Report on the Ad Campaign Data

Abhiruchi Shukla

Data

This data set is related to the ad campaign run over the internet on various platforms spanning various cities in the US. The information collected through these ad campaigns include region and city where the campaign was running, the device platform, browser platforms, time and day it appeared on the platform, click through ratio, cost per click, etc. These factors can help us infer important aspects of the ad campaign like cost incurred in each region per click, top performing cities, type of devices on which the ad video is viewed more frequently, etc. These insights can help make important business decisions in order to target the ads in the right way while optimizing cost.

In the figure 1 below, we can see different factors over which the ad campaign data was collected.

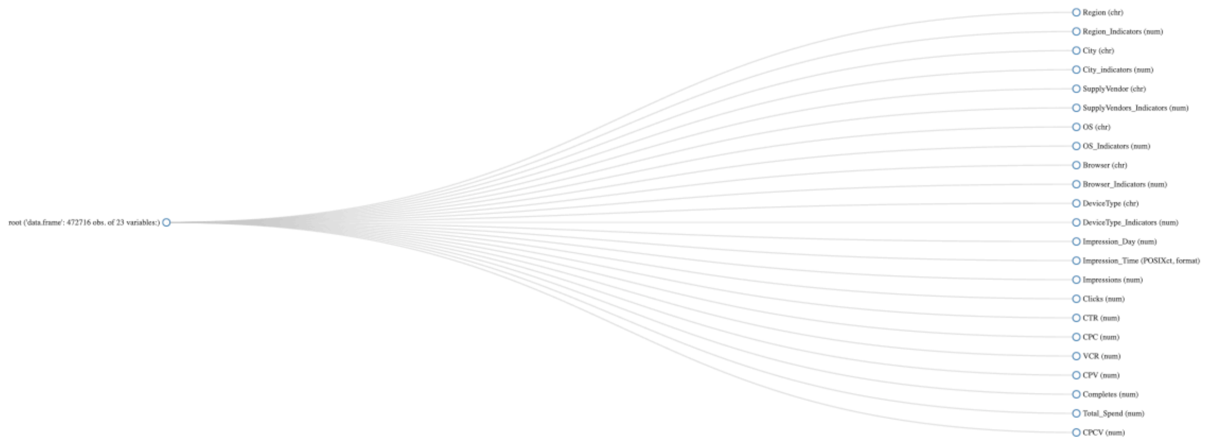

Figure 1

This dataset has 472716 observations over 23 variables. Most of the variables are numeric type. The character type variables like region, city, OS, browser, device, supply vendor have been transformed to corresponding indicator variables. Therefore, we can easily get rid of all these character type variables for analysis purposes.

Exploratory Data Analysis

Missing Values

It is important to see if the dataset has any missing values. For that, plot shown in figure 2 was obtained in R which shows percentage of missing rows of data for each variable.

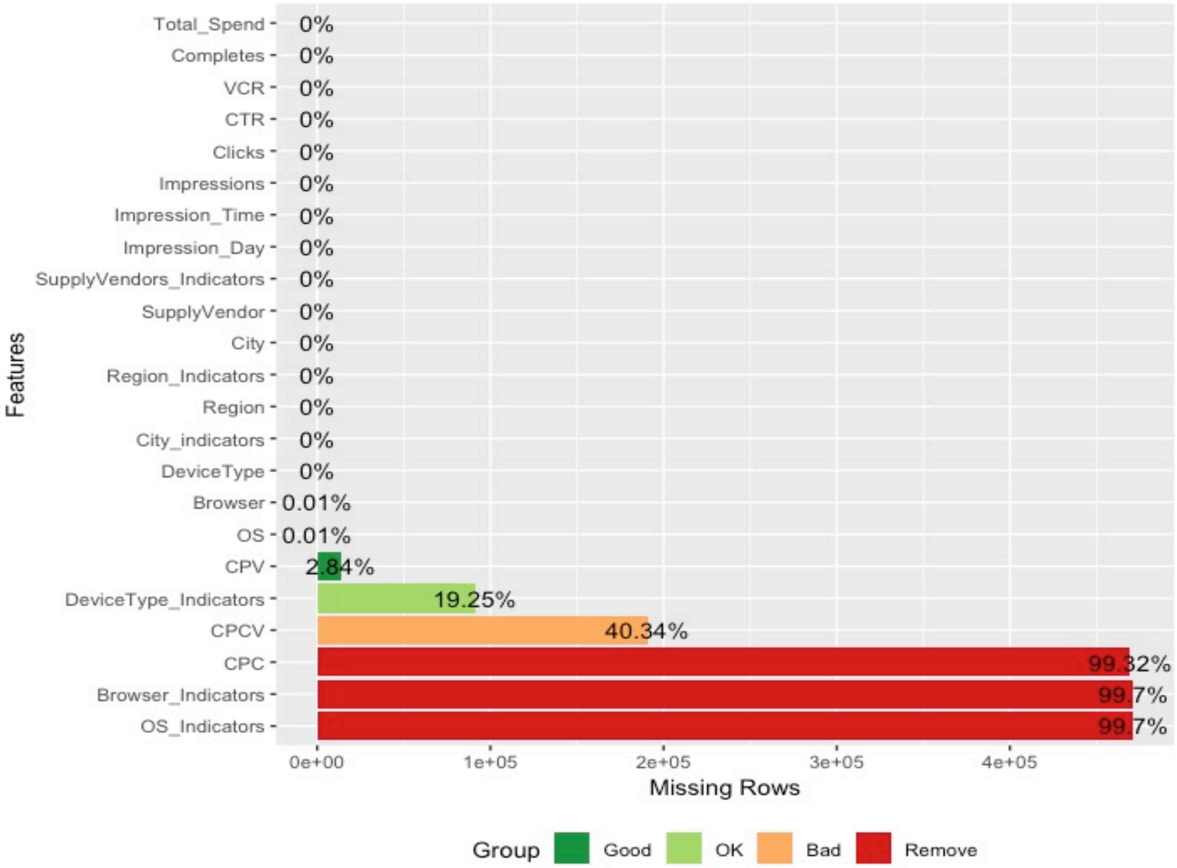

Figure 2

As it can be seen clearly, there are variables which have no missing values at all. But, our target variable (CPC) has 99.32% data missing. Therefore, I decided to remove only the missing rows from the CPC column. Variables like Browser_Indicators, and OS_Indicators had to be removed for the analysis since they have high number of missing rows of data. Removing the missing rows for these two variables reduces the dataset size, significantly.

Histogram

A histogram plot was obtained to check for continuous and discreet variables

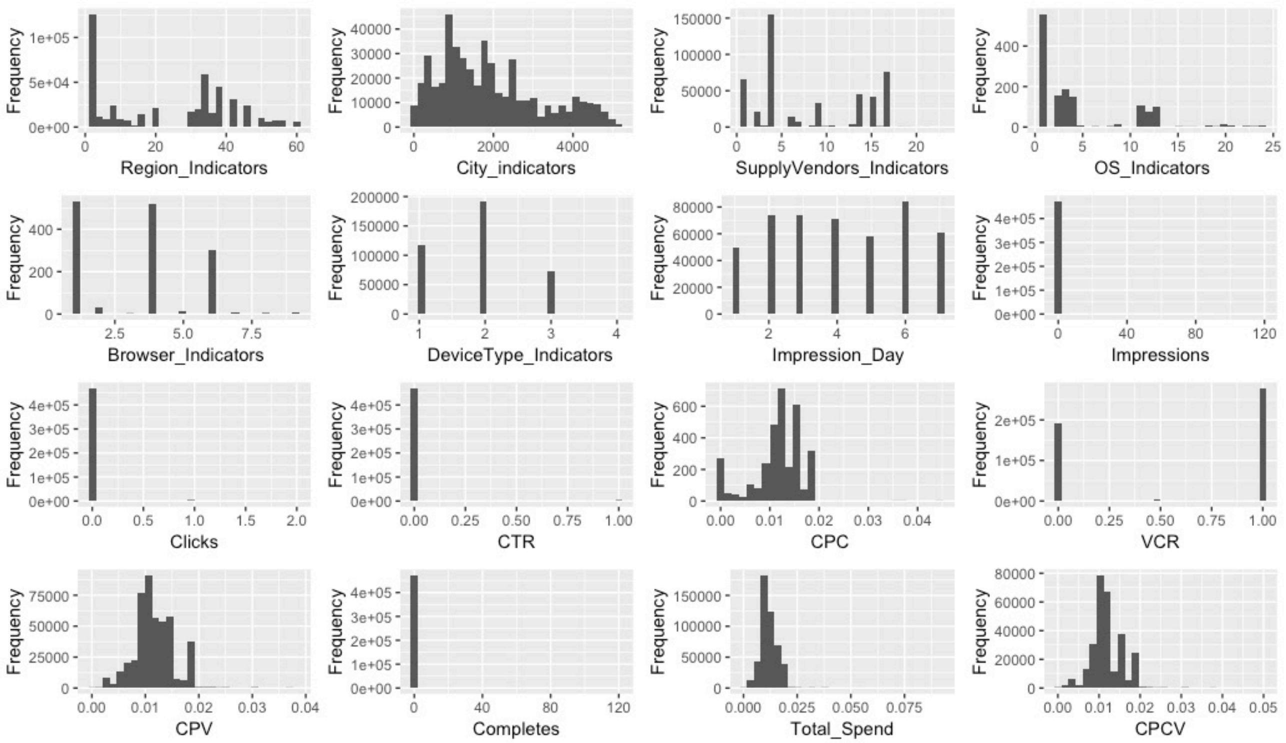

Figure 3

In figure 3, the indicator variables aka categorical variables are seen as discreet values on the histogram. CPC, CPV, Total_Spend and CPCV are some of continuous variables. Variables like Impressions, Clicks, CTR, and Completes have a very high frequency of 0 and 1 values.

Bar Plot

A bar plot for categorical variables was obtained as shown in figure 4.

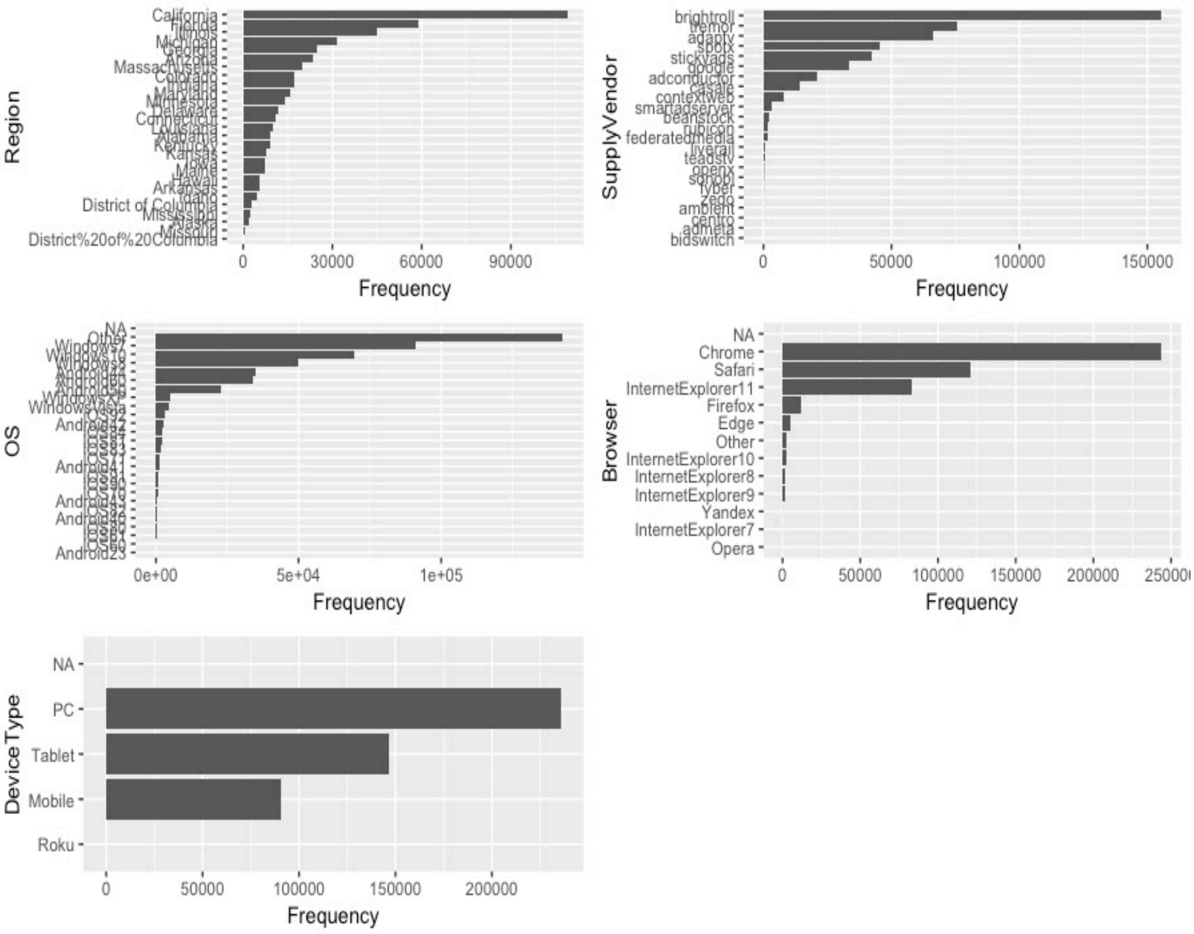

Figure 4

The plot shows, for each categorical variable, rectangular bars with heights proportional to the frequency with which they appear in the data. It gives some very useful information about the data set. For e.g., the dataset has more records from "Chrome" Browser category, from the "California" region, with Supply Vendor as "brightroll" , Device Type as "PC" and "Other" OS platform.

Correlation Plot

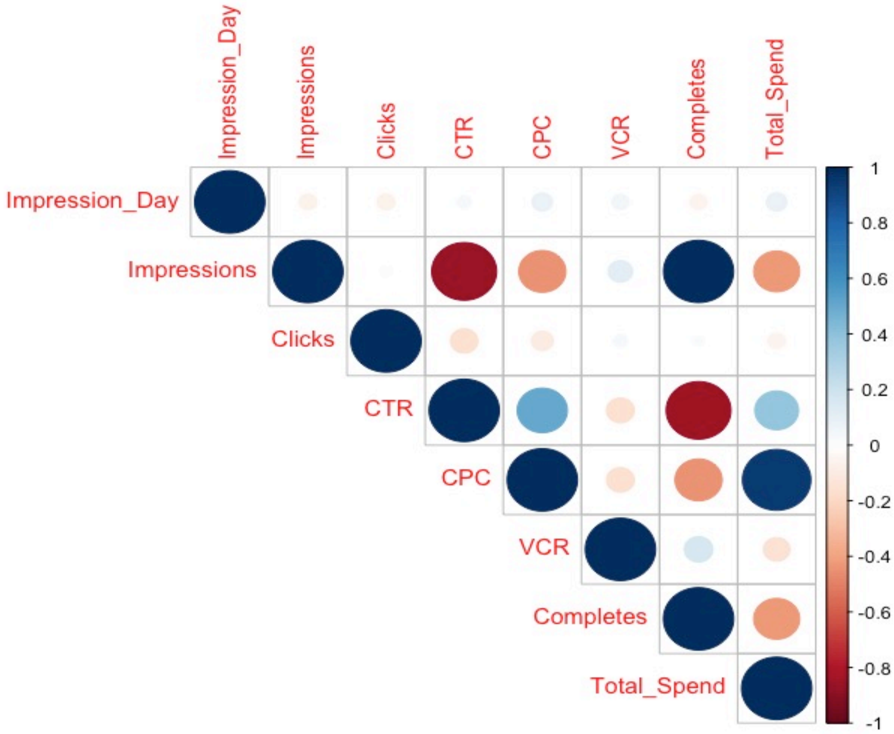

Figure 5

The correlation map in figure 5, shows correlation between the continuous variables varying between -1 to 1. Certain correlation like Impressions and Completes, CPC and TotalSpend, CTR and Completes, Impressions and CTR have a correlation factor of 1 or -1 which shows that these variables are functions of each other and have a prediction accuracy of 100%.
An interpretation of correlation value of -1 between CTR and Impressions could be that more users are clicking on the ad video link to view the ad campaign, but the ad server is not called because of broken link or fraud videos.
A correlation value of 1 between Impressions and Completes indicates that once the ad server receives the request to view the ad, the video is viewed completely.
Correlation like CPC and Completes have a correlation factor of 0.4 which implies that CPC can predict Completes with 40% accuracy.

Performance Measurement

The performance metric used to measure performance of cities and devices is taken as

$$\frac{CTR}{1 + CPC}$$

A weighted average of this metric over the clicks was obtained to measure "performance" after data was grouped by cities and devices.

For this analysis, some data cleaning was done to obtain sensible results. Rows of data with missing CPC (target value) were removed and the data was subset to include only relevant columns like City_indicators, DeviceType_Indicators, Clicks, CTR, and CPC.

To measure the top 5 performing cities, the aggregated city data was sorted by performance in decreasing order. To judge a reasonable number of clicks, the following graph was studied.

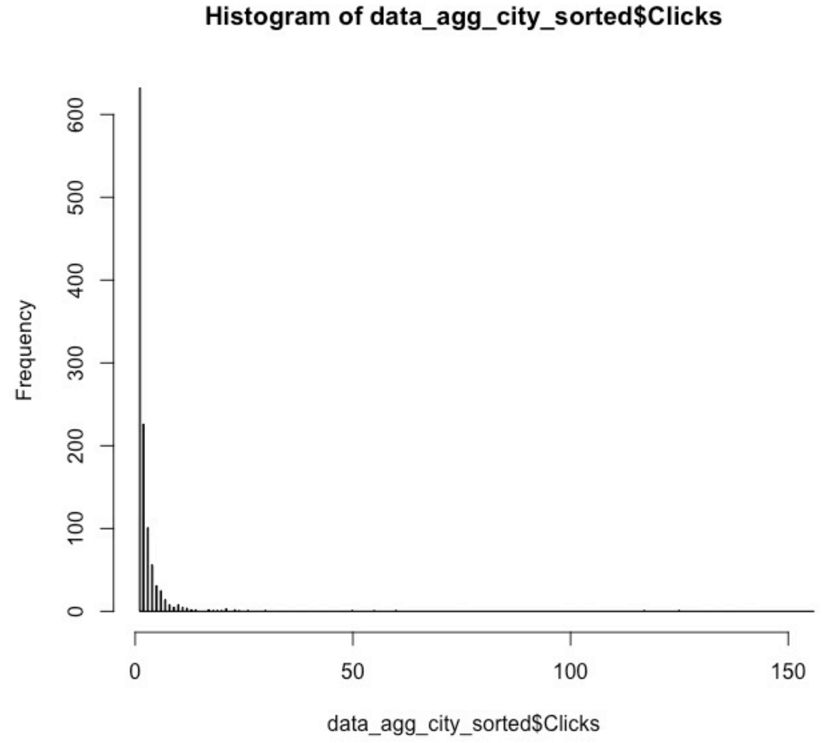

Figure 6

To get a balanced data subset, value of Clicks greater than 5 was used to measure performance by cities. The following output was obtained from R.

```
            Group.1 City_indicators   Clicks    ob1        performance
480         1686        13488         8       7.976033   0.9970042
678         2265        15855         7       6.969711   0.9956730
534         1825        10950         6      5.971865  0.9953109
535         1828        18280         10      9.934361  0.9934361
397         1406        14060         10      9.925930  0.9925930
511         1766        10596         6       5.954528   0.9924213
```

Clearwater, Wailuku, North Miami Beach, Ocala, and Littleton are the top performing cities by this metric.

To find the best and worst performing device, the same performance metric was used as in case of the cities, but the metric was aggregated over Clicks grouped by devices. The following output was obtained from R.

```
Group.1 DeviceType_Indicators Clicks       ob1      performance
1     1              448             448   436.784311    0.9749650
2     2             3238            1619   1448.850840   0.8949048
3     3             1803             607   581.472327    0.9579445
4     4               16             4     3.997351      0.9993378
```

Considering a reasonable number of clicks from each device to be at least 100, Tablet was the best performing device and PC was the worst.

Data Model

The dataset has both continuous and categorical variables. To deal with these two types of variables appropriately, I segregated them in to different data frames and analyzed them separately.

Categorical Variable

A couple of categorical variables like Os_Indicators and Browser_Indicators were removed before this analysis since they had 99% rows of data with missing values. Region_Indicators, City_indicators, DeviceType_indicators and SupplyVendor_Indicators variables were then analyzed by finding the mean CPC (over Clicks) aggregated by region, city, supply vendor, and device type. A histogram plot was obtained for each of these categorical variables as shown in figure 7.

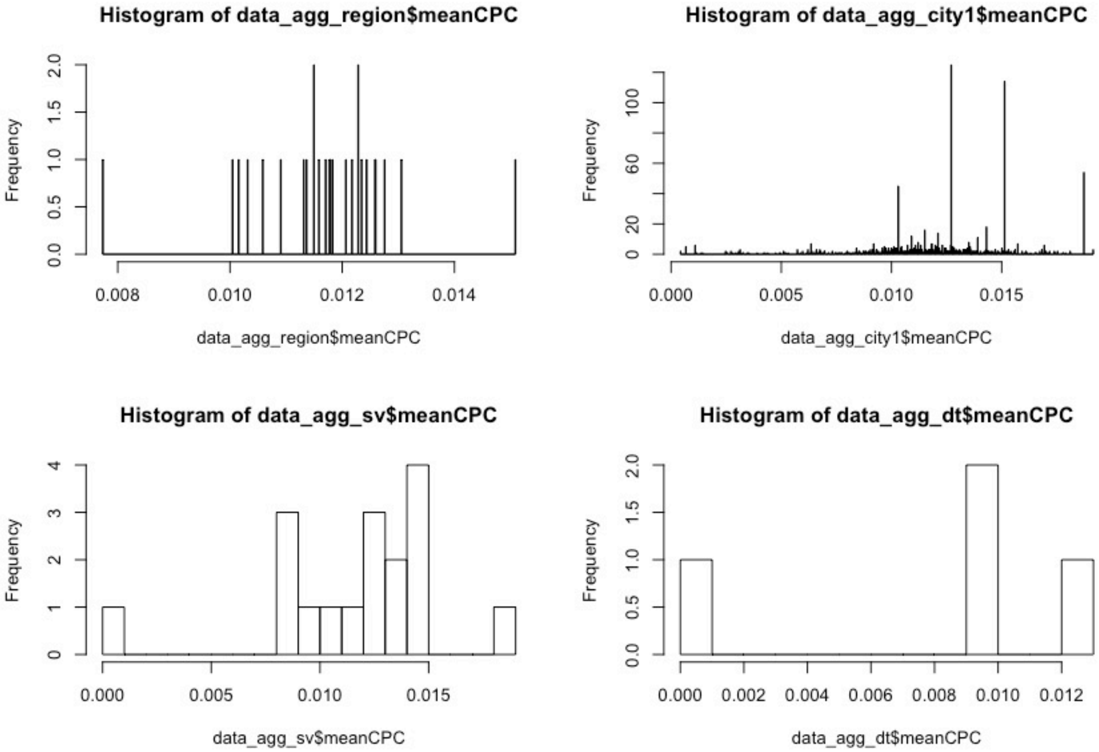

Figure 7

We can see from the figure that the mean CPC value for all the indicator variables is concentrated between the values of 0.010 and 0.0125. The height of the bar indicates values which lie close together. For e.g., a look at the graph of DeviceType_Indicators graph shows that there are at least 2 device type categories whose mean CPC values are close together.

Continuous Variables

To begin with, for this analysis, only numeric variables were considered, all categorical variables were removed, Impression_Time a POSIXCT type of data(date-time) variable was removed, and OS_Indicators and Browser_Indicators were removed since these two had a large number of missing rows of data. The missing rows from the CPC column were also omitted. The data was standardized.
To obtain a model to predict the factors affecting CPC, subset selection process was used for selecting a subset of relevant features. *The final model was selected based on least AIC value.*

The following models were obtained after regressing CPC on various factors for subset selection.

```
> step(fit)
Start:  AIC=-3236.94
CPC ~ Impressions + Clicks + CTR + VCR + CPV + Completes + Total_Spend + 
    CPCV

              Df Sum of Sq     RSS     AIC
- Clicks       1    0.0082  97.348 -3238.8
- CPCV         1    0.1103  97.450 -3237.5
<none>                      97.340 -3236.9
- Completes    1    0.8730  98.213 -3227.6
- CTR          1    1.7992  99.139 -3215.7
- Impressions  1    2.4431  99.783 -3207.5
- VCR          1    9.5575 106.897 -3120.2
- CPV          1   10.4290 107.769 -3109.9
- Total_Spend  1   22.7756 120.115 -2972.4

Step:  AIC=-3238.84
CPC ~ Impressions + CTR + VCR + CPV + Completes + Total_Spend + 
    CPCV

              Df Sum of Sq     RSS     AIC
- CPCV         1    0.1056  97.453 -3239.5
<none>                      97.348 -3238.8
- Completes    1    0.8743  98.222 -3229.5
- CTR          1    1.9167  99.264 -3216.1
- Impressions  1    2.4697  99.817 -3209.1
- VCR          1    9.5493 106.897 -3122.2
- CPV          1   10.4331 107.781 -3111.7
- Total_Spend  1   22.8291 120.177 -2973.7

Step:  AIC=-3239.46
CPC ~ Impressions + CTR + VCR + CPV + Completes + Total_Spend

              Df Sum of Sq     RSS     AIC
<none>                      97.453 -3239.5
- Completes    1    0.9624  98.416 -3229.0
- CTR          1    1.8111  99.264 -3218.1
- Impressions  1    2.6201 100.073 -3207.8
- VCR          1   12.5072 109.961 -3088.4
- CPV          1   28.3397 125.793 -2917.8
- Total_Spend  1   30.5854 128.039 -2895.4

Call:
lm(formula = CPC ~ Impressions + CTR + VCR + CPV + Completes + 
    Total_Spend, data = data_num_scaled1_subset)

Coefficients:
(Intercept)  Impressions          CTR          VCR          CPV    Completes  Total_Spend  
   -0.87536      0.03161      0.02832      0.62965      0.34885     -0.01984      0.33549  
```

The model CPC ~ Impressions + CTR + VCR + CPV + Completes + Total_Spend has the least AIC value of -3239.46

Therefore, this model, a more parsimonious model, can be used for prediction purposes.