STAT350 Project: Questions Language: Rstudio

Group name: Infinity, 8 Group members: Youngjun Suh, Fiona Crenshaw, Jorge Mejia

Situation	Inference 1	Inference 2	Inference 3
Likes of the videos	One-sample t procedure: interest in if true population mean likes in America is more than Youtube's most liked videos in 2017.	ANOVA: do different regions in America have different numbers of likes in North, South, Caribbean, Central America.	ANOVA: Is the mean number of likes in different quality videos from 144p to 480p different from each other?
Death		Two-sample t: procedure(independent): 7day average death in North Dakota and South Dakota in 2021.	

Specification:

"YouTube Top 500 Most Liked Music Video of All Times." *Youtube Top 500 Most Liked Music Video of All Times*, https://www.popsonner.com/p/youtube-top-500-most-liked-music-video.html.

Null value:_ 14,830,718

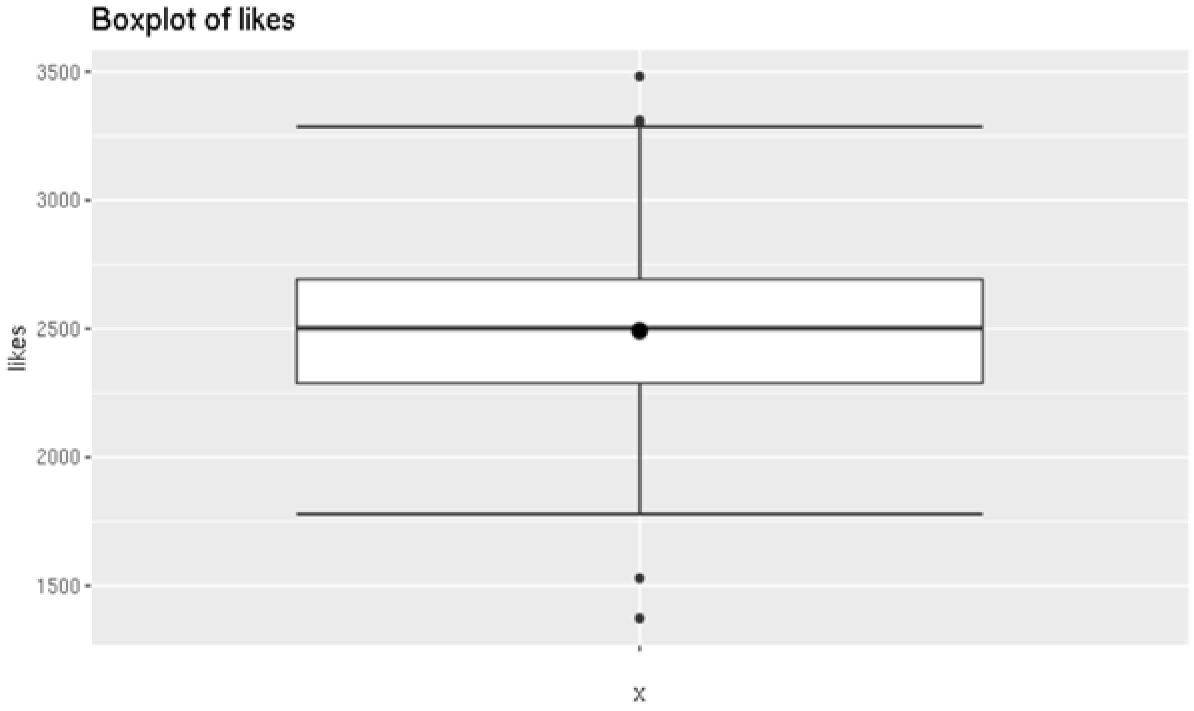
Figure was found by taking the top 20 most liked videos published in 2017 from the data set, adding them, and then dividing that figure by 20, giving us the average of the most liked videos of 2017.

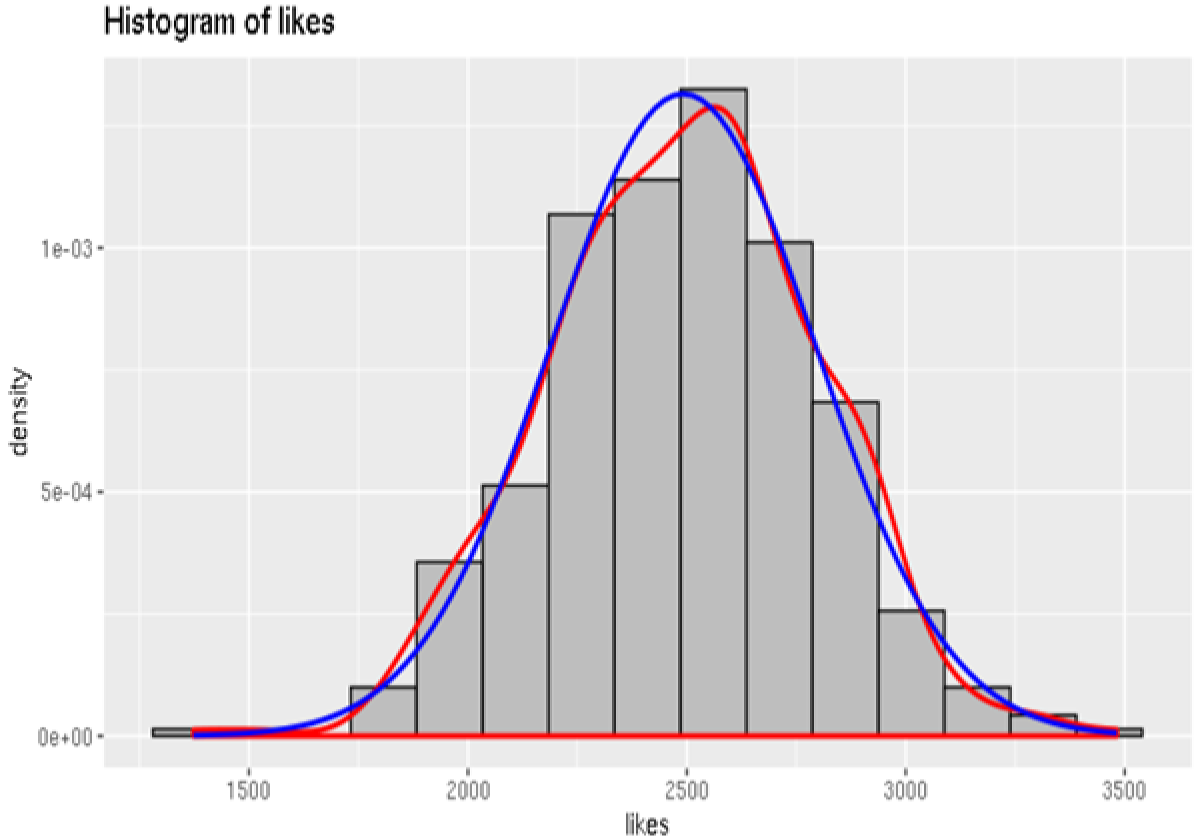
Q1: One-sample t procedure: Is the true population mean likes in America more than Youtube's most liked videos in 2017?

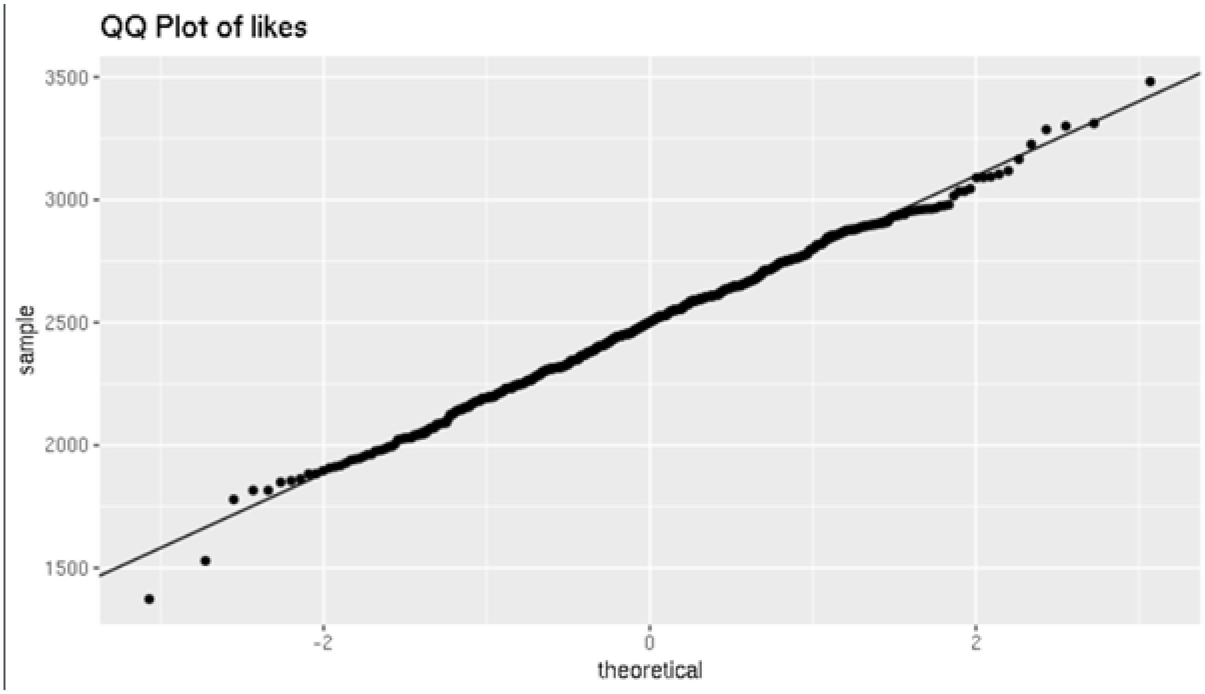
a) Code:

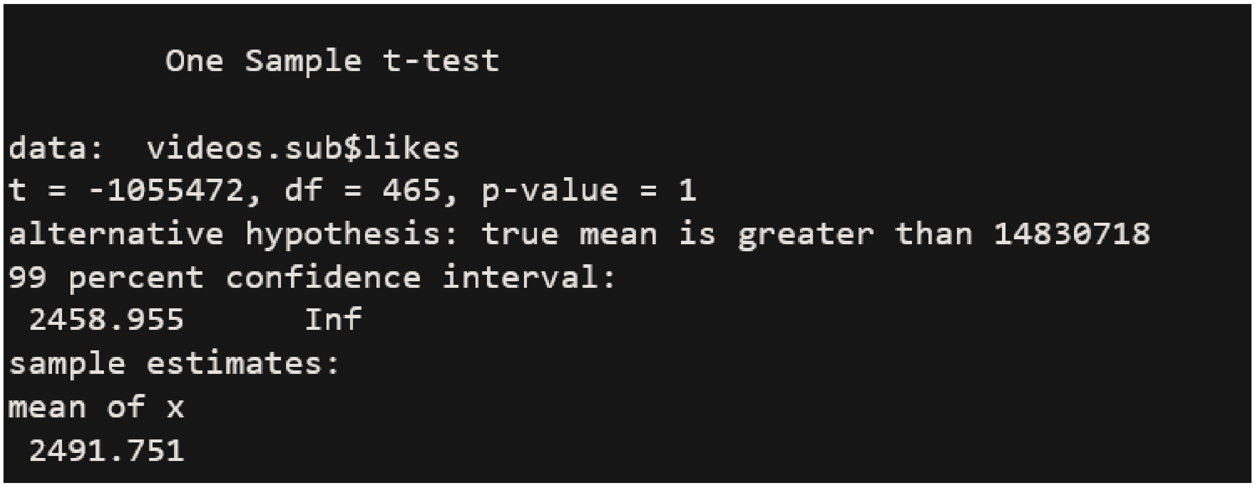
```
videos.sub<-videos.sub[complete.cases(videos.sub),]
videos.sub <- subset(videos, (shares > 1000) & (region == "North America" | region == "South America" | region == "Caribbean" | region == "Central America"), select = c("region", "likes", "views", "shares"))
library(ggplot2)
ggplot(videos.sub, aes(x = "", y = likes)) +
  stat_boxplot(geom = "errorbar") +
  geom_boxplot() +
  ggtitle("Boxplot of likes") +
  stat_summary(fun.y = mean, col = "black", geom = "point", size =3)
# Part B and show airbnb area
# numeric summary
xbar<-mean(videos.sub$likes)
s<-sd(videos.sub$likes)
quantile(videos.sub$likes)

ggplot(videos.sub, aes(likes))+
  geom_histogram(aes(y=..density..),bins=sqrt(nrow(videos.sub))-6, fill="grey", col="black")+
  geom_density(col="red", lwd=1)+
  stat_function(fun=dnorm, args=list(mean=xbar, sd=s),col="blue",lwd=1)+
  ggtitle("Histogram of likes")
```

```
# QQPlot
ggplot(videos.sub, aes(sample = likes)) +
  stat_qq() +
  geom_abline(slope = s, intercept = xbar) +
  ggtitle("QQ Plot of likes")
# d) critical value
t<-qt(0.01, 465, lower.tail=FALSE)
#Parts d) and e), the same code should be used
# Parameters for t.test():
# - mu: mu_0 of the null hypothesis
# - conf.level: confidence level of the confidence interval,
# which also indicates the significance level of the hypothesis
# test, alpha, by "alpha = 1 – confidence level"
# - alternative: form of the alternative hypothesis and confidence
# interval/bounds, possible options including
# - "two.sided" (not equal to, confidence interval)
# - "less" (<, upper confidence bound)
# - "greater" (>, lower confidence bound)
t.test(videos.sub$likes, conf.level = 0.99, mu = 14830718, alternative = "greater")
```


Boxplot of likes
3500
3000
2500
2000
1500
likes
x


Histogram of likes
1e-03
5e-04
0e+00
density
1500
2000
2500
3000
3500
likes





```
        One Sample t-test

data:  videos.sub$likes
t = -1055472, df = 465, p-value = 1
alternative hypothesis: true mean is greater than 14830718
99 percent confidence interval:
 2458.955      Inf
sample estimates:
mean of x
 2491.751
```

b) (5 pts.) What statistical procedure should be used and why? If this is for a one-sample, two sample paired, or two-sample independent, include what the null value is and why you chose that value. Besides the technique itself, be sure to state whether you are performing an inference procedure for a one-sided or two-sided hypothesis (except for ANOVA) with an explanation for the choice. Remember, this needs to be determined BEFORE you analyze the data**.**

We are going to use a one sample procedure because we want to find if the true mean population for the most liked videos in America in 2020 is more than the most liked videos of

2017. The null value we chose is 14,830,718 million and we chose this value because it is the mean of the sample population that is the top 20 most liked videos from 2017. This is going to be one-sided because we are analyzing the data to find a difference in one specific direction.

c) (10 pts.) Determine if the appropriate assumptions are satisfied. You may assume that the data set is from an SRS. Even though you are assuming this assumption is true, it must be explicitly stated. Please provide all of the diagnostic graphs to show that the assumptions are met and explain your decision. If the assumptions are not valid for your methodology and you still perform the analysis, you will lose 30 points. If a transformation is needed, state that you have performed a transformation and explain why you felt that a transformation is necessary. The explanation for the transformation should include at a minimum the histogram of the original data. You may include additional graphs of the untransformed variables, if necessary, for your explanation. If you transform your data, you will need to provide all of the diagnostic graphs for the transformed variables.

1. We are assuming the data is from an SRS data set.
2. Normality Assumption: The statistics measured form a normal distribution and we can see this by observing the histogram plots generated which are approximately normal. In addition the boxplot that was generated is not manipulated by the outliers, resulting in normality.

d) (7 pts.) Graphically display the data as appropriate for your answer in step b) with an interpretation of the output. The point of this part is to understand and explore your data, not merely to check the assumptions needed for inference as you did in step c). Therefore, you will need to state what you think that the conclusion for the inference is BASED ON THE GRAPHS in addition to a description of each of the graphs. Some of the graphs in step d) may have already been used in step c); however, the description of the graphs will be different. For one- and two-sample inferences and ANOVA, you will need to generate the appropriate histograms and boxplots. For what is meant by 'appropriate,' please see the computer assignments and tutorials. For ANOVA and linear regression where additional graphs are required, please see the computer assignments and tutorials to determine what graphs are needed. Remember that your discussion of the graphs should not include whether they are normal or not. Normality is an assumption and is discussed in part c).

The boxplot for the number of likes is roughly symmetrical and has outliers on both ends of the boxplot.

The histogram is approximately normal, unimodal, and symmetric.

The QQ plot is approximately the same to the provided line, which means it is good for inference.

e) (20 pts.) Perform the appropriate inference with a significance level of 0.01. This may consist of more than one step depending on the methodology in step b). The possible methodologies are 1) Confidence interval AND hypothesis test (Chapters 9, 10, and 11): This includes one sample, two-sample independent and two-sample paired t procedures. Each of these procedures is regarded as a different type of inference. 2) ANOVA (Chapter 12): Both the hypothesis test and the multiple comparison (if appropriate) need to be included. The multiple comparison test is not needed if the hypothesis test does not show a significant result. 3) Linear regression (Chapter 13): At least one inference needs to be included besides the equation of the line. Please see Computer Assignments 9 for possible inferences. All confidence intervals should include the interpretation. All hypothesis tests should consist of the four steps. Be careful about the interpretations when you transform the variables.

1. μ_0 = the true population mean number of likes in America

μ_A = the true population mean number of most liked videos in America in 2017

2. H0: μ_0 =< μ_a

Ha: $\mu_0 > \mu_a$

3. Fts = 0.78

Df = -1055472

p-value = 1

4. Since 1 > 0.01, we will fail to reject the null hypothesis.

The data does not provide evidence (p-value = 1.0) to support the claim that the population mean number of likes for 2020 is greater than the 2017 mean for most liked videos.

f) (8 pts.) The last part is the conclusion of your inference. This part is your final answer to each inference. This is split into two parts. a) If you obtained a significant result, determine the practicality of the answer. This is required even if you transformed your data, or you are performing ANOVA, although these explanations are more difficult. And b) A conclusion in words that relates to the context of the question. This part is a short paragraph explaining your conclusions of the Part and should be understandable to someone who has not taken a course in statistics.

a.) We did not obtain a significant result, so practically the population mean of likes was not greater than the mean number of most liked videos in 2017.

b.) In this analysis, the question asks if the mean likes of 2020 were greater than the mean most liked videos of 2017. We concluded that the assumptions to perform an T.test were satisfied. The results of the test indicate that the mean likes of 2020 were not greater than the mean most liked videos of 2017.

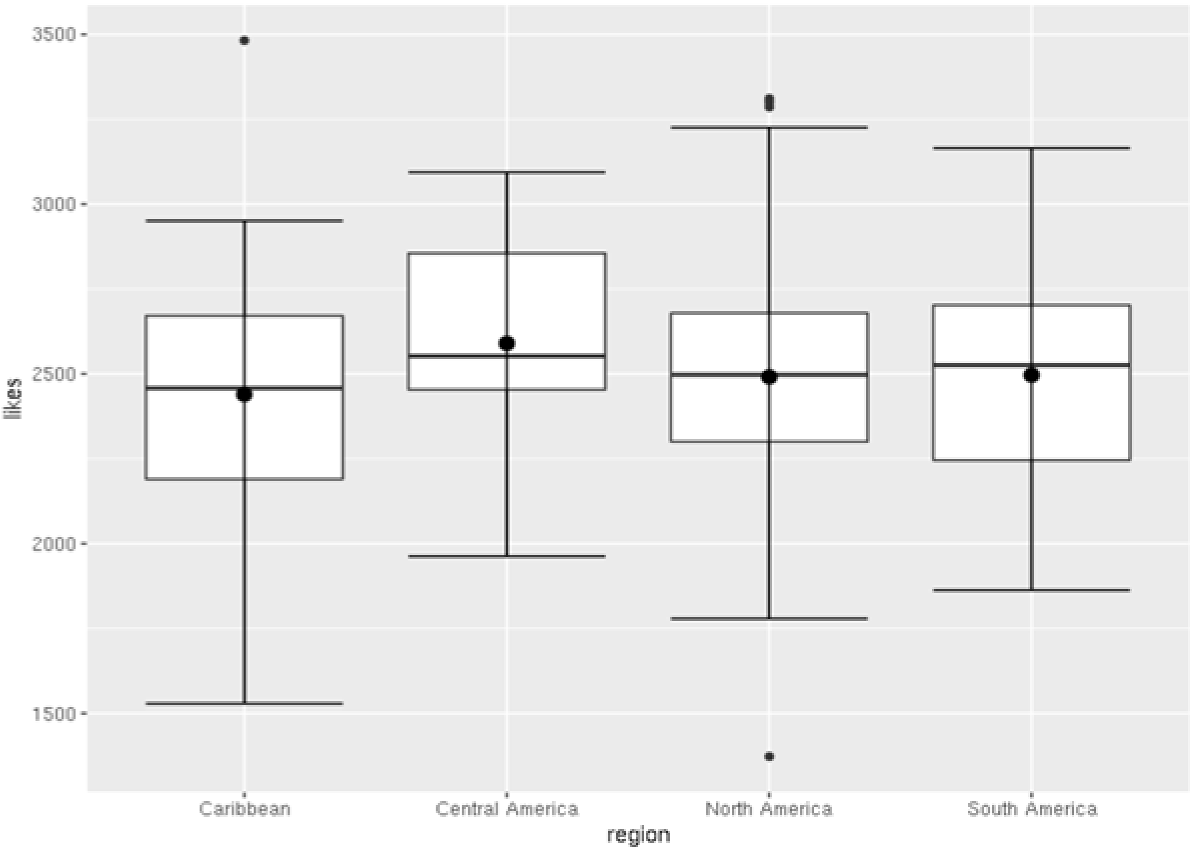
Q2: (ANOVA) do different regions in America have different numbers of likes in North, South, Caribbean, Central America.

Code:

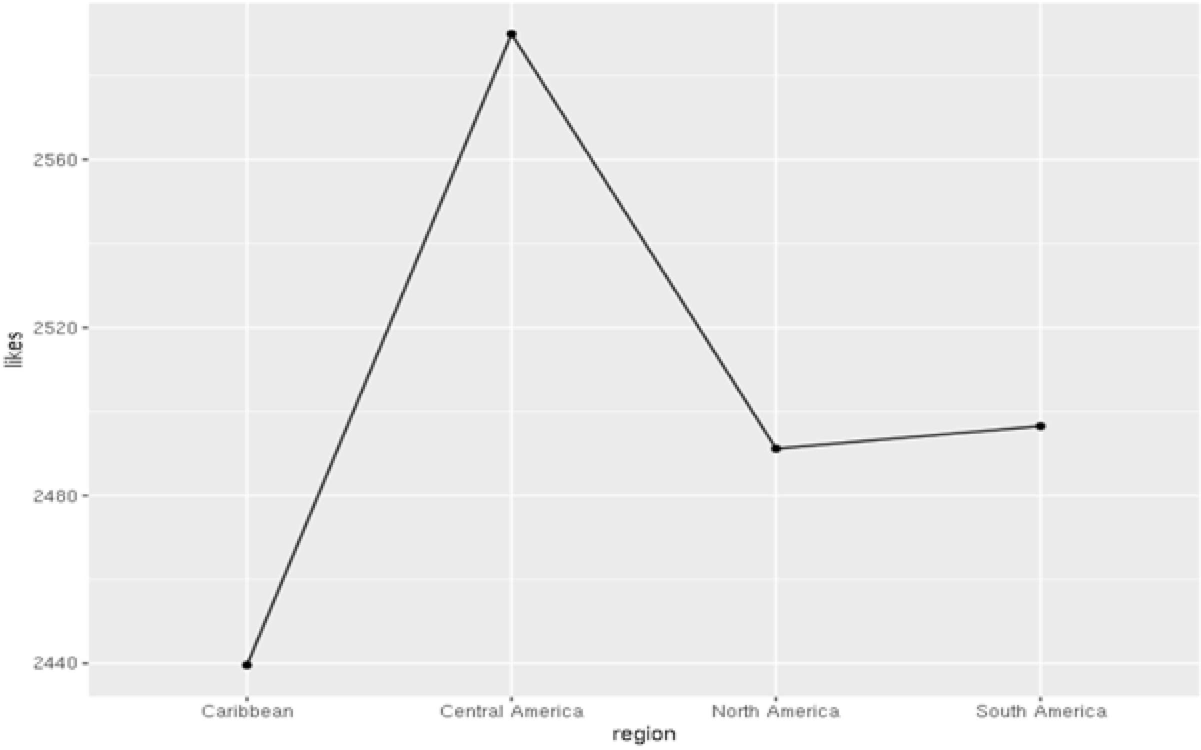
```
videos.sub<-videos.sub[complete.cases(videos.sub),]
videos.sub <- subset(videos, (shares > 1000) & (region == "North America" | region == "South America" | region == "Caribbean" | region == "Central America"), select = c("region", "likes", "quality", "views", "shares", "tags", "recommendations", "size"))library(ggplot2)
# a) Side-by-side boxplot
# See Computer Assignment 7 for the code. Remember to use
# x=GroupingVariable and y=NumericResponseVariable
# in this data set, the GroupingVariable is GroupName and
# the ResponseVariable is C
#
# BOXPLOT
# Specify an x variable (categorical) to graph multiple boxplots in
# the same command.
#
ggplot(videos.sub, aes(x = region, y = likes)) +
  geom_boxplot() +
  stat_boxplot(geom = "errorbar") +
  stat_summary(fun.y = mean, col = "black", geom = "point", size = 3) +
  ggtitle("Boxplots of number of likes by region")
#
# a) Effects plot
# We use two calls to stat_summary(): One plots the points ub the
# sample means of each category, the other connects the points with
# lines
ggplot(data = videos.sub, aes(x = region, y = likes)) +
  stat_summary(fun.y = mean, geom = "point") +
  stat_summary(fun.y = mean, geom = "line", aes(group = 1)) +
  ggtitle("Effects Plot of number of reviews by area")
#
# Calculating the sample size, sample mean and standard deviation
l <- tapply(videos.sub$likes, videos.sub$region, length)
m <- tapply(videos.sub$likes, videos.sub$region, mean)
s <- tapply(videos.sub$likes, videos.sub$region, sd)
#
# b) Histogram
# This is a little more complicated because we have three groups now .
# The code consists of two steps.
#
# (1) Make theoretical density curve
# You need to specify the name of the numeric response in the square
# brackets of as.numeric(x[]),
# and the name of the grouping variable in xbar[x[]] and s[x[]].
#
xbar <- tapply(videos.sub$likes, videos.sub$region, mean)
s <- tapply(videos.sub$likes, videos.sub$region, sd)

videos.sub$normal.density <- apply(videos.sub, 1, function(x){
  dnorm(as.numeric(x["likes"]),
        xbar[x["region"]], s[x["region"]])})
#
# (2) Make the histogram
# Remember that the number of bins in the histogram should be the
# maximum of the number of levels. See Tutorial 7c for details.
#
binlen <- as.numeric(max(tapply(videos.sub$likes,
                                videos.sub$region,length)))
ggplot(videos.sub, aes(x = likes)) +
  geom_histogram(aes(y = ..density..), bins = sqrt(binlen) + 2,
                        fill = "grey", col = "black") +
  facet_grid(region ~ .) +
  geom_density(col = "red", lwd = 1) +
  geom_line(aes(y = normal.density), col = "blue", lwd = 1) +
  ggtitle("Histograms of number of reviews by area")
#
# QQ plot
#
# (1) Calculate slope and intercept
# You need to put the name of the grouping variable in
# xbar[x[]] and s[x[]].
#
videos.sub$intercept <- apply(videos.sub, 1, function(x){xbar[x["region"]]})
videos.sub$slope <- apply(videos.sub, 1, function(x){s[x["region"]]})

#
# (2) Make the QQ plot
#
ggplot(videos.sub, aes(sample = likes)) +
  stat_qq() +
  facet_grid(region ~ .) +
  geom_abline(data = videos.sub, aes(intercept = intercept, slope = slope)) +
  ggtitle("QQ Plots of number of reviews by area")
```

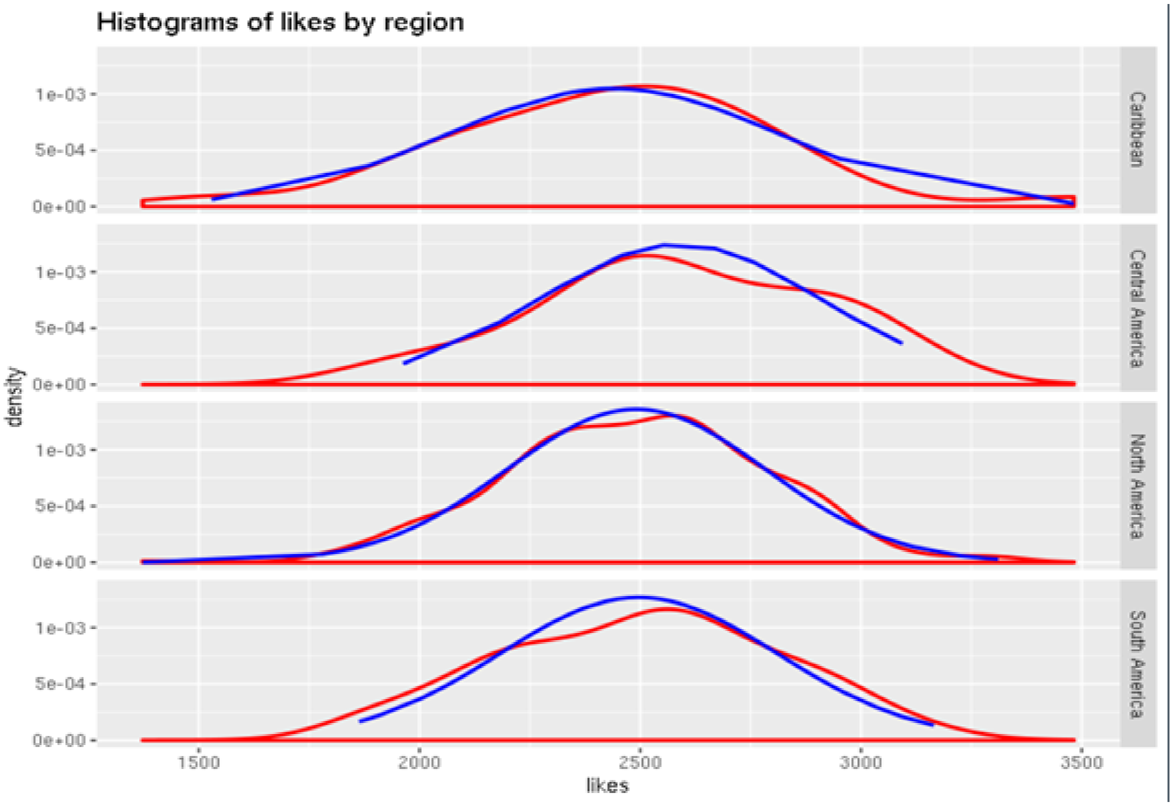
```r
#
# ANOVA
# c) hypothesis test
# The command is aov(numeric ~ categorical, data = datasetName)
# The function is called aov(): a as in analysis, o as in of, and
# v as in variance.
# The categorical variable must be of the class "factor" or "integer."
# To print out the results, you need to use the function summary().
# Note: this does not print out the "total" line in the ANOVA table.
# you may calculate it by hand or via R if required.
fit <- aov(likes ~ region, data = videos.sub)
summary(fit)
# d) Tukey method
TukeyHSD(fit, conf.level = 0.99)
# critical value:
# parameter: qtukey(ConfidenceLevel, #groups or k, dfe)
# though possible to have R determine what these values are, it is
# easier to put them in by hand after you see the results from aov().
# The critical value is the parameter/sqrt(2)
qtukey(0.99,4,466)/sqrt(2)
```

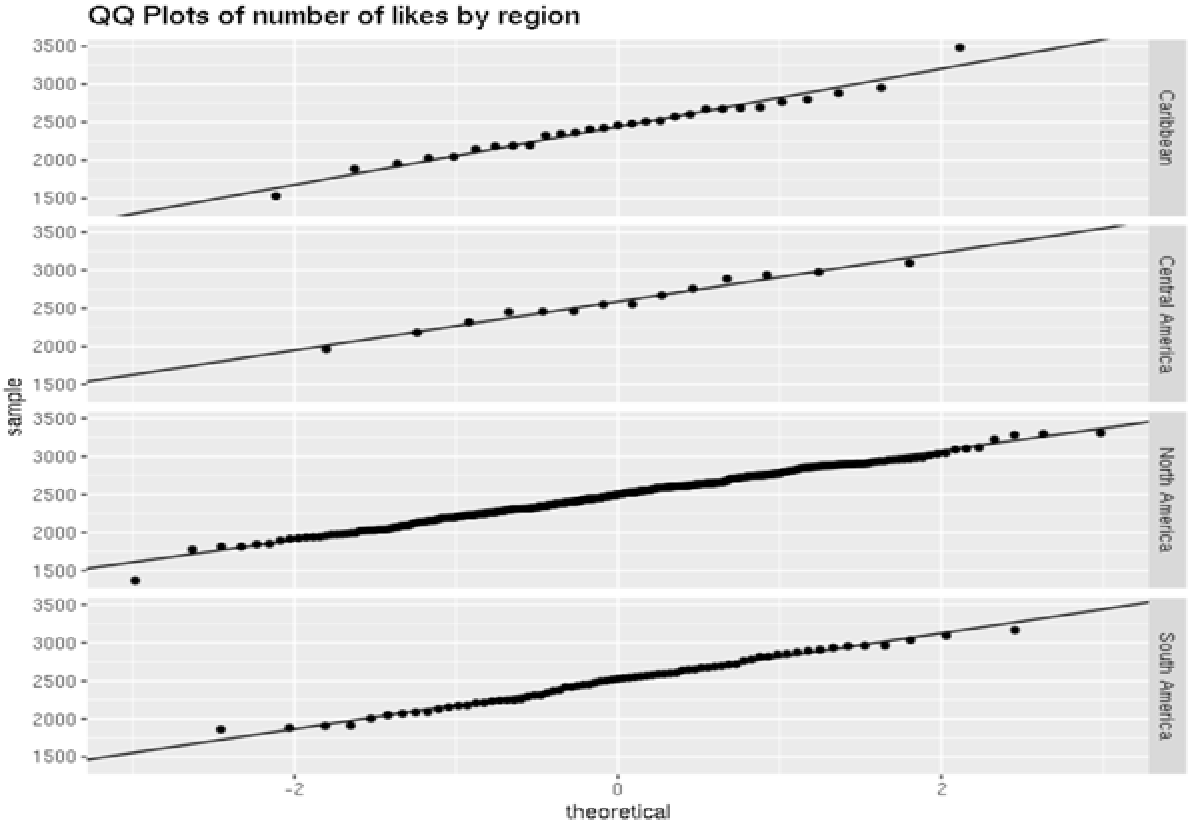
Boxplots of number of likes by region

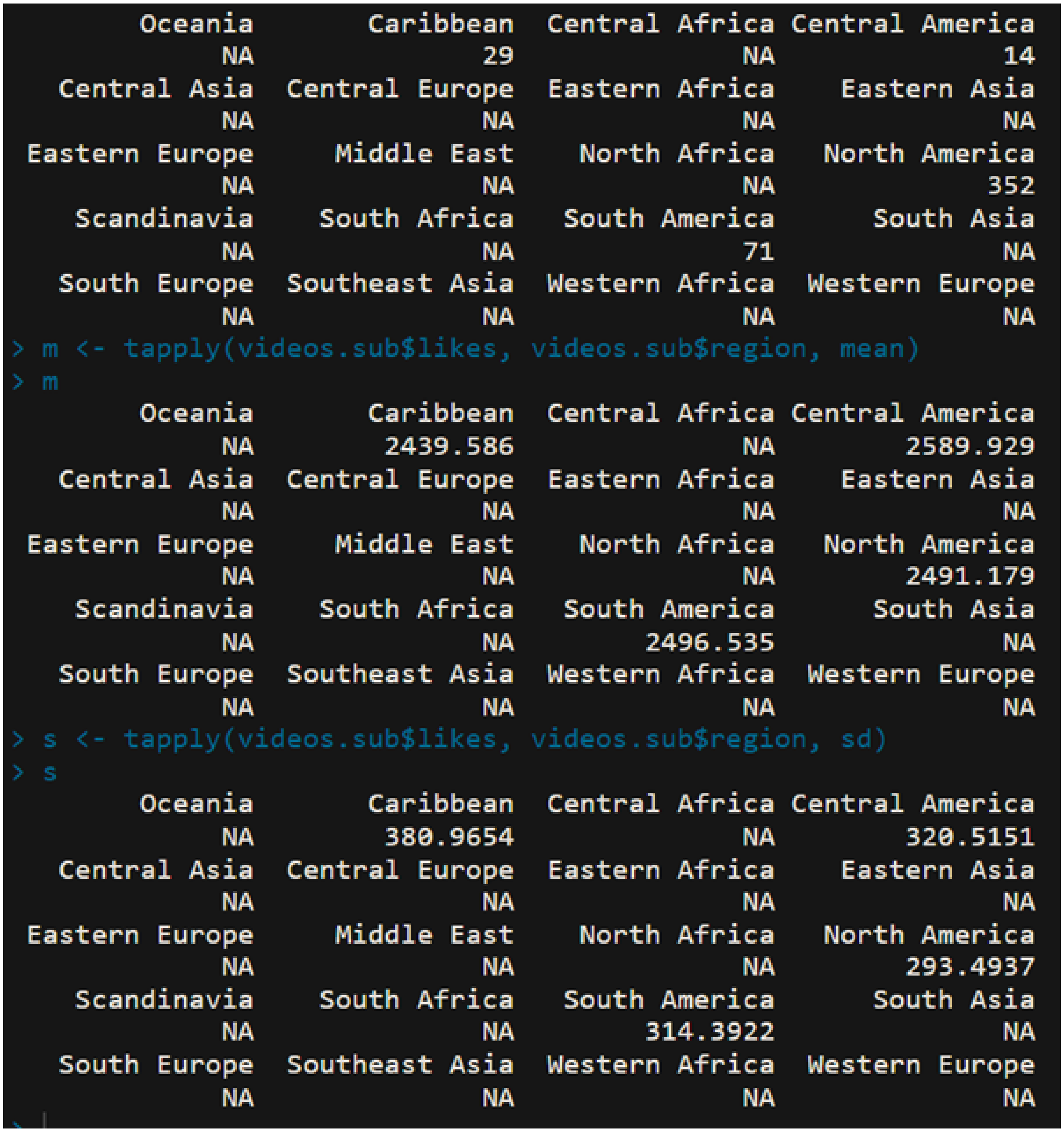


Effects Plot of likes by region

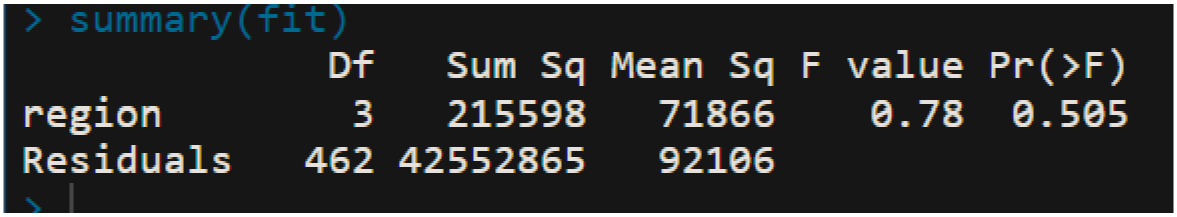



Histograms of likes by region
density
1e-03
5e-04
0e+00
Caribbean
Central America
North America
South America
1500
2000
2500
3000
3500
likes

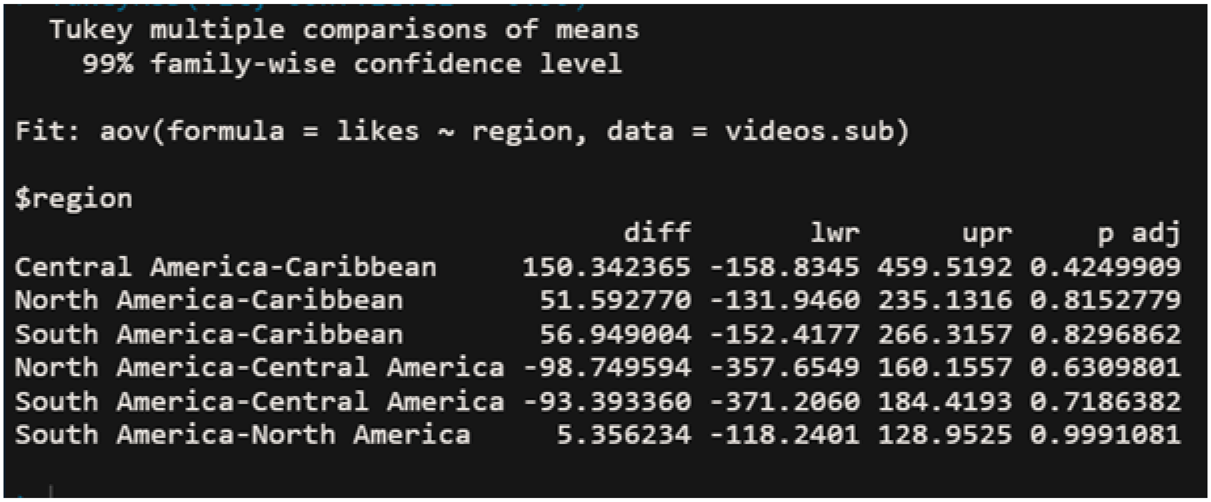

QQ Plots of number of likes by region
sample
3500
3000
2500
2000
1500
Caribbean
Central America
North America
South America
-2
0
2
theoretical



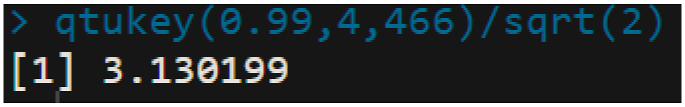
```
        Oceania       Caribbean  Central Africa Central America 
             NA              29              NA              14 
   Central Asia  Central Europe  Eastern Africa    Eastern Asia 
             NA              NA              NA              NA 
 Eastern Europe     Middle East    North Africa   North America 
             NA              NA              NA             352 
    Scandinavia    South Africa   South America      South Asia 
             NA              NA              71              NA 
   South Europe  Southeast Asia  Western Africa  Western Europe 
             NA              NA              NA              NA 
> m <- tapply(videos.sub$likes, videos.sub$region, mean)
> m
        Oceania       Caribbean  Central Africa Central America 
             NA        2439.586              NA        2589.929 
   Central Asia  Central Europe  Eastern Africa    Eastern Asia 
             NA              NA              NA              NA 
 Eastern Europe     Middle East    North Africa   North America 
             NA              NA              NA        2491.179 
    Scandinavia    South Africa   South America      South Asia 
             NA              NA        2496.535              NA 
   South Europe  Southeast Asia  Western Africa  Western Europe 
             NA              NA              NA              NA 
> s <- tapply(videos.sub$likes, videos.sub$region, sd)
> s
        Oceania       Caribbean  Central Africa Central America 
             NA        380.9654              NA        320.5151 
   Central Asia  Central Europe  Eastern Africa    Eastern Asia 
             NA              NA              NA              NA 
 Eastern Europe     Middle East    North Africa   North America 
             NA              NA              NA        293.4937 
    Scandinavia    South Africa   South America      South Asia 
             NA              NA        314.3922              NA 
   South Europe  Southeast Asia  Western Africa  Western Europe 
             NA              NA              NA              NA 
```



```
> summary(fit)
             Df   Sum Sq Mean Sq F value Pr(>F)
region        3   215598   71866    0.78  0.505
Residuals   462 42552865   92106               
>
```



```
  Tukey multiple comparisons of means
    99% family-wise confidence level

Fit: aov(formula = likes ~ region, data = videos.sub)

$region
                                     diff       lwr      upr     p adj
Central America-Caribbean      150.342365 -158.8345 459.5192 0.4249909
North America-Caribbean         51.592770 -131.9460 235.1316 0.8152779
South America-Caribbean         56.949004 -152.4177 266.3157 0.8296862
North America-Central America  -98.749594 -357.6549 160.1557 0.6309801
South America-Central America  -93.393360 -371.2060 184.4193 0.7186382
South America-North America      5.356234 -118.2401 128.9525 0.9991081
```



```
> qtukey(0.99,4,466)/sqrt(2)
[1] 3.130199
```

b) (5 pts.) What statistical procedure should be used and why? If this is for a one-sample, two-sample paired, or two-sample independent, include what the null value is and why you chose that value. Besides the technique itself, be sure to state whether you are performing an inference procedure for a one-sided or two-sided hypothesis (except for ANOVA) with an explanation for the choice. Remember, this needs to be determined BEFORE you analyze the data**.**

We will use an ANOVA to analyze the data because we are comparing the mean number of likes for 4 different regions.

c) (10 pts.) Determine if the appropriate assumptions are satisfied. You may assume that the data set is from an SRS. Even though you are assuming this assumption is true, it must be explicitly stated. Please provide all of the diagnostic graphs to show that the assumptions are met and explain your decision. If the assumptions are not valid for your methodology and you still perform the analysis, you will lose 30 points. If a transformation is needed, state that you have performed a transformation and explain why you felt that a transformation is necessary. The explanation for the transformation should include at a minimum the histogram of the original data. You may include additional graphs of the untransformed variables, if necessary, for your explanation. If you transform your data, you will need to provide all of the diagnostic graphs for the transformed variables.

Assumptions

1. SRS: We are assuming the data set is from an SRS, so this assumption is satisfied.
2. Normality: All the histograms are bell-shaped curves that appear to be approximately normal. Additionally, the red line on each histogram is similar to the blue curve, indicating normality. The normal probability plots all have points that are close to the line, in a linear pattern, showing normality. Because of the histograms and normal probability plots, the normality assumption is satisfied.
3. Constant standard deviation: Smax = 380.9654, Smin = 293.4937

Smax/Smin = 380.9654/293.4937 = 1.298 < 2

Because the maximum standard deviation divided by the minimum standard deviation is less than 2, the constant standard deviation assumption is satisfied.

d) (7 pts.) Graphically display the data as appropriate for your answer in step b) with an interpretation of the output. The point of this part is to understand and explore your data, not merely to check the assumptions needed for inference as you did in step c). Therefore, you will need to state what you think that the conclusion for the inference is BASED ON THE GRAPHS in addition to a description of each of the graphs. Some of the graphs in step d) may have already been used in step c); however, the description of the graphs will be different. For one- and two-sample inferences and ANOVA, you will need to generate the appropriate histograms and boxplots. For what is meant by 'appropriate,' please see the computer assignments and tutorials. For ANOVA and linear regression where additional graphs are required, please see the computer assignments and tutorials to determine what graphs are needed. Remember that your discussion of the graphs should not include whether they are normal or not. Normality is an assumption and is discussed in part c).

The histograms for number of likes in each region are bell shaped curves that appear to be relatively symmetric. The normal probability plots for number of likes in each region appear to be linear. The boxplots have a lot of overlap, indicating that all the means could be the same. There is one outlier shown in the boxplot for the Caribbean at approximately 3500 likes. The effects plot shows that Central America has the largest mean number of likes, North and South America could have the same mean, and the Caribbean has the lowest mean. However, the range on the y-axis of this plot is relatively small, which indicates that there is not a large difference in the mean number of likes. From these plots, there does not appear to be a significant difference in the mean number of likes for each region.

e) (20 pts.) Perform the appropriate inference with a significance level of 0.01. This may consist of more than one step depending on the methodology in step b). The possible methodologies are 1) Confidence interval AND hypothesis test (Chapters 9, 10, and 11): This includes one-sample, two-sample independent and two-sample paired t procedures. Each of these procedures is regarded as a different type of inference. 2) ANOVA (Chapter 12): Both the hypothesis test and the multiple comparison (if appropriate) need to be included. The multiple comparison test is not needed if the hypothesis test does not show a significant result. 3) Linear regression (Chapter 13): At least one inference needs to be included besides the equation of the line. Please see Computer Assignments 9 for possible inferences. All confidence intervals should include the interpretation. All hypothesis tests should consist of the four steps. Be careful about the interpretations when you transform the variables.

Hypothesis Test

1. μ_C = the true population mean number of likes in the Caribbean

μ_{CA} = the true population mean number of likes in Central America

μ_{NA} = the true population mean number of likes in North America

μ_{SA} = the true population mean number of likes in South America

2. H0: $\mu_C = \mu_{CA} = \mu_{NA} = \mu_{SA}$

Ha: at least two μ_i's are different

3. Fts = 0.78

df1 = 3, df2 = 462

p-value = 0.505

4. Since 0.505 > 0.01, we will fail to reject the null hypothesis.

The data does not provide evidence (p-value = 0.505) to support the claim that the population mean number of likes for at least one region is different from the rest.

Tukey Plot:

```
Caribbean      Central America        North America          South America

     ------------------------------------------------------------------------------------
```

f) (8 pts.) The last part is the conclusion of your inference. This part is your final answer to each inference. This is split into two parts. a) If you obtained a significant result, determine the practicality of the answer. This is required even if you transformed your data, or you are performing ANOVA, although these explanations are more difficult. And b) A conclusion in words that relates to the context of the question. This part is a short paragraph explaining your conclusions of the Part and should be understandable to someone who has not taken a course in statistics.

a.) We did not obtain a significant result, so practically there is no difference in the mean number of likes.

b.) In this analysis, the question asks if the number of likes in North America, South America, Caribbean, and Central America are different. We concluded that the assumptions to perform an ANOVA were satisfied. The results of the test indicate that none of the mean number of likes for any region are different.

Q3: (ANOVA) Are mean number of likes in videos of low qualities different from each other?

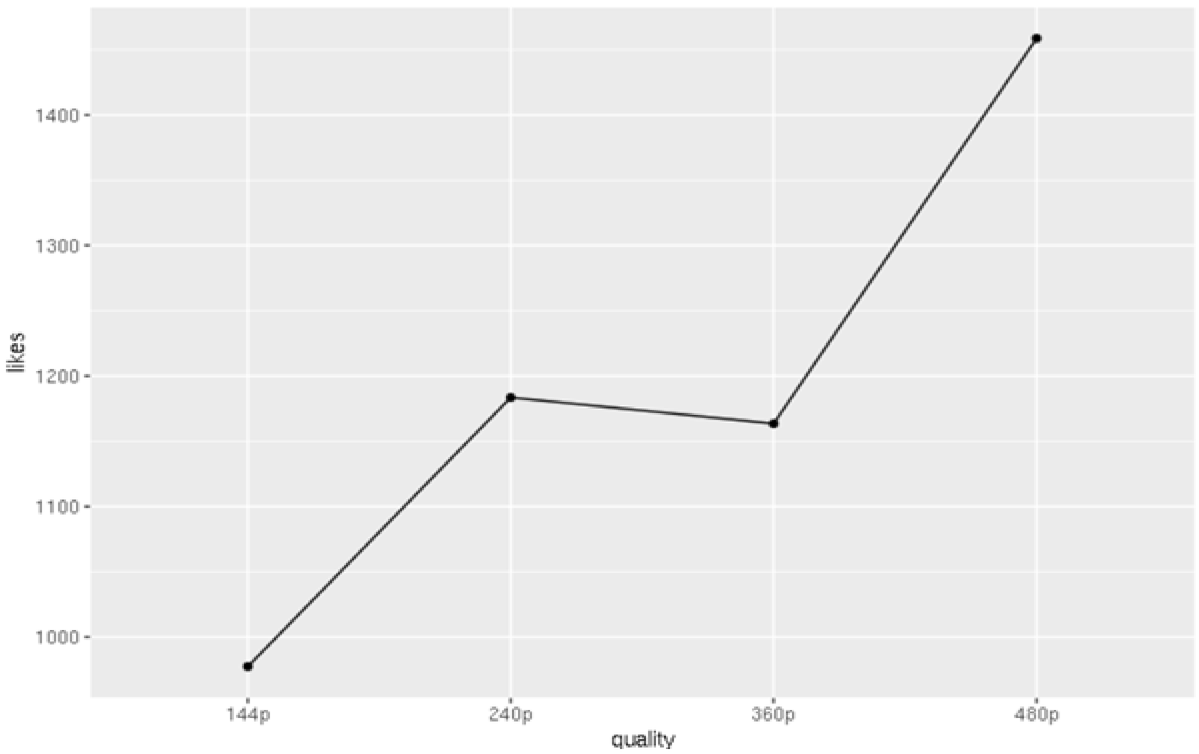
a) Code:

```
videos.sub<-videos[complete.cases(videos),]

videos.aov <- subset(videos, (quality == "144p" | quality == "240p" | quality == "360p" | quality == "480p"), select = c("quality", "likes", "shares"))

View(videos.aov)


library(ggplot2)

# BOXPLOT

# Specify an x variable (categorical) to graph multiple boxplots in

# the same command.

#

ggplot(videos.aov, aes(x = quality, y = likes)) +

  geom_boxplot() +

  stat_boxplot(geom = "errorbar") +

  stat_summary(fun.y = mean, col = "black", geom = "point", size = 3) +

  ggtitle("Boxplots of number of likes by quality")

# a) Effects plot

# We use two calls to stat_summary(): One plots the points ub the

# sample means of each category, the other connects the points with

# lines
```

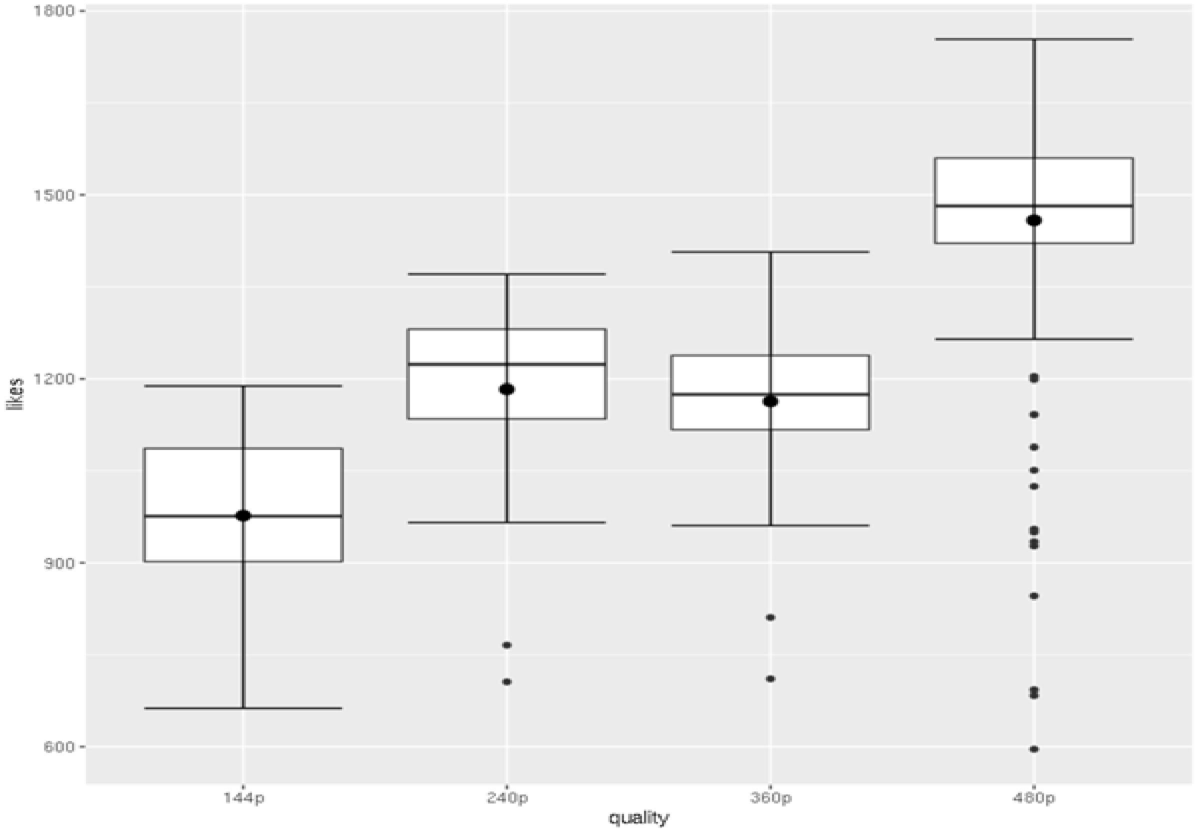
```
ggplot(data = videos.aov, aes(x = quality, y = likes)) +
  stat_summary(fun.y = mean, geom = "point") +
  stat_summary(fun.y = mean, geom = "line", aes(group = 1)) +
  ggtitle("Effects Plot of likes by quality")
#
# Calculating the sample size, sample mean and standard deviation
l <- tapply(videos.aov$likes, videos.aov$quality, length)
m <- tapply(videos.aov$likes, videos.aov$quality, mean)
s <- tapply(videos.aov$likes, videos.aov$quality, sd)
#
# b) Histogram
# This is a little more complicated because we have three groups now.
# The code consists of two steps.
#
# (1) Make theoretical density curve
# You need to specify the name of the numeric response in the square
# brackets of as.numeric(x[]),
# and the name of the grouping variable in xbar[x[]] and s[x[]].
#
xbar <- tapply(videos.aov$likes, videos.aov$quality, mean)
```

```
s <- tapply(videos.aov$likes, videos.aov$quality, sd)
videos.aov$normal.density <- apply(videos.aov, 1, function(x){
  dnorm(as.numeric(x["likes"]),
        xbar[x["quality"]], s[x["quality"]])})
#
# (2) Make the histogram
# Remember that the number of bins in the histogram should be the
# maximum of the number of levels. See Tutorial 7c for details.
#
binlen <- as.numeric(max(tapply(videos.aov$likes,
                                  videos.aov$quality,length)))
ggplot(videos.aov, aes(x = likes)) +
  geom_histogram(aes(y = ..density..), bins = sqrt(binlen) + 2,
                        fill = "grey", col = "black") +
  facet_grid(quality ~ .) +
  geom_density(col = "red", lwd = 1) +
  geom_line(aes(y = normal.density), col = "blue", lwd = 1) +
  ggtitle("Histograms of likes by quality")
#
# QQ plot
```

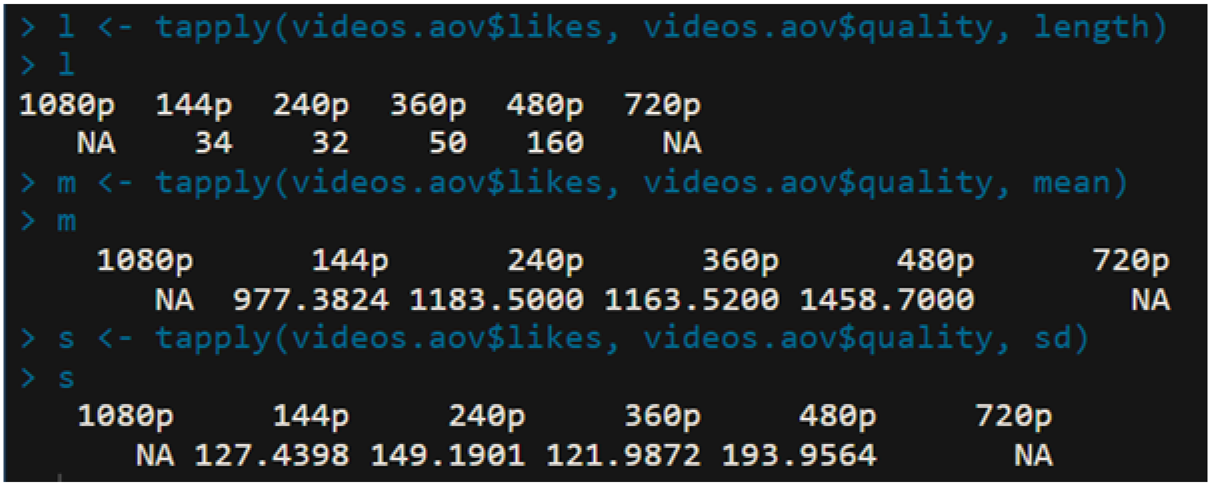
```
#
# (1) Calculate slope and intercept
# You need to put the name of the grouping variable in
# xbar[x[]] and s[x[]].
#
videos.aov$intercept <- apply(videos.aov, 1, function(x){xbar[x["quality"]]})
videos.aov$slope <- apply(videos.aov, 1, function(x){s[x["quality"]]})
#
# (2) Make the QQ plot
#
ggplot(videos.aov, aes(sample = likes)) +
  stat_qq() +
  facet_grid(quality ~ .) +
  geom_abline(data = videos.aov, aes(intercept = intercept, slope = slope)) +
  ggtitle("QQ Plots of likes by quality")
#
# ANOVA
# c) hypothesis test
# The command is aov(numeric ~ categorical, data = datasetName)
# The function is called aov(): a as in analysis, o as in of, and
# v as in variance.
# The categorical variable must be of the class "factor" or "integer."
# To print out the results, you need to use the function summary().
# Note: this does not print out the "total" line in the ANOVA table.
# you may calculate it by hand or via R if required.
fit <- aov(likes ~ quality, data = videos.aov)
summary(fit)
# d) Tukey method
TukeyHSD(fit, conf.level = 0.99)
# critical value:
# parameter: qtukey(ConfidenceLevel, #groups or k, dfe)
# though possible to have R determine what these values are, it is
# easier to put them in by hand after you see the results from aov().
# The critical value is the parameter/sqrt(2)
qtukey(0.99,6,1889)/sqrt(2)
```



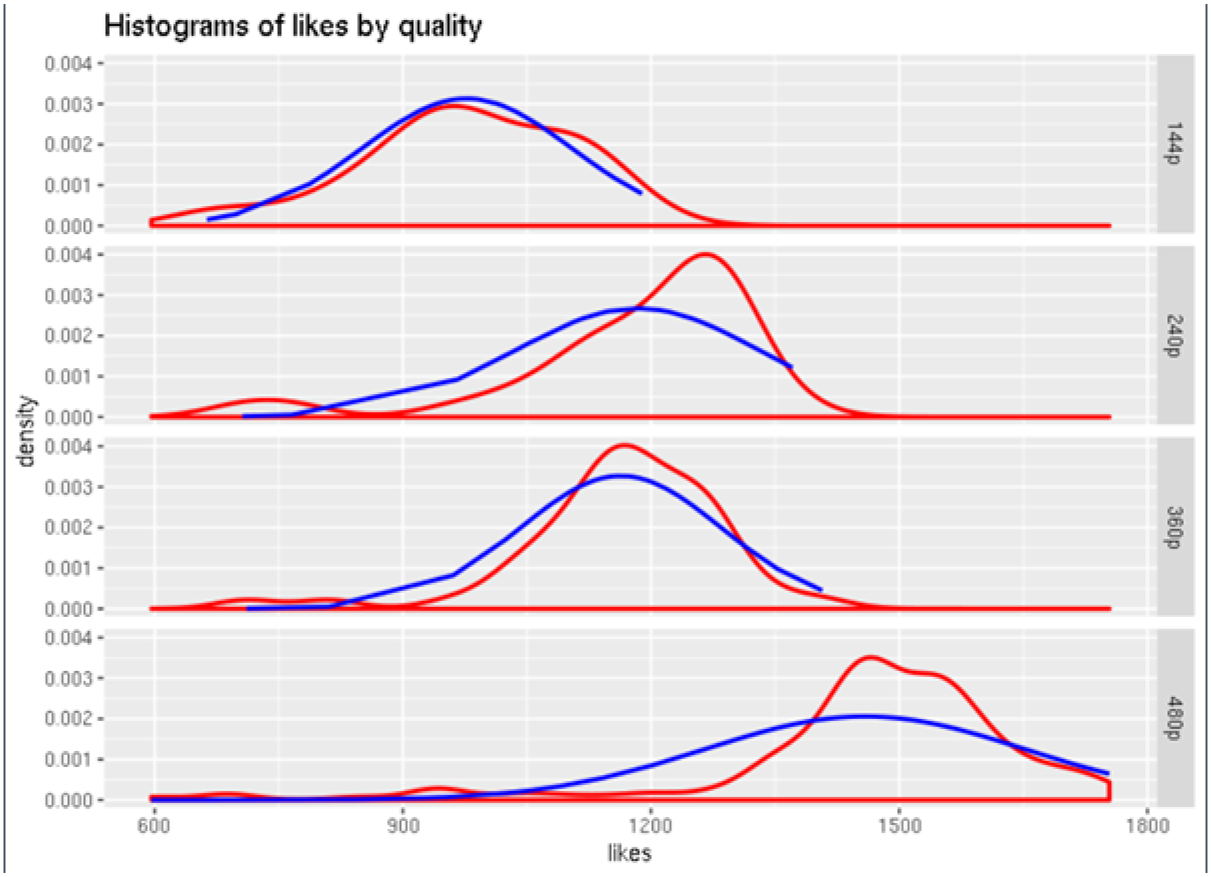
Effects Plot of likes by quality

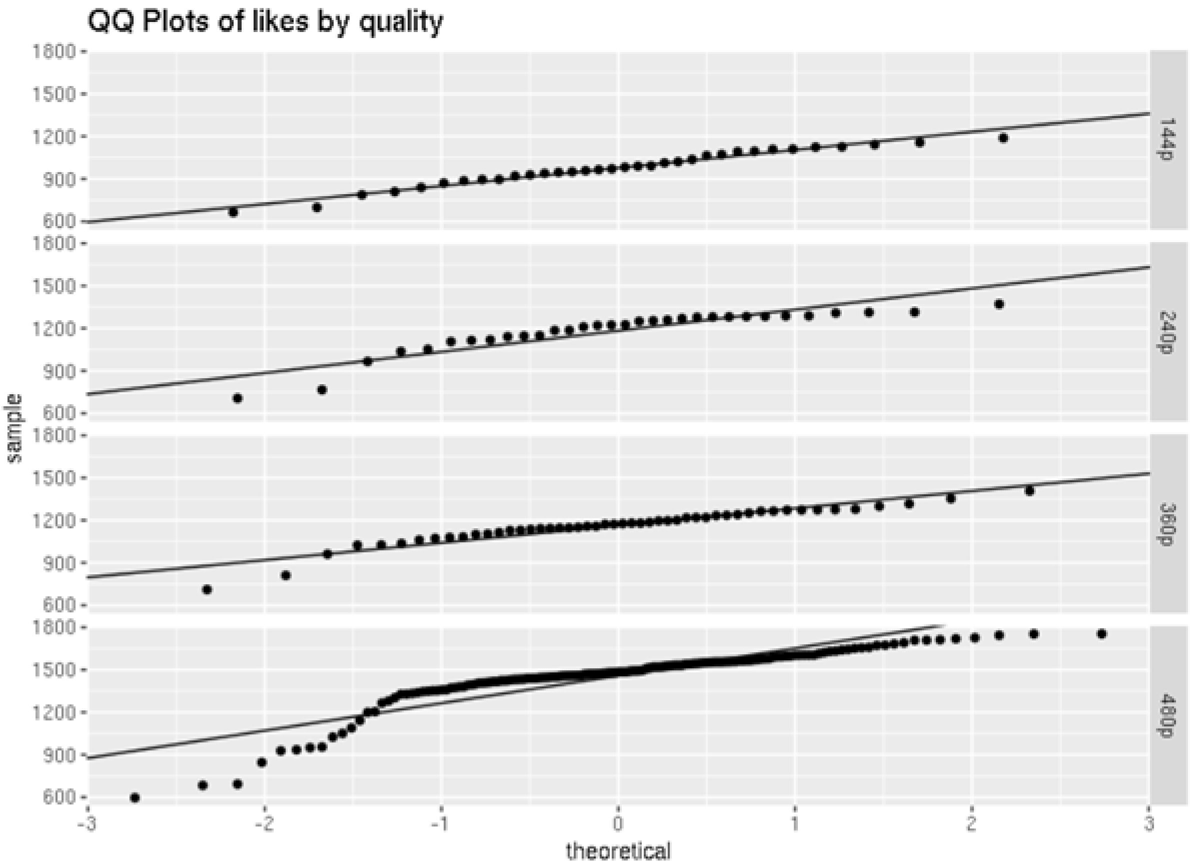


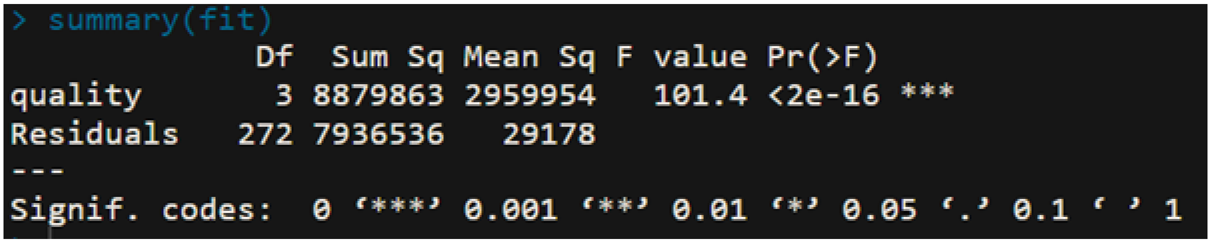
Boxplots of number of likes by quality



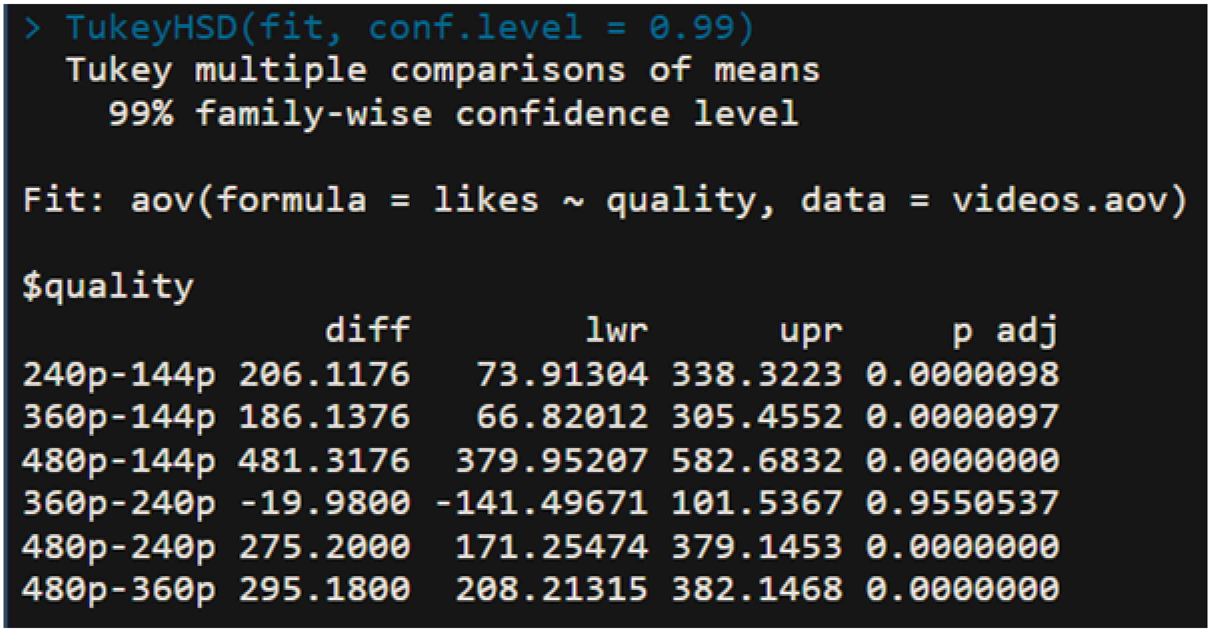
```
> l <- tapply(videos.aov$likes, videos.aov$quality, length)
> l
1080p  144p  240p  360p  480p  720p
   NA    34    32    50   160    NA
> m <- tapply(videos.aov$likes, videos.aov$quality, mean)
> m
    1080p      144p      240p      360p      480p      720p
       NA  977.3824 1183.5000 1163.5200 1458.7000        NA
> s <- tapply(videos.aov$likes, videos.aov$quality, sd)
> s
   1080p     144p     240p     360p     480p     720p
      NA 127.4398 149.1901 121.9872 193.9564       NA
```



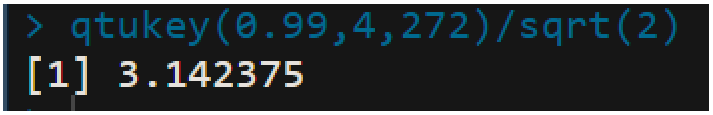




```
> summary(fit)
             Df  Sum Sq Mean Sq F value Pr(>F)    
quality       3 8879863 2959954   101.4 <2e-16 ***
Residuals   272 7936536   29178                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



```
> TukeyHSD(fit, conf.level = 0.99)
  Tukey multiple comparisons of means
    99% family-wise confidence level

Fit: aov(formula = likes ~ quality, data = videos.aov)

$quality
                diff        lwr      upr     p adj
240p-144p 206.1176   73.91304 338.3223 0.0000098
360p-144p 186.1376   66.82012 305.4552 0.0000097
480p-144p 481.3176  379.95207 582.6832 0.0000000
360p-240p -19.9800 -141.49671 101.5367 0.9550537
480p-240p 275.2000  171.25474 379.1453 0.0000000
480p-360p 295.1800  208.21315 382.1468 0.0000000
```



```
> qtukey(0.99,4,272)/sqrt(2)
[1] 3.142375
```

b) (5 pts.) What statistical procedure should be used and why? If this is for a one-sample, two-sample paired, or two-sample independent, include what the null value is and why you chose that value. Besides the technique itself, be sure to state whether you are performing an inference procedure for a one-sided or two-sided hypothesis (except for ANOVA) with an explanation for the choice. Remember, this needs to be determined BEFORE you analyze the data**.**

An ANOVA should be used because we are comparing the mean number of likes for four different video qualities.

c) (10 pts.) Determine if the appropriate assumptions are satisfied. You may assume that the data set is from an SRS. Even though you are assuming this assumption is true, it must be explicitly stated. Please provide all of the diagnostic graphs to show that the assumptions are met and explain your decision. If the assumptions are not valid for your methodology and you still perform the analysis, you will lose 30 points. If a transformation is needed, state that you have performed a transformation and explain why you felt that a transformation is necessary.

The explanation for the transformation should include at a minimum the histogram of the original data. You may include additional graphs of the untransformed variables, if necessary, for your explanation. If you transform your data, you will need to provide all of the diagnostic graphs for the transformed variables.

1. We are assuming that the data set is from an SRS.

2. The histogram for 155p appears to be slightly right skewed. The histograms for 240p and 480p appear to left skewed. The histogram for 360p appears to be somewhat symmetric. The red lines for 144p and 360p closely resemble the blue lines, indicating normality. The normal probability plots for 144p, 240p, and 360p all follow a linear pattern, appearing to be normally distributed. The normal probability plot for 480p has a slight curve facing down, appearing to be left skewed. We can assume normality for all the graphs given the large amount of data points.

3. Smax = 193.9564, Smin = 121.9872

Smax/Smin = 193.9564/121.9872 = 1.58997 < 2

The constant standard deviation assumption is satisfied because the maximum standard deviation divided by the minimum standard deviation is less than 2.

d) (7 pts.) Graphically display the data as appropriate for your answer in step b) with an interpretation of the output. The point of this part is to understand and explore your data, not merely to check the assumptions needed for inference as you did in step c). Therefore, you will need to state what you think that the conclusion for the inference is BASED ON THE GRAPHS in addition to a description of each of the graphs. Some of the graphs in step d) may have already been used in step c); however, the description of the graphs will be different. For one- and two-sample inferences and ANOVA, you will need to generate the appropriate histograms and boxplots. For what is meant by 'appropriate,' please see the computer assignments and tutorials. For ANOVA and linear regression where additional graphs are required, please see the computer assignments and tutorials to determine what graphs are needed. Remember that your discussion of the graphs should not include whether they are normal or not. Normality is an assumption and is discussed in part c).

The histogram for each quality is a bell-shaped curve. The histogram 144p is slightly left skewed, the ones for 240p and 480p are slightly right skewed, and the histogram for 480p is relatively symmetric. The normal probability plots for 144p, 240p, and 380p all have data points close to the line that follow a relatively linear pattern. The normal probability plot for 480p has a slight downward facing curve. The boxplots show that 240p and 360p could have the same mean number of likes because there is a lot of overlap. The mean number of likes for 480p appears to be greater than the other qualities and the mean number of likes for 144p appears to be less than the other qualities. The boxplot for 240p appears to have two outliers that are each below 900 views. With the effects plot, we can rank the mean number of likes for each video quality. 480p has the largest mean, 360p and 240p have relatively the same mean, and 144p has the lowest mean number of views.

e) (20 pts.) Perform the appropriate inference with a significance level of 0.01. This may consist of more than one step depending on the methodology in step b). The possible methodologies are 1) Confidence interval AND hypothesis test (Chapters 9, 10, and 11): This includes onesample, two-sample independent and two-sample paired t procedures. Each of these procedures is regarded as a different type of inference. 2) ANOVA (Chapter 12): Both the hypothesis test and the multiple comparison (if appropriate) need to be included. The multiple comparison test is not needed if the hypothesis test does not show a significant result. 3) Linear regression (Chapter 13): At least one inference needs to be included besides the equation of the line. Please see Computer Assignments 9 for possible inferences. All confidence intervals should include the interpretation. All hypothesis tests should consist of the four steps. Be careful about the interpretations when you transform the variables.

Hypothesis test

1. μ_{144} = the population mean number of likes for a video of quality 144p

μ_{240} = the population mean number of likes for a video of quality 240p

μ_{360} = the population mean number of likes for a video of quality 360p

μ_{480} = the population mean number of likes for a video of quality 480p

2. H0: $\mu_{144} = \mu_{240} = \mu_{360} = \mu_{480}$

Ha: at least two μ_i's are different

3. fts = 101.4

df1 = 3, df2 = 272

p-value = 2e-16

4. We will reject the null hypothesis because the p-value of 2e-16 is less than significance level of 0.01.

The data provides evidence (p-value = 2e-16) to support the claim that the population mean number of likes on a video of at least of the video qualities is different from the rest.

f) (8 pts.) The last part is the conclusion of your inference. This part is your final answer to each inference. This is split into two parts. a) If you obtained a significant result, determine the practicality of the answer. This is required even if you transformed your data, or you are performing ANOVA, although these explanations are more difficult. And b) A conclusion in words that relates to the context of the question. This part is a short paragraph explaining your conclusions of the Part and should be understandable to someone who has not taken a course in statistics.

Tukey Plot

```
144p          240p          360p          480p

                 ------------------------
```

The mean number of likes 240p and 360p are statistically the same. The mean number of likes for the rest of the qualities are statistically different.

360-480: 208.21315/121.9872 = 1.706

144-360: 66.82012/121.9872 = 0.54776

144-480: 379.95207/127.4398 = 2.981

144-240: 73.91304/127.4398 = 0.57998

Practically, these values are all the same because the effect size is relatively small.

This question asked if the number of likes for different video qualities are different. The different video qualities are 144p, 240p, 360p, and 480p. After finding that the assumptions for an ANOVA test are satisfied, we ran the ANOVA to compare the mean number of likes. After the test, we concluded that the mean number of likes for 240p and 360p are statistically the same. This means that 480p has the greatest number of likes, then 240p and 360p have the same number, and 144p has the lowest number of likes.

Q4: (Two-sample t procedure(independent)) 7day average death in North Dakota and South Dakota in 2021.

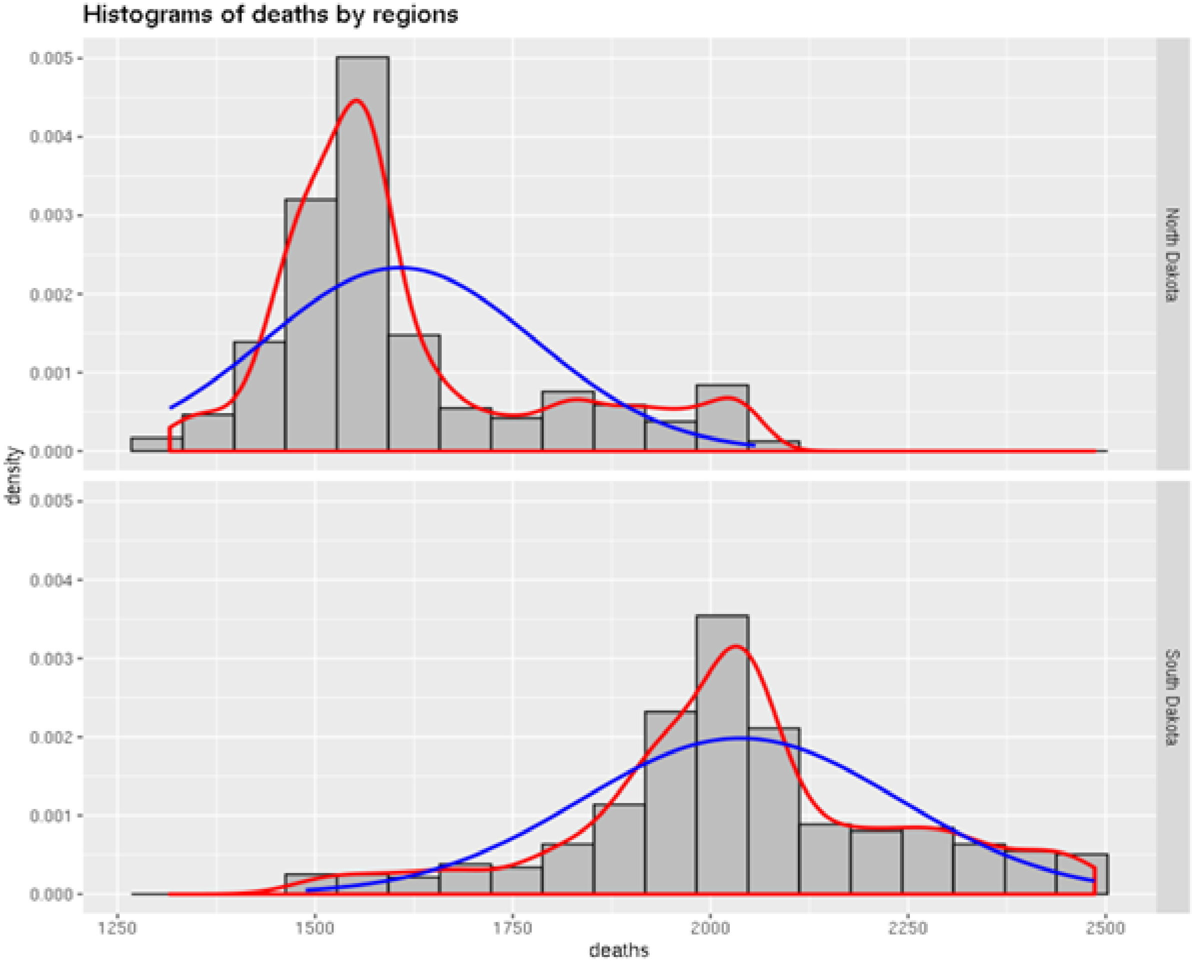
Code:

```
View(covid)

 covid.sub <- subset(covid, (grepl("2021", date, fixed=TRUE)) & (state == "North Dakota" | state == "South Dakota"))

covid.sub$state <- factor(covid.sub$state)


library(ggplot2)
```

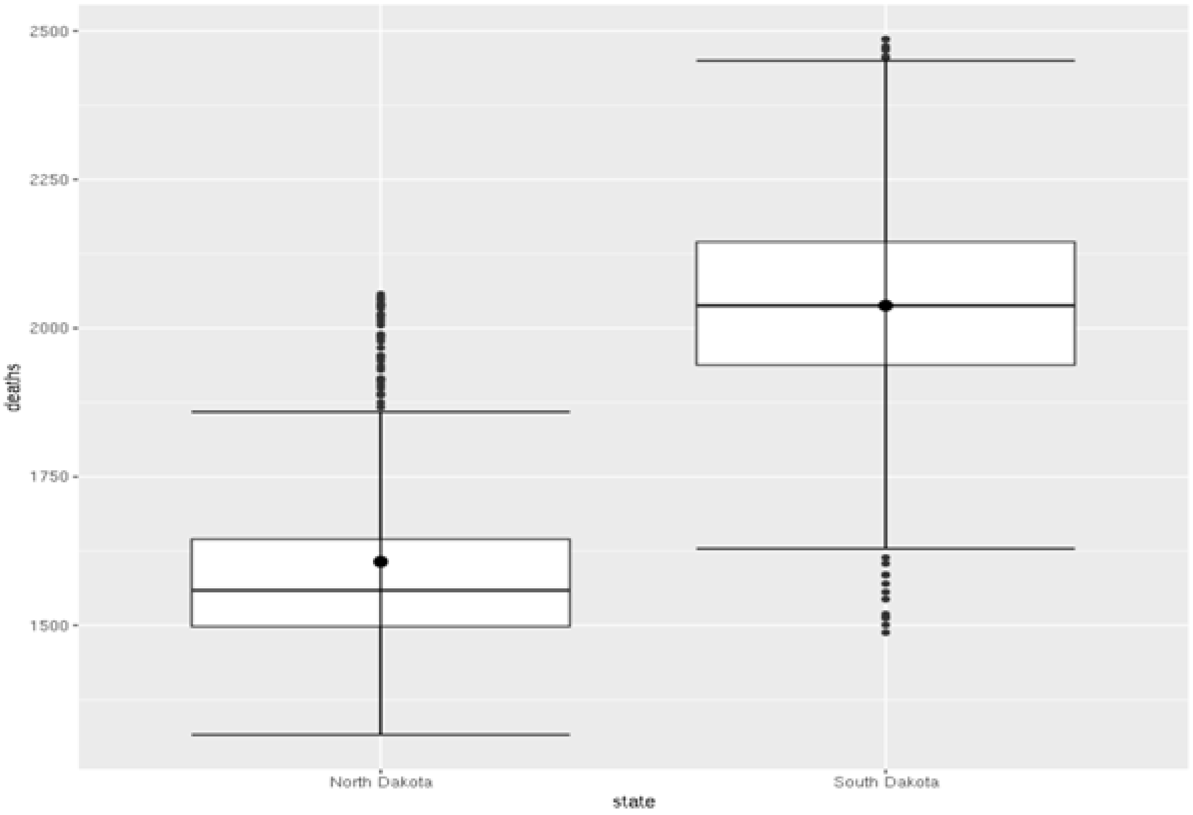
```
############################# BOXPLOT

## HISTOGRAM FOR EACH GROUP

# (1) Obtain sample mean and standard deviation for each group. Now

# xbar and s are vectors.

xbar <- tapply(covid.sub$deaths, covid.sub$state, mean)

s <- tapply(covid.sub$deaths, covid.sub$state, sd)

#

# (2) Create the estimated normal density curve for each group, based

# on xbar and s.

# You need to specify names of the categories in xbar[] and s[].

#

covid.sub$normal.density <- ifelse(covid.sub$state == "North Dakota",

                     dnorm(covid.sub$deaths, xbar["North Dakota"] , s["North Dakota"]),

                     dnorm(covid.sub$deaths, xbar["South Dakota"], s["South Dakota"]))

#

# (3) ggplot with facet_grid(),

# The number of bins has to be the same for each of the histograms.

# We will choose the maximum of the sizes of each of the levels

# facet_grid() tells R what variable contains the categories.

#
```

```
binlen <- as.numeric(max(tapply(covid.sub$deaths, covid.sub$state,length)))
ggplot(covid.sub, aes(x = deaths)) +
  geom_histogram(aes(y = ..density..),
                 bins = sqrt(binlen),
                 fill = "grey", col = "black") +
  facet_grid(state ~ .) +
  geom_density(col = "red", lwd = 1) +
  geom_line(aes(y = normal.density), col = "blue", lwd = 1) +
  ggtitle("Histograms of deaths by regions")
#
# BOXPLOT
# Specify an x variable (categorical) to graph multiple boxplots in
# the same command.
#
ggplot(covid.sub, aes(x = state, y = deaths)) +
  geom_boxplot() +
  stat_boxplot(geom = "errorbar") +
  stat_summary(fun.y = mean, col = "black", geom = "point", size = 3) +
  ggtitle("Boxplots of deaths by regions")
#
```

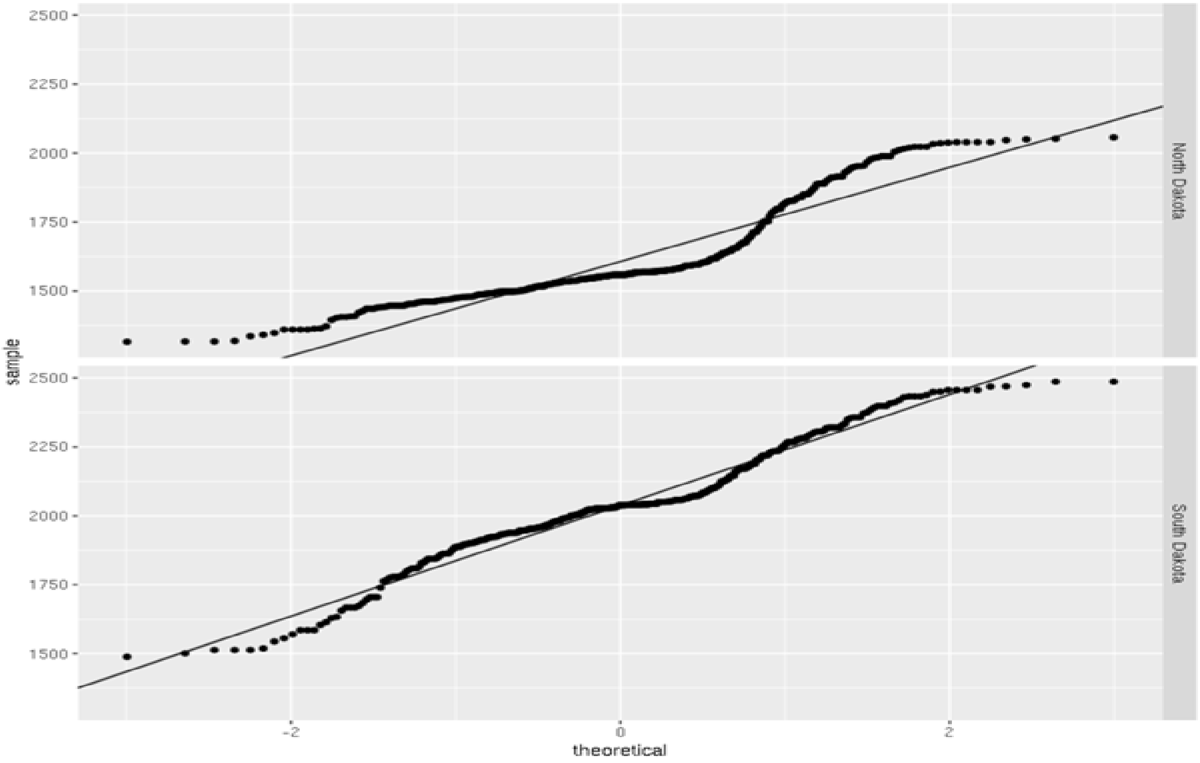
```
# NORMAL PROBABILITY PLOT FOR EACH GROUP
# (1) Calculate slope and intercept of the reference line in
# the normal probability plot for each group. These need to be
# vectors too.
#
covid.sub$intercept <- ifelse(covid.sub$state == "North Dakota",
                              xbar["North Dakota"], xbar["South Dakota"])
covid.sub$slope <- ifelse(covid.sub$state == "North Dakota",
                    s["North Dakota"], s["South Dakota"])
#
# (2) Make normal probability plots using facet_grid()
#
ggplot(covid.sub, aes(sample = deaths)) +
  stat_qq() +
  facet_grid(state~ .) +
  geom_abline(data= covid.sub, aes(intercept = intercept,
                              slope = slope)) +
  ggtitle("QQ Plots of deaths by regions")
#
# t TEST
# In t.test(), the first argument is
# quantitativeVariable ~ categoricalVariable,
# telling R that we want to compare the groups specified by
# the categoricalVariable in terms of the quantitativeVariable.
# The comparison will be based on the alphabetical order of the group
# names ("Men" being the first and "Women" being the second).
# Other parameters for t.test():
# conf.level = C = 1 - alpha
# mu: the null mean, mu_0. The default is 0, but we include it for
# completeness.
# alternative: form of the alternative hypothesis and confidence
# interval/bounds, possible options including
# - "two.sided" (not equal to, confidence interval)
# - "less" (<, upper confidence bound)
# - "greater" (>, lower confidence bound)
# paired: whether the two-sample t-test is paired, possible options
# being TRUE or FALSE. Use FALSE for the two-sample independent
# case.
# var.equal: FALSE means the variances of different groups are
# not assumed equal. In the output, you will see that R calls this
# the Welch approximation (we call it the Satterthwaite

# approximation).

t.test(covid.sub$deaths ~ covid.sub$state, mu = 0, conf.level = 0.99,

    paired = FALSE, alternative = "two.sided", var.equal = FALSE)
```

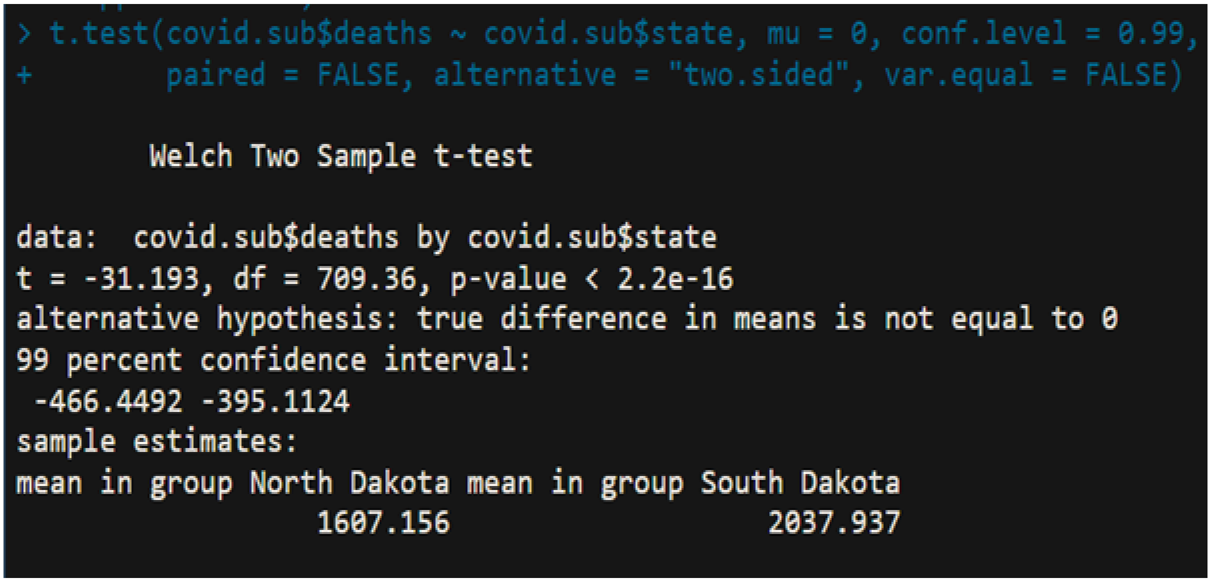


Boxplots of deaths by regions



QQ Plots of deaths by regions





```
> t.test(covid.sub$deaths ~ covid.sub$state, mu = 0, conf.level = 0.99,
+        paired = FALSE, alternative = "two.sided", var.equal = FALSE)

        Welch Two Sample t-test

data:  covid.sub$deaths by covid.sub$state
t = -31.193, df = 709.36, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
 -466.4492 -395.1124
sample estimates:
mean in group North Dakota mean in group South Dakota
                  1607.156                   2037.937
```

Done assuming the significance level is 1%.

b.) Should we use a two-sample independent or a two-sample paired t test?

We should use a two-sample independent t test. This is because selecting the subjects from the richest states does not impact the selection of subjects from the poorest states.

c.) Determine if the appropriate assumptions are satisfied

- SRS: We are assuming that the data is from an SRS
- Normality: We can assume normality due to the large number of data points.

d.) The histogram for North Dakota appears to be left skewed and the red line is not very similar to the blue line. The histogram for South Dakota is approximately symmetric and the red line is relatively similar to the blue line. The normal probability plot for both North and South Dakota has curves. The boxplots do not show any outliers. The boxplot for North Dakota looks slightly right skewed due to the data points above the top whisker. The boxplot for South Dakota appears to be relatively symmetric. From the graphs, it looks as though the number of deaths is higher in South Dakota than North Dakota.

e.) **Calculate the test statistic. Calculate the p-value. Write the four steps for the hypothesis.**

1. μr = the true mean population death rate in the midwest μp = the true mean population death rate in the east coast.

2. H0: μr - μp = 0 Ha: μr - μp != 0

3. tts = -31.193

p = 2.2e-16

df = 709.36

4. Since 2.2e-16 < 0.01, we reject the null hypothesis

The data does provide evidence (p-value = 2.2e-16) to support the claim that the population mean number of deaths for North Dakota is different than the population mean number of death for South Dakota.

99% confidence interval: (-466.4492, -395.1124)

We are 99% confident that the difference in the population mean number of deaths in North Dakota and South Dakota is covered by the interval from (-466.4492 to -395.1124)

f.) This question asks to compare the mean number of deaths due to COVID-19 in North and South Dakota. After determining that the assumptions for the t-test are satisfied, we performed the test to compare the number of deaths. We determined that the number of deaths in North and South Dakota are statistically different.