Probability Theory and Intro Statistics

ALY6010, WINTER 2022

Module 2 R-Practice

Week-2

Submitted by: Abhinav Jain

NUID: 002938209

Submitted To: Tom Beuer

Date: 03/7/2022

Module 2- R Practice

Dataset

Facebook advertisement analysis

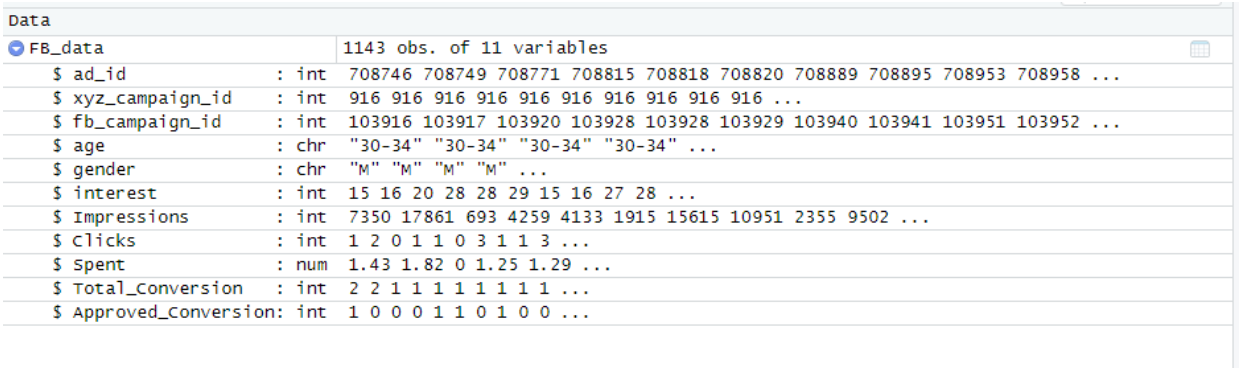
Introduction

When it comes to online advertising, Facebook is one of the major players, alongside Google's search and display networks. Facebook generates a profile of a user based on who they are and what they're interested in when they interact with the platform, adding demographic information, liking specific pages, and commenting on specific posts.

Facebook's advertising income was $26 billion in 2016, up from $17 billion the year before. This compared to Google's $79 billion, Twitter's $638 million in Q4 2016, and LinkedIn's $173 million in Q3 2016. These data demonstrate how large an advertising platform is, yet it faces future issues due to a drop in younger users in 2017, with generation Z migrating to Snapchat and Instagram.

In this dataset we have 1143 observations and 11 variables which will help in analyzing the Facebook advertising response through campaign in public. It contains ad_id is unique for each ad, xyz_campaign_id connected to xyz campaign company, Facebook_id track by campaign, age which show the person's age, gender give the specific male or female vote for the add, interest about the category person belongs with ad, impression ad shown number of times, clicks this show the number of click on ad, spent expenditure about the ad in campaign, total conversion – number of person approaches about the product after watching ad, approved conversion- customer bought the product after clicking the ad.

Task1: In this task imported dataset in R, which had 1143 observation and 11 variables.



	ad_id	xyz_campaign_id	fb_campaign_id	age	gender	interest	Impressions	Clicks	Spent	Total_Conversion	Approved_Conversion
1	708746	916	103916	30-34	M	15	7350	1	1.43	2	1
2	708749	916	103917	30-34	M	16	17861	2	1.82	2	0
3	708771	916	103920	30-34	M	20	693	0	0.00	1	0
4	708815	916	103928	30-34	M	28	4259	1	1.25	1	0
5	708818	916	103928	30-34	M	28	4133	1	1.29	1	1
6	708820	916	103929	30-34	M	29	1915	0	0.00	1	1
7	708889	916	103940	30-34	M	15	15615	3	4.77	1	0
8	708895	916	103941	30-34	M	16	10951	1	1.27	1	1
9	708953	916	103951	30-34	M	27	2355	1	1.50	1	0
10	708958	916	103952	30-34	M	28	9502	3	3.16	1	0
11	708979	916	103955	30-34	M	31	1224	0	0.00	1	0
12	709023	916	103962	30-34	M	7	735	0	0.00	1	0
13	709038	916	103965	30-34	M	16	5117	0	0.00	1	0
14	709040	916	103965	30-34	M	16	5120	0	0.00	1	0
15	709059	916	103968	30-34	M	20	14669	7	10.28	1	1
16	709105	916	103976	30-34	M	28	1241	0	0.00	1	1
17	709115	916	103978	30-34	M	30	2305	1	0.57	1	0
18	709124	916	103979	30-34	M	31	1024	0	0.00	1	1
19	709179	916	103988	35-39	M	15	4627	1	1.69	1	0
20	709183	916	103989	35-39	M	16	21026	4	4.63	2	1
21	709320	916	104012	35-39	M	15	1422	0	0.00	1	1
22	709323	916	104012	35-39	M	15	7132	2	2.61	1	0
23	709326	916	104013	35-39	M	16	12190	2	3.05	1	0
24	709327	916	104013	35-39	M	16	12193	2	3.06	1	1
25	709328	916	104013	35-39	M	16	3332	0	0.00	1	1

```
> KAG_data <-read.csv("C:\\Users\\abhin\\Downloads\\KAG_conversion_data.csv")
> KAG_data
    ad_id xyz_campaign_id fb_campaign_id   age gender interest Impressions Clicks Spent Total_Conversion Approved_Conversion
1  708746             916        103916 30-34      M       15        7350      1  1.43                2                   1
2  708749             916        103917 30-34      M       16       17861      2  1.82                2                   0
3  708771             916        103920 30-34      M       20         693      0  0.00                1                   0
4  708815             916        103928 30-34      M       28        4259      1  1.25                1                   0
5  708818             916        103928 30-34      M       28        4133      1  1.29                1                   1
6  708820             916        103929 30-34      M       29        1915      0  0.00                1                   1
7  708889             916        103940 30-34      M       15       15615      3  4.77                1                   0
8  708895             916        103941 30-34      M       16       10951      1  1.27                1                   1
9  708953             916        103951 30-34      M       27        2355      1  1.50                1                   0
10 708958             916        103952 30-34      M       28        9502      3  3.16                1                   0
11 708979             916        103955 30-34      M       31        1224      0  0.00                1                   0
12 709023             916        103962 30-34      M        7         735      0  0.00                1                   0
13 709038             916        103965 30-34      M       16        5117      0  0.00                1                   0
```

Task2: In this task, summaries the data variables which contain mean, median, mode, minimum value, maximum value, length, quartile range.

```
> summary(FB_data)
     ad_id          xyz_campaign_id fb_campaign_id      age                gender             interest
 Min.   : 708746   Min.   : 916    Min.   :103916   Length:1143        Length:1143        Min.   :  2.00
 1st Qu.: 777633   1st Qu.: 936    1st Qu.:115716   Class :character   Class :character   1st Qu.: 16.00
 Median :1121185   Median :1178    Median :144549   Mode  :character   Mode  :character   Median : 25.00
 Mean   : 987261   Mean   :1067    Mean   :133784                                         Mean   : 32.77
 3rd Qu.:1121805   3rd Qu.:1178    3rd Qu.:144658                                         3rd Qu.: 31.00
 Max.   :1314415   Max.   :1178    Max.   :179982                                         Max.   :114.00
  Impressions          Clicks            Spent         Total_Conversion Approved_Conversion
 Min.   :     87   Min.   :  0.00   Min.   :  0.00   Min.   : 0.000   Min.   : 0.000
 1st Qu.:   6504   1st Qu.:  1.00   1st Qu.:  1.48   1st Qu.: 1.000   1st Qu.: 0.000
 Median :  51509   Median :  8.00   Median : 12.37   Median : 1.000   Median : 1.000
 Mean   : 186732   Mean   : 33.39   Mean   : 51.36   Mean   : 2.856   Mean   : 0.944
 3rd Qu.: 221769   3rd Qu.: 37.50   3rd Qu.: 60.02   3rd Qu.: 3.000   3rd Qu.: 1.000
 Max.   :3052003   Max.   :421.00   Max.   :639.95   Max.   :60.000   Max.   :21.000
>
```

Task 3: This task is to see the columns of the dataset and display variable type

```
> glimpse(FB_data)
Rows: 1,143
Columns: 11
$ ad_id               <int> 708746, 708749, 708771, 708815, 708818, 708820, 708889, 708895, 708953, 708958, 708~
$ xyz_campaign_id     <int> 916, 916, 916, 916, 916, 916, 916, 916, 916, 916, 916, 916, 916, 916, 916, 916, 916~
$ fb_campaign_id      <int> 103916, 103917, 103920, 103928, 103928, 103929, 103940, 103941, 103951, 103952, 103~
$ age                 <chr> "30-34", "30-34", "30-34", "30-34", "30-34", "30-34", "30-34", "30-34", "30-34", "3~
$ gender              <chr> "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M"~
$ interest            <int> 15, 16, 20, 28, 28, 29, 15, 16, 27, 28, 31, 7, 16, 16, 20, 28, 30, 31, 15, 16, 15, ~
$ Impressions         <int> 7350, 17861, 693, 4259, 4133, 1915, 15615, 10951, 2355, 9502, 1224, 735, 5117, 5120~
$ Clicks              <int> 1, 2, 0, 1, 1, 0, 3, 1, 1, 3, 0, 0, 0, 0, 7, 0, 1, 0, 1, 4, 0, 2, 2, 2, 0, 0, 2, 4,~
$ Spent               <dbl> 1.43, 1.82, 0.00, 1.25, 1.29, 0.00, 4.77, 1.27, 1.50, 3.16, 0.00, 0.00, 0.00, 0.00,~
$ Total_Conversion    <int> 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1,~
$ Approved_Conversion <int> 1, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 1, 0, 1, 0,~
```

Task 4: In this task clean the dataset, particularly age and gender are the main variables on which some corrections need to be done.

Age

```
> unique(FB_data$age)
[1] "30-34" "35-39" "40-44" "45-49"
> FB_clean <- FB_data
> FB_clean$age [FB_clean$age =="30-34"] <- 32
> FB_clean$age [FB_clean$age =="35-39"] <- 37
> FB_clean$age [FB_clean$age =="40-44"] <- 42
> FB_clean$age [FB_clean$age =="45-49"] <- 47
> FB_clean$age <- as.integer(FB_clean$age)
> unique(FB_clean$age)
[1] 32 37 42 47
```

Gender

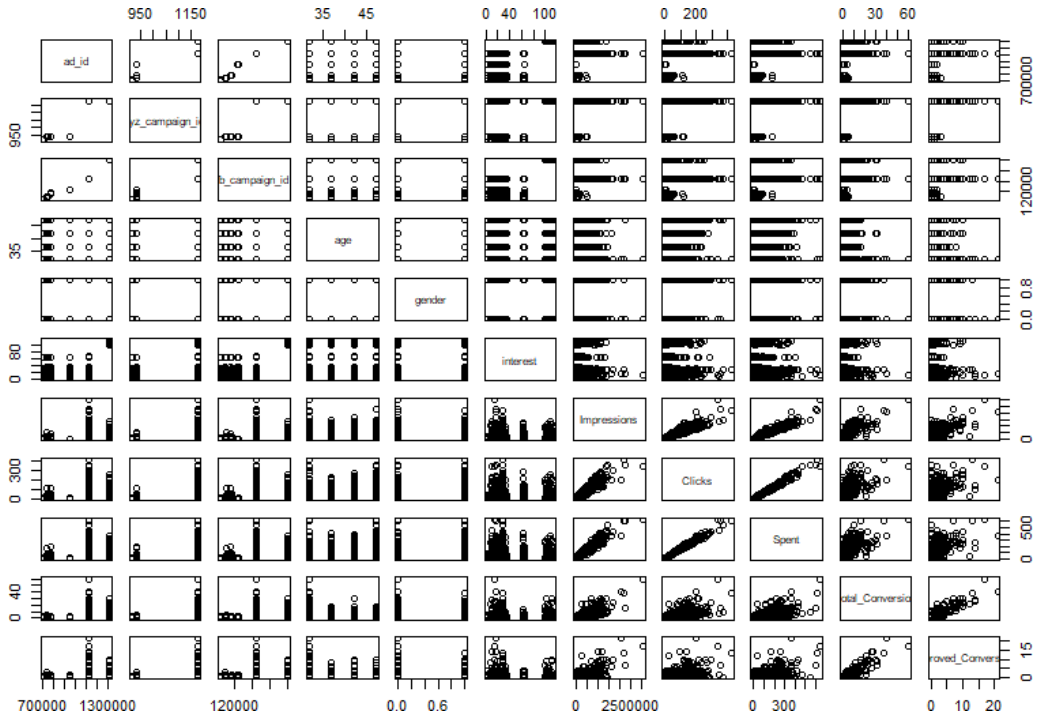
```
> FB_clean$gender[FB_clean$gender == 'M'] <- 0
> FB_clean$gender[FB_clean$gender == 'F'] <- 1
> FB_clean$gender<- as.integer(FB_clean$gender)
> unique(FB_clean$gender)
[1] 0 1
> str(FB_clean$gender)
 int [1:1143] 0 0 0 0 0 0 0 0 0 0 ...
> describe_all <- describe(FB_clean)
> view(describe_all)
> str(FB_clean)
'data.frame':   1143 obs. of  11 variables:
 $ ad_id               : int  708746 708749 708771 708815 708818 708820 708889 708895 708953 708958 ...
 $ xyz_campaign_id     : int  916 916 916 916 916 916 916 916 916 916 ...
 $ fb_campaign_id      : int  103916 103917 103920 103928 103928 103929 103940 103941 103951 103952 ...
 $ age                 : int  32 32 32 32 32 32 32 32 32 32 ...
 $ gender              : int  0 0 0 0 0 0 0 0 0 0 ...
 $ interest            : int  15 16 20 28 28 29 15 16 27 28 ...
 $ Impressions         : int  7350 17861 693 4259 4133 1915 15615 10951 2355 9502 ...
 $ Clicks              : int  1 2 0 1 1 0 3 1 1 3 ...
 $ Spent               : num  1.43 1.82 0 1.25 1.29 ...
 $ Total_Conversion    : int  2 2 1 1 1 1 1 1 1 1 ...
 $ Approved_Conversion: int  1 0 0 0 1 1 0 1 0 0 ...
```

Task 5 : Produce several descriptive statistics tables

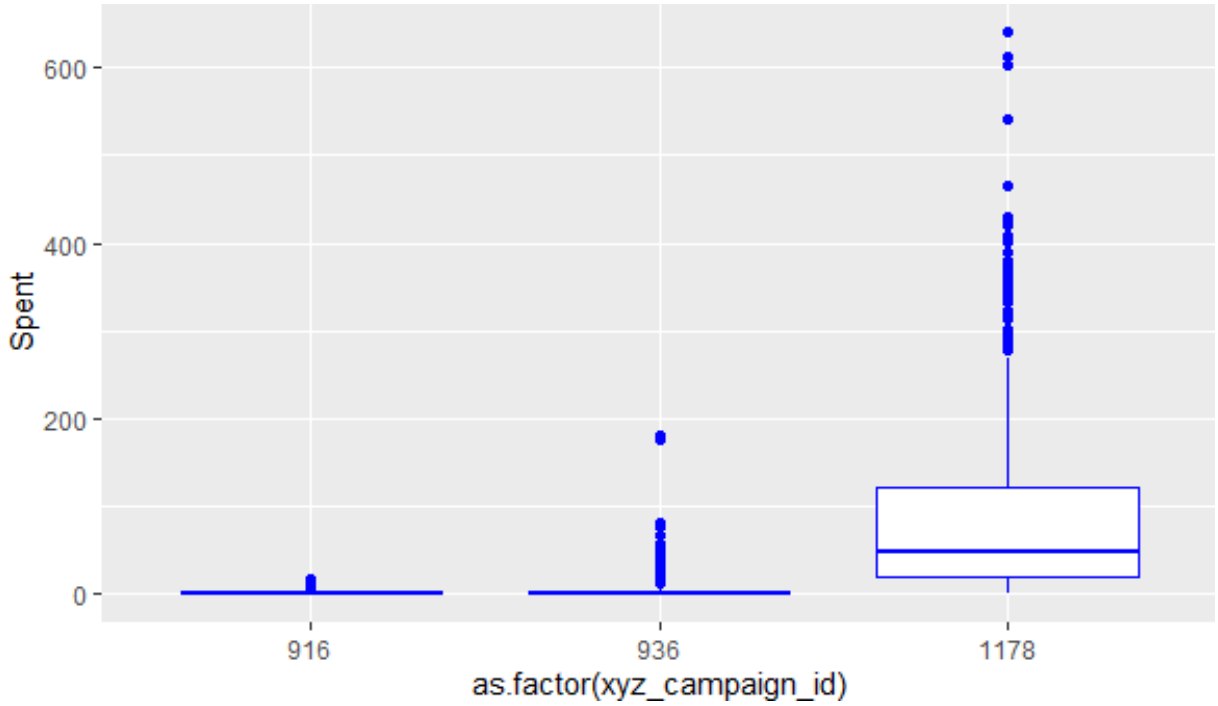
	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
ad_id	1	1143	987261.1303587	193992.8147382	1121185.00	983417.1737705	252016.79580	708746	1314415.00	605669.00	-0.10251519	-1.4142114	5738.02623126
xyz_campaign_id	2	1143	1067.3823272	121.6293929	1178.00	1071.1497268	0.00000	916	1178.00	262.00	-0.19139889	-1.9598703	3.59762112
fb_campaign_id	3	1143	133783.9895013	20500.3086219	144549.00	132157.3038251	31018.95720	103916	179982.00	76066.00	0.52110316	-0.2548283	606.36941002
age	4	1143	38.3210849	5.9038681	37.00	38.0273224	7.41300	32	47.00	15.00	0.30443687	-1.4237921	0.17462786
gender	5	1143	0.4820647	0.4998969	0.00	0.4775956	0.00000	0	1.00	1.00	0.07169304	-1.9966046	0.01478623
interest	6	1143	32.7664042	26.9521310	25.00	27.5114754	10.37820	2	114.00	112.00	1.76163972	2.2023391	0.79720496
Impressions	7	1143	186732.1329834	312762.1832082	51509.00	112606.6819672	74063.28300	87	3052003.00	3051916.00	3.00228887	13.0334127	9251.05197230
Clicks	8	1143	33.3902012	56.8924383	8.00	19.4338798	11.86080	0	421.00	421.00	2.70507297	8.4769813	1.68279585
Spent	9	1143	51.3606561	86.9084179	12.37	30.2847322	18.33976	0	639.95	639.95	2.70176128	8.7794575	2.57062501
Total_Conversion	10	1143	2.8556430	4.4835935	1.00	1.7978142	0.00000	0	60.00	60.00	5.08255157	38.3429148	0.13261820
Approved_Conversion	11	1143	0.9440070	1.7377080	1.00	0.5868852	1.48260	0	21.00	21.00	4.82484988	34.3715387	0.05139888

In the above table, data from the facebook ad compaign shows the variance defined with (vars), n define with number of observations, an average of the different variables, standard deviation, a median of each variable, mad, minimum, maximum, range, skew, kurtosis are the variable measures. There are two types of descriptive statistics measure of central tendency and measures of variability.

Task 6 : ScatterPlot: As one might assume, there are substantial correlations between the amount we spent and the number of impressions and clicks we received, but less so between our spend, clicks, and impressions, and our conversions. We could go on to quantify the importance of these correlations if we wanted to at this point, but for now, let's focus on a specific campaign and go a little more precise.

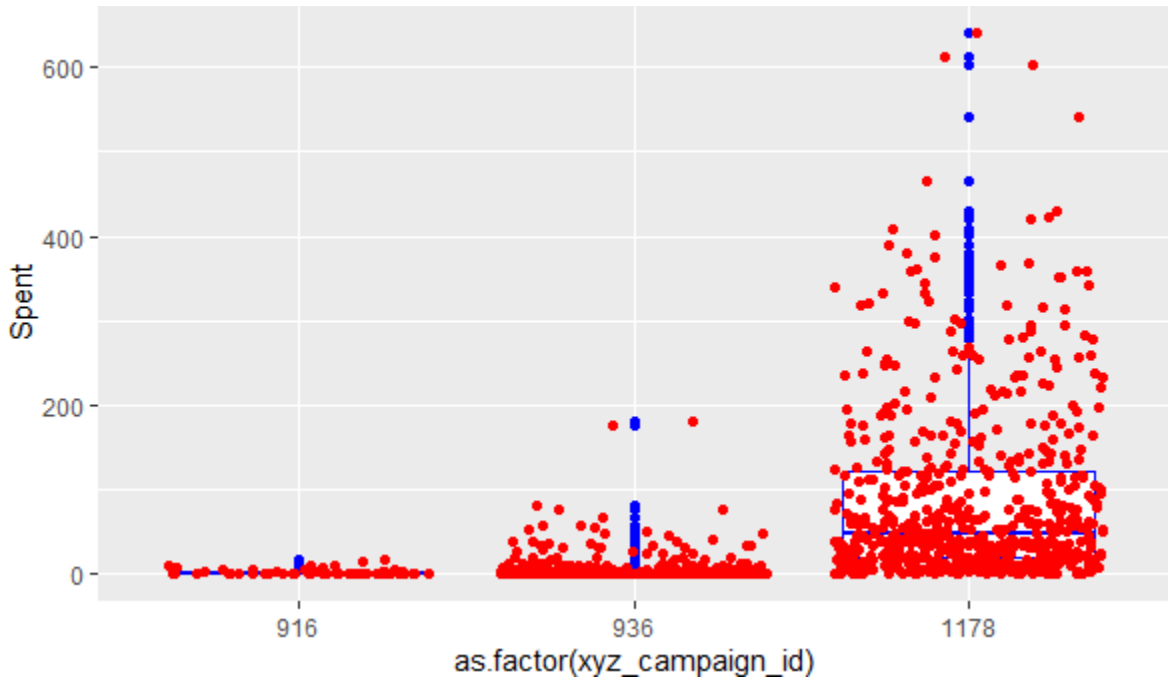


Task 7 Box Plot: After detecting the outlier of the xyz_campaign_id we can see above spent we can see the outlier on the boxplot graph.

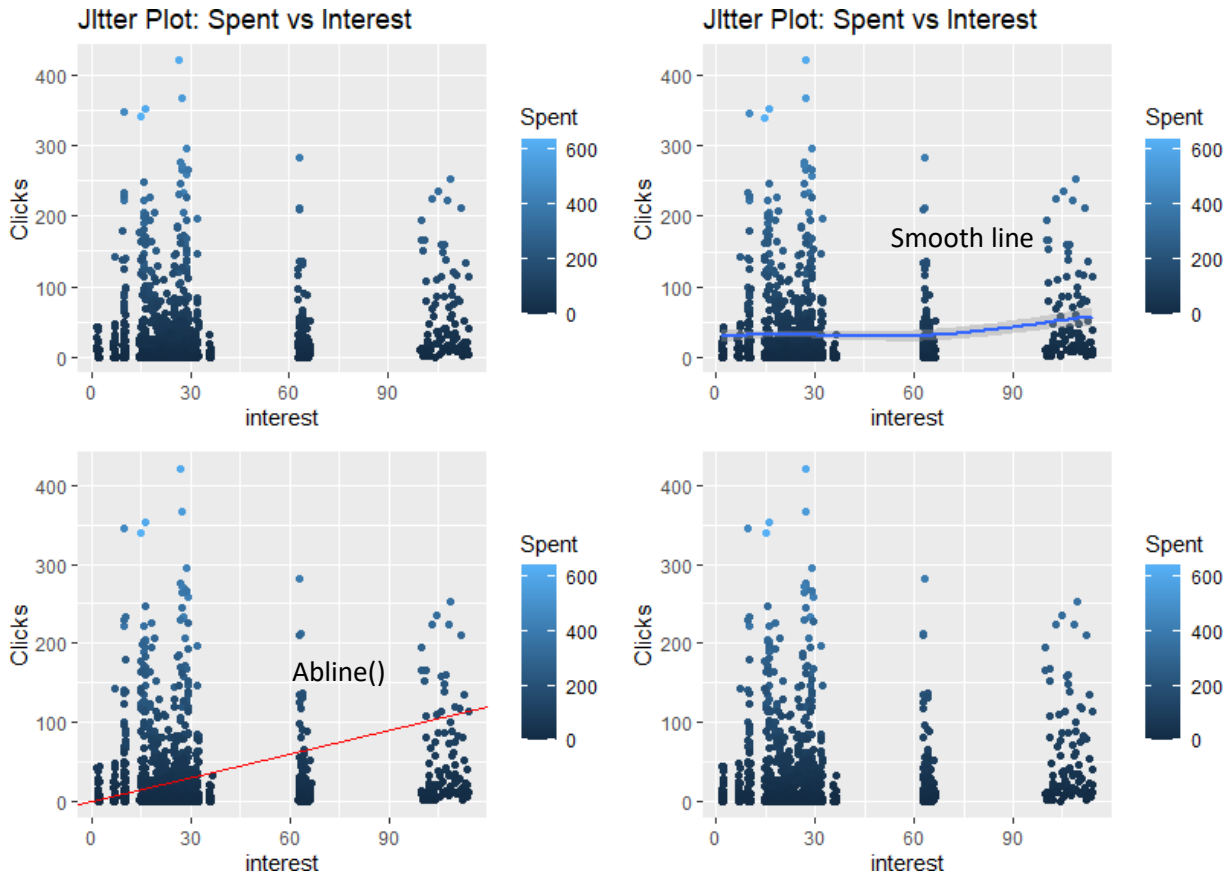


After Removing Outliers from the xyz_campaign_id =1178

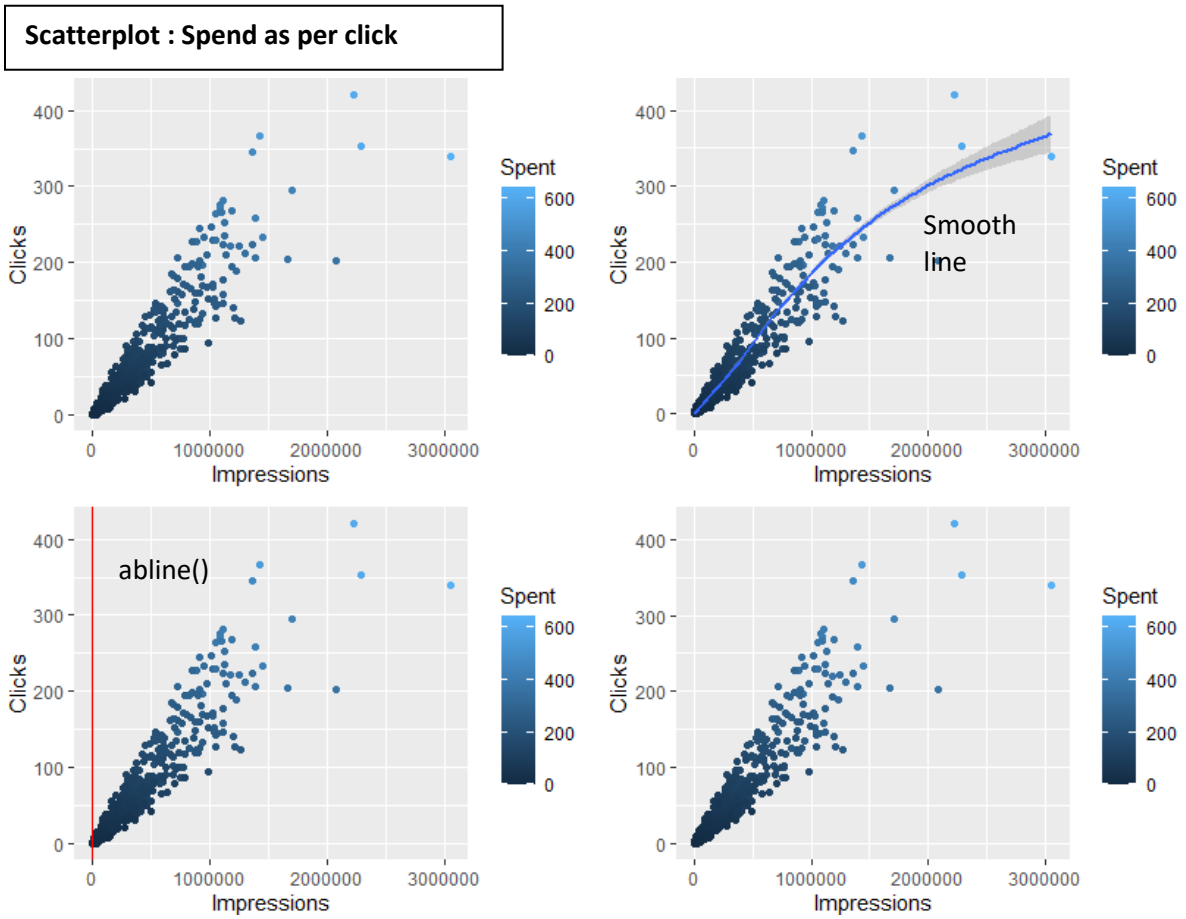
```
> ggplot( FB_clean, aes(as.factor(xyz_campaign_id), Spent))+
+   geom_boxplot(color = "Blue")+
+   geom_jitter(color = 'red')
```



Task 8: Jitter Plot: In the numeric vector we can a noise to the vector, It is a random variation that we can add to each point while handling the overplotting caused by smaller



Task 9: Scatterplot with abline() and par(): In this task, we are comparing the click of the advertisement with the impression as per the spent with abline and a smooth line to



Summary:

If you are new to pay-per-click advertising or have been seeking new ways to enhance ROI from your digital campaigns, maybe this data analysis has been helpful. This analysis is only a taste of the types of studies you can perform with your digital advertising datasets, but it's simply a starting point: the right forms of analysis and success metrics will be determined by your business model and underlying marketing goals. The types of analyses you want to run on your own data will be determined by your campaign goals, the data you have, and the decisions you want to be able to make with the information you've gathered. We look at this dataset from the perspective of exploratory data analysis, utilizing tools you could apply to your own data because we don't know what it's about. This dataset helps in running many businesses around the world.

References:

[1] Ignore Outliers in ggplot2 Boxplot in R (Example) - YouTube

How to remove outliers from ggplot2 boxplots in the R programming language. More information: https://statisticsglobe.com/ignore-outliers-in-ggplot2-boxplot-...

https://www.youtube.com/watch?v=QvdHb23t_8c

[2] par Function in R (3 Examples) | How to ... - youtube.com

How to set or query graphical parameters using the par function in the R programming language. More details: https://statisticsglobe.com/par-function-in-r/R ...

https://www.youtube.com/watch?v=B9KTX4X0V5U

[3] ggplot basics, creating scatterplot in ... - youtube.com

In this video,You will learn the basics of ggplot and different variations of scatterplot.1. a basic scatterplot of two numerical variables2. a scatterplot s...

https://www.youtube.com/watch?v=kaW6FmIcnkk

[4] YouTube

Enjoy the videos and music you love, upload original content, and share it all with friends, family, and the world on YouTube.

https://www.youtube.com/?gl=NL

Appendix:

```r
install.packages("ggplot2")
install.packages("tidyverse")
install.packages("dplyr")
install.packages("modeest")
install.packages("tableone")
install.packages("plyr")
install.packages("epiDisplay")
install.packages("gmodels")
install.packages("gridExtra")



library(ggplot2)
library(tidyverse)
library(dplyr)
library(tidyverse)
library(dplyr)
library(plyr)
library(psych)
library(epiDisplay)
library(gmodels)
library(modeest)
library(tableone)
library(DataExplorer)
library(gridExtra)

FB_data <-read.csv("C:\\Users\\abhin\\Downloads\\FB_conversion_data.csv")
FB_data

str(FB_data)
summary(FB_data)
glimpse(FB_data)


unique(FB_data$age)

FB_clean <- FB_data

FB_clean$age [FB_clean$age =="30-34"] <- 32
FB_clean$age [FB_clean$age =="35-39"] <- 37
FB_clean$age [FB_clean$age =="40-44"] <- 42
FB_clean$age [FB_clean$age =="45-49"] <- 47

FB_clean$age <- as.integer(FB_clean$age)

unique(FB_clean$age)

summary(FB_clean)

FB_clean$gender[FB_clean$gender == 'M'] <- 0
FB_clean$gender[FB_clean$gender == 'F'] <- 1

FB_clean$gender<- as.integer(FB_clean$gender)

unique(FB_clean$gender)
```

```
un1que(rB_c1ean$aged

summary(KB_c1 ean}

KB_c1ean$gender [rB_c1eanlgender        'M'   <- 0
KB_c1ean$gender [rB_c1eanlgender        'r'   <- 1

KB_c1ean$gender<- as . 1nteger {rB_c1 ean$gender

unique(FB_clean$gender)
s*r(rB_cleanJ

optl ons (sc? pen = 9$
FB_clean


describe_all    describe(rB_clean)

v1 ew(desertbe_a11
####P Sa ttg>1ot


ggpl ot (rB_c1ean}
ggp1  <-ggp1ot (KB_c1 ean , aes Cx=  Impr essl ons , y    c11cks ,  color    spent}}+
   geors_po1 nt (}

ggp2<-ggp1ot (KB_c1ean , aes Cx= Impr essl ons , y    c11cks ,  color    spent}}+
   geors_po1 nt (}+
   geors_smooth (}

ggp3<-ggp1 ot (KB_c1 ean , aes  Cx=  Impr essl ons,  y    c11cks ,  color    spent}}+
   geom_point(J+
   geom_abline(color = "red")


ggp4<-ggp1 ot (KB_c1 ean ,  aes Cx= Impr essl ons ,  y    c11cks ,  color    spent}}+
   geors_po1 nt (}

#grid.arrange(ggpl, ggp2, ggg3, ggp4, ncol =2, nrow =2)



   c   p on  between  Int er est  and  spent  use ng  georrt_ tier C)
ggpl ot (rB_c1ean}
ggp 5 <- ggp1at (KB_c1 ean , aes {x= 1 nter est , y - c11cks , color = spent)}+
   geors_j Otter (}+
   1 abs (t 1t1 e = " 3Itt er P1ot : spent vs Interest "}

ggp6 <- ggp1at (KB_c1 ean , aes {x= 1 nter est , y - c11cks , color = spent)}+
```

```
* aom      on between Interest and spenc using geom_ i te ()
ggplot(re_clem)
ggp5 <- ggploi(rB clean, aes(x= imerest, y = Clicks, color = Spent))+
   geof¥_jOtter {}+
   1 abs {tlt1e = "3IEE er P1ot : spent vs Interest"}

ggp6 <- ggpl ot (FB_c1ean, aes {x- 1merest , y - c1lcks , co1or - spent} }+
   geofa_j1tter {}+
   geof¥_seooth(}+
   1abs {t1t1e - "3Itt er P1ot: spent vs Interest wlth  s«ooth{} "}

ggp7<-ggp1ot (rB_c1ean, aes (x= Inter est, y - c1lcks , co1or - spent} }+
   geof¥_jltter {}+
   geon_ab11ne (co1or = "Red"}
   1abs {t1t1e - "3Itt er P1ot: spent vs Irrcerest w1th ab1lne {} " }

ggp8<-ggp1oE (FB_c1ean,   aes (x=   Int  er esE ,   y  =  C11cks,   co1or      Sperrc} }+
   geofit_j1tt er {}
   geofa_count {}


grld. arrange (ggp1 , ggp2 , ggp3, ggp4 , nco1 =2, nrow -2 }


ggp9 <- ggplot ra_c1ean , aes {as. factor (xyz_caapa1gn_1d} , spent , mat n - "xyz_C acpal gn vs spent " } }+
   geom_boxploz(color  = "Blue") +
   geofX_j ltter (co1or -  ' red' }

ggp10 <- ggp1ot ( rB_c1ean, aes {as. factor {xyz_caapa1gn_1d} , spent , maln - "xyz_campa1 gn vs  spent"}}+
   geof¥_boxp1 ot (co1or - "B1ue"}+
   geofa_j1tter {co1or = ' red'}+
   coord_cartes1an(y1im =quant11e{re_c1ean5sperrc , c(0. 1, 0. 9} }}+
   1abs {t1t1e - "aox P1ot: spent w1thout out11ers"}
```