STAT35000 Final Project

An analysis of Trending Videos and COVID-19

By Justin L, Sam S, Karim E, and Lamya F. B

A. Introduction and questions.

For our video data set, we were interested in three main aspects of the given data. The first question we asked about the data set was, "Is the mean duration of a video less than 11.7 minutes?" We used 11.7 for our H_0 in this case because it is the average based on 2018 data (Ceci). We wanted to see if this was consistent with our data set of more recent, 2020 data, or if there is some sort of discrepancy among the two figures, and if so, look for potential causes in possible lurking variables, such as the year. For our second question, we wanted to see if there is a linear association between daytime views of 1080p videos and the number of recommendations of 1080p videos. The reason we wanted to look at these two variables against each other to identify a possible association is that we wanted to see how the number of times a video is recommended to users affects the number of daytime views it gets. This will help us get an idea of if recommendations truly do help put videos in front of other people. We are only looking at 1080p videos, as the quality of a video may impact a user's chance to watch it, in that very low-quality videos will deter potential viewers. Doing this removes a potential lurking variable. As for our third question about the video data set, we wanted to see if there was a significant difference in the number of views between videos published in each region. We hope that determining this would help us understand the demographic of YouTube's userbase in 2020 much better, such as seeing the amount of high-viewership videos published in America versus those of Africa or Asia. These three questions will serve to deepen our understanding of YouTube's demographic and inner workings as of 2020.

For the COVID-19 question we were primarily interested in what state-level variables influence the deaths due to COVID in that state. Some variables under consideration for this question were GDP, government, and region (titled state.gdp.2017, state.gov, and region in the

provided dataset). We ultimately chose GDP because we wanted to see if the wealth of a state influenced the impact of COVID, being the deaths in that state. Specifically, the relationship between the state deaths due to COVID by July 10th, 2020, and state GDP in 2017. However, it was observed that the Vermont and Utah state GDP in 2017 mismatched data collected by U.S. Bureau of Economics (U.S. BEA), so the decision was made to substitute the 2017 GDP data collected by the BEA for the 2017 GDP data in the COVID-19 data set provided. The first question drafted was "Is there a correlation between the number of deaths due to COVID from March 31, 2020, and state/territory GDP?" which we intended to answer using linear regression. However, we received the feedback to rephrase "correlation" to "linear association" as correlation is an inappropriate term for linear regression as we are specifically looking for a "linear" association, nothing else. The question was then changed to "Is there a linear association between the number of deaths per capita due to COVID from March 31, 2020, and state GDP per capita? If so, what is it?" which has two more revisions besides those suggested to us. We changed "deaths" and "GDP" to "deaths per capita" and "gdp per capita", realizing that states with a higher population were predisposed to have a larger number of deaths and GDP. In order to avoid a potential common response fallacy, we decided to divide both "deaths" and " GDP" variables by their state's population (state.pop.2010). Furthermore, we decided to omit the "territories" and District of Colombia from the COVID data set as they are not in the same situation as the other 50 states and we omitted New York, New Jersey, Connecticut, and Massachusetts from the data set as they have a much larger Deaths per Capita by July 10th, 2020. Overall, this question will serve to increase our understanding of the impact of state GDP on statewide deaths due to COVID-19.

B. Data

1. Table of Variables used in Project

Name of Variable	Description of Variable	Type of Variable (numerical/categorical)
Video Data Set		
duration	Duration of the given video in seconds.	Numerical (discrete)
recommendations	Size of the given video in megabytes.	Numerical (discrete)
day_views	Views of the given video in daytime, based on the user's timezone.	Numerical (discrete)
quality	Quality of the video (144p, 240p, 360p, 480p,720p, 1080p).	Categorical
views	Views of the given video.	Numerical (discrete)
region	Region of the publisher's IP (Africa, America, Asia, Europe, Oceania). Each category has subsections, such as Eastern Europe and South Africa.	Categorical
reg_new *(Defined in preparations, determined from region)*	More general region of the publisher's IP, separated into Africa, America, Asia, Europe, and Oceania.	Categorical
Covid/BEA Data Set		
state.deaths.7.10	A given state's number of reported deaths from COVID-19 on 7/10/20.	Numerical (discrete)
stateGDP2017 (BEA)	GDP of the state evaluated in 2017..	Numerical (continuous)
deathsPerCapita	A given state's number of reported deaths from COVID-19 on 7/10/20 divided by the state population in 2010.	Numerical (discrete)
gdpPerCapita	GDP of the state evaluated in 2017 divided by the 2010 population.	Numerical (continuous)
state.pop.2010	The population of a given state.	Numerical (discrete)

Note regarding data from BEA: The BEA estimated state level GDP using data compiled from various government agencies, primarily the Bureau of Labor Statistics, and is laid out in great detail in the documentation tab under the Citation (U.S. BEA). This data does not greatly differ from any of the state.gdp.2017 data points provided with the exception of Utah and Vermont, which is discussed in Section F Assumption I.

2. Video Data Set Cleaning.

```
library(ggplot2)
videos <- read.table("videos", header = TRUE, sep = "\t")
videos_cleaned <- videos[complete.cases(videos),]
write.table(videos_cleaned,file="videos_cleaned",
            row.names=FALSE, sep="\t")
videos_cleaned <- subset(videos_cleaned,
                         select = c("duration", "recommendations",
                                    "day_views", "quality", "views", "region"))

# Bunching together all the various sub-regions into their main region
videos_cleaned$reg_main <- as.character(videos_cleaned$region)
videos_cleaned$reg_main[videos_cleaned$reg_main =="North Africa"] <-"Africa"
videos_cleaned$reg_main[videos_cleaned$reg_main =="Eastern Africa"] <-"Africa"
videos_cleaned$reg_main[videos_cleaned$reg_main =="Central Africa"] <-"Africa"
videos_cleaned$reg_main[videos_cleaned$reg_main =="Western Africa"] <-"Africa"
videos_cleaned$reg_main[videos_cleaned$reg_main =="South Africa"] <-"Africa"
videos_cleaned$reg_main[videos_cleaned$reg_main =="North America"] <-"America"
videos_cleaned$reg_main[videos_cleaned$reg_main =="Central America"] <-"America"
videos_cleaned$reg_main[videos_cleaned$reg_main =="South America"] <-"America"
videos_cleaned$reg_main[videos_cleaned$reg_main =="Caribbean"] <-"America"
videos_cleaned$reg_main[videos_cleaned$reg_main =="Eastern Asia"] <-"Asia"
videos_cleaned$reg_main[videos_cleaned$reg_main =="Southeast Asia"] <-"Asia"
videos_cleaned$reg_main[videos_cleaned$reg_main =="Central Asia"] <-"Asia"
videos_cleaned$reg_main[videos_cleaned$reg_main =="South Asia"] <-"Asia"
videos_cleaned$reg_main[videos_cleaned$reg_main =="Middle East"] <-"Asia"
videos_cleaned$reg_main[videos_cleaned$reg_main =="Eastern Europe"] <-"Europe"
videos_cleaned$reg_main[videos_cleaned$reg_main =="Central Europe"] <-"Europe"
videos_cleaned$reg_main[videos_cleaned$reg_main =="Western Europe"] <-"Europe"
videos_cleaned$reg_main[videos_cleaned$reg_main =="South Europe"] <-"Europe"
videos_cleaned$reg_main[videos_cleaned$reg_main =="Scandinavia"] <-"Europe"
```

3. Covid Data Set Cleaning.

```
#clean overall dataset
covid_cleaned <- covidF21[complete.cases(covidF21),]

#import new GDP Data
newGDPData <- updatedGDPDataBEA

#remove row 9, D.C. from GDP Data
newGDPData <- newGDPData[-c(9),]

#Remove non-states and territories as they do not have the same situation as the rest
of the US
covid_cleaned <- covid_cleaned[-c(48,37,42,12,9), ]

#replace faulty 2017 GDP data set with typo-free data from U.S. BEA
covid_cleaned$state.gdp.2017 <- newGDPData$X1

#setup data subset
covidGDPDeaths.sub <- subset(covid_cleaned, select =
c("state.gdp.2017","state.death.7.10", "state.pop.2010", "name"))

#in order to remove potentially spurious variable of population divide by population
to give
#deaths per capita and gdp per capita
covidGDPDeaths.sub$deathsPerCapita = covidGDPDeaths.sub$state.death.7.10 /
covidGDPDeaths.sub$state.pop.2010
covidGDPDeaths.sub$gdpPerCapita = covidGDPDeaths.sub$state.gdp.2017 /
covidGDPDeaths.sub$state.pop.2010
```

C. Inference 1 (video data set).

Is the mean duration of a video equal to 11.7 minutes?

A. Code.

a. Relevant code is listed in the Appendix under Code for Part C.

B. Statistical Procedure.

In order to answer the question “is the mean duration of a video less than 11.7 minutes?”, we will use a 1-sample t-test. However, there is a little problem, the dataset includes values in seconds, not minutes. So we multiplied 11.7 by 60 to get into the correct units (seconds), then took the natural log of it to normalize the data. We ended with a value of 6.553933 which is what will be used for mu. We are using a 1-sample t-test because we want to know the relationship between the data and our mean.

H_0 - mu = 6.553933

H_a - mu != 6.553933

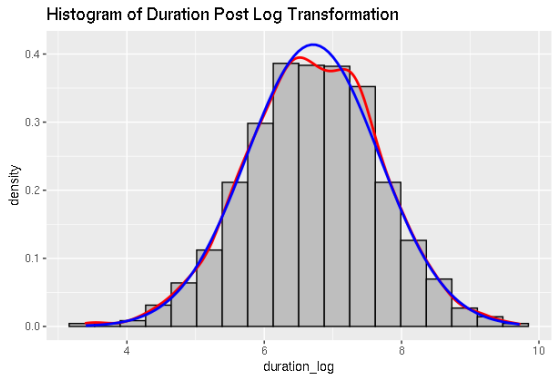
C. Assumptions for Inference.

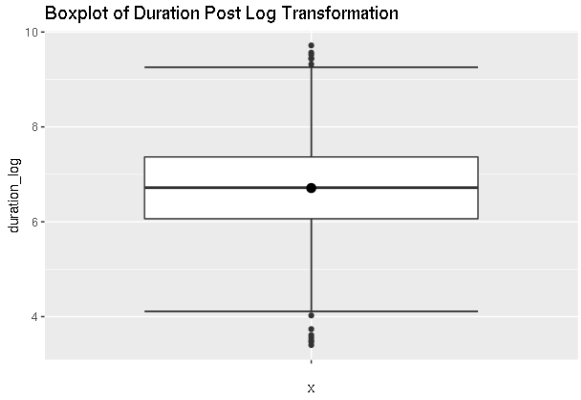
Assumption 1: Data is obtained from an SRS

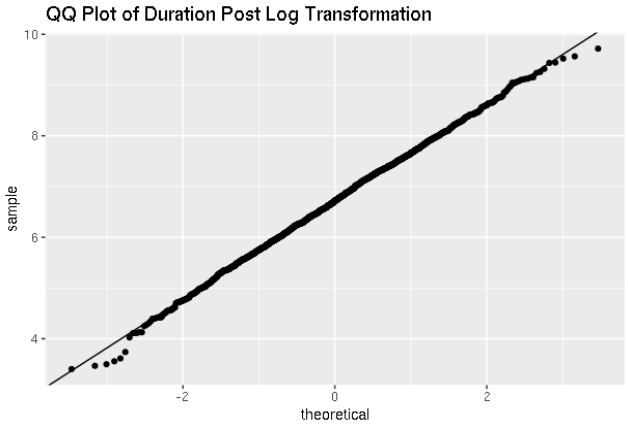
Assumption 2: Data is normal

We’re assuming that the videos that were sampled were sampled using an SRS. The histogram, boxplot, and probability plot are all shown below. The general description is that this distribution is approximated by a normal distribution so it is symmetric. There are small deviations, however, with the large sample size, it is appropriate to overlook them and call this distribution normal. The histogram is approximately symmetric and unimodal and the estimated density curve closely follows the theoretical normal density. This suggests that the distribution of the duration is

approximately normal. The boxplot reveals an exceptionally symmetric shape that has some outliers. However, none of these outliers seem 'real' so there's no need to question the validity of a normal model for the data for t-procedures.The points on the probability plot mostly follow the line, but there are some minor fluctuations on both ends. However, with such a large sample size, this can be overlooked and assumed normal.

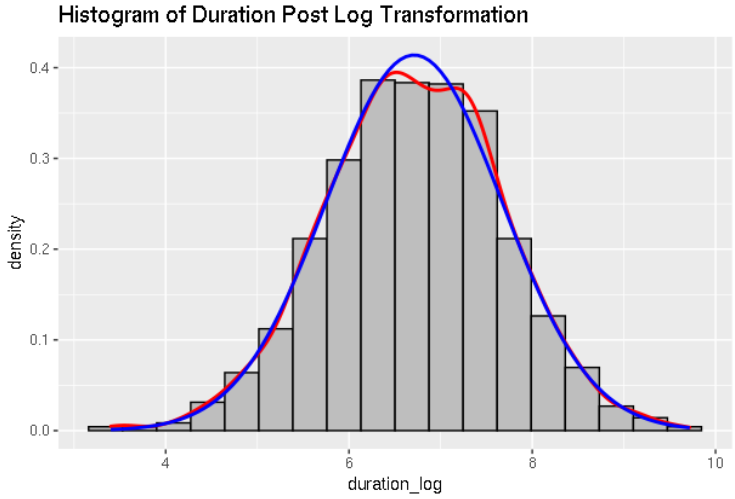


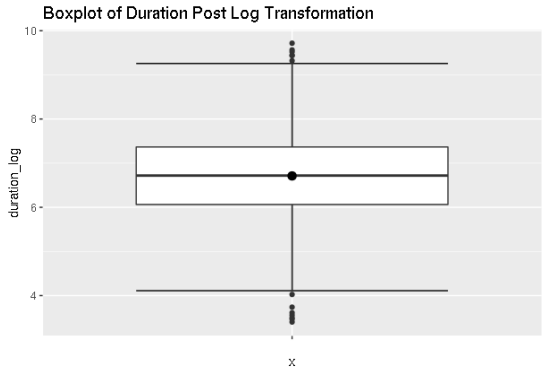




D. Appropriateness Check Graphs for Statistical Inference.

The histogram is approximately symmetric and unimodal and the estimated density curve closely follows the theoretical normal density. The boxplot reveals an exceptionally symmetric shape that has some outliers. However, none of these outliers seem 'real' so there's no need to question the validity of a normal model for the data for t-procedures.





E. Statistical Inference.

```
        One Sample t-test

data:  videos_cleaned$duration_log
t = 7.0913, df = 1894, p-value = 1.868e-12
alternative hypothesis: true mean is not equal to 6.553933
99 percent confidence interval:
 6.653854 6.768033
sample estimates:
mean of x
 6.710944
```

Step 1: Mu is the mean duration of a video

Step 2: H0: mu = 6.553933 Ha: mu != 6.553933

Step 3: tts = 7.0913 df = 1894 p-value = 1.868e-12

Step 4: Since 1.868e-12 is < 0.01 (p-value <= alpha), we reject the null hypothesis.

The data provides evidence (p = 1.868e-12) that the true mean duration of a video is not 6.553933.

The 99% confidence interval is between 6.653854 and 6.768033. R reports that as (6.653854,6.768033). We are 99% confident that the true mean of the duration of videos is covered by the interval (6.653854,6.768033).

The critical value is 2.328318

Effect size: $|(6.553933 - 6.653854)/0.9638403| = 0.222132235$

F. Conclusion.

The question being put forward is: “Is the mean duration of a video equal to 11.7 minutes”? Since we used a log function to normalize the data and the data was in seconds not minutes, the 11.7 minutes will be changed to 6.553933 seconds. It is apparent that the mean duration of a youtube video is not 6.553933 seconds and so we reject the null hypothesis. We are 99% confident that the true mean is captured within the interval (6.653854,6.768033) with an alpha of 0.01, a p-value of 1.868e-12, and an effect size of 0.222132235 (where s = 0.9638403 and mu = 6.653854). This conclusion is consistent with the data which is that since the mean of 6.553933 is not contained in the interval, we would reject the null hypothesis. Meaning, the true mean duration of a youtube video is not 6.553933 seconds, but slightly higher. This makes sense because our sample size is 1875 youtube videos, but there’s a much larger size of videos on youtube, so this isn’t exactly accurate. We can also say that the true mean duration of youtube videos is not 11.7 minutes, since we can just change the 6.553933 seconds back to minutes. However, even with all of this, I wouldn’t say that this is very significant. The difference between the mean values isn’t very large and the effect size is quite small. It also wouldn’t be noticeable to most people, if any people, whether the average was in between (6.653854,6.768033) or 6.553933 seconds. When thinking about this, it makes sense that the true average is not 11.7 minutes. Because even though the average is 11.7 minutes (according to L. Ceci), there are so many videos in the film and animation, entertainment, and gaming industry that make that average higher. Not to mention that some categories like gaming are more popular than others, which brings that average higher or lower depending on the category (higher for gaming because gaming is very popular and getting more popular). There are also so many videos in general, that looking at such a small set, compared to the amount of videos on YouTube, is like looking at a small rock and comparing it to the Earth. Therefore, the conclusion is consistent with the data and makes perfect sense.

D. Inference 2 (video data set).

Is there a linear association between the daytime views of 1080p videos and the number of recommendations of 1080p videos?

A. Code.

a. Relevant code is listed in the Appendix under Code for Part D.

B. Statistical Procedure.

To answer the question of if there is a linear association between the daytime views of 1080p videos and the number of recommendations of 1080p videos, the statistical procedure of Linear Regression should be used. This is because linear regression is best used when we are looking to see if there is an association between the two variables and if said association is linear or not, and in this case, that is the crux of our question. Our inference procedure is for the following set of null and alternative hypotheses:

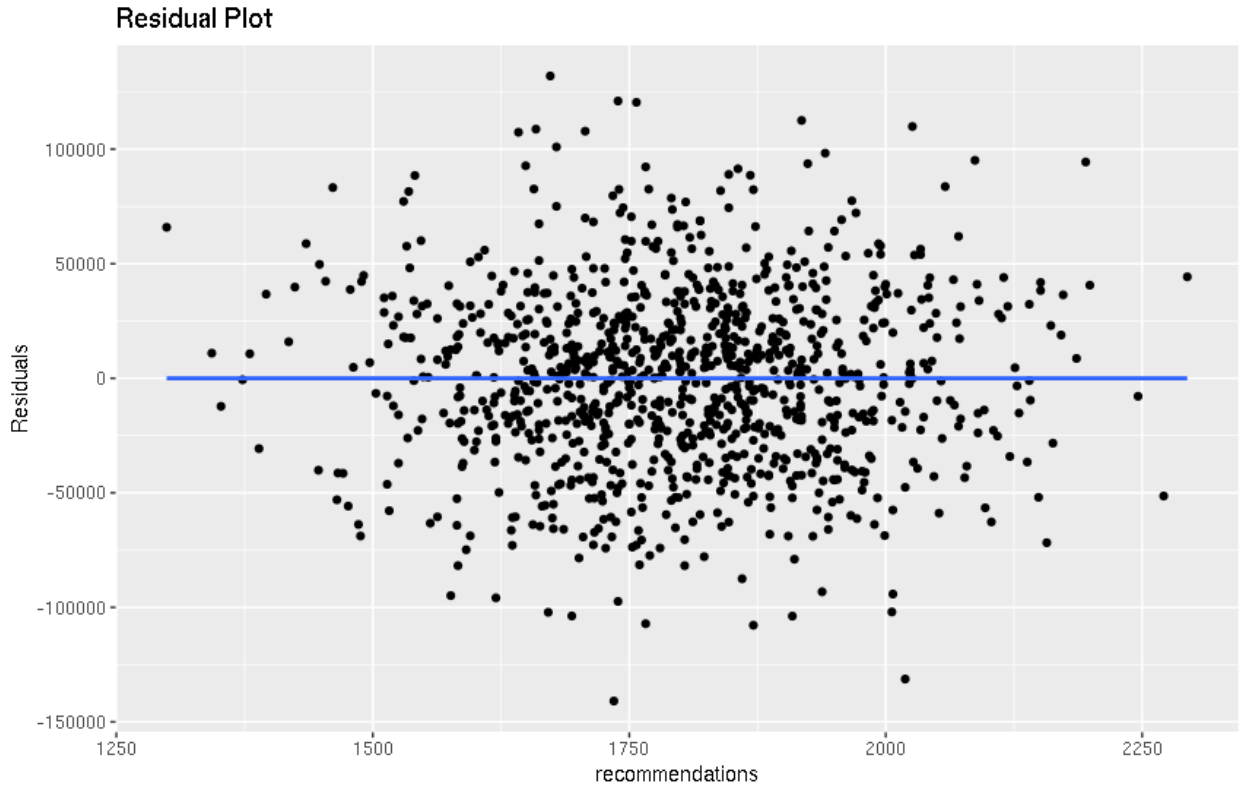
H_0 - There is no association between the video recommendations and daytime views of 1080p videos.
H_a - There is an association between the video recommendations and daytime views of 1080p videos.

C. Assumptions for Inference.

Assumption I: Data is obtained from a SRS.

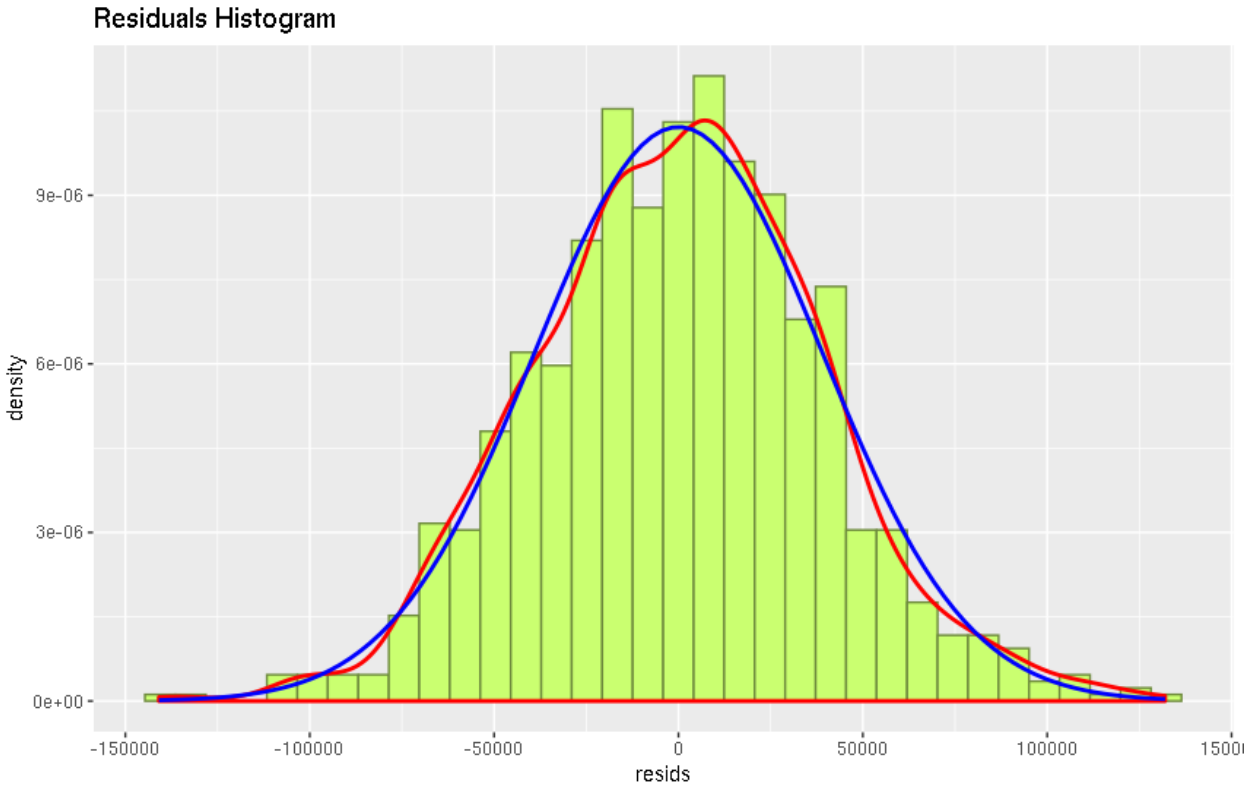
- This is assumed true, as the data given is confirmed to be from a valid simple random sample.

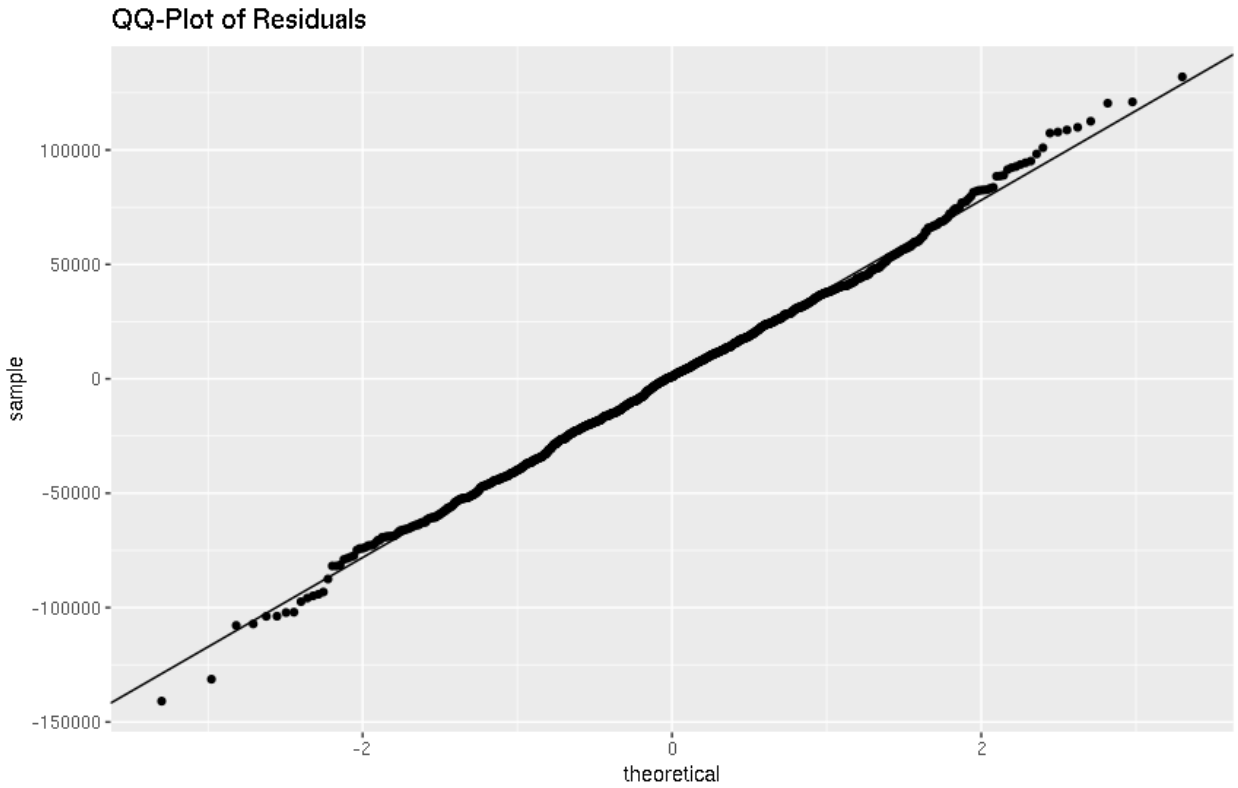
Assumption II: Relationship between X and Y is linear.



- We know this is true as well, as there is no pattern in the above graph, so the association between the video recommendations and daytime views of 1080p videos seems to be linear.
- The variability does sway slightly near the ends of the graph, which may be a sign of a non-constant standard deviation, however, there are less points overall at these ends, so it is significantly harder to assess the variability at these points.
- Additionally, there are no clear outliers present in the plot.

Assumption III: Residuals have a normal distribution.



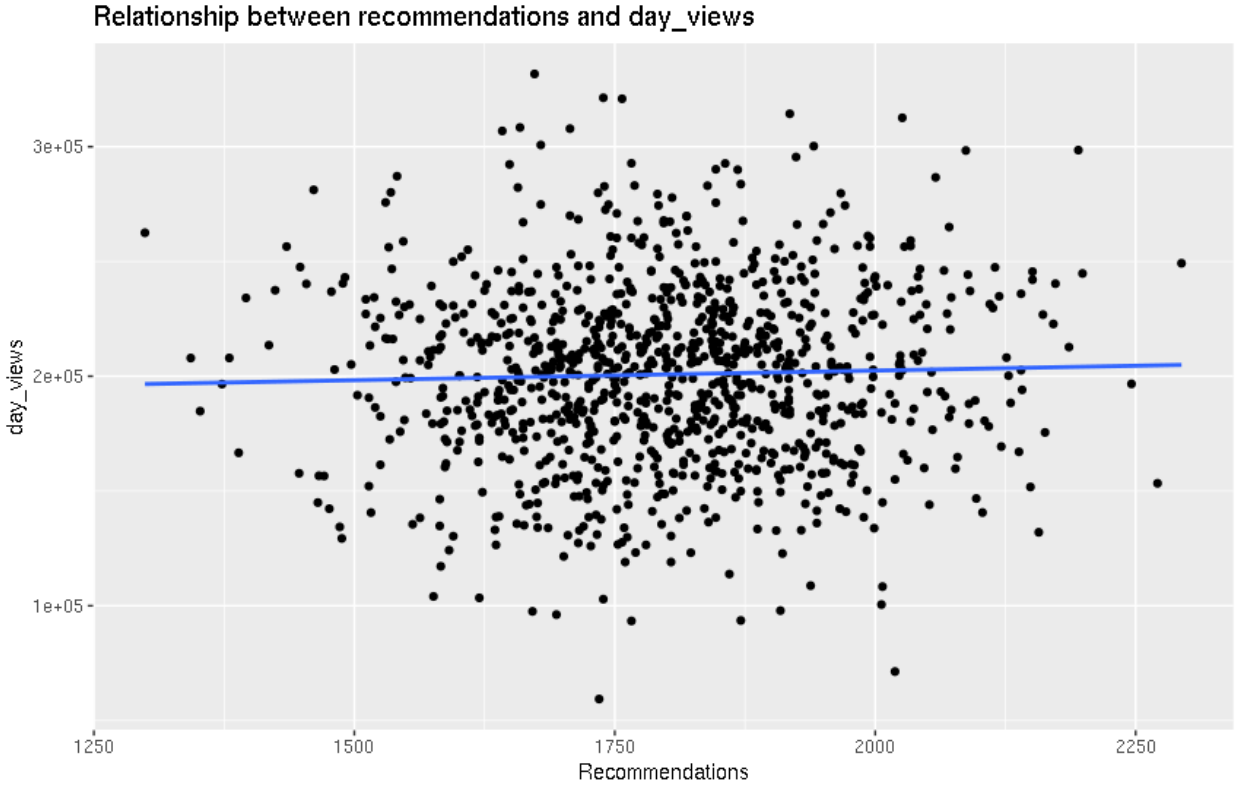


- The above histogram and QQ plot of the residuals show clear normality.
- The histogram's distribution is symmetric and unimodal, as well as the fact that the red estimated kernel density curve and the blue estimated normal density curve are very close together. All of these are indicators of a normal distribution.
- As for the QQ plot, all points are close to the line, with no systematic deviation. At the very left endpoint, the points dip slightly below the line, but there is no overarching pattern that indicates a non-normal distribution.

Assumption IV: Standard deviation of the residuals is constant.

- As seen in the Residual Plot of Assumption II, the variation between the points is constant throughout. The endpoints of the plot have slight abnormalities in their variations, however, this is due to the lack of data points at those ends as opposed to a non-constant standard deviation of the residuals.

D. Appropriateness Check Graphs for Statistical Inference.



- Pictured above is the scatterplot of the daytime views versus recommendations of the 1080p videos. In blue is the linear regression line of best fit.
- The first notable comparison that can be made from this graph is that there is visually no positive effect of recommendations on daytime views at all, despite the fact that one would think that an increasing number of recommendations would help boost its daytime views significantly.
- A common line of thinking is that more recommendations leads to more views, but this graph does not mirror this pattern.

E. Statistical Inference.

```
> cor(videos.sub$recommendations, videos.sub$day_views)
[1] 0.03272429
```

- Correlation Coefficient: 0.033. This correlation coefficient is very small, not quite zero, but it is classified as Weak. We cannot quite say that there is no association between the two either, it just means that the points are going to fall farther away from the line than if we had a Moderate or Strong correlation.

```
                    0.5 %       99.5 %
(Intercept)     148560.51339 222846.42510
recommendations    -12.19231     28.95553
```

- Slope of the Fit Line -
  - We are 99% confident that the population slope of daytime views versus recommendations is covered by the interval from:
    (-12.192 to 28.956).

- Intercept of the Fit Line -
  - We are 99% confident that the population y-intercept of daytime views versus recommendations is covered by the interval from:
    (148560.513 to 222846.425).

- Examples of Population Mean View Estimations based off of Recommendation Count -

```
> newdata <- data.frame(recommendations = 1250)
> predict(videos.sub.lm, newdata, interval = "confidence", level = 0.99)
       fit    lwr    upr
1 196180.5 184460 207901
```

  - We are 99% confident that the population mean daytime views when there are 1250 recommendations is covered by the interval from:
    (184460 to 207901).

```
> newdata <- data.frame(recommendations = 1750)
> predict(videos.sub.lm, newdata, interval = "confidence", level = 0.99)
       fit      lwr      upr
1 200371.3 197076.8 203665.8
```

  - We are 99% confident that the population mean daytime views when there are 1750 recommendations is covered by the interval from:
    (197076.8 to 203665.8).

```
> newdata <- data.frame(recommendations = 2250)
> predict(videos.sub.lm, newdata, interval = "confidence", level = 0.99)
       fit      lwr      upr
1 204562.1 194765.1 214359.1
```

  - We are 99% confident that the population mean daytime views when there are 2250 recommendations is covered by the interval from:
    (194765.1 to 214359.1).
- The similarities between the intervals even though the recommendation number varies wildly is telling that recommendations don't affect views that heavily, if at all.

F. Conclusion.

The question being posed is, "Is there a linear association between the daytime views of 1080p videos and the number of recommendations of 1080p videos"? There is such a small correlation between the daytime views of 1080p videos and the number of recommendations of 1080p videos that there is essentially not a linear association between the two quantities at all. If

one were to try to fit the data despite the very small correlation, the line's slope would be very small in context - we are 99% sure that the number of daytime views a recommendation would add would be between ~-12 and ~29. Considering the high amounts of daytime views videos have, these numbers are insignificant in comparison. This very much goes against what one would expect out of comparing these two values against each other - YouTube touts that its recommendation feature helps put videos in front of viewers, but this test shows that this is clearly not the case. It prompts the question of why some videos get more exposure than others to viewers despite being recommended less? This is likely due to the lurking variable of the Subscriber + Notification system. Users can "Subscribe" to channels to be notified when they upload, so it's natural that they will prefer to watch those channels that they already know that they like (hence their subscription) instead of those recommended by the algorithm. Factoring in the Subscriber system, it is more obvious why the correlation is so weak. Further testing will hopefully show that Subscribers and Daytime Views are strongly correlated, however, we cannot conclude this with our current data set that lacks subscriber data.

E. Inference 3 (video data set).

Is there a difference in the mean duration of trending videos published in 2020 in the four major continents? (ANOVA)

A. Code.

a. Relevant code is listed in the Appendix under Code for Part E.

B. Statistical Procedure.

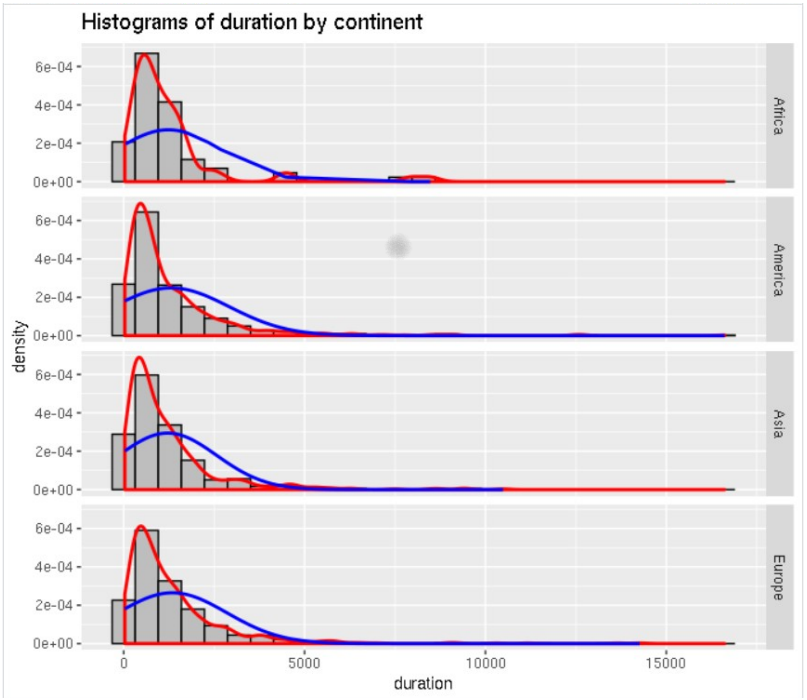
To answer the question on whether or not there was a difference in the mean duration of the trending videos published in 2020 in each continent. Based on the factor of continents with 5 different levels we want to evaluate what the mean of the duration of the trending videos published in 2020 in all continents. As there are more than 2 independent groups that are being compared we chose to use ANOVA.
Our inference is for the following inference hypotheses:

H_0 - There is no difference between the mean duration of trending videos published in 2020 among Africa, America, Asia and Europe.
H_a - There is a difference between the mean duration of trending videos published in 2020 among Africa, America, Asia and Europe.

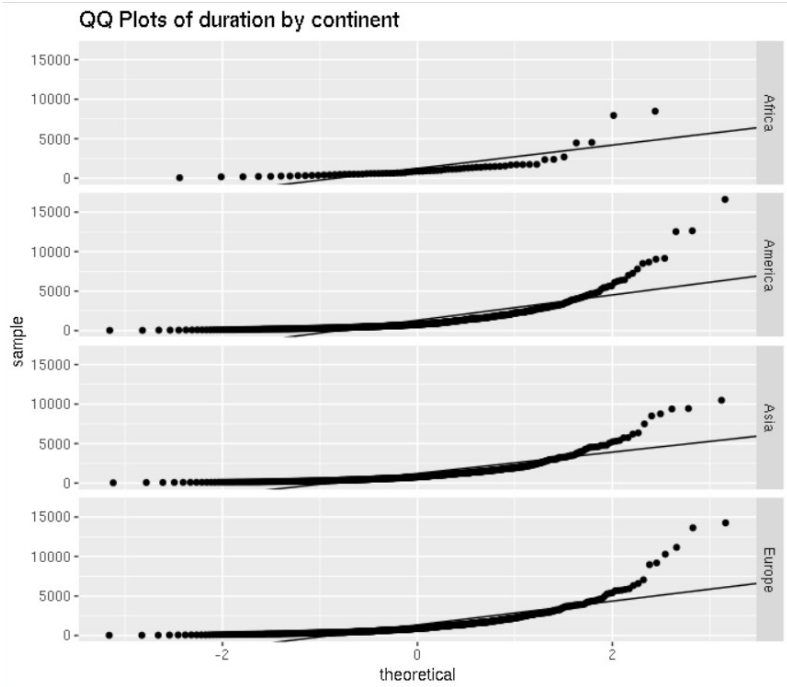
C. Assumptions for Inference.

Assumption 1: We assume that the data is a simple random sample (SRS) and we assume all the populations are independent.
Assumption 2: To continue the ANOVA test we need each population to be normal and that is checked by creating a histogram and a normality plot. When creating these graphs we realized that there were only 9 data points from Oceania. Due to the few points (less than 30) compared to the hundreds of points from the other locations the team chose to leave out the data from Oceania.

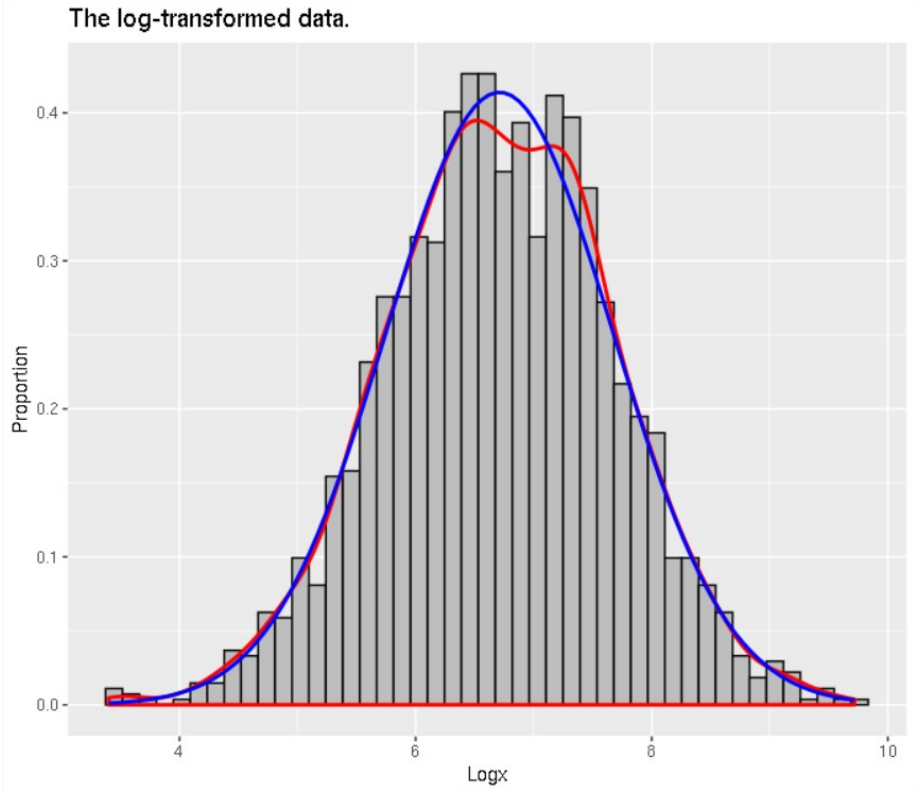
- Normality check:



This graph was created of the duration of the videos based on the continents from which they were released. The data is unfortunately largely skewed. This skew was also reflected on the normality plots created for the 4 continents:



As we can see here most of the points lie on the left of line of best fit which show that the data is heavily right skewed.
To fix this skew we chose to use a log transformation to transform the duration graph



Assumption 3:
To perform this test we will be pooling in ANOVA which means that all of the populations have to have the same variance and so to check that we found the standard deviations for all the data for each continent individually and then checked to see if the ratio of smax/smin is less than two.

Continent	Standard deviation
Asia	`> sd(videosasia$duration)` `[1] 1354.768`
Africa	`> sd(videosafrica$duration)` `[1] 1476.573`
America (not a continent but large body)	`> sd(videosamerica$duration)` `[1] 1605.527`
Europe	`> sd(videoseurope$duration)` `[1] 1507.483`

Here S max = 1605.527 and S min = 1354.768
Hence the ratio = 1605.527/1354.768 = 1.1850
As 1.18 is less than 2 we can assume that the populations have the same variance.

D. Appropriateness Check Graphs for Statistical Inference.

```
                Df    Sum Sq Mean Sq F value Pr(>F)
continent        3 5.638e+06 1879273   0.838  0.473
Residuals     1882 4.220e+09 2242143
```

E. Statistical Inference.

Step 1:
μ_{AS} = the population mean for the duration of the trending videos from Asia.
μ_{AF} = the population mean for the duration of the trending videos from Africa.
μ_{AM} = the population mean for the duration of the trending videos from America.
μ_{EU} = the population mean for the duration of the trending videos from Europe.

Step 2:
H_0 : $\mu_{AS} = \mu_{AF} = \mu_{AM} = \mu_{EU}$
H_a : at least two μ_i's are different

Step 3:
ANOVA table shown in Part D
F(ts) = 0.838
Dfa = 3 dfe = 1882
P = 0.473

Step 4:
We fail to reject H_0 as 0.473 > 0.01 and so we can say:
This data does not give strong support to reject the claim that there is a difference between the means of video durations of 4 large regions.

F. Conclusion.

The question being answered here is “Is there a difference in the mean duration of trending videos published in 2020 in the four major areas?” The results from the test proved that there wasn’t enough information to reject the claim that there is a difference between the means of the video durations of the 4 large regions. This is not exactly what we would expect. The reason we are interested in video durations is due to the fact that it can be reasonably concluded that videos with a relatively shorter duration of about 11.7 minutes have a higher retention rate. This means the audience receives the most information possible in that small window of opportunity which lets them get the most out of the videos. It is reasonable to assume that locations with large variation in content being produced would have similar means in video durations. The content can definitely change from one region to another based on culture, resources and viewership in that specific area. In the future a better approach to understanding the audience based on location would be to break down the information from each location based on the content that is popular in that region and the average video durations based on content category.

F. Inference 4 (COVID data set).

a. Code

a. Relevant code is listed in the Appendix under Code for Part F.

b. Statistical Procedure

To answer the question of if there is a linear association between the number of deaths per capita due to COVID from July 10, 2020, and 2017 state GDP per capita, we should use the procedure of Linear Regression. This is because Linear Regression is well suited to examining if there exists an association between two variables and, if it is linear, what that association is.

The following Null and Alternative Hypotheses are what we will use in our

H_0 - There is no association between the 2017 GDP per Capita of a state and the state Deaths per Capita by July 10th, 2020, due to COVID-19.
H_a - There is an association between the 2017 GDP per Capita of a state and the state Deaths per Capita by July 10th, 2020, due to COVID-19.

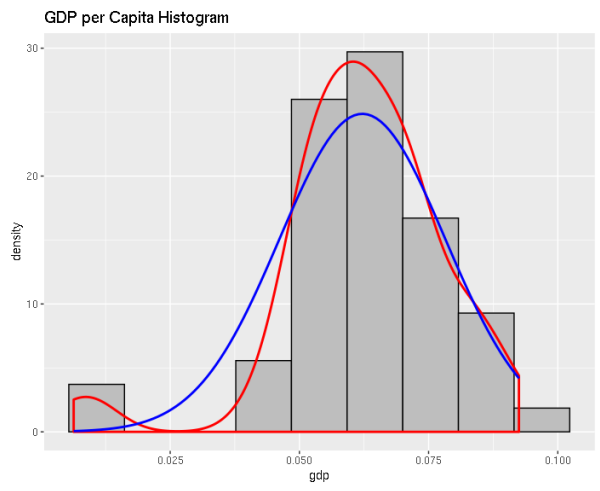
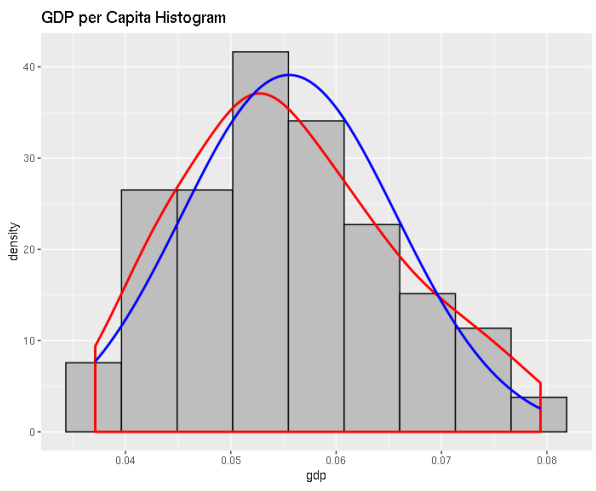
c. Assumptions for Inference

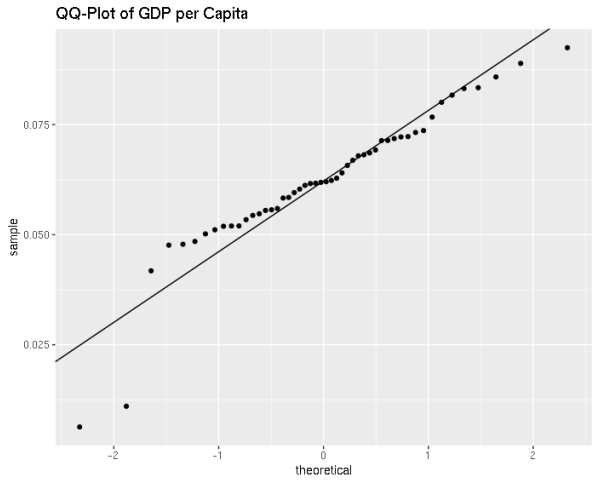
Assumption I: Data is obtained from a SRS.

- We assume this to be true for the COVID-19 provided data set because it is confirmed to be a valid random sample. The additional data for the 2017 state GDP drawn from the U.S. Bureau of Economics (BEA) is assumed to be sampled appropriately.

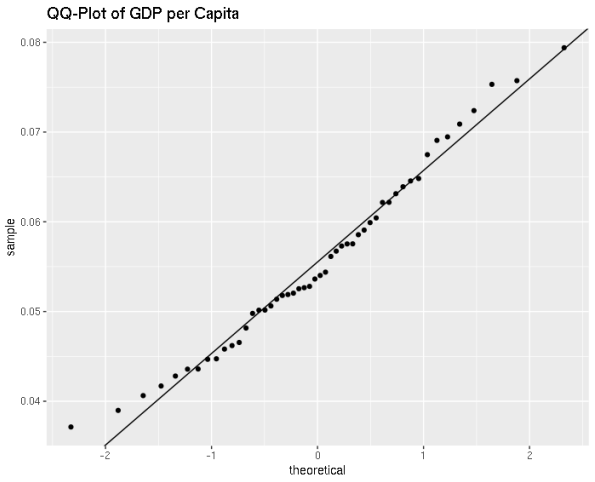
Reasoning for using BEA Data Set:

- As seen in the plots below, there are two distinct outliers in the QQ plot made from the provided dataset, this is because the GDP for Vermont and Utah in 2017 had been entered as 4068 and 36089 million dollars. When these are plotted over their population they are very far from the distribution and suspiciously low considering the population of each state. These two entries for GDP are presumably errors in the data set and are corrected by using the BEA estimated GDP instead.

GDP per Capita Plots using state.2017.gdp	GDP per Capita Plots using BEA Data
	

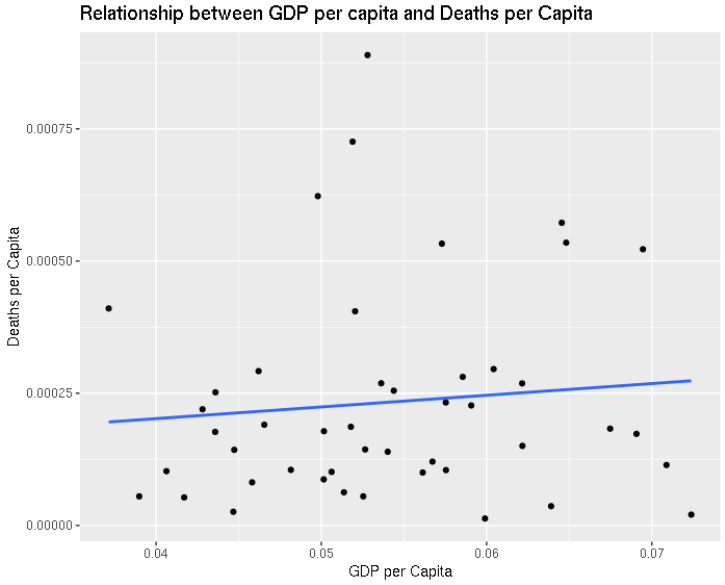


name	deathsPerCapita	gdpPerCapita
Vermont	8.708133e-05	0.006325837
Utah	6.266017e-05	0.011030940
Mississippi	4.102957e-04	0.041799844
Arkansas	1.026059e-04	0.047629705
WestVirginia	5.296268e-05	0.047833102
Alabama	2.197330e-04	0.048475612
Idaho	5.491799e-05	0.050187428
NewMexico	2.517093e-04	0.051116824
Oklahoma	1.050677e-04	0.051917321
SouthCarolina	1.768122e-04	0.051984546
Kentucky	1.429258e-04	0.052001671



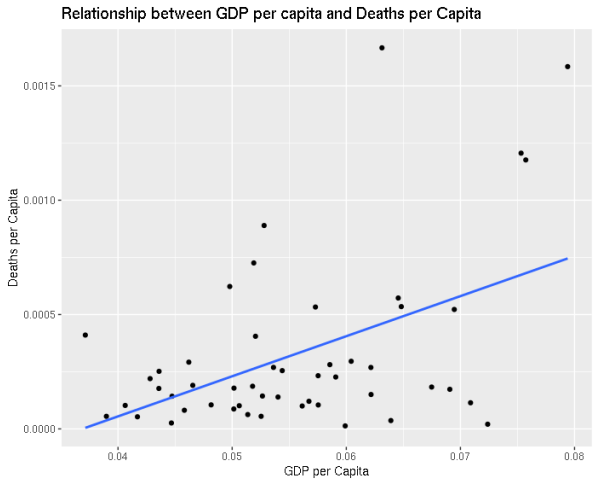
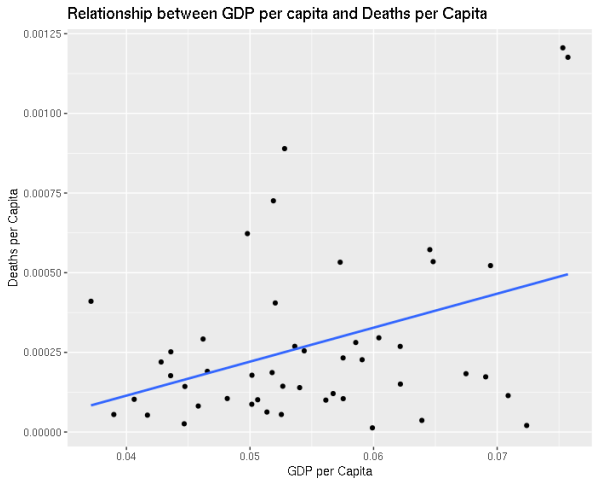
name	deathsPerCapita	gdpPerCapita
Vermont	8.708133e-05	0.05014485
Missouri	1.782316e-04	0.05015863
Tennessee	1.012901e-04	0.05062163
Utah	6.266017e-05	0.05136850
Nevada	1.864966e-04	0.05178905
Louisiana	7.256712e-04	0.05189279
Indiana	4.049795e-04	0.05204459
Oregon	5.498842e-05	0.05253730
NorthCarolina	1.436866e-04	0.05265357
RhodeIsland	8.893919e-04	0.05279944
Georgia	2.688597e-04	0.05362294

Assumption II: The relationship between X and Y is linear



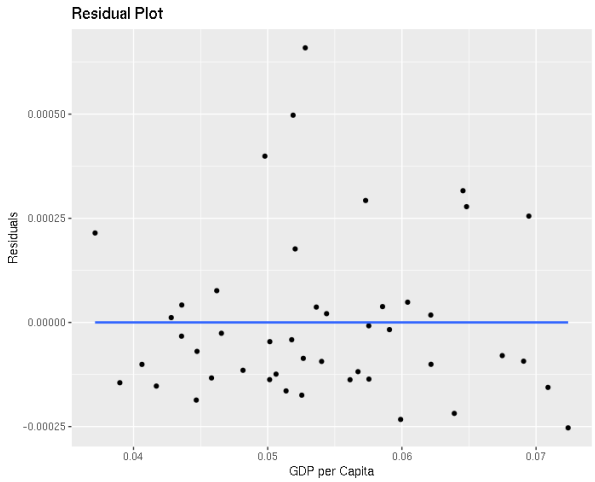
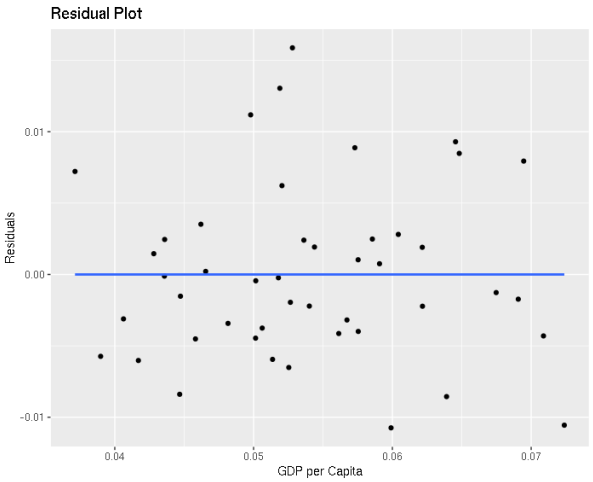
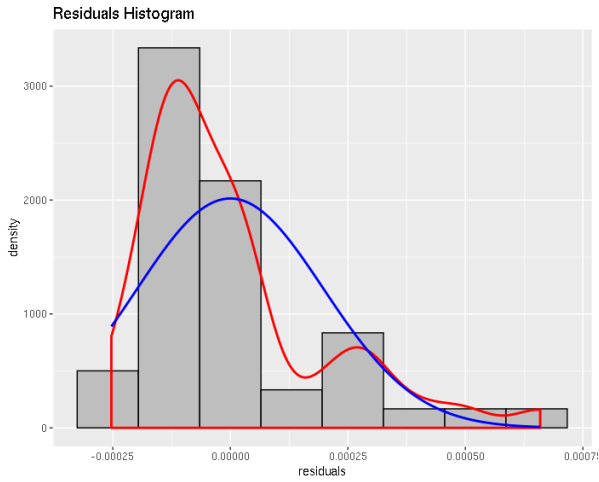
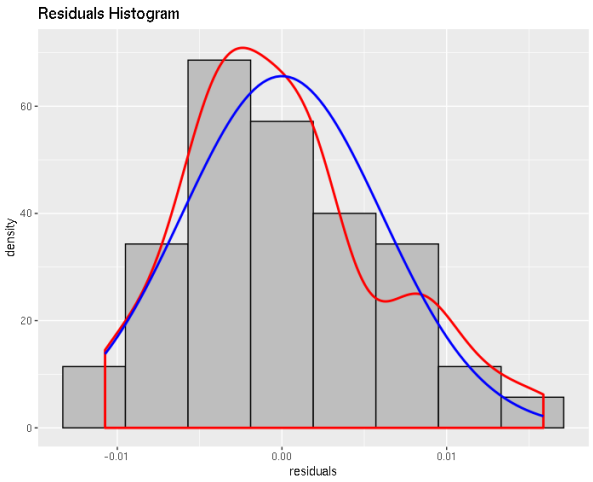
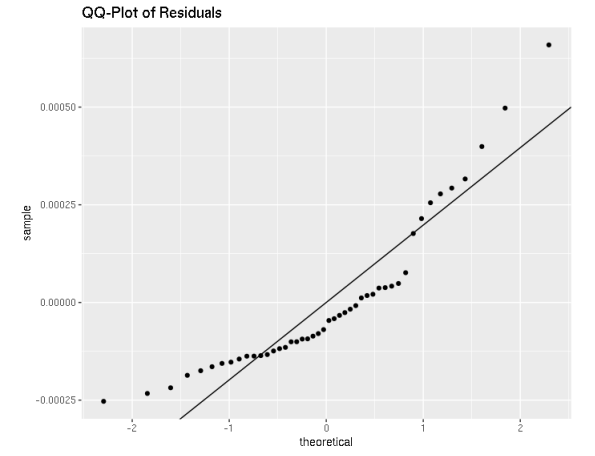
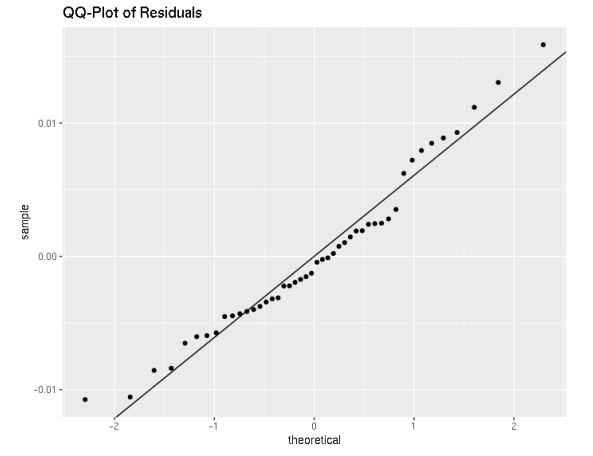
- The linear regression plot shown above appears to be linear with no other apparent pattern, however, the data is not very close to the trend line, indicating there is likely a weak correlation. There appear to be no outliers in the data set after omitting the data points New Jersey, New York, Massachusetts, and Connecticut.

The plot containing all four points can be seen below as the group of four are clearly far from the rest of the data as outliers and greatly influence the trendline.

Linear Regression with NY, NJ, MA, & CT	**Linear Regression with MA, & CT**
	

Assumption III: Residuals have a normal distribution.

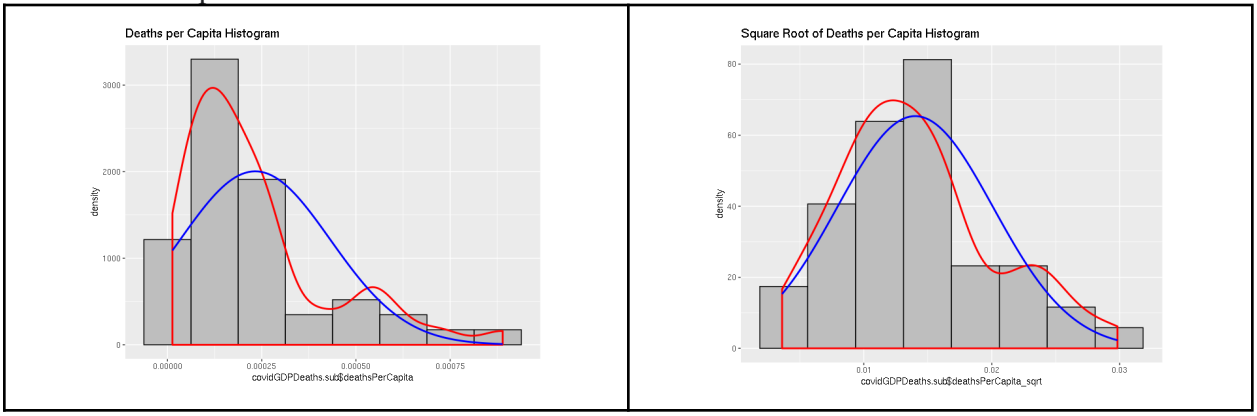
Residuals Plot

Plots Pre Sqrt Transformation	Plots Post Sqrt Transformation
	
	
	

The above histogram and QQ plot of the residuals pre sqrt transformation show non-normal behaviour.

- The histogram's distribution is asymmetric and right skewed and the normal density curve and geometric density curves have disimillar shapes. Furthermore, the QQ plot shows a concave-up pattern, indicating a right skew, and does not follow the trend line closely.

To account for the Residual abnormality a square root transformation of the deaths per capita data set was required.



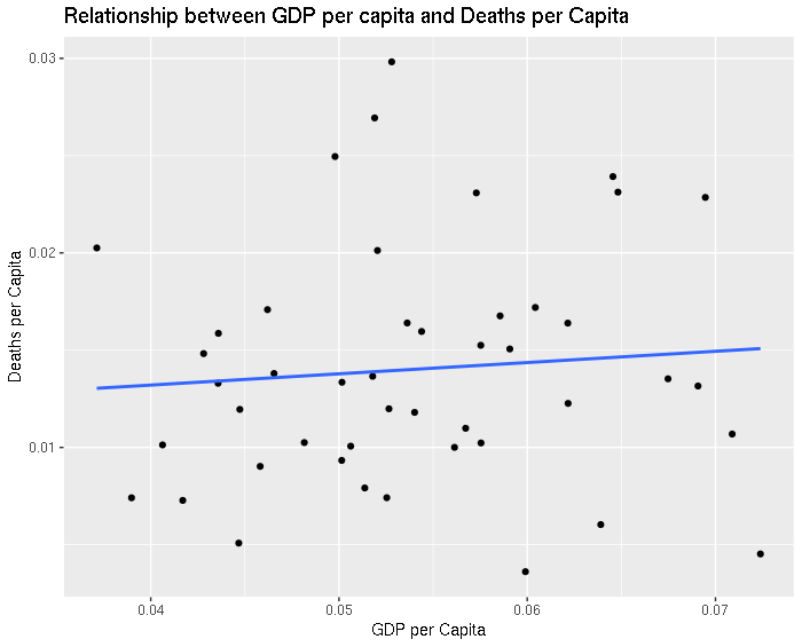
The above histogram and QQ plot of the residuals post sqrt transformation show normal behaviour.

- The histogram's distribution is symmetric and the normal density curve and geometric density curves have simillar shapes. Furthermore, the QQ plot shows a linear pattern and follows the trend line closely.

<u>Assumption IV: Standard deviation of the residuals is constant</u>

- As seen in the Post-transformation Residual Plot of Assumption II, the variation between the points seems constant. The points are all collected near each other and grouped generally close to the trend line.

d. Appropriateness Check Graphs for Statistical Inference



Exploring the data:

- Above is the scatterplot of the GDP per Capita versus the Square Root of the Deaths per Capita. In blue is the linear regression line of best fit.
- One notable comparison we can make from this graph is that the trend line indicates a slight positive association between GDP per Capita and the Square root of Deaths per Capita. Possibly indicating that the states that make more money per person had a higher
- A common line of thinking is that more recommendations leads to more views, but this graph does not mirror this pattern.

e. Statistical Inference

```
> cor(covidGDPDeaths.sub$deathsPerCapita_sqrt, covidGDPDeaths.sub$gdpPerCapita)
[1] 0.08437822
```

- Correlation Coefficient: 0.084. The correlation question, although non-zero, is quite small, indicating a Weak association. However, it isn't appropriate to completely disregard the association even though it is small, this small association purely indicates that the points in our Linear Regression won't be extremely close to the trend line.

```
> confint(covidGDPDeaths.sub.lm, level = 0.99)
                    0.5 %     99.5 %
(Intercept)  -0.004284208 0.02606305
gdpPerCapita -0.219607098 0.33540166
```

- Slope of the Fit Line -

We are 99% confident that the population slope of GDP per Capita versus Square Root of Deaths per Capita is covered by the interval:

(-0.0043, 0.0261)

- Intercept of the Fit Line -

We are 99% confident that the population y-intercept of GDP per Capita versus Square Root of Deaths per Capita is covered by the interval:

(-0.220, 0.335)

- Examples of Population Mean View Estimations based off of Recommendation Count -

```
> pointEstimate <- data.frame(gdpPerCapita = 0.045)
> predict(covidGDPDeaths.sub.lm, pointEstimate, interval = "confidence", level = 0.99)
       fit        lwr        upr
1 0.0134948 0.01000918 0.01698042
> pointEstimate <- data.frame(gdpPerCapita = 0.055)
> predict(covidGDPDeaths.sub.lm, pointEstimate, interval = "confidence", level = 0.99)
        fit        lwr        upr
1 0.01407377 0.01161614 0.01653141
> pointEstimate <- data.frame(gdpPerCapita = 0.065)
> predict(covidGDPDeaths.sub.lm, pointEstimate, interval = "confidence", level = 0.99)
        fit        lwr        upr
1 0.01465274 0.01073712 0.01856837
```

- We are 99% confident that the population mean Square Root of Deaths per Capita when the GDP per Capita equals 0.045 is covered by the interval:

  (0.0100, 0.0170)

- We are 99% confident that the population mean Square Root of Deaths per Capita when the GDP per Capita equals 0.055 is covered by the interval:

  (0.0116, 0.0165)

- We are 99% confident that the population mean Square Root of Deaths per Capita when the GDP per Capita equals 0.065 is covered by the interval:

  (0.0107, 0.0186)

The three point estimates at vastly different points in the linear regression yielded very similar intervals, indicating that the GPA per capita of a state does not have a significant impact on the square root of deaths per Capita due to COVID in that state.

f. Conclusion

Now that we have performed a sufficient statistical analysis it is appropriate to answer the question of "Is there a linear association between the number of deaths per capita due to COVID from July 10, 2020, and 2017 state GDP per capita? If so, what is it?" We can no longer compare the GDP per capita of a state directly with the deaths per capita because we had to transform deaths so it would fit our assumptions. That being said, the correlation between the two is small

enough that it wouldn't be correct to claim there is a linear association between them. Should we attempt to fit a line to the data anyways, the slope would be insignificantly small. We are 99% sure that if the GDP per Capita of a state increased by one the square root of the deaths per capita would increase by a value between -0.0043 and 0.0261, which tells us that there either could be and increase or decrease in the square root of deaths per capita, an insignificant result. This was a surprising outcome to me as one would expect there to be a negative associatio between GDP per Capita and deaths per capita as states with more GDP per capita make more money. Those states with more money should have a greater ability to treat their residents as they would have more resources to allocate to the issue. However, our data says this is not the case and that they have little to do with eachother. This drives the curiosity of where the extra wealth these states have was allocated to during the pandemic, if not towards their hospitals. One potential reason for this could be economic funding, as every state during the pandemic was struggling to keep it's local economies running, which is a lot of spending that would be diverted away from life-saving resources like ventilators.

G. (30 points) Conclusion.

Our first three questions tried mainly to understand the correlation between the mean duration of videos and their viewership or the locations from which they were released or even the mean duration worldwide. The group was focused on understanding how the duration of a video changed and what it affected. Based on the region the video was released we expected there to be a difference between the mean duration. We soon realized that there are many factors that affected the mean duration of videos and location of release did not give us enough evidence to support that there was a strong relationship between the two. We also chose to research if our data supported the claim that most videos that have short view times correlate to higher view. We concluded that if shorter videos resulted in higher view then it was reasonable to assume that a larger number of the videos online did not have long video durations. To test this we checked to see if the mean of all videos were closer to 11.7 minutes and found that to not be true. We found evidence to support that the average mean would be higher and this could be affected by many factors. While short videos might get more views, longer videos such as movies and gaming streams are many and largely available. Lastly we also tried to understand if there was an association between the daytime views of 1080p videos and the number of recommendations of 1080p videos. Our hypothesis test concluded that there was little to no correlation between the two which prompted us to question why certain videos get more exposure than others despite being recommended less. We considered that all the unaccounted factors such as subscriber counts, notifications, algorithmic differences and content significance are things that can be connected to why these differences arise. For the COVID data set we tried to determine if there was “a linear association between the number of deaths per capita due to COVID from July 10, 2020, and 2017 state GDP per capita? If so, what is it?” and could not find strong evidence of a

linear association between the two. Our process looked for an association a state's GDP per person and the square root of the deaths per person. But we could not find a strong correlation as we were 99% certain that, should we increase the GDP per person by 1 million dollars, the square root of the deaths per capita could increase by 0.0261 or decrease by 0.0043. When we square that number again, reversing the adaptations we did to the data, this increase/decrease would only become smaller, meaning the true effect on deaths per capita is extremely insignificant. Ultimately the deaths per capita could either increase or decrease, meaning we cannot say whether increasing GDP per person would improve the overall deaths a state suffered during COVID. This tells us that state GDP really didn't effect how well they could combat the pandemic as, regardless of GDP, deaths would liekly still occur at the same scale.

H. Appendix

1. Code for Part C:

```
library(ggplot2)
videos_cleaned$duration_log <- log(videos_cleaned$duration) #this is used to
normalize the histogram
xbar <- mean(videos_cleaned$duration_log) #this code is used to make the
histogram
s <- sd(videos_cleaned$duration_log)
ggplot(videos_cleaned, aes(duration_log))+
  geom_histogram(aes(y=..density..), bins=18, 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 Duration Post Log Transformation")

ggplot(videos_cleaned, aes(x ="", y =duration_log))+ #this is to make the
boxplot
  stat_boxplot(geom="errorbar")+
  geom_boxplot()+
  ggtitle("Boxplot of Duration Post Log Transformation")+
  stat_summary(fun.y=mean, color="black", geom="point", size =3)

ggplot(videos_cleaned, aes(sample = duration_log)) + #this is to make the
QQPlot
  stat_qq() +
  geom_abline(slope = s, intercept = xbar) +
  ggtitle("QQ Plot of Duration Post Log Transformation")

xbar
s
nrow(videos_cleaned)
s/sqrt(nrow(videos_cleaned)) #SE

qt(0.01, 1894, lower.tail=FALSE)

t.test(videos_cleaned$duration_log, conf.level = .99, mu = log(702),
       alternative = 'two.sided') #this was used for the one-sample t-test
df <- length(videos_cleaned$duration_log)-1
qt(0.01/2, df, lower.tail = FALSE) #code to find the critical value

The code below is used to make the plots BEFORE the transformation:

 xbar <- mean(videos$duration) #this code is used to make the actual histogram
 s <- sd(videos$duration)
 ggplot(videos, aes(duration))+
     geom_histogram(aes(y=..density..), bins=18, 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 Duration before transformation")

 xbar <- mean(videos$duration) #this code is used to make the histogram
 s <- sd(videos$duration)
 ggplot(videos_cleaned, aes(duration))+
     geom_histogram(aes(y=..density..), bins=18, 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 Duration before transformation")

 ggplot(videos_cleaned, aes(x ="", y =duration))+ #this is to make the boxplot
     stat_boxplot(geom="errorbar")+
     geom_boxplot()+
     ggtitle("Boxplot of Duration before transformation")+
```

```
    stat_summary(fun.y=mean, color="black", geom="point", size =3)

 ggplot(videos_cleaned, aes(sample = duration)) + #this is to make the QQPlot
    stat_qq() +
    geom_abline(slope = s, intercept = xbar) +
    ggtitle("QQ Plot of Duration before transformation")
```

2. Code for Part D:

```
videos.sub <- subset(videos_cleaned, quality == "1080p",
                     select = c("recommendations", "day_views"))

recommendations <- videos.sub$recommendations
day_views <- videos.sub$day_views

ggplot(videos.sub, aes(recommendations)) +
  geom_histogram(aes(y =..density..),
                 bins=sqrt(nrow(videos.sub))+2, fill="grey", col="black") +
  geom_density(col="red", lwd=1) +
  stat_function(fun=dnorm, args = list(mean = mean(videos.sub$recommendations),
                 sd = sd(videos.sub$recommendations)), col = "blue", lwd = 1) +
  ggtitle("Recommendations Histogram")


ggplot(videos.sub, aes(day_views)) +
  geom_histogram(aes(y =..density..),
                 bins=sqrt(nrow(videos.sub))+2, fill="grey", col="black") +
  geom_density(col="red", lwd=1) +
  stat_function(fun=dnorm, args = list(mean = mean(videos.sub$day_views),
                                       sd = sd(videos.sub$day_views)), col =
"blue", lwd = 1) +
  ggtitle("Day_views Histogram")


#
# a) Scatterplot of the data
#
ggplot(videos.sub, aes(x=recommendations, y=day_views))+
  geom_point() +
  geom_smooth(method = lm, se = FALSE) +
  ggtitle("Relationship between recommendations and day_views") +
  xlab("Recommendations") +
  ylab("day_views")


#
# b) Residual Plots
#
videos.sub.lm <- lm(day_views ~ recommendations, data = videos.sub)
summary(videos.sub.lm)

# Residuals Histogram
videos.sub$resids = videos.sub.lm$residuals
ggplot(videos.sub, aes(resids)) +
  geom_histogram(aes(y =..density..),
                 bins=sqrt(nrow(videos.sub)) + 2,
                 fill="darkolivegreen1", col="darkolivegreen4") +
  geom_density(col="red", lwd=1) +
  stat_function(fun=dnorm, args = list(mean = mean(videos.sub$resids),
                 sd = sd(videos.sub$resids)), col = "blue", lwd = 1) +
  ggtitle("Residuals Histogram")

# Residuals QQ Plot
```

```
ggplot(data = videos.sub, aes(sample= videos.sub$resids)) +
  stat_qq() +
  geom_abline(data= videos.sub, slope = sd(videos.sub$resids), intercept =
mean(videos.sub$resids)) +
  ggtitle("QQ-Plot of Residuals")

# Residuals Plot
ggplot(data = videos.sub, aes(x=recommendations, y=resids)) +
  geom_point() +
  geom_smooth(method = lm, se = FALSE) +
  ggtitle("Residual Plot") +
  xlab("Recommendations") +
  ylab("Residuals")

# Correlation
cor(videos.sub$recommendations, videos.sub$day_views)


#
# c) Generate the 2-sided Confidence Interval (CI) for the parameters
#
confint(videos.sub.lm, level = 0.99)


#
# d) Population mean confidence at several points based off recommendation
count.
#
newdata <- data.frame(recommendations = 1250)
predict(videos.sub.lm, newdata, interval = "confidence", level = 0.99)

newdata <- data.frame(recommendations = 1750)
predict(videos.sub.lm, newdata, interval = "confidence", level = 0.99)

newdata <- data.frame(recommendations = 2250)
predict(videos.sub.lm, newdata, interval = "confidence", level = 0.99)
```

3. Code for Part E:

```
library(ggplot2)
videos <- videos_cleaned

xbar <- tapply(videos$duration, videos$reg_main, mean)
s <- tapply(videos$duration, videos$reg_main, sd)

videos.sub <- subset(videos, reg_main != "Oceania",
                     select = c("reg_main", "duration"))
videos.sub$reg_main <- factor(videos.sub$reg_main)

videos.sub$normal.density <- apply(videos.sub, 1, function(x){
  dnorm(as.numeric(x["duration"]),
        xbar[x["reg_main"]], s[x["reg_main"]])})

binlen <- as.numeric(max(tapply(videos.sub$duration,
                                videos.sub$reg_main,length)))
ggplot(videos.sub, aes(x = duration)) +
  geom_histogram(aes(y = ..density..), bins = sqrt(binlen) + 2,
                 fill = "grey", col = "black") +
  facet_grid(reg_main ~ .) +
  geom_density(col = "red", lwd = 1) +
  geom_line(aes(y = normal.density), col = "blue", lwd = 1) +
  ggtitle("Histograms of duration by main region")
```

```
Logx <- log(videos$duration)
ggplot(videos, aes(Logx)) +
  geom_histogram(aes(y = ..density..),
                 bins = sqrt(nrow(videos))+2,
                 fill = "grey", col = "black") +
  geom_density(col = "red", lwd = 1) +
  stat_function(fun = dnorm, args = list(mean = mean(Logx), sd = sd(Logx)),
                col="blue", lwd = 1) +
  ggtitle("The log-transformed data.") +
  xlab("Logx") +
  ylab("Proportion")

videos.sub$intercept <- apply(videos.sub, 1, function(x){xbar[x["reg_main"]]})
videos.sub$slope <- apply(videos.sub, 1, function(x){s[x["reg_main"]]})

ggplot(videos.sub, aes(sample = duration)) +
  stat_qq() +
  facet_grid(reg_main ~ .) +
  geom_abline(data = videos.sub, aes(intercept = intercept, slope = slope)) +
  ggtitle("QQ Plots of duration by main region")


fit <- aov(duration ~ reg_main, data = videos.sub)
summary(fit)
```

4. Code for Part F:

```
library(ggplot2)

#clean overall dataset
covid_cleaned <- covidF21[complete.cases(covidF21),]

#import new GDP Data
newGDPData <- updatedGDPDataBEA

#remove row 9, D.C. from GDP Data
newGDPData <- newGDPData[-c(9),]

#Remove non-states and territories as they do not have the same situation as the rest
of the US
covid_cleaned <- covid_cleaned[-c(48,37,42,12,9), ]

#replace faulty 2017 GDP data set with typo-free data from U.S. BEA
covid_cleaned$state.gdp.2017 <- newGDPData$X1

#remove NY, NJ, MA, CT, outliers
covid_cleaned <- covid_cleaned[-c(32,30), ]
#MA CT
covid_cleaned <- covid_cleaned[-c(21,7), ]
#setup data subset
covidGDPDeaths.sub <- subset(covid_cleaned, select =
c("state.gdp.2017","state.death.7.10", "state.pop.2010", "name"))

#in order to remove potentially spurious variable of population divide by population
to give
#deaths per capita and gdp per capita
covidGDPDeaths.sub$deathsPerCapita = covidGDPDeaths.sub$state.death.7.10 /
covidGDPDeaths.sub$state.pop.2010
covidGDPDeaths.sub$gdpPerCapita = covidGDPDeaths.sub$state.gdp.2017 /
covidGDPDeaths.sub$state.pop.2010

#data looks complete so no need to factor
```

```
#Linear regression
ggplot(covidGDPDeaths.sub, aes(x=gdpPerCapita, y=deathsPerCapita))+
  geom_point() +
  geom_smooth(method = lm, se = FALSE) +
  ggtitle("Relationship between GDP per capita and Deaths per Capita") +
  xlab("GDP per Capita") +
  ylab("Deaths per Capita")

# Calculate linear regression and get results
covidGDPDeaths.sub.lm <- lm(deathsPerCapita ~ gdpPerCapita, data = covidGDPDeaths.sub)

#Residual Plots
covidGDPDeaths.sub$resids = covidGDPDeaths.sub.lm$residuals

#test residuals
#Residuals plot
ggplot(data = covidGDPDeaths.sub, aes(x=gdpPerCapita, y=resids))+
  geom_point() +
  geom_smooth(method = lm, se = FALSE) +
  ggtitle("Residual Plot") +
  xlab("GDP per Capita") +
  ylab("Residuals")

residuals <- covidGDPDeaths.sub$resids

# Histogram of Residuals
ggplot(data = covidGDPDeaths.sub, aes(residuals)) +
  geom_histogram(aes(y =..density..),
             bins=sqrt(nrow(covidGDPDeaths.sub))+2, fill="grey", col="black") +
  geom_density(col="red", lwd=1) +
  stat_function(fun=dnorm, args = list(mean = mean(residuals), sd = sd(residuals)),
             col = "blue", lwd = 1) +
  ggtitle("Residuals Histogram")

#QQ Plot
ggplot(data = covidGDPDeaths.sub, aes(sample= residuals)) +
  stat_qq() +
  geom_abline(data= covidGDPDeaths.sub, slope = sd(residuals), intercept =
mean(residuals)) +
  ggtitle("QQ-Plot of Residuals")

#Shapiro Test
shapiro.test(covidGDPDeaths.sub$resids)

#Normality Violeted, fix by SQRT Deaths and recalculate regression
covidGDPDeaths.sub$deathsPerCapita_sqrt <- sqrt(covidGDPDeaths.sub$deathsPerCapita)

#Linear regression
ggplot(covidGDPDeaths.sub, aes(x=gdpPerCapita, y=deathsPerCapita_sqrt))+
  geom_point() +
  geom_smooth(method = lm, se = FALSE) +
  ggtitle("Relationship between GDP per capita and Deaths per Capita") +
  xlab("GDP per Capita") +
  ylab("Deaths per Capita")

# Calculate linear regression and get results
covidGDPDeaths.sub.lm <- lm(deathsPerCapita_sqrt ~ gdpPerCapita, data =
covidGDPDeaths.sub)

#Residual Plots
covidGDPDeaths.sub$resids = covidGDPDeaths.sub.lm$residuals
```

```
#test residuals
#Residuals plot
ggplot(data = covidGDPDeaths.sub, aes(x=gdpPerCapita, y=resids))+
  geom_point() +
  geom_smooth(method = lm, se = FALSE) +
  ggtitle("Residual Plot") +
  xlab("GDP per Capita") +
  ylab("Residuals")

residuals <- covidGDPDeaths.sub$resids

# Histogram of Residuals
ggplot(data = covidGDPDeaths.sub, aes(residuals)) +
  geom_histogram(aes(y =..density..),
             bins=sqrt(nrow(covidGDPDeaths.sub))+2, fill="grey", col="black") +
  geom_density(col="red", lwd=1) +
  stat_function(fun=dnorm, args = list(mean = mean(residuals), sd = sd(residuals)),
             col = "blue", lwd = 1) +
  ggtitle("Residuals Histogram")

#QQ Plot
ggplot(data = covidGDPDeaths.sub, aes(sample= residuals)) +
  stat_qq() +
  geom_abline(data= covidGDPDeaths.sub, slope = sd(residuals), intercept =
mean(residuals)) +
  ggtitle("QQ-Plot of Residuals")

#Shapiro Test
shapiro.test(covidGDPDeaths.sub$resids)

#normality works now
# Correlation
cor(covidGDPDeaths.sub$deathsPerCapita_sqrt, covidGDPDeaths.sub$gdpPerCapita)

#results of LR
summary(covidGDPDeaths.sub.lm)

# Generate the 2-sided Confidence Interval (CI) for the parameters
confint(covidGDPDeaths.sub.lm, level = 0.99)

#Population mean confidence at several points based off GDP Per Capita.
pointEstimate <- data.frame(gdpPerCapita = 0.045)
predict(covidGDPDeaths.sub.lm, pointEstimate, interval = "confidence", level = 0.99)

pointEstimate <- data.frame(gdpPerCapita = 0.055)
predict(covidGDPDeaths.sub.lm, pointEstimate, interval = "confidence", level = 0.99)

pointEstimate <- data.frame(gdpPerCapita = 0.065)
predict(covidGDPDeaths.sub.lm, pointEstimate, interval = "confidence", level = 0.99)
```

Works Cited

Published by L. Ceci, & 23, A. (2021, August 23). *YouTube average video length by Category 2018*. Statista. Retrieved December 3, 2021, from https://www.statista.com/statistics/1026923/youtube-video-category-average-length/

U.S BEA. “GDP by State.” *GDP by State | U.S. Bureau of Economic Analysis (BEA)*, U.S Bureau of Economics Analysis, 1 Oct. 2021. Retrieved December 3, 2021, from https://www.bea.gov/data/gdp/gdp-state