Stroke Mortality - ANOVA

Author: Marissa Munoz-Ruiz

Goal of Script: Explore ethnicity in stroke mortality

Is there is a difference in stroke mortality among different ethnic groups? Does stroke mortality differ in various regions of the U.S.?

The U.S. government has publicly accessible data on stroke mortality on data.gov. I downloaded the excel file for 2016 (Stroke Mortality (https://chronicdata.cdc.gov/Heart-Disease-Stroke-Prevention/Stroke-Mortality-Data-Among-US-Adults-35-by-State-/v246-z5tb)). Since the stroke dataset gives the stroke mortality rate across ethnicity and gender at the state level, I also downloaded an additional dataset that breaks down states into different regions (Regions (https://www.kaggle.com/omer2040/usa-states-to-region)).

I will use ggplot2 to visually explore the data and then conduct an ANOVA analysis to determine if there is a statistical difference between groups. Note: All plots use the same color schemes for consistent representation across ethnic groups and regions.

```
knitr::opts_chunk$set(echo = TRUE, message = FALSE, warning = FALSE)

library(readr)
library(dplyr)
library(ggplot2)
library(rstatix)
```

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

```r
# Load Data (only 2016)
file <- "C:/Users/mmuno/Desktop/GitHub/ores5310-2020 (Stats)/Data"
Data <- read_csv(paste(file,'/StrokeMortality_2016.csv',sep=''))
USRegions <- read_csv(paste(file,'/State_Regions.csv',sep=''))

# Define colors for plotting
EthnList = c("Asian and Pacific Islander","Black","Hispanic","White")
EthnColor = c("darkorange","purple","red","darkgreen")
RegList = c("Midwest","Northeast","South","West")
RegColor= c("gray30","darkgoldenrod3","deeppink","darkcyan")

List = c("Midwest"="gray30","Northeast"="darkgoldenrod3","South"="deeppink","West"="darkcyan","Asian and Pacific Islander"="darkorange","Black"="purple","Hispanic"="red","White"="darkgreen")

# Trim Data
USRegions <- USRegions %>%
  select(-State) %>%
  rename(StateCode='State Code')

StrokeData <- Data %>%
  filter(GeographicLevel=="State",!is.na(Data_Value)) %>%
  select(LocationAbbr, Data_Value, Stratification1, Stratification2) %>%
  rename(StateCode=LocationAbbr, MortalityRate=Data_Value, Sex=Stratification1, Ethnicity=Stratification2)%>%
  left_join(USRegions,"StateCode")

#Basic statistics
Stroke_Stats <- StrokeData %>%
  select(Region, Ethnicity, MortalityRate) %>%
  filter(Ethnicity!="Overall", Ethnicity!="American Indian and Alaskan Native") %>%
  mutate(Region=factor(Region,levels=RegList), Ethnicity=factor(Ethnicity,levels=EthnList)) %>%
  group_by(Region, Ethnicity) %>%
  summarize(MR=round(mean(MortalityRate, na.rm=TRUE),3),
            sd=round(sd(MortalityRate),3),
            skew=round(e1071::skewness(MortalityRate),3),
            N=n()) %>%
  arrange(Region, Ethnicity)

Stroke_Stats
```

```
## # A tibble: 16 x 6
## # Groups:   Region [4]
##    Region    Ethnicity                     MR    sd   skew     N
##    <fct>     <fct>                      <dbl> <dbl>  <dbl> <int>
##  1 Midwest   Asian and Pacific Islander  66.7 19.3   1.32     26
##  2 Midwest   Black                      100.   9.80  0.436    30
##  3 Midwest   Hispanic                    54.2 11.8   0.105    28
##  4 Midwest   White                       68.8  5.95  0.143    36
##  5 Northeast Asian and Pacific Islander  45.8  8.95 -0.001    14
##  6 Northeast Black                       74.2 17.5   0.593    18
##  7 Northeast Hispanic                    48.8 10.1   0.711    17
##  8 Northeast White                       57.6  6.94  0.742    27
##  9 South     Asian and Pacific Islander  60.2  9.26  0.152    37
## 10 South     Black                      109.  17.2   0.111    51
## 11 South     Hispanic                    47.3 13.4   0.951    37
## 12 South     White                       78.4  8.98 -1.21     51
## 13 West      Asian and Pacific Islander  70.9 14.9   1.72     29
## 14 West      Black                       96.1  8.24  0.178    21
## 15 West      Hispanic                    67.8 15.2   1.84     31
## 16 West      White                       66.6  5.51 -0.378    39
```

At first glance, the data shows that Black ethnic group has the highest mortality rates in the South and in the Midwest.

Note: The American Indian and Alaskan Native ethnicity group was removed from the data because mortality rate for this ethnic group was not listed at all levels of region. There were two states that were not listed in the Regions file and those were also removed.

Data Visualization

```
# Bar graph - Ethnicity & Region
Stroke_Stats %>%
ggplot(aes(x = Region, y=MR, fill=Ethnicity)) +
  geom_col(width=0.65, position=position_dodge(0.8), colour="black") +
  geom_errorbar(aes(ymin=MR-sd, ymax=MR+sd),width=0.2,position=position_dodge(0.8))+
  scale_fill_manual(values=EthnColor) +
  labs(x="Region", y="Mortality Rate", title="Fig 1: Stroke Mortality among Ethnic Groups & Regi
on") +
  theme(plot.title = element_text(hjust=0.5))
```

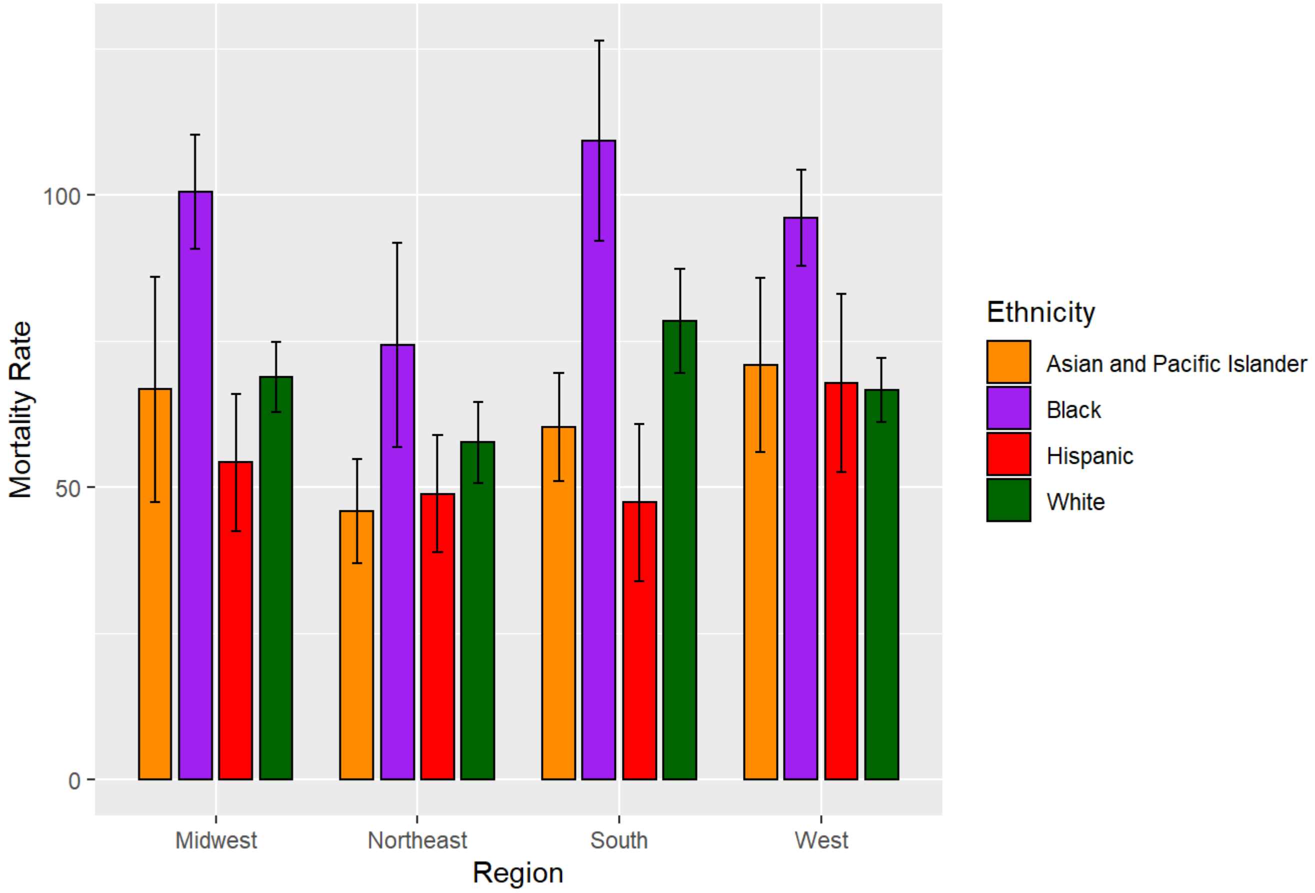
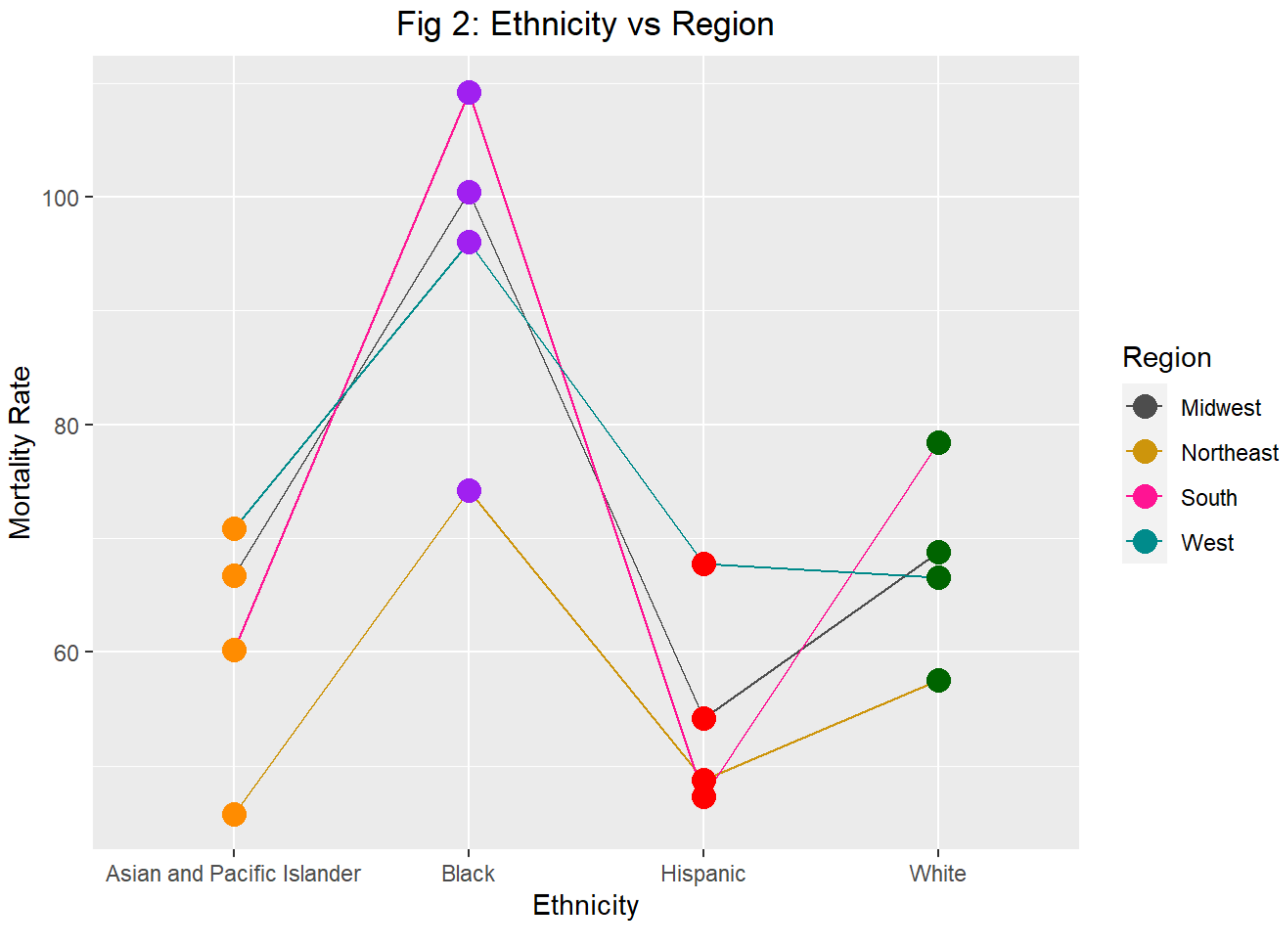
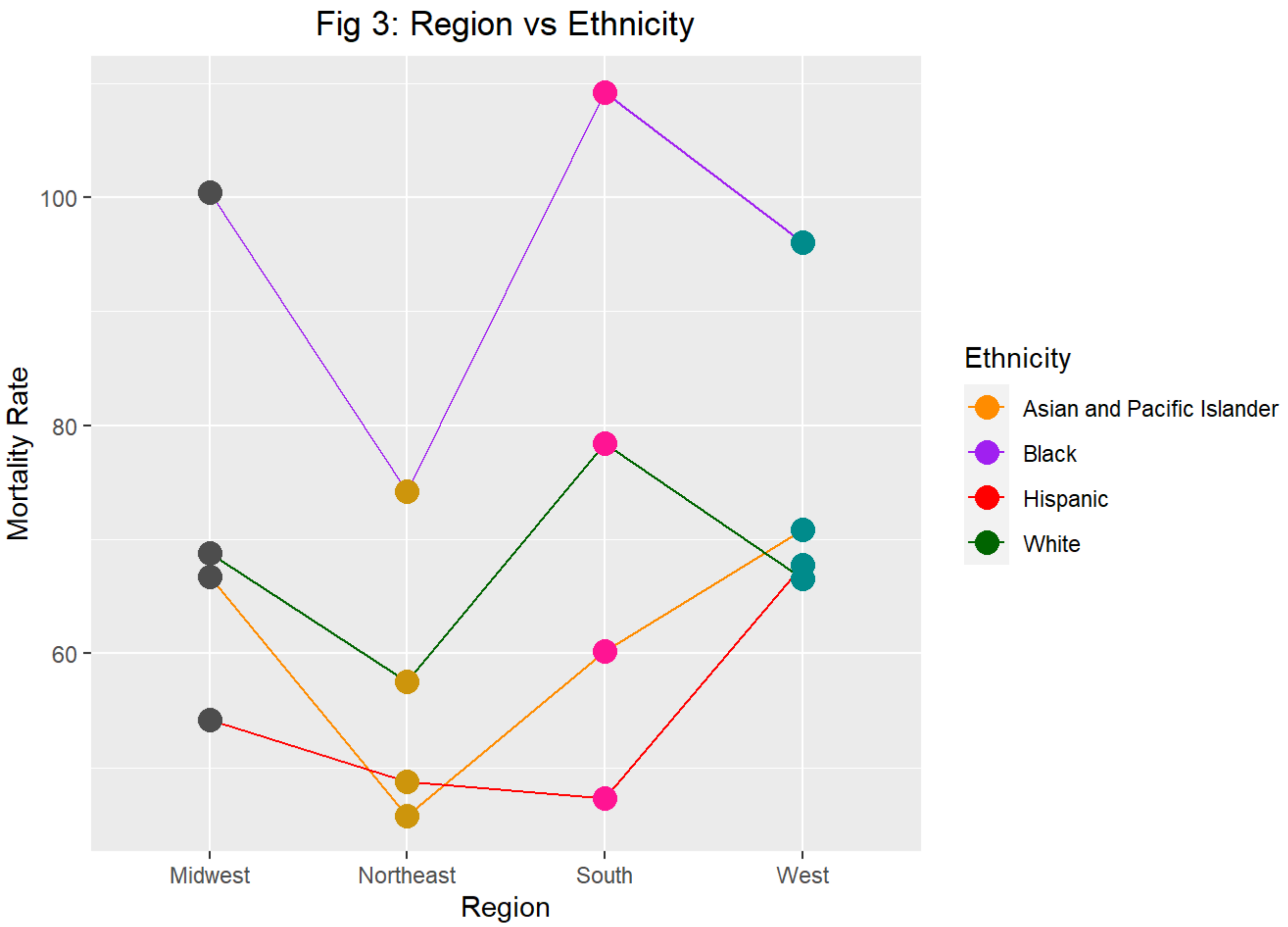


Fig 1: Stroke Mortality among Ethnic Groups & Region

```
# Interaction Plot
Stroke_Stats %>%
  ggplot(aes(x = Ethnicity, y = MR)) +
  geom_line(aes(group=Region,color=Region)) +
  geom_point(aes(color=Ethnicity),size=4) +
  scale_color_manual(values=List,breaks = RegList) +
  labs(y="Mortality Rate", title="Fig 2: Ethnicity vs Region") +
  theme(plot.title = element_text(hjust=0.5))
```



```
# Interaction Plot
Stroke_Stats %>%
  ggplot(aes(x = Region, y = MR)) +
  geom_line(aes(group=Ethnicity,color=Ethnicity)) +
  geom_point(aes(color=Region),size=4) +
  scale_color_manual(values=List, breaks=EthnList) +
  labs(y="Mortality Rate", title="Fig 3: Region vs Ethnicity") +
  theme(plot.title = element_text(hjust=0.5))
```



There seem to be a difference in stroke mortality rate among various ethnic groups and regions (see Fig. 1). The interaction plots indicate there is a complex interaction occurring between ethnic groups and region (see Fig. 2 and Fig. 3).

```
# Trim data for stats
Stroke_Analysis <- StrokeData %>%
  select(Region, Ethnicity, MortalityRate) %>%
  filter(Ethnicity!="Overall", Ethnicity!="American Indian and Alaskan Native") %>%
  rename(MR=MortalityRate) %>%
  arrange(Region, Ethnicity)

# Ethnicity distribution
Stroke_Analysis %>%
ggplot(aes(x=MR)) +
    geom_histogram(binwidth=5) +
    labs(x="Mortality Rate", title = "Fig 4: Distribution of Ethnicity") +
    theme(plot.title = element_text(hjust=0.5)) +
    facet_wrap(~Ethnicity)
```

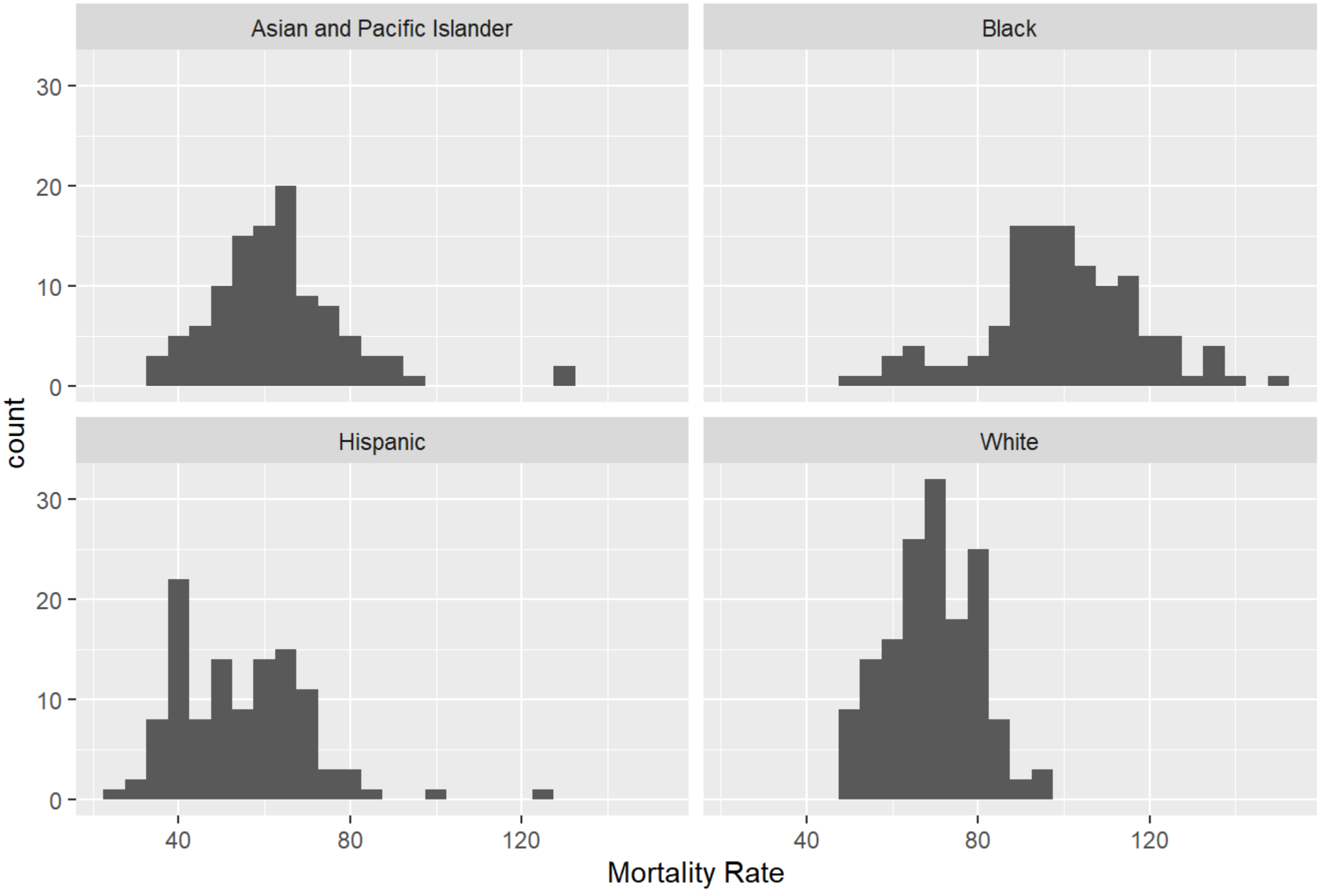
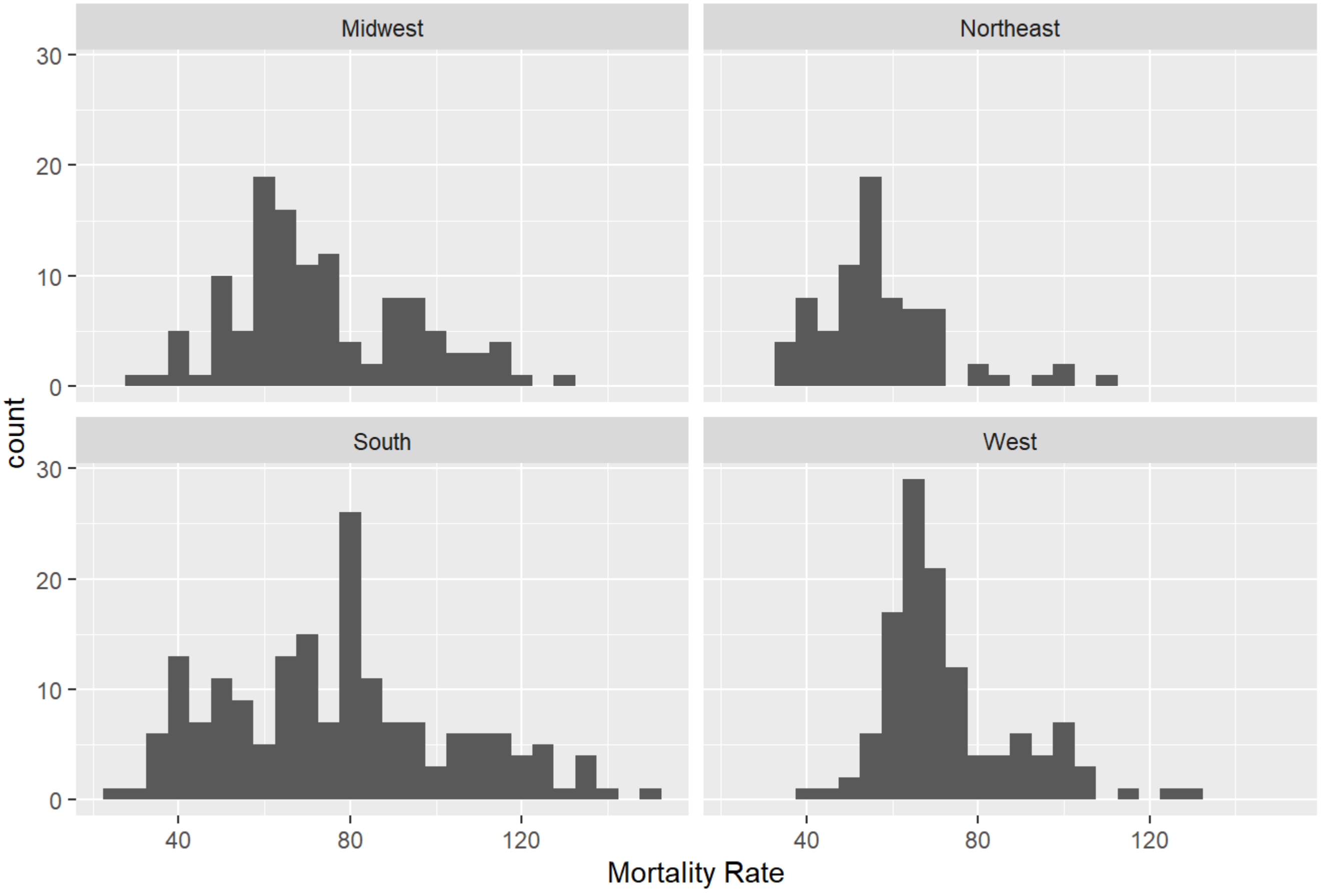


Fig 4: Distribution of Ethnicity

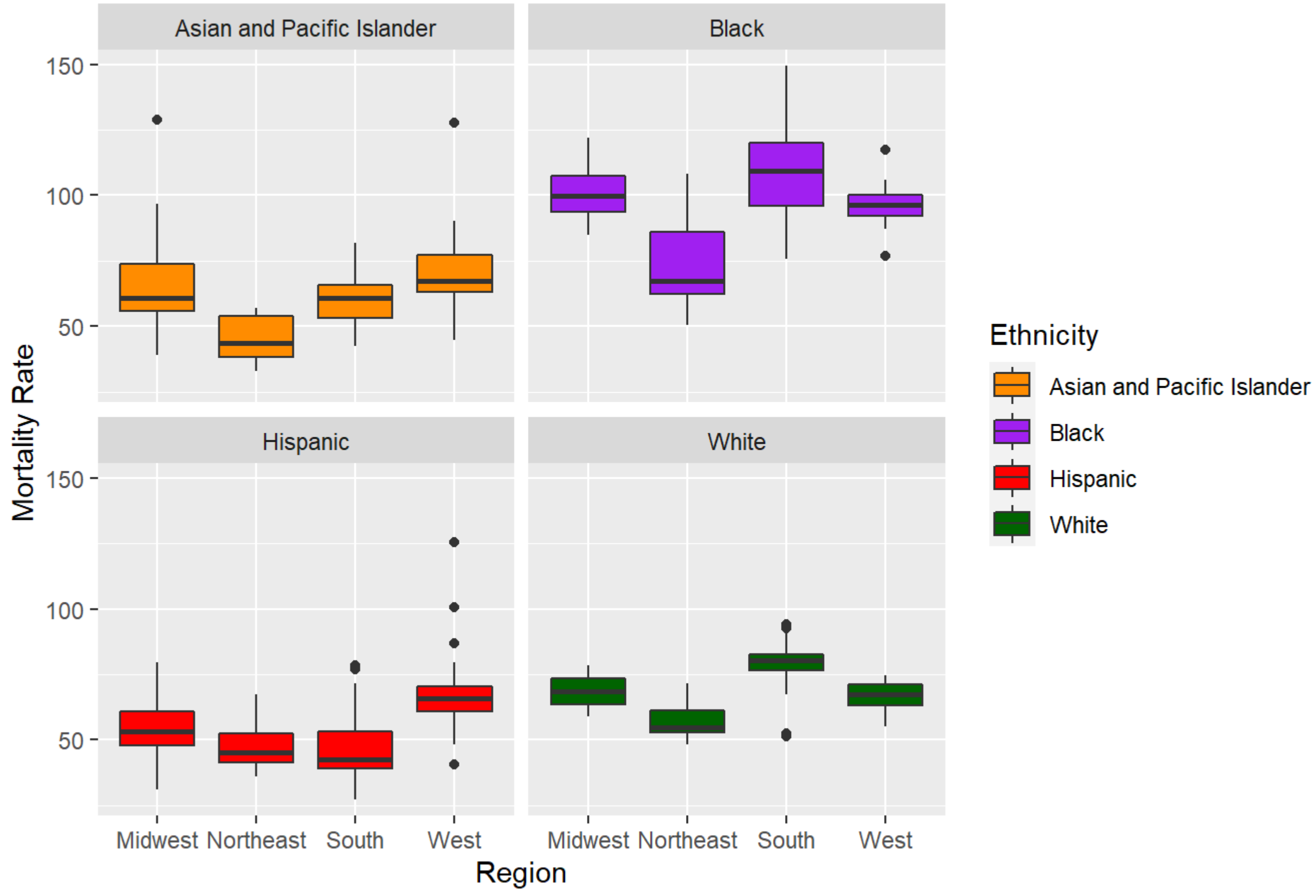
```
# Region distribution
Stroke_Analysis %>%
ggplot(aes(x=MR)) +
    geom_histogram(binwidth=5) +
    labs(x="Mortality Rate", title = "Fig 5: Distribution of Region") +
    theme(plot.title = element_text(hjust=0.5)) +
    facet_wrap(~Region)
```

Fig 5: Distribution of Region



```
# Box plot of Ethnicity & Region
Stroke_Analysis %>%
ggplot(aes(x=Region, y=MR, fill=Ethnicity)) +
    geom_boxplot() +
    labs(y="Mortality Rate", title = "Fig 6: Ethnicity & Region") +
    scale_fill_manual(values=EthnColor) +
    theme(plot.title = element_text(hjust=0.5)) +
    facet_wrap(~Ethnicity)
```

Fig 6: Ethnicity & Region



From the box plots (see Fig 6.), it appears there is a higher stroke mortality rate among Black and White observations in the south region than any other ethnic group or region. Additionally, there seems to be outliers, namely Hispanic observations with really high stroke mortality rates in the West region. There also seems to be a lower stroke mortality rate among Asian and Pacific Islander observations in the Northeast regions compared to other groups and regions.

Hypotheses

Ethnicity and region are both independent categorical variables and since I am interested in seeing if there are differences between and among groups then a two way ANOVA is appropriate.

With informal notation, the null hypothesis for the two way ANOVA with interaction tests are:

1. Group means at any level of Ethnicity are all equal:

$$H_0 : \mu_{Alaskan} = \mu_{Black} = \mu_{Hispanic} = \mu_{White}$$

2. Group means at any level of Region are all equal:

$$H_0 : \mu_{Midwest} = \mu_{Northeast} = \mu_{South} = \mu_{West}$$

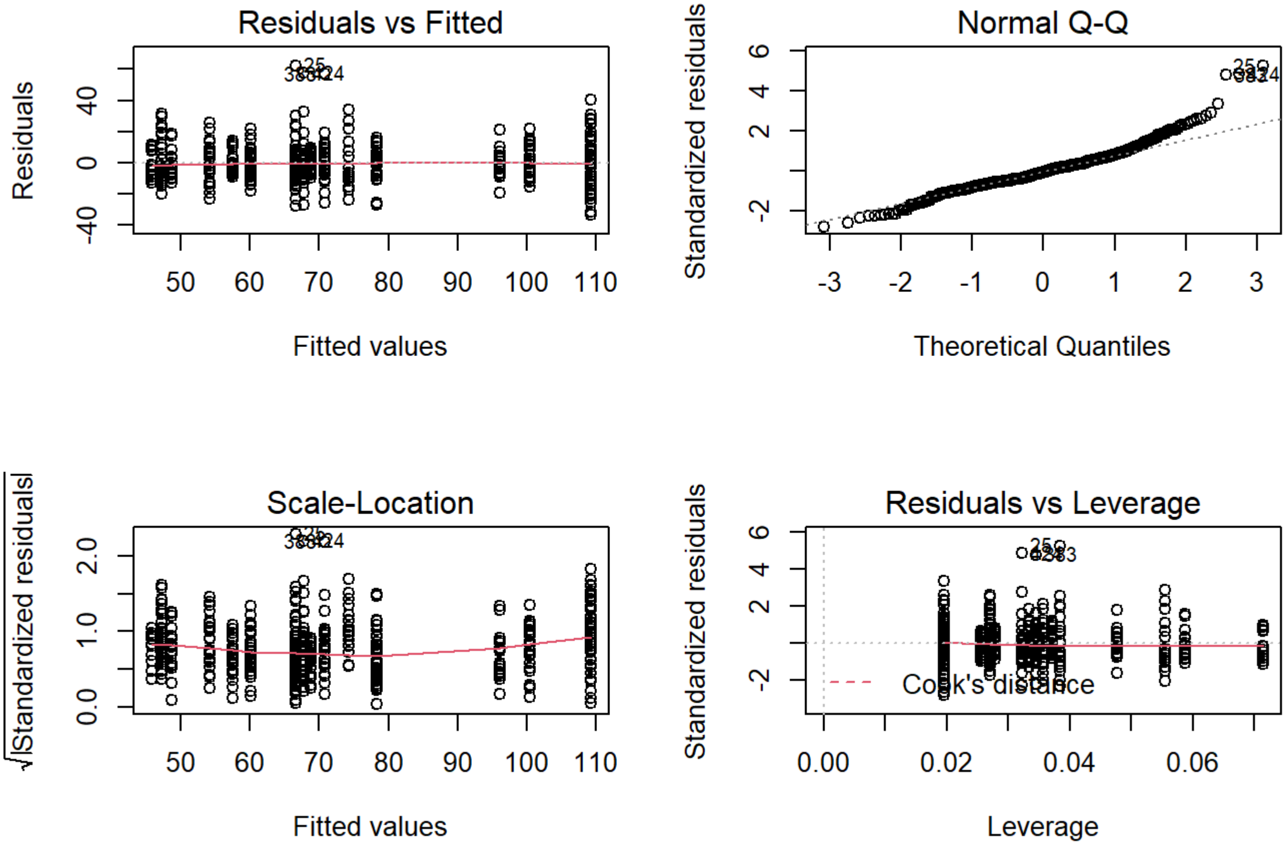
3. There is no interaction effect i.e. effect of Ethnicity does not depend on the effect of Region and vice versa.

The alternative hypothesis for the two way ANOVA with interaction tests are: 1) Group means at any level of Ethnicity are not equal 2) Group means at any level of Region are not equal 3) There is an interaction effect between Ethnicity and Region

```
# Two Way ANOVA Model
Model <- lm(MR ~ Region*Ethnicity + Region + Ethnicity, data=Stroke_Analysis)
anova(Model)
```

```
## Analysis of Variance Table
##
## Response: MR
##                   Df Sum Sq Mean Sq F value    Pr(>F)
## Region             3  20795    6932  47.017 < 2.2e-16 ***
## Ethnicity          3 133141   44380 301.027 < 2.2e-16 ***
## Region:Ethnicity   9  19221    2136  14.486 < 2.2e-16 ***
## Residuals        476  70176     147
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Assumptions
par(mfrow=c(2,2))
plot(Model)
```



```
shapiro.test(residuals(Model))
```

```
## 
##  Shapiro-Wilk normality test
## 
## data:  residuals(Model)
## W = 0.94949, p-value = 6.47e-12
```

```
Stroke_Analysis %>%
  levene_test(MR ~ Region*Ethnicity) #rstatix function
```

```
## # A tibble: 1 x 4
##     df1   df2 statistic           p
##   <int> <int>     <dbl>       <dbl>
## 1    15   476      4.14 0.000000350
```

```
# Outliers
Stroke_Analysis %>%
  group_by(Region, Ethnicity) %>%
  identify_outliers(MR)
```

```
## # A tibble: 16 x 5
##    Region  Ethnicity                     MR is.outlier is.extreme
##    <chr>   <chr>                      <dbl> <lgl>      <lgl>     
##  1 Midwest Asian and Pacific Islander 129.  TRUE       TRUE      
##  2 South   Hispanic                    78.4 TRUE       FALSE     
##  3 South   Hispanic                    76.8 TRUE       FALSE     
##  4 South   White                       94   TRUE       FALSE     
##  5 South   White                       92.6 TRUE       FALSE     
##  6 South   White                       51.4 TRUE       TRUE      
##  7 South   White                       52.5 TRUE       TRUE      
##  8 South   White                       94.2 TRUE       FALSE     
##  9 South   White                       52   TRUE       TRUE      
## 10 West    Asian and Pacific Islander 128.  TRUE       TRUE      
## 11 West    Black                      117.  TRUE       FALSE     
## 12 West    Black                       76.7 TRUE       FALSE     
## 13 West    Hispanic                   125.  TRUE       TRUE      
## 14 West    Hispanic                   101.  TRUE       TRUE      
## 15 West    Hispanic                    86.7 TRUE       FALSE     
## 16 West    Hispanic                    40.8 TRUE       FALSE
```

Model Decision

Since the p-value of the interaction and each factor are all less than our set alpha level of 0.05, we reject all of the null hypotheses.

Interpretation

A two-way ANOVA test indicated that the effect of Ethnicity on stroke mortality rate is dependent on the levels of Region and vice versa (F(9,476) = 14.49, p<0.05). Furthermore, there was a main effect for Ethnicity (F(3,476) = 301.03, p<0.05) and Region (F(3,476) = 47.01, p<0.05).

Assumptions

In the QQ plot, the residuals look fairly normal. However, the p-value in the Shapiro-Wilk test of normality for the residuals of the model (p=6.47e-12) is significant so normality can’t be assumed. In the Residuals vs Fitted plot, there is no relationship between residuals and fitted values so we can assume the homogeneity of variances. The p-value in the Levene Test is significant (p=3.5e-07), which means we can not assume the homogeneity of variances between Ethnicity and Region. There are three potential outliers in the top right hand corner of the QQ plot.

Pairwise Comparisons

```
# Pairwise comparisons
PWC <- Stroke_Analysis %>%
  group_by(Region) %>%
  emmeans_test(MR ~ Ethnicity, p.adjust.method = "bonferroni") #rstatix function

PWC
```

```
## # A tibble: 24 x 10
##    Region term  .y.   group1 group2    df statistic        p    p.adj
##  * <chr>  <chr> <chr> <chr>  <chr>  <dbl>     <dbl>    <dbl>    <dbl>
##  1 Midwe~ Ethn~ MR    Asian~ Black    476   -10.4   6.78e-23 4.07e-22
##  2 Midwe~ Ethn~ MR    Asian~ Hispa~   476     3.79  1.73e- 4 1.04e- 3
##  3 Midwe~ Ethn~ MR    Asian~ White    476    -0.678 4.98e- 1 1.00e+ 0
##  4 Midwe~ Ethn~ MR    Black  Hispa~   476    14.5   9.26e-40 5.55e-39
##  5 Midwe~ Ethn~ MR    Black  White    476    10.5   1.64e-23 9.84e-23
##  6 Midwe~ Ethn~ MR    Hispa~ White    476    -4.78  2.30e- 6 1.38e- 5
##  7 North~ Ethn~ MR    Asian~ Black    476    -6.57  1.29e-10 7.74e-10
##  8 North~ Ethn~ MR    Asian~ Hispa~   476    -0.683 4.95e- 1 1.00e+ 0
##  9 North~ Ethn~ MR    Asian~ White    476    -2.94  3.40e- 3 2.04e- 2
## 10 North~ Ethn~ MR    Black  Hispa~   476     6.20  1.24e- 9 7.46e- 9
## # ... with 14 more rows, and 1 more variable: p.adj.signif <chr>
```

There was a significant difference of stroke mortality rate between all ethnic groups for all levels of Region (p < 0.05).