Huiyi Huang

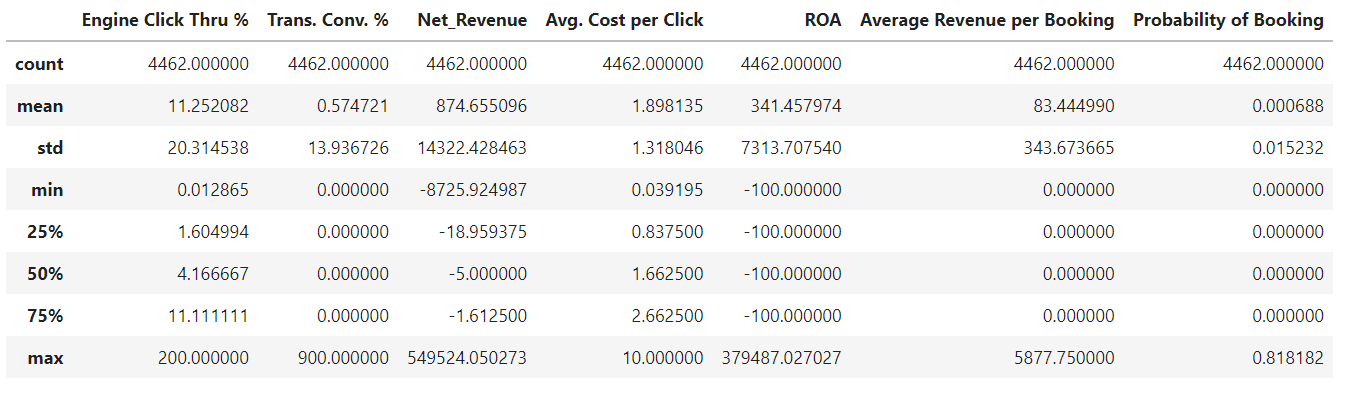
ISYS 622

May 1st , 2022

Professor Wenqi Zhou

Air France Online Marketing Analysis

1. *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.*



For cleaning data, since there are only 48 missing values in the Match type, comparing to the total number(4510), I choose to drop them. The descriptive table is shown above.

* Engine Click Thru %; Trans.Conv.% ; Probability of booking: Considering the purpose of online marketing, we should pay attention on rates that customers will finally click on the ads. Engine Click Thru rate has an average as 11.25%. It means on average 11 clicks will happen with 100 impression. For Trans.Conv.%, on average, 0.57% probability when viewers completed the transaction when sees the ads. Probability of booking’s average is really small, 0.00068%, means only small part of viewers will become the buyer. These three variables can be used to evaluate how good the campaign is; they are good indicators, and affect the total costs at the end of the month.

As this boxplot shows below, we can know Yahoo engine reached the highest upper limit of the click through rate as 36%. On average, we can know that Google has the best clickthrough rate as 6%. We could summarize as Google provided the most interactive Air France advertisement among 4 publishers.

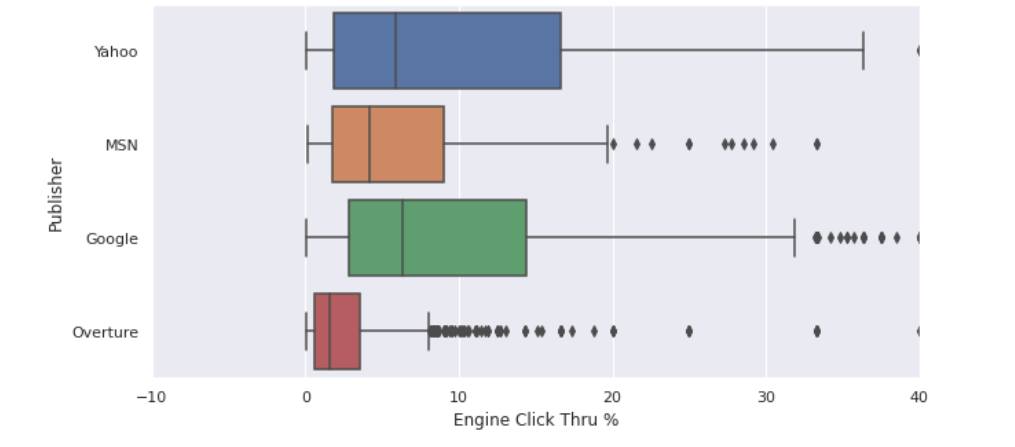


Figure 1Boxplot of engine click thru% by publishers

* Net\_Revenue: The range of this variable falls into a large range. Since net revenue can be used to determine if an ad make any returns, I log transform it.(both negative and positive values)
* ROA: I observed the minimum is -100 and maximum is 379,487; mean is 341. It means we have some campaign make huge profit, but most of them are making small profit or making no profit.

1. *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.*

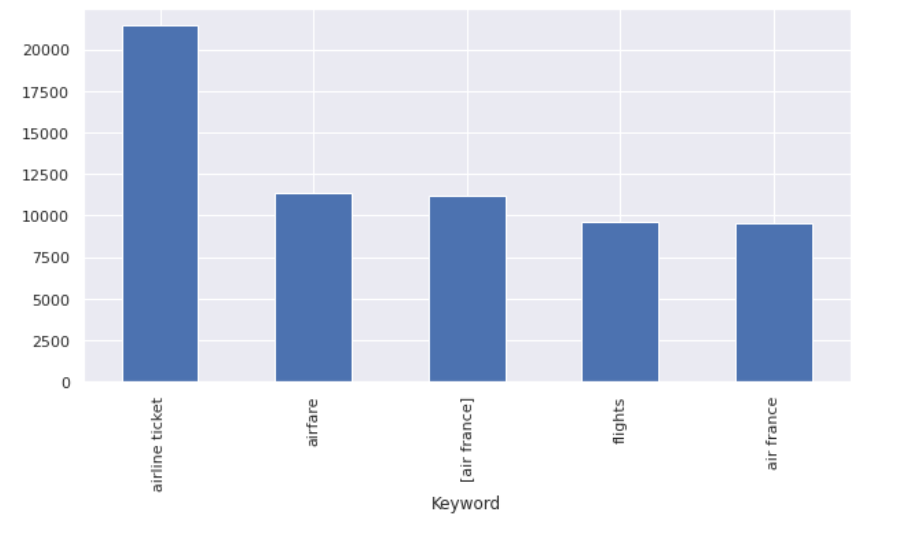


Figure 2Histogram of Total Cost by top5 Key words

The top 5 key words generated the most costs are: ‘airline ticket’, ‘airfare’, ‘[air france]’, ‘flights’, and ‘air france’. From these, ‘airline ticket’ generates over $21,250‬ advertisement fee, and other 4 keywords only generate costs that are half of it. It means publishers charged a lot on users searching ‘airline ticket’.

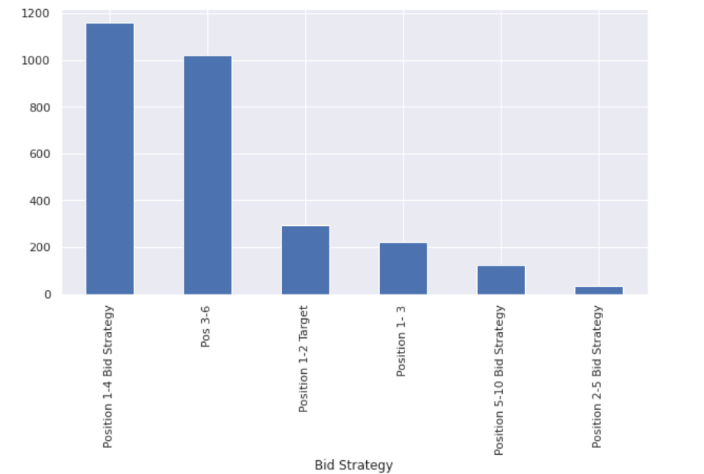
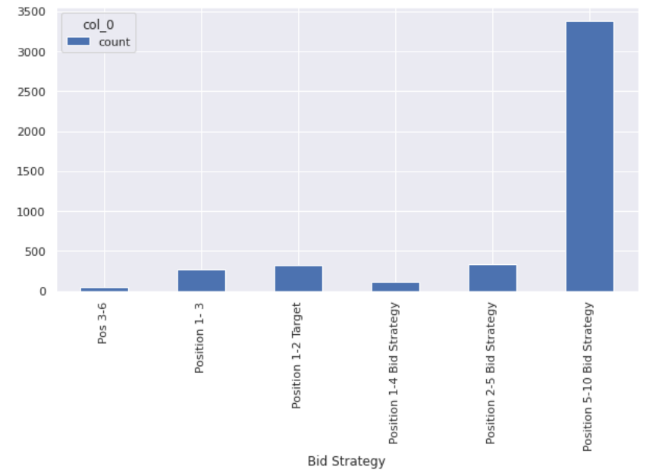


Figure 3Histogram on frequency of Big Strategy Figure 4Histogram on Total Cost by Bid Strategy

As this group of histogram shows, we can see the most frequently bid happened 3384 times that uses Position 5-10 bid strategy, however, it only costs $122 per bid on average. 1- 4 bid strategy happens only 111 times, and costs $1161 on average every time it happens. We can see how the marketer project advertisement. They project less-quality advertisement for 3000 times, to cover the disadvantage of that projecting high-quality advertisement being too expensive.

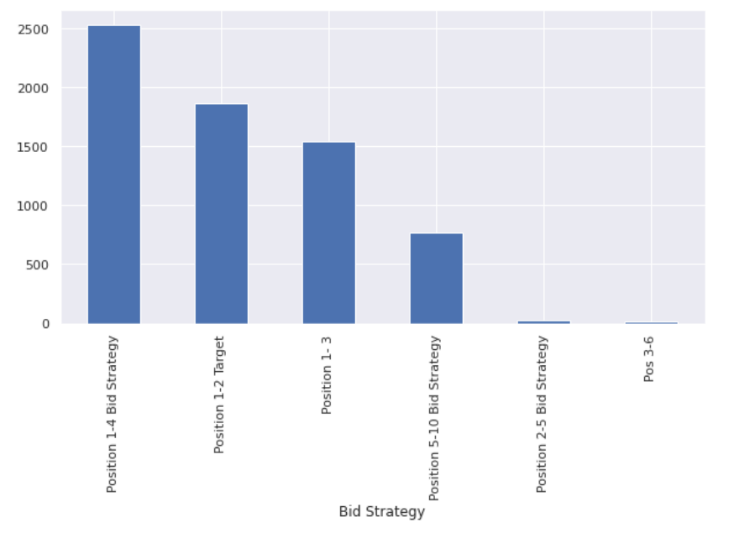
In this histogram, we can see that 1-4 position bid strategy advertisements generated $2500 net revenue on average, which is almost 2 times of the costs. And 5-10 position bid strategy advertisements generate 7 times of the profit.(122-> 767). For advertisement using 3-6 bid strategy, it doesn’t make any revenue, it spent the second most money and has the worst effective.

Figure 5 Histogram on Net Revenue by Bid Strategy

1. *(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). You should also check the multicollinearity issue for a legitimate behavioral analytics model. Please report 1) the final set of independent variables you have chosen and why you have chosen them; and 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. but keep in mind our priority #1: does the IV possibly influence DV?)*
2. Search Engine Bid, Clicks, Impressions, Avg. Pos, and Total Volume of Bookings I finally choose in regression analysis.

Search Engine Bid: According to research, bid is one of the reasons affects costs. Bid is the maximum price an advertiser is willing to pay for each click on the advertisement. The higher the bid is, the higher the final cost should be.

Clicks: Number of clicks happens to Air France’s ads. It will be charged each time click happens, so it is one of the reasons affects total costs.

Impressions: It is the number of times this ad shown to viewers. It is one of the indicators to see if an ad is effective. Usually, the more impression, the more effective the ad is. It is possible that website like Google will rate the ad for higher grade. The higher the grade is, the more expensive it will cost.

Total Volume of Bookings: It shows how many people engaged in booking through the click of advertisement. It reflects if an ad is effectively influence users’ behaviors. The more people book a trip through the ads, the more expensive it could cause.

Avg. Pos. : It shows how competitive an ad is comparing to other ads. An average of 1-4 ad usually shown on the first page. The less the average position is, the higher the cost could be.

* I choose these IVs, because they all causations of the final cost. As I check the Pearson correlation coefficient below, they all have the same sign of coefficient as I assumed. Search Engine Bid, Clicks, Impressions, and Total Volume of Booking are all affecting total cost in the same direction. IV goes up, DV goes up too. Among them, clicks as a variable is highly relative to the total cost, because the magnitude is 0.86, larger than other variables. For Avg. Pos. variable, it has negative correlation with total cost, meaning the least the average position it is, the more it will cost finally. It has very small magnitude as 0.018; meaning it has little impact on total cost. As the heatmap shows, the lighter the box is, the more correlation two variables are.

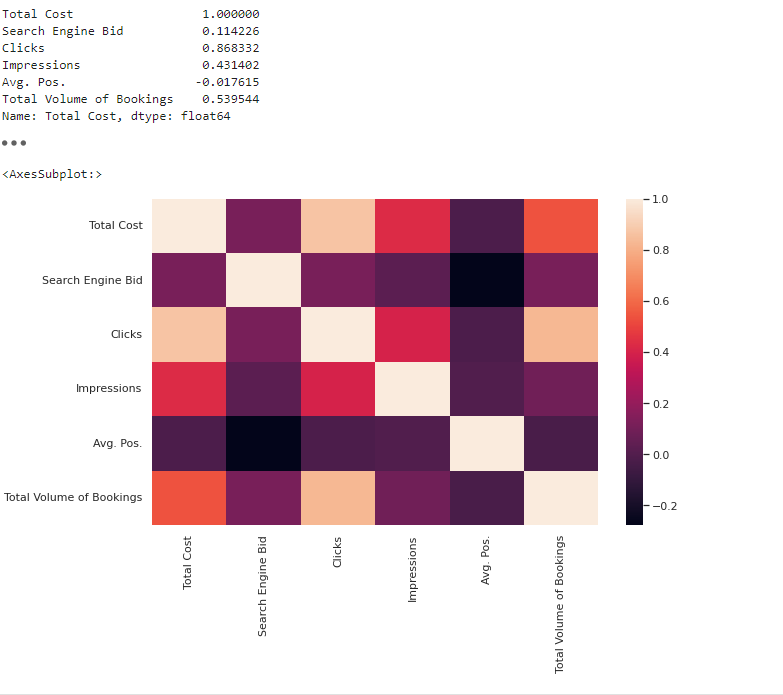


Figure 6 Pearson Coefficient

1. The OLS regression results shows below

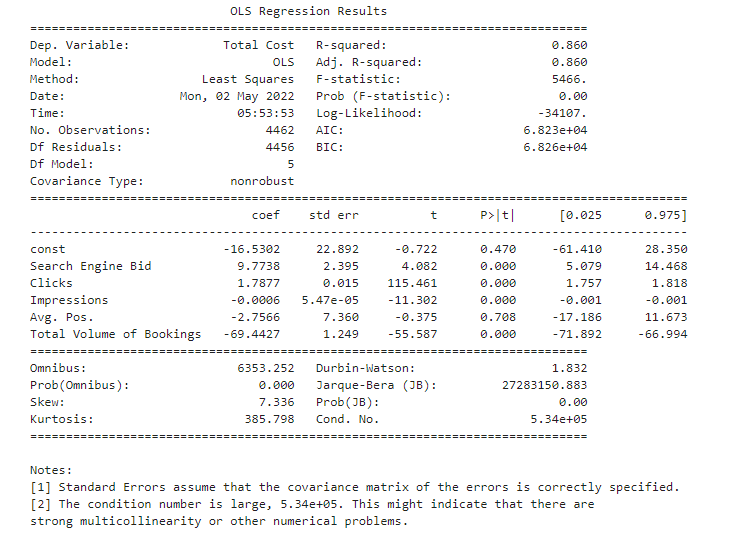


Figure 7 Final OLS regression results

Equation:

Total Cost = -16.5 + 9.8\*Search Engine Bid+ 1.8\*Clicks- 0.0006\* Impressions-69.4\*Total Volume of Booking – 2.8 \*Avs.Pos.

* Search Engine Bid, Clicks, Impressions, and Total Volume of Booking are statistically significant, since their p-value are all less than 0.05. Avg.Pos’s p-value is 0.7 and it is not statistically significant. I think it should be involved because it is the only indicator here to determine what’s the quality of advertisement is.
* The R-squared is 0.86. It shows this model captured 86% of data variation. It is a good model.
* The VIF table shows that all variables I choose. They are all in the regular area. Not too much collinearity to other variables.

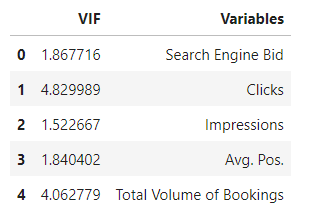


Figure 8 Multicollinearity for IVs

Search Engine Bid: the coefficient is 9.8 and it is positive, meaning that $1 more search engine bid will cause $9.8 on the total cost.

Clicks: the coefficient is 1.8 and it is positive, meaning that 1 more click on the ad will cause $1.8 more on the total cost.

Impressions: the coefficient is 0.0006 and it is negative, meaning that 1 more impression the ad show to viewers will cause $0.0006 less on the total cost. This IV affect DV in a small magnitude for every 1 time impression changes. The sign is different from what I expected.

Total Volume of Booking: the coefficient is 69.4 and it is negative, meaning that 1 more case of booking through the ad will cause $69.4 less on the total cost. The sign is different from what I expected. Maybe this variable and Impressions are counteracting a positive sign coefficient variable’s change on Total costs. That’s why it is not that reasonable.

Avg.Pos.: the coefficient is 2.8 and it is negative, meaning that 1 more grade add to the average position rating will cause $2.8 less on the total cost. It makes senses because usually, the more quality advertisement it shows, the more expensive it will cost. It is reasonable, since it’s expensive when projecting high quality advertisement.

*Bonus questions*

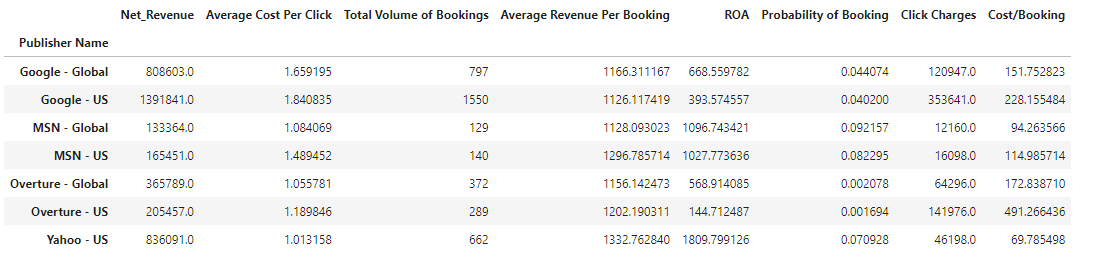
1. *****(5 points) Summarize metrics for each publisher. Please report the summary table including the variables as shown below (The answer for Google-global is already provided☺, and please fill-in others.) In addition, discuss Key Observations and Takeaways.*

Figure 8 Summary table of metrics for each publishers

* We can see for MSN – US, it has the highest probability of booking. (0.092%) And it only costs $1.48 per click.
* Advertisement on Google – US generates $1,391,941 of net revenue and the most total volume of booking. But the ROA shows only 393%, comparing to MSN or Yahoo advertisement, Google - US’s advertisement is not making huge profit.
* Yahoo-US have the highest ROA and Average cost per click is the lowest. For these reasons, I think projecting advertisement on Yahoo – US is a very good deal. Air France should continue work with it.

1. *(15 points) Based on the one-week summary data provided for Kayak in “kayak” sheet of the excel file, please calculate the following metrics and clearly show your calculation process.*

*a. Kayak Trans. Conv. Rate 7.33%*

*b. Average Publisher TCR 7.68%*

*c. Kayak CPC $1.26*

*d. Average Publisher CPC $1.47*

*Compare the calculations with what you have derived from the Bonus question #1, what recommendation you would like to make about marketing in Kayak relative to other publishers?*

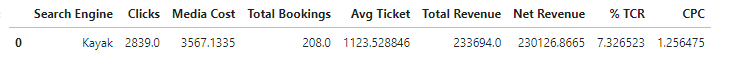


Figure 9 Summary table of metrics for Kayak

Kayak has TCR slightly lower than the average TCR of publishers, and the Kayak CPC is lower than publishers too. Kayak also generated $230,127 net revenue. In my opinion, Kayak is very competitive in booking hotel/ trip area. My recommendation for marketing in Kayak is to raise more advertisements bid to help raising impression and total bookings volumes. Since the ROA is very high as 6451%(Net Revenue/Media Cost), more advertisement will generate more profits. However, there is no impression number for Kayak, mainly because it has smaller market share and data is not integrated with Double Click. It’s hard to detect the real click-through rate.