Monika Pawar

This project will analyze the data that comes with the Air France case, interpret the results, and make recommendation from the Case: Air France Internet Marketing: Optimizing Google, Yahoo!, MSN, and Kayak Sponsored Search.

**SP20: Behavioral Analytics with Vsulztn**

Air France Internet Marketing

**1.** (**50 Points**) Please add four additional metrics into your data using the formula provided (This part is done in Jupiter Note Book)

Please provide descriptive statistics (Count, Max, Min, Mean, and Std.) for variables (CTR,

TCR, Net Revenue, Avg. Cost per Click, ROA, Average Revenue per Booking, Probability

of Booking). **Please report a summary statistics table and provide short descriptions of**

**your observations and thoughts.**

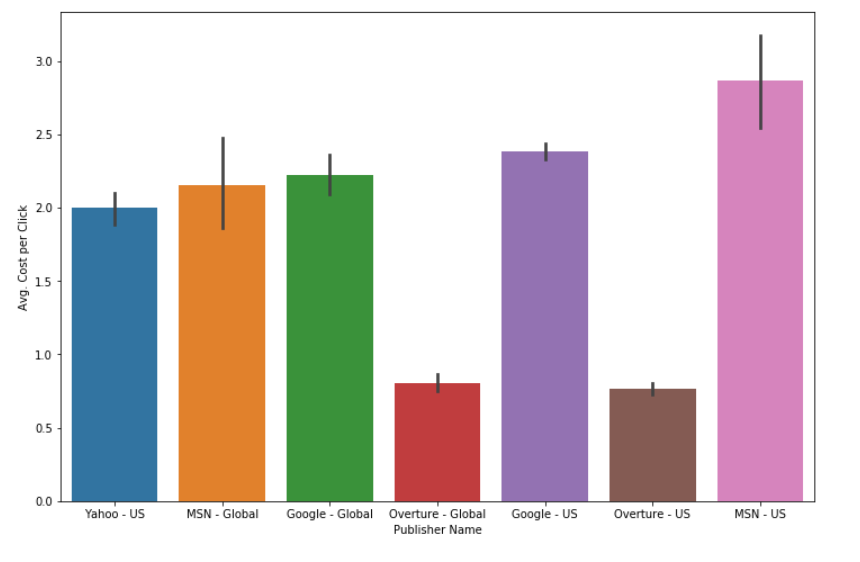
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Summary Statistics** | **Engine Click Thru %** | **Trans. Conv. %** | **Net Revenue** | **Avg. Cost per Click** | **ROA** | **Average Revenue per Booking** | **Probability of Booking** |
|  |  |  |  |  |  |  |  |
| **Count** | 4510 | 4510 | 4510 | 4510 | 4510 | 4510 | 4510 |
| **Maximum** | 200 | 900 | 549524.0503 | 10 | 379487.027 | 5877.75 | 0.818182 |
| **Minimum** | 0 | 0 | -8725.924987 | 0 | -100 | 0 | 0 |
| **Mean** | 11.141451 | 0.569255 | 866.207678 | 1.89024 | 341.413864 | 83.575942 | 0.000681 |
| **Standard deviation** | 20.234582 | 13.862485 | 14246.2989 | 1.322375 | 7277.915278 | 344.981225 | 0.015151 |

* The above summary table is showing the summary statistics (N, Max, Min, Mean, and Std.) for variables: Engine Click Thru %, Trans. Conv. %, Net Revenue, Avg. Cost per Click, ROA, Average Revenue per Booking, Probability of Booking.
* The count indicates the number of observations in the AirFrance\_df dataset. Here, total of 4510 observation are available.
* The average of all variables is shown as the mean in the below table.
* The standard deviation shows how the numbers are spread out from the mean (average) value. Here, Probability of Booking have low standard deviation, Hence the data is closer to the mean value.
* Engine Click Thru % --Its ratio showing how often people who see your ad end up clicking it, the mean value for Engine Click Thru % is 11.14 which is a good indication that users find your ads helpful and relevant.
* Trans. Conv. % -- The transaction conversion rate shows that how manu clicks were converted into transactions. The mean Trans. Conv. % is 0.57 which is low.
* Net Revenue -- The maximum amount of net revenue is 549524 and the minimum amount of net revenue is -8725.9 Net revenue is obtained by total revenue minus cost incurred.
* Avg. Cost per Click -- It is calculated by dividing the total cost to an advitiser by the total number of clicks, it’s the amount that a website publisher receives when a paid advertisement on the site is clicked. Here the mean of Avg cost per click is 1.89$ and Maximum cost is 10$.
* ROA for return on ad spend. It's a calculation that divides the amount of revenue generated from ads by the amount spent on advertising. The goal of measuring ROA is to determine if the cost of advertising yields an acceptable amount of incoming business revenue. So, we have average ROA as 341% which is good.
* Average Revenue per Booking -- It is calculated by dividing the Amount by Total Volume of Bookings which is the revenue received for every booking. The mean value for this 83.57
* Probability of Booking --This shows the actual number of transactions took place after the views on the ad. Probability of Booking can be calculated by dividing the Engine Click Thru % Trans. Conv. %. The average value shows 0.000681 which is very low.

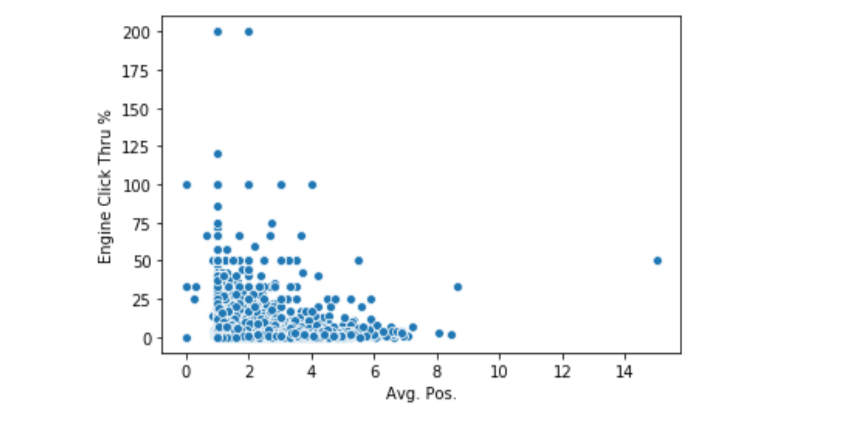
e. Please make a Histogram for any of the variables of your own interests in the data. Then

report any insights you may be able draw from the charts.

* Histogram plot:



* The above plot is drawn between two variables Publisher Name and Average Cost per Click.
* Average Cost per Click (CPC) is based on actual cost-per-click (actual CPC), which is the actual amount you're charged for a click on your ad.
* The above plot shows that which publisher charged more for Average Cost per Click, we can see that MSN-US charges the most and Overture-US charges the least.
* Scatter plot:

****

* The above scattered plot is drawn between two variables Avg. Pos and Engine click thru %
* Average position is a statistic that describes how your ad typically ranks against other ads. This rank determines in which order ads appear on the page. Positions 1-8 generally appear on the first page while 9-16 on the second
* From the above scatterplot, we can interfere that as Average position value increases, the engine click thru % will decrease

**2.** (60 points) Please conduct regression analysis to study what factors influence the Total Cost. Basically, Total Cost is your dependent variable (Y) and your task is to determine what the important independent (explanatory) variables are. You should use the domain knowledge you have learnt from the case, personal experiences, and external research to guide your variable selections. You may try different set of independent variables in the data set to see which one(s) has significant results and thus support your belief (you may need to create dummy variables for some of the non-numerical variables).

Please report 1) the final set of independent variables you have chosen and why you have chosen them; 2) the estimated regression equation with simple explanations for each estimated coefficient (β) and its associated relationship (include significance, direction of the impact, magnitude of the impact, and justification of the identified relationship). (Hint: feel free to explore the data in any way you want, e.g. correlation matrix, scatter plots, etc.)

In my final regression model, I have chosen (**Keyword, Clicks, Impressions, Total Cost/ Trans., Amount**) as

my final set of independent variables after doing 3-4 sets of combinations of independent variables in my OLS model testing. I have

chosen these variables due to below reasons:

* By creating Pearson's correlation matrix between various independent variables and dependent variables, check the significant variables which has correlation with Dependent variable.
* During my testing for OLS models I made sure there is no multicollinearity between the various independent variables.
* During my first and second attempt of OLS model I dropped few columns because their P-values in the result summary was not significant.
* In my final attempt to OLS with the above-mentioned independent variables, I was achieving the optimum result summary with good f-stats and adj. R-Squared.

After reading the given case study, I have understanding that 𝑇𝑜𝑡𝑎𝑙𝑐𝑜𝑠𝑡=Impressions\*CPC\*CTR and with variables understanding, I interpret the importance of my chosen independent variables as below.

* **Keyword**: By determining the keyword searches that are most relevant to business offerings, then bid to place ads in the search results for relevant keywords.
* **Click**: How many clicks has happened for your ad will determine CPC (cost per click) and it has impact on Total cost.
* **Impressions:** The number of times a sponsored link is displayed on search result pages will also affect the total cost.
* **Total Cost/ Trans:** This variable is related to Total cost as increase in total cost per transaction will directly increase total cost.
* **Amount:** It is kinda gross revenue which impact the total cost.

A linear regression model with five predictor variables can be expressed with the following equation:

Y = B0 + B1\*X1 + B2\*X2 + B3\*X3 + B4\*X4  + B5\*X5 +e.

The variables in the model are:

* Y, the response variable;
* X1, the first predictor variable;
* X2, the second predictor variable; and
* e, the residual error, which is an unmeasured variable.

The parameters in the model are:

* B0, the Y-intercept;
* B1, the first regression coefficient; and
* B2, the second regression coefficient.
* B3, the third regression coefficient.
* B4, the fourth regression coefficient.
* B5, the fifth regression coefficient.

We have below values from our OLS model Result summary ( x1 = Keyword , x2 =Clicks, x3 = Impressions ,x4 = Total Cost/ Trans. , x5 = Amount )

const 167.4758

X1 19.8571

X2 1799.0207

X3 -91.8418

X4 146.6864

X5 -750.0293

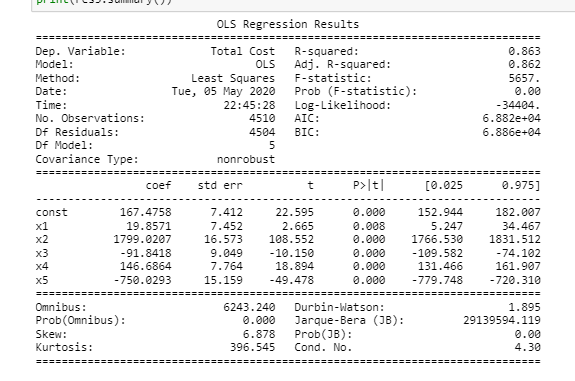
Hence from above result, the regression equation is

Y = 167.48 + 19.85(X1) +1799.03(X2) + (-91.84) (X3) + 146.68 (X4) + (-750) (X5) + e

x1 = Keyword, x2 =Clicks, x3 = Impressions, x4 = Total Cost/ Trans., x5 = Amount

The value of KEYWORD in regression result is 19.86, positive value indicating that there is positive correlation between KEYWORD and TOTAL COST. The p - value is ‘0.008’ which is less than 0.05, implies that the relationship is statistically significant. Hence, we can say that with every 1 keyword increase in the campaign, the TOTAL COST increases by 19.86 dollars.

The value of CLICKS in regression result is 1799.02, positive value indicating that there is positive correlation between CLICKS and TOTAL COST. The p - value is ‘0.000’ which is less than 0.05, implies that the relationship is statistically significant. Hence, we can say that with each 1 unit increase in CLICKS, the TOTAL COST increases by 1799.02 dollars.



The value of IMPRESSIONS in regression result is -91.84, negative value indicating that there is negative correlation between IMPRESSIONS and TOTAL COST. The p - value is ‘0.000’ which is less than 0.05, implies that the relationship is statistically significant. Hence, we can say that with each 1 unit increase in IMPRESSIONS, the TOTAL COST decreases by 91.84 dollars.

The value of Total Cost/ Trans in regression result is 146.69, positive value indicating that there is positive correlation between Total Cost/ Trans and TOTAL COST. The p - value is ‘0.000’ which is less than 0.05, implies that the relationship is statistically significant. Hence, we can say that with each 1 unit increase in total cost per transactions, the TOTAL COST increases by 146.69 dollars.

The value of AMOUNT in regression result is -750.02, negative value indicating that there is negative correlation between AMOUNT and TOTAL COST. The p - value is ‘0.000’ which is less than 0.05, implies that the relationship is statistically significant. Hence, we can say that with each 1 unit increase in AMOUNT, the TOTAL COST decreases by 750.02 dollars.

adj. R-square has value 86.3% and F-stat value at 5657. So this indicates that our regression line has a good fit using the above independent variables.