NATIONAL UNIVERSITY OF SINGAPORE

SCHOOL OF COMPUTING

BT1101 - BUSINESS ANALYTICS

GROUP PROJECT

DEATH RATES IN USA (2014)

STUDENT NAME	MATRICULATION NUMBER
LEE SHI MING	A0158677J
NG AI HIANG	A0157049Y
SARAH GUI POH CHENG	A0160664E
SIM TENG KUAN JOEL	A0162624J
TEO JIN MING	A0155030X

Table of Contents

Table of Figures

1. Motivation of Analytics

a. Main Objective

The main objective of our analysis is to provide insights to governments and insurance companies so that these parties will have a clearer understanding on the factors affecting life expectancy, or the age an American will die, in the United States of America (USA). The descriptive analysis in our report aims to show the relation between the variables (i.e. gender and age of death) of the deceased in USA. The predictive analysis in our report aims to show which variables or sectors (education, race, resident status, and healthcare) are significant for policy makers to take note of, in relation to life expectancy.

b. Justification and Relevance of Analysis

There is a saturation of data in the world today, and effective use of such data can provide more meaning to the data and is a powerful tool in our informative world. This report will be especially useful to the government of USA. They will have a clearer visual of which variables can potentially affect the life expectancy of their citizens and which areas they should allocate a higher budget for in terms of education, healthcare, or addressing racial disparity. These efforts can then increase the social welfare and have dynamic effects on the country. On the other hand, insurance companies are able to utilize the analysis and tailor their products or policies to their advantage, generating higher profits. After all, insurance companies are profit driven.

2. Data Sources

a. Source of Data

The data was taken from Kaggle[1], titled 'Death in the United States'[2]. The number of observations was scaled down to 10,000, using random sampling. Random sampling was performed using the formula (=RAND()) to generate random numbers for all the observations, and sort it from largest to smallest. From there, we took the first 10,000 observations and used it as a dataset for our analysis in this project.

b. Justification and Suitability of Data Collected

Kaggle is the world's largest community of data scientists[3] *(Dean Dacosta, 2015*). It is a platform which allows data scientists to share their ideas and compete with one another to solve complex data science problems. It also provides a wide array of datasets, for data scientists to work on. Kaggle is a credible website to source datasets from[4] *(Dean Dacosta, 2015)*, and it is suitable as the origin of our dataset is from the USA government health records. There are many variables available in the dataset which we hope to find some relationships, and give us more interesting and better analysis in this project.

c. Unbiased Data

Statistical bias in choosing subjects occurs when there is a bias in selection of sample data. Bias includes choosing non-random data for statistical analysis which limits the ability of the result to be generalised to the rest of the population dataset.

Mean Age	POPULATION	SAMPLE	DIFFERENCE (%)
ALL GENDERS	73.3	73.0	0.49%
FEMALE	76.5	76.3	0.29%
MALE	70.3	69.9	0.54%

Table 2.1 Mean Age Comparison for Population and Sample

To test for biasness in our dataset after random sampling was performed, we compared the mean age of death from the sample and population. As shown in Table 2.1, the population mean age of death is 73.3 years old, whereas the sample mean age of death is 73.0 years old. To further check for biasness, we went on to compare the mean age of death for females and males for both population and sample. The difference in the mean age for population and sample is relatively small, with 0.49% difference for all genders, 0.29% difference for females and 0.54% for males.

Resident_Status	POPULATION	SAMPLE	DIFFERENCE (%)
RESIDENTS	83.43%	83.73%	0.30%
INTRASTATE NON-RESIDENTS	13.54%	13.32%	0.22%
INTERSTATE NON-RESIDENTS	2.77%	2.71%	0.06%
FOREIGN RESIDENTS	0.26%	0.24%	0.02%

Table 2.2 Resident Status Comparison for Population and Sample

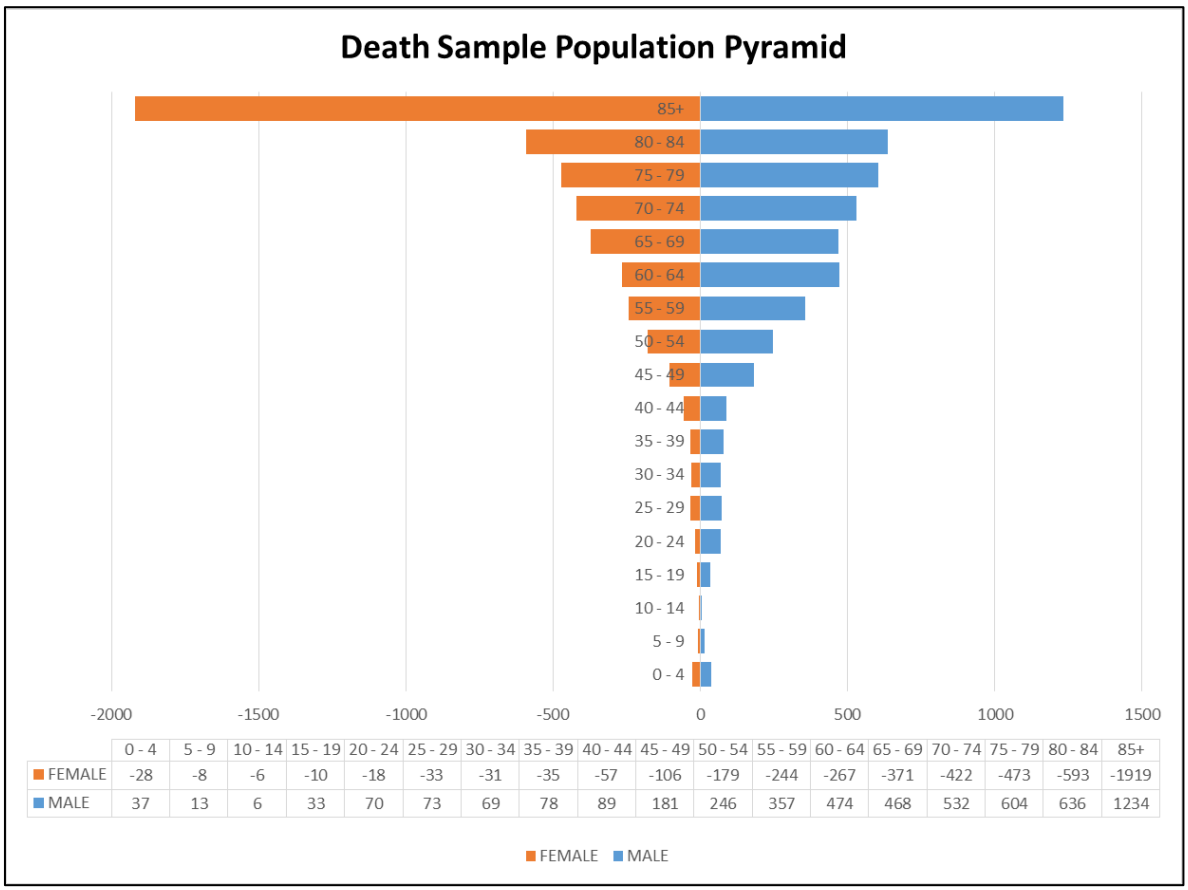
Beyond that, we again tested for biasness in the random sample dataset by comparing the percentage of resident status between the population and sample. The difference in the percentage for resident status between population and sample is also relatively small, ranging from 0.02% to 0.30% difference.

Therefore, from the above two tests, we are able to conclude that the data collected from random sampling is unbiased.

3. Descriptive Analytics

a. Chart Analysis

i. Females have Higher Life Expectancy



	0 - 4	5 - 9	10 - 14	15 - 19	20 - 24	25 - 29	30 - 34	35 - 39	40 - 44	45 - 49	50 - 54	55 - 59	60 - 64	65 - 69	70 - 74	75 - 79	80 - 84	85+
FEMALE	-28	-8	-6	-10	-18	-33	-31	-35	-57	-106	-179	-244	-267	-371	-422	-473	-593	-1919
MALE	37	13	6	33	70	73	69	78	89	181	246	357	474	468	532	604	636	1234

Figure 3.1 Death Sample Population Pyramid

Figure 3.1 shows the mortality population pyramid for the United States in 2014. The population pyramid is a representative of a developed country, where life expectancy is high due to higher standards of living. The chart shows that females generally have a higher life expectancy than males, where the orange chart is significantly longer than the blue chart at age 85 and over. Also, another significant age group is the age between 0 and 4, which it shows that the mortality rate for children below 4 years old is comparatively high for a developed country. According to The Washington Post, the higher infant mortality rate in the United States is "almost entirely due to the high mortality among the less advantaged groups"[5] *(Christopher Ingraham, 2014)*.

ii. Higher Mortality Rate for Ages 0-4 Years American Indians

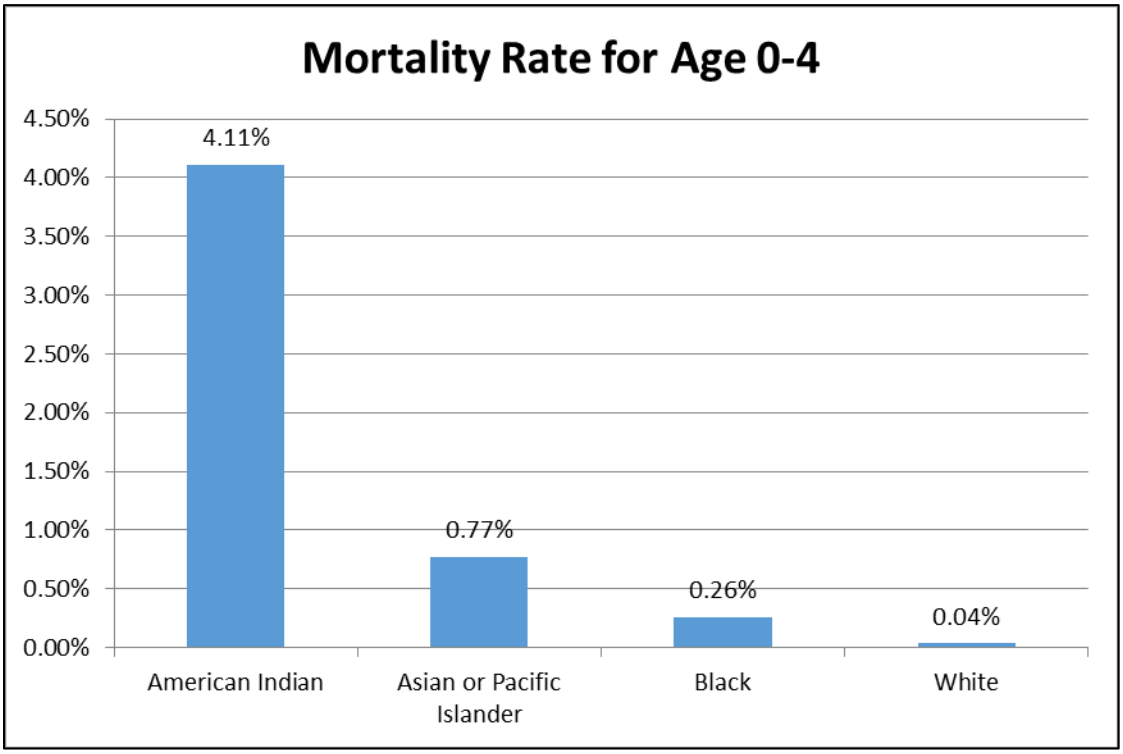


Figure 3.2 Mortality Rate for Age 0 to 4

To further explore the previous trend identified - that mortality rate for ages 0 to 4 year old is high in the USA, we divided the data based on race. As shown in Figure 3.2, the mortality rate is the highest among American Indian, at 4.11%, more than 4 times the rate for the next highest race – Asian or Pacific Islanders, at 0.77%. The Whites, on the other hand, have the lowest mortality rate of 0.04%. This is equivalent to saying that an American Indian child is 11,453 times more likely to die within the age of 4 as compared to a White child. Rates were used instead of absolute values to account for the vast difference in the numbers in each race group (Whites, American Indians, et cetera).

iii. Married Individuals have Longer Life Expectancy

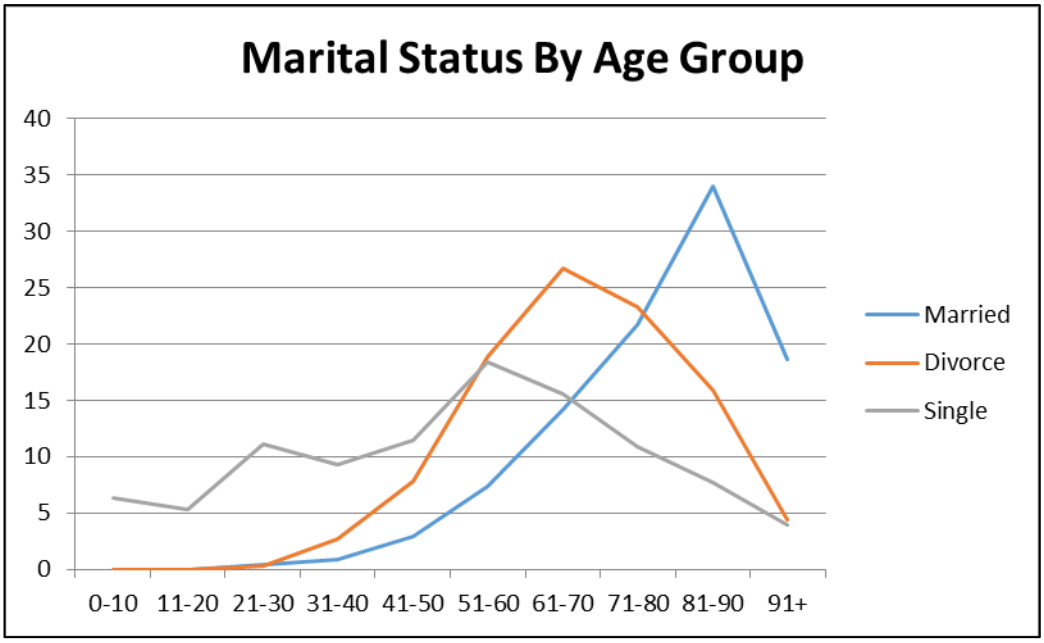


Figure 3.3 Marital Status by Age Group

Figure 3.3 shows the death statistics for three main groups of Americans: the married, divorced and singles. The line graph clearly shows 3 different peaks, the peak death age for singles is the youngest, at around 53 years old, followed by divorcees, at around 65 years old and lastly, married individuals with the highest peak with the death age at about 84 years old. Hence, we can infer that marriage can possibly increase life expectancy since married individuals tend to have a higher mortality age. This may be due to qualitative factors such as having social support, and a higher combined income leading to better quality of life.

iv. Common Diseases Contracted

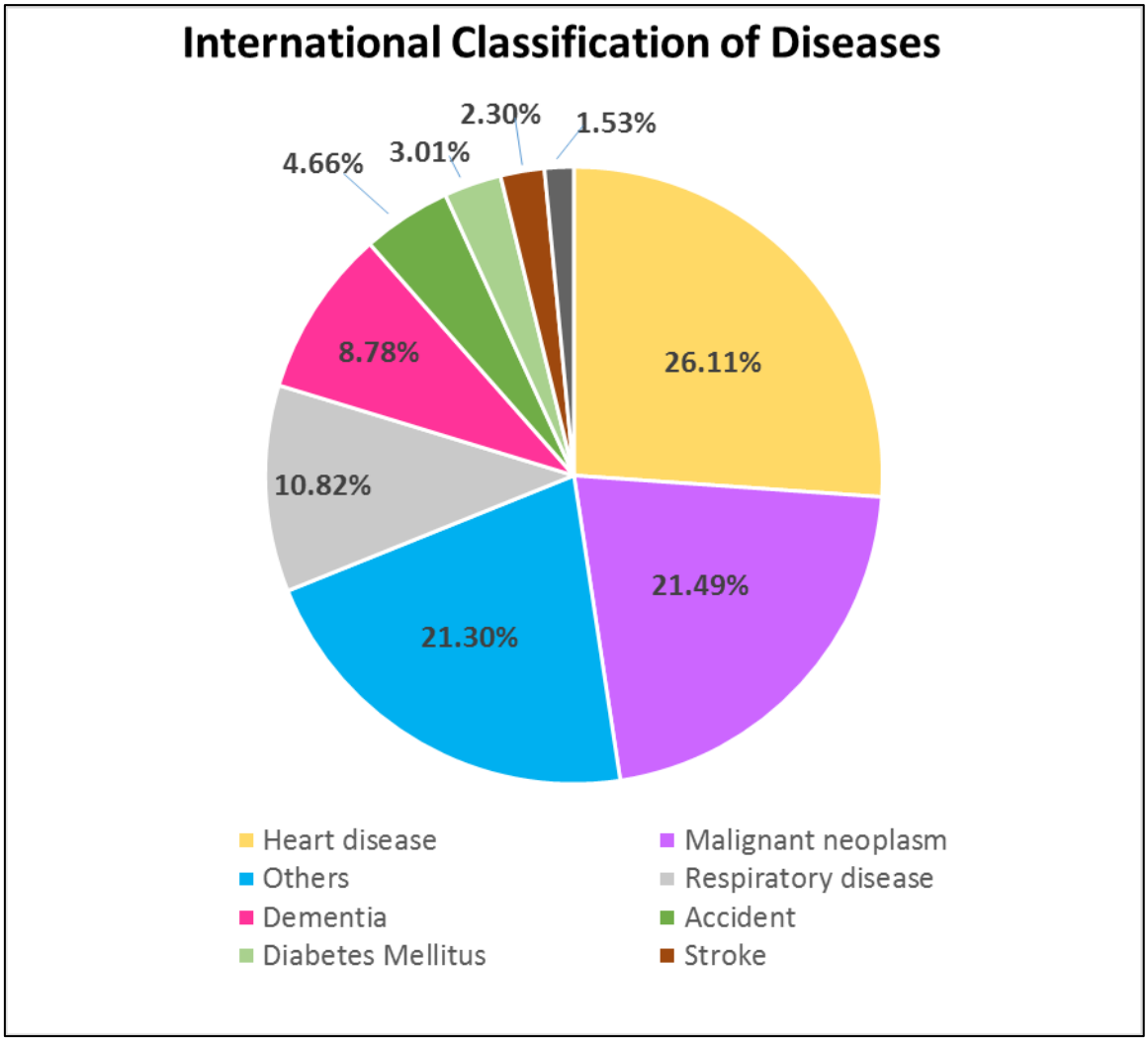


Figure 3.4 International Classification of Diseases

Figure 3.4 shows the percentage of various diseases classified among Americans in terms of the International Classification of Diseases (ICD). Based on the chart above, the highest ICD from the sample is heart disease, at 26.11%, followed by malignant neoplasm at 21.49%. The third highest ICD with 21.30% represents other types of diseases, including tooth disease, skin disease, and more. Based on the Centers for Disease Control and Prevention, every 1 in 4 deaths in USA is caused by heart disease[6] *(Heart Disease Fact Sheet, 2016).*This source also mentioned that heart disease is the leading cause of death for both men and women in USA[7] *(Heart Disease Fact Sheet, 2016)*. Therefore, it supports our findings that heart disease is the leading cause of death in the USA.

v. Breakdown of Heart Disease Cases into Age Groups

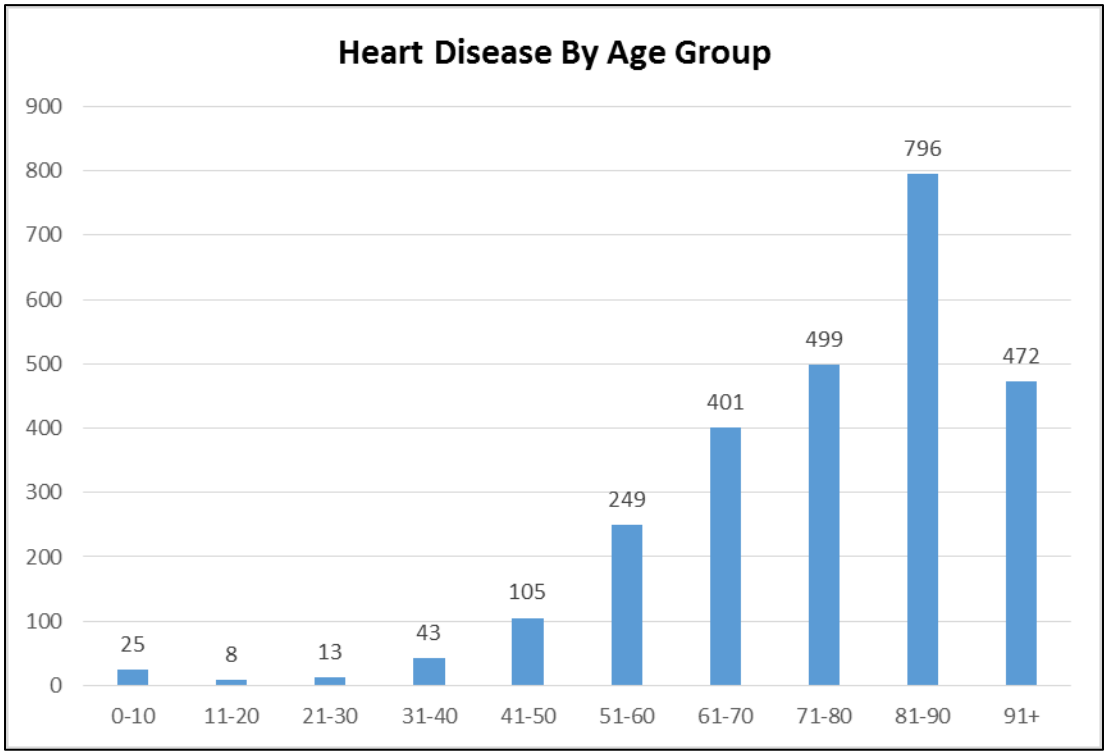


Figure 3.5 Heart Disease By Age Group

Figure 3.5 shows the percentage of ICD cases for each age group. Since heart disease is the most common ICD (as seen from Figure 3.4), we want to probe deeper and find out at what age range do Americans contract this disease. From the chart above, we see that the number of heart disease cases is the highest at the age range of 81 to 90 years old. Additionally, Americans above 40 years old are more likely to get heart disease as the number of cases increases gradually from the age range 41-50 and beyond.

vi. Difference in Diseases Contracted by Males and Females

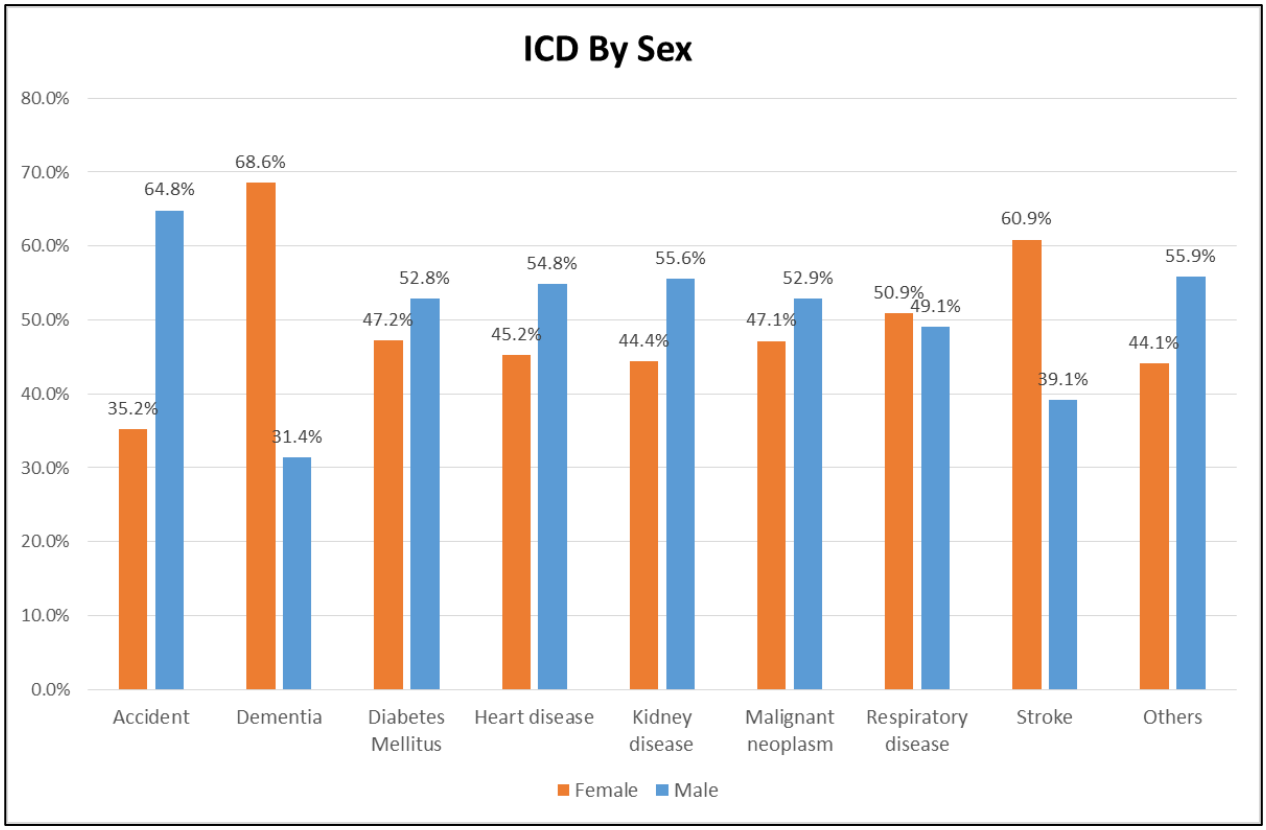


Figure 3.6 ICD By Sex

Figure 3.6 shows the comparison between the genders for the different categories under the ICD, of which, there is a significant difference of at least 20% in the percentage of males and females in cases of accidents, dementia and stroke. For accident cases, the percentage of males who died from accident is at 64.8%, almost twice the percentage of females who died from accident, at 35.2%.

Based on Alzheimers.net, a website for dementia care in the USA, 1 in 9 Americans aged 65 and above has Alzheimer's disease, also known as dementia[8] *(2016 Alzheimer's Statistics, n.d.)*. For dementia cases, about 13.8% out of the death population aged 70 and above suffered from dementia. From Figure 3.6, 68.6% of people with dementia are women and 31.4% are men. This is mainly because women tend to live longer than men (as seen from Figure 3.1), and the likelihood of developing dementia increases with age.

As for stroke cases, 60.9% of the death population with stroke are females and 39.1% are males. Based on the Centers for Disease Control and Prevention, stroke is the fifth leading cause of death in the United States[9] *(Stroke, 2016)*. From Figure 3.4, it can be seen that stroke is one of the diseases that is relatively more common among the people.

vii. Whites have the Highest Life Expectancy

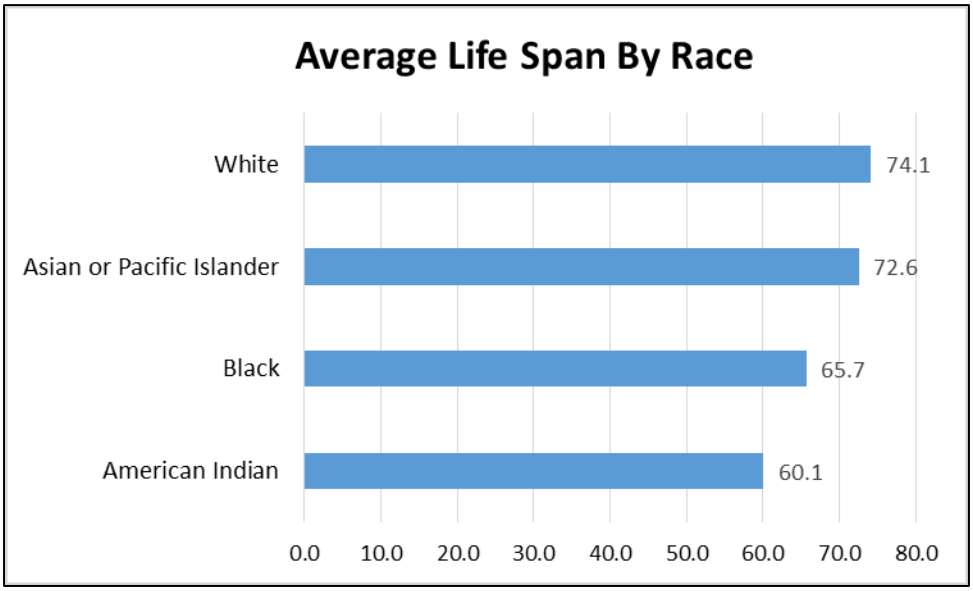


Figure 3.7 Average Life Span By Race

Figure 3.7 shows the average life span of the Americans in the sample, grouped according to their race, namely White, Black, Asian or Pacific Islander and American Indian. The American Indians generally have the shortest average life span of 60.1 years old while Whites generally have the longest average life span of 74.1 years old. This can probably be attributed to the Whites having better access and ability to afford healthcare services due to their higher socio-economic status.

b. Descriptive Statistics

Row Labels	Count of Sex
F	4800
M	5200
Grand Total	**10000**

Row Labels	Average of Age
F	76.30
M	69.91
Grand Total	**72.9786**

Row Labels	Count of Age Range
0-10	87
11-20	73
21-30	188
31-40	229
41-50	489
51-60	1071
61-70	1638
71-80	2048
81-90	2747
91+	1430
Grand Total	**10000**

Age	
Mean	72.9786
Standard Error	0.181919388
Median	77
Mode	85
Standard Deviation	18.19193878
Sample Variance	330.9466367
Kurtosis	1.575046664
Skewness	-1.14686161
Range	112
Minimum	1
Maximum	113

Figure 3.8 Descriptive Statistics for Age

The mortality sample of the United States in 2014 consists of 4,800 females and 5,200 males. In terms of age of death, majority of the people falls in the range of 81 years old to 90 years old, followed by in the range of 71 years old to 80 years old. The mean age of death in the sample is 72.98 years old, in which females have a higher average death age of 76.3 years old as compared to the males, who have a lower average death age of 69.91 years old.

Row Labels	Count of Resident_status
RESIDENTS	83.73%
INTRASTATE NONRESIDENTS	13.32%
INTERSTATE NONRESIDENTS	2.71%
FOREIGN RESIDENTS	0.24%
Grand Total	**100.00%**

Row Labels	Count of Method_of_disposition
Cremation	52.29%
Burial	44.26%
Burial at sea	2.51%
Donation for study	0.94%
Grand Total	**100.00%**

Row Labels	Count of Hispanic_origin
Non-Hispanic white	75.11%
Non-Hispanic black	11.63%
Mexican	5.03%
Non-Hispanic other races	4.53%
Cuban	1.31%
Puerto Rican	0.81%
Central or South American	0.81%
Other or unknown Hispanic	0.51%
Hispanic origin unknown	0.26%
Grand Total	**100.00%**

Row Labels	Count of Education
High school graduate or GED completed	40.29%
Some college credit, but no degree	14.54%
8th grade or less	12.17%
Bachelor's degree	10.41%
9 - 12th grade, no diploma	9.80%
Associate degree	5.31%
Master's degree	4.01%
Unknown	1.91%
Doctorate or professional degree	1.56%
Grand Total	**100.00%**

Figure 3.9 Descriptive Statistics for other variables

c. Nature of Data

The nature of our dataset lies towards categorical, where the majority of our variables are categorical and only one variable is a numerical variable which is the age of death.

The variables of our dataset and its description is summarized in the table as follow:

Variable Name	Description	Values
ID	A unique identifier for the record	Integer (any values)
ResidentStatus	A code to identify the resident status of the person	Integer (1 to 4) 1 – Residents 2 – Intrastate Resident 3 – Interstate Resident 4 – Foreign Residents
EducationCode	A code to identify the education level of the person	Integer (1 to 9) 1 – 8th grade or less 2 – 9 – 12th grade, no diploma 3 – High school graduate or GED completed 4 – Some college credit, but no degree 5 – Associate degree 6 – Bachelor's degree 7 – Master's degree 8 – Doctorate or professional degree 9 – Unknown
MonthOfDeath	A value to identify the month of death of the person	Integer (1 to 12) 1 being January and 12 being December
DayOfWeekOfDeath	A value to identify the day of the week of death of the person	Integer (1 to 7) 1 being Sunday and 7 being Saturday
Sex	A code to identify the gender of the person	String M – Male F – Female
Age	A value to identify the age of the person	Integer (any values)
Age Range	A category to classify the age of the person	Range of values (0-10, 11-20, 21-30, etc)
PlaceOfDeathAndDecedentsStatus	A code to identify the place of death of the person	Integer (1 to 7) 1 – Hospital, clinic or Medical Center – Inpatient 2 – Hospital, clinic or Medical Center – Outpatient or admitted to Emergency Room 3 – Hospital, clinic or Medical Center – Dead on Arrival 4 – Decedent's home 5 – Hospice facility 6 – Nursing home/long term care 7 – Other

MaritalStatus	A code to identify for the marital status of the person	String (M, S or D) M – Married S – Single D – Divorced
InjuryAtWork	A code to identify whether the person is injured at work	String (Y, N, U) Y – Yes N – No U – Unknown
MannerOfDeath	A code to identify the manner of death of the person	Integer (0 to 4) 0 – Natural 1 – Accident 2 – Suicide 3 – Homicide 4 – Pending investigation
MethodOfDisposition	A code to identify the method of disposition of the person	String (B, C, D or S) B – Burial C – Cremation D – Donation for study S – Burial at sea
Autopsy	A code to identify if autopsy is done for the person	String (Y, N, or U) Y – Yes N – No U – Unknown
PlaceOfInjury	A code to identify the place of injury	Integer (0 to 8) 0 – Home 1 – Residential institution 2 – School, other institution and public administrative area 3 – Sports and athletics area 4 – Street and highway 5 – Trade and service area 6 – Industrial and construction area 7 – Farm 8 – Others
Icd10Code	A modified code to identify the diseases of the person	Integer (0 to 8) 0 – Others 1 – Accident 2 – Dementia 3 – Diabetes Mellitus 4 – Heart disease 5 – Stroke 6 – Kidney disease 7 – Malignant neoplasm 8 – Respiratory disease
RaceCode	A code to identify the race of the person	Integer (1 to 4) 1 – White 2 – Black 3 – American Indian 4 – Asian or Pacific Islander

HispanicOriginRaceRecode	A code to identify the Hispanic origin of the person	Integer (1 to 9) 1 – Mexican 2 – Puerto Rican 3 – Cuban 4 – Central or South American 5 – Other or unknown Hispanic 6 – Non-Hispanic white 7 – Non-Hispanic Black 8 – Non-Hispanic other races 9 – Hispanic origin unknown

Table 3.1 Dataset Variables

4. Predictive Analytics (Regression)

a. Dependent and Independent Variables

Multiple linear regression was conducted to predict the age of death of a person given the different attributes of the person. Looking at how insurance coverage is segmented[10] *(Jessica C. Barnett, 2016)*, we decided to look into individuals aged 19 and above. In addition, since accidents, homicide and suicides are random incidents, we decided to only focus on "Natural" cases under the "MannerOfDeath" variable. This is because we are only able to predict the age of natural death while deaths resulting from random incidents cannot be predicted.

```
data <- data[Manner_of_death == "Natural" & Age > 18]
```

Next, we have shortlisted 6 variables out of the remaining 15 variables to be included in our regression model as independent variables. The dependent and independent variables for the regression model are as follow:

Dependent variable: Age of death

Independent variables: Resident Status, Marital Status, Gender, Education Level, Race and Hispanic Origin

The reason for shortlisting the above 6 independent variables is mainly because we believe that these 6 variables can possibly affect the death of age of a person. The rest of the variables such as month of death, day of death, place of death, method of disposition, autopsy, place of injury, the type of disease a person has (ICD) and whether the person is injured at work (InjuryAtWork) are excluded from our regression model as it does not help us to determine the age of death of a person as all these variables can only be derived upon the death of a person. Hence, we are focusing on what determines the age of death of a person so that government agencies and relevant policy makers can act accordingly.

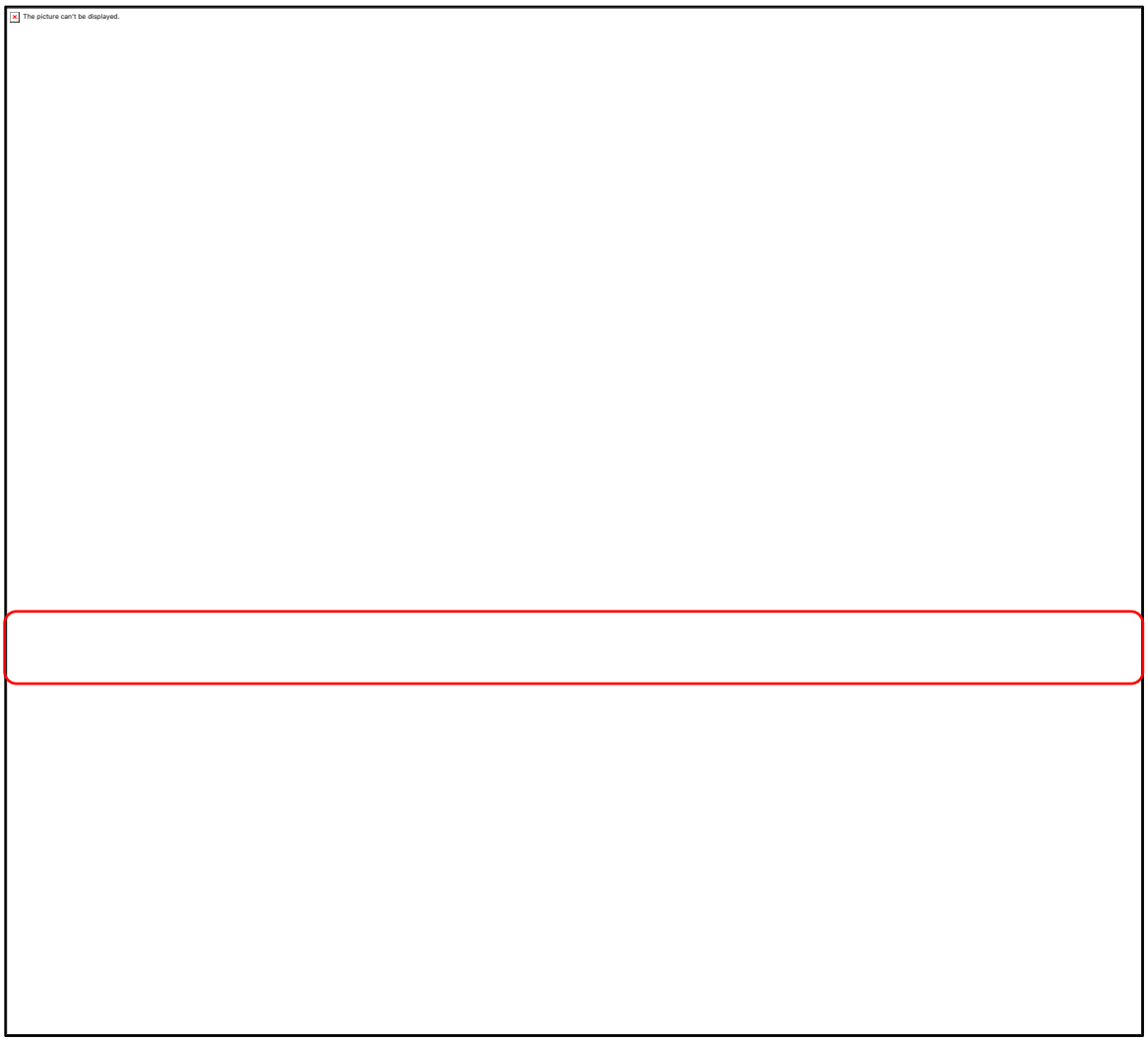


Figure 4.1 First Regression Analysis

After running the first regression with all the above mentioned 6 independent variables using R Studio, the p-value for all the race variables were greater than 0.05 (refer to the red box). This means that the race variable is not statistically significant at 95% confidence level and thus, we removed it from our regression model. Hence, our regression model now comprises the dependent variable and 5 independent variables, as follow:

Dependent variable: Age of death

Independent variables: Resident status, Marital Status, Gender, Education Level and Hispanic Origin

b. Causality

Our analysis above does not prove any causality but we logically infer that a causal relationship exists between the variables. For instance, we know that females generally have a higher life expectancy than males, hence we expect a correlation between the gender and age of death. This can be supported by our regression results below.

c. Interpretation

We decided to run stepwise regression based on Akaike Information Criterion (stepAIC) using R Studio to get a better fit of a linear equation to the sampling population, as shown in Figure 4.2 Step AICFigure 4.2 below.

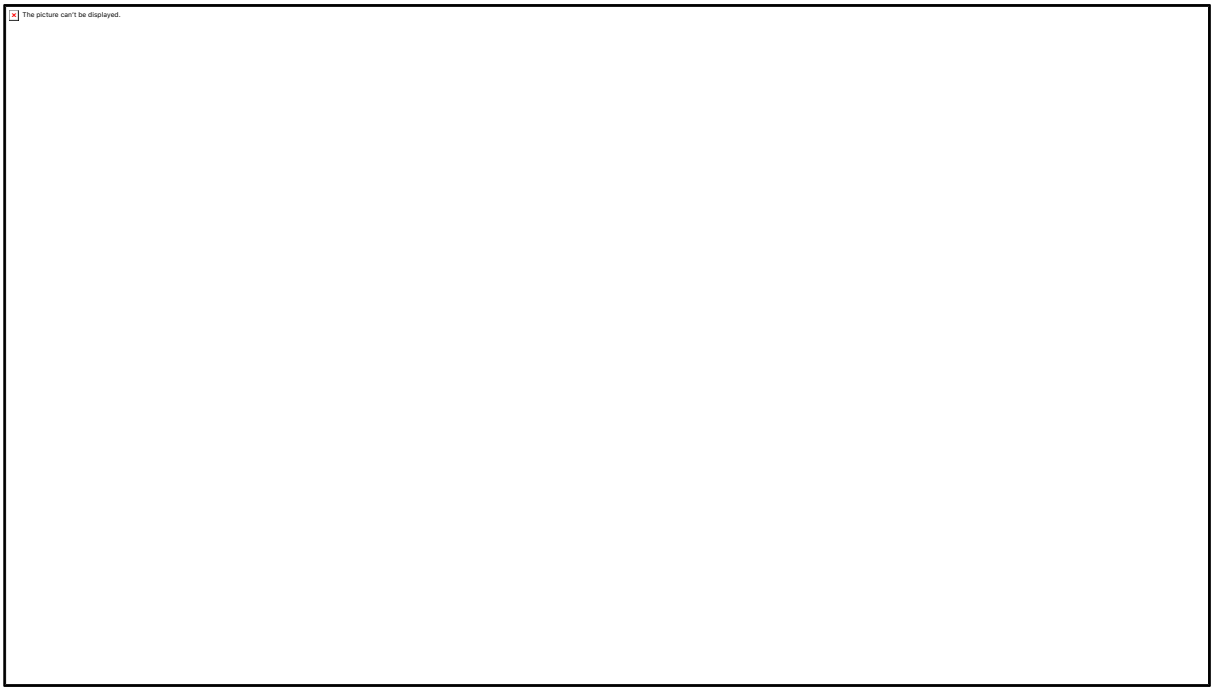


Figure 4.2 Step AIC

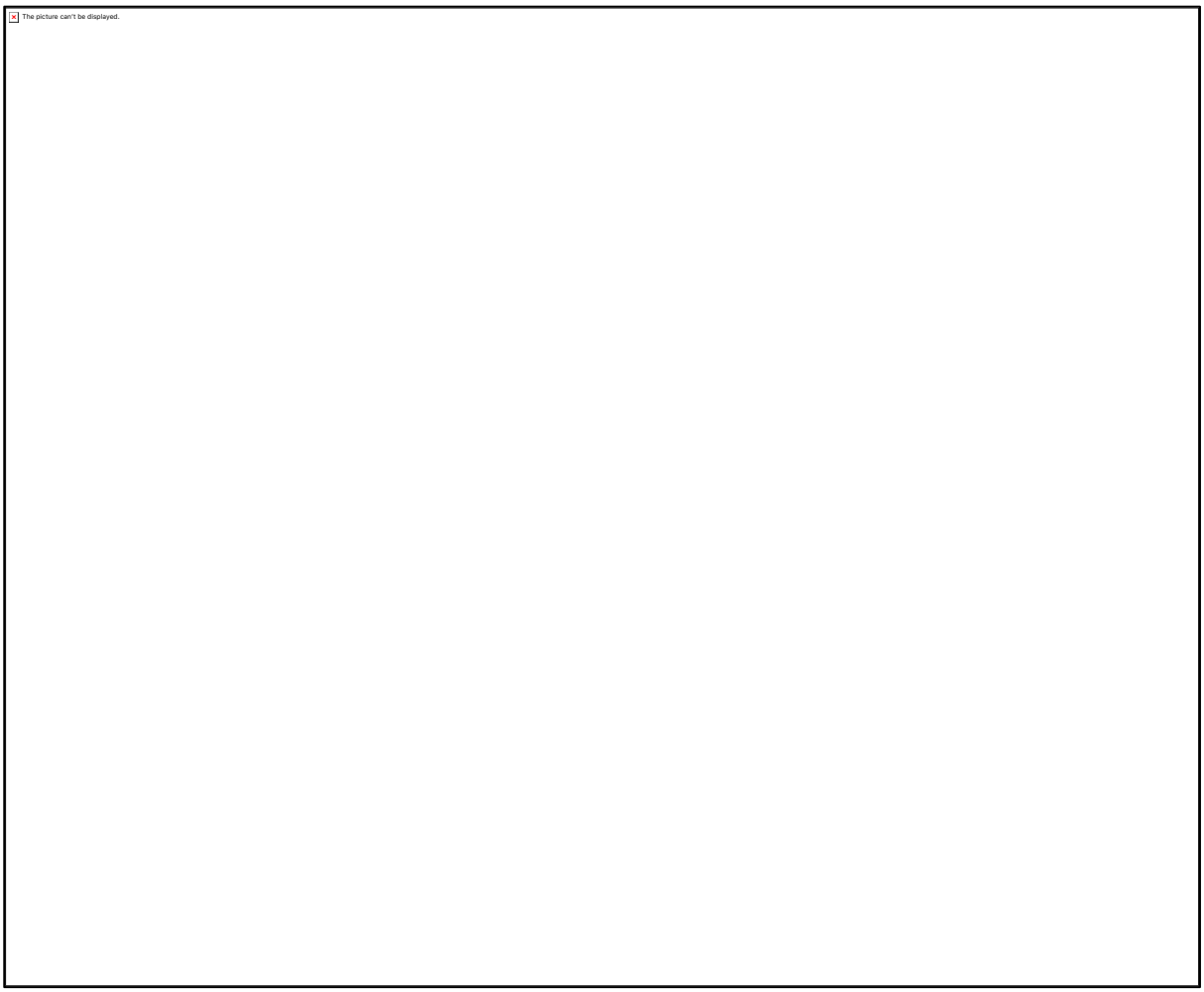


Figure 4.3 Final Regression Analysis

After removing all the statistically insignificant variables, we are now left with most of the variables that are statistically significant, except for Hispanic_origin_unknown, Hispanic_originMexican, other or unknown Hispanic, and unknown education. Since these variables are categorical variables, we cannot remove the individual variable. Although our R-squared values is not very high, it is still acceptable given that our project is one of a social science application where there is bound to be many extraneous variables that we are unable to control for. The equation of the regression model is as follow:

Base Case: Foreign Resident, 8th Grade or less, Female, Divorced, Central or South American

Age = 61.5 + 1.58 Interstate Nonresidents + 3.50 Intrastate Nonresidents + 7.50 Residents – 5.37 12th grade, no diploma – 5.94 Associate Degree – 4.15 Bachelor's Degree – 2.36 Doctorate or Professional degree – 4.98 High school grad/ GED completed – 4.37 Master's degree – 6.89 College credit but no degree – 0.919 Unknown – 3.47 SexM + 8.94 Married – 7.94 Single + 9.24 Cuban + 5.88 Hispanic_Unknown + 2.11 Mexican + 4.14 Non-Hispanic Black + 6.44 Non-Hispanic Other Races + 8.94 Non-Hispanic White + 2.04 Other or unknown Hispanic + 5.17 Puerto Rican

Our regression analysis backs our causality, whereby we see that a female generally lives about 3.5 years (3.4699) longer than a male.

The Standard Error (Std. Error) in the regression output measures the average amount that the coefficient estimates vary from the actual average value of our dependent variable. Ideally, the lower number is preferred. From our regression analysis, the age can vary by 0.28 years old for sex variable, and so on for the rest of the independent variables.

The coefficient t-value is a measure of how many standard deviations our coefficient estimate is far away from 0. Ideally, we would want the t-value to be as far away from 0 as possible. The t-value in the regression output is also used to calculate the p-value.

As for Pr (>|t|), it refers to the probability of observing any value equal or larger than the absolute value of t (|t|). Depending on the significance level of the test, the variable which exceeds the significance level will be deemed at statistically insignificant. An alternative way to determine if the variable is statistically significant is the asterisk (*) sign at the end of each variable row. The asterisks represent the significant code, in which three asterisks refer to a highly significant p-value. In our case, we are testing at 95% confidence level, which means any p-value that exceed 0.05 will be deemed as statistically insignificant.

The residual standard error is a measure of the quality of a linear regression fit. The Residual Standard Error is the average amount that the actual age will deviate from the true regression line. In this case, the actual age can deviate from the true regression line by 13.14 years old, with 9,074 degrees of freedom.

d. Evaluation

The hypothesis test associated with regression analysis is predicated on the following key assumptions: linearity, normality of errors, homoscedasticity, multicollinearity and independence of error.

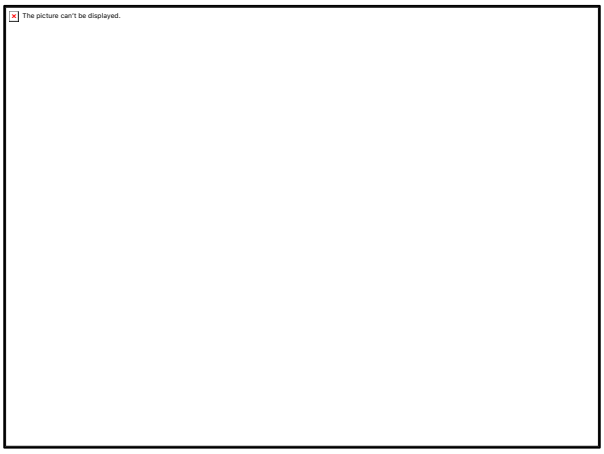


Figure 4.4 Residuals vs Fitted Plot

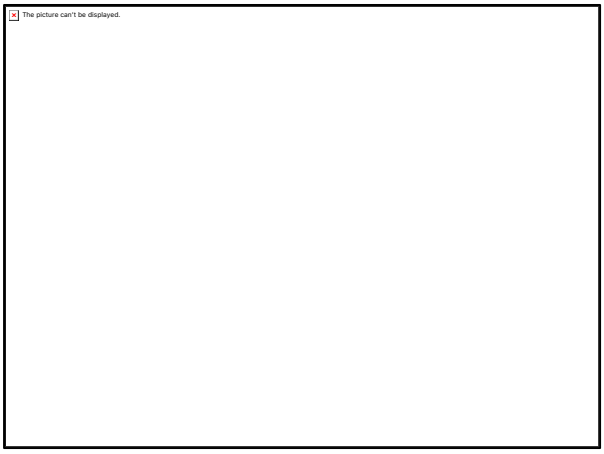


Figure 4.5 Scale Location Plot

The plot in Figure 4.4 tests for assumptions for linearity and homoscedasticity. Since the red line is straight, the assumption for linearity has been met. Based on Figure 4.4 and Figure 4.5, the data points appear evenly distributed and there is no distinct change in variance. Hence, the assumption for homoscedasticity is considered valid.

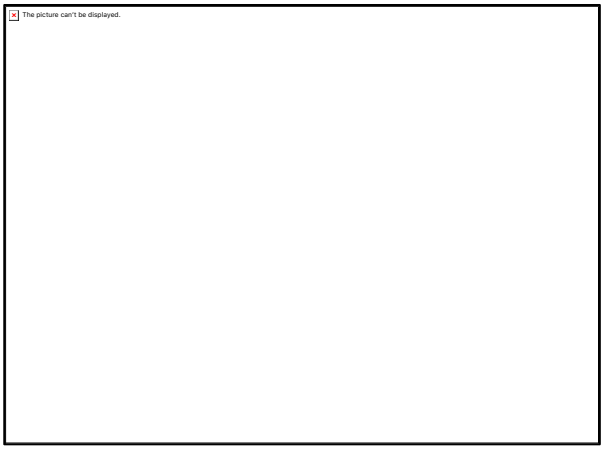


Figure 4.6 Normal Q-Q Plot

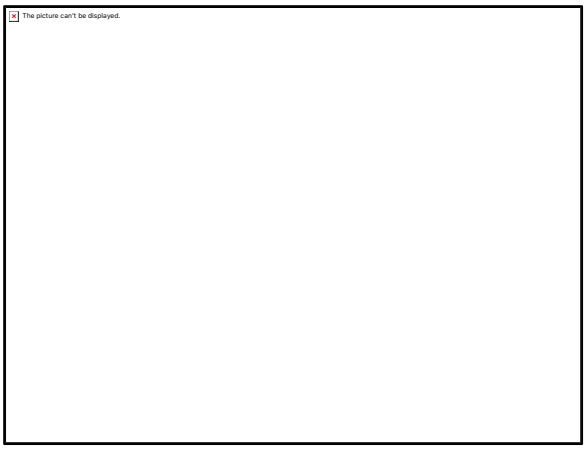


Figure 4.7 Residuals vs. Leverage Plot

The plot in Figure 4.6 is used to test for normality. Normality of errors refers that the errors for each individual value of X is normally distributed, with a mean of 0. Since the data points on the Q-Q plot are generally closely aligned to the line on the chart, we can hold the assumption for normality and we can assume that the data points follow a normal distribution.

The last plot in Figure 4.7 measures the Cook's distance which is the acceptable range for outliers. As the red dotted line that represents the Cook's distance is outside the range of the plot above, we conclude that there are no potential outliers in the regression model.

Figure 4.8 Variation Inflation Factor (VIF) Test

Lastly, to ensure that no two variables in our model are highly correlated, the Variance Inflation Factor (VIF) test was carried out. Independence of errors means that the distribution of errors is random and is not influenced by or correlated to the errors in prior observations. Since all the VIF values are below 10, as seen in Figure 4.8, the regression assumption for multicollinearity is considered valid.

In conclusion, all the key regression assumptions - linearity, normality of errors, homoscedasticity, multicollinearity and independence of error - were met for our regression analysis.

5. Predictive Analytics (Data Mining)

a. Data Mining Objectives

Our data mining objectives are to observe if the variables can be grouped into clusters, and if a particular cluster may lead to a higher or lower life expectancy. Also, since heart disease is the most common disease contracted, we hope to segment this group of deceased and observe if there are any similarities among them. By doing so, the government and insurance companies can gain insights and cater policies that better fit each segment.

b. Justification

We conducted k-means clustering because our data contains 10,000 observations while hierarchical clustering only supports data with less than 5,000 observations. The elbow method was first conducted to determine the optimal number of clusters before clustering is done. Association rule approach was also used to observe if any statistical link can be seen between variables.

c. Findings and Implications

i. K-Means Clustering

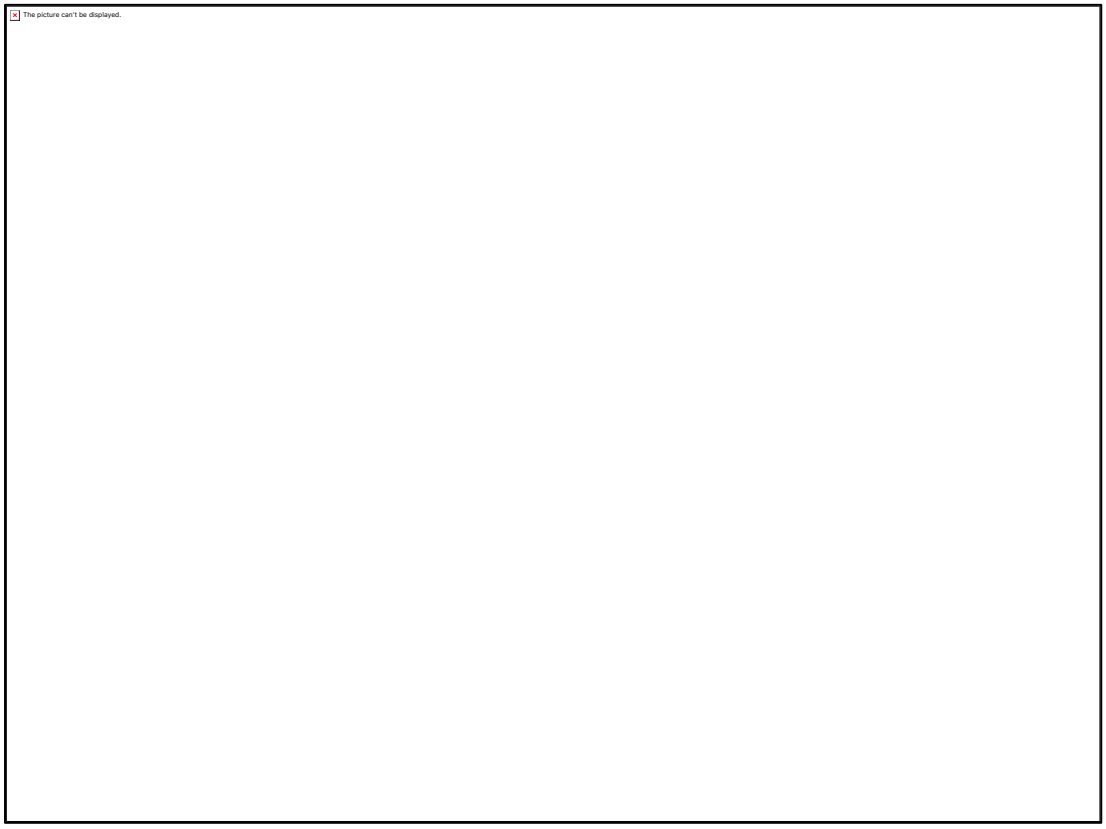


Figure 5.1 Elbow Method

Elbow method was performed first using R Studio to determine the optimal number of clusters for k-means clustering. Based on the chart above in Figure 5.1, the optimal number of clusters is 3 as we can see that the within-group distances decline dramatically between k=1 to k=3, where k denotes the number of clusters. Any clusters more than 3 is considered quite levelled off as the gradient of the line after k=3 is more gentle compared to gradient of the line when k=1 and k=2.

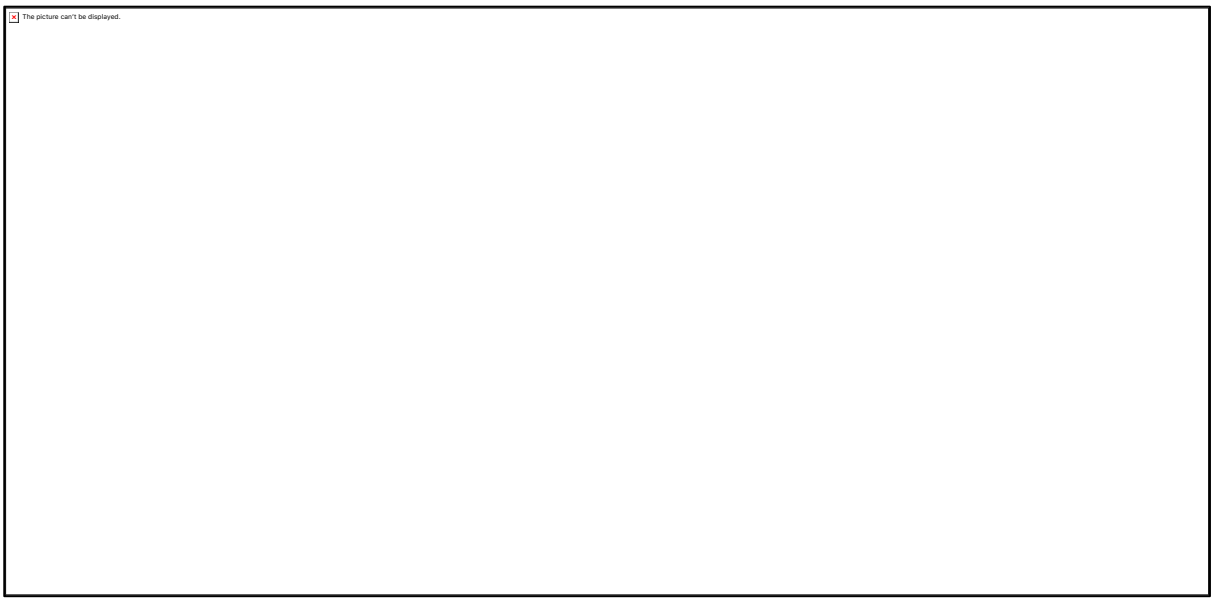


Figure 5.2 Clustering Analysis

Our first cluster is characterized by a majority of 83.94% of married deceased with residential status as intrastate non-residents, interstate non-residents, and foreign residents, and an average age of 72.61 years old. The second cluster has a 97.16% majority of married deceased with residential status as residents of the US, and an average age of 80 years old. Our last cluster shows a 64.82% majority of divorced deceased with residential status as residential and intrastate non-resident, with an average age of 62.39 years old.

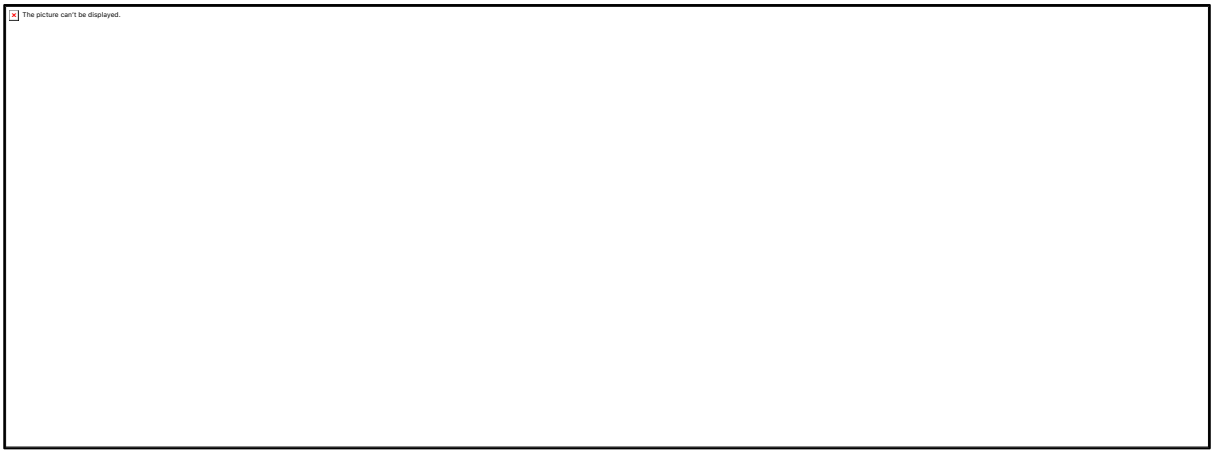


Figure 5.3 Clustering Analysis

From the above k-means clustering result, it shows that the clustering was done by grouping observations in terms of marital status and resident status. Cluster 1 mainly consists of intrastate non-residents, interstate non-residents and foreign residents who are either married, single or divorced. Cluster 2 mainly consists of residents who are married and single. Cluster 3 consists of residents and intrastate non-residents who are either married, single or divorced.

Further analysis is done to analyse the similarities within the clusters. For example, we would like to find out why is intrastate non-residents who are single (Marital Status 2, Resident Status 2) is clustered into Cluster 3 instead of Cluster 1, and so on.

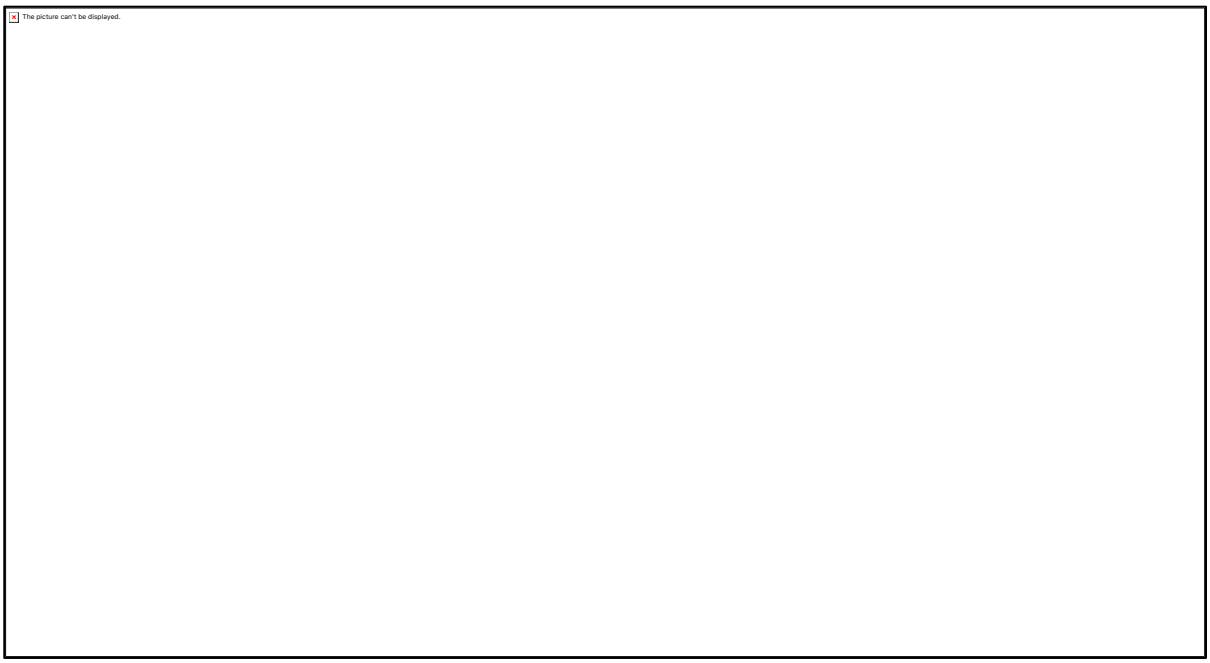


Figure 5.4 Clustering Chart 1

Figure 5.4 is a further analysis on the married deceased (Marital Status 0) who are clustered into three different clusters. Cluster 1 consists of intrastate non-residents, interstate non-residents and foreign residents with an age range of 29 to 102 years old. Cluster 2 consists of residents with an age range of 32 to 103 years old. Cluster 3 consists of residents with an age range of 26 to 36 years old. Further analysis is done to understand the overlapping of deceased aged 26 to 32 years old, as found in Cluster 2 and Cluster 3. From Figure 5.4, it shows that deceased aged between 32 to 36 years old in Cluster 2 are all females whereas deceased aged between 32 to 36 years old in Cluster 3 are mostly males, except 1 female. We believed that the female deceased aged 32 to 36 years old in Cluster 3 is an outlier.

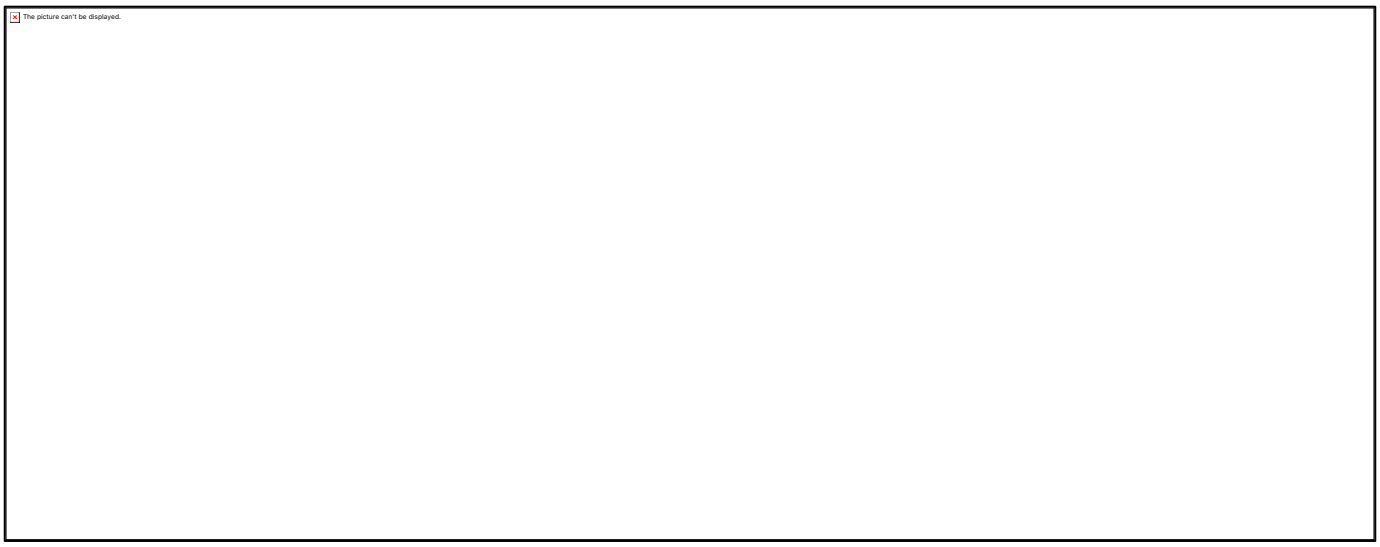


Figure 5.5 Clustering Chart 2

Figure 5.5 is a further analysis on the single deceased (Marital Status 1) who are clustered into three different clusters. Cluster 1 consists of intrastate non-residents, interstate non-residents and foreign residents with an age range of 5 to 93 years old. Cluster 2 consists of residents with an age range of 75 years old to 105 years old, whereas Cluster 3 consists of residents with an age range of 1 to 81 years old. Further analysis was conducted to understand the overlapping of deceased aged 75 to 81 years old, as found in Cluster 2 and Cluster 3. However, we were unable to rule out any similarities between the two clusters as both clusters consist of both males and females, with different ICD code and different education code. As such, we ruled out that this could be an outlier from the clustering.

Figure 5.6 Clustering Chart 3

Figure 5.6 is the last analysis on divorced deceased (Marital Status 2) who are clustered into three different clusters. Cluster 1 consists of intrastate non-residents and interstate non-residents with an age range of 36 to 98 years old, whereas Cluster 2 consists of residents and intrastate non-residents with an age range of 33 to 104 years old. Further analysis was done to understand the clustering of divorced deceased who are interstate non-residents (Resident Status 2) for both clusters 1 and 3. Cluster 1 consists of female interstate non-residents who are divorced, with an age range of 91 to 98 years old, whereas Cluster 3 consists of both male and female residents and interstate non-residents who are divorced, with an age range of 34 to 94 years old. There was an overlapping of interstate non-residents aged between 91 to 94 years old, and we found out that Cluster 1 consists of male interstate non-residents aged between 91 to 94 years old and Cluster 3 consists of 1 male and 1 female interstate non-residents deceased. As such, we believed that the male interstate non-resident in this cluster is an outlier from the clustering.

In conclusion, there are many other variables to consider in the clustering of these three groups, although no clear and significant links can be established using the other variables such as the ICD code, gender, and education code. However, we observed that the most significant factor in this dataset affecting life expectancy is the marital status, with the second factor being residential status. As such, on a micro level, insurance companies can consider the marital status of a client before setting the premium, and on a macro level, the government should provide more social services, especially to the elderly living alone, and have social support for single families. The findings are a rough guideline rather than conclusive statistical results. As such, we narrowed the scope and conducted another clustering analysis, by doing k-means clustering for the heart disease group.

ii. K-Means Clustering for Heart Disease Group

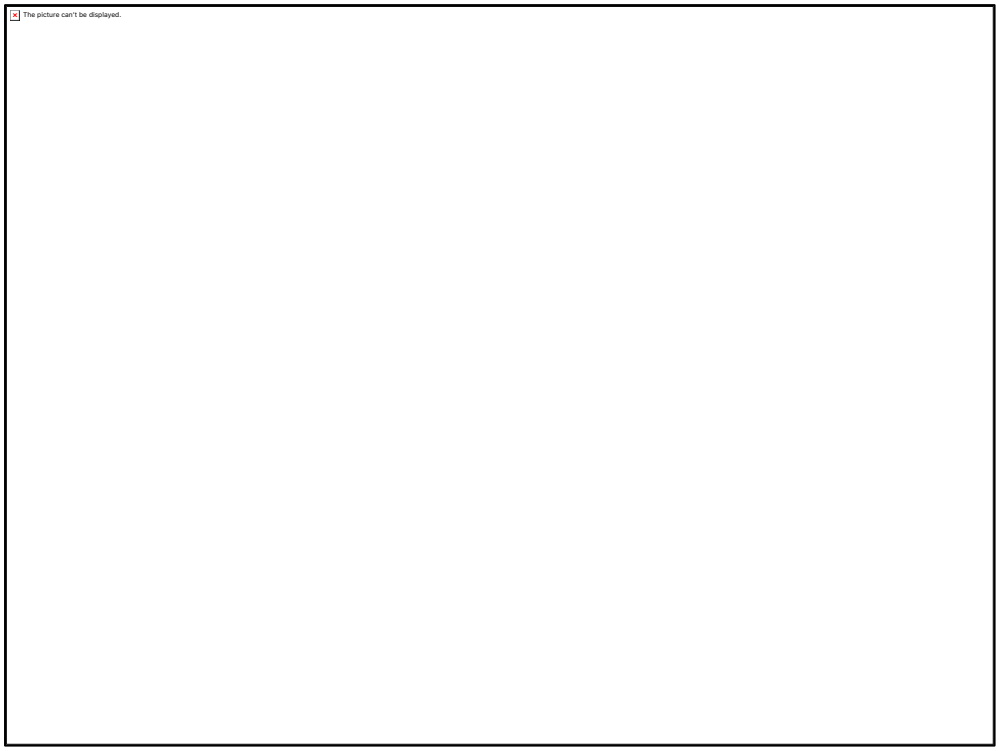


Figure 5.7 Elbow Method

Similarly, elbow method is first performed using R Studio to determine the optimal number of clusters for k-means clustering. Based on the chart above in Figure 5.7, the optimal number of clusters is 3 as we can see that the within-group distances decline dramatically between k=1 to k=3, where k denotes the number of clusters. Any clusters more than 3 is considered quite levelled off as the gradient of the line after k=3 is more gentle compared to gradient of the line when k=1 and k=2.

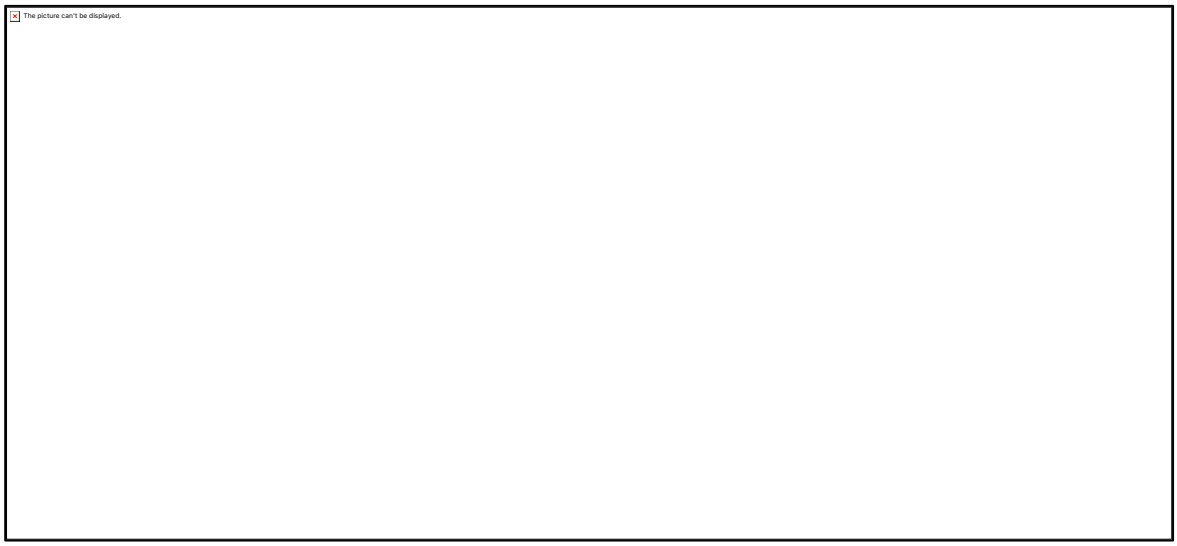


Figure 5.8 Clustering Analysis

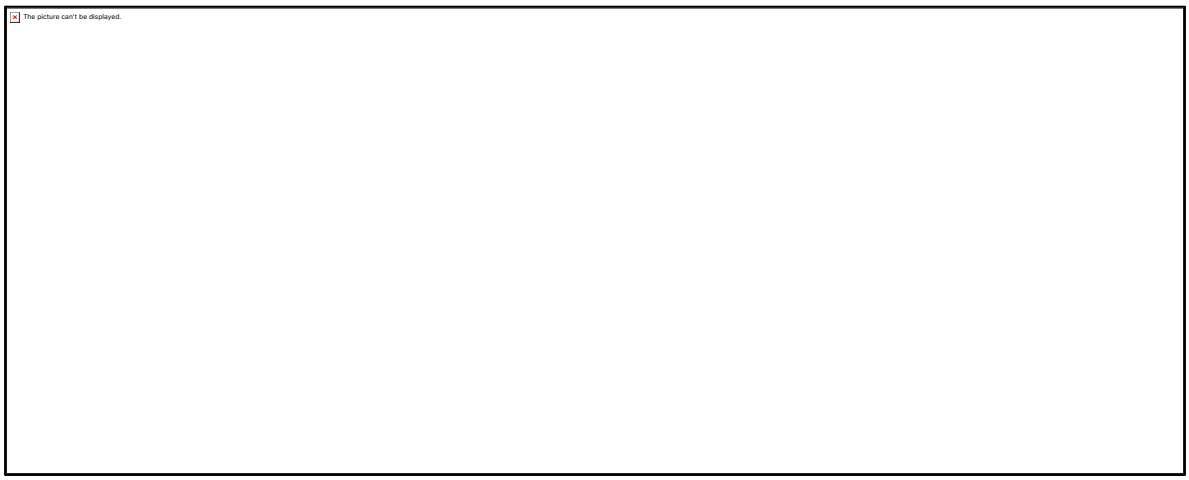


Figure 5.9 Clustering Chart

From our second clustering analysis, there are three different clusters where the first cluster has the largest count of 1,197, and the majority are male residents, with an average age of 74.18 years old. The second cluster shows the second largest count of 1,007, consisting of only female residents, with an average age of 80.38 years old. The last cluster consists of both genders that are either intra-state or inter-state non-residents or foreign residents, and with an average age of 70.39 years old. From this analysis, we can conclude from the count that males have a slightly higher chance of having a heart disease, and that they tend to have a significantly lower life expectancy than females with heart disease. The last cluster shows that the residential status has a significant impact on people who have heart disease, and could signify that the US government's healthcare system does not significantly support the welfare of non-residents and foreign residents. These findings can also be significant to insurance companies, who should consider the gender and the residential status of a client before insuring them. Not to mention, the findings will also be beneficial for heart disease researchers and scientists.

iii. Association Rules

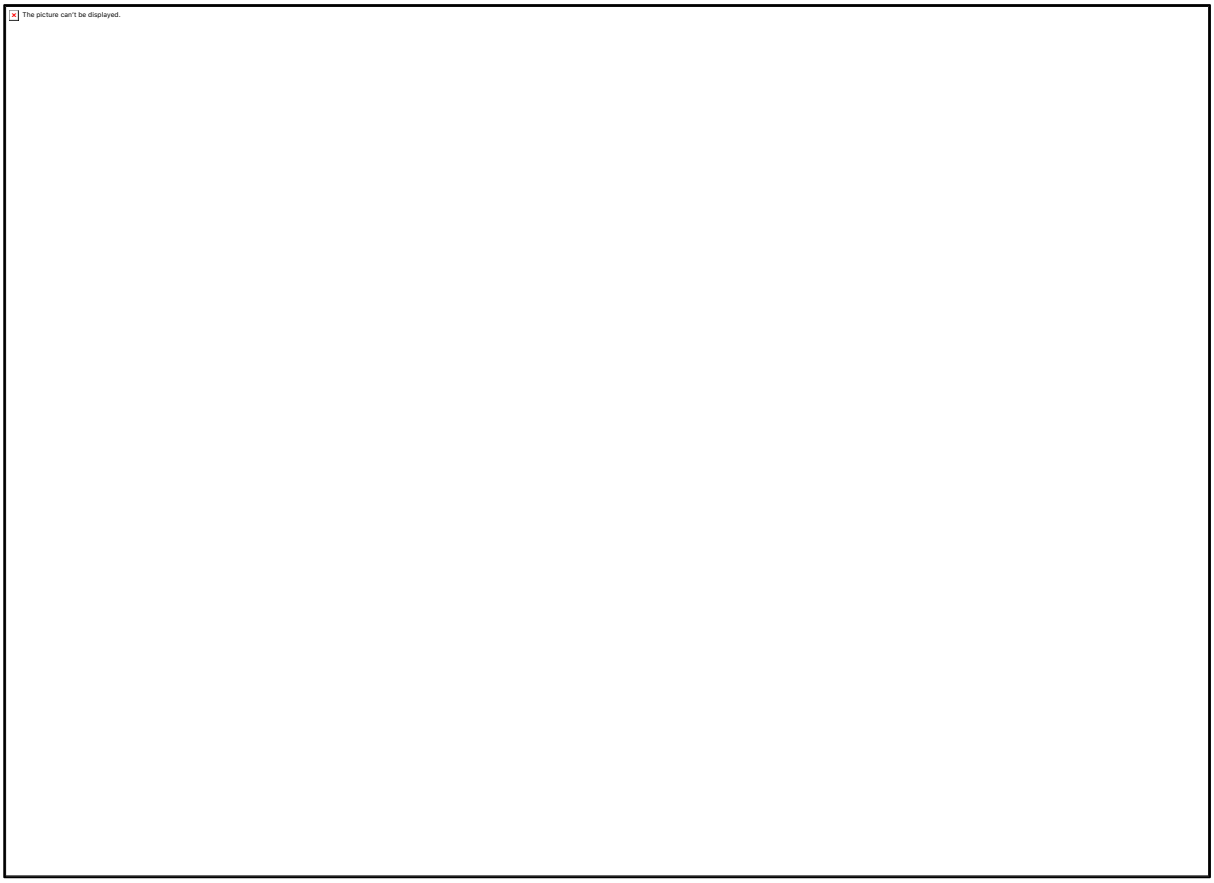


Figure 5.10 Association Rules Output

Association rule approach was also conducted as part of data mining, as seen in Figure 5.10. Minimum support was set to 50% and minimum confidence of 80%. From the above output as shown in Figure 5.10, there is no clear association between any of the variables in the dataset, especially when the lift ratios are near 1 or below 1, and there are no meaningful relationships and results that we can draw from the association rule.

6. Discussions and Conclusion

a. Relevance of Insights for Government of USA

Insufficient Resources Allocated to Non-Whites

We previously identified that American Indians have the highest morality rate for age 0 to 4 years, with 4.11%, more than 4 times next highest race – Asian or Pacific Islanders, with 0.77%. The Whites also have the lowest mortality rate of 0.04%. All these can be reflective of the social discrimination against non-whites in USA, and the lack of resources being channelled to post-natal care, which is a perennial issue. It is important for the government to bridge the inequality between races. Therefore, the government should not relent in their efforts in promoting social cohesion, and they can possible make use of the existing Obamacare, and expand post-natal home visits, which are important to close the gap for inequality, as well as reduce infant mortality.

Need to Expand Geriatric Care for Single Individuals

Based on our findings from both descriptive and predictive analysis, we identified that single individuals tend to have lower age of death, with a mean death age ranging from 51 to 60 years. This contrasts with the mean death age for married individuals, ranging from 81 to 90 years old.

This is supported by a research published in the American Journal of Epidemiology, where the researchers found the risk of death was 32% higher across a lifetime for single men compared to married men[11] *(Joan Raymond, 2011).* Single women face a 23% higher mortality risk, compared to married women[12] *(Joan Raymond, 2011). The* researchers speculated that their longevity findings could be tied to poorer health benefits, meagre public assistance and less income for singles. Also, some singles may not have the same social support that married couples have by default, resulting in lower life expectancy. This study excludes divorcees and widows, whereas our previous graph (as seen in Figure 3.3) shows that they still have a higher life expectancy than singles. However, the quality of marriage could very well affect a person's life expectancy as well.

The difference in life expectancy is shocking, and it is advisable that the government provide more healthcare support to the single geriatrics community, especially singles who live alone. It is certainly a social issue that needs to be addressed, with higher levels of public assistance to singles and divorcees alike. Providing higher quality healthcare services to this specific group will ensure that the life expectancy of the general population is rather consistent. With a healthier general population, people can work till an older age and hence, reducing the dependency ratio.

Heart Disease is the Most Common Disease

From Figure 3.4, we see that in the USA, one of the top diseases that people who suffer from is heart disease, with 26.1% of the sample suffering from heart disease. This percentage is the highest compared with the other diseases and health conditions classified under the ICD, followed by malignant neoplasm and respiratory disease being the second and third top diseases respectively. This brings our attention to the issue of heart disease being a common disease that many Americans suffer from due to their lifestyle and dietary habits and the leading cause of death for Americans.

In particular, USA is known for its high obesity rate for decades, with 36% of American adults and 17% of American teenagers with obesity in 2011 to 2014[13] *(Cynthia L. Ogden, Margaret D. Carroll, Cheryl D. Fryar, 2015).* Also, 80% of American adults do not engage in a sufficient level of physical activity that meets the government's recommended level of national physical activity[14] *(Physical Inactivity in the United States, n.d.).* Given the statistics of obesity rate and level of physical inactivity in the United States, it is clear that in order for the government to effectively reduce the rate of heart disease among the Americans, it has to work on its current measures to lower obesity rates and encourage Americans to be more active as obesity and physical inactivity are factors that can increase the risk of individuals getting heart disease.

In general, the analysis we have shown earlier will aid the USA government in addressing social issues and ensure more efficient allocation of resources. Death rates alone might seem rather insignificant for the government however, with effective and apt use of analysis, it will bring about dynamic effects on the country.

b. Relevance of Insights for Insurance Companies

Males are more Accident Prone

From Figure 3.6, we notice that males make up a greater percentage among those that died from accident, with males taking up 64.8% and females taking up only 35.2%. This could be an important observation for insurance companies when it comes to determining the premiums for their policies. Therefore, since males are more likely to be involved in accidents, when it comes to automobile insurance for males, insurance companies can consider raising the premium collected for this group of clients so as to cover their expenses in the event of an accident occurring and make a profit.

Previously, we identified that for accident cases, the percentage of males who died from accident is almost twice as many as the percentage of females who died from accident. Based on the Insurance Institute for Highway Safety Highway Loss Data Institute, much more males died of motor vehicle accidents as they tend to engage in more risky driving practices such as not putting on seat belts and speeding[15] *(General Statistics: Crashes took 32,675 lives in the U.S. in 2014, n.d.)*. A point to note would be that accidents do not only include motor vehicle accidents, but also include any other accidental deaths such as drowning, electrocution, fires, et cetera.

Males, Non-Residents and Foreign Residents Susceptible to Heart Diseases

From our clustering results as shown in Figure 5.9, we observed that males, non-residents and foreign residents are more susceptible to contracting heart diseases. Hence, insuring them comes with greater risk and to cover their potential losses, insurance companies should increase premiums, with regards to medical or life insurance.

Higher Life Expectancy for Females and hence prone to dementia

From our descriptive analysis, we observe that females have higher life expectancy and it can be backed with our regression analysis, which also shows that married individuals are likely to die at age 70, much higher than single individuals at age 54, ceteris paribus. Females tend to have a higher life expectancy, which also puts them at a higher risk of getting dementia as dementia comes with old age. In USA, a person with dementia may have Medicare, private insurance, a group employee plan or retiree health coverage to pay for healthcare[16] *(Insurance, n.d.)*. With so many forms of insurance coverage for dementia, insurance companies should set competitive premiums but enough to generate profit.

In general, the insights above will allow insurance companies to tailor their policies according to the vulnerability of each group. Given that the insurance industry in USA is the largest in the world by revenue, worth more than $1.2 trillion[17] *(Sean Millard, 2015)*, insurance companies should segment their market. One possible way is to segment based on the type of disease one is likely to contract and conduct targeted marketing. This will ensure that each of their product will reach its targeted audience more effectively.

c. Possible Improvements of Analysis

As the dataset only contains data for the year 2014, our findings may be limited and may not be as accurate making it difficult to compare between the years. As our analysis is only limited to the year 2014, changes and trends over the years which can be important in our analysis might be missed out. Our analysis can be improved if the death records of USA for other years are available, as such we will be able to better identify significant trends, areas of improvements, patterns and possible outliers.

Furthermore, there is an apparent lack of geographical data provided. For instance, the origin of state was not given. This limits our ability to further analyse on some of the given variables such as resident status. With the availability of the geographical data, we can possibly harness more insights and achieve more accurate results, in terms of the different area and states in United States. This is especially so as different states have different levels of urbanisation and development, and we feel that trends associated with geographical location might be present. This can be improved with more holistic data collected. With more accurate analysis, it can strengthen our findings and trends, so that the relevant policymakers can better target and address the various issues.

Lastly, there is difficulty in representing and determining the individual behaviours, such as smoking or exercising, which can affect the cause of death. One possible improvement is to identify common behaviours, and collect qualitative data on it.

The findings in our report might not be fully relatable to the relevant policy makers in USA due to the aforementioned limitations. However, we believe that our analysis is accurate to a large extent and will still prove to be useful for the USA government and insurance companies.

d. Lessons Learnt

This project has allowed us to consolidate and make use of what we have learnt over the past few months by using it in real life applications. Each analysis produces its own meaningful result which allows us to better understand and interpret the data.

From descriptive analytics, we are able to better identify the summary of the data in order for us to yield useful information in the later part of the analysis. As descriptive analysis provides us with information about what has happened, we can attempt to find out deeper and more information with the aid of subsequent analytics tools. Descriptive analytics are useful because the results from the analysis allow us to understand how the past events might influence future outcomes. It also allows us to explore and fully understand the dataset in order for us to proceed on with further analytics tools. For our case, we are able to use the current data to understand what affects the life expectancy, and hopefully able to predict and provide useful insights to the different relevant policy makers.

Predictive analytics allows us to understand and analyse what could happen in the future. Although predictive analytics only provide an estimate about the likelihood of future outcome, it is the only way to predict the future with the use of past data. Through the use of regression, we are able to predict the life expectancy of a person, given certain statistically significant variables. And through data mining, we are able to discover patterns in huge datasets, especially in our use of clustering. It allows us to discover similarities within clusters as well as identify underlying differences between clusters. This allows us to understand what groups the cluster together, and discover patterns which can potentially be useful to various policy makers.

In conclusion, this project allows us to learn more in depth about the skills and tools that we have learnt in this introduction module. It also allows us to have a try in applying it to real life applications, to further enhance our learning. Overall, from this project, all of the members have gained a better understanding of the tools and concepts that were taught through putting them to use and have benefitted greatly. It has been an enjoyable and interesting project for all of us.

7. References

[1] Kaggle, www.kaggle.com, accessed October 2016.

[2] Kaggle. *Death in United States*. Retrieved from https://www.kaggle.com/cdc/mortality. Accessed on October 2016.

[3] Dean Dacosta. (10 February, 2015). Recruiting Tools. Retrieved from Kaggle: The Home Of Data Science: http://recruitingtools.com/kaggle-home-data-science/

[4] Ibid.

[5] Christopher Ingraham. (29 September, 2014). The Washington Post. Retrieved from Our infant mortality rate is a national embarrassment: https://www.washingtonpost.com/news/wonk/wp/2014/09/29/our-infant-mortality-rate-is-a-national-embarrassment/

[6] Heart Disease Fact Sheet. (16 June, 2016). Retrieved from Centers for Disease Control and Prevention: http://www.cdc.gov/dhdsp/data_statistics/fact_sheets/fs_heart_disease.htm

[7] Ibid.

[8] 2016 Alzheimer's Statistics. (n.d.). Retrieved from Alzheimers.net: http://www.alzheimers.net/resources/alzheimers-statistics/

[9] Stroke. (5 May, 2016). Retrieved from Centers for Disease Control and Prevention: https://www.cdc.gov/stroke/

[10] Jessica C. Barnett, M. V. (13 September, 2016). Health Insurance Coverage in the United States: 2015. Retrieved from United States Census Bureau: http://www.census.gov/library/publications/2016/demo/p60-257.html

[11] Joan Raymond. (18 August, 2011). Single people may die younger, new study finds. Retrieved from NBC News: http://www.nbcnews.com/id/44122528/ns/health-behavior/t/single-people-may-die-younger-new-study-finds/

[12] Ibid.

[13] Cynthia L. Ogden, Margaret D. Carroll, Cheryl D. Fryar. (November, 2015). Prevalence of Obesity Among Adults and Youth: United States, 2011-2014. Retrieved from NCHS: https://www.cdc.gov/nchs/data/databriefs/db219.pdf

[14] Physical Inactivity in the United States. (n.d.). Retrieved from The State of Obesity: http://stateofobesity.org/physical-inactivity/

[15] General Statistics: Crashes took 32,675 lives in the U.S. in 2014. (n.d.). Retrieved from Insurance Institute for Highway Safety, Highway Loss Data Institute: http://www.iihs.org/iihs/topics/t/general-statistics/fatalityfacts/gender

[16] Insurance. (n.d.). Retrieved from Alzheimer's Association: http://www.alz.org/care/alzheimers-dementia-insurance.asp#longterm

[17] Sean Millard. (11 Feb, 2015). The US insurance industry: Largest in the world. Retrieved from Market Realist: http://marketrealist.com/2015/02/us-insurance-industry-largest-world/