NCHS

## R Markdown

This is an R Markdown document.

Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot.

The Research Questions had been modified based on the comments given for the previous assignments as below:

1. What is the trend of all the five-leading cause of deaths over time for each year?

2. What is the maximum and minimum number of expected deaths, observed deaths, potentially excess deaths for a particular cause of death in each age group?

3. What is the trend of the deaths over various regions of US? Which region has the maximum and minimum number of deaths?

4. What is the trend of expected deaths, observed deaths, potentially excess deaths in each region for each benchmark in each locality?

5. What is the ratio between the observed deaths and the Population?

What is the trend of the ratios over time by region?

6. What is the ratio between the expected deaths and the Population?

What is the trend of the ratios over time by region?

7. Are the ratios in questions 5 and 6 correlated?

This document consists of R script and R plots which will help in providing answers to the above questions.

NCHS\_New1 <- read.csv("NCHS.csv")  
str(NCHS\_New1)

summary(NCHS\_New1)

sum(is.na(NCHS\_New1))

aft\_mis\_NCHS\_New1<-na.omit(NCHS\_New1)  
sum(is.na(aft\_mis\_NCHS\_New1))

str(aft\_mis\_NCHS\_New1)

#duplicated(NCHS\_New1)  
str(NCHS\_New1)

aft\_mis\_NCHS\_New2<- aft\_mis\_NCHS\_New1[!(aft\_mis\_NCHS\_New1$State == "United States"),]  
dat <- aft\_mis\_NCHS\_New2  
head(dat)

## Code to include regions column in the data

library(noncensus)

data("states")  
states

Reg <- states[,c(1,3)]  
Reg

Health\_data <- merge(dat,Reg,by.x = "State.FIPS.Code",by.y = "state")  
head(Health\_data)

# write.csv(Health\_data,"Health data.csv")  
  
# 1.trend of all the five leading cause of Observed deaths over time for each region  
  
Data\_all <- Health\_data[,c(2:3,9,14)]  
head(Data\_all)

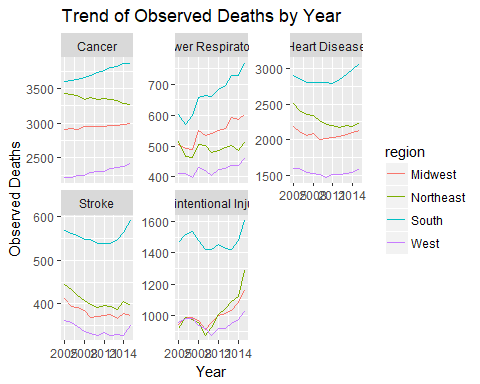
library(dplyr)

Data\_allNew <- Data\_all %>%  
 group\_by(region, Year, Cause.of.Death) %>%  
 summarise\_all(funs(mean))  
Data\_allNew

library(ggplot2)

ggplot(data = Data\_allNew, aes(x=Year, y = Observed.Deaths, colour = region, group = region)) + labs(title="Trend of Observed Deaths by Year", x="Year",y="Observed Deaths") +  
 geom\_line()+scale\_x\_continuous(breaks=seq(2005, 2015, 3)) + facet\_wrap(~Cause.of.Death,scales = "free\_y")

This plot provides the overview of trend for the five leading cause of deaths (observed deaths) over time. From the plot we could see that “Cancer” has the highest number of deaths and that is in the southern region of United States. The lower number of deaths could be seen for “Stroke” in the western region of United States. The order of the trends in descending order are Cancer, Heart Disease, Unintentional Injury, Chronic Lower Respiratory Disease, Stroke.



# 2. Max and Min value of all types of deaths in Each age group  
  
Data\_sec <- Health\_data[,c(3,6,9,11,12)]  
str(Data\_sec)

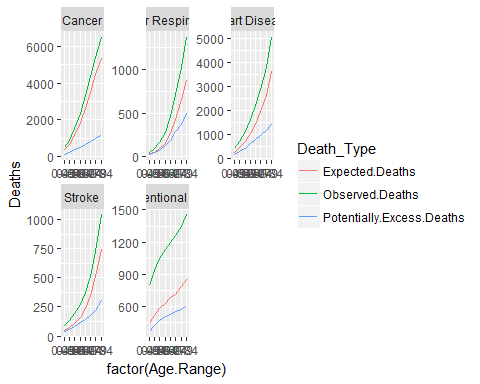
library(dplyr)  
Data\_sec\_New <- Data\_sec %>%  
 group\_by(Age.Range, Cause.of.Death) %>%  
 summarise\_all(funs(mean))  
Data\_sec\_New

library(tidyr)

long\_data\_sec\_New <- gather(Data\_sec\_New,"Death\_Type", "Deaths", Observed.Deaths, Expected.Deaths, Potentially.Excess.Deaths)  
long\_data\_sec\_New

library(ggplot2)  
ggplot(data = long\_data\_sec\_New, aes(x=factor(Age.Range), y = Deaths, colour = Death\_Type, group = Death\_Type)) +   
 geom\_line() + facet\_wrap(~Cause.of.Death,scales = "free")

This plot provides the trend of all types of deaths for each age group. In this plot we could see that the observed deaths are always greater than the expected deaths. As the age increases the numbers of deaths kept increasing. This behavior is consistent for all the causes of death.

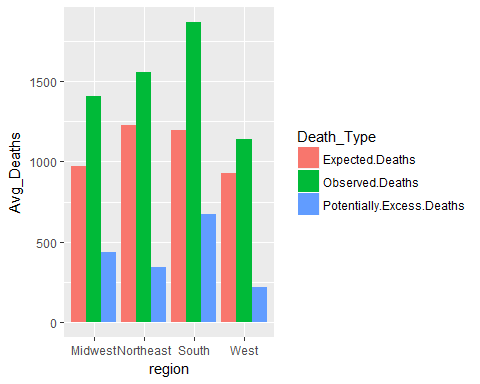


# 3. Max and Min Deaths by region  
library(dplyr)  
Data\_region <- Health\_data[,c(9,11,12,14)] %>%  
 group\_by(region) %>%  
 summarise\_all(funs(mean))  
Data\_region

library(tidyr)  
long\_data\_region <- gather(Data\_region,"Death\_Type", "Avg\_Deaths", Observed.Deaths, Expected.Deaths, Potentially.Excess.Deaths)  
long\_data\_region

library(ggplot2)  
ggplot(long\_data\_region,aes(x = region,y = Avg\_Deaths,fill = Death\_Type)) + geom\_bar(position = "dodge", stat = "identity")

This plot gives the mean of all the deaths over various regions. The observed deaths are more in the southern part of the United States followed by Northeast, Midwest and West. The Expected Deaths average is almost equal in both South and Northeast. Midwest and West also have equal averages but lower than South and Northeast. South region also has the maximum average for potentially excess deaths followed by Midwest, Northeast and West

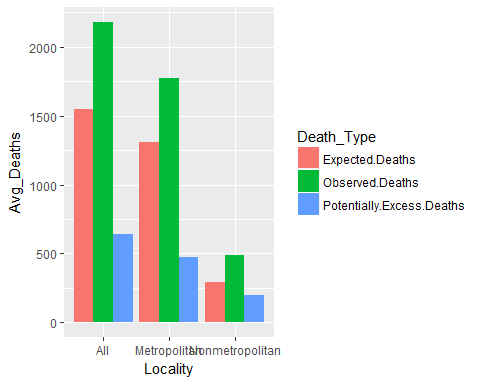


#4. Trend of Deaths by locality  
library(dplyr)  
Data\_Locality <- Health\_data[,c(8,9,11,12)] %>%  
 group\_by(Locality) %>%  
 summarise\_all(funs(mean))  
Data\_Locality

library(tidyr)  
long\_Data\_Locality <- gather(Data\_Locality,"Death\_Type", "Avg\_Deaths", Observed.Deaths, Expected.Deaths, Potentially.Excess.Deaths)  
long\_Data\_Locality

library(ggplot2)  
ggplot(long\_Data\_Locality,aes(x = Locality,y = Avg\_Deaths,fill = Death\_Type)) + geom\_bar(position = "dodge", stat = "identity")

Below plot provides the means of all the deaths by locality. For all the death parameters the highest and lowest are consistent which are found in All localities and Non-Metropolitan.

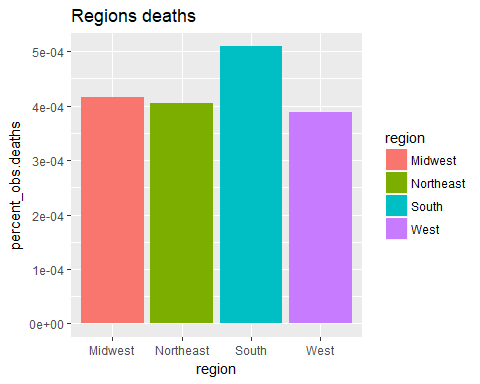


#5. Ratio of observed deaths and population over various regions  
Health\_data$percent\_obs.deaths <- Health\_data$Observed.Deaths / Health\_data$Population  
head(Health\_data)

Obs.Deaths\_region <- Health\_data[,c(14:15)] %>%  
 group\_by(region) %>%  
 summarise\_all(funs(mean))  
Obs.Deaths\_region

library(ggplot2)  
ggplot(Obs.Deaths\_region,aes(x = region,y = percent\_obs.deaths,fill = region)) + labs(title="Regions deaths") + geom\_bar(stat = "identity")

This plot provides the fact that the observed deaths percentage is found highest in the Northeast and the lowest in West. Based on different regions (Midwest, Northeast, West and South) in the united states, the average for Observed deaths in south of united states is larger than other regions and so percentage of observed deaths in south of united states is more than other regions.

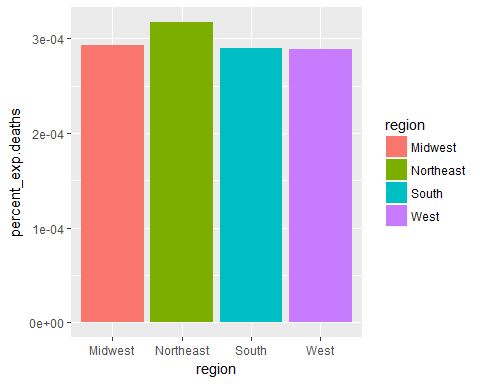


#6. Ratio of Expected Deaths to Population over regions  
Health\_data$percent\_exp.deaths <- Health\_data$Expected.Deaths / Health\_data$Population  
head(Health\_data)

exp.Deaths\_region <- Health\_data[,c(14,16)] %>%  
 group\_by(region) %>%  
 summarise\_all(funs(mean))  
exp.Deaths\_region

library(ggplot2)  
ggplot(exp.Deaths\_region,aes(x = region,y = percent\_exp.deaths,fill = region)) + geom\_bar(stat = "identity")

This plot is between ratio of expected deaths to population over regions. The expected death in northeast has highest rate compared with other regions. So even the expected death in northeast is greater than other regions. But as observed deaths in south of the united states is much greater than other regions, the potentially excess deaths (difference between expected death and observed death) in south of America is greater than other regions.



7. What is the correlation between the two ratios in Question 5 and 6

*cor(Health\_data$percent\_obs.deaths,Health\_data$percent\_exp.deaths)*

*This gives a 0.94 positive correlation value*

#Standard deviation:  
  
sd(Health\_data$Potentially.Excess.Deaths)

sd(Health\_data$Observed.Deaths)

sd(Health\_data$Expected.Deaths)

#Anova table  
Anova.aov= aov(Health\_data$Expected.Deaths~Health\_data$Cause.of.Death)  
Anova.aov

summary(Anova.aov)

Anova.aov= aov(Health\_data$Population~Health\_data$Observed.Death)  
Anova.aov

summary(Anova.aov)

Anova.aov= aov(Health\_data$Observed.Deaths~Health\_data$Cause.of.Death)  
Anova.aov

summary(Anova.aov)

#Linear Regression  
Regression<-lm(Health\_data$Expected.Deaths ~ Health\_data$Cause.of.Death + Health\_data$Population)  
summary(Regression)

#Correlation  
cor.test(Health\_data$percent\_obs.deaths,Health\_data$Population)

# partial correlations  
library(ggm)

pcor(c("Observed.Deaths", "percent\_obs.deaths"), var(Health\_data))