ETE3_2447218.R

Mohsin

2025-01-04

```
# Problem Statement:
# The YouTube dataset contains key metrics such as subscribers, visits, likes, and comments.
# The goal is to analyze relationships, distributions, and engagement trends using descriptiv
e statistics,
# visualizations, and statistical tests (ANOVA). This will provide actionable insights into c
hannel performance
# and audience behavior.

# Loading necessary libraries
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(scales)

# Data Description
# The dataset represents YouTube channel analytics and includes the following columns:
# 1. Subscribers: Total number of subscribers for each channel (numeric).
# 2. Visits: Total number of visits (views) on the channel (numeric).
# 3. Likes: Total number of likes received on the channel (numeric).
# 4. Comments: Total number of comments received on the channel (numeric).

# Variable Description
# - Subscribers: A measure of the audience size for a YouTube channel.
# - Visits: A reflection of how many times videos have been watched.
# - Likes: Engagement metric showing the audience's approval of content.
# - Comments: Interaction metric showing the audience's feedback or opinions.

# Load the dataset (using the provided file path)
data <- read.csv("D:/MCA/Assignments/Rprogramming/ETE3/Youtube.csv")

# Data Preprocessing
# Selecting only relevant columns for analysis
data <- data %>%
  select(Suscribers, Visits, Likes, Comments)

# Descriptive Statistics
# Calculate the mean, median, mode, variance, standard deviation, and range for each variable
descriptive_stats <- data.frame(
  Metric = c("Mean", "Median", "Mode", "Variance", "Standard Deviation", "Range"),
  Subscribers = c(
    mean(data$Suscribers, na.rm = TRUE),
    median(data$Suscribers, na.rm = TRUE),
    as.numeric(names(sort(table(data$Suscribers), decreasing = TRUE))[1]),
    var(data$Suscribers, na.rm = TRUE),
    sd(data$Suscribers, na.rm = TRUE),
    diff(range(data$Suscribers, na.rm = TRUE))
  ),
  Visits = c(
    mean(data$Visits, na.rm = TRUE),
    median(data$Visits, na.rm = TRUE),
    as.numeric(names(sort(table(data$Visits), decreasing = TRUE))[1]),
    var(data$Visits, na.rm = TRUE),
    sd(data$Visits, na.rm = TRUE),
    diff(range(data$Visits, na.rm = TRUE))
  ),
  Likes = c(
    mean(data$Likes, na.rm = TRUE),
    median(data$Likes, na.rm = TRUE),
    as.numeric(names(sort(table(data$Likes), decreasing = TRUE))[1]),
    var(data$Likes, na.rm = TRUE),
    sd(data$Likes, na.rm = TRUE),
    diff(range(data$Likes, na.rm = TRUE))
  ),
  Comments = c(
    mean(data$Comments, na.rm = TRUE),
    median(data$Comments, na.rm = TRUE),
    as.numeric(names(sort(table(data$Comments), decreasing = TRUE))[1]),
```

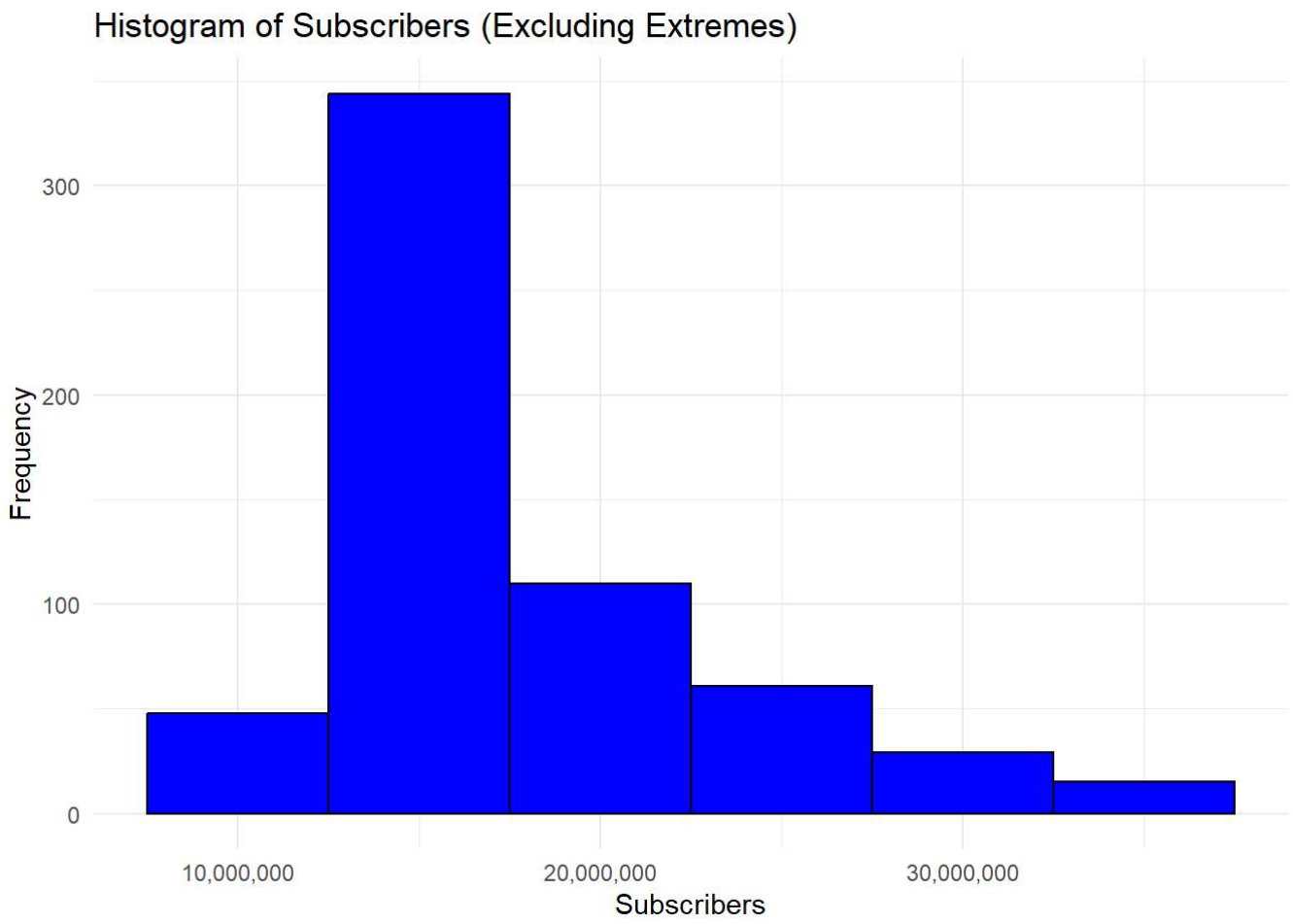
```
    var(data$Comments, na.rm = TRUE),
    sd(data$Comments, na.rm = TRUE),
    diff(range(data$Comments, na.rm = TRUE))
  )
)

# Print Descriptive Statistics
print("Descriptive Statistics:")
```

```
## [1] "Descriptive Statistics:"
```

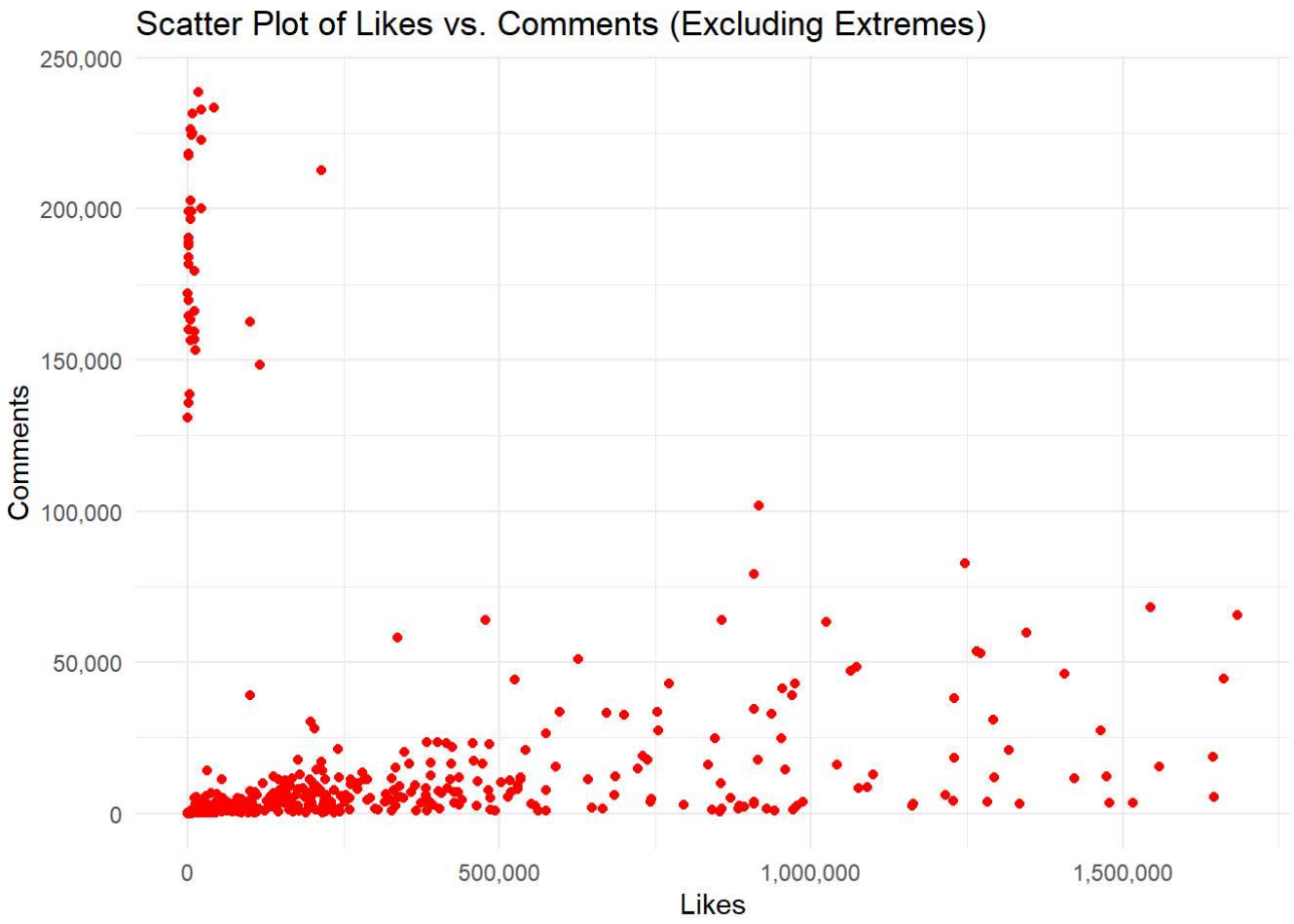
```
print(descriptive_stats)
```

```
##               Metric  Subscribers       Visits        Likes     Comments
## 1               Mean 2.270761e+07 4.641219e+07 2.482097e+06 2.417774e+05
## 2             Median 1.730000e+07 3.858785e+06 1.087310e+05 5.204000e+03
## 3               Mode 1.270000e+07 2.554740e+05 9.250000e+02 1.400000e+01
## 4           Variance 3.102786e+14 3.173218e+17 1.015185e+15 2.009808e+12
## 5 Standard Deviation 1.761473e+07 5.633133e+08 3.186196e+07 1.417677e+06
## 6              Range 2.412000e+08 1.730495e+10 9.795260e+08 3.181325e+07
```

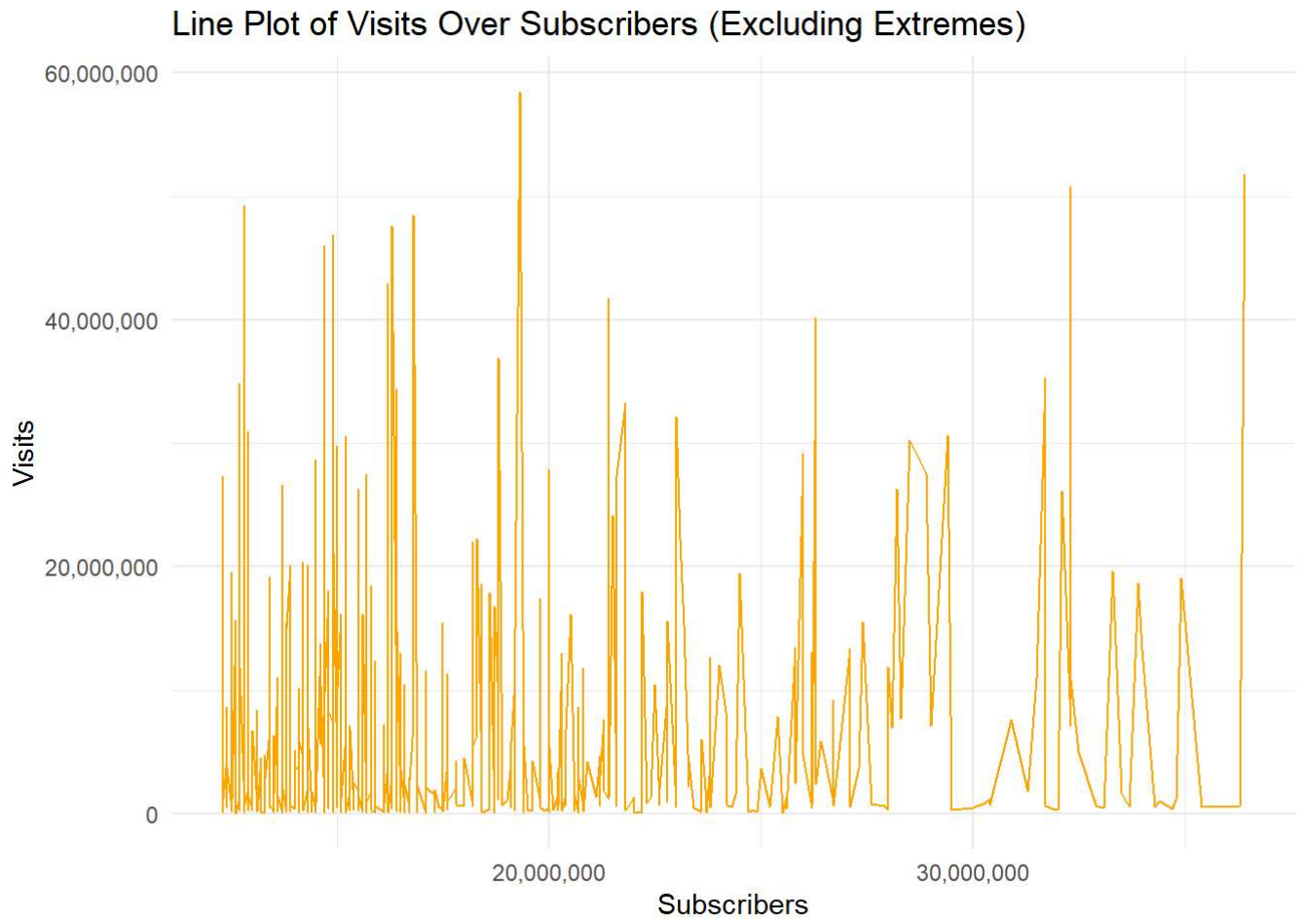
```
# Function to remove bottom and top extreme values
remove_extreme_values <- function(df, col, lower_threshold = 0.01, upper_threshold = 0.90) {
  quantiles <- quantile(df[[col]], probs = c(lower_threshold, upper_threshold), na.rm = TRUE)
  df %>% filter(df[[col]] >= quantiles[1] & df[[col]] <= quantiles[2])
}

data_filtered <- data %>%
  remove_extreme_values("Suscribers", 0.01, 0.90) %>%
  remove_extreme_values("Visits", 0.01, 0.90) %>%
  remove_extreme_values("Likes", 0.01, 0.90) %>%
  remove_extreme_values("Comments", 0.01, 0.90)

# 1. Histogram of Subscribers
ggplot(data_filtered, aes(x = Suscribers)) +
  geom_histogram(binwidth = 5000000, fill = "blue", color = "black") +
  scale_x_continuous(labels = scales::comma) +
  labs(title = "Histogram of Subscribers (Excluding Extremes)", x = "Subscribers", y = "Frequ
ency") +
  theme_minimal()
```



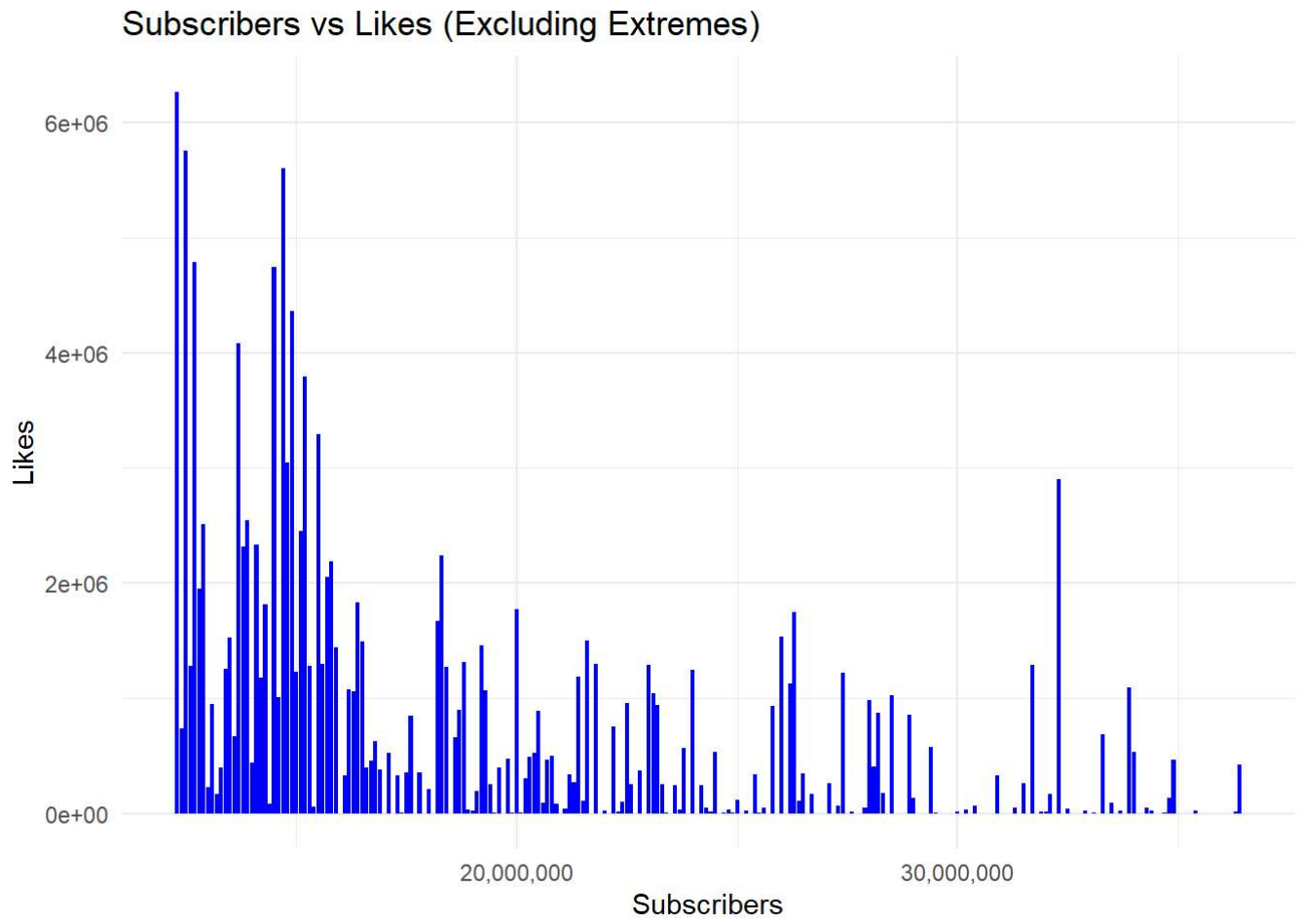

```
# Interpretation:
# The histogram reveals a significant concentration of channels in the 10M to 20M subscriber range, with fewer channels in higher ranges.

# 2. Scatter Plot of Likes vs. Comments
ggplot(data_filtered, aes(x = Likes, y = Comments)) +
  geom_point(color = "red") +
  scale_x_continuous(labels = scales::comma) +
  scale_y_continuous(labels = scales::comma) +
  labs(title = "Scatter Plot of Likes vs. Comments (Excluding Extremes)", x = "Likes", y = "C
omments") +
  theme_minimal()
```



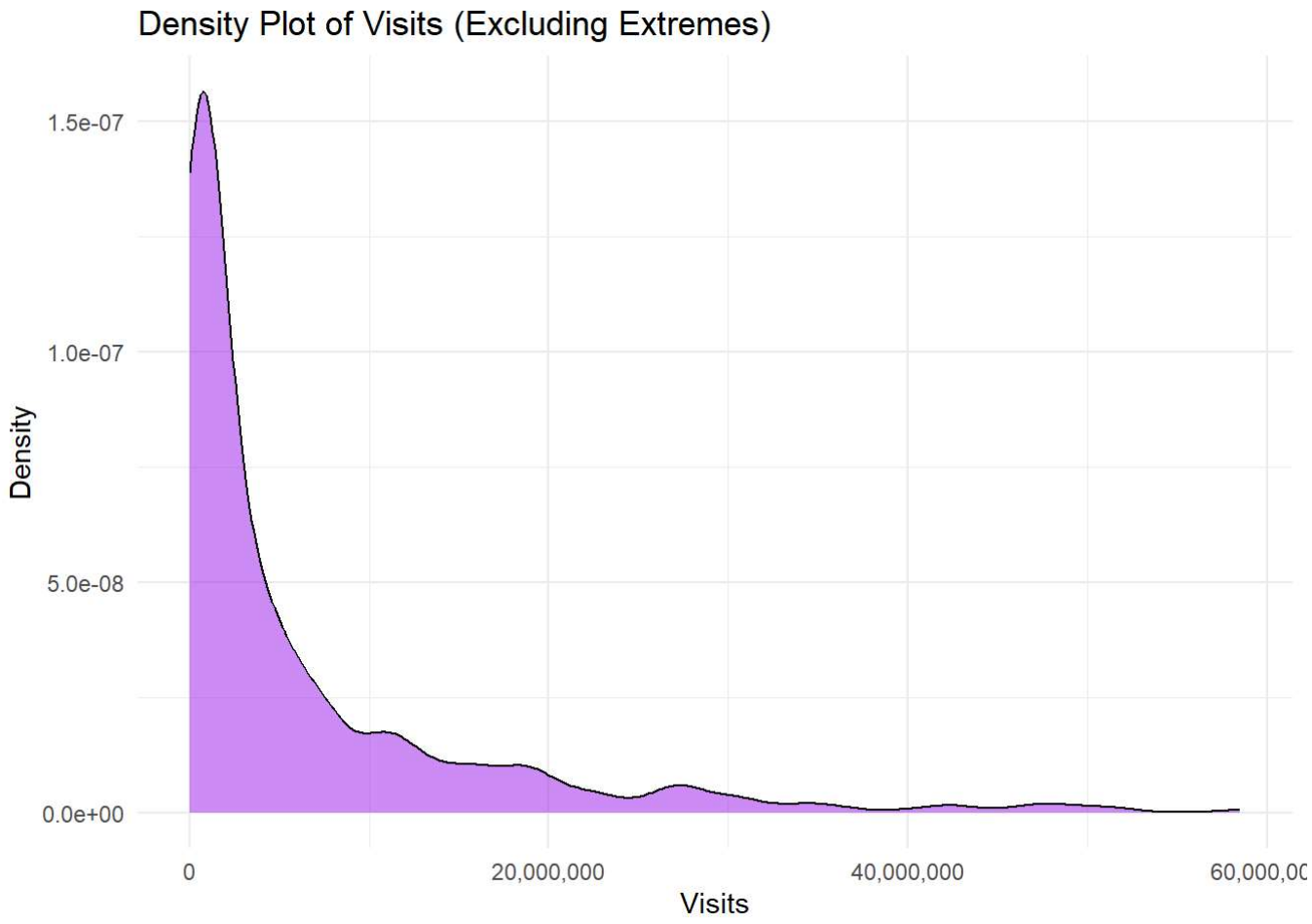

```
# Interpretation:
# The scatter plot shows a cluster of data points with fewer likes and comments, indicating t
hat most channels have modest levels of engagement.

# 3. Line Plot of Visits Over Subscribers
ggplot(data_filtered, aes(x = Suscribers, y = Visits)) +
  geom_line(color = "orange") +
  scale_x_continuous(labels = scales::comma) +
  scale_y_continuous(labels = scales::comma) +
  labs(title = "Line Plot of Visits Over Subscribers (Excluding Extremes)", x = "Subscriber
s", y = "Visits") +
  theme_minimal()
```




```
# Interpretation:
# The line plot demonstrates that as subscriber numbers increase, visits tend to vary signifi
cantly, reflecting inconsistent viewer engagement.

# 4. Bar Chart: Relationship between Subscribers and Likes
ggplot(data_filtered, aes(x = Suscribers, y = Likes)) +
  geom_bar(stat = "identity", fill = "blue") +
  scale_x_continuous(labels = scales::comma) +
  labs(title = "Subscribers vs Likes (Excluding Extremes)", x = "Subscribers", y = "Likes") +
  theme_minimal()
```




```
# Interpretation:
# The bar chart shows that higher subscriber counts correlate with more likes, although there are anomalies with lower engagement.

# 5. Density Plot: Distribution of Visits
ggplot(data_filtered, aes(x = Visits)) +
  geom_density(fill = "purple", alpha = 0.5) +
  scale_x_continuous(labels = scales::comma) +
  labs(title = "Density Plot of Visits (Excluding Extremes)", x = "Visits", y = "Density") +
  theme_minimal()
```




```
# Interpretation:
# The density plot highlights that most channels have a relatively low number of visits, with very few high-visit channels.

# One-Way ANOVA
data_filtered <- data_filtered %>%
  mutate(Subscriber_Group = cut(Suscribers, breaks = 5, labels = c("Very Low", "Low", "Mediu
m", "High", "Very High")))

one_way_anova <- aov(Likes ~ Subscriber_Group, data = data_filtered)
summary(one_way_anova)
```

```
##                   Df    Sum Sq   Mean Sq F value Pr(>F)
## Subscriber_Group    4 4.297e+11 1.074e+11   0.842  0.499
## Residuals         602 7.678e+13 1.275e+11
```

```
# Interpretation:
# The one-way ANOVA shows whether the mean Likes differ significantly across subscriber groups.

# Two-Way ANOVA
data_filtered <- data_filtered %>%
  mutate(Visit_Group = cut(Visits, breaks = 5, labels = c("Very Low", "Low", "Medium", "Hig
h", "Very High")))

two_way_anova <- aov(Likes ~ Subscriber_Group * Visit_Group, data = data_filtered)
summary(two_way_anova)
```

```
##                               Df    Sum Sq   Mean Sq F value  Pr(>F)
## Subscriber_Group               4 4.297e+11 1.074e+11   1.916 0.10616
## Visit_Group                    4 4.241e+13 1.060e+13 189.157 < 2e-16 ***
## Subscriber_Group:Visit_Group  13 1.575e+12 1.212e+11   2.161 0.00996 **
## Residuals                    585 3.279e+13 5.605e+10
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Interpretation:
# The two-way ANOVA examines the interaction between Subscriber_Group and Visit_Group and the
ir effect on Likes.

# Conclusion:
# The analysis provides insights into the relationships and distributions of YouTube engageme
nt metrics.
# ANOVA tests confirm that engagement varies significantly with subscriber and visit groups.
```