**WEB ANALYTICS ASSIGNMENT #1 – Air France**

**By Navya Manepalli, Meghana Rega**

**Question 1a**

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) summary report and observations.

**Summary Report:**

*Table1.*Statistics Summary Report for the set of variables

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Count | Max | Min | Mean | Std |
| CTR | 4510 | 200.00% | 0.00% | 11.14% | 20.23 |
| TCR | 4510 | 900.00% | 0.00% | 0.57% | 13.86 |
| Net Revenue | 4510 | $549,524 | -$8,726 | $866 | 14246.3 |
| Avg. Cost per Click | 4510 | $10.00 | $0.00 | $1.89 | 1.32 |
| ROA | 4509 | 379487% | -100% | 341% | 72.79 |
| Average Revenue per Booking | 368 | 587775% | 3400% | 102426% | 704.32 |
| Probability of Booking | 4510 | 82% | 0% | 0% | 0.015 |

**Observations:**

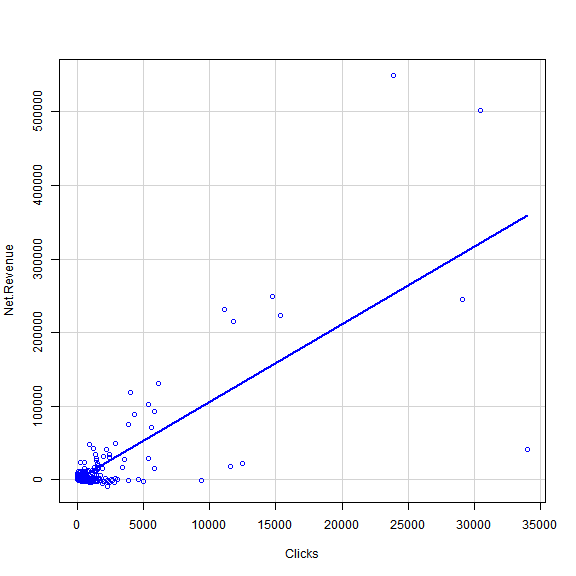
The Max and Min values of the descriptive statistics of CTR, TCR, Net Revenue, Average Cost per Click, Average Revenue per booking and ROA define a data set with extreme values.

The mean is on the very lower bound side. Therefore, clearly there are more keywords/ publishers/ campaigns with lower Click through rate than the others. Since the data is extremes and very diverse set, we have a very high standard deviation.

Clearly, few publishers charge more for the ads since the click through rate is very high for few publishers. For ROA, the amount spent on Ads for few publishers is very negligible (too positive MAX) whereas for few it is more than the profits (negative MIN).For Probability of booking, max is 81% and the min is 0% with mean as 0.07% and standard deviation as 0.01 which says the data across the publishers is on par with the others.

The strategies used by few publishers is clearly different from that of other publishers.

**Question 1b**

**Scatter Plot (with Trend Line): Clicks vs. Net Revenue**

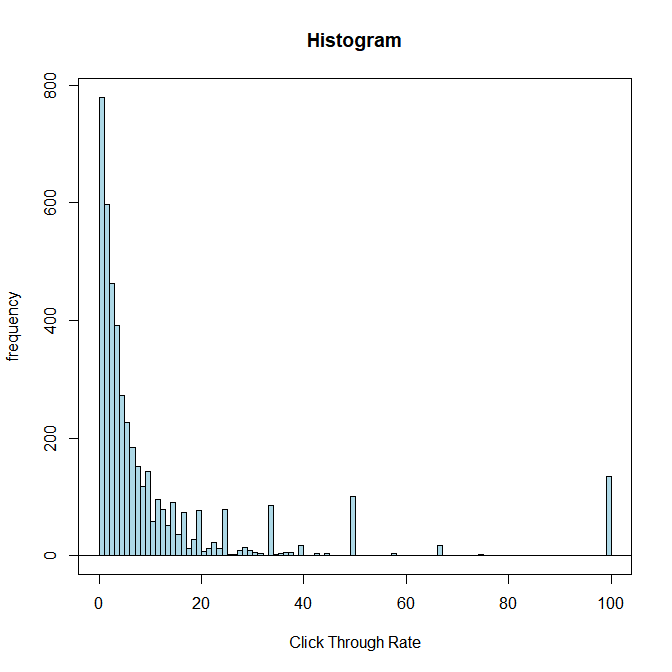
*Figure1.* Scatter plot Clicks vs Net Revenue

At the first impression, there is no negative impact on the revenue due to number of clicks i.e, as the clicks are increasing the revenue is increasing but not decreasing which shows that clicks make a positive impact.

It also proves that both are directly proportional to each other by the representation on the trendline to be linear (exponentially increasing)

It also says that the correlation between the two variables, clicks and net revenue are strongly correlated.

Also, there are a greater number of publishers/clicks having the range of 0-5000. And there are a very few on the higher range.

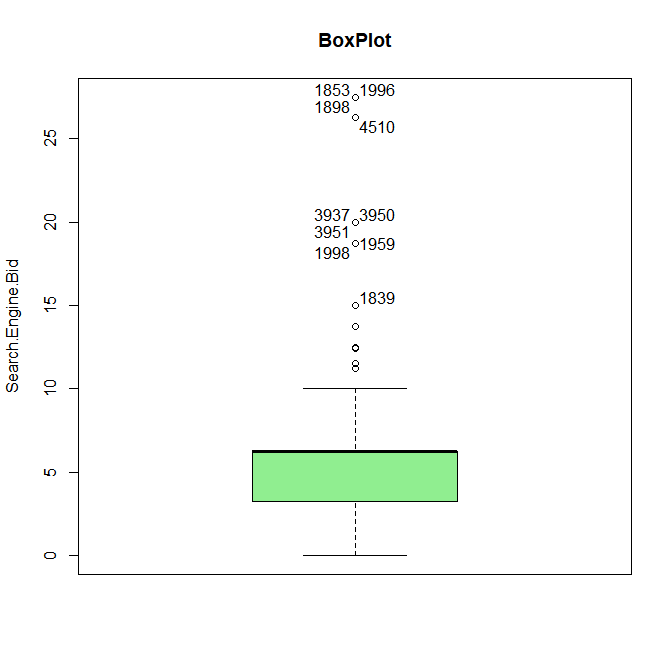
**Histogram: Engine Click Through Rate (CTR ) **

*Figure2.* Histogram for CTR

Click through rate is derived by calculating number of clicks divided by number of impressions.

When we have more number of clicks when compared to number of impressions, we get a value greater 1 or more. There are around 200 with 100 as CTR and about 150 with more than 20 CTR. Which means that there are a lot of clicks directly going to the website and having some productive activity done instead of leaving the landing page. However, this is less than 1% of the total.

On the other side, we have more than 90% of the publishers have less than 1 CTR represented by the first bin in *Figure2.* That is the number of impressions are more than number of successful click to the landing page. That says more numbers of customers are having a look at it but not clicking on it. The reason can be either they are not getting the right result that is relevant or perhaps they have to increase the readability of the website, improve key word search or improve the look and feel of the ads etc.

**Box Plot: Search Engine Bid**

*Figure3.* Boxplot for Search Engine Bid

A boxplot tells how the data is distributed among the given values. The line represents the value of the median and colored box represents most of the data is in that range.

*Table2.* Statistics of observation

|  |  |
| --- | --- |
| Measure | Search Engine Bid |
| Min | $0.00 |
| Max | $27.50 |
| Average | $5.430042 |
| Median | $6.25 |

By the above representation, the bid rate is highly concentrated between $0 to $6.25. The range is between $0 and $10 on higher side. A very few above $10. Which means a lot of websites are spending about $5 per bid and that they don’t usually bid above $6.25. Surprisingly there are also a lot which do not bid at all.

**Question 2**

1. **Pivot Table with the following variables mapped against each publisher.**

*Table3.* Pivot table for variables plotted against publishers

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Publisher Name | Sum of Click Charges | Sum of Net Revenue | Average of Avg. Pos. | Sum of Total Volume of Bookings | Average Revenue Per Booking | ROA | Probability of Booking |
| Google - Global | $ 120,947 | $ 808,603 | 1.528 | 797 | $ 1,166 | 669% | 0.04407% |
| Google - US | $ 353,641 | $ 1,391,841 | 1.883 | 1550 | $ 1,126 | 394% | 0.04020% |
| MSN - Global | $ 12,160 | $ 133,364 | 1.785 | 129 | $ 1,128 | 1097% | 0.09216% |
| MSN - US | $ 16,098 | $ 165,451 | 1.759 | 140 | $ 1,297 | 1028% | 0.08229% |
| Overture - Global | $ 64,296 | $ 365,789 | 1.929 | 372 | $ 1,156 | 569% | 0.00208% |
| Overture - US | $ 141,976 | $ 205,457 | 2.498 | 289 | $ 1,202 | 145% | 0.00169% |
| Yahoo - US | $ 46,198 | $ 836,091 | 1.789 | 662 | $ 1,333 | 1810% | 0.07093% |
| Grand Total | $ 755,316 | $ 3,906,597 | 1.930 | 3939 | $ 1,184 | 517% | 0.00941% |

From the table, we see that there are no negative values which means that there are at least no losses in terms of ROA or Revenue incurred per booking. Also, the sum of net revenue is very high compared to sum of click charges which means that the charges paid for clicks has gotten back fully. Google-US has the greatest number of bookings and Google-Global has the highest revenue per booking on an average. The ROA is way too high than 100% which is a good sign and that the publishers are doing good. However, MSN is in lead for getting its ROA. Also, on the probability of booking is more in MSN, hence the greater value of ROA.

1. **Bubble Chart (X=Probability of Booking, Y=Avg. Cost Per Click, Bubble Size=Total Costs)**



*Figure 4.* Bubble Chart showing

With the above Bubble Chart drawn, below is the classification

*Table4.* Category for publishers

|  |  |
| --- | --- |
| Publisher Name | Category |
| Google - Global | Low Probability of Booking and High CPC |
| Google - US | Low Probability of Booking and High CPC |
| MSN - Global | High Probability of Booking and Low CPC |
| MSN - US | High Probability of Booking and High CPC |
| Overture - Global | Low Probability of Booking and Low CPC |
| Overture - US | Low Probability of Booking and Low CPC |
| Yahoo - US | High Probability of Booking and Low CPC |

From the bubble chart, it is clear that Google and Overture stand out with low probability of booking. Within which Overture has low CPC as well. Where as the others like Yahoo MSN have high probability of booking with low CPC. Probably the strategies, key words and other factors of Yahoo, MSN are better.

**Question 3**

**Bubble Chart: Overture US and Overture Global (**X=CTR, Y=TCR, Bubble size=Net Revenue**)**

*Figure4.* BubbleChart representing Overture Campaigns.

The bubble chart generated has all the bubbles lying at the lower level closer to the X axis which means even though the CTR is increasing, the TCR hasn’t shown a great increase. This infers that customers are clicking on the links but not booking tickets through these links/ads. Hence, the reason may be clearly because of the poor website-copy or bad ad campaign.

Air France needs to spend more on creating a relevant and better website. IT should probably create more search capability and work on tags/indexes and get the customers to book after they click on the links. Suggestions include user friendly / interactive website with very less number of links and that reduces the effort for customer. As a marketing strategy they could also start giving some bonus when they buy from that publisher etc.

Also, ad campaigns must be made more relevant and better matches with key words etc. This would increase the overall TCR indirectly.

**Question 4**

Performing a regression analysis with Total Cost as the dependent variable, allows us to approach the data to solve for the relation between multiple independent variables. As a first step, the dependent variable was regressed with various independent variables to study the magnitude and structure of relationships between each of the data points available.

Based on this analysis we determined the following independent variables:

* Clicks
* Impressions
* Total Volume of Bookings
* Net Revenue

The above independent variables stratified the metrics for single variable linear regression modelling. The R2 value for Clicks indicated promising 75% of total variation in the Total Cost, while the other independent variables stayed below 30%. None the less, all the above variables exhibited a low confidence interval (p-value) far below the 0.05 to reject any null hypothesis and conclude that we have sufficient evidence to be 95% confidence in a linear relationship. The data points from single variable regression analysis is summarized below

|  |  |  |
| --- | --- | --- |
| Independent Variables | R2 | p-value |
| Clicks | 75.49% | 0 |
| Impressions | 18.59% | 1.0849E-203 |
| Total Volume of Bookings | 29.11% | 0 |
| Net Revenue | 23.28% | 8.508E-262 |

As we identified the independent variables for the Total Cost, multiple regression was used to investigate the relation between them together. The results can be summarized as follows

|  |  |
| --- | --- |
| *Regression Statistics* | |
| Multiple R | 0.931457001 |
| R Square | 0.867612144 |
| Adjusted R Square | 0.867494597 |
| Standard Error | 488.6421286 |
| Observations | 4510 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 23.57133759 | 7.331298312 | 3.215165526 | 0.001312899 |
| Clicks | 1.671187825 | 0.016508797 | 101.2301409 | 0 |
| Impressions | -0.000523206 | 5.3152E-05 | -9.843571528 | 1.2343E-22 |
| Total Volume of Bookings | 29.54945374 | 5.957315188 | 4.960196466 | 7.30144E-07 |
| Net Revenue | -0.077445206 | 0.004571572 | -16.94060693 | 1.86872E-62 |

Adjusted R2:

In the case of multiple regression we look to the Adjusted R2 value to make a quick determination of the model. An adjusted R2 of 86.75% is a great indicator of the relationship between the variables.

Residual Plots:

As it is difficult identify view data in a scattered plot, the residual plot gives the underlying relationship is linear for each of the independent variables in the multiple regression.

p-Value:

Similar to single variable regression, the all the above variables exhibited a low confidence interval (p-value) far below the 0.05 to reject any null hypothesis and conclude that we have sufficient evidence to be 95% confidence in relationship variables.

**Estimated Regression Equation**

The structure of the multiple regression equation is

*𝐲=Slope+Intercept1x1+Intercept2x2+…+Interceptkxk*

Based on the multiple regression model for Total Cost, the following equation can be structured

***Total Cost = 23.5713 + Clicks \* 1.67 – Impression \* 0.00052***

***+ Total Volume of Booking \* 29.5494 – Net Revenue \* 0.074***

The above equation can be used for forecasting the Total Cost for a particular campaign based on target values of Clicks, Impression, Total Volume of Booking and Net Revenue. In short, the campaign spending threshold can be forecasted.

**Question 5**

Summary from Kayak and Pivot table from Question2

* Kayak Trans. Conv. Rate: Total no. of Booking / no. of Clicks - 7.33%
* Average Publisher TCR: (From Q2) - 0.77%
* Kayak CPC: Media Cost / No. of Clicks - $1.26
* Average Publisher CPC (From Q2) - $1.47

From the above statistics, it is inferred that transaction conversion rate for Kayak is may too high compared to that of all the other publishers. On the other hand, Cost per click is a little less for Kayak when compared to other publishers.

In this case, even if the CPC is same or little higher, we recommend going with Kayak due to its high TCR. It clearly is the go-to solution. Also, Air France could shift majority of its investments from other publishers to Kayak.

1. Appendix

Script used for getting charts using R

Scatter plot with trendline:

names(RawData\_Updated\_AirFrance) <-

+ make.names(names(RawData\_Updated\_AirFrance))

scatterplot(Net.Revenue~Clicks, regLine=TRUE, smooth=FALSE, boxplots=FALSE,

+ data=RawData\_Updated\_AirFrance)

Histogram :

with(RawData\_Updated\_AirFrance, Hist(Engine.Click.Thru.., scale="frequency",

breaks="FD", xlim=c(0,100), col="lightblue",xlab="Click Through Rate",main="Histogram"))

Boxplot :

Boxplot( ~ Search.Engine.Bid, data=RawData\_Updated\_AirFrance, main="BoxPlot", col="lightgreen",

id=list(method="y",location="r"))

Excel Sheet used for all the calculations:

