Philip Dowhy

Behavioral Analytics with Visualizations Spring 2021

Internet Marketing Project-Air France

Philip Dowhy & Bevon Gomez

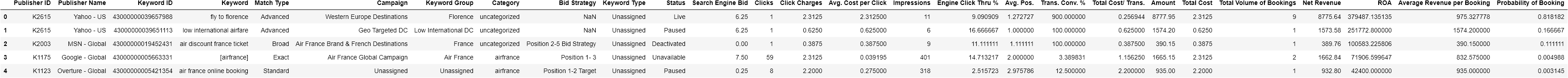
1. **(50 Points) Please add four additional metrics into your data using the formula provided**

**a. Net Revenue (Amount (total revenue) – Total Cost))**

**b. Return on Ad $ Spent (ROA) (Net Revenue / Total Cost) (Note: Set this variable as percentage; if Total Cost is 0, then set ROA as 0 for that observation.)**

**c. Average Revenue per Booking (Amount /Total Volume of Bookings) (Note: if Total Volume of Bookings is 0, then set null value for that observation)**

**d. Probability of Booking (Engine Click Thru % (CTR) \* Trans. Conv. % (TCR) / 10000)**

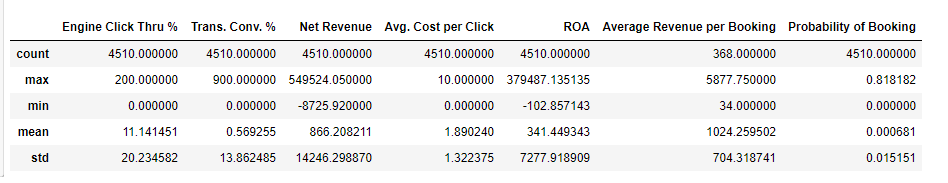


**Please provide descriptive statistics (Count, Max, Min, Mean, and Std.) for variables (CTR,**

**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.**



***Observations***

1: Total number of transactions are 4510

2: The largest net revenue is 549524.00 and the smallest is 866.20

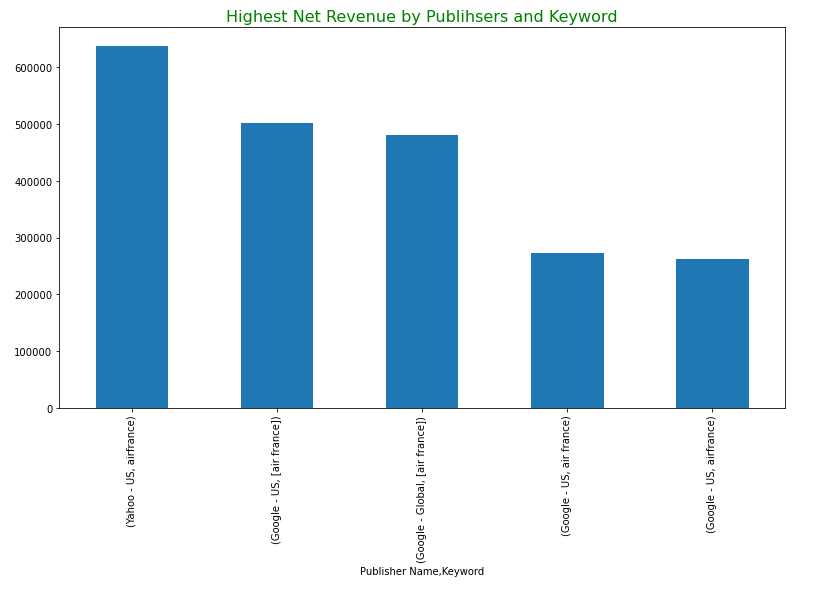
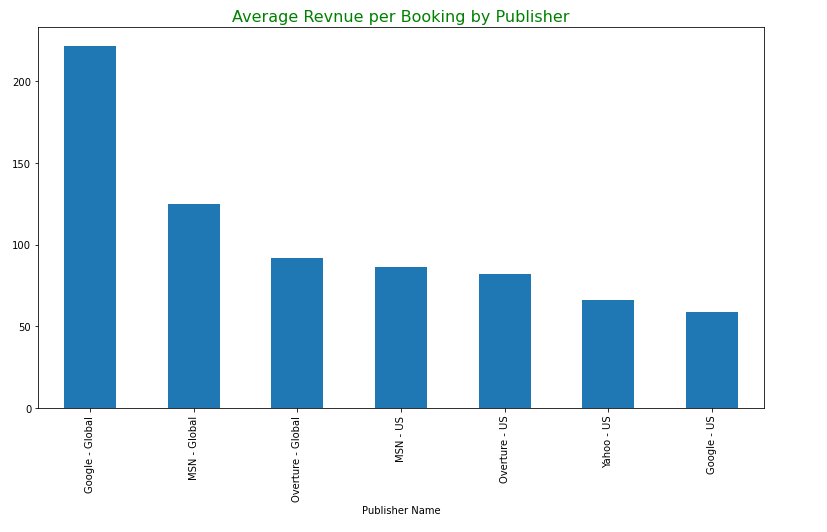
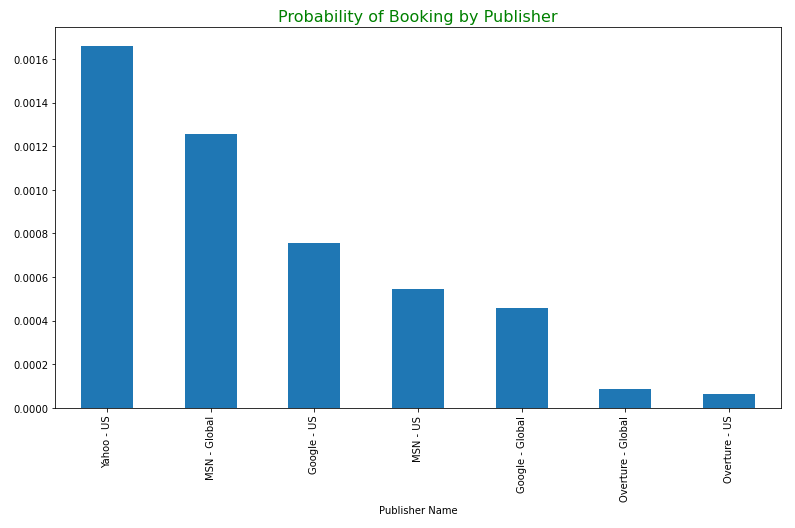
3: The data points for Engine Click Thru %, Trans. Conv. %, Net Revenue, ROA and Probability of Booking are widely distributed with a strong positive skewness because the stdev is higher than the mean.

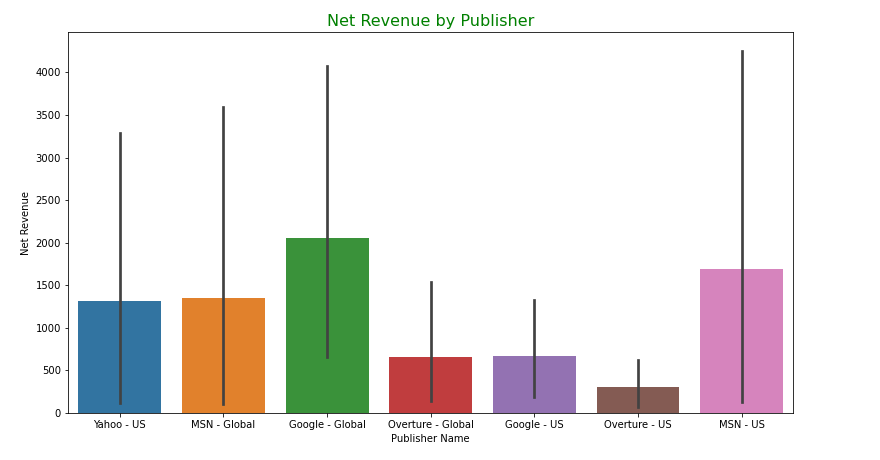
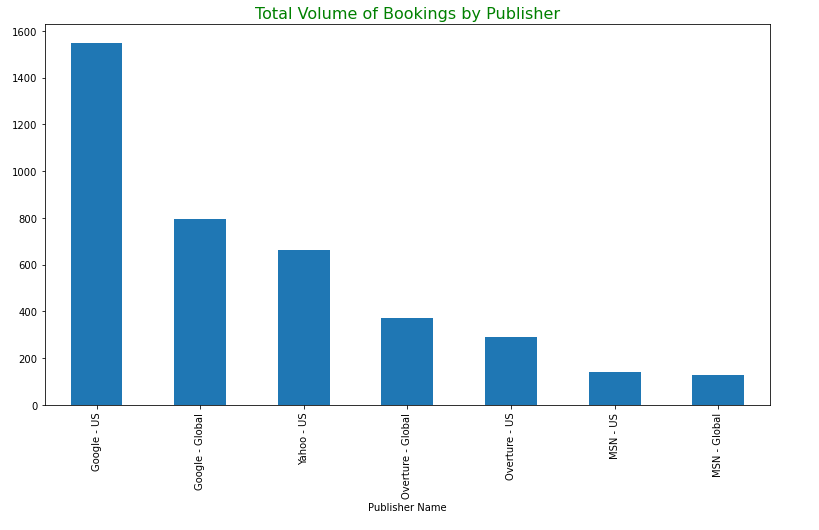
4: The data points for Avg. Cost per Click and Average Revenue per Booking are clustered about the mean because the stdev is less than the mean

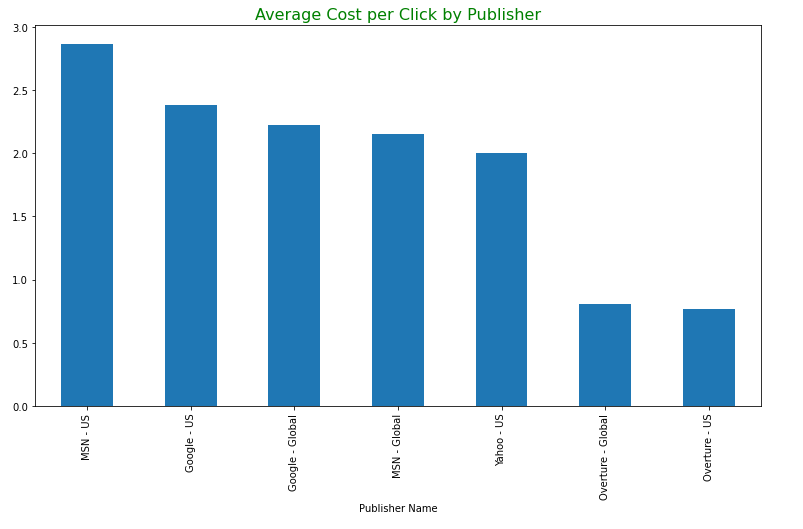
5: The mean Probability of Booking is .000681 which would indicate that the in general there is not a high likelihood of booking.

6: The max Probability of Booking is .818182 which is much greater than the mean and stdev a could potentially be an outlier.

**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.**





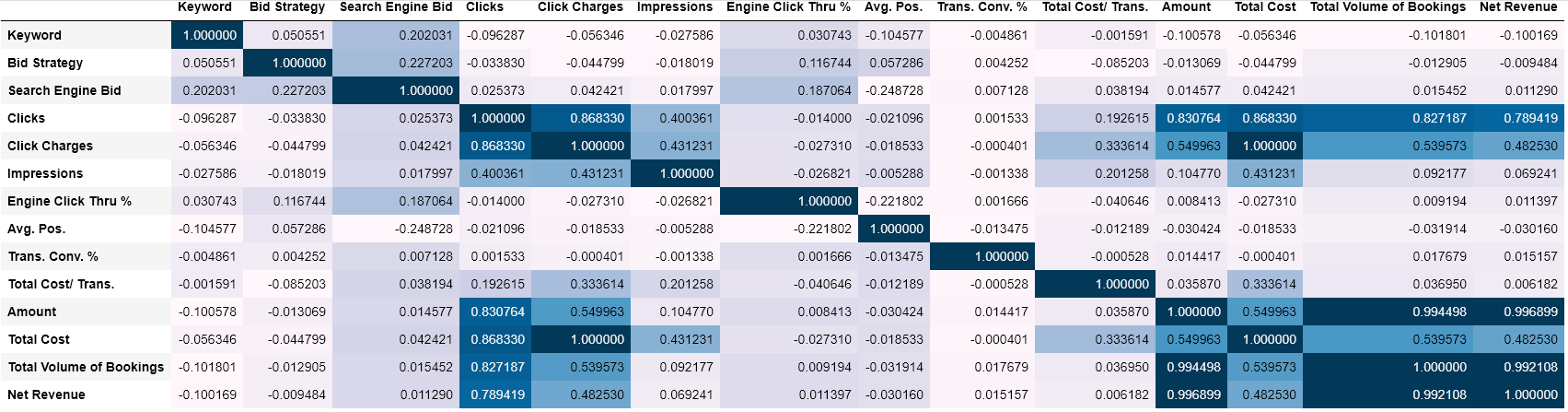


**Insight from the charts:**

* Yahoo has the highest probability of booking by a publisher and Overture-US has the lowest.
* Google US generated the highest volume of bookings by a publisher whereas MSN-US generated the lowest.
* Google Global generated the highest average revenue per booking by publisher whereas Google US generated the lowest.
* Google Global generated the highest net revenue by publisher whereas Overture US generated the lowest.
* Yahoo US generated the highest net revenue by publisher and keyword whereas Google US generated the lowest
* MSN US has the charges the most cost per click whereas Overture US charges the lowest.

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.**

***Correlation Matrix:***



***Interpretation:***

Clicks, Impressions, Total Cost/Trans., Amount, Total Volume of Bookings, and Net Revenue appeared to have the strongest correlations to Total Cost.

***Multicollinearity* (Initial)**

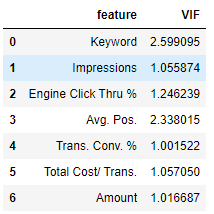


***Interpreting Multicollinearity Results:***

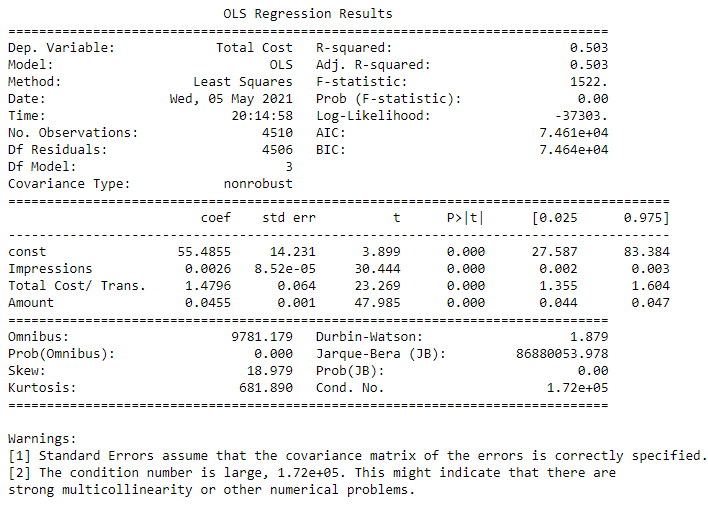
Using vif and pearson correlation it was determined that the independent variables Click Charges, Net Revenue, Total Volume of Bookings, and Search Engine Bid are highly correlated to other variables as such they will be removed.

After removing different variables through trial and error we revised the dataset to also remove Clicks and Bid Strategy to limit multicollinearity.

***Multicollinearity* (adjusted)**



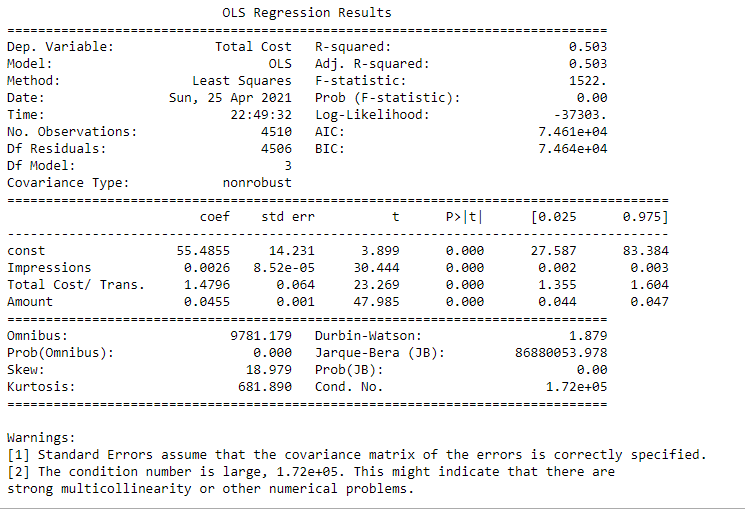
***Initial Regression Results:***



***Adjustments to independent variables used:***

Keyword, Engine Click Thru%, Avg. Pos., and Trans. Conv% all had p-values of greater .05 so they were removed due to no statistical significance.

***Final Regression Results***:



**Please Report**

**1) The final set of independent variables you have chosen and why you have chosen them.**

The final set of independent variables that we used were ‘Impressions’, ‘Total Cost/Trans’, and ‘Amount’.

Statistically:

Clicks, Impressions, Total Cost/Trans., Amount, Total Volume of Bookings, and Net Revenue had the strongest correlations to Total Cost.

Using vif and pearson correlation it was determined that the independent variables Clicks, Bid Strategy, Click Charges, Net Revenue, Total Volume of Bookings, and Search Engine Bid are highly correlated to other variables and as were removed to preserve the integrity of our regression results.

Of the remaining variables Keyword, Engine Click Thru %, Avg. Pos., and Trans. Conv. % returned a p-value of greater than .05 and were subsequently not reported due to the negligible impact on the dependent variable.

From a logical perspective:

Impressions was chosen because it records the number of times that the content was displayed. Most search engines establish their cost per a click based off the ad quality metric which utilizes, in part, the ratio of the impressions to clicks. Having too many impressions and not enough clicks would lead to a higher total cost.

Total Cost per a transaction was chosen because this metric determines whether the money that was spent on the ad was worth it.

Amount was chosen because this is a quantifiable metric showing the amount of money generated per campaign.

**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?)**

Total Cost = 55.4855 + 0.0026(Impressions) + 1.4796(Total Cost/Trans) + 0.045(Amount)

The value for Impressions is 0.0026 indicating a positive correlation between Impression and Total Cost. The p-value for Impressions is 0.00 which is less than 0.05 hence we can infer that the relationship both variables is statistically significant. Therefore, we can say for every 1 unit increase in Impressions the Total Cost increases by .0026 cents.

The value for Total Cost/Trans is 1.4796 indicating a positive correlation between Total Cost/Trans and Total Cost. The p-value for Total Cost/Trans is 0.00 which is less than 0.05 hence we can infer that the relationship both variables is statistically significant. Therefore, we can say for every 1 unit increase in Total Cost/Trans the Total Cost increases by 1.4796 dollars.

The value for Amount is 0.045 indicating a positive correlation between Amount and Total Cost. The p-value for Amount is 0.00 which is less than 0.05 hence we can infer that the relationship both variables is statistically significant. Therefore, we can say for every 1 unit increase in Amount the Total Cost increases by 0.045 cents.