Report

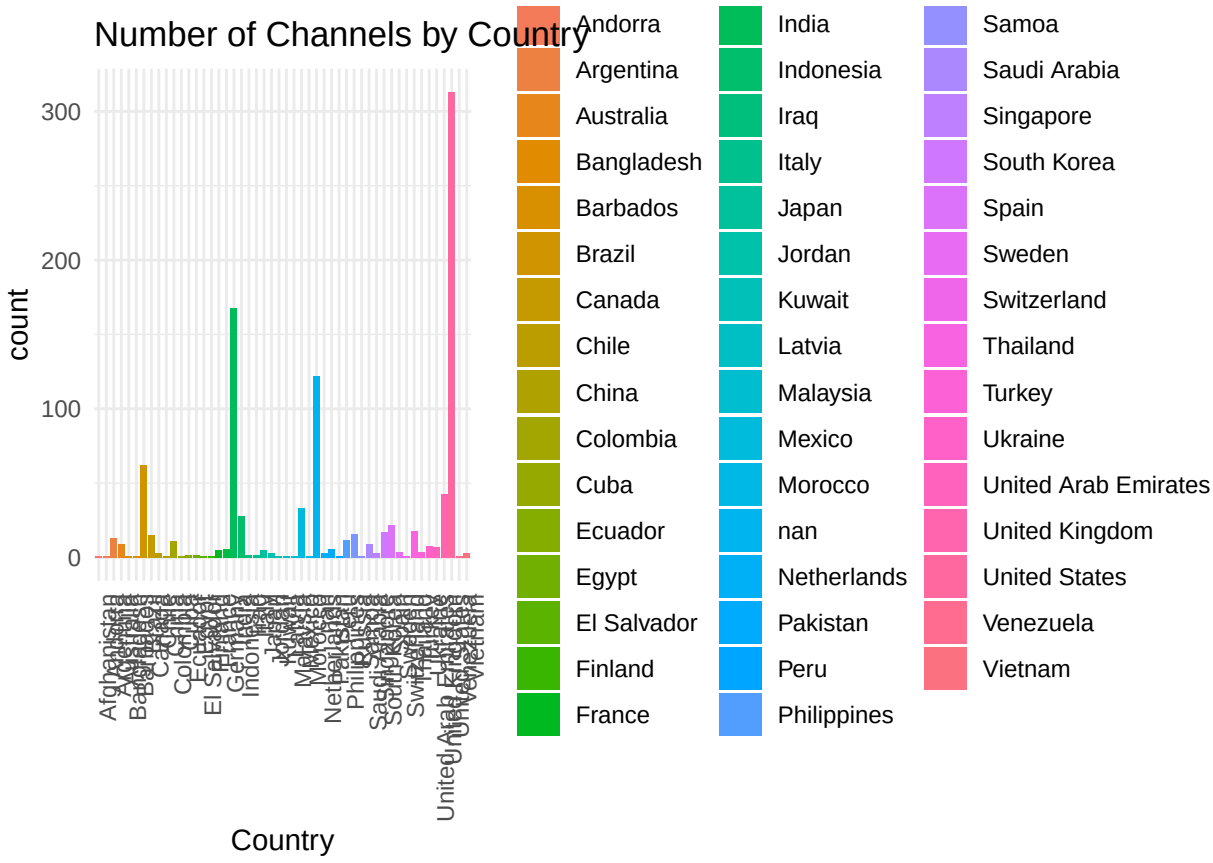
2023-12-14

This dataset examines data from 995 YouTube channels and contains values for a variety of variables, such as a channel's number of uploads or its country of origin. We can categorize most of the dataset's variables into two categories: predictors of video success (e.g. date of channel creation, education level of country of origin) or measurements of video success (e.g. total views, highest yearly earnings). The high number of variables in the first category allows us to examine highly detailed models for predicting the success of a channel, while the high number in the second category enables us to determine if our results are robust to different measures of success.

First, we will do some exploratory data analysis. Here, I will plot a histogram of the number of channels by country and by channel category. Below, we plot the number of channels by category and by country (see produced figure in repository if this one is difficult to read).

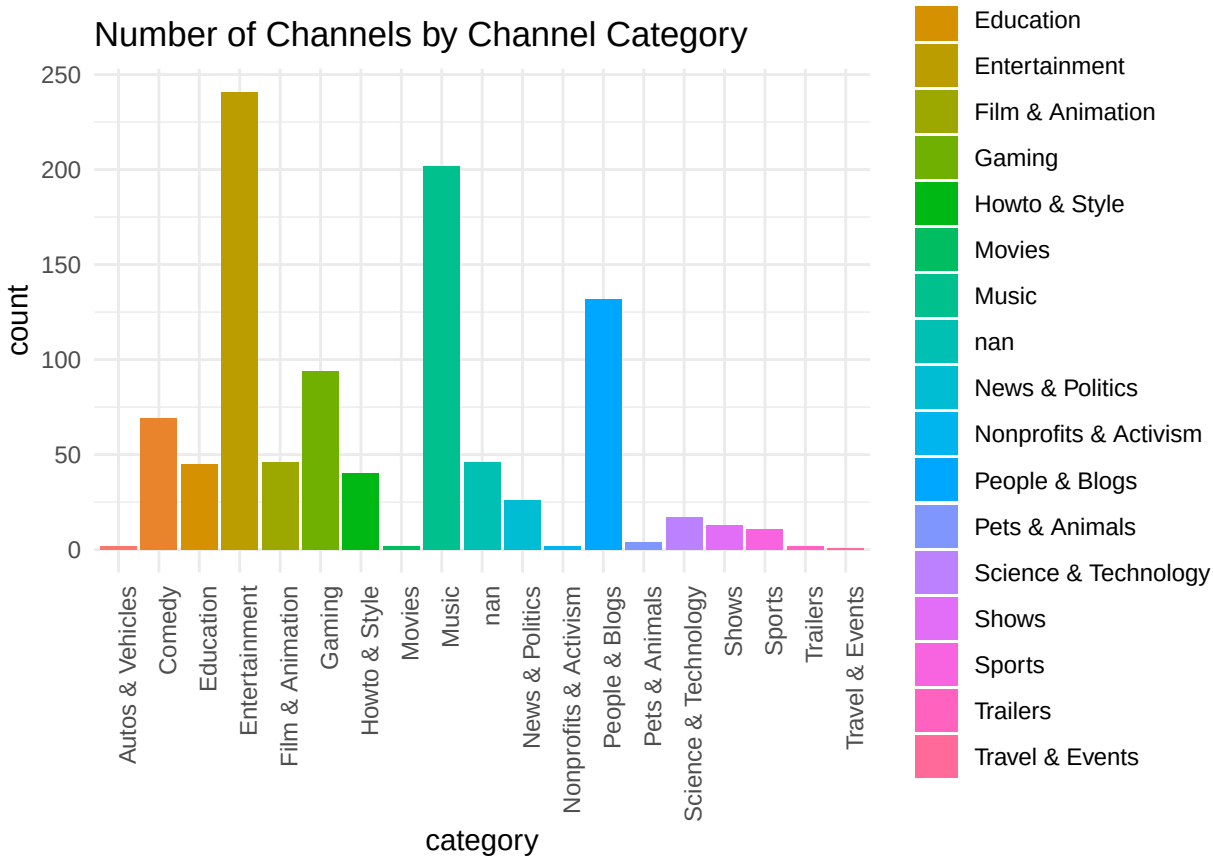
```
library(ggplot2)
youtube<- read.csv("~/Desktop/Global YouTube Statistics.csv")
ggplot(youtube, aes(x = Country, fill = Country)) +
  geom_histogram(stat = "count") +
  labs(title = "Number of Channels by Country", xlab = "Country", ylab = "Number of Channels") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

```
## Warning in geom_histogram(stat = "count"): Ignoring unknown parameters:
## `binwidth`, `bins`, and `pad`
```




```
ggplot(youtube, aes(x = category, fill = category)) +
  geom_histogram(stat = "count") +
  labs(title = "Number of Channels by Channel Category", xlab = "Country", ylab = "Number of Channels") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

```
## Warning in geom_histogram(stat = "count"): Ignoring unknown parameters:
## `binwidth`, `bins`, and `pad`
```



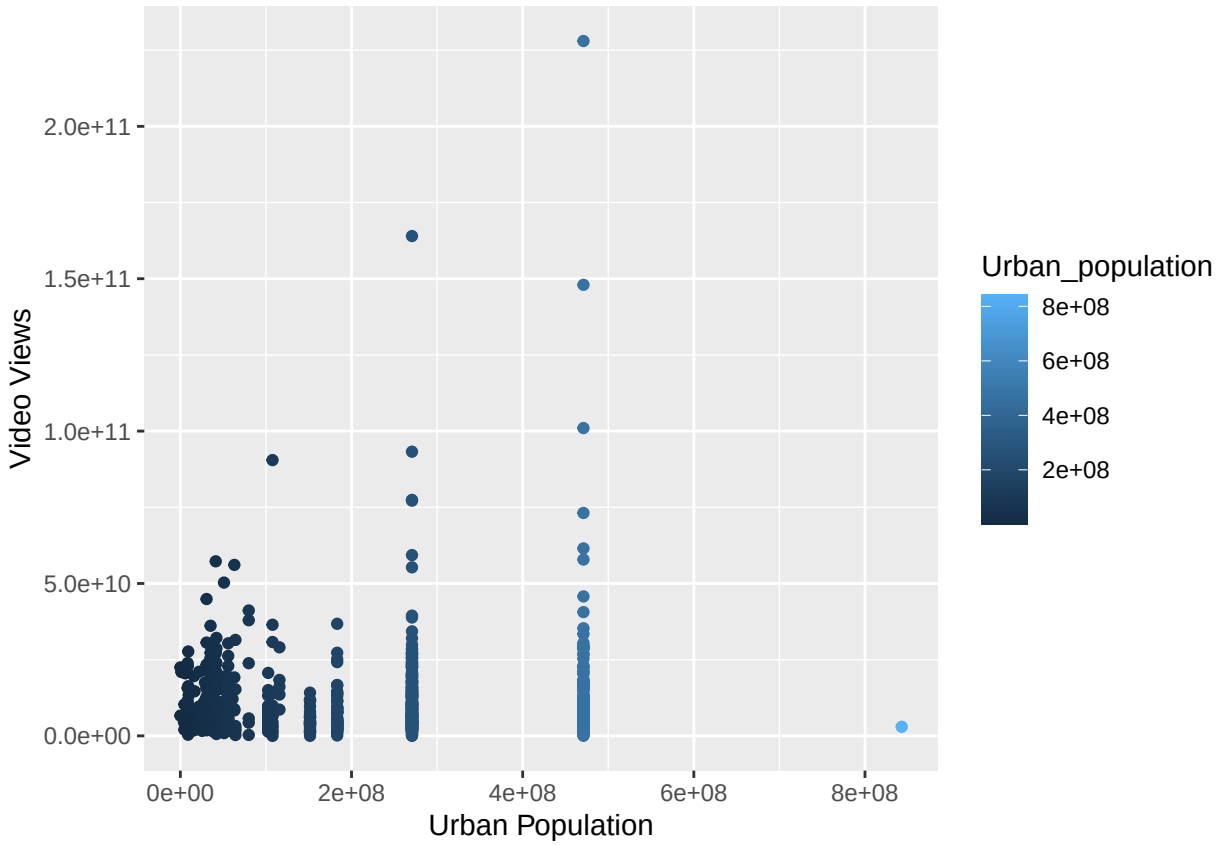
Based on the first histogram, the most popular country of origin for YouTube channels in this dataset is the United States. India was the second most common source of YouTube channels, followed by Brazil and the UK. Based on the second histogram, the most common channel categories are entertainment and music. A large number of channels do not have an associated country because they are Youtube's own channels. Although these channels are actually channels, unlike most of the other channels, they merely aggregate videos across various categories rather than creating their own videos. Aggregator channels are also missing values for variables such as "category", "Country", and "video.views". Therefore, we will restrict our analysis to channels with a focus on creation.

```
channels_only<- subset(youtube, video.views != 0)
```

Next, we will attempt to determine the impact of the urban population of a channel's home country on the number of views that channel receives. We create a scatterplot to examine the relationship between the urban population of a channel's associated country and video views.

```
ggplot(channels_only,aes(Urban_population, video.views)) +
  geom_point(aes(color = Urban_population)) +
  labs(x = "Urban Population", y = "Video Views")
```

```
## Warning: Removed 117 rows containing missing values (`geom_point()`).
```



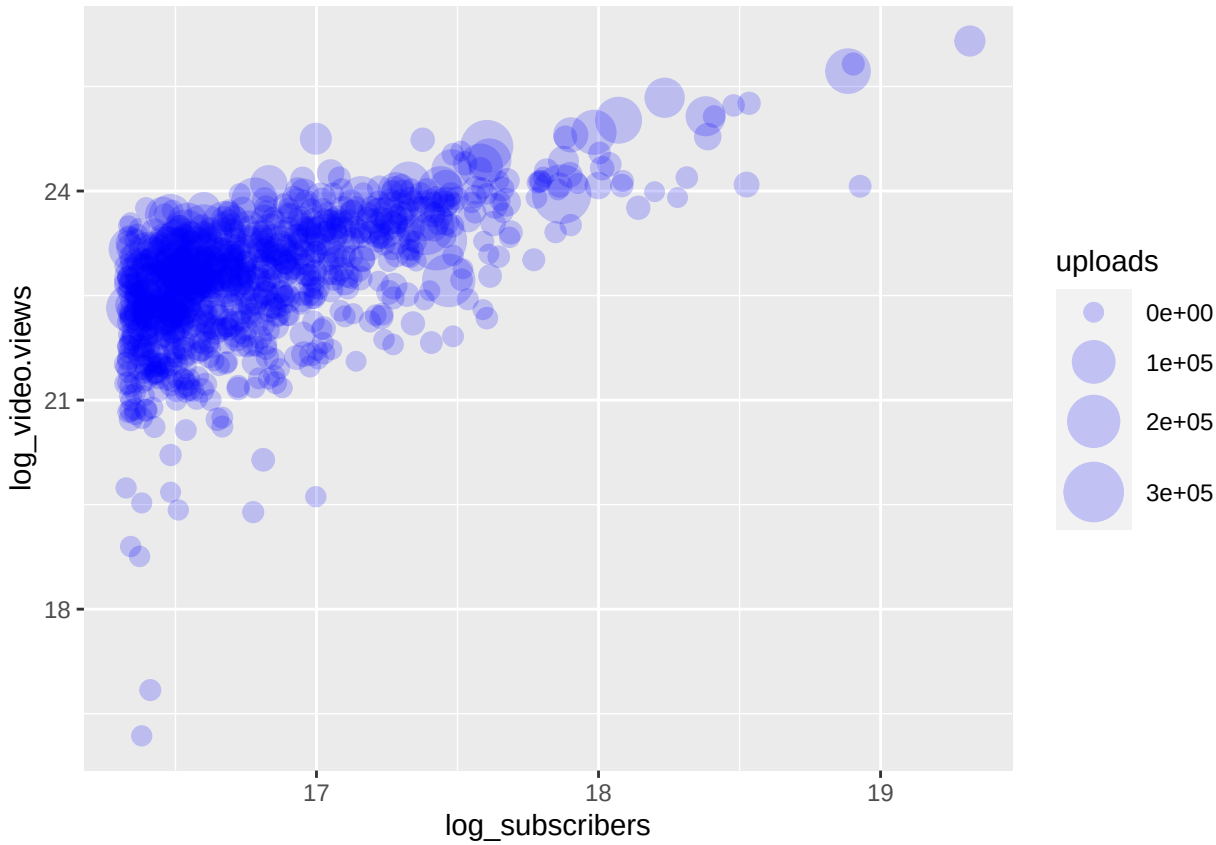

```
cor(channels_only$Urban_population, channels_only$video.views, use = "complete.obs")
```

```
## [1] 0.07723294
```

Although we do not observe a strong effect of urban population on video views, with countries showing mostly similar distributions regardless of urban population, we observe that the channels with the highest numbers of video views are almost exclusively created by channels that hail from countries with large urban populations. Thus, a country's urban population may not have any effect on the success of its channels unless those channels become extremely popular, in which case a high urban population may be modestly associated with more total video views.

Other factors which could plausibly impact the overall success of a channel are its subscriber count and its total number of uploads. To illustrate the impact of these two factors simultaneously, we will create a bubble plot. We apply to a log transformation to the variables representing total video views and total subscribers due to the exponential growth rate typically associated with both.

```
channels_only$log_video.views<- log(channels_only$video.views)
channels_only$log_subscribers<- log(channels_only$subscribers)
channels_only<- subset(channels_only, log_video.views > 15) #Remove outliers
ggplot(data = channels_only, aes(x = log_subscribers, y = log_video.views, size = uploads)) +
  geom_point(alpha = 0.2, color = "blue") +
  scale_size_continuous(range = c(3, 10))
```




```
cor(channels_only$log_subscribers, channels_only$log_video.views, use = "complete.obs")
```

```
## [1] 0.6048599
```

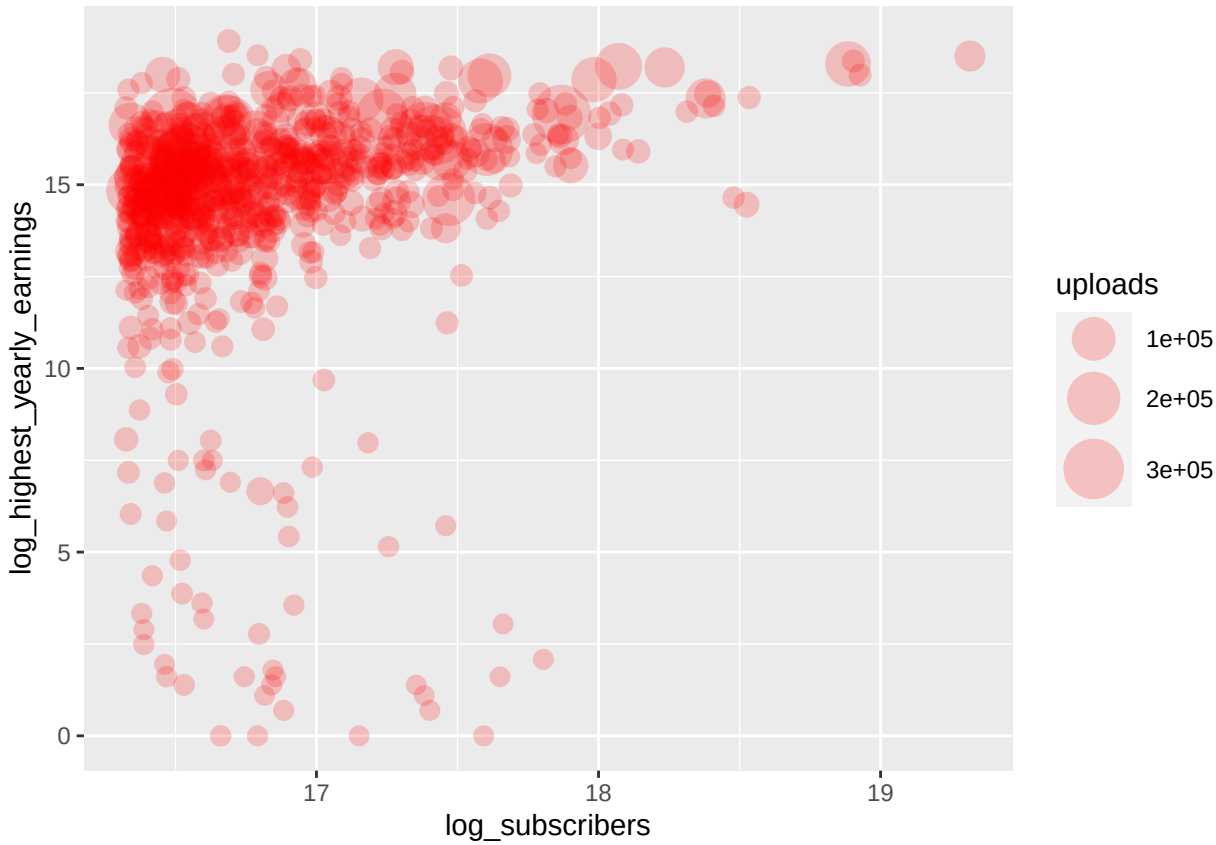
```
cor(channels_only$uploads, channels_only$log_video.views, use = "complete.obs")
```

```
## [1] 0.178092
```

Total number of subscribers and the total number of uploads seem to be positively correlated with channel success (as measured by the total number of video views), albeit to varying degrees. A channel's total number of uploads modestly predicted the total number of video views it received, possibly because a greater number of uploads gives a channel more opportunities to receive views. A channel's total number of subscribers was a substantially stronger predictor of the number of video views it received. There are two plausible (non-mutually exlusive) explanations: 1) channels which have more subscribers receive more views because more potential viewers are exposed to their content or 2) high viewership leads to a high number of subscriptions because people often subscribe to a channel after watching a video.

To ensure that our analysis is robust, we should examine the relationship of our predictors to a different measure of channel success, log transformed yearly earnings (specifically, the highest yearly earnings a channel has received).

```
nonzero<- subset(channels_only, highest_yearly_earnings >= 1) #Remove values equal to zero, as they can
nonzero$log_highest_yearly_earnings<- log(nonzero$highest_yearly_earnings)
ggplot(data = nonzero, aes(x = log_subscribers, y = log_highest_yearly_earnings, size = uploads)) +
  geom_point(alpha = 0.2, color = "red") +
  scale_size_continuous(range = c(3, 10))
```




```
cor(nonzero$uploads, nonzero$log_highest_yearly_earnings, use = "complete.obs")
```

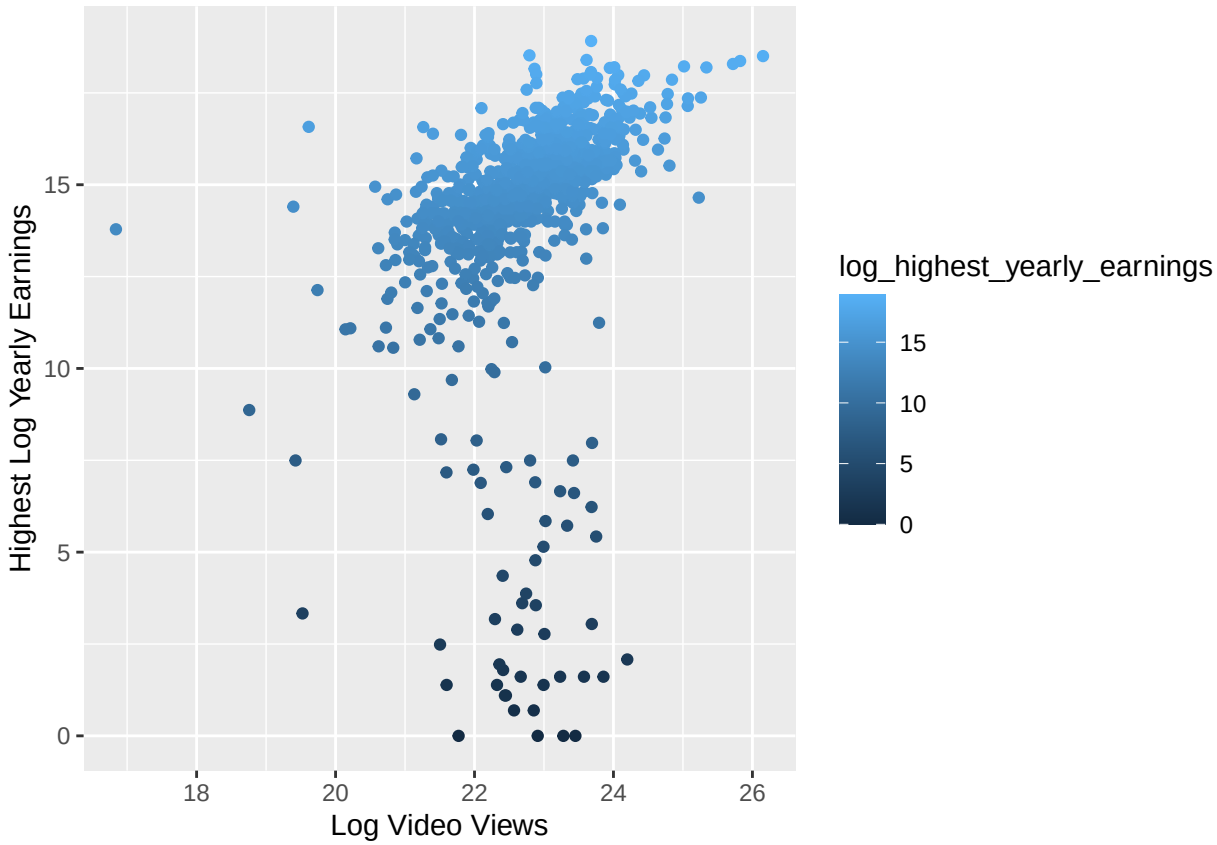
```
## [1] 0.1457022
```

```
cor(nonzero$log_subscribers, nonzero$log_highest_yearly_earnings, use = "complete.obs")
```

```
## [1] 0.1743523
```

We observe similar results when using the highest amount of log-transformed yearly earnings as our measure of channel success; therefore, our results are robust to different measures of channel success. However, the strength of the relationship between log-transformed subscriptions and log-tranformed yearly earnings is significantly lower than the strength of the one between log-transformed subscriptions and log-transformed video views. We might observe this effect because some channels may mostly contain high numbers of highly-watched videos that are unable to be monetized for various reasons (e.g. explicit content, copyright violations). To test this hypothesis, we will examine the relationship between log-transformed video views and the highest amount of log-transformed yearly earnings a channel has received.

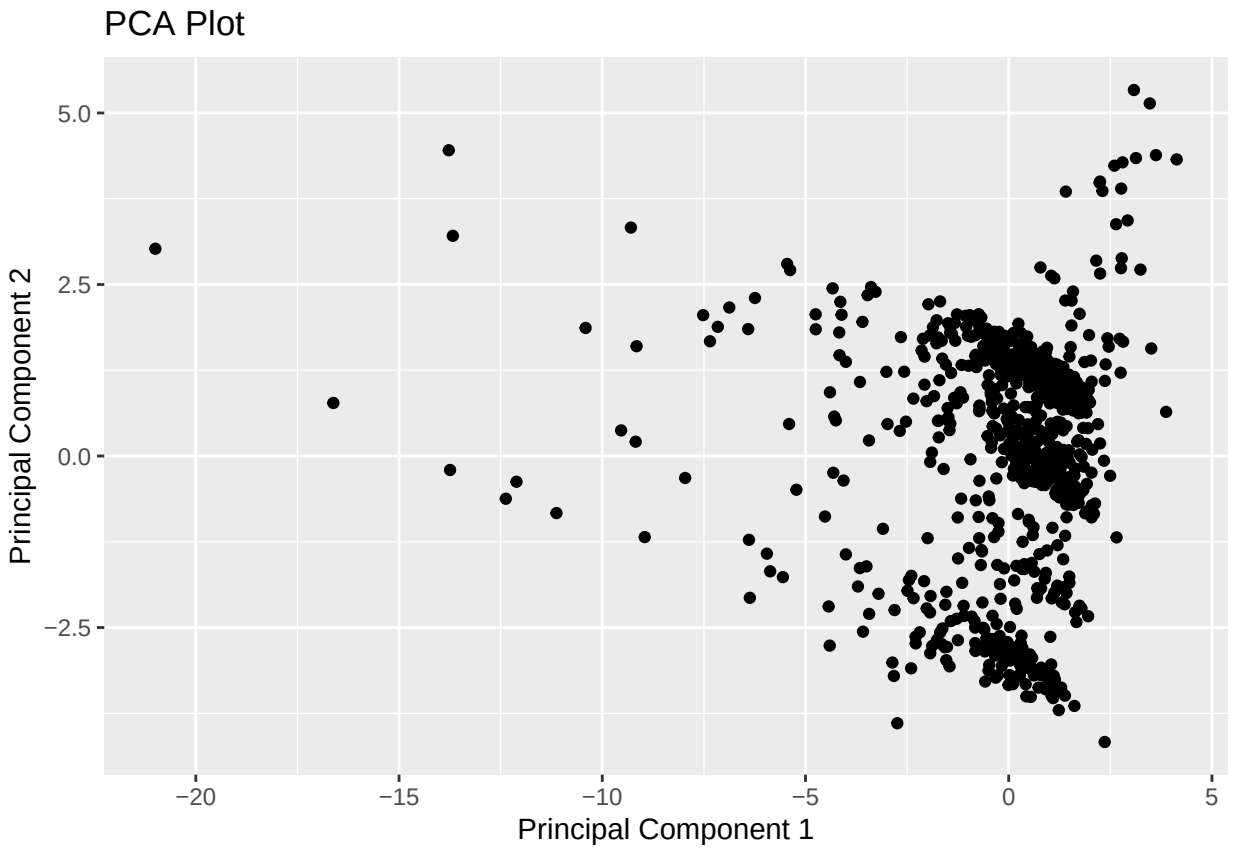
```
ggplot(nonzero, aes(log_video.views, log_highest_yearly_earnings)) +
  geom_point(aes(color = log_highest_yearly_earnings)) +
  labs(x = "Log Video Views", y = "Highest Log Yearly Earnings")
```



We observe a similar effect: a strong, positive, linear correlation between log transformed video views and log transformed yearly earnings for the majority of datapoints but a lack of correlation among channels which earn little compared to their peers. This effect provides further evidence that large-scale video demonetization creates the heterogeneous relationship between subscriptions and earnings.

Finally, because many of our variables are related to each other, we will explore the possibility of dimemsionality reduction for our dataset. We conduct principal component analysis (PCA) on all numerical variables.

```
channels_only<- subset(channels_only, select = -c(subscribers_for_last_30_days)) #Remove the "subscribe
channels_only<- channels_only[complete.cases(channels_only$country_rank), ] #Remove channels without an
youtube_numeric<- channels_only[sapply(channels_only, is.numeric)]
youtube_numeric <- na.omit(youtube_numeric)
youtube_numeric <- subset(youtube_numeric, select = -c(subscribers, video.views)) #Remove variables tha
pca_result <- prcomp(youtube_numeric, scale = TRUE)
scores <- as.data.frame(pca_result$x[, 1:2])
ggplot(scores, aes(x = PC1, y = PC2)) +
  geom_point() +
  labs(title = "PCA Plot",
       x = "Principal Component 1",
       y = "Principal Component 2")
```




```
summary(pca_result)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.345 1.7461 1.6521 1.36084 1.27346 1.05807 0.99933
## Proportion of Variance 0.275 0.1524 0.1365 0.09259 0.08109 0.05598 0.04993
## Cumulative Proportion  0.275 0.4274 0.5639 0.65650 0.73758 0.79356 0.84349
##                           PC8     PC9   PC10    PC11    PC12    PC13    PC14
## Standard deviation     0.96215 0.83195 0.6856 0.57751 0.54630 0.36921 0.35364
## Proportion of Variance 0.04629 0.03461 0.0235 0.01668 0.01492 0.00682 0.00625
## Cumulative Proportion  0.88978 0.92438 0.9479 0.96456 0.97949 0.98630 0.99255
##                           PC15    PC16    PC17    PC18     PC19     PC20
## Standard deviation     0.30253 0.19596 0.13714 0.01118 0.006868 0.001263
## Proportion of Variance 0.00458 0.00192 0.00094 0.00001 0.000000 0.000000
## Cumulative Proportion  0.99713 0.99905 0.99999 1.00000 1.000000 1.000000
```

The first two principal components explain approximately 43% of the variance in the data. We observe three major clusters of data in regions centered around (0, -3), (1.25, 0), and (1, 1.25). An increase in the value of PC1 is associated with a decrease in the value of PC2 within each of these clusters. We will examine some of the clusters below.

```
cluster1_indices <- which(pca_result$x[, 1] > -2 & pca_result$x[, 2] < -2.5)
cluster1_data <- youtube_numeric[cluster1_indices, ]
summary(cluster1_data)
```

```
##       rank          uploads       video_views_rank  country_rank
##  Min.   : 79.0   Min.   :    29   Min.   :    88   Min.   :  1.00
##  1st Qu.:450.5   1st Qu.:   477   1st Qu.:   527   1st Qu.: 75.00
```

```
##  Median :655.0   Median :  1476   Median :  1033   Median :100.00
##  Mean   :636.3   Mean   : 16732   Mean   : 11928   Mean   : 91.91
##  3rd Qu.:828.5   3rd Qu.:  5673   3rd Qu.:  2886   3rd Qu.:113.00
##  Max.   :995.0   Max.   :296272   Max.   :772571   Max.   :125.00
##  channel_type_rank video_views_for_the_last_30_days lowest_monthly_earnings
##  Min.   :  1.00    Min.   :8.378e+05                Min.   :    0
##  1st Qu.: 28.50    1st Qu.:3.086e+07                1st Qu.: 7050
##  Median : 47.00    Median :7.865e+07                Median :18200
##  Mean   : 72.29    Mean   :1.219e+08                Mean   :25153
##  3rd Qu.:125.00    3rd Qu.:1.404e+08                3rd Qu.:34800
##  Max.   :172.00    Max.   :2.292e+09                Max.   :94800
##  highest_monthly_earnings lowest_yearly_earnings highest_yearly_earnings
##  Min.   :      0          Min.   :      0        Min.   :       0
##  1st Qu.: 113150          1st Qu.:  84850        1st Qu.: 1350000
##  Median : 291600          Median : 218700        Median : 3500000
##  Mean   : 402050          Mean   : 300440        Mean   : 4828413
##  3rd Qu.: 556500          3rd Qu.: 417350        3rd Qu.: 6700000
##  Max.   :1500000          Max.   :1100000        Max.   :18200000
##   created_year   created_date  Gross.tertiary.education.enrollment....
##  Min.   :2006   Min.   : 1.0   Min.   :28.10
##  1st Qu.:2012   1st Qu.: 9.5   1st Qu.:28.10
##  Median :2015   Median :18.0   Median :28.10
##  Mean   :2014   Mean   :16.6   Mean   :28.49
##  3rd Qu.:2017   3rd Qu.:23.0   3rd Qu.:28.10
##  Max.   :2022   Max.   :31.0   Max.   :50.60
##    Population        Unemployment.rate Urban_population       Latitude
##  Min.   :1.081e+08   Min.   :2.150     Min.   : 50975903   Min.   :-0.7893
##  1st Qu.:1.366e+09   1st Qu.:5.360     1st Qu.:471031528   1st Qu.:20.5937
##  Median :1.366e+09   Median :5.360     Median :471031528   Median :20.5937
##  Mean   :1.338e+09   Mean   :5.313     Mean   :465256765   Mean   :20.2978
##  3rd Qu.:1.366e+09   3rd Qu.:5.360     3rd Qu.:471031528   3rd Qu.:20.5937
##  Max.   :1.398e+09   Max.   :5.360     Max.   :842933962   Max.   :35.8617
##    Longitude      log_video.views log_subscribers
##  Min.   : 78.96   Min.   :16.18   Min.   :16.33
##  1st Qu.: 78.96   1st Qu.:21.82   1st Qu.:16.43
##  Median : 78.96   Median :22.52   Median :16.54
##  Mean   : 80.12   Mean   :22.25   Mean   :16.61
##  3rd Qu.: 78.96   3rd Qu.:22.93   3rd Qu.:16.74
##  Max.   :121.77   Max.   :23.86   Max.   :17.48
```

```
cluster2_indices <- which(pca_result$x[, 1] > 0 & pca_result$x[, 2] < 0.75 & pca_result$x[, 2] > -0.75)
cluster2_data <- youtube_numeric[cluster2_indices, ]
summary(cluster2_data)
```

```
##       rank          uploads      video_views_rank   country_rank
##  Min.   : 53.0   Min.   :    3   Min.   :    113   Min.   :   1.00
##  1st Qu.:459.0   1st Qu.:  377   1st Qu.:    648   1st Qu.:   6.00
##  Median :672.0   Median :  825   Median :   1472   Median :  17.00
##  Mean   :642.4   Mean   : 2414   Mean   :  50706   Mean   :  90.15
##  3rd Qu.:840.0   3rd Qu.: 1888   3rd Qu.:   2897   3rd Qu.:  34.00
##  Max.   :994.0   Max.   :64496   Max.   :4039216   Max.   :4651.00
##  channel_type_rank video_views_for_the_last_30_days lowest_monthly_earnings
##  Min.   :   1.0    Min.   :        7                Min.   :    0
##  1st Qu.:  35.0    1st Qu.: 21072000                1st Qu.: 4100
##  Median :  87.0    Median : 43007000                Median : 9900
```

```
##  Mean   : 180.7    Mean   : 60158691                Mean   :12515
##  3rd Qu.: 143.0    3rd Qu.: 76903000                3rd Qu.:18100
##  Max.   :7638.0    Max.   :757789000                Max.   :57200
##  highest_monthly_earnings lowest_yearly_earnings highest_yearly_earnings
##  Min.   :     0           Min.   :     0          Min.   :       0
##  1st Qu.: 65600           1st Qu.: 49200          1st Qu.:  787600
##  Median :158000           Median :118500          Median : 1900000
##  Mean   :200205           Mean   :150152          Mean   : 2401309
##  3rd Qu.:289000           3rd Qu.:216700          3rd Qu.: 3500000
##  Max.   :915600           Max.   :686700          Max.   :11000000
##   created_year   created_date   Gross.tertiary.education.enrollment....
##  Min.   :2005   Min.   : 1.00   Min.   :  9.00
##  1st Qu.:2011   1st Qu.: 8.00   1st Qu.: 51.30
##  Median :2013   Median :17.00   Median : 60.00
##  Mean   :2013   Mean   :16.44   Mean   : 62.17
##  3rd Qu.:2015   3rd Qu.:25.00   3rd Qu.: 81.90
##  Max.   :2022   Max.   :31.00   Max.   :113.10
##    Population        Unemployment.rate Urban_population       Latitude
##  Min.   :5.520e+06   Min.   : 3.040    Min.   :  4694702   Min.   :-38.42
##  1st Qu.:5.171e+07   1st Qu.: 3.850    1st Qu.: 42106719   1st Qu.:-14.23
##  Median :1.004e+08   Median : 5.930    Median : 64324835   Median : 23.89
##  Mean   :1.318e+08   Mean   : 7.711    Mean   :105530510   Mean   : 22.58
##  3rd Qu.:2.126e+08   3rd Qu.:12.080    3rd Qu.:183241641   3rd Qu.: 52.13
##  Max.   :1.366e+09   Max.   :14.720    Max.   :471031528   Max.   : 61.92
##    Longitude       log_video.views log_subscribers
##  Min.   :-106.35   Min.   :18.90   Min.   :16.33
##  1st Qu.: -74.30   1st Qu.:21.85   1st Qu.:16.42
##  Median : -51.93   Median :22.34   Median :16.53
##  Mean   : -21.99   Mean   :22.29   Mean   :16.60
##  3rd Qu.:  10.45   3rd Qu.:22.82   3rd Qu.:16.73
##  Max.   : 133.78   Max.   :23.79   Max.   :17.65
```

```
cluster3_indices <- which(pca_result$x[, 1] < 0 & pca_result$x[, 2] > 0.75 & pca_result$x[, 2] < 1.875)
cluster3_data <- youtube_numeric[cluster3_indices, ]
summary(cluster3_data)
```

```
##       rank          uploads         video_views_rank    country_rank
##  Min.   :  1.0   Min.   :    15.0   Min.   :    1.00   Min.   :  1.00
##  1st Qu.:122.5   1st Qu.:   386.8   1st Qu.:   95.75   1st Qu.:  6.00
##  Median :214.0   Median :   765.0   Median :  195.50   Median : 51.50
##  Mean   :270.6   Mean   :  6428.5   Mean   :  704.74   Mean   : 59.12
##  3rd Qu.:363.8   3rd Qu.:  2659.2   3rd Qu.:  376.75   3rd Qu.: 94.75
##  Max.   :971.0   Max.   :200933.0   Max.   :40117.00   Max.   :176.00
##  channel_type_rank video_views_for_the_last_30_days lowest_monthly_earnings
##  Min.   :  1.00    Min.   :4.564e+07                Min.   : 11400
##  1st Qu.: 16.00    1st Qu.:1.451e+08                1st Qu.: 36250
##  Median : 42.50    Median :2.068e+08                Median : 51700
##  Mean   : 46.60    Mean   :3.293e+08                Mean   : 82321
##  3rd Qu.: 68.75    3rd Qu.:3.687e+08                3rd Qu.: 92225
##  Max.   :163.00    Max.   :2.258e+09                Max.   :564600
##  highest_monthly_earnings lowest_yearly_earnings highest_yearly_earnings
##  Min.   : 182600          Min.   : 136900        Min.   :  2200000
##  1st Qu.: 580200          1st Qu.: 435200        1st Qu.:  6925000
##  Median : 827350          Median : 620550        Median :  9900000
##  Mean   :1315489          Mean   : 988962        Mean   : 15807143
```

```
##  3rd Qu.:1475000          3rd Qu.:1100000        3rd Qu.: 17675000
##  Max.   :9000000          Max.   :6800000        Max.   :108400000
##   created_year   created_date   Gross.tertiary.education.enrollment....
##  Min.   :2005   Min.   : 1.00   Min.   :23.90
##  1st Qu.:2009   1st Qu.: 8.00   1st Qu.:88.20
##  Median :2013   Median :15.00   Median :88.20
##  Mean   :2013   Mean   :15.35   Mean   :82.26
##  3rd Qu.:2016   3rd Qu.:22.00   3rd Qu.:88.20
##  Max.   :2021   Max.   :31.00   Max.   :94.30
##    Population        Unemployment.rate Urban_population       Latitude
##  Min.   :2.870e+05   Min.   : 3.04     Min.   :    89431   Min.   :-38.42
##  1st Qu.:1.444e+08   1st Qu.:10.00     1st Qu.:107683889   1st Qu.: 37.09
##  Median :3.282e+08   Median :14.70     Median :270663028   Median : 37.09
##  Mean   :2.621e+08   Mean   :12.57     Mean   :209317064   Mean   : 31.14
##  3rd Qu.:3.282e+08   3rd Qu.:14.70     3rd Qu.:270663028   3rd Qu.: 37.09
##  Max.   :1.366e+09   Max.   :14.72     Max.   :471031528   Max.   : 61.52
##    Longitude       log_video.views log_subscribers
##  Min.   :-106.35   Min.   :19.61   Min.   :16.33
##  1st Qu.: -95.71   1st Qu.:23.12   1st Qu.:16.85
##  Median : -95.71   Median :23.51   Median :17.09
##  Mean   : -68.08   Mean   :23.47   Mean   :17.14
##  3rd Qu.: -74.30   3rd Qu.:23.85   3rd Qu.:17.33
##  Max.   : 127.77   Max.   :26.15   Max.   :19.32
```

Interestingly, we observe that channels appear to cluster by country. Cluster 1 (the one centered around (0, -3) is entirely composed of channels associated with India (although it does not include every channel associated with India). Cluster 2 (the one centered around (1.25, 0)) contains a high number of channels associated with Brazil, while cluster 3 is mostly comprised of channels associated with the United States.