**Sergio Rosendo**

**COM SCI X 450.2**

**Bianca Cung**

**Jan 22, 2023**

**Discussion 2**

**Data Gathering:**

**Source:** Kaggle

**Source URL:** https://www.kaggle.com/datasets/

**Dataset:** Most Subscribed YouTube Channels

**Dataset URL:** https://www.kaggle.com/datasets/surajjha101/top-youtube-channels-data

**Last Updated:** 5 months ago

**Dataset summary:**

The dataset contains information about the top 1,000 YouTube channels sampled based on subscription count.

**Data Attributes:**

Rank: Position among 1,000 channels based on subscription count

YouTuber: Channel name

Subscribers: Number of subscribers

VideoViews: Number channel collective views

VideoCount: Number of videos in channel

Category: Genre of channel

Started: Year the channel started

**Data Cleanup:**

The original dataset contains observations in which the VideoViews, VideoCount, and Category are 0 or empty.

Upon further inspection, the YouTuber value of these contains what appears to be a category (e.g. Music, Sports).

Knowing that this dataset was the product of a webscrape, it is safe to assume that these are erroneous entries made by the scrapper and thus can be removed.

19 channels were detected with blank Category. 14 of them corresponding to infant education. The Category 'Education' was given to these channels.

based on the categorization of a similar channel known as 'Coco Melon'. There were also 2 gaming, 1 music, and 2 unknown channels.

It is also worth noting that some of the YouTuber names were not properly displayed as they contain foreign/abnormal characters.

This could be a result of bad encoding of special characters.

Due to the possible usage of multiple encodings, large quantity of data, and relevance of this attribute the problem was not addressed.

Since all Subscribers values are in the millions, a column named SubscribersFormatted was added for readability.

**Stem and Leaf Plot:**

1 | 11111111111111111111111111111111111111111111111111111111111111111111+596

2 | 00000000000000000000000000000000000011111111111111111111222222222222+102

3 | 00000000001111111111222222223333333333333444445555666666777777888889

4 | 00001111223333344456666678

5 | 0222233333568889

6 | 1368

7 | 001467

8 | 57

9 | 09

10 | 02

11 | 1

12 |

13 | 9

14 | 0

15 |

16 |

17 |

18 |

19 |

20 |

21 |

22 | 2

**Code:**   
stem(data$SubscribersFormatted, scale = 1)

**Analysis:**

The stem and leaf plot brings insights on the following:

* 664 channels with at least 11 million subscribers.
* 170 channels with at least 20 million subscribers.
* 68 channels with at least 30 million subscribers.
* 26 channels with at least 40 million subscribers.
* 16 channels with at least 50 million subscribers.
* 4 channels with at least 60 million subscribers.
* 6 channels with at least 70 million subscribers
* 2 channels with at least 80 million subscribers
* 2 channels with at least 90 million subscribers
* 2 channels with at least 100 million subscribers
* 1 channel with at least 110 million subscribers
* 1 channel with at least 130 million subscribers
* 1 channel with at least 140 million subscribers
* 1 channel with at least 220 million subscribers

It is very lonely in the top with only 6 channels with over 100 million subscribers.

**Histogram:**  
**Chart, histogram

Description automatically generated**

**Code:**

dens <- density(data$SubscribersFormatted)

stem(data$SubscribersFormatted, scale = 1)

hist(data$SubscribersFormatted, freq = FALSE, main = 'Top Youtube Channels Subscriber Counts', xlab = 'Number of Subscribers (millions)', col = 'lightblue')

lines(dens, lwd = 2, lty = 1)

**Analysis:**

The histogram above is the graphical representation of the stem and leaf plot previously shown.

It is interesting to see how the number of channels lowers almost exponentially every increase of 10 million subscribers.

**Bar Plot  
Chart, bar chart

Description automatically generated**

**Code:**

category\_counts <- table(data$Category)

category\_counts <- category\_counts[order(category\_counts, decreasing = TRUE)]

par(mar=c(4, 7, 4, 1))

barplot(category\_counts, cex.names=0.7, las = 1, horiz = TRUE, main = "Youtube Top Channels Category Count", col = 'lightpink')

**Analysis:**

We can see the two most popular categories are music and entertainment, followed by gaming and blogs.

If you use YouTube long enough, you will notice that these are indeed the most common categories of videos suggestions.