GROUP WEB ANALYTICS ASSIGNMENT #1

Ziwei Yang G45808017 | Yuxiang Fan G34111808 | Kexin Zhang G43097333

1. (8 Points) Please first add four additional metrics (columns) into your data using the formula provided

- Net Revenue (Amount (total revenue) – Total Cost))
- 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.)
- Average Revenue per Booking (Amount /Total Volume of Bookings) (Note: if Total Volume of Bookings is 0, then set “NA” for that observation)
- 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, ROA, Net Revenue, Avg. Cost per Click, Average Revenue per Booking, and Probability of Booking.

Engine Click Thru % (CTR)		*Trans. Conv. % (TCR)*		*Return on Ad $ Spent (ROA)*	
Mean	11.14145058	Mean	0.569255075	Mean	3.414138636
Standard Error	0.301304742	Standard Error	0.206420496	Standard Error	1.083724083
Median	4.105613718	Median	0	Median	-1
Mode	100	Mode	0	Mode	-1
Standard Deviation	20.23458199	Standard Deviation	13.862485	Standard Deviation	72.77915278
Sample Variance	409.4383084	Sample Variance	192.1684903	Sample Variance	5296.80508
Kurtosis	17.52822411	Kurtosis	3934.548318	Kurtosis	1959.88253
Skewness	3.770376525	Skewness	60.92945983	Skewness	41.51899995
Range	200	Range	900	Range	3795.87027
Minimum	0	Minimum	0	Minimum	-1
Maximum	200	Maximum	900	Maximum	3794.87027
Sum	50247.94212	Sum	2567.340387	Sum	15397.76525
Count	4510	Count	4510	Count	4510
Confidence Level(9	0.590705006	Confidence Level(95.0%)	0.404685369	Confidence Level(95.0%)	2.124630491

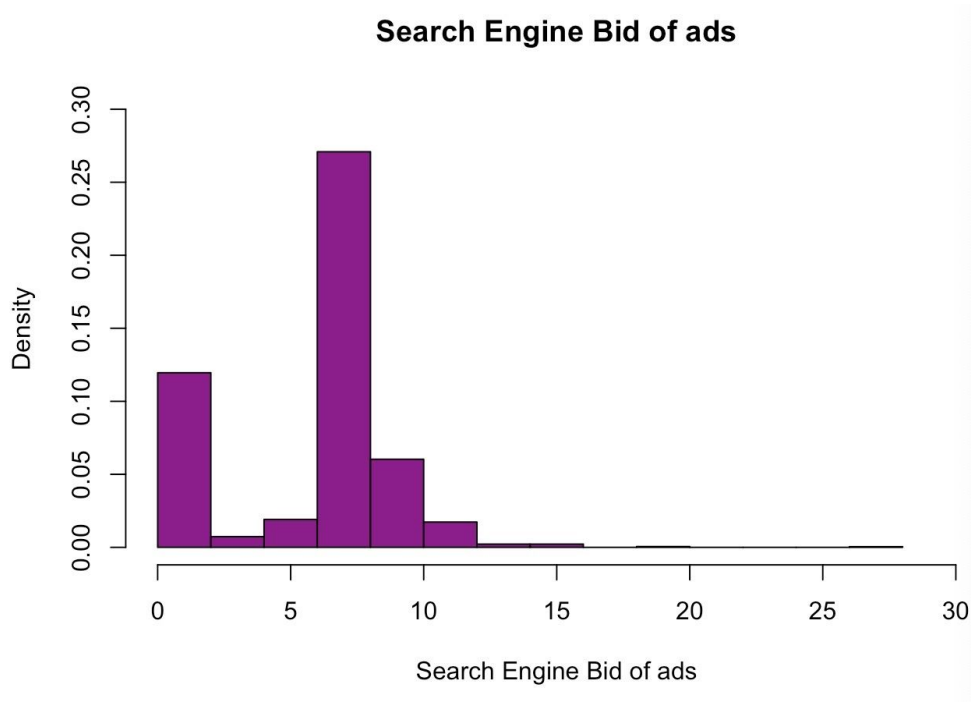
Net Revenue		*Avg. Cost per Click*		*Average Revenue per Booking*	
Mean	866.2076781	Mean	1.89023958	Mean	1024.259502
Standard Error	212.1357096	Standard Error	0.01969093	Standard Error	36.71515237
Median	-4.9875	Median	1.650493419	Median	899.725
Mode	-0.125	Mode	0.125	Mode	935
Standard Deviation	14246.2989	Standard Deviation	1.322374609	Standard Deviation	704.3187405
Sample Variance	202957032.4	Sample Variance	1.748674608	Sample Variance	496064.8882
Kurtosis	909.942747	Kurtosis	1.598030882	Kurtosis	8.892682027
Skewness	27.75691787	Skewness	1.091461565	Skewness	2.209180654
Range	558249.9753	Range	10	Range	5843.75
Minimum	-8725.924987	Minimum	0	Minimum	34
Maximum	549524.0503	Maximum	10	Maximum	5877.75
Sum	3906596.628	Sum	8524.980504	Sum	376927.4966
Count	4510	Count	4510	Count	368
Confidence Level(9	415.8899888	Confidence Level(95.0%)	0.038603876	Confidence Level(95.0%)	72.1984727

Probability of Booking	
Mean	6.80949E-06
Standard Error	2.25602E-06
Median	0
Mode	0
Standard Deviation	0.000151506
Sample Variance	2.29541E-08
Kurtosis	2150.549636
Skewness	43.73895451
Range	0.008181818
Minimum	0
Maximum	0.008181818
Sum	0.030710806
Count	4510
Confidence Level(9	4.4229E-06

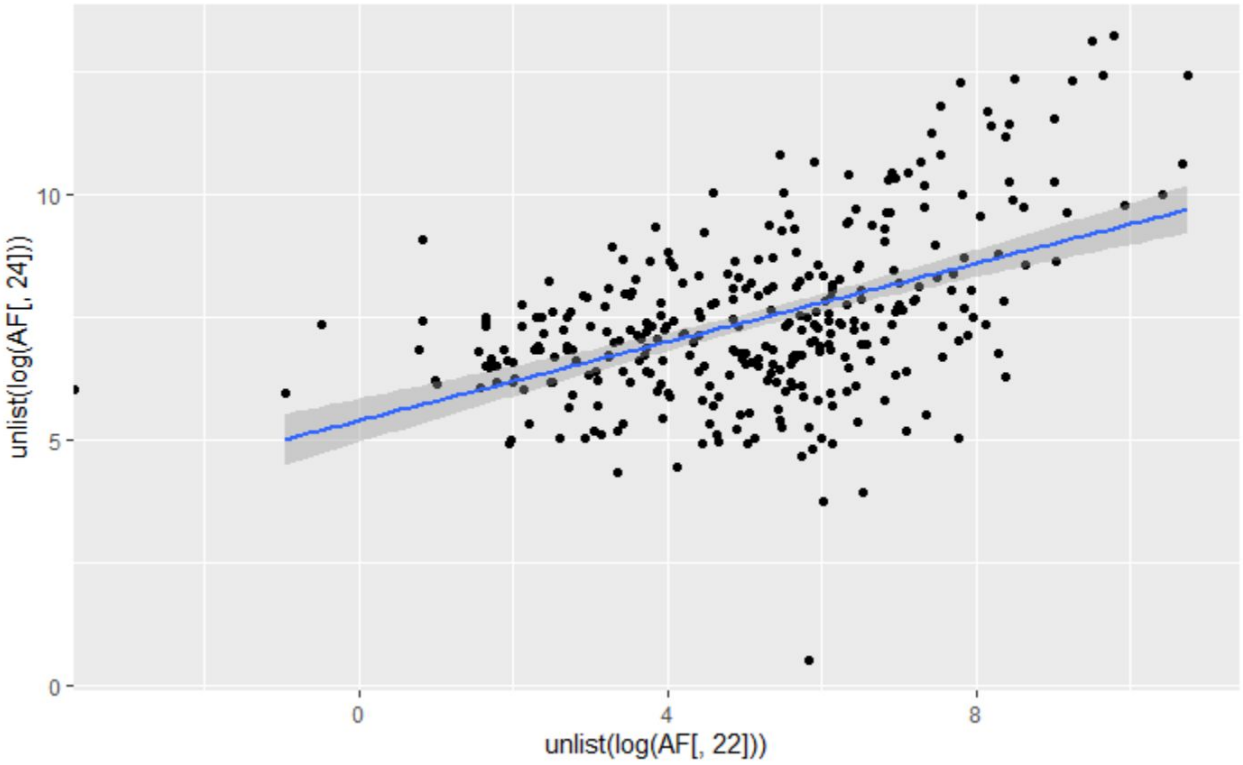
Please ONLY report a summary statistics table and provide short descriptions of your observations and thoughts.

From analyzing the summary of the dataset, we found something interesting.
For variables, CTR, TCR, ROA, the values are very small. Moreover, the frequency of Numbers with larger values is much lower than that of Numbers with smaller values.This means only a small percentage of the records have yielded significant achievements among all advertising records records. And also there is a huge difference between highest value and lowest value. Here the probability of booking is really low that the largest one is only around 0.008. It surprised us that the company is spending so much for such a small percentage of booking.

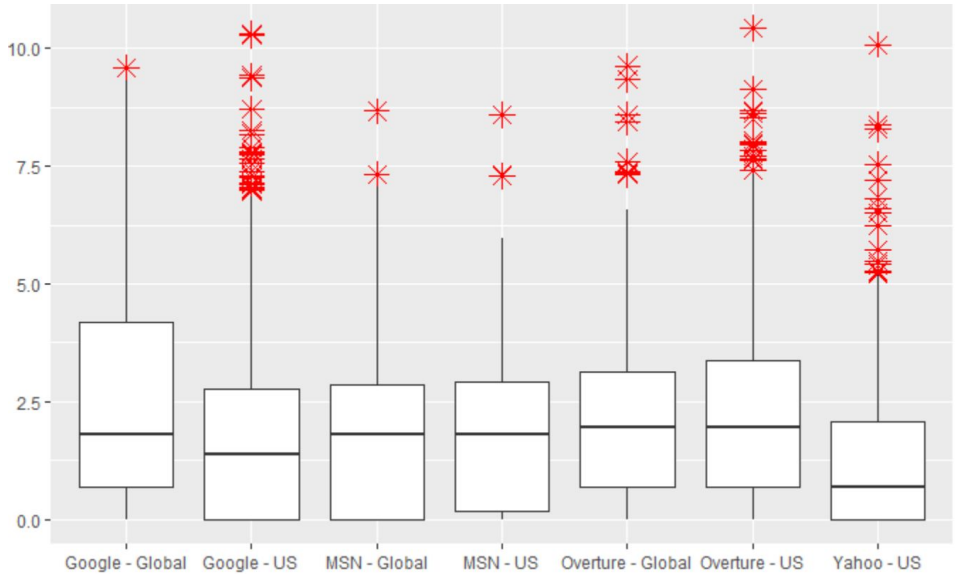
Please make a **Scatter Plot** (with Trend line), a **Histogram**, and a **Box Plot** for any of the variables **of your own interests** in the data. Then report any insights you may be able to draw from the charts.



Answer: From the figure, we can see that the bidding price of most advertisements is controlled between 0 and 12. The most frequent bid is around $7. That means it's a price most advertisers can afford. The number of ads bidding more than $10 has dropped significantly, while there are also some ads bid less than $0 to $2.



We drew a scarlett plot of log-transformed Total Cost versus log-transformed Net Revenue. We found that these 2 variables seem to have a positive linear relationship and an increasing trend. We can see clearly that when the Total Cost goes up, the Net Revenue also goes up comparatively. This is understandable because with more investments on ads campaigns, there could be more clicks transferred to bookings, which causes Total Revenue increasing, and Net Revenue increasing simultaneously in general.



Since the number of Clicks is too large to see a significant boxplot, we make a log transform. By using publisher name as x and log of clicks as y, and from the boxplot we can see that the

Clicks of Google-Global is the largest, and the Clicks of Yahoo is the smallest. However, the difference between clicks of different companies is small.

2.

a) (10 points) Use pivotTables to summarize metrics for each publisher. Please report the summary table including the variables as shown below. Please report your summary excel table and discuss Key Observations and Takeaways.

Publisher Name	Sum of Net Revenue	Sum of Click Charges	Sum of Clicks
Google - Global	$ 808,603.09	$ 120,946.71	72895
Google - US	$ 1,391,841.20	$ 353,640.60	192109
MSN - Global	$ 133,363.89	$ 12,160.36	11217
MSN - US	$ 165,451.31	$ 16,098.49	10808
Overture - Global	$ 365,788.84	$ 64,295.86	60899
Overture - US	$ 205,457.18	$ 141,976.07	119323
Yahoo - US	$ 836,091.13	$ 46,197.82	45598
Grand Total	$ 3,906,596.63	$ 755,315.92	512849

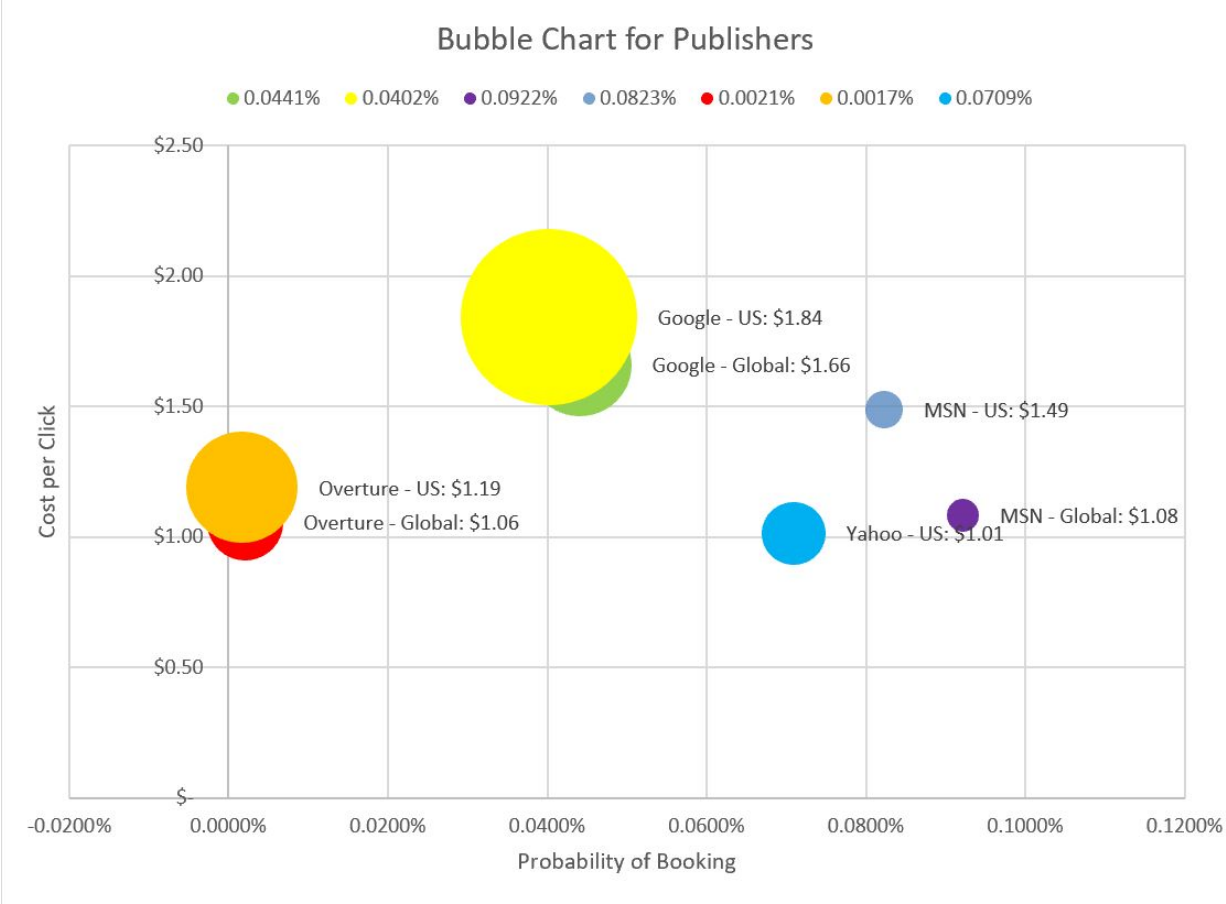
Publisher Name	Total Volume of Booking	Avg Cost per Click	Average Position
Google - Global	797	$ 1.66	1.52782221
Google - US	1550	$ 1.84	1.882806467
MSN - Global	129	$ 1.08	1.784683389
MSN - US	140	$ 1.49	1.758796362
Overture - Global	372	$ 1.06	1.928808169
Overture - US	289	$ 1.19	2.497783393
Yahoo - US	662	$ 1.01	1.789350217
Grand Total	3939	$ 1.47	1.929639723

Publisher Name	Avrg of Revenue per Booking	Return on Ad Spent (ROA)	Cost / Booking
Google - Global	$ 1,014.56	669%	$ 151.75
Google - US	$ 897.96	394%	$ 228.16
MSN - Global	$ 1,033.83	1097%	$ 94.27
MSN - US	$ 1,181.80	1028%	$ 114.99
Overture - Global	$ 983.30	569%	$ 172.84
Overture - US	$ 710.92	145%	$ 491.27
Yahoo - US	$ 1,262.98	1810%	$ 69.79
Grand Total	$ 991.77	517%	$ 191.75

Publisher Name	Probability of Bookings.	Sum of Total Cost
Google - Global	0.0441%	$ 120,946.71
Google - US	0.0402%	$ 353,640.60
MSN - Global	0.0922%	$ 12,160.36
MSN - US	0.0823%	$ 16,098.49
Overture - Global	0.0021%	$ 64,295.86
Overture - US	0.0017%	$ 141,976.07
Yahoo - US	0.0709%	$ 46,197.82
Grand Total	0.0094%	$ 755,315.92

Answer: From the table we can find that Yahoo-US with less clicks have the highest Net Revenue, since it has the highest probability of booking among all companies. Also, Yahoo-US owns the highest ROA, which means it has achieved a high efficiency in website advertising.
From the Cost/Booking and total volume of bookings variables, we can find that Google - US is the one that costs most on one booking and also the one that achieves the largest volume of bookings.
From avg of revenue per booking, we can see that MSN-Global has achieved the highest value.

b) (10 points) Based on your results in a), graph publishers on a bubble chart using the following dimensions: X=Probability of Booking, Y=Avg. Cost Per Click, Bubble Size=Total Costs or Total Funding (Sum of Click Charges). Base on your bubble chart, can you categorize different publishers into four different types 1) for search engines with High probability of booking and Low CPC; (2) for search engines with Low probability of booking and Low CPC; (3) for search engines with High probability of booking and High CPC, and (4) for search engines with Low probability of booking and High CPC. Please summarize and report your recommendations for each different type (keep your answers brief).



Answer: In our bubble plot, we draw the circle size based on the total cost of each branch of different companies.
From the Bubble chart we can see that the search engine with High probability of booking and Low CPC is Yahoo US and MSN Global. Both of the models are the one that we are looking

forward to. In between these two, the Yahoo-US has a slightly small probability per booking but also a smaller cost per click, also, the total cost of Yahoo-US is slightly larger than MSN-global.It could be something that Yahoo-US could work on. In general, They are both very competitive models.

The search engine with Low probability of booking and Low CPC is the Overture company, with both US and Global line. Although the cost per click is small for Overture company, they should work more on the probability of booking.Their probability of booking is less than 0.01% now and their total cost is not small.With cost per click similar than Yahoo-US, the probability of booking is around 0.07% smaller. This is something big that Overture could work on.
The search engines with High probability of booking and High CPC is MSN-US.We can see that when compared to the global line,the MSN-US line has a higher cost per click and a lower probability of booking.When it comes to total cost, both of the lines are similar. This is something that MSN would work on. They need to pay more attention to their US lines.

The search engines with Low probability of booking and High CPC are Google US and Google Global. Here we can see that Google has a large total cost compared with all other companies, while a 0.04% probability of booking, which is not as high as we expected, and the cost per click is also the highest among all companies. Our recommendation for google is to focus more on pushing up the probability of booking, while if possible, try to get a lower cost per click.

3. (5 points) Use the following pivot tables to study tactics of campaigns with a high ROA (last column) for Google U.S., Based on this table, please answer what are the characteristics of best campaigns such as campaign category, keyword combination, match type, bid strategy, etc.? And make your own recommendations. (note: You don't have to reproduce this pivot table, but you are more than welcome to)

Bid Strategy Data to improve campaigns within high CPC publishers

Most common characteristics of these high ROA campaigns should be considered for improving the performance of other campaigns

TOP 10 Campaigns

Publisher Name	Google - US

Average of Return on Ad Dollar Spe					
Campaign	**Keyword**	Match Type	Avg. Pos.	Bid Strategy	Total
Geo Targeted San Francisco	paris cheap airline	Broad	1.00	Position 5-10 Bid St	32237%
Air France Branded	air france us	Broad	1.02		23081%
Geo Targeted New York	france airline ticket	Broad	1.54	Position 5-10 Bid St	18307%
Geo Targeted Miami	france airfare sale	Broad	1.00	Position 5-10 Bid St	16915%
Geo Targeted DC	france flights	Broad	1.20	Position 5-10 Bid St	15214%
Geo Targeted Detroit	international airfares	Broad	3.01	Position 5-10 Bid St	13899%
Geo Targeted Boston	paris cheap ticket	Broad	1.83	Position 5-10 Bid St	13070%
Geo Targeted Houston	paris cheap flights	Broad	1.32	Position 5-10 Bid St	12895%
Google_Yearlong 2006	rabat flights	Broad	1.14	Postiion 1-4 Bid Stra	9864%
Geo Targeted Philadelphia	paris flight	Broad	1.71	Position 5-10 Bid St	8787%
Geo Targeted Chicago	paris ticket	Broad	1.26	Position 5-10 Bid St	6290%
Geo Targeted Seattle	paris tickets	Broad	1.69	Position 2-5 Bid Stra	4124%
Geo Targeted Los Angeles	france air flight	Broad	1.14	Position 5-10 Bid St	3793%
French Destinations	air france to nice	Broad	1.08	Position 2-5 Bid Stra	2312%
Paris & France Terms	air france tickets paris	Broad	1.01	Position 2-5 Bid Stra	1467%

Answer:

Here we could see from the campaign category that Google targeted in large cities tends to have a high ROA. All of them are Geo targeted, which means Geo targeted campaign worked. In the keyword column, all of them are no more than 3 words. Then for the keyword combination we could see that the key word “cheap” , “france” and “paris” has shown up a few times in the top ROA’s, which means low-priced air tickets and the tourist attraction "Paris, France" are very attractive to users. For the bid strategy, we could see that the position 5-10 bids is the most popular type around the high ROA’s. However, we cannot be sure that the position 5-10 bids has a strong correlation with a high ROA. Since the position 5-10 bids owns a very high frequency in our records compared with other levels. Last, for the match type, we cannot be sure if the Broad type helps in ROA because we have only provided with the type “Broad”.

4. (10 points) Please conduct regression analysis to study what factors influence the Total Revenue (Total Amount in the data). Basically Total Revenue is your dependent variable (Y) and your task is to determine what the important independent (explanatory) variables are. You should try different sets of independent variables in the data set to see which one(s) has significant results (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; and 2) the estimated regression equation with simple explanations for each estimated coefficient (β). (Hint: you may start with doing pair-wise correlations between Y and other variables to see what variables are significantly correlated with Y.)

Answer:
We use two functions to build models and then make a comparison.
Lasso:

```
Coefficients 67.29247 0 0 0 0 0 0 0 0 0 0 0.6563705 0.2518725 0 0 0 0 0 0 0 239.2484 0 0.7653163 0 0
> cat("Number of Zero Coefficients",sum(abs(coef(cv.out))<1e-8), fill=TRUE)
Number of Zero Coefficients 20
>
> lasso.bestlamuda<- cv.out$lambda.min
> lasso.bestlamuda
[1] 411.1642
> predict(cv.out, type = "coefficients")
25 x 1 sparse Matrix of class "dgCMatrix"
                                      1
(Intercept)                  67.2924708
(Intercept)                   .
as.factor(`Match Type`)Broad  .
as.factor(`Match Type`)Exact  .
as.factor(`Match Type`)N/A    .
as.factor(`Match Type`)Standard .
as.factor(Status)Live         .
as.factor(Status)Paused       .
as.factor(Status)Sent         .
as.factor(Status)Unavailable  .
`Search Engine Bid`           .
Clicks                        0.6563705
`Click Charges`               0.2518725
`Avg. Cost per Click`         .
Impressions                   .
`Engine Click Thru %`         .
`Avg. Pos.`                   .
`Trans. Conv. %`              .
`Total Cost/ Trans.`          .
`Total Cost`                  .
`Total Volume of Bookings`  239.2483628
`Average Revenue per Booking` .
`Net Revenue`                 0.7653163
`Return on Ad $ Spent (ROA)`  .
`Probability of Booking`      .
```

1. From the lasso function, we find that the independent variables we have chosen are: Clicks ,Click Charge,Total Volume of Booking, Net Revenue. Because these variables are statistically significant since their p-value is much less than 0.05.

```
Coefficients:
                            Estimate Std. Error    t value Pr(>|t|)
(Intercept)                -7.338e-11  1.485e-11 -4.941e+00 8.04e-07 ***
`Total Volume of Bookings`  8.970e-11  1.188e-11  7.553e+00 5.14e-14 ***
Clicks                     -4.101e-14  5.454e-14 -7.520e-01    0.452
`Click Charges`             1.000e+00  2.983e-14  3.353e+13  < 2e-16 ***
`Net Revenue`               1.000e+00  9.518e-15  1.051e+14  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.888e-10 on 4505 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 2.574e+29 on 4 and 4505 DF,  p-value: < 2.2e-16
```

2. The estimated regression equation is: Total Revenue = -7.338e-11+ (8.970e-11)*Total Volume of Bookings + (-4.101e-14)*Clicks + (1.000e+00)Click_Charge + (1.000e+00)* Net Revenue

Linear regression:

```
Call:
lm(formula = Amount ~ as.factor(`Match Type`) + as.factor(Status) +
    `Search Engine Bid` + Clicks + `Click Charges` + `Avg. Cost per Click` +
    Impressions + `Trans. Conv. %` + `Total Cost/ Trans.` + `Return on Ad $ Spent (ROA)`,
    data = AF)

Residuals:
    Min      1Q  Median      3Q     Max
-225642    -334     -45     246  207508

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                   -5.642e+02  6.507e+02  -0.867   0.3860
as.factor(`Match Type`)Broad   4.093e+02  2.599e+02   1.575   0.1154
as.factor(`Match Type`)Exact   5.202e+03  1.315e+03   3.956 7.72e-05 ***
as.factor(`Match Type`)N/A     4.537e+02  9.108e+02   0.498   0.6184
as.factor(`Match Type`)Standard -3.977e+02  2.821e+02  -1.410   0.1587
as.factor(Status)Live          8.853e+02  6.774e+02   1.307   0.1913
as.factor(Status)Paused        3.164e+00  6.379e+02   0.005   0.9960
as.factor(Status)Sent          6.266e+02  8.134e+02   0.770   0.4411
as.factor(Status)Unavailable  -3.140e+02  6.434e+02  -0.488   0.6255
`Search Engine Bid`            3.901e+01  2.980e+01   1.309   0.1905
Clicks                         2.038e+01  1.751e-01 116.419  < 2e-16 ***
`Click Charges`               -6.962e+00  1.430e-01 -48.680  < 2e-16 ***
`Avg. Cost per Click`          6.984e+01  7.752e+01   0.901   0.3677
Impressions                   -1.817e-02  5.761e-04 -31.539  < 2e-16 ***
`Trans. Conv. %`              -1.170e+01  1.223e+01  -0.956   0.3389
`Total Cost/ Trans.`           3.994e-01  4.196e-01   0.952   0.3412
`Return on Ad $ Spent (ROA)`   5.149e+00  2.333e+00   2.207   0.0273 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5854 on 4492 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.8471,    Adjusted R-squared:  0.8465
F-statistic:  1555 on 16 and 4492 DF,  p-value: < 2.2e-16
```

3. The independent variables we have chosen are: Match type, Clicks ,Click Charge, impression, return on Ad $ Spent (ROA). Because these variables are statistically significant since their p-value is much less than 0.05.

```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                   -3.361e+01  1.889e+02  -0.178 0.858807
as.factor(`Match Type`)Broad   2.720e+02  2.215e+02   1.228 0.219426
as.factor(`Match Type`)Exact   4.792e+03  1.294e+03   3.704 0.000215 ***
as.factor(`Match Type`)N/A    -6.820e-01  8.670e+02  -0.001 0.999372
as.factor(`Match Type`)Standard -5.838e+02  2.731e+02  -2.138 0.032585 *
Clicks                         2.038e+01  1.698e-01 120.026  < 2e-16 ***
`Click Charges`               -6.917e+00  1.349e-01 -51.259  < 2e-16 ***
Impressions                   -1.819e-02  5.729e-04 -31.758  < 2e-16 ***
`Return on Ad $ Spent (ROA)`   3.354e+00  1.201e+00   2.792 0.005259 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5863 on 4501 degrees of freedom
Multiple R-squared:  0.8463,    Adjusted R-squared:  0.846
F-statistic:  3098 on 8 and 4501 DF,  p-value: < 2.2e-16
```

4. The estimated regression equation is: Total Revenue = -3.361e+(2.720e+02)*(1if(match type == Broad))+(4.792e+03)*(1(if match type == Exact)) + (-6.820e-01)*(1 (if match type == N/A)) + (-5.838e+02)*(1 (if match type == standard)) + (2.038e+01)*Clicks + (-6.917e+00)Click_Charge + (-1.819e-02)*Impressions + (3.354e+00)*Return on Ad $ Spent (ROA)

From the lasso and linear regression function, we find that variables clicks and variable click charge are significant in these two models. Meanwhile, our adjusted R-squared keeps comparatively high: 80.95% of Total Revenue can be explained by variable Clicks and Click Charges. So, eventually, we choose these two variables to make a model.

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      37.1022    97.8471   0.379    0.705
Clicks           20.1857     0.1842 109.570   <2e-16 ***
`Click Charges`  -7.7552     0.1458 -53.174   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6520 on 4507 degrees of freedom
Multiple R-squared:  0.8096,    Adjusted R-squared:  0.8095
F-statistic:  9583 on 2 and 4507 DF,  p-value: < 2.2e-16
```

The estimated regression equation is: Total Revenue = 37.1022+20.1857*Clicks-7.7552*Click Charges

With holding all variables in this model (Clicks and Click Charges), the Total Revenue equals to $37.1022.

β1: By increasing additional Clicks (Clicks+1), the Total Revenue will increase $20.1857 with holding other variables constant.

β2: By increasing additional Click Charges (Click Charges+1), the Total Revenue will decrease $7.7552 with holding other variables constant.

5. (7 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.
Kayak Trans. Conv. Rate Average Publisher TCR Kayak CPC Average Publisher CPC
Compare the calculations with what you have derived from Q2, what recommendation you would like to make about marketing in Kayak relative to other publishers?

Kayak Trans. Conv. Rate =	0.073265
Average Publisher TCR =	0.569255
Kayak CPC =	$1.26
Average Publisher CPC =	$1.89

Answer:

Compare Kayak’s TCR with publishers’ average TCR, Kayak has a comparatively low TCR which shows Kayak has lower successful transaction rate after customer viewing Kayak website. Since Kayak CPC is lower than the average publisher CPC, we can understand Kayak did not invest as much as other publishers on each click for their revenue. Therefore, the reason for Kayak's lower TCR may be caused by its CPC.