Appendix

```
---
title: "Final Project: Final Project Ethnicity vs SexAge"
output: html_notebook
---

First, we need to install and load packages

```{r}
install.packages("dunn.test", dependencies = TRUE)
install.packages("reshape2")

```

```{r}
packages <- c('tidyverse','car','dunn.test','tableone','lmtest','broom')
```

```{r}
purrr::walk(packages,library,character.only=T)

```

We will load the "cancer.csv" file into R Studio. Then assign the file to a new variable "cancer"

```{r}
#load packages
library(package = tidyverse)
library(readr)
library("reshape2")
library(ggplot2)
library(car)
```

```{r}
cancer <- read_csv("cancer.csv")
head(cancer)
```

A tibble: 6 × 75

State <chr>	Total.Rate <dbl>	Total.Number <dbl>	Total.Population <dbl>	Rates.Age.< 18 <dbl>	Rates.Age.18-45 <dbl>
Alabama	214.2	71529	33387205	2.0	18.5
Alaska	128.1	6361	4966180	1.7	11.8
Arizona	165.6	74286	44845598	2.5	13.6
Arkansas	223.9	45627	20382448	2.3	17.6
California	150.9	393980	261135696	2.6	13.7
Colorado	139.0	49035	35267734	1.9	11.7

6 rows | 1-6 of 75 columns

```{r}
cancer.cleaned <- cancer %>%

#rename ethnicty
  rename(Types.Colorectal.Race.White.non.Hispanic = "Types.Colorectal.Race.White non-Hispanic") %>%
  rename(Types.Colorectal.Race.Black.non.Hispanic = "Types.Colorectal.Race.Black non-Hispanic") %>%

#rename age/gender
  rename(Types.Colorectal.Age.Male.Above.64 = "Types.Colorectal.Age and Sex.Male.> 64") %>%
  rename(Types.Colorectal.Age.Female.Above.64 = "Types.Colorectal.Age and Sex.Female.> 64") %>%

#recode zero value into NA
  mutate(Types.Colorectal.Race.White = na_if(x = Types.Colorectal.Race.White, y = 0)) %>%
  mutate(Types.Colorectal.Race.White.non.Hispanic = na_if(x = Types.Colorectal.Race.White.non.Hispanic, y = 0)) %>%
  mutate(Types.Colorectal.Race.Black = na_if(x = Types.Colorectal.Race.Black, y = 0)) %>%
  mutate(Types.Colorectal.Race.Black.non.Hispanic = na_if(x = Types.Colorectal.Race.Black.non.Hispanic, y = 0)) %>%
  mutate(Types.Colorectal.Race.Asian = na_if(x = Types.Colorectal.Race.Asian, y = 0)) %>%
  mutate(Types.Colorectal.Race.Indigenous = na_if(x = Types.Colorectal.Race.Indigenous, y = 0)) %>%
  mutate(Types.Colorectal.Race.Hispanic = na_if(x = Types.Colorectal.Race.Hispanic, y = 0))


```
```

ANOVA One-way Test on Ethnicty

```{r}
#select variable of interest into a new variable to melt
colorectal.2019.ethnicity <- cancer.cleaned %>%
  select(State,
         Types.Colorectal.Race.White,
         Types.Colorectal.Race.White.non.Hispanic,
         Types.Colorectal.Race.Black,
         Types.Colorectal.Race.Black.non.Hispanic,
         Types.Colorectal.Race.Asian,
         Types.Colorectal.Race.Indigenous,
         Types.Colorectal.Race.Hispanic
         ) %>%

#rename
  rename(White = Types.Colorectal.Race.White) %>%
  rename(White.non.Hispanic = Types.Colorectal.Race.White.non.Hispanic) %>%
  rename(Black = Types.Colorectal.Race.Black) %>%
  rename(Black.non.Hispanic = Types.Colorectal.Race.Black.non.Hispanic) %>%
  rename(Asian = Types.Colorectal.Race.Asian) %>%
  rename(Indigenous = Types.Colorectal.Race.Indigenous) %>%
  rename(Hispanic = Types.Colorectal.Race.Hispanic)
```

```{r}
colorectal.2019.ethnicity
```

A tibble: 51 × 8

State <chr>	White <dbl>	White.non.Hispanic <dbl>	Black <dbl>	Black.non.Hispanic <dbl>	Asian <dbl>	Indigenous <dbl>	Hispanic <dbl>
Alabama	15.9	16.0	24.4	24.5	*NA*	*NA*	5.7
Alaska	13.6	13.8	*NA*	*NA*	12.5	34.7	*NA*
Arizona	13.8	13.9	18.7	19.7	10.6	10.1	13.1
Arkansas	17.7	17.9	26.3	26.4	*NA*	*NA*	8.1
California	14.4	15.0	21.2	22.2	11.6	7.7	11.7
Colorado	13.4	13.1	17.2	17.9	10.7	7.6	14.8
Connecticut	12.8	12.9	15.5	16.4	6.6	*NA*	10.2
Delaware	15.4	15.4	16.9	17.0	*NA*	*NA*	*NA*
District of Columbia	10.0	9.8	23.0	23.1	*NA*	*NA*	9.1
Florida	14.3	14.5	18.5	19.2	9.2	5.6	13.4

1-10 of 51 rows

Previous 1 2 3 4 5 6 Next

```{r}
#melt different ethnicity columns into a new variable
colorectal.2019.ethnicity.melt <- melt(colorectal.2019.ethnicity, id = "State")


colorectal.2019.ethnicity.cleaned <- colorectal.2019.ethnicity.melt %>%
  #rename
  rename(Ethnicity = variable) %>%
  rename(Colorectal.Rate = value)

colorectal.2019.ethnicity.cleaned
```

Description: df [357 × 3]

State <chr>	Ethnicity <fctr>	Colorectal.Rate <dbl>
Alabama	White	15.9
Alaska	White	13.6
Arizona	White	13.8
Arkansas	White	17.7
California	White	14.4
Colorado	White	13.4
Connecticut	White	12.8
Delaware	White	15.4
District of Columbia	White	10.0
Florida	White	14.3

1-10 of 357 rows

Previous 1 2 3 4 5 6 ... 36 Next

Use graphics and descriptive statistics to examine the total mortality rate of colorectal cancer on its own.

```{r}
#calculate descriptive stats for Total Mortality Rates of Colorectal Cancer
colorectal.rate <- colorectal.2019.ethnicity.cleaned %>%
    drop_na(Colorectal.Rate) %>%
    summarize(mean.colorectal.rate = mean(x = Colorectal.Rate), sd.colorectal.rate = sd(x = Colorectal.Rate))
```

```{r}
colorectal.rate
```

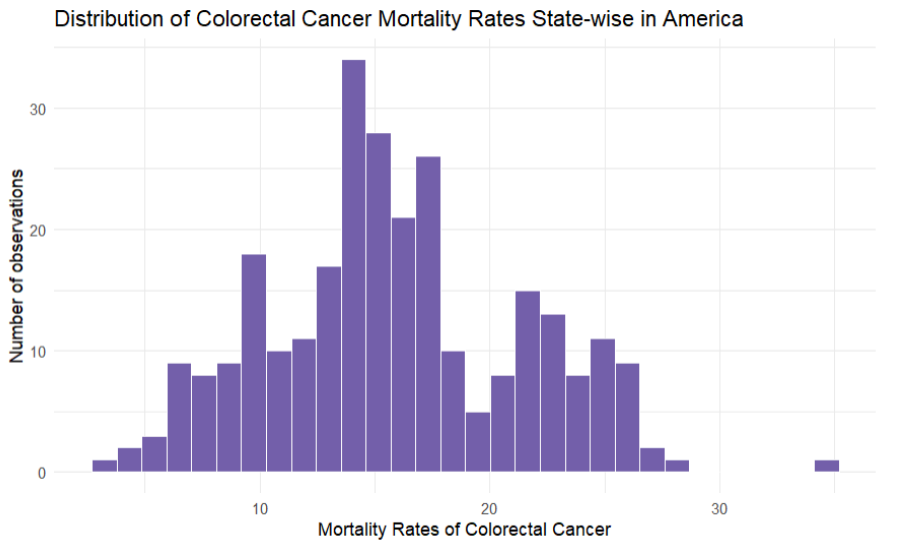
Description: df [1 × 2]

mean.colorectal.rate <dbl>	sd.colorectal.rate <dbl>
15.875	5.488778

1 row

```
colorectal.2019.ethnicity.cleaned %>%
  drop_na(Colorectal.Rate) %>%
  ggplot(aes(x = Colorectal.Rate)) +
  geom_histogram(fill = "#7463AC", color = "white") +
  theme_minimal() +
  labs(x = "Mortality Rates of Colorectal Cancer", y = "Number of observations") +
  ggtitle("Distribution of Colorectal Cancer Mortality Rates State-wise in America")
```

```
[38;5;232m`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.[39m
```



Interpretation:
* The graph depicts a distribution that closely resembles a normal distribution.

```{r}
#calculate descriptive stats for Colorectal.Rate after grouping by Ethnicity
colorectal.2019.ethnicity.cleaned %>%
  drop_na(Colorectal.Rate) %>%
  group_by(Ethnicity) %>%
  summarize(mean.usetech = mean(x = Colorectal.Rate), sd.usetech = sd(x = Colorectal.Rate))
```

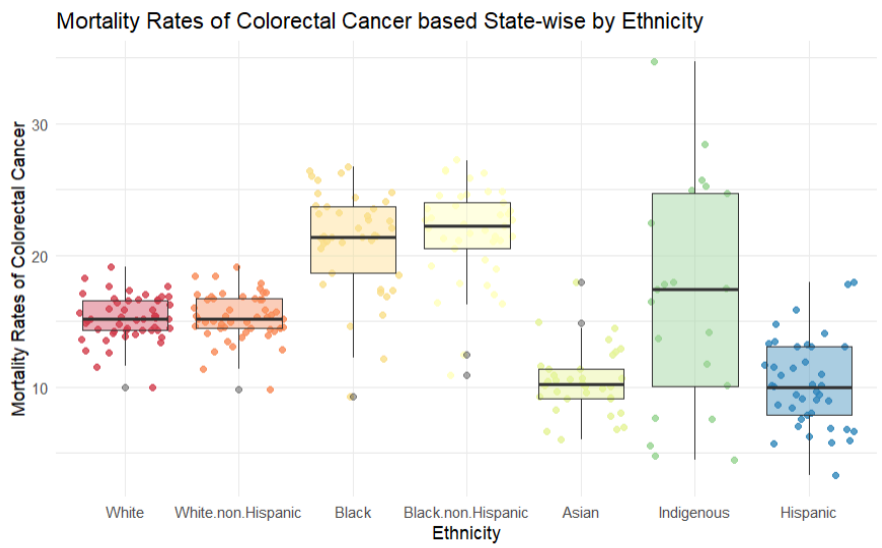
A tibble: 7 × 3

Ethnicity <fctr>	mean.usetech <dbl>	sd.usetech <dbl>
White	15.23333	1.686377
White.non.Hispanic	15.35098	1.767074
Black	21.16098	3.856869
Black.non.Hispanic	21.70000	3.580433
Asian	10.42647	2.509414
Indigenous	16.83333	8.504019
Hispanic	10.24146	3.329638

7 rows

Interpretation:
* Black and non-Hispanic black are at a higher risk of colorectal cancer mortalities in comparison to other ethnic groups.
* Standard deviations of all categorical variables are smaller than their respective means.

```
#graph Mortality Rates vs. Ethnicity
colorectal.2019.ethnicity.cleaned %>%
              drop_na(Colorectal.Rate) %>%
              ggplot(aes(y = Colorectal.Rate, x = Ethnicity)) +
              geom_jitter(aes(color = Ethnicity), alpha = .8) +
              geom_boxplot(aes(fill= Ethnicity), alpha = .4) +
              scale_fill_brewer(palette = "Spectral", guide = FALSE) +
              scale_color_brewer(palette = "Spectral", guide = FALSE) +
              theme_minimal() +
              labs(x = "Ethnicity", y = "Mortality Rates of Colorectal Cancer") +
              ggtitle("Mortality Rates of Colorectal Cancer based State-wise by Ethnicity")
```



Interpretation:
* Black and non-Hispanic black have a higher mean in comparison to other ethnic groups.
* Black and non-Hispanic black datapoints cluster indicates they are 23-25 percentage more likely to die from colorectal cancer.
* Indigenous groups have a wider range but less cluster, indicating disparity in cancer mortality rates

###The F-Test Statistic for ANOVA Across Ethnic Groups###

```{r}
#average colorectal cancer mortality rate by ethnicity
cancer.by.eth <- oneway.test(formula = Colorectal.Rate ~ Ethnicity,
                  data = colorectal.2019.ethnicity.cleaned,
                  var.equal = TRUE)

cancer.by.eth
```

```
        One-way analysis of means

data:  Colorectal.Rate and Ethnicity
F = 63.653, num df = 6, denom df = 273, p-value < 2.2e-16
```

Interpretation:
* F-Statistic = 63.653
* The probability of an F-statistic this large or larger if the null were true was reported in the output as < 2.2e-16, which is < .001.
* With a p-value this small, the F-statistic is considered to be statistically significant.

We will perform the NHST process in order to determine if the evidence and data are statistically significant enough to reject the null hypothesis.

    H0: The average mortality rate of colorectal cancer in America is equal across ethnic groups
    HA: The average mortality rate of colorectal cancer in America is NOT equal across ethnic groups

* The average mortality rate of colorectal cancer is significantly different across ethnicity [F(6, 273) = 63.653; p < .05], indicating the disparity in vulnerability to cancer mortality among the ethnic groups in America.

We will use a post hoc test, in this case the "bonf" pairwise.t.test(), to determine which mean/average are significantly different from each other.

```{r}
#differences in the average rate of colorectal cancer in America across ethnic groups using the pairwise.t.test()
bonf.tech.by.eth <- pairwise.t.test(x = colorectal.2019.ethnicity.cleaned$Colorectal.Rate,
                                    g = colorectal.2019.ethnicity.cleaned$Ethnicity,
                                    p.adj = "bonf")

bonf.tech.by.eth
```

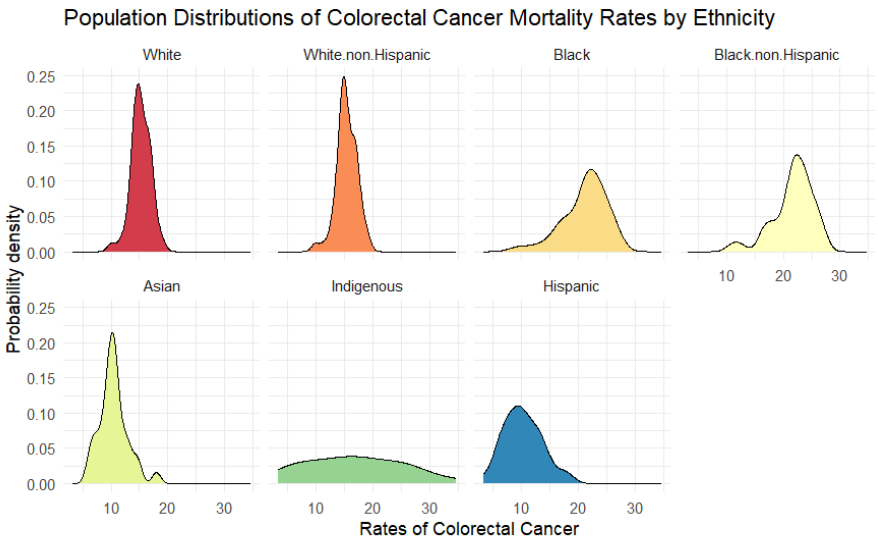
```
        Pairwise comparisons using t tests with pooled SD

data:  colorectal.2019.ethnicity.cleaned$Colorectal.Rate and colorectal.2019.ethnicity.cleaned$Ethnicity

                  White   White.non.Hispanic Black   Black.non.Hispanic Asian   Indigenous
White.non.Hispanic 1.00000 -                  -       -                  -       -
Black              1.6e-12 4.3e-12           -       -                  -       -
Black.non.Hispanic 1.3e-14 3.7e-14           1.00000 -                  -       -
Asian              9.4e-08 4.2e-08           < 2e-16 < 2e-16            -       -
Indigenous         1.00000 1.00000           0.00021 1.6e-05            1.1e-08 -
Hispanic           3.5e-09 1.4e-09           < 2e-16 < 2e-16            1.00000 9.8e-10

P value adjustment method: bonferroni
```

Interpretation:
* We observe that two of the t-tests fall below .05
* There is a significant difference in average rate of colorectal cancer between African American and non-Hispanic African American and the rest of the ethnic groups (p < .05).
* There is a significant difference in average rate of colorectal cancer between white and non-Hispanic white the rest of the ethnic groups, except for the Indigenous (p < .05).
* There is a significant difference in average rate of colorectal cancer between Asian and the Indigenous (p < .05)
* there are no significant differences among the remaining groups.

Check assumptions for the ANOVA and conduct an appropriate alternate analysis if it does not pass assumptions.

First, we will observe the assumptions of normality through plotting a density plot.

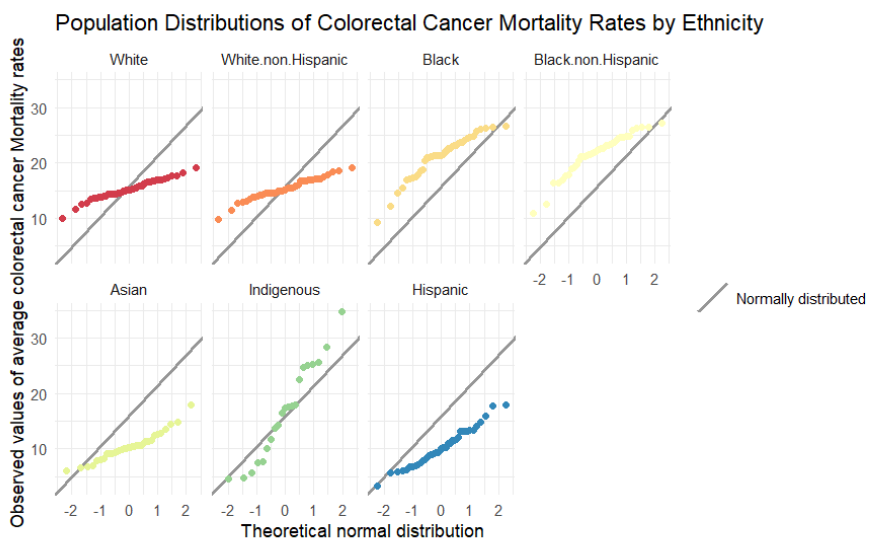
```{r}
colorectal.2019.ethnicity.cleaned %>%
  drop_na(Colorectal.Rate) %>%
  ggplot(aes(x = Colorectal.Rate)) +
  geom_density(aes(fill = Ethnicity)) +
  facet_wrap(facets = vars(Ethnicity), nrow = 2) +
  scale_fill_brewer(palette = "Spectral", guide = FALSE) +
  theme_minimal() +
  labs(x = "Rates of Colorectal Cancer",
   y = "Probability density") +
  ggtitle("Population Distributions of Colorectal Cancer Mortality Rates by Ethnicity")
```



Interpretation:
* The graphs for "Black", "Black non-Hispanic", "Asian", and "Hispanic" are skewed
* None of the graphs display a similarity to a normal distribution.

Next, we use Q-Q plots to confirm our assumptions.

```{r}
colorectal.2019.ethnicity.cleaned %>%
    drop_na(Colorectal.Rate) %>%
    ggplot(aes(sample = Colorectal.Rate)) +
    geom_abline(aes(intercept = mean(Colorectal.Rate), slope = sd(Colorectal.Rate),
    linetype = "Normally distributed"),
    color = "gray60", size = 1) +
stat_qq(aes(color = Ethnicity)) +
scale_color_brewer(palette = "Spectral", guide = FALSE) +
scale_linetype_manual(values = 1, name = "") +
labs(x = "Theoretical normal distribution",
y = "Observed values of average colorectal cancer Mortality rates") +
ggtitle("Population Distributions of Colorectal Cancer Mortality Rates by Ethnicity") +
theme_minimal() +
facet_wrap(facets = vars(Ethnicity), nrow = 2)
```



Interpretation:
* None of the groups display any similarity to a normal distribution in either of the graph types.

```{r}
#observing equal variances for Colorectal.Rate group by Ethnicity
car::leveneTest(y = Colorectal.Rate ~ Ethnicity, data = colorectal.2019.ethnicity.cleaned, center = mean)
```

```
Levene's Test for Homogeneity of Variance (center = mean)
       Df F value    Pr(>F)
group   6  19.215 < 2.2e-16 ***
      273
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interpretation:
* p-value = < 2.2e-16
* The p-value for Levene's test indicates we should reject the null hypothesis.
* The variances of Colorectal.Rate are statistically significantly different across ethnicity (p < .05).
* The ANOVA fails the assumption of homogeneity of variances.

Since we've failed the ANOVA test, we will use the Welch t-test and the Kruskal-Wallis test to compare the categorical variable in the ethnic group.

```{r}
welch.colorectal.eth <-oneway.test(formula = Colorectal.Rate ~ Ethnicity,
                                   data = colorectal.2019.ethnicity.cleaned,
                                   var.equal=FALSE)

welch.colorectal.eth
```

```
        One-way analysis of means (not assuming equal variances)

data:  Colorectal.Rate and Ethnicity
F = 72.307, num df = 6.00, denom df = 102.42, p-value < 2.2e-16
```

Interpretation:
* [Fw(6, 102.42) = 72.307]
* Since p < 0.5, we reject the null hypothesis

```{r}
#Kruskal Test for Colorectal Rate among Ethnicity
kw.cancer.by.eth <- kruskal.test(formula = Colorectal.Rate ~ Ethnicity,
                                 data = colorectal.2019.ethnicity.cleaned)
kw.cancer.by.eth
```

```
        Kruskal-Wallis rank sum test

data:  Colorectal.Rate by Ethnicity
Kruskal-Wallis chi-squared = 170.95, df = 6, p-value < 2.2e-16
```

Interpretation:
* The p-value is 2.2e-16
* The Kruskal-Wallis chi-squared is 170.95
* Since p < 0.5, we reject the null hypothesis.

```
dunn.cancer.by.eth <- dunn.test::dunn.test(x = colorectal.2019.ethnicity.cleaned$Colorectal.Rate, g =
colorectal.2019.ethnicity.cleaned$Ethnicity, method = "bonferroni")
```

```
  Kruskal-Wallis rank sum test

data: x and group
Kruskal-Wallis chi-squared = 170.9517, df = 6, p-value = 0


                           Comparison of x by group
                                 (Bonferroni)
Col Mean-|
Row Mean |      Asian      Black   Black.no   Hispanic   Indigeno      White
---------+------------------------------------------------------------------
   Black |  -8.836149
         |    0.0000*
         |
Black.no |  -9.190036  -0.371656
         |    0.0000*     1.0000
         |
Hispanic |  -0.054544   9.222529   9.594186
         |     1.0000    0.0000*    0.0000*
         |
Indigeno |  -4.297325   3.193159   3.499052  -4.397489
         |    0.0002*    0.0148*    0.0049*    0.0001*
         |
   White |  -4.498136   5.023253   5.414587  -4.687574   0.758996
         |    0.0001*    0.0000*    0.0000*    0.0000*     1.0000
         |
White.no |  -4.699937   4.810247   5.201581  -4.900580   0.586676  -0.225620
         |    0.0000*    0.0000*    0.0000*    0.0000*     1.0000     1.0000

alpha = 0.05
Reject Ho if p <= alpha/2
```

Interpretation:
* "Asian", "Black", "Black.no.Hispanic", and "Hispanic groups seem to have a statistically significant differences between the mean ranks.

```{r}
#compute mode using aov
colorectal.eth.aov <- aov(formula = Colorectal.Rate ~ Ethnicity,
                          data = colorectal.2019.ethnicity.cleaned)

summary(colorectal.eth.aov)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
Ethnicity     6   4902   816.9   63.65 <2e-16 ***
Residuals   273   3504    12.8
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
 77 個の観測値が欠損のため削除されました
```

```{r}
summ.colorectal.eth <- summary(colorectal.eth.aov)

#compute effect size
k.om <- summ.colorectal.eth[[1]][1, 1] + 1
n.om <- summ.colorectal.eth[[1]][2, 1] + summ.colorectal.eth[[1]][1, 1] + 1
omega.sq <- (summ.colorectal.eth[[1]][1, 2])/(summ.colorectal.eth[[1]][1, 2] + (n.om - k.om + 1)/(k.om - 1))
omega.sq
```

```
[1] 0.9907693
```

Interpretation:
The strength of the relationship is large (ω2 = .99).

The average colorectal cancer rates was significantly different among the ethnic groups [F(6, 273) = 63.65; p < .05].

T-Test on Male vs Female above 64

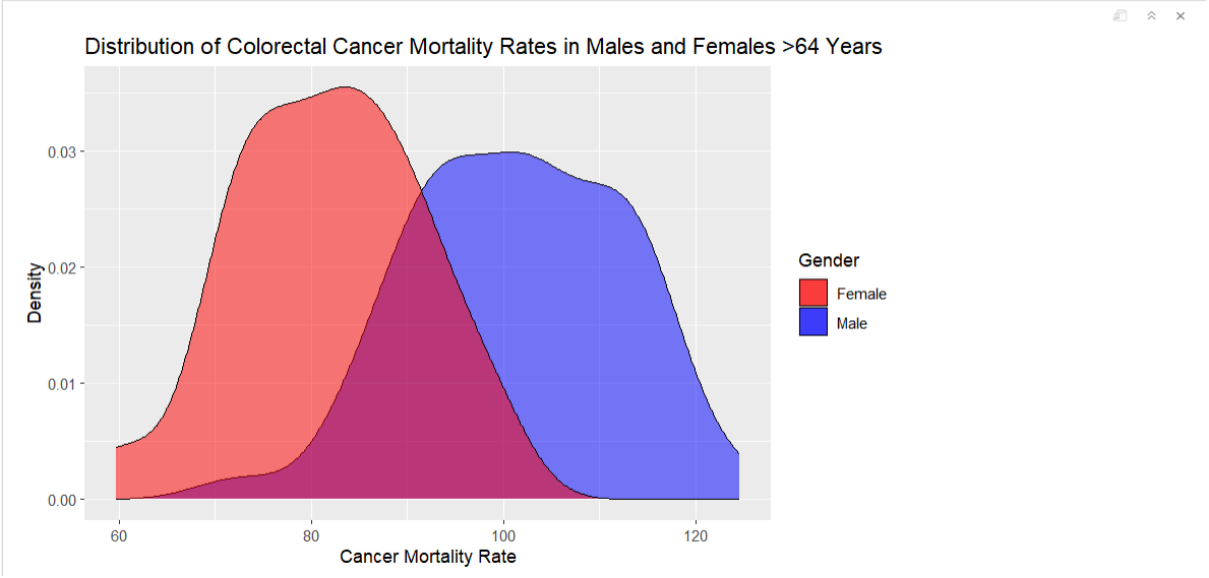
```{r}
# summarize the data
summary(object = cancer.cleaned$Types.Colorectal.Age.Male.Above.64)
summary(object = cancer.cleaned$Types.Colorectal.Age.Female.Above.64)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  72.40   93.35  102.10  101.45  110.70  124.60 
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  59.70   74.55   82.40   81.80   88.70   99.80 
```

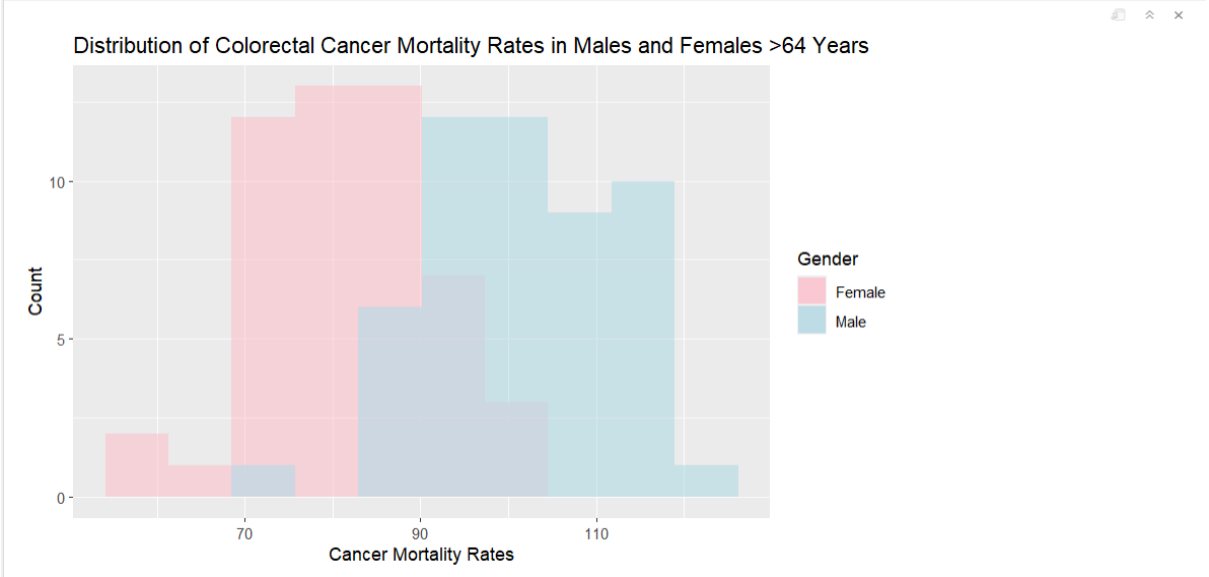
```{r}
# conduct independent sample t-test
t.test(cancer.cleaned$Types.Colorectal.Age.Female.Above.64,
                cancer.cleaned$Types.Colorectal.Age.Male.Above.64,
                var.equal = FALSE)
```

```
        Welch Two Sample t-test

data:  cancer.cleaned$Types.Colorectal.Age.Female.Above.64 and cancer.cleaned$Types.Colorectal.Age.Male.Above.64
t = -9.7026, df = 98.266, p-value = 5.219e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -23.67004 -15.63192
sample estimates:
mean of x mean of y 
   81.800   101.451 
```

```
# Create density plot
ggplot(cancer.cleaned, aes(x = Types.Colorectal.Age.Male.Above.64)) +
  geom_density(aes(fill = "Male"), alpha = 0.5) +
  geom_density(aes(x = Types.Colorectal.Age.Female.Above.64, fill = "Female"), alpha = 0.5) +
  ggtitle("Distribution of Colorectal Cancer Mortality Rates in Males and Females >64 Years") +
  xlab("Cancer Mortality Rate") +
  ylab("Density") +
  scale_fill_manual(values = c("red","blue"), name = "Gender")
```




```
# Plot histograms using ggplot
ggplot(cancer.cleaned, aes(x=Types.Colorectal.Age.Female.Above.64, fill = "Female")) +
  geom_histogram(alpha = 0.5, bins = 10) +
  geom_histogram(aes(x=Types.Colorectal.Age.Male.Above.64, fill = "Male"), alpha = 0.5, bins = 10) +
  labs(title = "Distribution of Colorectal Cancer Mortality Rates in Males and Females >64 Years", x = "Cancer
Mortality Rates", y = "Count", fill = "Gender") +
  scale_fill_manual(values = c("Female" = "pink", "Male" = "lightblue"))
```



Interpretation:
The t-value shows how the means of the two groups differ in relation to the degree of data variability. The difference between the means of the two groups is bigger when the absolute t-value is larger. The t-value in this instance is -9.7. A warning symbol denotes that women are more likely than males to develop cancer on average.

The amount of evidence contradicting the null hypothesis is indicated by the p-value. Given that the null hypothesis is true, it shows the likelihood of observing a t-value that is equally extreme or more extreme than the one that was actually observed. Strong evidence is presented against the null hypothesis when the p-value is less than the significance level, which is often set at 0.05. The observed difference in cancer mortality rate between males and females over the age of 64 is statistically significant in this instance since the p-value is zero. Since there is a significant difference in the mean cancer mortality rate between males and females over the age of 64, the null hypothesis can be rejected.

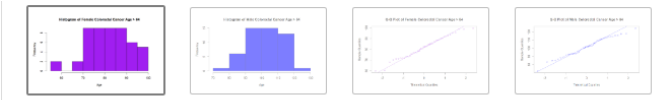
```{r}
# Check for normality using Shapiro-Wilk test
shapiro.test(cancer.cleaned$Types.Colorectal.Age.Male.Above.64)
shapiro.test(cancer.cleaned$Types.Colorectal.Age.Female.Above.64)
```

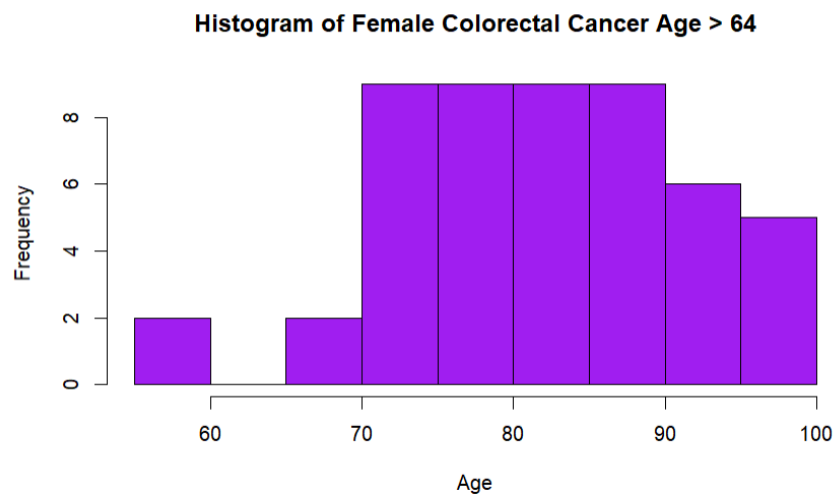
```
        Shapiro-Wilk normality test

data:  cancer.cleaned$Types.Colorectal.Age.Male.Above.64
W = 0.9825, p-value = 0.6491


        Shapiro-Wilk normality test

data:  cancer.cleaned$Types.Colorectal.Age.Female.Above.64
W = 0.98276, p-value = 0.6609
```

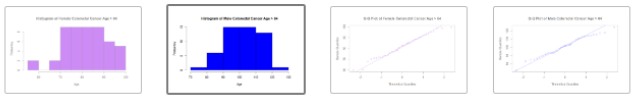
```
# Perform F-test
var.test(cancer.cleaned$Types.Colorectal.Age.Female.Above.64,
                  cancer.cleaned$Types.Colorectal.Age.Male.Above.64,
                  alternative = "two.sided")
```

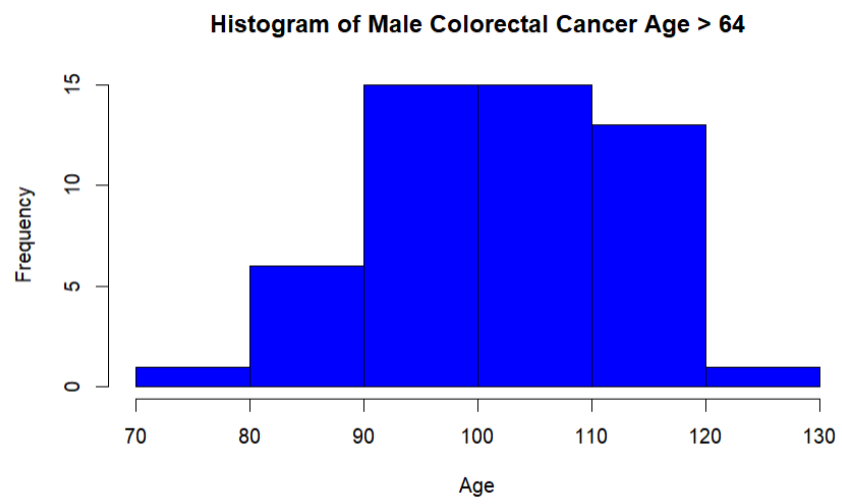
```
        F test to compare two variances

data:  cancer.cleaned$Types.Colorectal.Age.Female.Above.64 and cancer.cleaned$Types.Colorectal.Age.Male.Above.64
F = 0.76549, num df = 50, denom df = 50, p-value = 0.3479
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.4369328 1.3410956
sample estimates:
ratio of variances
         0.7654859
```

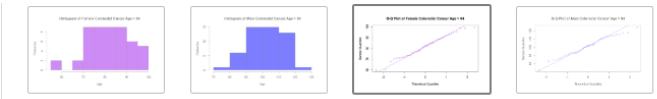
```{r}
# Create a colorful histogram for female data
hist(cancer.cleaned$Types.Colorectal.Age.Female.Above.64,
     col = "purple",
     main = "Histogram of Female Colorectal Cancer Age > 64",
     xlab = "Age")

# Create a colorful histogram for male data
hist(cancer.cleaned$Types.Colorectal.Age.Male.Above.64,
     col = "blue",
     main = "Histogram of Male Colorectal Cancer Age > 64",
     xlab = "Age")

# Create a colorful Q-Q plot for female data
qqnorm(cancer.cleaned$Types.Colorectal.Age.Female.Above.64,
       col = "purple",
       main = "Q-Q Plot of Female Colorectal Cancer Age > 64")
qqline(cancer.cleaned$Types.Colorectal.Age.Female.Above.64, col = "purple")

# Create a colorful Q-Q plot for male data
qqnorm(cancer.cleaned$Types.Colorectal.Age.Male.Above.64,
       col = "blue",
       main = "Q-Q Plot of Male Colorectal Cancer Age > 64")
qqline(cancer.cleaned$Types.Colorectal.Age.Male.Above.64, col = "blue")
```

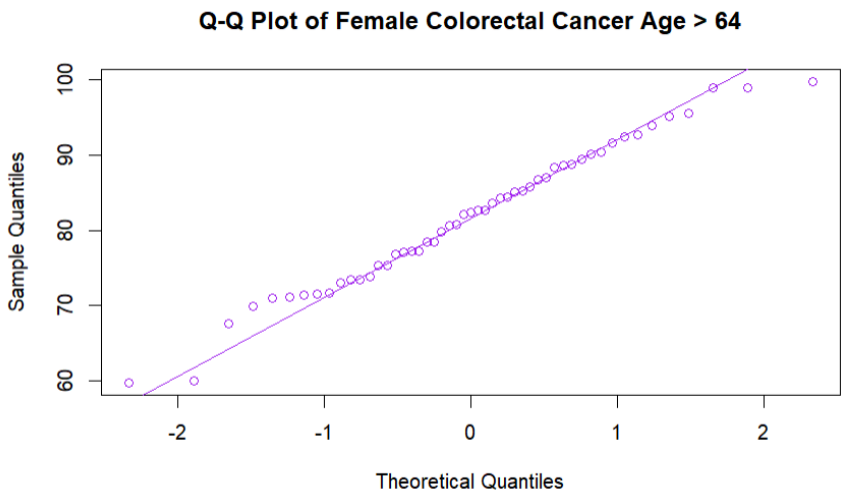


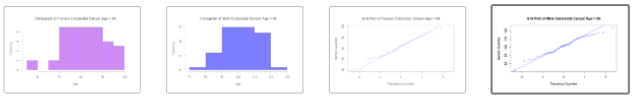

Histogram of Female Colorectal Cancer Age > 64
Frequency
0
2
4
6
8
60
70
80
90
100
Age

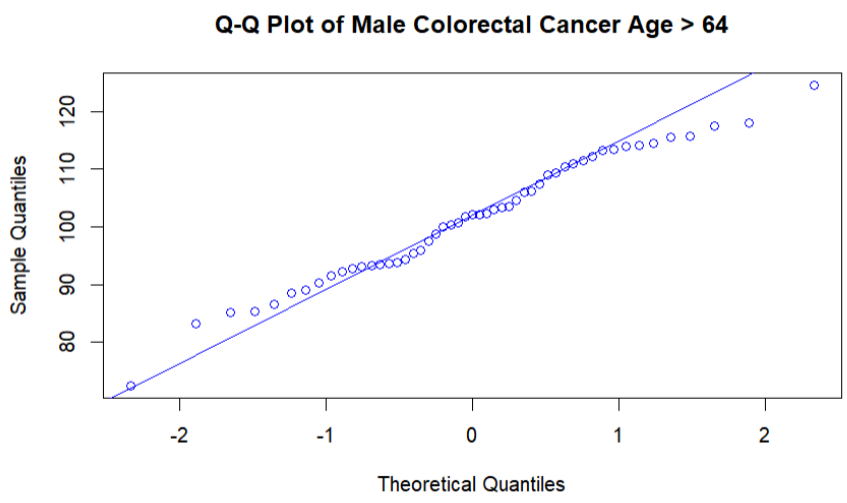



Histogram of Male Colorectal Cancer Age > 64
Frequency
0
5
10
15
70
80
90
100
110
120
130
Age



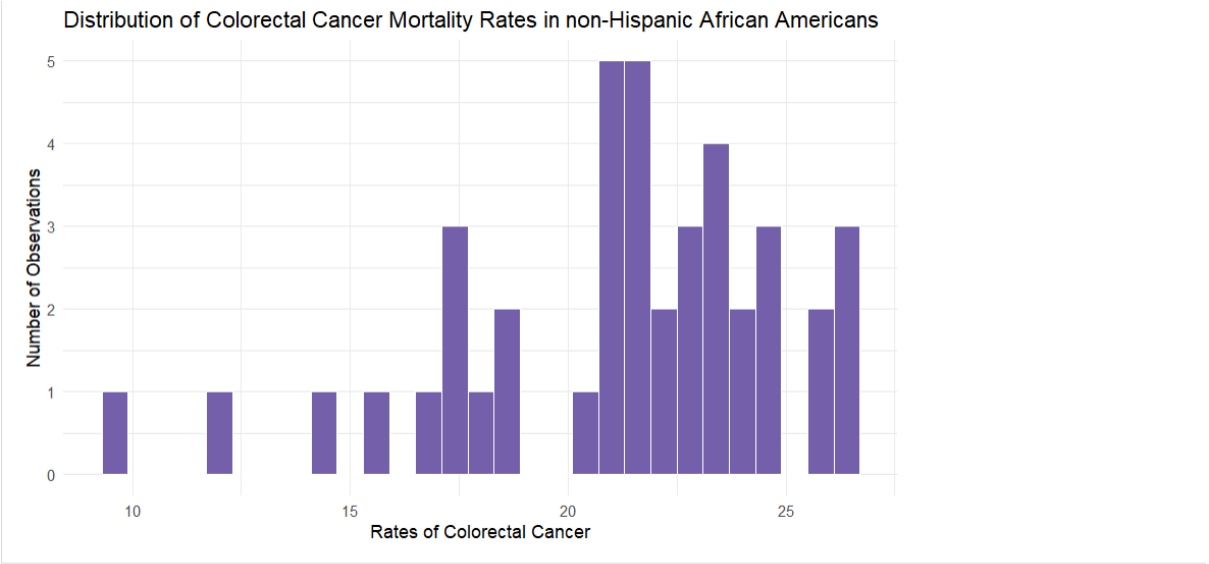






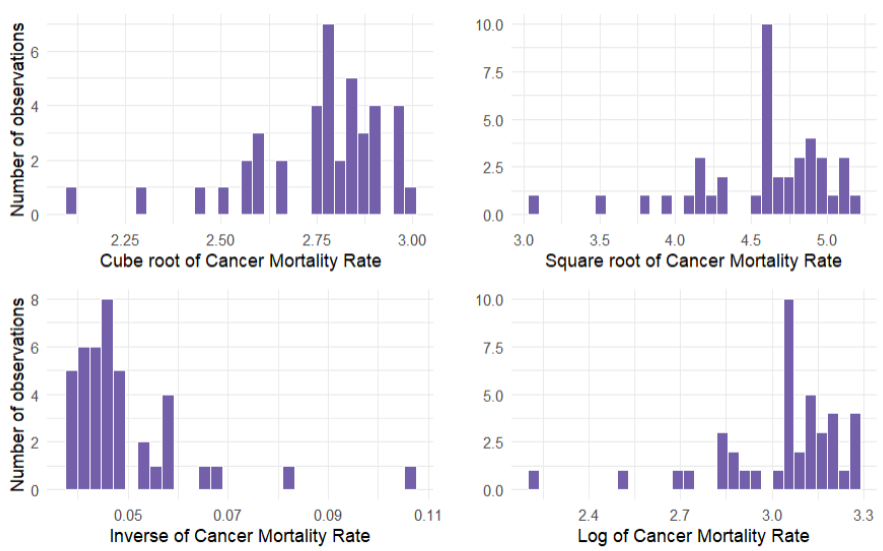
Linear Regression of Colorectal Cancer Mortality Rate Between non-Hispanic African American and Male Above 64

```{r}
# check distribution of black patients with colorectal cancer
cancer.cleaned %>%
    ggplot(aes(x = Types.Colorectal.Race.Black)) +
    geom_histogram(fill = "#7463AC", color = "white") +
    labs(x = "Rates of Colorectal Cancer", y = "Number of Observations") +
    ggtitle("Distribution of Colorectal Cancer Mortality Rates in non-Hispanic African Americans") +
theme_minimal()
```



Interpretation:
* Distribution is left-skewed.

```{r}
# histograms of square root of Types.Colorectal.Race.Black
# cube root
cube.root.cancer <- cancer.cleaned %>%
ggplot(aes(x = (Types.Colorectal.Race.Black)^(1/3))) +
geom_histogram(fill = "#7463AC", col = "white") +
labs(x = "Cube root of Cancer Mortality Rate", y = "Number of observations") +
theme_minimal()

# square root
sq.root.cancer <- cancer.cleaned %>%
ggplot(aes(x = sqrt(x = Types.Colorectal.Race.Black))) +
geom_histogram(fill = "#7463AC", col = "white") +
labs(x = "Square root of Cancer Mortality Rate", y = "")+
theme_minimal()

# inverse
inverse.cancer <- cancer.cleaned %>%
ggplot(aes(x = 1/Types.Colorectal.Race.Black)) +
geom_histogram(fill = "#7463AC", col = "white") +
labs(x = "Inverse of Cancer Mortality Rate", y = "Number of observations")+
theme_minimal()

# log
log.cancer <- cancer.cleaned %>%
ggplot(aes(x = log(x = Types.Colorectal.Race.Black))) +
geom_histogram(fill = "#7463AC", col = "white") +
labs(x = "Log of Cancer Mortality Rate", y = "")+
theme_minimal()

# view options for transformation
gridExtra::grid.arrange(cube.root.cancer, sq.root.cancer,
             inverse.cancer, log.cancer)
```

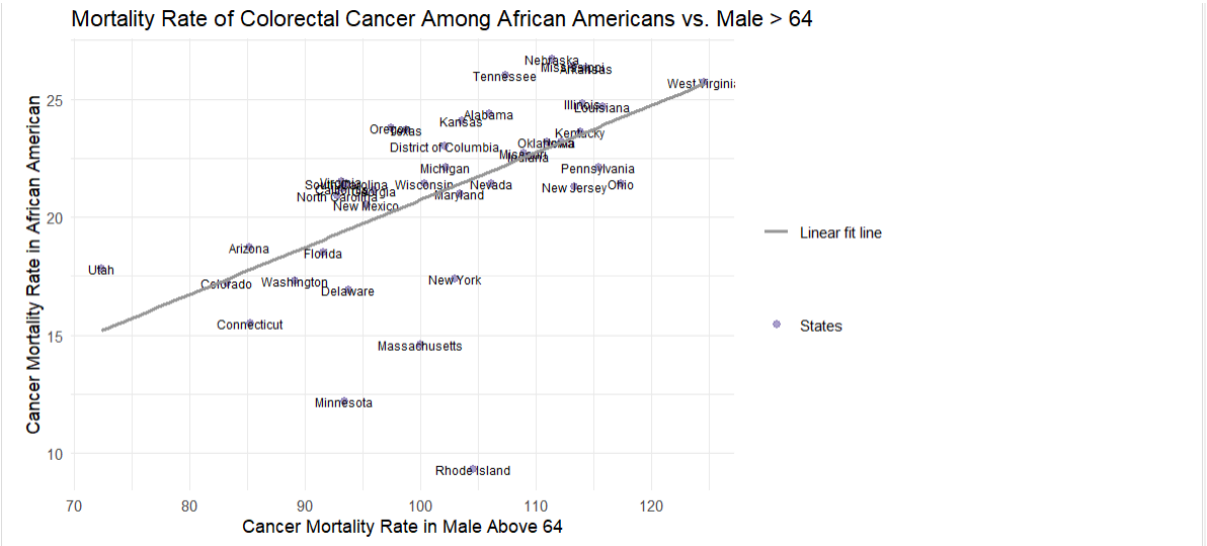



```
Interpretation:
* Other alternatives are also skewed. Taking median would be better.
```

```
# descriptive statistics for black patients
table.eth <- tableone::CreateTableOne(data = cancer.cleaned,
                      vars =
c("Types.Colorectal.Race.Black","Types.Colorectal.Age.Male.Above.64","Types.Colorectal.Age.Female.Above.64"))

print(x = table.eth, nonnormal = c("Types.Colorectal.Race.Black"))
```

```
                                              Overall
  n                                              51
  Types.Colorectal.Race.Black (median [IQR])   21.40 [18.70, 23.70]
  Types.Colorectal.Age.Male.Above.64 (mean (SD))  101.45 (10.89)
  Types.Colorectal.Age.Female.Above.64 (mean (SD))  81.80 (9.52)
```

```
#colorectal cancer mortality rate in black vs male > 64
cancer.cleaned %>%
  ggplot(aes(x = Types.Colorectal.Age.Male.Above.64, y = (Types.Colorectal.Race.Black), label=State)) +
  geom_point(aes(size = "States"), color = "#7463AC", alpha = .6) +
  geom_text(size = 2.5) +
  geom_smooth(aes(linetype = "Linear fit line"), method = "lm",
      se = FALSE, color = "gray60") +
theme_minimal() +
labs(x = "Cancer Mortality Rate in Male Above 64", y = "Cancer Mortality Rate in African American") +
ggtitle("Mortality Rate of Colorectal Cancer Among African Americans vs. Male > 64") +
scale_size_manual(values = 2, name = "") +
scale_linetype_manual(values = 1, name = "")
```




```{r}
# correlation coefficient between African American vs Male above 64
cancer.cleaned %>%
  drop_na(Types.Colorectal.Race.Black) %>%
  summarize(eth.age.gender = cor(x = Types.Colorectal.Race.Black,
                                 y = Types.Colorectal.Age.Male.Above.64),
                                 samp.n = n())
```

A tibble: 1 × 2

eth.age.gender <dbl>	samp.n <int>
0.5750488	41

1 row

Interpretation:
* The correlation coefficient was positive (r = 0.575).
* This correlation is moderately strong.

```
# linear regression cancer Mortality rates in blacks vs cancer Mortality rates in male above 64
# note the function arguments
# na.action = deal with missing values
#na.action = na.exclude -> options for excluding obervations
# with missing values
black.male.above.64 <- lm(formula = (Types.Colorectal.Race.Black) ~ Types.Colorectal.Age.Male.Above.64,
                  data = cancer.cleaned, na.action = na.exclude)

summary(black.male.above.64)
```

```
Call:
lm(formula = (Types.Colorectal.Race.Black) ~ Types.Colorectal.Age.Male.Above.64,
    data = cancer.cleaned, na.action = na.exclude)

Residuals:
     Min       1Q   Median       3Q      Max
-12.3478  -0.5685   0.5939   1.9567   3.7910

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                      0.68474    4.69138   0.146    0.885
Types.Colorectal.Age.Male.Above.64  0.20041    0.04566   4.390  8.4e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.196 on 39 degrees of freedom
  ( 10 個の観測値が欠損のため削除されました )
Multiple R-squared:  0.3307,    Adjusted R-squared:  0.3135
F-statistic: 19.27 on 1 and 39 DF,  p-value: 8.402e-05
```

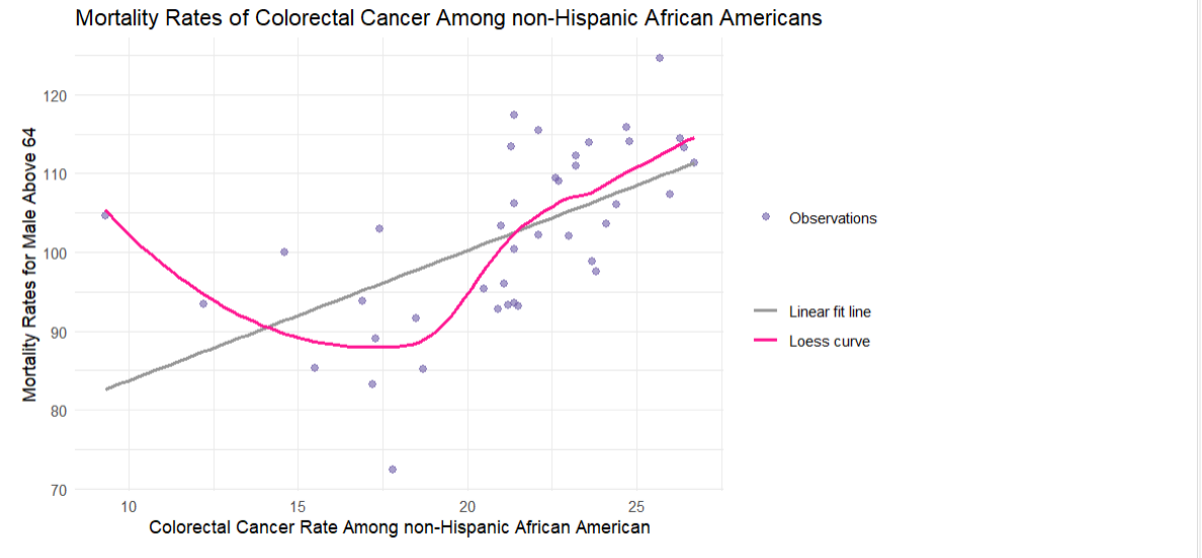
* p-value: 8.402e-05

```
# confidence interval for regression parameters
ci.black.male.above.64 <- confint(object = black.male.above.64)
#object created in line 472
ci.black.male.above.64
```

```
                                       2.5 %     97.5 %
(Intercept)                        -8.8044831 10.1739579
Types.Colorectal.Age.Male.Above.64  0.1080629  0.2927611
```

Interpretation:
The colorectal cancer mortality rates of African American in a state is a statistically significant predictor for male above 64 (b = 0.20041 ; p < .05). For every 1% increase in cancer rates of African Americans in a state, the predicted cancer mortality rate for male above 64 increases by 0.20041 percentage. The value of the slope in the sample is 0.20041, and the value of the slope is likely between 0.11 and 0.29 in the population that the sample came from (95% CI: 0.11-0.29). With every 1% increase in colorectal cancer mortality rate of African American, the cancer mortality rates for male above 64 increases 0.11 and 0.29 more. These results suggest that state with an older population and with a higher African American population can experience a significant impact on their overall colorectal cancer mortality rate, and more resources should be diverted to by the federal government into the funding of healthcare for those states.

```
#checking for linear assumptions for colorectal cancer mortality rates in African American.
cancer.cleaned %>%
ggplot(aes(x = Types.Colorectal.Race.Black, y =Types.Colorectal.Age.Male.Above.64 )) +
geom_point(aes(size = "Observations"), color = "#7463AC", alpha = .6) +
geom_smooth(aes(color = "Linear fit line"), method = "lm", se = FALSE) +
geom_smooth(aes(color = "Loess curve"), se = FALSE) +
theme_minimal() +
labs(y = "Mortality Rates for Male Above 64", x = "Colorectal Cancer Rate Among non-Hispanic African American") +
ggtitle("Mortality Rates of Colorectal Cancer Among non-Hispanic African Americans") +
scale_color_manual(values = c("gray60", "deeppink"), name = "") +
scale_size_manual(values = 2, name = "")
```



Interpretation:
* The Loess curve does not fit well at any given point.
* Doesn't seem this would be good enough to meet linearity and we suggest that this assumption failed.

```
# testing for equal variance
const.var.test <- lmtest::bptest(formula = black.male.above.64)
const.var.test
```

```
        studentized Breusch-Pagan test

data:  black.male.above.64
BP = 0.0039768, df = 1, p-value = 0.9497
```

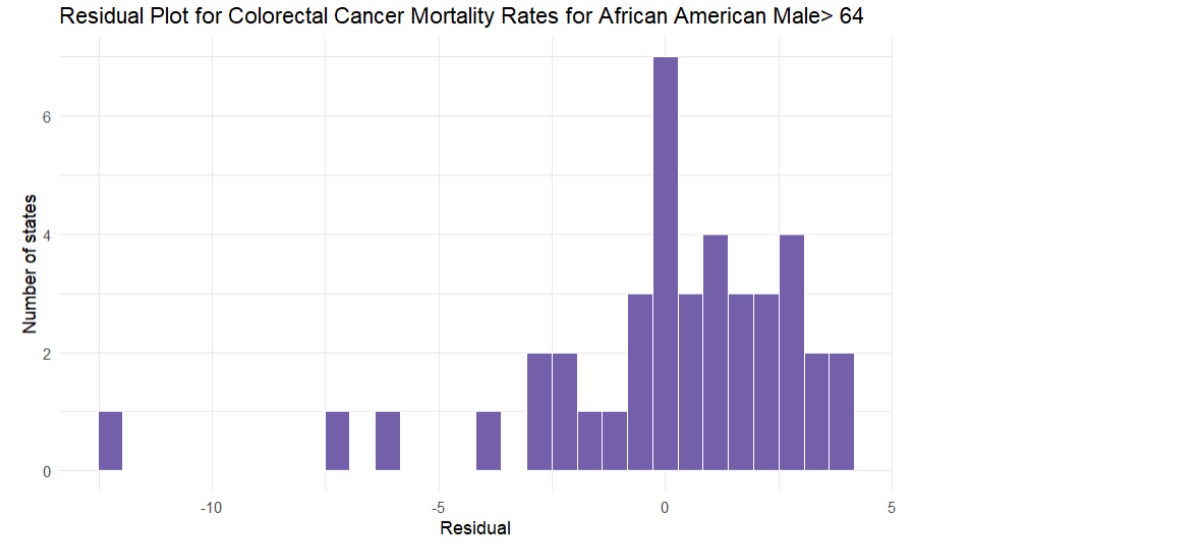
Interpretation:
* The Breusch-Pagan test statistic has a big p-value (BP = 0.0039768; p > .05), indicating that the null hypothesis of constant variance would be accepted.
* Even though the observations look cluster more right-sided, the test incidicates the variance spreads out enough.

```{r}
# test independence of residuals
lmtest::dwtest(formula = black.male.above.64)
```

```
        Durbin-Watson test

data:  black.male.above.64
DW = 2.052, p-value = 0.5584
alternative hypothesis: true autocorrelation is greater than 0
```

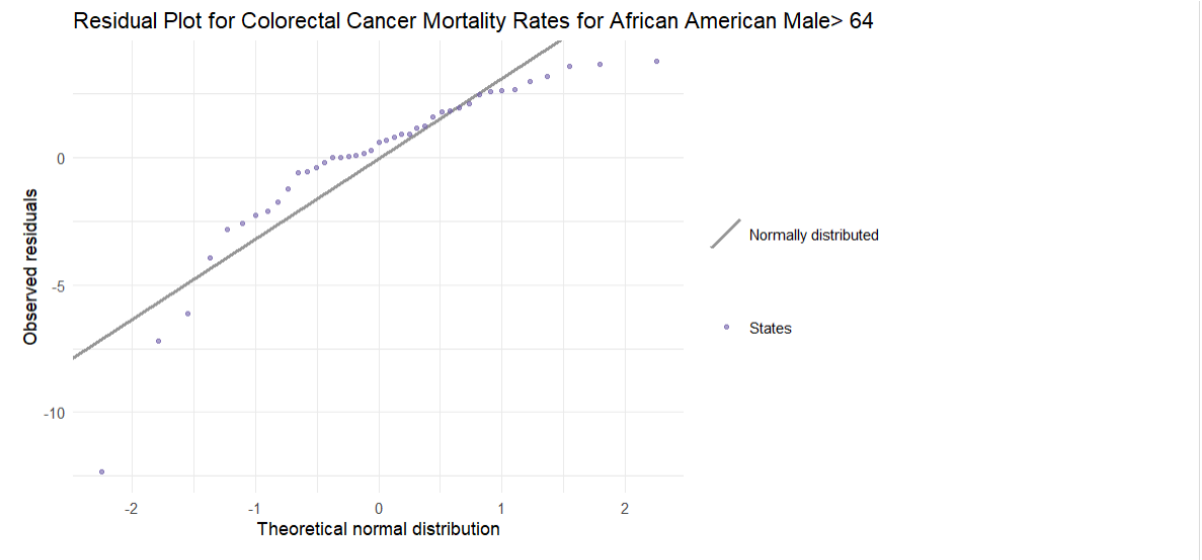
Interpretation:
* The D-W statistic is near 2 and the p-value was high, so we conclude that the null hypothesis is accepted.
* Since the null hypothesis was that the residuals were independent, we found that this assumption was met.

```{r}
#Check residual plot
data.frame(black.male.above.64$residuals) %>%
    ggplot(aes(x = black.male.above.64.residuals)) +
    geom_histogram(fill = "#7463AC", col = "white") +
  theme_minimal() +
labs(x = "Residual",
y = "Number of states") +
ggtitle("Residual Plot for Colorectal Cancer Mortality Rates for African American Male> 64")
```



Interpretation:
* The histogram suggests the residuals are left-skewed.

```r
# check residual plot
data.frame(black.male.above.64$residuals) %>%
ggplot(aes(sample = black.male.above.64.residuals)) +
geom_abline(aes(intercept = mean(x = black.male.above.64.residuals),
      slope = sd(x = black.male.above.64.residuals),
      linetype = "Normally distributed"),
      color = "gray60", size = 1) +
stat_qq(aes(size = "States"), color = "#7463AC", alpha = .6) +
theme_minimal() +
labs(x = "Theoretical normal distribution",
     y = "Observed residuals") +
ggtitle("Residual Plot for Colorectal Cancer Mortality Rates for African American Male> 64") +
scale_size_manual(values = 1, name = "") +
scale_linetype_manual(values = 1, name = "")
```



Interpretation:
* The Q-Q plot suggests the residuals are different from the values you'd expect from a normal distribution.
* Both graphs suggest some non-normality in the distribution of residuals.

Final verdict:
* The linear regression analysis met some assumptions and failed some assumptions.
* Because it does not meet all the assumptions, we know that the model is considered biased and should be interpreted with caution.
* Specifically, the results of a biased model are not usually applicable to the general population.