Biostatistics SDS 328M – Preliminary Analysis Report

Write in complete sentences using single-spaced, 12-pt font. Include figures in text. ***3-page limit.***

Name: Karla Castello
Lab day and time: Friday 4pm

Introduction

The main goal of this study is to confirm that the popularity of music is influencing the number of videos made on TikTok made for those popular songs. This study will show if there is a relationship between the number of TikTok videos made for a song and the number of streams for that song on Spotify. Also, this study will also try to show if a song being on the Billboard Top 100 chart has any effect on the number of videos made on TikTok for that song. The population of interest for my study are popular songs worldwide. Since I will be looking at how popularity of a song influences the number of videos for that song on TikTok, my sample would be taken from the most streamed songs on Spotify from 2018 to 2019. The stakeholders for this study are music artists, music streaming services, and TikTok users. These people would be involved in my study because they are the ones influenced by popular songs and help increase its popularity on TikTok.

Research Questions

RQ1 – Is the popularity of a song on Spotify increasing the number of videos made on TikTok?

RQ2 – Does a song having a spot on the Billboard chart influence the amount of TikTok videos made for that song?

Data Collection Summary

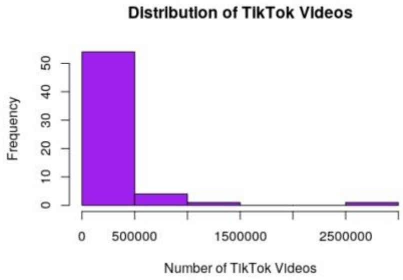
In order to collect my data, I used two data sets: A Top 100 on Spotify list from 2018 and a Top 100 on Spotify list from 2019. I generated randomly picked out 60 numbers from 200 songs using a number generator and randomly selected from both datasets. Then, I used Spotify to record the number of streams and I also recorded the number of videos made for each song directly from TikTok. Once I had all of my data collected from Spotify and TikTok, I used two Billboard Worldwide Top 100 lists: one from 2018 and one from 2019 in order to keep the dates the same as the Spotify datasets. I went through each data set and put ‘yes’ if the songs were on the list and ‘no’ if the song wasn’t on the list. The sampling units in my data are popular songs on Spotify from 2018 to 2019. My sample is not considered a representative sample because I am only using the top songs from Spotify. Even though Spotify is one of the most used music streaming services in the world, it is not representative of all the popular music from other streaming services as well. My final sample size is 60 songs. The range of my data is large but there weren’t any cases that were needed to be removed from my data; they are all important to the analysis.

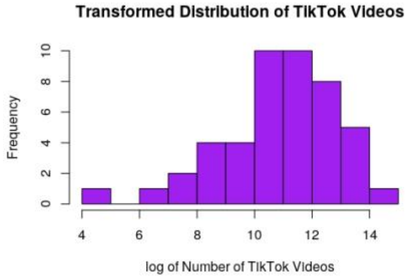
Descriptive Analysis of Response Variable *(include graph(s) of distribution here)*

Biostatistics SDS 328M – Preliminary Analysis Report

Write in complete sentences using single-spaced, 12-pt font. Include figures in text. ***3-page limit.***

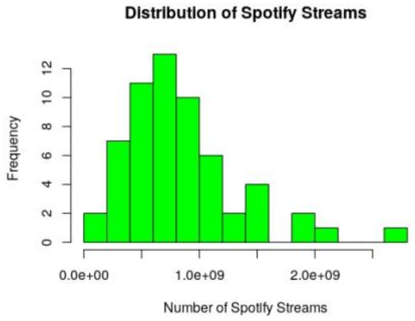
The distribution of the number of TikTok videos is a positively skewed unimodal distribution. There is one outlier in the distribution of TikTok videos. The median of this distribution is 31,250 videos and the IQR is 2,700,000. This shows most of the data is grouped to the left of the graph. If I take the outlier out of the distribution, the distribution would still be extremely positively skewed, therefore there is no reason to take it out. Since the distribution is positively skewed, I used the logarithm function on the data to make the graph relatively normal. The 1/x function nor the exponential function made the graph normal.

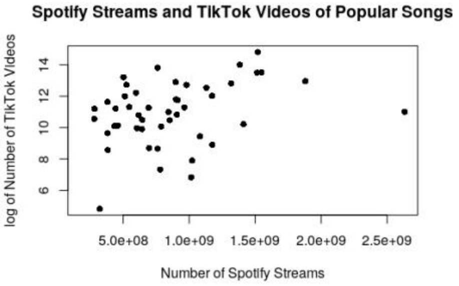




Investigation of Explanatory Variable 1 *(include univariate and bivariate graphs here)*

The distribution of Spotify streams has a unimodal shape and is positively skewed. The median of the data is 766,966,733 Spotify streams and the IQR is 2,812,997,573 Spotify streams. The range of Spotify streams is much larger than the range of TikTok videos. Based on the plot, there seems to be a relative positive relationship between the log of the number of TikTok videos and the number of Spotify streams because the points are all moving in an upward slope and are not spread out throughout the graph. I did expect this outcome since the more streams a song has, the more people are listening to it and using it in their videos. Also, their raw data both showed to be positively skewed.



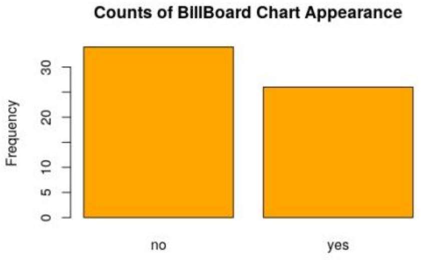


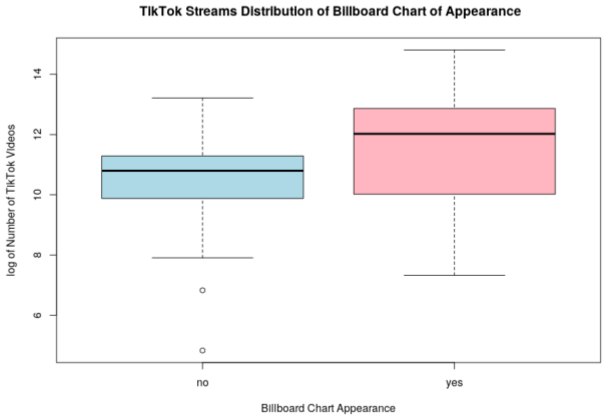
Investigation of Explanatory Variable 2 *(include univariate and bivariate graphs here)*

Biostatistics SDS 328M – Preliminary Analysis Report

Write in complete sentences using single-spaced, 12-pt font. Include figures in text. ***3-page limit.***

Based off of the frequency table, there are 26 songs that were in the Billboard Top 100 chart compared to 26 songs that were not on the chart in 2018 nor 2019. There does appear to be a relationship between the log of number TikTok videos and the appearance of songs on the Billboard chart. There is a higher number of videos to songs that did appear on the Billboard chart than songs that did not. This relationship is what I expected because the more popular songs will have the most videos made with that songs and the less popular songs with have less videos made with those songs.





R Code *(organized by variable without output or extraneous syntax)*

Response variable:

```
hist(myproject$tiktok, main = "Distribution of TikTok Videos",xlab = 'Number of TikTok Videos', breaks = 8, col = 'purple')
tiktoklog <- log(myproject$tiktok)
tiktoklog[is.infinite(tiktoklog)] <- NA
hist(tiktoklog, main = 'Transformed Distribution of TikTok Videos',xlab = 'log of Number of TikTok Videos', breaks = 10, col = 'purple')
median(myproject$tiktok,na.rm = TRUE)
fivenum(myproject$tiktok)
```

Explanatory variable 1:

```
hist(myproject$spotify, main = 'Distribution of Spotify Streams', xlab = 'Number of Spotify Streams',breaks= 10, col = 'green')
plot(myproject$spotify,tiktoklog, main = 'Spotify Streams and TikTok Videos of Popular Songs',ylab = 'log of Number of TikTok Videos', xlab = 'Number of Spotify Streams',pch = 16)
median(myproject$spotify)
fivenum(myproject$spotify)
```

Explanatory Variable 2:

```
table(myproject$billboard)
barplot(table(myproject$billboard), main = "Counts of BillBoard Chart Appearance", ylab = "Frequency", col='orange')
boxplot(tiktoklog~myproject$billboard, main = ' TikTok Streams Distribution of Billboard Chart of Appearance', xlab = "Billboard Chart Appearance", ylab = 'log of Number of TikTok Videos', col = c('light blue','light pink'))
```

Biostatistics SDS 328M – Preliminary Analysis Report

Write in complete sentences using single-spaced, 12-pt font. Include figures in text. ***3-page limit.***