STATISTICAL DATA, CALCULATIONS & RESULTS IN R PROGRAMMING

The steps performed in this project to obtain the calculations and derive inferential conclusions are as follows:

1. First, the datasets are loaded for Canada & US into a MS excel sheet and saved them in a .csv format. Now we import them into the RStudio program.
2. The sample distributions like mean and the standard deviation are calculated for the Views, Likes & the dislikes column of Canada dataset. Similar operation is done for the US dataset as well.
3. The Confidence Interval of the Difference of Means are obtained between all the 3 columns (Views, Likes & Dislikes) of Canada and US datasets.
4. By looking at the P-values of all three C.I. calculations, we can accept/reject the hypotheses tests.
5. The boxplots comparing the 3 columns for both the datasets individually are obtained graphically.
6. Now the correlations are studied between the Views & Likes (combined) and Likes & Dislikes (combined) for Canada & US datasets.
7. Finally, the linear regressions are found out between the same combinations (as mentioned in step 6) to infer how one data can influence the other.

DATASET FILE:-

For Canada: Viewing Trends in Canada 2016.csv

For US: Viewing Trends in the US 2016.csv

R Code for importing these datasets into RStudio:

setwd("C:\\Users\\skyeshwin\\Documents\\Course Material\\IE6200 Eng Probs & Stats SEC 07 - Fall 2017\\Proj & PPT")

dataCan<- read.csv("Viewing Trends in Canada 2016.csv")

dataUS <-read.csv("Viewing Trends in the US 2016.csv")

SAMPLE DISTRIBUTIONS FOR CANADA & US:-

For Canada & US datasets, we’re finding the mean and the standard deviations separately.

R code for finding the same:

viewsCan <- (dataCan[ ,2])

likesCan <- (dataCan[ ,3])

dislikesCan <- (dataCan[ ,4])

nrow(dataCan)

meanviewsCan <- mean(viewsCan)

meanlikesCan <- mean(likesCan)

meandislikesCan <- mean(dislikesCan)

stdevviewsCan <- sd(viewsCan)

stdevlikesCan <- sd(likesCan)

stdevdislikesCan <- sd(dislikesCan)

viewsUS <- (dataUS[ ,2])

likesUS <- (dataUS[ ,3])

dislikesUS <- (dataUS[ ,4])

nrow(dataUS)

meanviewsUS <- mean(viewsUS)

meanlikesUS <- mean(likesUS)

meandislikesUS <- mean(dislikesUS)

stdevviewsUS <- sd(viewsUS)

stdevlikesUS <- sd(likesUS)

stdevdislikesUS <- sd(dislikesUS)

The actual data obtained is stated below:

|  |  |  |
| --- | --- | --- |
| Parameter | For Canada | For US |
| No. of samples taken | 300 | 300 |
| Mean of Views | 1317834 | 868429.5 |
| Mean of Likes | 47839.47 | 34279.68 |
| Mean of Dislikes | 2931.83 | 1349.247 |
| Standard Deviation of Views | 3575581 | 2682614 |
| Standard Deviation of Likes | 156929.7 | 128340.3 |
| Standard Deviation of Dislikes | 15736.06 | 4480.233 |

C.I. LEVEL OF DIFFERENCE OF MEANS:-

Using the concept of unknown variances in the difference of means, we are calculating the Confidence Interval level with the help of t-Distribution values, taking the C.I. level default value as 95% (α – 0.05).

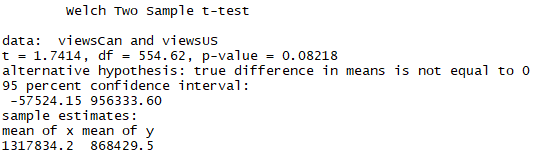
R Code for the same:

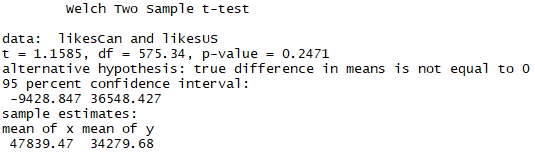
t.test(viewsCan, viewsUS, conf.level = 0.95)

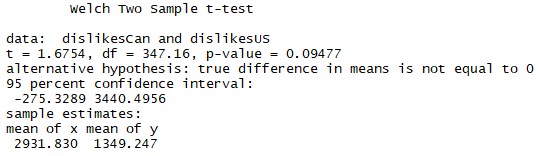
t.test(likesCan, likesUS, conf.level = 0.95)

t.test(dislikesCan, dislikesUS, conf.level = 0.95)

Complete result as shown in RStudio:







USING P-VALUES FOR TEST OF HYPOTHESES:-

The test of hypothesis between Canada & US is as follows:

H0 – More people in the US watched, liked and disliked youtube videos last year than in Canada.

H1 – Otherwise

Looking at the P-values (0.08218, 0.2471, 0.09477), since all are greater than 0.05, we can say that we can’t reject H0 hypothesis.

BOXPLOTS FOR VIEWS, LIKES, DISLIKES FOR BOTH DATASETS:-

We generate boxplots graphically comparing using the values of the views of Canada & US datasets. We obtain similar boxplots for likes and dislikes as well.

R code for generating boxplots:

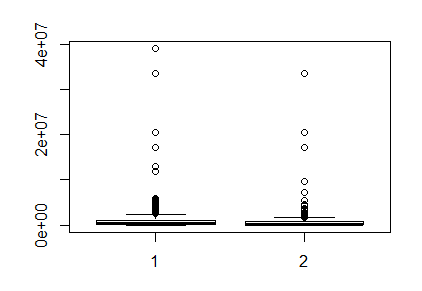
boxplot(viewsCan,viewsUS)

boxplot(likesCan,likesUS)

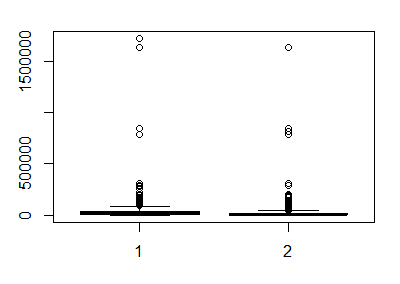
boxplot(dislikesCan,dislikesUS)

Result:

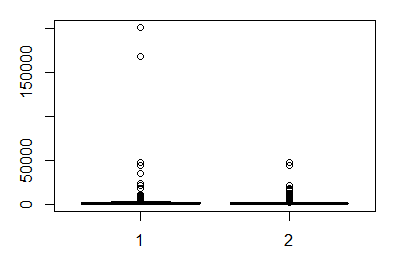
For Views in Canada & US:



For Likes in Canada & US:



For dislikes in Canada & US:



STUDY OF CORRELATIONS:-

Looking at the data for the both countries separately, it is clear that we can derive a relation between the views and the likes. A similar correlation can be done between likes and dislikes.

The correlation is represented mathematically in R as cor. The cor value ranges between 0 and 1. If the value is closer to 1, then the variables representing it are very much related to each other. The below result is obtained in Rstudio:

From the above result, it’s clearly inferred that views and likes are almost proportional in both the countries. In addition, the correlation value for likes and dislikes is high in the US. However, the same does not hold true for Canada.

LINEAR REGRESSION:-

The linear regression test analyzes the relationship between the dependent variable and a set of independent variables.

The below R code shows a list of operations related to it:

modCan1 <- lm(viewsCan ~ likesCan)

summary(modCan1)

plot(modCan1)

abline(modCan1, col=3)

modCan2 <- lm(likesCan ~ dislikesCan)

summary(modCan2)

plot(modCan2)

abline(modCan2, col=3)

modUS1 <- lm(viewsUS ~ likesUS)

summary(modUS1)

plot(modUS1)

abline(modUS1, col=5)

modUS2 <- lm(likesUS ~ dislikesUS)

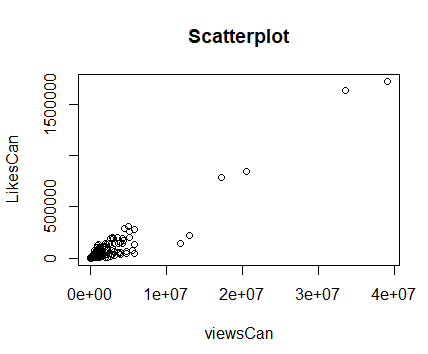
summary(modUS2)

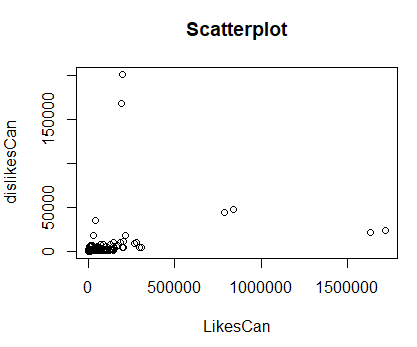
plot(modUS2)

abline(modUS2, col=5)

Since the list of graphs obtained in the above code are too subjective to understand, we draw a scatter plot of the same:

For Canada:





For US:-

