Final Project

Cherubin Frenel

8/8/2019

## Registered S3 methods overwritten by 'ggplot2':  
## method from   
## [.quosures rlang  
## c.quosures rlang  
## print.quosures rlang

## -- Attaching packages ----------------- tidyverse 1.2.1 --

## v ggplot2 3.1.1 v purrr 0.3.2  
## v tibble 2.1.3 v dplyr 0.8.3  
## v tidyr 0.8.3 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.4.0

## -- Conflicts -------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

## 'data.frame': 15 obs. of 3 variables:  
## $ Year : int 1900 1900 1950 1950 2000 2000 2000 1950 2000 1900 ...  
## $ Cause : Factor w/ 10 levels "Accidents","Cancer",..: 9 7 3 1 1 4 8 10 6 5 ...  
## $ Number.of.Deaths: int 38820 40362 60989 91249 97900 122009 167661 156751 710760 28491 ...

## Year Cause Number.of.Deaths  
## Min. :1900 Heart Disease :3 Min. : 21353   
## 1st Qu.:1900 Accidents :2 1st Qu.: 39591   
## Median :1950 Cancer :2 Median : 97900   
## Mean :1950 Stroke :2 Mean :190887   
## 3rd Qu.:2000 Certain Diseases of Infancy :1 3rd Qu.:189197   
## Max. :2000 Chronic Respiratory Diseases:1 Max. :710760   
## (Other) :4

## Year Cause Number.of.Deaths  
## 1 1900 Tuberculosis 38820  
## 2 1900 Influenza/Pneumonia 40362  
## 3 1950 Certain Diseases of Infancy 60989  
## 4 1950 Accidents 91249  
## 5 2000 Accidents 97900  
## 6 2000 Chronic Respiratory Diseases 122009

### Section 3 – Week 10

### Overall, write a coherent narrative that tells a story with the data as you complete this section.

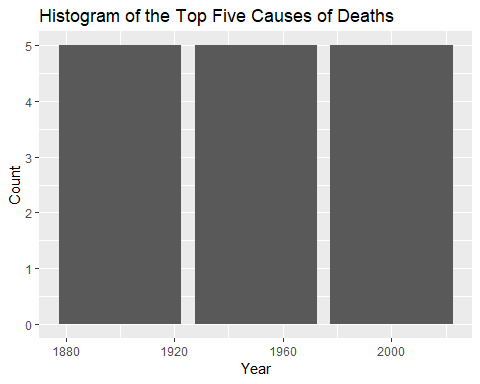
### For more than 10 years, diseases such as, heart disease, cancer, chronic lower respiratory disease, stroke, diabetes, etc. have been the focus of studies by researchers and doctors because those diseases have claimed important spots in leading causes of deaths. For instance, heart disease and cancer respectively have claimed spots number 1 and 2 in America and perhaps in the world. According to studies, both of those diseases are strongly related to behaviors like smoking. Therefore, American health experts and doctors strongly encourage Americans to curb smoking, and they focus on the treatment of heart disease. As a consequence of their efforts, the number of heart disease related deaths has decreased over time; whereas, the number of cancer related deaths has been rising. Base on the number of deaths associated with cancer over the last five decades, I suspect that cancer may at some point overtake heart disease as the leading cause of deaths in the years to come. This can be seeing in the Exploratory Data Analysis part of this project.

### Summarize the problem statement you addressed.

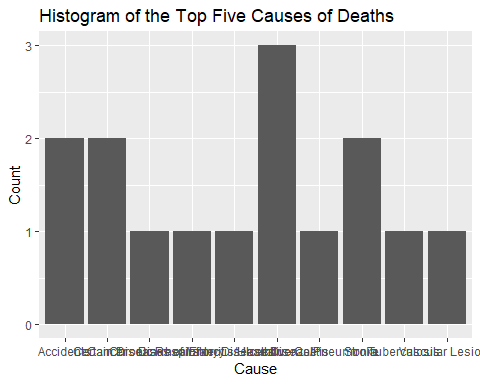
### The Top Five Leading Causes of Deaths is the problem statement I addressed. They are heart disease, cancer, accidents, chronic lower respiratory disease, and stroke as one can see in the dataset. For the last few decades, they lead to significant amount of deaths in America with heart disease claiming the number spot as the leading cause of deaths followed by cancer as the second. Third comes accidents followed by chronic lower respiratory disease and stroke as fourth and fifth respectively according to the data.

### Histogram of the Top Five Leading Causes of Deaths

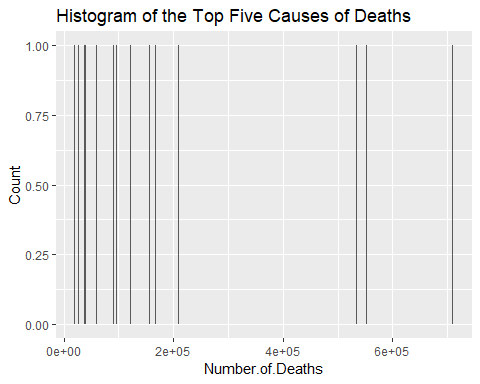
### Histogram of the Top Five Causese of Deaths for variable Year



### Histogram of the Top Five Causese of Deaths for variable Cause

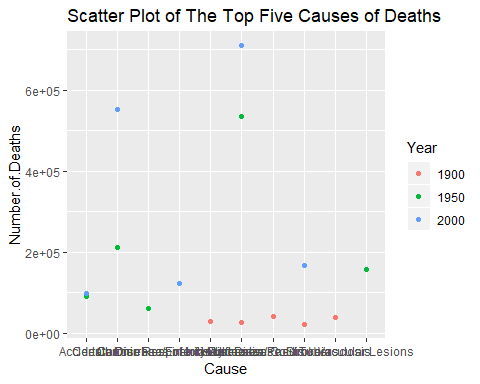


### Histogram of the Top Five Causese of Deaths for variable Number of Deaths

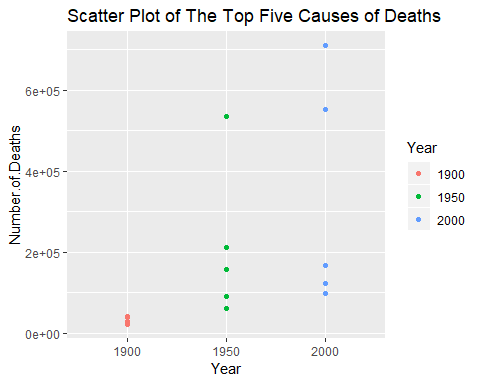


### Scatter plot of the Top Five Leading Causes of Deaths

### Number of Deaths Vs Cause

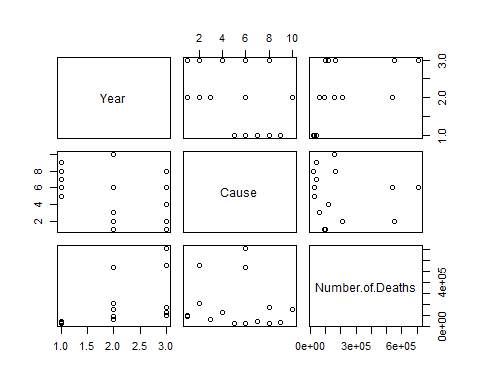


### Number of Deaths Vs Year



### Summarize how you addressed this problem statement (the data used, and the methodology employed).

### To address this problem statement, I used The Top Five Leading Causes of Deaths dataset. With this dataset, I created histograms and Scatter plots that displays the distribution of the variables. To further look at the distribution, I performed a correlation test between the variables, and I fitted a binary logistic regression model to the data set. Those are exploratory data analysis methodologies.



##   
## Pearson's product-moment correlation  
##   
## data: df$Number.of.Deaths and df$Number.of.Deaths  
## t = Inf, df = 13, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 1 1  
## sample estimates:  
## cor   
## 1

The correlation test has a p-value of p-value = p-value < 2.2e-16, which can be considered as small, so variables Year and Number of Deaths are correlated. Therefore, the null hypothesis is true.

##   
## Pearson's product-moment correlation  
##   
## data: df$Number.of.Deaths and df$Number.of.Deaths  
## t = Inf, df = 13, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 99 percent confidence interval:  
## 1 1  
## sample estimates:  
## cor   
## 1

I performed the test at the 99% confidence interval, and I think the results are still significant.

### Fit a binary logistic regression model to the data set

##   
## Call:  
## glm(formula = Cause ~ ., family = binomial, data = df)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.46919 0.00007 0.00009 0.20843 1.37016   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) 1.845e+01 4.793e+03 0.004 0.997  
## Year1950 -2.112e+01 4.793e+03 -0.004 0.996  
## Year2000 -2.228e+01 4.793e+03 -0.005 0.996  
## Number.of.Deaths 3.657e-05 3.592e-05 1.018 0.309  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 11.7802 on 14 degrees of freedom  
## Residual deviance: 6.3362 on 11 degrees of freedom  
## AIC: 14.336  
##   
## Number of Fisher Scoring iterations: 18

### Summarize the interesting insights that your analysis provided.

### I performed a correlation test between Year and Number of Deaths, which are 2 critical variables in the dataset, and I suspect that there is positive correlation between them. The correlation test returns a p-value of p-value < 2.2e-16. This value is small, and it’s an indication that the 2 variables are correlated. In addition, I performed the correlation test at the 99% confidence interval, and I think the results are still significant.

### Summarize the implications to the consumer (target audience) of your analysis.

### Base on the results of the correlation test, I suspect that the variables are correlated. It has been known that smoking causes cancer, and studies have proven it. For the past few years, governmental as well as private agencies have been encouraging people to stop smoking in order to prevent cancer. There are clear indications that this leading cause of deaths will take the number spot in the next few years if officials ignore it.

### Discuss the limitations of your analysis and how you, or someone else, could improve or build on it.

### I think the Scatter plots restrict my analysis when it comes to study the relatioship between the variables. I tried different ways of plotting them; they still display some limitations, however. I think I or someone else could improve on it by learning other ways to better manipulate the variables when creating them.