan focuses on customers’ health status, any death that caused by suicide or accidents should not be included.
4. Despite the fact that we found a certain pattern that relates the death with natural month in a year, we decide not to use this as a standard for this insurance plan because we cannot predict the date of death.

# Appendix

See attachment