

Exploratory Data Analysis + Statistical Testing

Patrick Asztabski

4/7/2022

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.5
```

```
## -- Attaching packages ----- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5      v purrr   0.3.4
## v tibble  3.1.5      v dplyr   1.0.7
## v tidyr   1.1.3      v stringr 1.4.0
## v readr   1.4.0      v forcats 0.5.1
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
## Warning: package 'tibble' was built under R version 4.0.5
```

```
## Warning: package 'tidyr' was built under R version 4.0.5
```

```
## Warning: package 'readr' was built under R version 4.0.3
```

```
## Warning: package 'purrr' was built under R version 4.0.5
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```
## Warning: package 'forcats' was built under R version 4.0.5
```

```
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(dplyr)
library(agricolae)
```

```
## Warning: package 'agricolae' was built under R version 4.0.5
```

```
library(FSA)
```

```
## Warning: package 'FSA' was built under R version 4.0.5
```

```
## ## FSA v0.9.3. See citation('FSA') if used in publication.  
## ## Run fishR() for related website and fishR('IFAR') for related book.
```

```
library(ggpubr)
```

```
## Warning: package 'ggpubr' was built under R version 4.0.5
```

```
## Registered S3 methods overwritten by 'car':  
##   method      from  
##   hist.boot   FSA  
##   confint.boot FSA
```

```
library(scales)
```

```
## Warning: package 'scales' was built under R version 4.0.3
```

```
##  
## Attaching package: 'scales'
```

```
## The following object is masked from 'package:purrr':  
##  
##   discard
```

```
## The following object is masked from 'package:readr':  
##  
##   col_factor
```

```
require(gridExtra)
```

```
## Loading required package: gridExtra
```

```
## Warning: package 'gridExtra' was built under R version 4.0.5
```

```
##  
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':  
##  
##      combine
```

Cleaning our data

This part is a portion of our data that is cleaned via R Programming

```
#####
heartData <- read.csv("heart_updated.csv")
strokeData <- read.csv("stroke_updated.csv")
spending <- read.csv("regionspending.csv")
strokeData[strokeData == "American Indian and Alaskan Native"] <- "AI/AN"
strokeData[strokeData == "Asian and Pacific Islander"] <- "A/PI"
heartData[heartData == "American Indian and Alaskan Native"] <- "AI/AN"
heartData[heartData == "Asian and Pacific Islander"] <- "A/PI"

Race <- c("White", "Black", "Asian", "Hispanic")
AnyHealthInsurance <- c(87.2,84.1,85.5,75.7)
PrivateHealthInsurance <- c(66.8,49.6,68.1,45.6)
GovtHealthInsurance <- c(33.2,43.8,25.8,36.4)
Uninsured <- c(12.8, 15.9, 14.5, 24.3)
Insurance <- data.frame(Race, AnyHealthInsurance, PrivateHealthInsurance, GovtHealthInsurance, U
ninsured)
Insurance$fraction = Insurance$AnyHealthInsurance / sum(Insurance$AnyHealthInsurance)
Insurance$ymax = cumsum(Insurance$fraction)
Insurance$ymin = c(0, head(Insurance$ymax, n = -1))
Insurance$labelPosition <- (Insurance$ymax + Insurance$ymin) / 2
Insurance$label <- paste0(Insurance$Race, "\n Percentage: ", round(Insurance$fraction * 100))

Race <- c("White", "Asian", "Hispanic", "Black", "Alaskan Natives", "Native Americans")
DiabetesPrevalence <- c(7.1,8.4,11.8,12.6,5.5,33)
Diabetes <- data.frame(Race,DiabetesPrevalence)
Diabetes$fraction = Diabetes$DiabetesPrevalence / sum(Diabetes$DiabetesPrevalence)
Diabetes$ymax = cumsum(Diabetes$fraction)
Diabetes$ymin = c(0, head(Diabetes$ymax, n = -1))
Diabetes$labelPosition <- (Diabetes$ymax + Diabetes$ymin) / 2
Diabetes$label <- paste0(Diabetes$Race, "\n Percentage: ", round(Diabetes$fraction * 100))
labels = paste0(DiabetesPrevalence, "%")

#SUBSETTING DATA FOR INDIVIDUAL ANALYSIS
#####
whiteheartData = subset(x = heartData, subset = Race.Ethnicity == "White")
whitestrokeData = subset(x = strokeData, subset = Race.Ethnicity == "White")
BlackheartData = subset(x = heartData, subset = Race.Ethnicity == "Black")
BlackstrokeData = subset(x = strokeData, subset = Race.Ethnicity == "Black")
APIheartData = subset(x = heartData, subset = Race.Ethnicity == "A/PI")
APIstrokeData = subset(x = strokeData, subset = Race.Ethnicity == "A/PI")
HispanicheartData = subset(x = heartData, subset = Race.Ethnicity == "Hispanic")
HispanicstrokeData = subset(x = strokeData, subset = Race.Ethnicity == "Hispanic")
AIANheartData = subset(x = heartData, subset = Race.Ethnicity == "AI/AN")
AIANstrokeData = subset(x = strokeData, subset = Race.Ethnicity == "AI/AN")

#ADD REGIONS FOR MERGED DATA
#####
mergedheartData <- rbind(whiteheartData,BlackheartData, APIheartData, HispanicheartData, AIANhea
rtData)
mergedstrokeData <- rbind(whitestrokeData,BlackstrokeData, APIstrokeData, HispanicstrokeData, AI
ANstrokeData)
mergedheartData$region <- with(mergedheartData,
```

```

        ifelse(LocationAbbr %in% c("AR","TN","LA","MS","AL","FL","GA","SC","NC",
        ,"KY","WV","DE","MD", "DC","TX","OK", "VA"),'South',
        ifelse(LocationAbbr %in% c("HI","AK","OR","WA","ID","UT","NV","C
        A","AZ","MT","CO","NM"), 'West',
        ifelse(LocationAbbr %in% c("IN","OH","MI","IL","MO","IA",
        "MN","ND","SD","NE","KS", "WI"), 'MidWest',
        ifelse(LocationAbbr %in% c("PA","RI","CT",
        "ME", "NH", "VT", "MA", "NY", "NJ"), 'Northeast',
        ifelse(LocationAbbr %in% c("US"), 'O
        verall', 'Overall')
    )
  )
)

mergedstrokeData$region <- with(mergedstrokeData,
    ifelse(LocationAbbr %in% c("AR","TN","LA","MS","AL","FL","GA","S
    C","NC","KY","WV","DE","MD", "DC","TX","OK", "VA"),'South',
    ifelse(LocationAbbr %in% c("HI","AK","OR","WA","ID","UT",
    "NV","CA","AZ","MT","CO","NM"), 'West',
    ifelse(LocationAbbr %in% c("IN","OH","MI","IL","MO"
    ,"IA","MN","ND","SD","NE","KS", "WI"), 'MidWest',
    ifelse(LocationAbbr %in% c("PA","RI","CT",
    "ME", "NH", "VT", "MA", "NY", "NJ"), 'Northeast',
    ifelse(LocationAbbr %in% c("US"), 'Ov
    erall', 'Overall')
  )
)

mergedheartData$region <- gsub("Overall", "U.S. Overall", mergedheartData$region)
mergedstrokeData$region <- gsub("Overall", "U.S. Overall", mergedstrokeData$region)

```

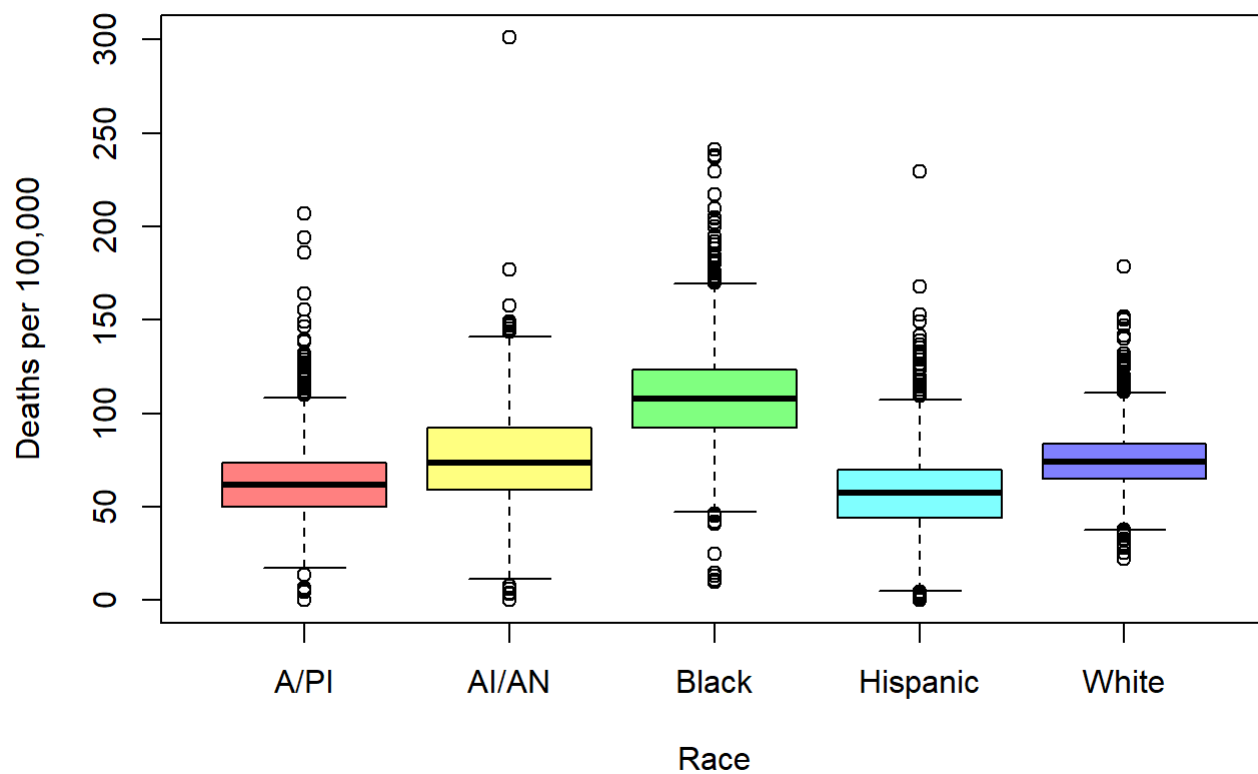
##Exploratory Data Analysis

```

rainbowcols <- rainbow(6,s = 0.5)
boxplot(mergedstrokeData$Deaths.per.100.000 ~ mergedstrokeData$Race.Ethnicity, xlab = "Race"
    , ylab = "Deaths per 100,000", col=c(rainbowcols), main = "Heart fatalities per 100,000
    by Race (2013)")

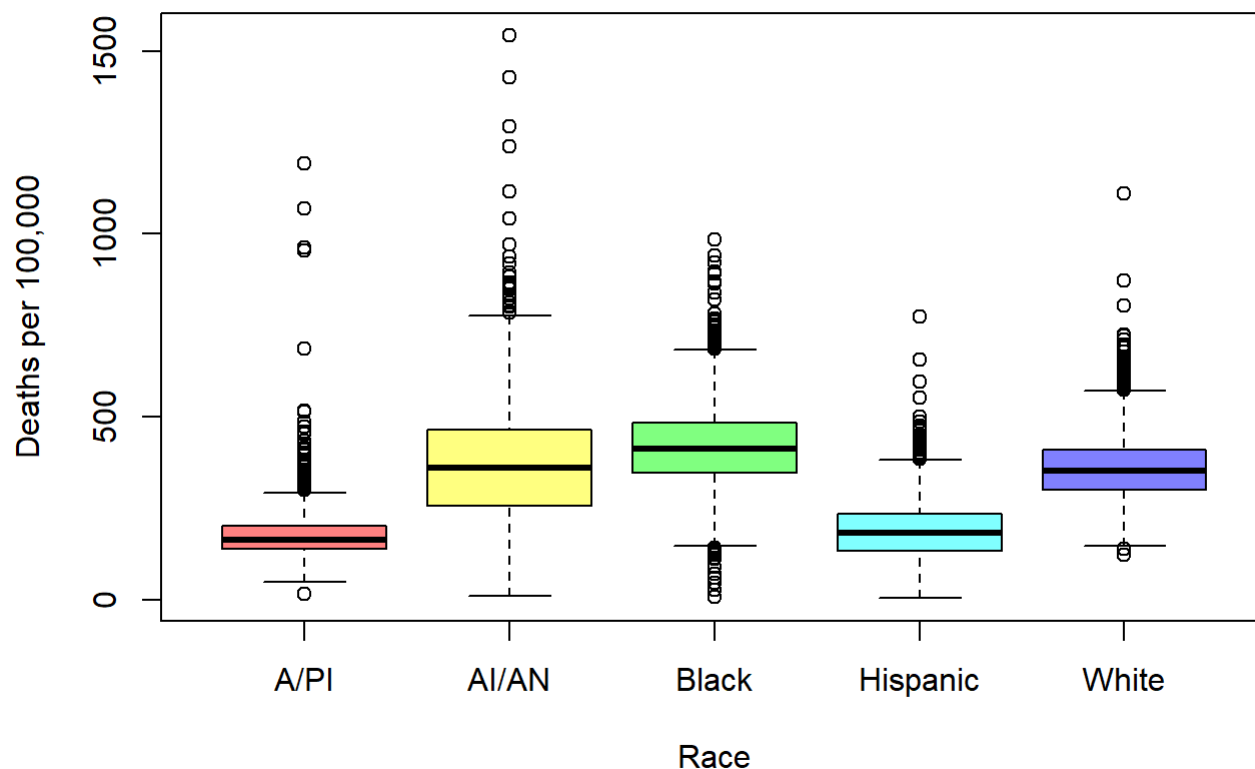
```

Heart fatalities per 100,000 by Race (2013)



```
boxplot(mergedheartData$Deaths.per.100.000 ~ mergedheartData$Race.Ethnicity, xlab = "Race",  
        ylab = "Deaths per 100,000", col=c(rainbowcols), main = "Stroke fatalities per 100,000  
by Race (2013)")
```

Stroke fatalities per 100,000 by Race (2013)



```
#Boxplots to distinguish outliers and present visualization on significant
#differences between multiple race/ethnicities
```

```
#Number of deaths
```

```
case.vector = tapply(mergedheartData$Deaths.per.100.000, mergedheartData$Race.Ethnicity, sum)
case.vector2 = tapply(mergedstrokeData$Deaths.per.100.000, mergedstrokeData$Race.Ethnicity, sum)
```

```
case.vector #Total number of deaths for this year from heart disease
```

```
##      A/PI      AI/AN      Black Hispanic      White
## 568707.6 300737.5 894024.6 327920.0 1150341.5
```

```
case.vector2 #stroke disease
```

```
##      A/PI      AI/AN      Black Hispanic      White
## 43708.2 18039.4 176585.9 62319.0 237908.1
```

```
#Average number of deaths
```

```
aggregate(mergedheartData$Deaths.per.100.000, list(mergedheartData$Race.Ethnicity), FUN=median)
```

```
#Average number of deaths
```

```
##      Group.1      x
## 1      A/PI 164.15
## 2      AI/AN 360.45
## 3      Black 413.15
## 4 Hispanic 182.10
## 5      White 351.30
```

```
aggregate(mergedstrokeData$Deaths.per.100.000, list(mergedstrokeData$Race.Ethnicity), FUN=median)
```

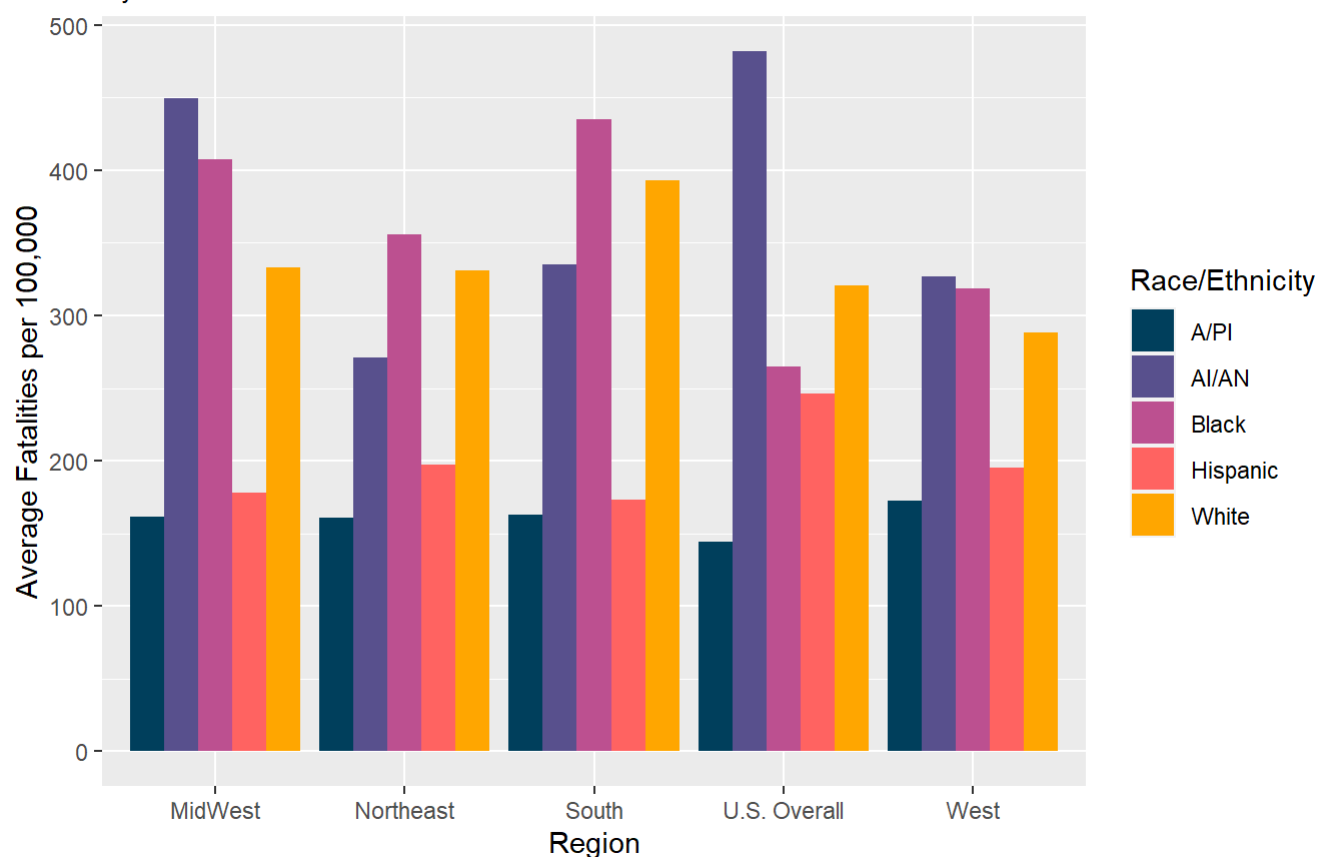
```
##      Group.1      x
## 1      A/PI  61.5
## 2      AI/AN  73.6
## 3      Black 108.1
## 4 Hispanic  57.4
## 5      White  74.3
```

##Additional Data Visualization

```
cbp1 <- c("#003f5c", "#58508d", "#bc5090", "#ff6361", "#ffa600")
x <- ggplot(data = mergedheartData) +
  geom_bar(mapping = aes(x=region, y = Deaths.per.100.000, fill=Race.Ethnicity), position = "dodge", stat = "summary", fun = "median")
x + scale_fill_manual(values = cbp1) + labs(title = "Average Heart Fatalities for Race/Ethnicity by region in the U.S. (2013)",
                                             x = "Region", y = "Average Fatalities per 100,000", subtitle = "By Patrick Asztabski", fill = "Race/Ethnicity")
```


Average Heart Fatalities for Race/Ethnicity by region in the U.S. (2013)

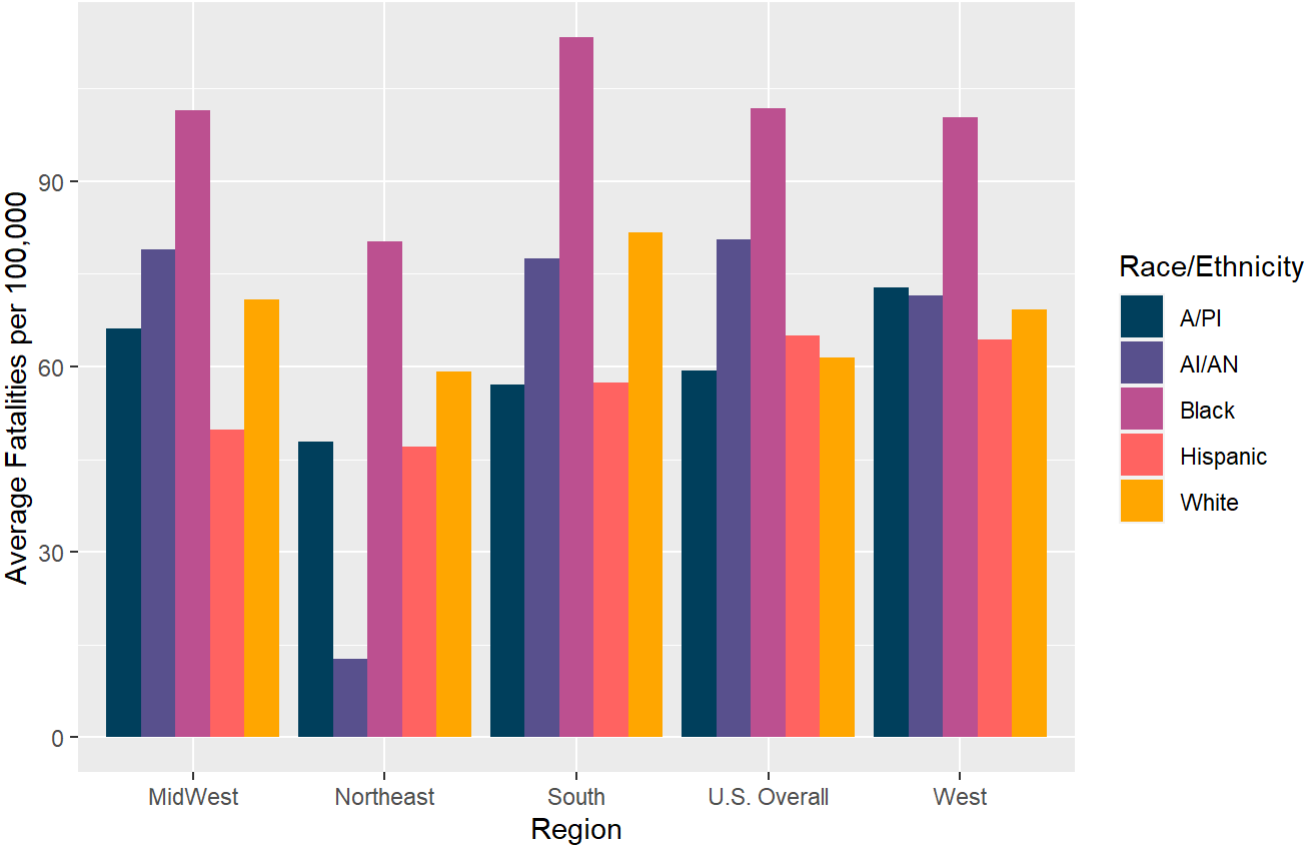
By Patrick Asztabski



```
x <- ggplot(data = mergedstrokeData) +
  geom_bar(mapping = aes(x=region, y = Deaths.per.100.000, fill=Race.Ethnicity), position = "dodge", stat = "summary", fun = "median")
x + scale_fill_manual(values = cbp1) + labs(title = "Average Stroke Fatalities for Race/Ethnicity by region in the U.S. (2013)",
  x = "Region", y = "Average Fatalities per 100,000",
  subtitle = "By Patrick Asztabski", fill = "Race/Ethnicity")
```

Average Stroke Fatalities for Race/Ethnicity by region in the U.S. (2013)

By Patrick Asztabski



```
#####
#RACE/ETHNICITY DENSITY PLOTS FOR HEART
par(mfrow = c(3,2))
hearthist <- hist(whiteheartData$Deaths.per.100.000, breaks = 100, xlab="Deaths per 100,000",
                  ylab="Frequency",main="White",xlim=(c(0,900)), ylim=(c(0,350)), col = "#0bda5
1", cex.lab = 1.25, cex.axis = 1.25)
xfit <- seq(min(whiteheartData$Deaths.per.100.000), max(whiteheartData$Deaths.per.100.000), leng
th = 40)
yfit <- dnorm(xfit, mean=mean(whiteheartData$Deaths.per.100.000), sd = sd(whiteheartData$Deaths.
per.100.000))
yfit <- yfit*diff(hearthist$mids[1:2])*length(whiteheartData$Deaths.per.100.000)
lines(xfit,yfit,lwd=2, col = "black")

hearthist <- hist(BlackheartData$Deaths.per.100.000, breaks = 100, xlab="Deaths per 100,000",
                  ylab="Frequency",main="Black",xlim=(c(0,900)), ylim=(c(0,350)), col = "#ff534
9", cex.lab = 1.25, cex.axis = 1.25)
xfit <- seq(min(BlackheartData$Deaths.per.100.000), max(BlackheartData$Deaths.per.100.000), leng
th = 40)
yfit <- dnorm(xfit, mean=mean(BlackheartData$Deaths.per.100.000), sd = sd(BlackheartData$Deaths.
per.100.000))
yfit <- yfit*diff(hearthist$mids[1:2])*length(BlackheartData$Deaths.per.100.000)
lines(xfit,yfit,lwd=2, col = "black")

hearthist <- hist(APIheartData$Deaths.per.100.000, breaks = 100, xlab="Deaths per 100,000",
                  ylab="Frequency",main="A/PI",xlim=(c(0,900)), ylim=(c(0,350)), col = "#008ecc"
, cex.lab = 1.25, cex.axis = 1.25)
xfit <- seq(min(APIheartData$Deaths.per.100.000), max(APIheartData$Deaths.per.100.000), length =
40)
yfit <- dnorm(xfit, mean=mean(APIheartData$Deaths.per.100.000), sd = sd(APIheartData$Deaths.per.
100.000))
yfit <- yfit*diff(hearthist$mids[1:2])*length(APIheartData$Deaths.per.100.000)
lines(xfit,yfit,lwd=2, col = "black")

hearthist <- hist(AIANheartData$Deaths.per.100.000, breaks = 100, xlab="Deaths per 100,000",
                  ylab="Frequency",main="AI/AN",xlim=(c(0,900)), ylim=(c(0,350)), col = "#9370D
B", cex.lab = 1.25, cex.axis = 1.25)
xfit <- seq(min(AIANheartData$Deaths.per.100.000), max(AIANheartData$Deaths.per.100.000), length
= 40)
yfit <- dnorm(xfit, mean=mean(AIANheartData$Deaths.per.100.000), sd = sd(AIANheartData$Deaths.pe
r.100.000))
yfit <- yfit*diff(hearthist$mids[1:2])*length(AIANheartData$Deaths.per.100.000)
lines(xfit,yfit,lwd=2, col = "black")

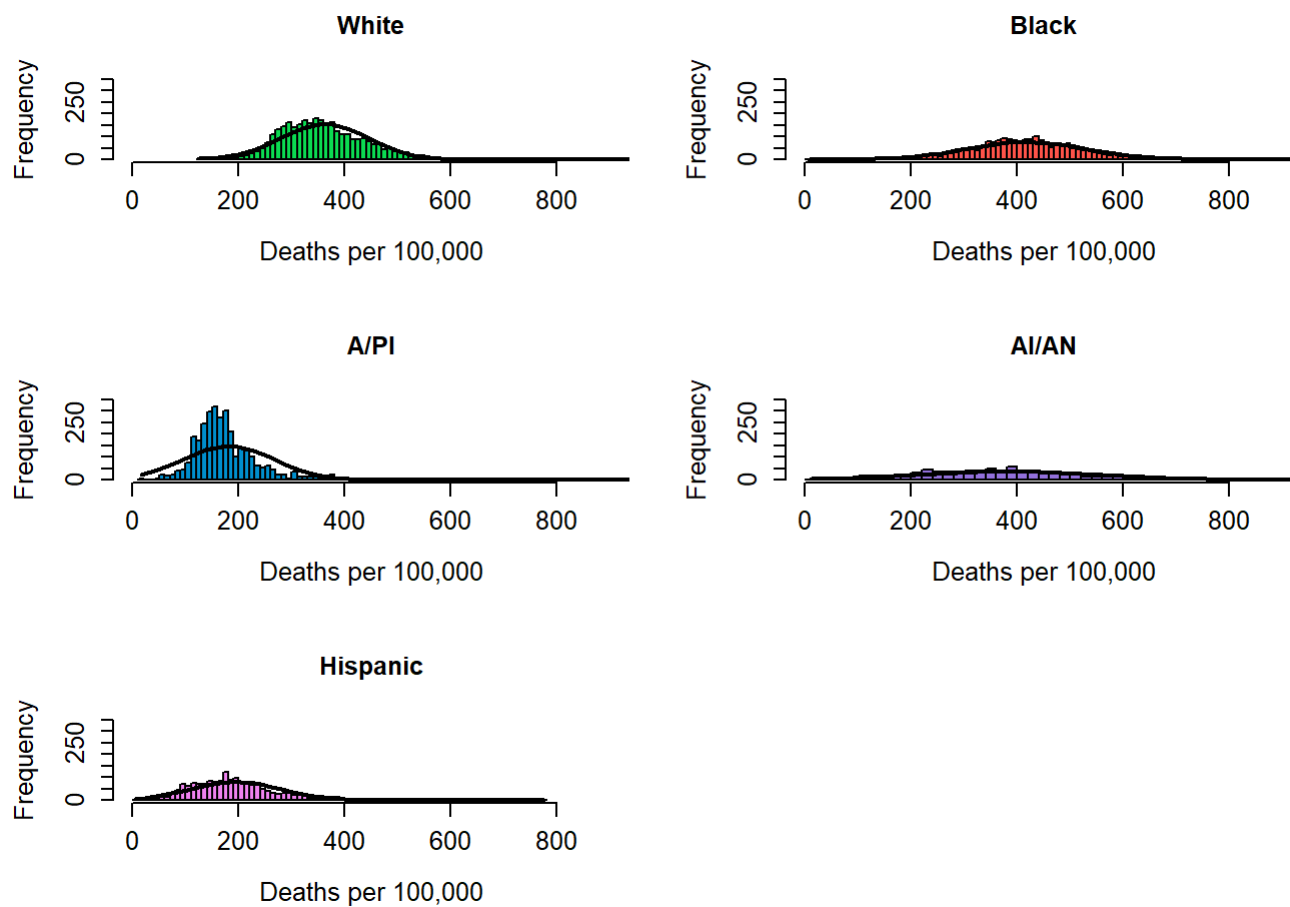
hearthist <- hist(HispanicheartData$Deaths.per.100.000, breaks = 100, xlab="Deaths per 100,000",
                  ylab="Frequency",main="Hispanic",xlim=(c(0,900)), ylim=(c(0,350)), col = "#EE8
2EE", cex.lab = 1.25, cex.axis = 1.25)
xfit <- seq(min(HispanicheartData$Deaths.per.100.000), max(HispanicheartData$Deaths.per.100.00
0), length = 40)
yfit <- dnorm(xfit, mean=mean(HispanicheartData$Deaths.per.100.000), sd = sd(HispanicheartData$D
eaths.per.100.000))
yfit <- yfit*diff(hearthist$mids[1:2])*length(HispanicheartData$Deaths.per.100.000)
lines(xfit,yfit,lwd=2, col = "black")
```

#Histograms for heart fatalities demonstrate wide variance in data, but some unclear results. We can further check this

#by creating qqplots, which we will do later on.

#RACE/ETHNICITY DENSITY PLOTS FOR STROKE

```
par(mfrow = c(3,2))
```



```

strokehist <- hist(whitestrokeData$Deaths.per.100.000, breaks = 15, xlab=" White Deaths per 100,000",
ylab="Frequency",main="Histogram of Deaths per 100,000",xlim=(c(0,250)), ylim=(c(0,1000)), col = "#0bda51")
xfit <- seq(min(whitestrokeData$Deaths.per.100.000), max(whitestrokeData$Deaths.per.100.000), length = 40)
yfit <- dnorm(xfit, mean=mean(whitestrokeData$Deaths.per.100.000), sd = sd(whitestrokeData$Deaths.per.100.000))
yfit <- yfit*diff(strokehist$mids[1:2])*length(whitestrokeData$Deaths.per.100.000)
lines(xfit,yfit,lwd=2, col = "black")

strokehist <- hist(BlackstrokeData$Deaths.per.100.000, breaks = 15, xlab="Black Deaths per 100,000",
ylab="Frequency",main="Histogram of Deaths per 100,000",xlim=(c(0,250)), ylim=(c(0,1000)), col = "#ff5349")
xfit <- seq(min(BlackstrokeData$Deaths.per.100.000), max(BlackstrokeData$Deaths.per.100.000), length = 40)
yfit <- dnorm(xfit, mean=mean(BlackstrokeData$Deaths.per.100.000), sd = sd(BlackstrokeData$Deaths.per.100.000))
yfit <- yfit*diff(strokehist$mids[1:2])*length(BlackstrokeData$Deaths.per.100.000)
lines(xfit,yfit,lwd=2, col = "black")

strokehist <- hist(APIstrokeData$Deaths.per.100.000, breaks = 15, xlab="A/PI Deaths per 100,000",
ylab="Frequency",main="Histogram of Deaths per 100,000",xlim=(c(0,250)), ylim=(c(0,1000)), col = "#008ecc")
xfit <- seq(min(APIstrokeData$Deaths.per.100.000), max(APIstrokeData$Deaths.per.100.000), length = 40)
yfit <- dnorm(xfit, mean=mean(APIstrokeData$Deaths.per.100.000), sd = sd(APIstrokeData$Deaths.per.100.000))
yfit <- yfit*diff(strokehist$mids[1:2])*length(APIstrokeData$Deaths.per.100.000)
lines(xfit,yfit,lwd=2, col = "black")

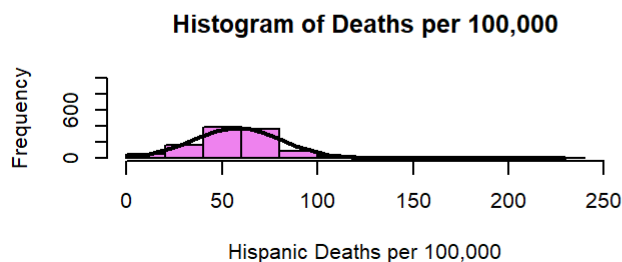
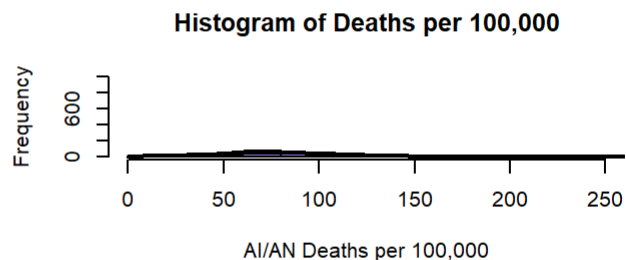
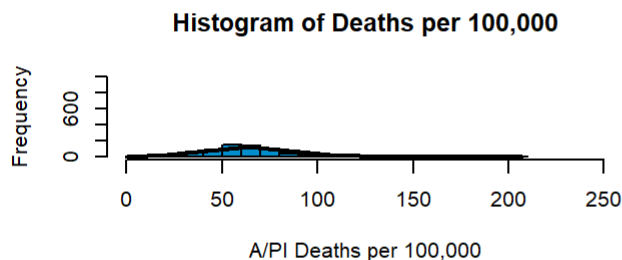
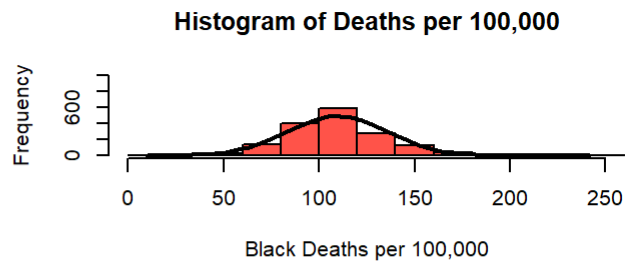
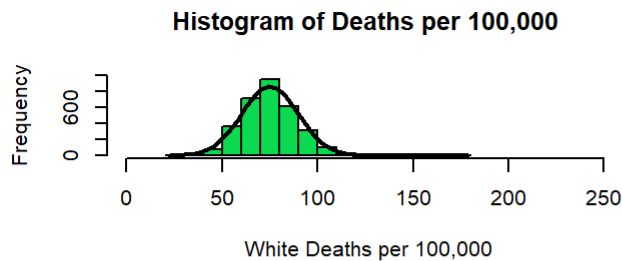
strokehist <- hist(AIANstrokeData$Deaths.per.100.000, breaks = 15, xlab="AI/AN Deaths per 100,000",
ylab="Frequency",main="Histogram of Deaths per 100,000",xlim=(c(0,250)), ylim=(c(0,1000)), col = "#9370DB")
xfit <- seq(min(AIANstrokeData$Deaths.per.100.000), max(AIANstrokeData$Deaths.per.100.000), length = 40)
yfit <- dnorm(xfit, mean=mean(AIANstrokeData$Deaths.per.100.000), sd = sd(AIANstrokeData$Deaths.per.100.000))
yfit <- yfit*diff(strokehist$mids[1:2])*length(AIANstrokeData$Deaths.per.100.000)
lines(xfit,yfit,lwd=2, col = "black")

strokehist <- hist(HispanicstrokeData$Deaths.per.100.000, breaks = 15, xlab="Hispanic Deaths per 100,000",
ylab="Frequency",main="Histogram of Deaths per 100,000",xlim=(c(0,250)), ylim=(c(0,1000)), col = "#EE82EE")
xfit <- seq(min(HispanicstrokeData$Deaths.per.100.000), max(HispanicstrokeData$Deaths.per.100.000), length = 40)
yfit <- dnorm(xfit, mean=mean(HispanicstrokeData$Deaths.per.100.000), sd = sd(HispanicstrokeData$Deaths.per.100.000))

```

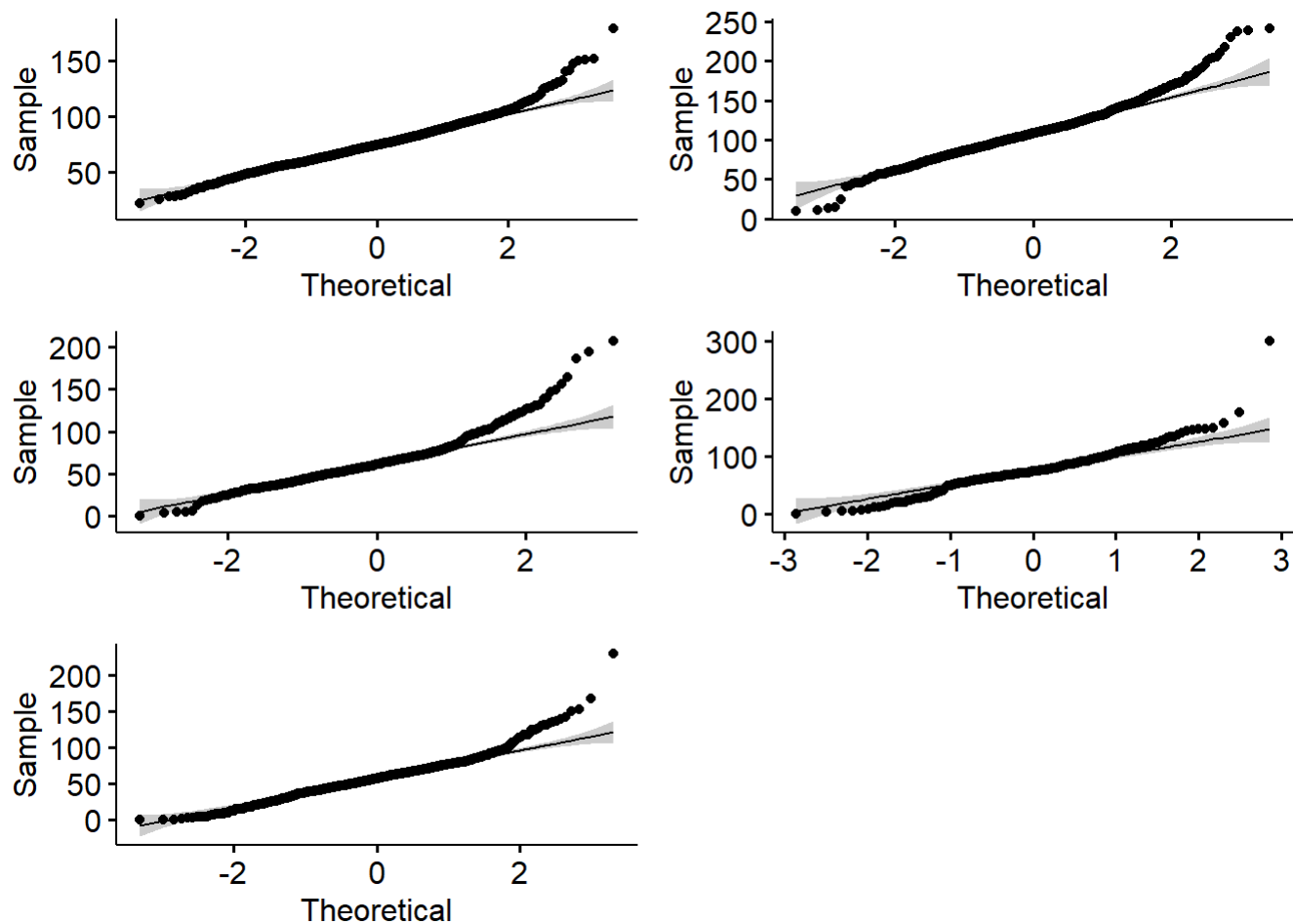
```
yfit <- yfit*diff(strokehist$mids[1:2])*length(HispanicstrokeData$Deaths.per.100.000)
lines(xfit,yfit,lwd=2, col = "black")
```

#Similar results in stroke data as seen in heart data. Let's continue by creating qqplots.



ANOVA TEST

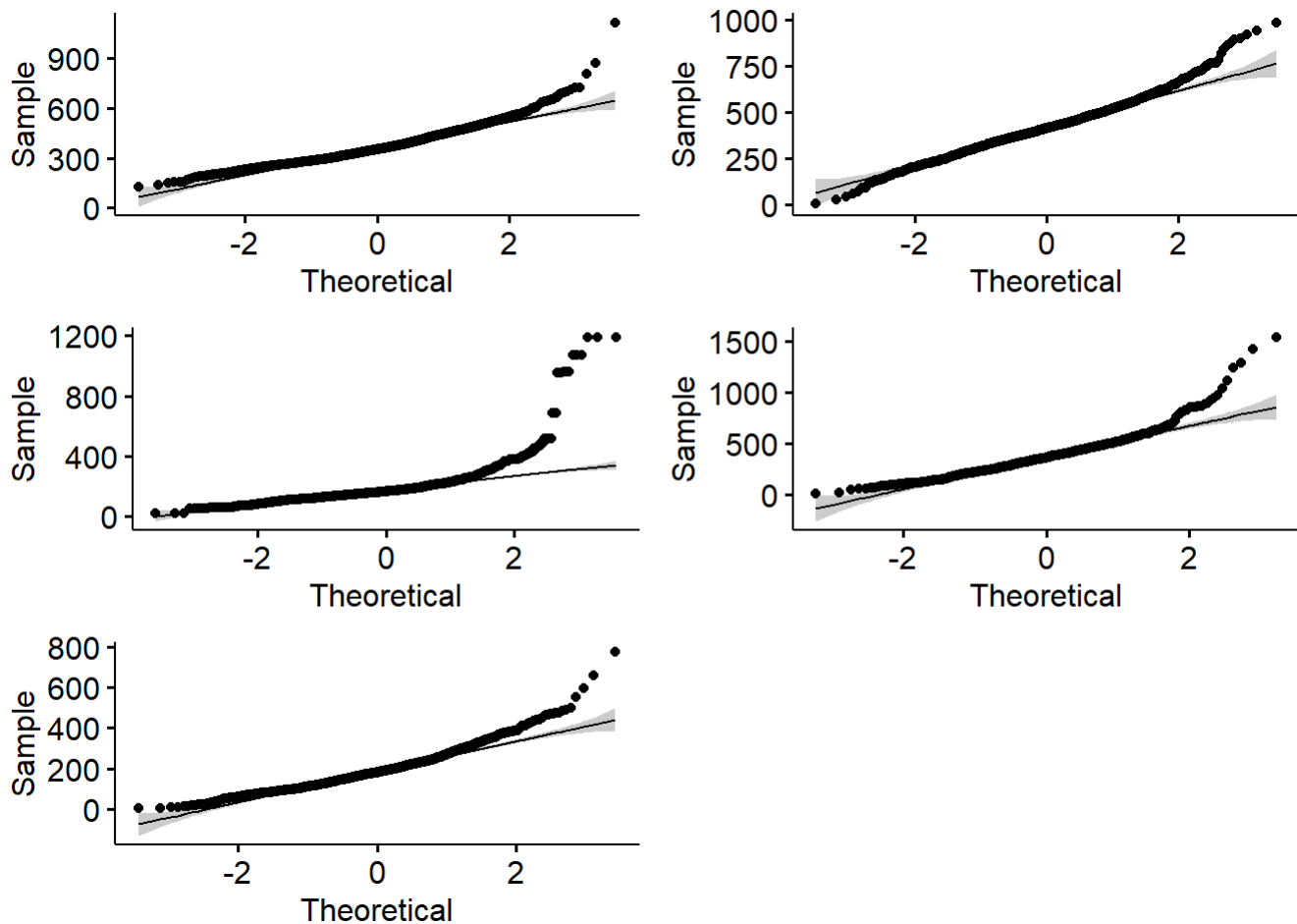
```
#TEST FOR NORMALITY
h1 <- ggqqplot(whitestrokeData$Deaths.per.100.000)
h2 <- ggqqplot(BlackstrokeData$Deaths.per.100.000)
h3 <- ggqqplot(APIstrokeData$Deaths.per.100.000)
h4 <- ggqqplot(AIANstrokeData$Deaths.per.100.000)
h5 <- ggqqplot(HispanicstrokeData$Deaths.per.100.000)
grid.arrange(h1,h2,h3,h4,h5, ncol=2, nrow=3)
```



#Here we can see that distribution along the line starts to trail off above, which signals a right tailed

#distribution, which means our median will reflect our distribution more accurately

```
s1 <- ggqqplot(whiteheartData$Deaths.per.100.000)
s2 <- ggqqplot(BlackheartData$Deaths.per.100.000)
s3 <- ggqqplot(APIheartData$Deaths.per.100.000)
s4 <- ggqqplot(AIANheartData$Deaths.per.100.000)
s5 <- ggqqplot(HispanicheartData$Deaths.per.100.000)
grid.arrange(s1,s2,s3,s4,s5, ncol=2, nrow=3)
```



#Here we can see similar results

#Kruskal Test

Hypotheses from Median Model

H_0 : $me_{A/PI} = me_{AI/AN} = me_{Black} = me_{Hispanic} = me_{White}$

H_1 : at least one of the medians is different

Kruskal test is used to find significant differences in the medians

#of the independent groups (race/ethnicity groups) (deaths per 100,000)

```
kruskal.test(mergedheartData$Deaths.per.100.000 ~ mergedheartData$Race.Ethnicity)
```

```
##
```

```
## Kruskal-Wallis rank sum test
```

```
##
```

```
## data: mergedheartData$Deaths.per.100.000 by mergedheartData$Race.Ethnicity
```

```
## Kruskal-Wallis chi-squared = 6462.3, df = 4, p-value < 2.2e-16
```

```
kruskal.test(mergedstrokeData$Deaths.per.100.000 ~ mergedstrokeData$Race.Ethnicity)
```



```
##
## Kruskal-Wallis rank sum test
##
## data: mergedstrokeData$Deaths.per.100.000 by mergedstrokeData$Race.Ethnicity
## Kruskal-Wallis chi-squared = 2867.4, df = 4, p-value < 2.2e-16
```

```
# In both datasets, p-value is small for both tests, reject H_0.
# Conclude at least one median is different statistically. There is a significant
# difference in the average mortalities for stroke and heart between races.
```

```
#Use poc-host to find exact differences
```

```
dunnTest(mergedheartData$Deaths.per.100.000 ~ mergedheartData$Race.Ethnicity)
```

```
## Warning: mergedheartData$Race.Ethnicity was coerced to a factor.
```

```
## Dunn (1964) Kruskal-Wallis multiple comparison
```

```
## p-values adjusted with the Holm method.
```

##	Comparison	Z	P.unadj	P.adj
## 1	A/PI - AI/AN	-35.126126	2.691385e-270	1.614831e-269
## 2	A/PI - Black	-64.002246	0.000000e+00	0.000000e+00
## 3	AI/AN - Black	-9.631743	5.872472e-22	1.761742e-21
## 4	A/PI - Hispanic	-4.517915	6.245153e-06	1.249031e-05
## 5	AI/AN - Hispanic	29.268923	2.579465e-188	1.289732e-187
## 6	Black - Hispanic	50.977150	0.000000e+00	0.000000e+00
## 7	A/PI - White	-58.341778	0.000000e+00	0.000000e+00
## 8	AI/AN - White	-1.873899	6.094436e-02	6.094436e-02
## 9	Black - White	11.641725	2.528504e-31	1.011402e-30
## 10	Hispanic - White	-44.274655	0.000000e+00	0.000000e+00

```
dunnTest(mergedstrokeData$Deaths.per.100.000 ~ mergedstrokeData$Race.Ethnicity)
```

```
## Warning: mergedstrokeData$Race.Ethnicity was coerced to a factor.
```

```
## Dunn (1964) Kruskal-Wallis multiple comparison
```

```
## p-values adjusted with the Holm method.
```

##	Comparison	Z	P.unadj	P.adj
## 1	A/PI - AI/AN	-7.08720223	1.368503e-12	4.105510e-12
## 2	A/PI - Black	-37.04456202	2.197016e-300	1.757613e-299
## 3	AI/AN - Black	-16.68971186	1.557205e-62	9.343230e-62
## 4	A/PI - Hispanic	4.07014593	4.698370e-05	9.396739e-05
## 5	AI/AN - Hispanic	10.23398627	1.396500e-24	5.586000e-24
## 6	Black - Hispanic	48.12822610	0.000000e+00	0.000000e+00
## 7	A/PI - White	-12.65524429	1.046495e-36	5.232477e-36
## 8	AI/AN - White	-0.01184307	9.905508e-01	9.905508e-01
## 9	Black - White	37.89042074	0.000000e+00	0.000000e+00
## 10	Hispanic - White	-20.82441659	2.600738e-96	1.820517e-95

In both datasets, p-value is small for every comparison, reject H_0 .
 # Conclude median is different statistically for each group. There is a significant
 # difference in the median mortalities for stroke and heart between races.

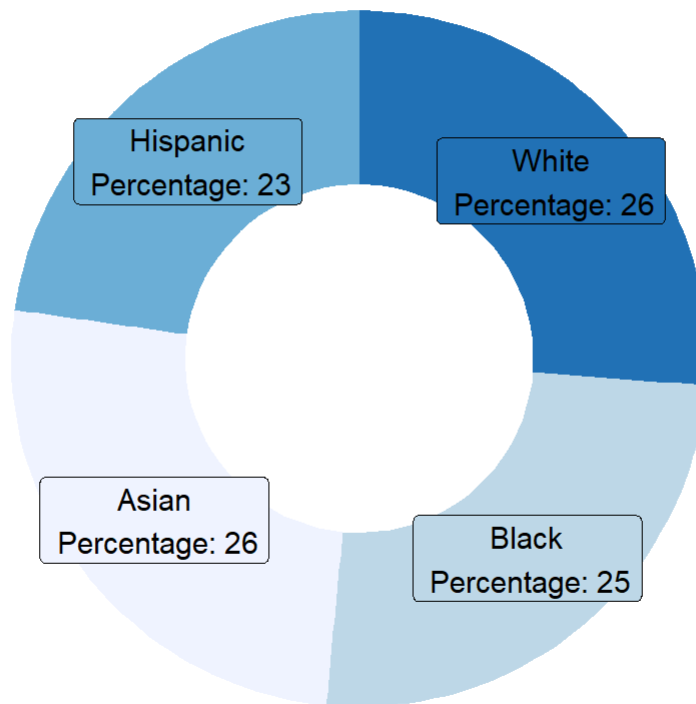
Why do we see differences in median mortalities?

There are a multitude of factors including health, socioeconomic, and financial problems. Here we will demonstrate just a few factors we think are significant enough to include

```
#INSURANCE RATES PIE GRAPH
#####
ggplot(Insurance, aes(ymax=ymax, ymin=ymin, xmax=4, xmin=3, fill = Race)) +
  geom_rect() +
  geom_label(x=3.5, aes(y=labelPosition, label=label), size=4) +
  coord_polar(theta="y") +
  scale_fill_brewer("Blues") +
  xlim(c(2,4)) +
  theme_void() +
  theme(legend.position = "none") +
  labs(title = "Uninsured by Race/Ethnicity (2013)", subtitle = "By Patrick Asztabski")
```

Uninsured by Race/Ethnicity (2013)

By Patrick Asztabski



Here we can see significant portion of the population by race who still are uninsured and therefore tend to avoid Healthcare solutions due to inheriting costs they would have to cover for themselves.

```
#DIABETES PREVALENCE PIE GRAPH
```

```
#####
```

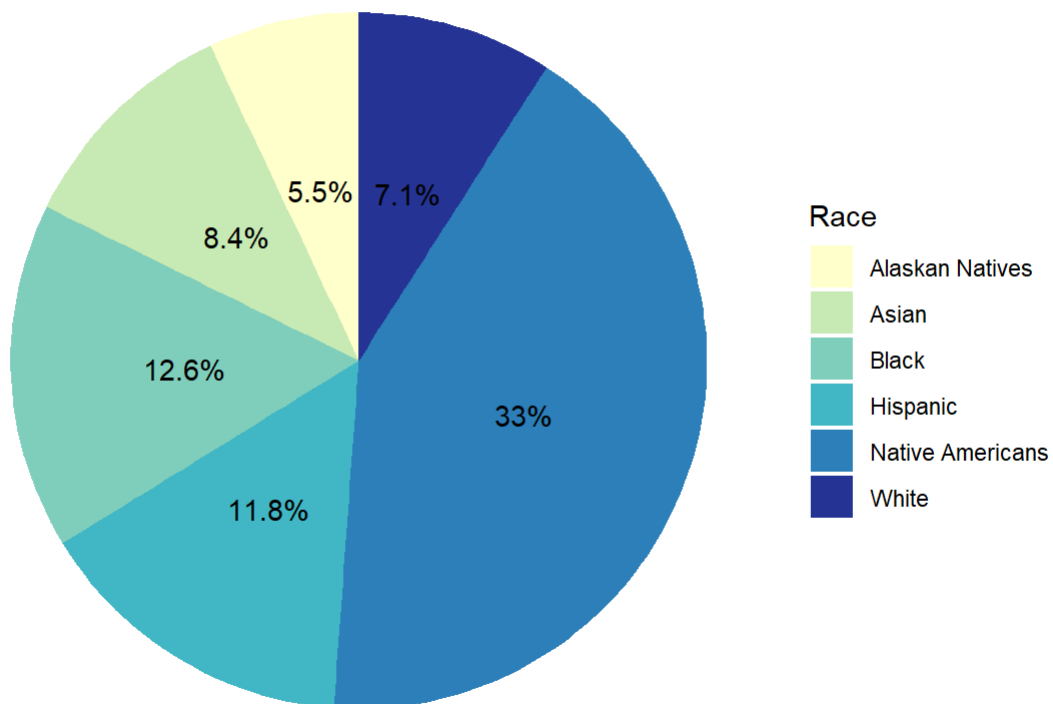
```
x <- ggplot(data = Diabetes, aes(x="", y = DiabetesPrevalence, fill = Race)) +
  geom_col(stat = "identity") +
  geom_text(aes(label=labels), position = position_stack(vjust = 0.5)) +
  coord_polar(theta = "y") +
  theme_void() +
  scale_fill_brewer(palette = "YlGnBu") +
  labs(title = "Diabetes Prevalence by Race/Ethnicity in Adults >= 20 years of age (2013)", subtit
  itle = "By Patrick Asztabski")
```

```
## Warning: Ignoring unknown parameters: stat
```

```
x
```

Diabetes Prevalence by Race/Ethnicity in Adults ≥ 20 years of age (2013)

By Patrick Asztabski



Here we can see a significant difference between American Indians/American Natives and any other group, which is a major factor in explaining the significant differences in overall heart and stroke fatalities between AI/AN and any other group. People with diabetes are twice as likely to have heart disease or a stroke.