Exploratory Data Ananlysis

Code

Exploratory Data Analysis on Multivatiate Dataset.

Steps Involved:

1. Loading and Reading the dataset
2. Insights about dataset

a. Structure
b. Dimensions
c. Data types of predicting variables
d. Summary of the dataset
e. Removing duplicate columns such as Region

3. Data Cleansing

a. Solved mapping issues between variable and its corresponding indicators using Excel.
b. Converted numeric variables with NA's to 0. c)Computed the summary of the new dataset.

4. Data Visualization

a. Plotted Box Plots and Strip charts to understand the data distribution and to detect outliers. Skewness in data and outliers were observed.
b. Plotted Scatter Plot Matrix using GGally library to understand the correlation between other variables as well as CPC.

–Loading and reading the dataset

Hide

```
Data<-read.csv('D:/Rutgers Study Material/MultivariateData1.csv')
# top 5 columns of the dataset
head(Data)
```

	ï..Region <fctr>	Region_Indicators <int>	City <fctr>	City_indicators <int>	SupplyVendor <fctr>	SupplyVendors_Indi
1	Hawaii	60	'Aiea	2208	beanstock	
2	Hawaii	60	'Aiea	2208	brightroll	
3	Hawaii	60	'Aiea	2208	brightroll	
4	Hawaii	60	'Aiea	2208	brightroll	
5	Hawaii	60	'Aiea	2208	brightroll	
6	Hawaii	60	'Aiea	2208	brightroll	

6 rows | 1-7 of 23 columns

Hide

```
#Names of the columns in dataset
names(Data)
```

```
 [1] "ï..Region"                 "Region_Indicators"
 [3] "City"                      "City_indicators"
 [5] "SupplyVendor"              "SupplyVendors_Indicators"
 [7] "OS"                        "OS_Indicators"
 [9] "Browser"                   "Browser_Indicators"
[11] "DeviceType"                "DeviceType_Indicators"
[13] "Impression_Day"            "Impression_Time"
[15] "Impressions"               "Clicks"
[17] "CTR"                       "CPC"
[19] "VCR"                       "CPV"
[21] "Completes"                 "Total_Spend"
[23] "CPCV"
```

There are 23 columns present.

Hide

```
#Structure of the data
str(Data)
```

```
'data.frame':	472666 obs. of  23 variables:
 $ ï..Region              : Factor w/ 27 levels "Alabama","Alaska",..: 13 13 13 13 13 13 13 13 13 13 ...
 $ Region_Indicators      : int  60 60 60 60 60 60 60 60 60 60 ...
 $ City                   : Factor w/ 5146 levels "'Aiea","'Ewa Beach",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ City_indicators        : int  2208 2208 2208 2208 2208 2208 2208 2208 2208 2208 ...
 $ SupplyVendor           : Factor w/ 23 levels "adaptv","adconductor",..: 5 7 7 7 7 7 7 7 7 7 ...
 $ SupplyVendors_Indicators: int  3 4 4 4 4 4 4 4 4 4 ...
 $ OS                     : Factor w/ 26 levels "Android23","Android40",..: 21 3 4 4 6 7 7 7 7 8 ...
 $ OS_Indicators          : int  1 8 9 9 11 12 12 12 12 13 ...
 $ Browser                : Factor w/ 12 levels "Chrome","Edge",..: 11 1 1 1 1 1 1 1 1 1 ...
 $ Browser_Indicators     : int  4 1 1 1 1 1 1 1 1 1 ...
 $ DeviceType             : Factor w/ 4 levels "Mobile","PC",..: 4 4 4 4 4 4 4 4 4 4 ...
 $ DeviceType_Indicators  : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Impression_Day         : int  1 3 5 5 4 4 5 5 5 4 ...
 $ Impression_Time        : Factor w/ 84836 levels "0:00:04","0:00:07",..: 83360 3857 75392 75506 39916 73326 8081 8637 14189 74577 ...
 $ Impressions            : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Clicks                 : int  0 0 0 0 0 0 0 0 0 0 ...
 $ CTR                    : num  0 0 0 0 0 0 0 0 0 0 ...
 $ CPC                    : num  NA NA NA NA NA NA NA NA NA NA ...
 $ VCR                    : num  0 1 1 0 1 1 1 0 0 1 ...
 $ CPV                    : num  0.01271 0.00997 0.01128 0.00941 0.01025 ...
 $ Completes              : int  0 1 1 0 1 1 1 0 0 1 ...
 $ Total_Spend            : num  0.01271 0.00997 0.01128 0.00941 0.01025 ...
 $ CPCV                   : num  NA 0.00997 0.01128 NA 0.01025 ...
```

There are 7 factors or categorical variables. Lots of NA’s or missing values were oobserved.

Hide

```
# Dimension of the data
dim(Data)
```

```
[1] 472666     23
```

Hide

```
# Removing the extra columns
drops <- c("ï..Region","City","SupplyVendor","OS","Browser","DeviceType","Impression_Time")
New_data<-Data[ , !(names(Data) %in% drops)]
dim(New_data)
```

```
[1] 472666     16
```

Since we have column names as well as their indicators, it’s always better to remove redundant information.

Hide

```
# Checking the datatype of each column
attach(New_data)
```

```
The following objects are masked from New_data (pos = 3):

    Browser_Indicators, City_indicators, Clicks,
    Completes, CPC, CPCV, CPV, CTR, DeviceType_Indicators,
    Impression_Day, Impressions, OS_Indicators,
    Region_Indicators, SupplyVendors_Indicators,
    Total_Spend, VCR

The following objects are masked from New_data (pos = 4):

    Browser_Indicators, City_indicators, Clicks,
    Completes, CPC, CPCV, CPV, CTR, DeviceType_Indicators,
    Impression_Day, Impressions, OS_Indicators,
    Region_Indicators, SupplyVendors_Indicators,
    Total_Spend, VCR

The following objects are masked from New_data (pos = 5):

    Browser_Indicators, City_indicators, Clicks,
    Completes, CPC, CPCV, CPV, CTR, DeviceType_Indicators,
    Impression_Day, Impressions, OS_Indicators,
    Region_Indicators, SupplyVendors_Indicators,
    Total_Spend, VCR

The following objects are masked from New_data (pos = 6):

    Browser_Indicators, City_indicators, Clicks,
    Completes, CPC, CPCV, CPV, CTR, DeviceType_Indicators,
    Impression_Day, Impressions, OS_Indicators,
    Region_Indicators, SupplyVendors_Indicators,
    Total_Spend, VCR

The following objects are masked from New_data (pos = 7):

    Browser_Indicators, City_indicators, Clicks,
    Completes, CPC, CPCV, CPV, CTR, DeviceType_Indicators,
    Impression_Day, Impressions, OS_Indicators,
    Region_Indicators, SupplyVendors_Indicators,
    Total_Spend, VCR

The following objects are masked from New_data (pos = 10):

    Browser_Indicators, City_indicators, Clicks,
    Completes, CPC, CPCV, CPV, CTR, DeviceType_Indicators,
    Impression_Day, Impressions, OS_Indicators,
    Region_Indicators, SupplyVendors_Indicators,
    Total_Spend, VCR

The following objects are masked from New_data (pos = 11):

    Browser_Indicators, City_indicators, Clicks,
    Completes, CPC, CPCV, CPV, CTR, DeviceType_Indicators,
    Impression_Day, Impressions, OS_Indicators,
```

```
    Region_Indicators, SupplyVendors_Indicators,
    Total_Spend, VCR
```

Hide

```
class(Region_Indicators)
```

```
[1] "integer"
```

Hide

```
class(City_indicators)
```

```
[1] "integer"
```

Hide

```
class(SupplyVendors_Indicators)
```

```
[1] "integer"
```

Hide

```
class(OS_Indicators)
```

```
[1] "integer"
```

Hide

```
class(Browser_Indicators)
```

```
[1] "integer"
```

Hide

```
class(DeviceType_Indicators)
```

```
[1] "integer"
```

Hide

```
class(Impression_Day)
```

```
[1] "integer"
```

Hide

```
class(Impressions)
```

```
[1] "integer"
```

Hide

```
class(Clicks)
```

```
[1] "integer"
```

Hide

```
class(CTR)
```

```
[1] "numeric"
```

Hide

```
class(CPC)
```

```
[1] "numeric"
```

Hide

```
class(VCR)
```

```
[1] "numeric"
```

Hide

```
class(CPV)
```

```
[1] "numeric"
```

Hide

```
class(Completes)
```

```
[1] "integer"
```

Hide

```
class(Total_Spend)
```

```
[1] "numeric"
```

Hide

```
class(CPCV)
```

```
[1] "numeric"
```

Hide

```
# Analyzing missing values
sapply(New_data,function(x) sum(is.na(x)))
```

```
       Region_Indicators         City_indicators 
                       0                       0 
SupplyVendors_Indicators           OS_Indicators 
                       0                       0 
      Browser_Indicators    DeviceType_Indicators 
                       0                       0 
          Impression_Day             Impressions 
                       0                       0 
                  Clicks                     CTR 
                       0                       0 
                     CPC                     VCR 
                  469436                       0 
                     CPV               Completes 
                   13443                       0 
             Total_Spend                    CPCV 
                       0                  190689 
```

A lot of indicators were missing from the data. On analyzing the file “VLookUP” was not working properly. Steps Taken: Mapped the indicators using excel.

Hide

```
summary(New_data)
```

```
Region_Indicators City_indicators SupplyVendors_Indicators
Min.   : 2.00     Min.   :   1    Min.   : 1.000
1st Qu.: 3.00     1st Qu.: 946    1st Qu.: 4.000
Median :31.00     Median :1672    Median : 4.000
Mean   :24.09     Mean   :1895    Mean   : 8.146
3rd Qu.:39.00     3rd Qu.:2590    3rd Qu.:15.000
Max.   :60.00     Max.   :5151    Max.   :23.000

OS_Indicators    Browser_Indicators DeviceType_Indicators
Min.   : 1.000   Min.   : 1.000     Min.   :1.000
1st Qu.: 1.000   1st Qu.: 1.000     1st Qu.:1.000
Median : 3.000   Median : 1.000     Median :2.000
Mean   : 4.703   Mean   : 2.801     Mean   :1.883
3rd Qu.: 5.000   3rd Qu.: 4.000     3rd Qu.:2.000
Max.   :25.000   Max.   :12.000     Max.   :4.000

Impression_Day   Impressions          Clicks
Min.   :1.000   Min.   :  1.000   Min.   :0.000000
1st Qu.:2.000   1st Qu.:  1.000   1st Qu.:0.000000
Median :4.000   Median :  1.000   Median :0.000000
Mean   :4.081   Mean   :  1.055   Mean   :0.006846
3rd Qu.:6.000   3rd Qu.:  1.000   3rd Qu.:0.000000
Max.   :7.000   Max.   :120.000   Max.   :2.000000

     CTR               CPC              VCR
Min.   :0.00000   Min.   :0        Min.   :0.0000
1st Qu.:0.00000   1st Qu.:0        1st Qu.:0.0000
Median :0.00000   Median :0        Median :1.0000
Mean   :0.00645   Mean   :0        Mean   :0.5925
3rd Qu.:0.00000   3rd Qu.:0        3rd Qu.:1.0000
Max.   :1.00000   Max.   :0        Max.   :1.0000
                  NA's   :469436
     CPV          Completes          Total_Spend
Min.   :0.000   Min.   :  0.0000   Min.   :0.000000
1st Qu.:0.010   1st Qu.:  0.0000   1st Qu.:0.009459
Median :0.011   Median :  1.0000   Median :0.011098
Mean   :0.012   Mean   :  0.6258   Mean   :0.011581
3rd Qu.:0.014   3rd Qu.:  1.0000   3rd Qu.:0.013244
Max.   :0.038   Max.   :120.0000   Max.   :0.092112
NA's   :13443
     CPCV
Min.   :0.00
1st Qu.:0.01
Median :0.01
Mean   :0.01
3rd Qu.:0.01
Max.   :0.05
NA's   :190689
```

Hide

```
New_data$CPV[ is.na(New_data$CPV)] <- 0
New_data$CPC[ is.na(New_data$CPC)] <- 0
New_data$CPCV[ is.na(New_data$CPCV)] <- 0
```

Hide

```
summary(New_data)
```

```
 Region_Indicators City_indicators SupplyVendors_Indicators
 Min.   : 2.00     Min.   :   1    Min.   : 1.000
 1st Qu.: 3.00     1st Qu.: 946    1st Qu.: 4.000
 Median :31.00     Median :1672    Median : 4.000
 Mean   :24.09     Mean   :1895    Mean   : 8.146
 3rd Qu.:39.00     3rd Qu.:2590    3rd Qu.:15.000
 Max.   :60.00     Max.   :5151    Max.   :23.000
 OS_Indicators    Browser_Indicators DeviceType_Indicators
 Min.   : 1.000   Min.   : 1.000     Min.   :1.000
 1st Qu.: 1.000   1st Qu.: 1.000     1st Qu.:1.000
 Median : 3.000   Median : 1.000     Median :2.000
 Mean   : 4.703   Mean   : 2.801     Mean   :1.883
 3rd Qu.: 5.000   3rd Qu.: 4.000     3rd Qu.:2.000
 Max.   :25.000   Max.   :12.000     Max.   :4.000
 Impression_Day   Impressions          Clicks
 Min.   :1.000   Min.   :  1.000   Min.   :0.000000
 1st Qu.:2.000   1st Qu.:  1.000   1st Qu.:0.000000
 Median :4.000   Median :  1.000   Median :0.000000
 Mean   :4.081   Mean   :  1.055   Mean   :0.006846
 3rd Qu.:6.000   3rd Qu.:  1.000   3rd Qu.:0.000000
 Max.   :7.000   Max.   :120.000   Max.   :2.000000
      CTR               CPC                  VCR
 Min.   :0.00000   Min.   :0.000e+00   Min.   :0.0000
 1st Qu.:0.00000   1st Qu.:0.000e+00   1st Qu.:0.0000
 Median :0.00000   Median :0.000e+00   Median :1.0000
 Mean   :0.00645   Mean   :7.941e-05   Mean   :0.5925
 3rd Qu.:0.00000   3rd Qu.:0.000e+00   3rd Qu.:1.0000
 Max.   :1.00000   Max.   :4.450e-02   Max.   :1.0000
      CPV             Completes          Total_Spend
 Min.   :0.000000   Min.   :  0.0000   Min.   :0.000000
 1st Qu.:0.009381   1st Qu.:  0.0000   1st Qu.:0.009459
 Median :0.011073   Median :  1.0000   Median :0.011098
 Mean   :0.011272   Mean   :  0.6258   Mean   :0.011581
 3rd Qu.:0.013164   3rd Qu.:  1.0000   3rd Qu.:0.013244
 Max.   :0.038316   Max.   :120.0000   Max.   :0.092112
      CPCV
 Min.   :0.000000
 1st Qu.:0.000000
 Median :0.009233
 Mean   :0.007089
 3rd Qu.:0.012036
 Max.   :0.049803
```

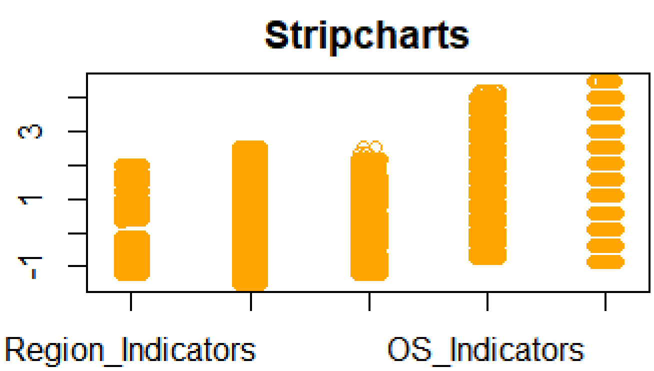
We can observe skewness in data as mean is either greater than or less than median.

Visualizing the Data

Plotting stripcharts and boxplots side-by-side can be useful to visualize the spread and distribution of data as well as analyzing outliers.

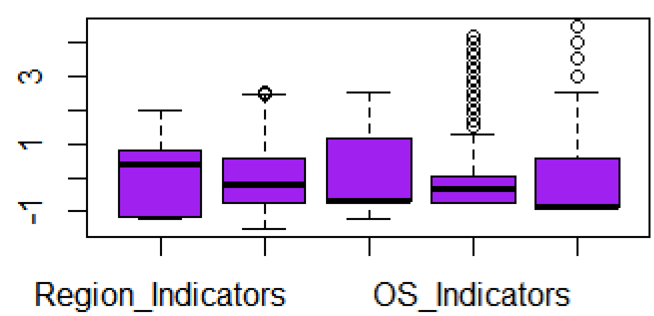
Hide

```
#New_data$City_indicators<-as.numeric(levels(New_data$City_indicators))[New_data$City_indicator
s]
## Stripcharts
numeric_data <- New_data[,c(1:5)]
numeric_data <- data.frame(scale(numeric_data ))
strip<-stripchart(numeric_data,
           vertical = TRUE,
           method = "jitter",
           col = "orange",
           pch=1,
           main="Stripcharts")
```



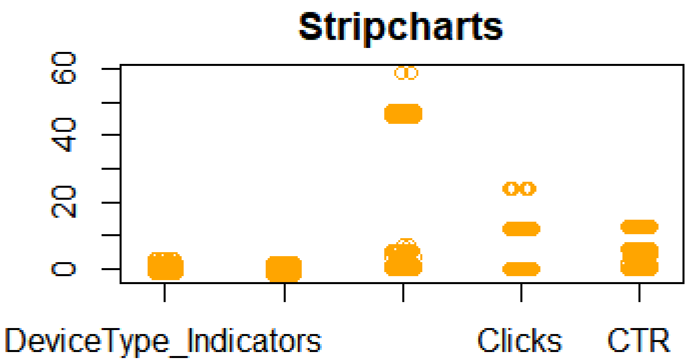
Hide

```
box<-boxplot(numeric_data,col='Purple')
```



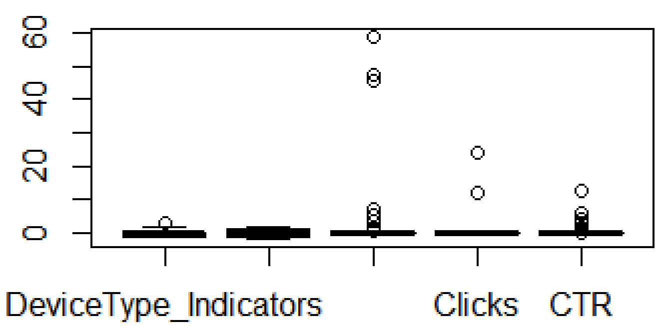
Hide

```
## Stripcharts
numeric_data <- New_data[,c(6:10)]
numeric_data <- data.frame(scale(numeric_data ))
strip<-stripchart(numeric_data,
           vertical = TRUE,
           method = "jitter",
           col = "orange",
           pch=1,
           main="Stripcharts")
```



Hide

```
box<-boxplot(numeric_data)
```



Hide

```
## Stripcharts
numeric_data <- new[,c(10:16)]
numeric_data <- data.frame(scale(numeric_data ))
strip<-stripchart(numeric_data,
           vertical = TRUE,
           method = "jitter",
           col = "orange",
           pch=1,
           main="Stripcharts")
box<-boxplot(numeric_data)
```

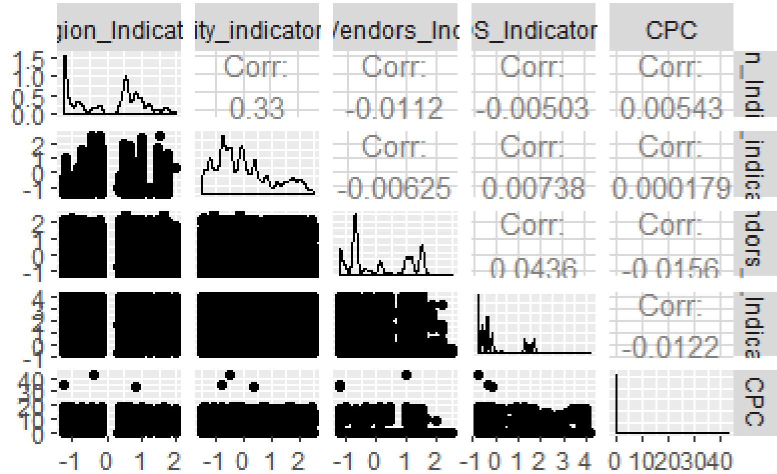
From the above plots we can confirm about skewness and presence of outliers as well.

Scatter Plot matrix is another important way to visualize data, its distribution and correlation with other variables.

Hide

```
numeric_data <- New_data[,c(1,2,3,4,11)]
numeric_data <- data.frame(scale(numeric_data ))
library("GGally")
ggpairs(numeric_data)
```

```
 plot: [1,1] [=-----------------------------------]  4% est: 0s
 plot: [1,2] [===---------------------------------]  8% est: 4s
 plot: [1,3] [====--------------------------------] 12% est: 5s
 plot: [1,4] [======------------------------------] 16% est: 4s
 plot: [1,5] [=======-----------------------------] 20% est: 3s
 plot: [2,1] [=========---------------------------] 24% est: 3s
 plot: [2,2] [==========--------------------------] 28% est: 6s
 plot: [2,3] [============------------------------] 32% est: 5s
 plot: [2,4] [=============-----------------------] 36% est: 5s
 plot: [2,5] [===============---------------------] 40% est: 4s
 plot: [3,1] [================--------------------] 44% est: 4s
 plot: [3,2] [==================------------------] 48% est: 5s
 plot: [3,3] [===================-----------------] 52% est: 5s
 plot: [3,4] [=====================---------------] 56% est: 5s
 plot: [3,5] [======================--------------] 60% est: 4s
 plot: [4,1] [========================------------] 64% est: 3s
 plot: [4,2] [=========================-----------] 68% est: 4s
 plot: [4,3] [===========================---------] 72% est: 3s
 plot: [4,4] [============================--------] 76% est: 3s
 plot: [4,5] [==============================------] 80% est: 3s
 plot: [5,1] [===============================-----] 84% est: 2s
 plot: [5,2] [=================================---] 88% est: 2s
 plot: [5,3] [==================================--] 92% est: 1s
 plot: [5,4] [===================================-] 96% est: 1s
 plot: [5,5] [====================================]100% est: 0s
```

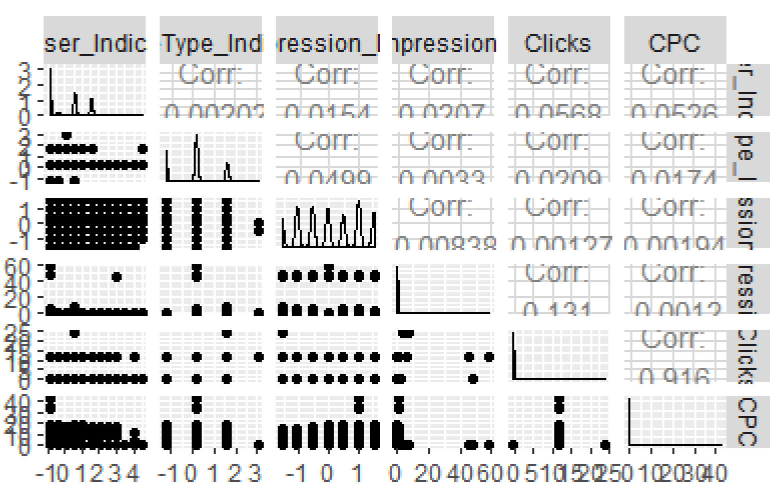


CPC, Region Indicators, City Indicators are positively correlated while CPC, Vendor Indicator and OS Indiator are negatively correlated.

Hide

```
numeric_data <- New_data[,c(5,6,7,8,9,11)]
numeric_data <- data.frame(scale(numeric_data ))
library("GGally")
ggpairs(numeric_data)
```

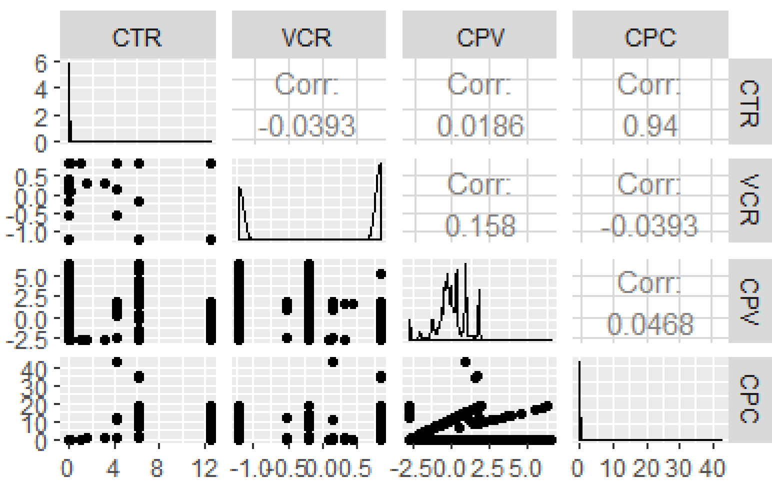
```
plot: [1,1] [=-----------------------------------]  3% est: 0s
plot: [1,2] [==----------------------------------]  6% est: 5s
plot: [1,3] [===---------------------------------]  8% est: 4s
plot: [1,4] [====--------------------------------] 11% est: 4s
plot: [1,5] [=====-------------------------------] 14% est: 4s
plot: [1,6] [======------------------------------] 17% est: 4s
plot: [2,1] [=======-----------------------------] 19% est: 4s
plot: [2,2] [========----------------------------] 22% est: 7s
plot: [2,3] [=========---------------------------] 25% est: 7s
plot: [2,4] [==========--------------------------] 28% est: 7s
plot: [2,5] [===========-------------------------] 31% est: 6s
plot: [2,6] [============------------------------] 33% est: 6s
plot: [3,1] [=============-----------------------] 36% est: 5s
plot: [3,2] [==============----------------------] 39% est: 7s
plot: [3,3] [===============---------------------] 42% est: 8s
plot: [3,4] [================--------------------] 44% est: 7s
plot: [3,5] [=================-------------------] 47% est: 7s
plot: [3,6] [==================------------------] 50% est: 6s
plot: [4,1] [===================-----------------] 53% est: 6s
plot: [4,2] [====================----------------] 56% est: 6s
plot: [4,3] [=====================---------------] 58% est: 6s
plot: [4,4] [======================--------------] 61% est: 6s
plot: [4,5] [=======================-------------] 64% est: 6s
plot: [4,6] [========================------------] 67% est: 5s
plot: [5,1] [=========================-----------] 69% est: 5s
plot: [5,2] [==========================----------] 72% est: 5s
plot: [5,3] [===========================---------] 75% est: 5s
plot: [5,4] [============================--------] 78% est: 4s
plot: [5,5] [=============================-------] 81% est: 4s
plot: [5,6] [==============================------] 83% est: 3s
plot: [6,1] [===============================-----] 86% est: 3s
plot: [6,2] [================================----] 89% est: 2s
plot: [6,3] [=================================---] 92% est: 2s
plot: [6,4] [==================================--] 94% est: 1s
plot: [6,5] [===================================-] 97% est: 1s
plot: [6,6] [====================================]100% est: 0s
```



Hide

```
numeric_data <- New_data[,c(10,12,13,11)]
numeric_data <- data.frame(scale(numeric_data ))
library("GGally")
ggpairs(numeric_data)
```

```
plot: [1,1] [==---------------------------------]  6% est: 0s
plot: [1,2] [====-------------------------------] 12% est: 2s
plot: [1,3] [======-----------------------------] 19% est: 1s
plot: [1,4] [=========--------------------------] 25% est: 1s
plot: [2,1] [===========------------------------] 31% est: 1s
plot: [2,2] [=============----------------------] 38% est: 3s
plot: [2,3] [===============--------------------] 44% est: 3s
plot: [2,4] [=================------------------] 50% est: 2s
plot: [3,1] [===================----------------] 56% est: 2s
plot: [3,2] [=====================--------------] 62% est: 2s
plot: [3,3] [========================-----------] 69% est: 2s
plot: [3,4] [==========================---------] 75% est: 2s
plot: [4,1] [============================-------] 81% est: 1s
plot: [4,2] [==============================-----] 88% est: 1s
plot: [4,3] [================================---] 94% est: 1s
plot: [4,4] [==================================]100% est: 0s
```

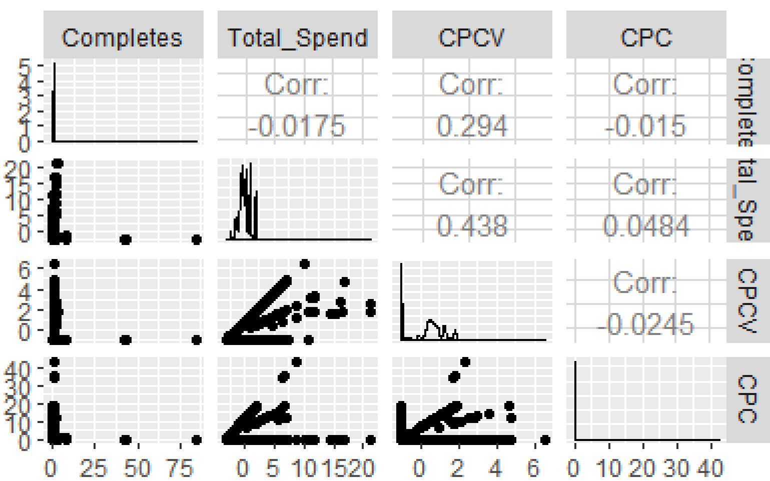


CTR,CPC,CPV are positively correlated and VCR negatively.

Hide

```
numeric_data <- New_data[,c(14,15,16,11)]
numeric_data <- data.frame(scale(numeric_data ))
library("GGally")
ggpairs(numeric_data)
```

```
plot: [1,1] [==--------------------------------]  6% est: 0s
plot: [1,2] [====------------------------------] 12% est: 3s
plot: [1,3] [======----------------------------] 19% est: 2s
plot: [1,4] [========--------------------------] 25% est: 2s
plot: [2,1] [==========------------------------] 31% est: 2s
plot: [2,2] [============----------------------] 38% est: 3s
plot: [2,3] [==============--------------------] 44% est: 3s
plot: [2,4] [================------------------] 50% est: 2s
plot: [3,1] [==================----------------] 56% est: 2s
plot: [3,2] [====================--------------] 62% est: 2s
plot: [3,3] [=======================-----------] 69% est: 2s
plot: [3,4] [=========================---------] 75% est: 2s
plot: [4,1] [===========================-------] 81% est: 1s
plot: [4,2] [=============================-----] 88% est: 1s
plot: [4,3] [===============================---] 94% est: 1s
plot: [4,4] [==================================]100% est: 0s
```



CPC, Completes, CPCV are negatively correlated and total spend is positively correlated.