STATISTICAL ANALYSIS OF THE POPULAR WEBSITES

Siva Manda, Manuj Jha and
Krishna Charan Bajanthri Chikalaguriki

Dr. Alireza Sheikh-Zadeh
ISQS 5347 Adv Statistical Methods

INTRODUCTION

In this study, we are looking at a data set containing information about the most popular websites across the globe. For each website, the data set contains attributes/information such as the average daily visitors, average daily page views, reach per day, social media likes (Facebook, Twitter, Tumblr, etc.) and many more.

Our aim is to analyze the popular websites based on these various parameters provided in the dataset, using R. With these statistical analyses, we attempt to gain some insights into the distribution of the data, construct relationships between variables in the data by using correlation and covariance, proving the central limit theorem with the data, and calculation of probabilities and conditional probabilities for different variables in the data.

The data set contains data entries for websites from each of the 191 countries in the globe with a total of 9541 observations and 31 variables of interest. Each country has entries for its top 50 websites, which are ranked from 1 to 50.

In this study, we are doing 5 different kinds of statistical analyses/tests for this data, which are plotting the distribution, applying central limit theorem on our dataset, correlation and covariance, establishing various probabilities/conditional probabilities and applying multivariate analysis.

By finding the distribution of the data, we are looking to find a listing of the outcomes or possible values of and how they often they occur, and figure out what kind of function or equation, called a probability distribution function, links each of these values with its probability of occurrence. In this case, we are looking at whether the variables average daily visitors, average daily page views, and daily reach fits the normal distribution function.

Then we look to prove the central limit theorem for the data, which states that given a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from the same population will be approximately normally distributed.

Following which, we will measure the correlation and covariance between select variables from the data set, such as Country rank vs (Average Page views, Average Visitors, Traffic Rank) to see the relationships between these variables. Correlation and covariance measure a linear relationship between these variables. Finally, we will conduct multivariate analysis across various fields and try to understand how one variable influences other.

DATA SOURCE

The dataset is obtained from Kaggle, with the following URL:

https://www.kaggle.com/bpali26/popular-websites-across-the-globe

HYPOTHESES:

1. Investigation of conditional probability and Bayes' theorem on the dataset
2. Understanding the distributions of various variables, testing for normality
3. Establishing the Central Limit Theorem for our dataset
4. Analyzing the relation between the variables of dataset using correlation and covariance
5. Performing multivariate data analysis for understanding the impact of social media on the traffic ranking

ASSUMPTIONS:

- The data collected is random and free from any sampling biases.
- The observations for each website are independent of the other websites, i.e. independence of data.
- We are considering the “country” column for the location of the website, instead of the “location” column.
- By omitting the missing values, we are not losing any significant data pattern.

DATA IMPORTING AND CLEANSING

We are importing the dataset containing the top 50 ranked sites for each of the 191 countries in the globe along with their traffic rank into our R environment as follows:

```
data <- read.csv("data/Web_Scrapped_websites.csv", na.strings = c("", "NA", "-"),   stringsAsFactors = F)
```

The numeric data in variable “Traffic_Rank” is read into R environment as characters due to white space character. In order to remove the space we used gsub() function and converted the variable into numeric data type.

```
data$Traffic_Rank<- gsub(" ", "", data$Traffic_Rank, fixed = TRUE)
data$Traffic_Rank <- as.numeric(data$Traffic_Rank)
```

We use the same process for converting the Avg_Daily_Visitors and Avg_Daily_Pageviews variables into numeric data type as follows:

```
data$Avg_Daily_Visitors<- gsub(" ", "", data$Avg_Daily_Visitors, fixed = TRUE)
```

```
data$Avg_Daily_Visitors<- as.numeric(data$Avg_Daily_Visitors)
data$Avg_Daily_Pageviews<- gsub(" ", "", data$Avg_Daily_Pageviews, fixed = TRUE)
data$Avg_Daily_Pageviews<- as.numeric(data$Avg_Daily_Pageviews)
```

PROBABILITY ANALYSES OVER THE DATA:

1. Finding the probability that a website from dataset has the country as "United States":

   P(*Website country is United States*) =

   Number of websites from US / Total number of websites

   Solving in R:

```
#data : Contains website data from all the countries
#dataUS : Contains website data from United States filtered from data
dataUS<- data %>%  filter(country == "United States")
total_Country = length(data$country) = 9540
total_US = length(dataUS$country) = 50
P(Website country is United States):
P1 = total_US / total_Country = 0.00524109
```

2. Probability that a website has 'excellent' trustworthiness from across the world:

   P(*Website has excellent trust from over the world*) =

   Number of websites with 'excellent' trustworthiness / Total no of websites

   Solving in R:

```
no_of_excellent_trust <- data %>% filter(Trustworthiness == 'Excellent') %>% tally()
Number of websites with 'excellent' trustworthiness = 6071
Total no of websites = 9540
P(Website has excellent trust from over the world):
P2 = no_of_excellent_trust / length(data$Trustworthiness) = 0.6363732
```

3. Probability that the website has ‘excellent’ trustworthiness given it is from United States:

P (*Website has ‘excellent’ trustworthiness / Website is from US*) =

Number of websites with ‘excellent’ trustworthiness / Total no of websites in US

Solving in R:

We find the number of websites with ‘Excellent’ trustworthiness from dataset having records only from United States

```
no_of_excellent_trust_US <- dataUS %>% filter(Trustworthiness == 'Excellent') %>% tally() = 46
P3 = no_of_excellent_trust_US / length (dataUS $ Trustworthiness) = 46 / 50 = 0.92
```

4. Given that a website selected randomly has ‘excellent’ trustworthiness, probability of that website belonging to United States:

P(*Website from US / Website has ‘excellent’ trustworthiness*)

We have to use Bayes’ Theorem to solve this conditional probability.

P(*Website from USA / Website has excellent trustworthiness*) =

(P(*Website has excellent trust / website is from US*) * P(*Website country is US*)) /

P(*Website has excellent trust from over the world*)

Or, P4 = (P1 * P3) / P2 = (0.00524109 * 0.92) / 0.6363732 = 0.007577005

5. Probability that a website chosen at random has ‘Very poor’ child safety:

P(*A website has 'Very poor' Child Safety*) =

No of websites with ‘Very poor’ child safety / Total no of websites

Using to R to solve for the following:

```
no_very_poor <- data %>% filter(Child_Safety == "Very poor") %>% tally() = 764
P5 = no_very_poor / length(data$Child_Safety) = (764 / 9540) = 0.08008386
```

6. Given that a website hails from US, probability that it has 'Very poor' child safety:

   P(*'Very poor' Child Safety / Website is from US*) =

   No of websites with 'Very poor' child safety / Total no of websites in US

   Solving in R:

   We find out the following by counting the number of websites with 'Very poor' child safety from data subset containing websites hailing only from United States.

   no_unsatisfactory_US <- dataUS %>% filter(Child_Safety == "Very poor") %>% tally() =3

7. We find the required probability by dividing number of interested websites by the number of websites in US.

   P6 = no_unsatisfactory_US / length(dataUS$Child_Safety) = 3 / 50 = 0.06

8. Given that a website has 'Very poor' child safety, probability that it hailed from US:

   P(*website is from US / 'Very poor' child safety*)

   Since here the condition is reversed from the last time, we need to use Bayes' Theorem

   To find this probability using the probabilities already evaluated.

   P(*website is from US / 'Very poor' child safety*) =

   (P(*'Very poor' child safety / Website is from US*) * P(*website is from US*)) /

   P(*'Very poor' child safety*)

   Or, P7 = (P6 * P1) / P5 = (0.06 * 0.005241) / 0.8008 = 0.003926702

EXPLORATORY ANALYSES:

We shall divide our exploratory analyses into two parts namely, Categorical data analysis and Numerical data analysis.

Categorical Data Analysis:

We have two types of data, i.e. numerical and categorical data. The categorical data is represented as factors in R using an equivalent integer. In the data we have three variables, namely, "Trustworthiness", "Child_Safety", "Privacy". These variables are read into our R environment as character type, hence we convert them into factorial data.

For further categorical data analyses, let us consider the websites data of the country United States only. We filter out the data pertaining only to the country USA from the complete dataset as follows:

```
dataUS<- data %>%  filter(country == "United States")
```

Let us select the required variables from the above US dataset we created omitting the “NA” value records.

```
dataCat <- na.omit(dataUS) %>% select("Trustworthiness", "Child_Safety", "Privacy")
```

We calculate the number of each type of factors occurring in the categorical variables of the US dataset.

```
dataTrust <- dataCat %>% group_by(Trustworthiness=as.factor(dataCat$Trustworthiness)) %>% tally()
dataChild <- dataCat %>% group_by(Child_Safety=as.factor(dataCat$Child_Safety)) %>% tally()
dataPrivacy <- dataCat %>% group_by(Privacy=as.factor(dataCat$Privacy)) %>% tally()
```

Following is the output of the above R code:

```
> dataTrust
# A tibble: 2 x 2
  Trustworthiness     n
            <fctr> <int>
1       Excellent    25
2            Good     1
> dataChild
# A tibble: 3 x 2
  Child_Safety     n
         <fctr> <int>
1    Excellent    23
2         Good     2
3         Poor     1
> dataPrivacy
# A tibble: 2 x 2
    Privacy     n
     <fctr> <int>
1 Excellent    25
2      Good     1
```

Visualizing the above data using ggplot() function from the "ggplot2" package

a) *Trustworthiness of US websites:*

```
ggplot(dataTrust, aes(x=Trustworthiness,y=n))
        + geom_bar(stat="identity", aes(fill = Trustworthiness))
        + ggtitle("Trustworthiness of websites in US")
        + labs(x="Trustworthiness",y="Count")
        + theme(plot.title = element_text(hjust = 0.5))
```

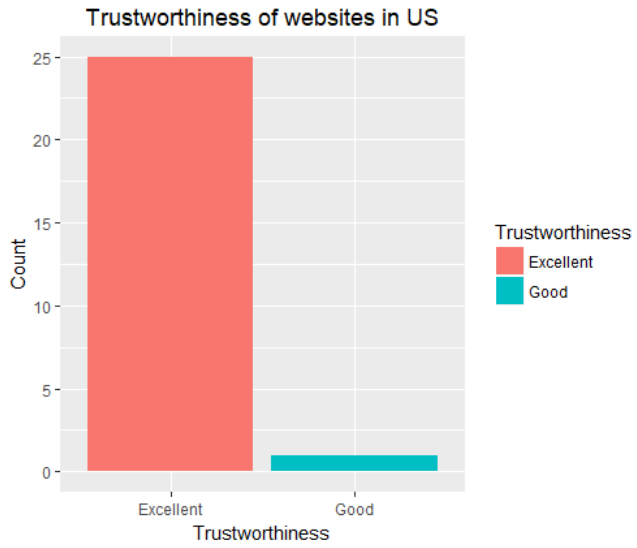


Figure 1: Categorical data summary for trustworthiness of US websites

b) *Child Safety of US websites:*

```
ggplot(dataChild, aes(x=Child_Safety,y=n))
        + geom_bar(stat="identity", aes(fill = Child_Safety))
        + ggtitle("Child Safety of websites in US")
        + labs(x="Child Safety",y="Count")
        + theme(plot.title = element_text(hjust = 0.5))
```

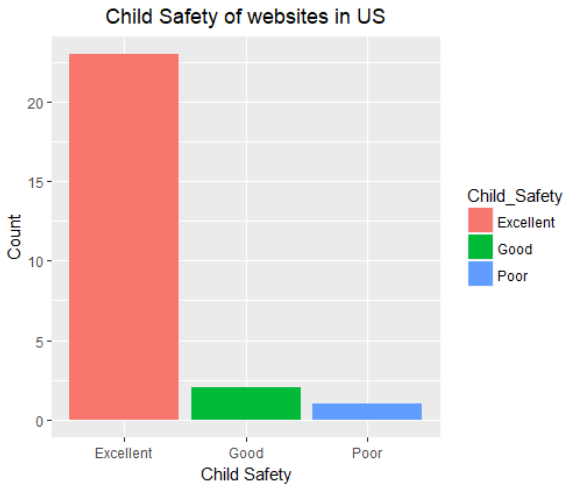


Figure 2: Categorical data summary for child safety of US websites

c) *Privacy of US websites:*

```
ggplot(dataPrivacy, aes(x=Privacy,y=n))
        + geom_bar(stat="identity", aes(fill = Privacy))
        + ggtitle("Privacy of websites in US")
        + labs(x="Privacy",y="Count")
        + theme(plot.title = element_text(hjust = 0.5))
```

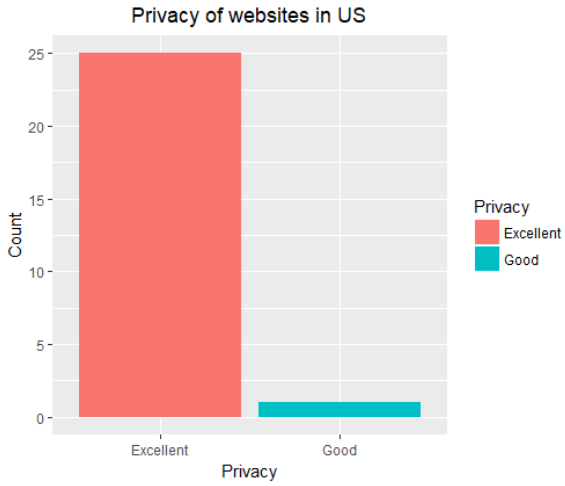


Figure 3:Categorical data summary for privacy of US websites

NORMALITY TESTS

a) Let us observe how the Avg_Daily_Visitors data is distributed. We calculate the average number of daily visitors by grouping country-wise a websites Avg_Daily_Visitors.

```
dataCLT<- data %>% group_by(country) %>% summarise(Average = mean (Avg_Daily_Visitors,
na.rm=TRUE))
qqnorm(dataCLT$average)
qqline(dataCLT$average)
```

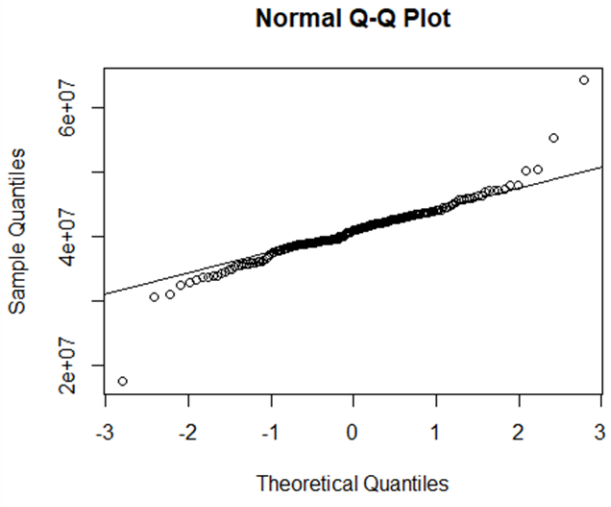


Figure 4: QQ plot for Avg_Daily_Visitors data

We can observe that the observations appear closer to the line, except for a few outliers. Hence, we can conclude that the average of Avg_Daily_Visitors data is approximately normally distributed with heavy tails.

The outliers present in the above QQ plot are identified and marked in “red” as follows:

```
> identify(qqnorm(dataCLT$average),col=”Red”)
[1] 82 92
```

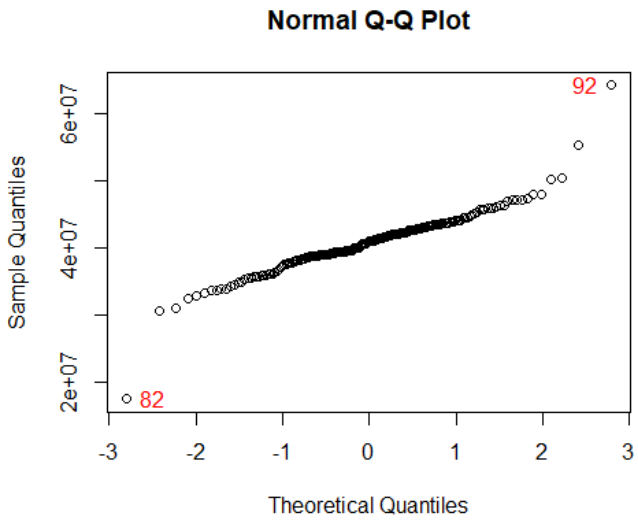


Figure 5: QQ plot for Avg_Daily_Visitors data with Outliers marked

LOG TRANSFORMATION OF DATA

Let us observe the data of Avg_Daily_visitors , Avg_Daily_Pageviews and Daily_Pageviews by plotting a histogram for the data.

```
par(mfrow=c(1,3))
hist(data$Avg_Daily_Visitors, main="Histogram")
hist(data$Avg_Daily_Pageviews, main="Histogram")
hist(data$Daily_Pageviews_per_user, main="Histogram")
```

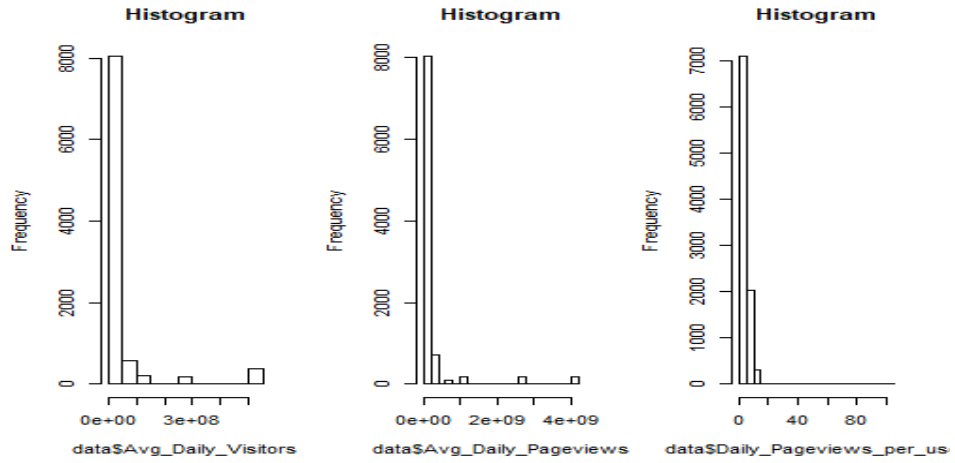


Figure 6: Histogram of variables of our interest

From the figure 6, we can observe the right skewness of the distributions. However, analysis of data is easier and more feasible with reduced variability. Hence to reduce the variability of the data, we use log transformation and make data conform more closely to the normal distribution.

a) *<u>Distribution of average daily visitors</u>*

```
hist(log10(data$Avg_Daily_Visitors) , main="Histogram")
qqnorm(data$Avg_Daily_Visitors)
qqline(data$Avg_Daily_Visitors)
```

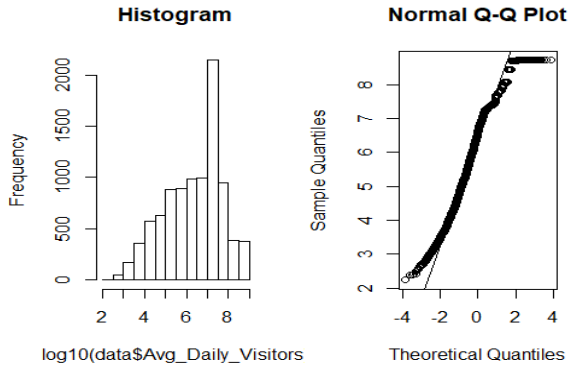


Figure 7: Histogram and QQ Plot of log(Avg_Daily_Visitors)

We observe that the log transformation of Avg_Daily_Visitors data is almost normally distributed with most of the points falling on the straight line.

b) *Distribution of Average Daily Page Views*

```
hist(log10(data$Avg_Daily_Pageviews) , main="Histogram")
qqnorm(data$Avg_Daily_Pageviews)
qqline(data$Avg_Daily_Pageviews)
```

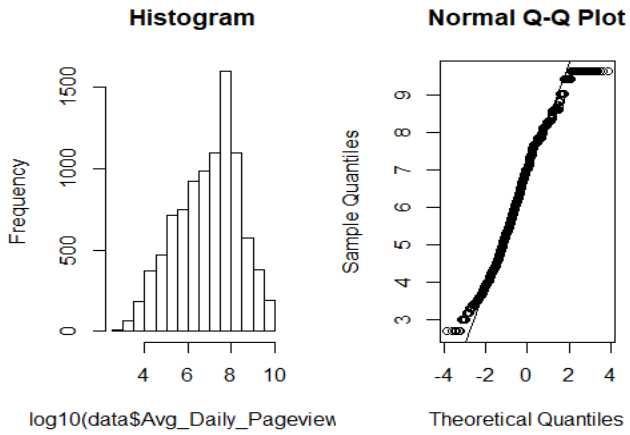


Figure 8: Histogram and QQ Plot of log(Avg_Daily_Pageviews)

We observe that the log transformation of Avg_Daily_Pageviews data is almost normally distributed with most of the points falling on the straight line.

c) *Distribution of Daily Reach*

```
hist(log(data$Daily_Pageviews_per_user) , main="Histogram")
qqnorm(data$Daily_Pageviews_per_user)
qqline(data$Daily_Pageviews_per_user)
```

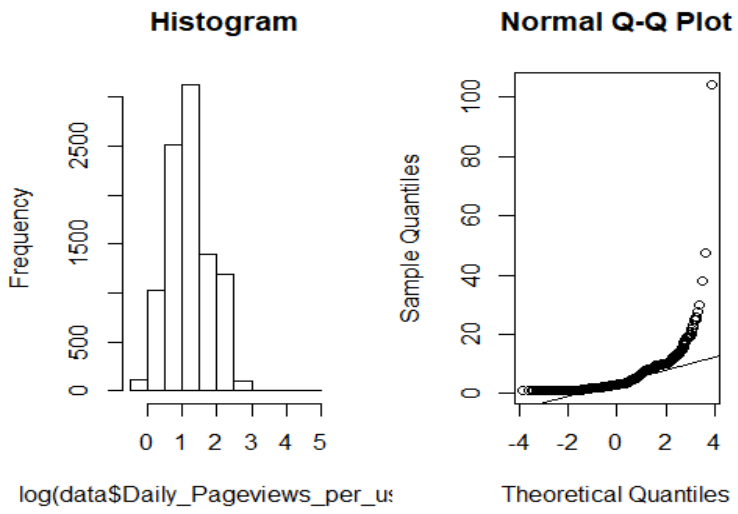


Figure 9: Histogram and QQ Plot of log(Daily_Pageviews_per_user)

We can observe that the log transformation of Daily_Pageviews_per_user data is right skewed and is not distributed normally.

CENTRAL LIMIT THEORAM

The central limit theorem (CLT) is a statistical theory that states that given a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from the same population will be approximately normally distributed. We will y to prove the theorem for the 'Avg_Daily_Visitors' column which has entries from 135 countries and each country having 50 observations. We will group the data by country and find mean of the 'Avg_Daily_Visitors' column using summarise() function in R. This will give us the mean value for each of the 135

samples corresponding to each country having 50 observations in each. Then we will plot the mean for each sample and judge the normality of distribution using Q-Q plot.

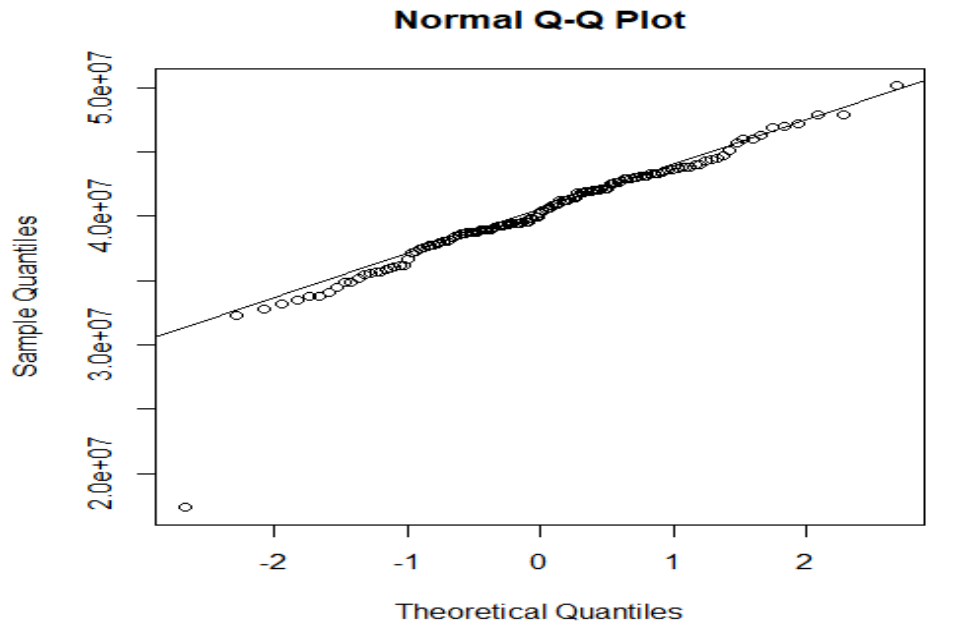
```
# Central Limit Theoram for Avg_Daily_Visitors: One Sample has 50 observations, total 135 samples

dataCLT<- data %>% group_by(country) %>% summarise(average = mean(Avg_Daily_Visitors))

dataCLT<- dataCLT %>% filter (average != "NA")

qqnorm(dataCLT$average)

qqline(dataCLT$average)
```



We can see that most of the points in the Q-Q plot fall in a straight line where each point corresponds to the mean value for each sample in 'Avg_Daily_Visitors' column.

So, we can confirm the Central Limit Theorem by saying that the distribution of mean score from sufficiently large number of samples follow approximately follow a normal distribution.

CORRELATION AND COVARIANCE

Let us utilize the data set containing only US websites for computing correlation and covariance among different variables of interest.

a) Country (Website) Rank vs Average Daily Visitors

```
cor(dataUS$Country_Rank, dataUS$Avg_Daily_Visitors)
cov(dataUS$Country_Rank, dataUS$Avg_Daily_Visitors)
plot(dataUS$Country_Rank, dataUS$Avg_Daily_Visitors, xlab = "Website Rank", ylab = "Average Daily Visitors", main = "Average Daily Visitors vs Website Rank")
abline(lm(dataUS$Avg_Daily_Visitors~dataUS$Country_Rank))
```

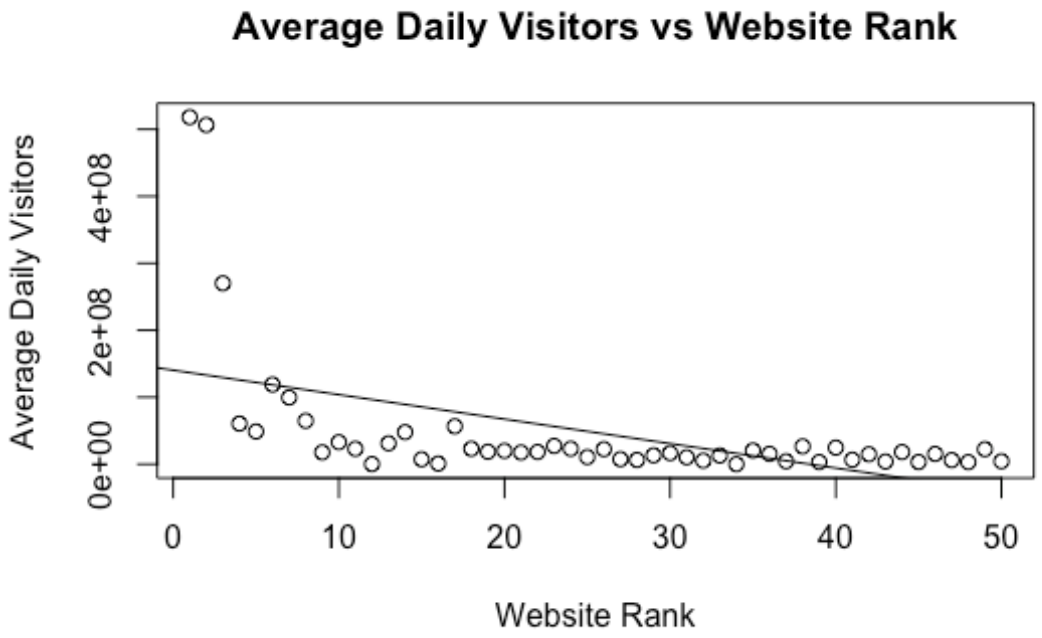


Figure 10: Plot of Avg_Daily_Visitor vs Website Rank

Correlation = -0.5091063

Covariance = -777764204

From the plot above and the correlation and covariance values, we can see that the variables Avg_Daily_Visitors and Country_Rank have a negative correlation. The first few data points have a very high number of average daily visitors than the lower ranked websites. The higher the rank number, the lower the number of average daily visitors.

Similarly, we can measure the correlation and covariance between Country Rank and the Average number of daily page views for each website.

b) Country (Website) Rank vs Average Daily Pageviews

```
cor(dataUS$Country_Rank, dataUS$Avg_Daily_Pageviews)
cov(dataUS$Country_Rank, dataUS$Avg_Daily_Pageviews)
plot(dataUS$Country_Rank, dataUS$Avg_Daily_Pageviews, xlab = "Website Rank", ylab =
"Average Daily Pageviews", main = "Average Daily Pageviews vs Website Rank")
abline(lm(dataUS$Avg_Daily_Pageviews ~ dataUS$Country_Rank))
```

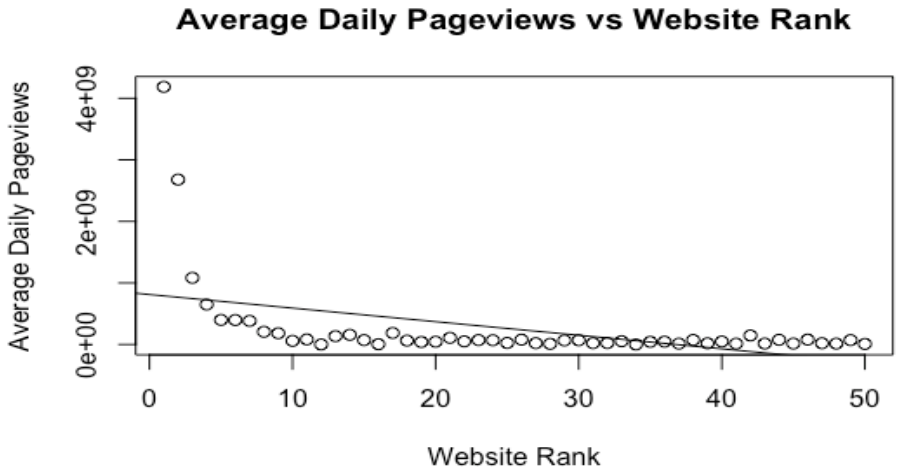


Figure 11: Plot of Avg_Daily_Pageviews vs Website Rank

Correlation = - 0.4609779 and Covariance = - 4699616735

Looking at the plot, we can see that this is similarly shaped as the previous one and state that the Average Daily Page views and Country Rank for the website have a negative correlation. This is confirmed by the computed values of the correlation and covariance, which are both negative. This makes sense because as a website as a website is less popular in a country, its average daily page views will be lower than a higher ranked one.

c) Country (Website Rank vs Traffic Rank)

```
cor(dataUS$Country_Rank, dataUS$Traffic_Rank)
cov(dataUS$Country_Rank, dataUS$Traffic_Rank)
plot(dataUS$Country_Rank, dataUS$Traffic_Rank, xlab = "Website Rank", ylab = "Traffic Rank",
main = "Average Daily Pageviews vs Traffic Rank")
abline(lm(dataUS$Traffic_Rank~ dataUS$Country_Rank))
```

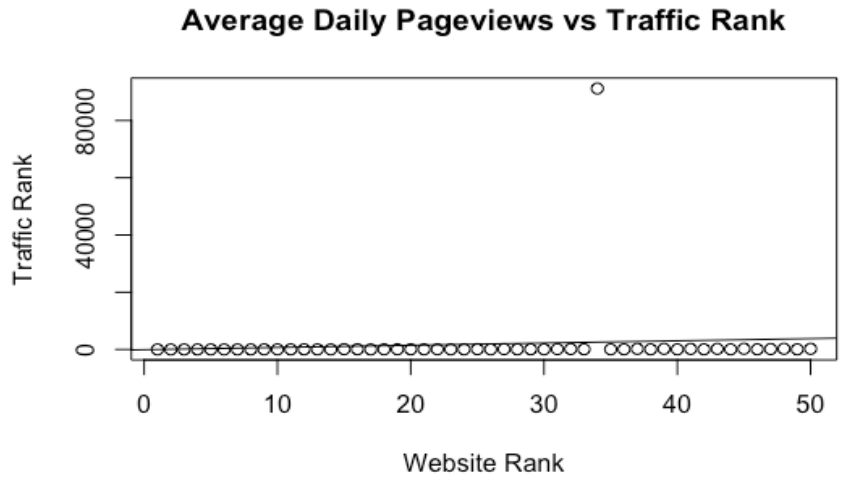


Figure 12: Plot of Avg_Daily_Pageviews vs Traffic Rank

Correlation = 0.08776685 and Covariance = 16488.16

By looking at this plot, we can infer that there is one website which has a significantly lower rank (higher rank number), than the other websites, making the other points seem constant. However, there is a positive correlation between traffic rank and country (website rank). This makes sense because the Country rank is the rank of a particular website for a specific country, and the traffic rank is the overall rank of all the websites in the data. This is confirmed by the measured correlation and covariance values, which are both positive.

MULTIVARIATE DATA COMPARISON USING GGALLY::GGPAIRS

The ggpairs() function produces a pairwise comparison of multivariate data. By default, ggpairs() provides two different comparisons of each pair of columns and displays either the density or count of the respective variable along the diagonal. With different parameter settings, the diagonal can be replaced with the axis values and variable labels. Let us understand the impact of social media on the traffic ranking using the ggpairs() function.

We create a new dataset with our required social media data and traffic rank from the original dataset.

```
dataSocial <- na.omit(data) %>%
select("Website", "Avg_Daily_Visitors", "Facebook_likes", "Twitter_mentions", 'Google_pluses", "LinkedIn_mentions", "Pinterest_pins", "StumbleUpon_views", "Traffic_Rank")
```

Let us observe how the data looks like:

```
> dataSocial[3:8] %>% head()
```

	Facebook_likes	Twitter_mentions	Google_pluses	LinkedIn_mentions	Pinterest_pins	StumbleUpon_views
2	94.2K	11.2K	11.7M	1.67K	10.8K	246K
3	13.5K	16.5K	19.3M	60K	47	329K
4	5.87M	64.4K	127K	6.23K	4.15K	23.1K
5	17.2K	1.11K	798K	7.5K	433	68.9K
7	9	9.37K	7.2K	1.12K	16	136
8	476	162	126K	1.24K	5	5

We can observe that our social media data is represented as characters with “K=1,000” and “M=1,000,000”. We use the “stringr” package to perform required operations to convert the data into numeric type as follows:

```
for(j in 3:8){
for(i in 1 : length(dataSocial$Website)){
 if(str_sub(dataSocial[i,j],-1,-1)=="K"){
  dataSocial[i,j]<-as.numeric(str_sub(dataSocial[i,j],1,-2))*1000
 }
 else if(str_sub(dataSocial[i,j],-1,-1)=="M"){
    dataSocial[i,j]<-as.numeric(str_sub(dataSocial[i,j],1,-2))*1000000
  }
 else
    dataSocial[i,j]<-as.numeric(str_sub(dataSocial[i,j],1,str_length(dataSocial[i,j])))
}}
for(i in 3:8){dataSocial[i]<-as.numeric(unlist(dataSocial[i]))}
```

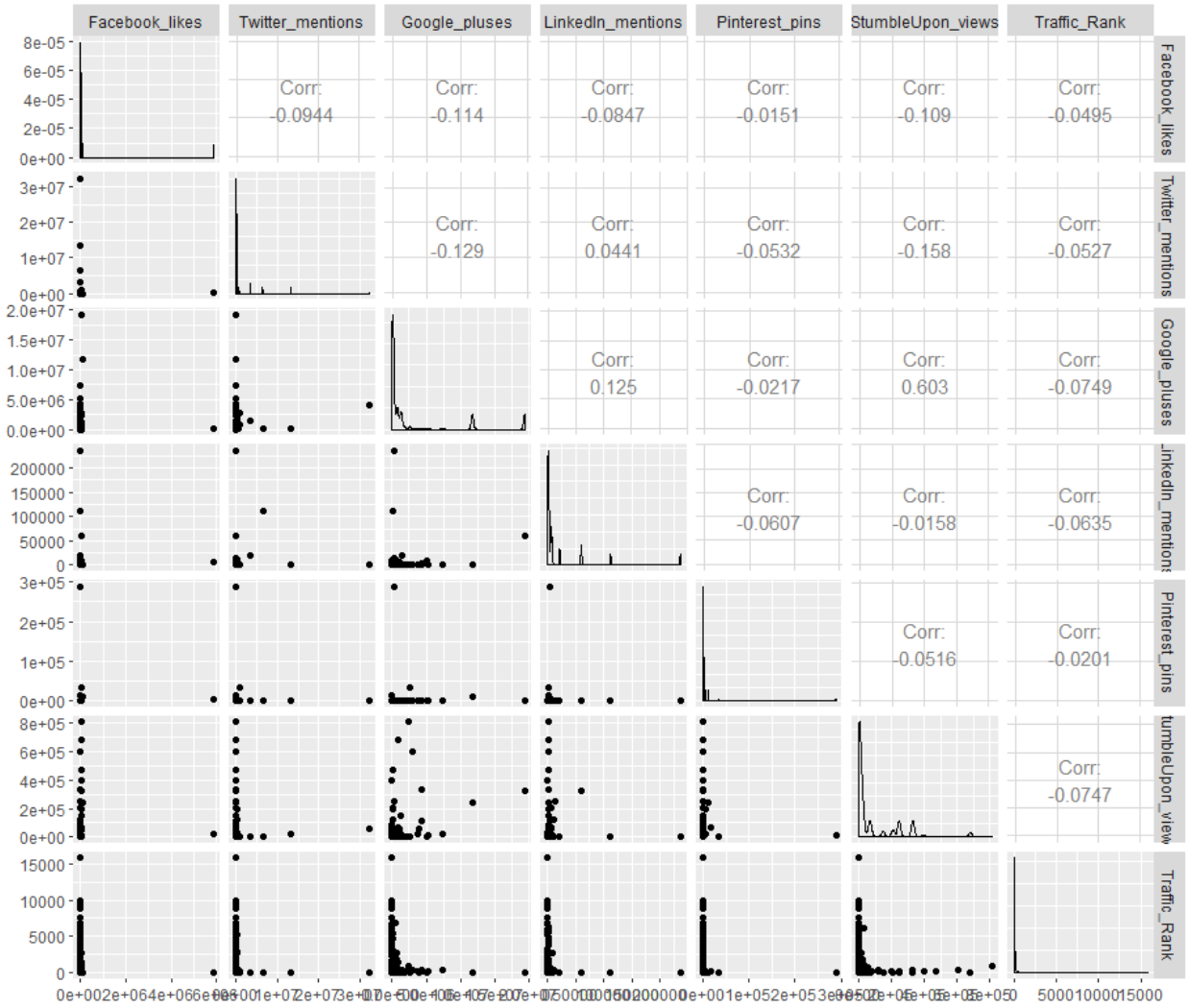
In the above code, we search in the data for “K”, “M” and according replace it by multiplying the rest of the value with “1,000” and “1,000,000”.

We can now observe the data as follows:

```
> dataSocial[3:8] %>% head()
  Facebook_likes Twitter_mentions Google_pluses LinkedIn_mentions Pinterest_pins StumbleUpon_views
2          94200             11200      11700000              1670          10800            246000
3          13500             16500      19300000             60000             47            329000
4        5870000             64400        127000              6230           4150             23100
5          17200              1110        798000              7500            433             68900
7              9              9370          7200              1120             16               136
8            476               162        126000              1240              5                 5
```

Visualizing a matrix plot for the above data using the ggpairs() function

> ggpairs(dataSocial, columns = 3:9)



From the last column in the above matrix we can observe the correlation coefficients for Traffic_Rank vs Social media data. These correlation coefficients can also be obtained as follows:

```
> cor(dataSocial[9],dataSocial[3:8])
             Facebook_likes Twitter_mentions Google_pluses LinkedIn_mentions Pinterest_pins StumbleUpon_views
Traffic_Rank    -0.04945581      -0.05272037   -0.07491811       -0.06353143     -0.0201417       -0.07473315
```

From the above correlation coefficients, we observe that the variable Traffic_Rank mostly does not rely or relate to the social media.

CONCLUSION

In our study, we performed various statistical analyses over the popular website dataset. We started off with establishing the probabilities of various occurrences by marginal and conditional probabilities for data pertaining to US and world. We also used Bayes' Theorem to determine conditional probabilities in cases of condition reversal using the previously computed probabilities.

Then we went ahead with the exploratory analysis of our data where we studied how our data is distributed for various fields containing data as factors. We also plotted the graphs to better visualize the data in form of factors for US. Following, we examined the normality of various records for different fields and using Q-Q plot in R, established whether the observations were normally distributed. We discovered few fields having its observation distributed normally while a few having skewed and not distributed normally.

We also wanted to verify Central Limit Theorem for our data. For that, we grouped our data by country hence generating 135 samples with 50 observations each and found mean value for each

sample. Then we plotted a Q-Q plot for the computed mean values for each sample and observed that the mean score was almost normally distributed hence proving the CLT.

In the next step, we conducted the correlation and covariance tests across various fields and observed both positive and negative correlation, covariance across different pairs of fields. This gave us insight on whether there is any relation between the observations of different fields.

Finally, we conducted a multivariate analysis of social media on the traffic ranking using the ggpairs() function in R. We observed that the variable Traffic_Rank mostly did not rely or relate to the social media.

In future, we hope to extend our analysis to advanced multivariate analyses and regression models where we could predict different parameters of a website by examining the past data. This would provide very valuable insights for the trends which websites follow and how different parameters affect each other.

REFERENCES

1. Popular Websites across the Globe. (2017, May 25). Retrieved from https://www.kaggle.com/bpali26/popular-websites-across-the-globe
2. ggpairs. Retrieved from https://www.rdocumentation.org/packages/GGally/versions/1.3.2/topics/ggpairs