Submitted by -- Monika Pawar

This assignment will analyze the data (HotelClickStream.xls) and interpret the results. This dataset includes clickstream data of online transactions for hotel booking in year 2011

***Click Stream Analytics Project Part 1***

Question 1.

1. ------------------------------------------------------------------------------------------------------------

* The below summary table is showing the top 10 domain names which generated the most volume of transactions. Here, Domain name is the website name where the transactions for booking hotels online was completed.
* The below observation table shows that ***marriott.com*** generated the most volume with 524 transactions, while ***orbitz.com*** ranks at 10 with 109 transactions.

|  |  |  |
| --- | --- | --- |
| **Rank** | **Domain Names** | **# of Transactions** |
| 1 | marriott.com | 524 |
| 2 | hilton.com | 438 |
| 3 | hotels.com | 400 |
| 4 | expedia.com | 380 |
| 5 | priceline.com | 309 |
| 6 | choicehotels.com | 297 |
| 7 | jetblue.com | 229 |
| 8 | hotwire.com | 217 |
| 9 | bestwestern.com | 120 |
| 10 | orbitz.com | 109 |

1. ------------------------------------------------------------------------------------------------------------

* The below summary table is showing the top 10 reference domain names that generated the most volume of transactions. Here, Reference domain names are the referring website name through which the final purchase website was reached.
* The below observation table shows that ***google.com*** generated the most volume with 620 transactions, while ***mywebsearch.com*** ranks at 10 with 17 transactions.

|  |  |  |
| --- | --- | --- |
| **Rank** | **REF\_DOMAIN\_NAME** | **# of Transactions** |
| 1 | google.com | 620 |
| 2 | yahoo.com | 222 |
| 3 | bing.com | 129 |
| 4 | aol.com | 53 |
| 5 | comfortinn.com | 48 |
| 6 | jetblue.com | 43 |
| 7 | qualityinn.com | 29 |
| 8 | comfortsuites.com | 22 |
| 9 | kayak.com | 20 |
| 10 | mywebsearch.com | 17 |

1. ------------------------------------------------------------------------------------------------------------

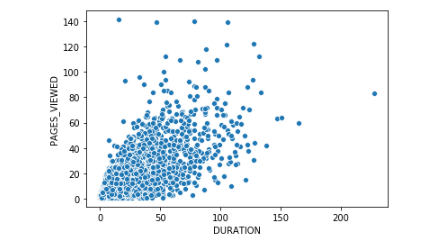
* The below summary table is showing the summary statistics (N, Max, Min, Mean, and Std.) for variables: *DIRECTP\_D; REF\_D; DURATION; PAGES\_VIEWED; LOG\_PRICE*; and *TRANS\_FREQ*
* The count indicates the number of observations in the HotelClickStream dataset. Here, total of 3749 observation are available.
* The maximum time spent at a site i.e. DURATION is 227.343 mins and the minimum time spent at a site is 0.02 mins.
* The maximum number of pages viewed at a site i.e. PAGE\_VIEWED is 141 and the minimum number of pages viewed at a site is 1.
* The maximum number of transactions for the household i.e. TRANS\_FREQ is 30 and the minimum number of transactions for the household is 1.
* 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, DIRECTTP\_D' and REF\_D' have low standard deviation, Hence the data is closer to the mean value.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Summary Statistics** | **DIRECTP\_D** | **REF\_D** | **DURATION** | **PAGES\_VIEWED** | **LOG\_PRICE** | **TRANS\_FREQ** |
|  |  |  |  |  |  |  |
| **Count** | 3749 | 3749 | 3749 | 3749 | 3749 | 3749 |
| **Maximum** | 1 | 1 | 227.34375 | 141 | 7.954386 | 30 |
| **Minimum** | 0 | 0 | 0.029998779 | 1 | 0 | 1 |
| **Mean** | 0.481462 | 0.449987 | 25.317508 | 18.331288 | 4.439926 | 2.981328 |
| **Standard deviation** | 0.499723 | 0.497559 | 22.806597 | 16.636523 | 1.615478 | 4.120927 |

1. -----------------------------------------------------------------------------------------------------------

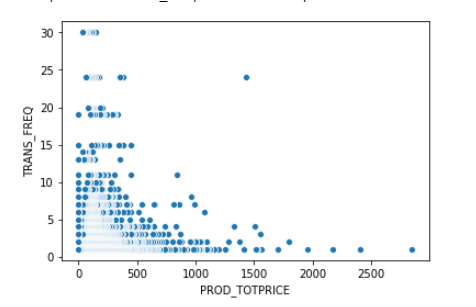
* Scatter plot:

The below scattered plot is drawn between two variables PAGES\_VIEWED and DURATION. It can be inferred from the graph that these two variables are positively correlated to each other, as the number of pages viewed increases the duration spent on the website will also increase and vise – versa.



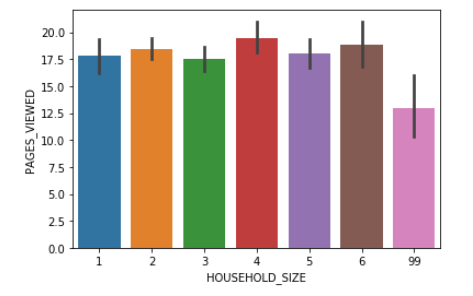
* Scatter plot:

The below scattered plot is drawn between two variables PROD\_TOTPRICE and TRANS\_FREQ. It can be inferred from the graph that these two variables are Negatively correlated to each other, as the PROD\_TOPRICE the total price paid for the transaction increases the TRNAS\_FREQ the total number of transactions for the household decreases. Here, we say that it is a negative correlation between the variables.



* Histogram plot:

The below plot is drawn between two variables HOUSEHOLD\_SIZE and PAGES\_VIEWED. It can be inferred from the graph that these two variables are not correlated to each other, as the HOUSEHOLD\_SIZE has no visual effect on the number of Pages viewed. When there is no clear relationship between the variables, we can conclude that there is no correlation between the two variables.



Question 2.

1. -----------------------------------------------------------------------------------------------------------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). Please ONLY report and interpret your final regression results.

Firstly, I have used whole dataset in my regression model. Then interpreted the P-values for each variable in the model and came to conclusion that only PAGES\_VIEWED and PROD\_QTY are statistically significant variable for final regression model.

The value of PAGE\_VIEWED in regression result is 0.9320, positive value indicating that there is positive correlation between Duration and PAGE\_VIWED. The p - value is ‘0’ which is less

than 0.05, implies that the relationship is statistically significant. Hence, we can say that with each 1 unit increase in the PAGE\_VIEWED, the DURATION increases by 0.93 minutes.

The value of PROD\_QTY in regression result is 0.6784, positive value indicating that there is positive correlation between Duration and PROD\_QTY. The p - value is ‘0.001’ which is less

than 0.05, implies that the relationship is statistically significant. Hence, we can say that with

each 1 unit increase in PROD\_QTY, the DURATION increases by 0.67 minutes.

A linear regression model with two predictor variables can be expressed with the following equation: Y = B0 + B1\*X1 + B2\*X2 + 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.

We have below values from our OLS model Result summary

const 7.0677

PAGES\_VIEWED 0.9320

PROD\_QTY 0.6784

Hence from above result, the regression equation is Y = 7.06 + 0.93(X1) +0.68(X2) + e

In the previous model, I used DOMAIN\_NAME\_CAT, PAGES\_VIEWED, PROD\_QTY as independent variable, where R-squared was same as Final model = 0.47, but the F statistic increases and the p value of DOMAIN\_NAME\_CAT was 0.586 which implies that the relationship is not statistically significant. Hence, I dropped DOMAIN\_NAME\_CAT and choose only 2 independent variables PAGES\_VIEWED, PROD\_QTY.

1. ------------------------------------------------------------------------------------------------------------Please use the linear regression technique to answer the question on “what are the factors that influence how many pages people views when visiting a website?”(Hint: use PAGES\_VIEWED as your DV). And decide your IVs by conducting the similar exercises in the above analysis. Please ONLY report and interpret your final regression results.

Firstly, I have used whole dataset in my regression model. Then interpreted the P-values for each variable in the model and came to conclusion that only DURATION and LOG\_PRICE are statistically significant variable for final regression model.

The value of DURATION in regression result is 0.4955, positive value indicating that there is

positive correlation between PAGE\_VIWED and DURATION. The p - value is ‘0’ which is less

than 0.05, implies that the relationship is statistically significant. Hence, we can say that with each 1 more minute on site will increase 0.5 pages viewed.

The value of LOG\_PRICE in regression result is -0.8132, negative value indicating that there is negative correlation between PAGE\_VIWED and LOG\_PRICE. The p - value is ‘0’ which is less than 0.05, implies that the relationship is statistically significant. Hence, we can say that with each 1 unit increase in LOG\_PRICE, there will be decrease of 0.6 pages viewed.

A linear regression model with two predictor variables can be expressed with the following equation: Y = B0 + B1\*X1 + B2\*X2 + e.

We have below values from our OLS model Result summary

const 9.3961

DURATION 0.4955

LOG\_PRICE -0.8132

Hence from above result, the regression equation is Y = 9.39 + 0.5(X1) + (-0.81) (X2) + e

In the previous model, I used TRANS\_FREQ, DOMAIN\_NAME\_CAT, DURATION, LOG\_PRICE as independent variable, where R-squared was almost same as Final model, but the F statistic are much higher than the previous model. Hence, I choose only 2 independent variables

DURATION, LOG\_PRICE.

1. ------------------------------------------------------------------------------------------------------------

Please summarize your observations by comparing the results from a) and b).

In these two models, we can see that DURATION and PAGE\_VIWED are strongly correlated. It means that more pages viewed can increase the duration and vice versa. The websites could focus on either increasing the pages viewed or duration.

To increase DURATION, main variables we should focus on improving are PAGES\_VIEWED, PROD\_QTY. Increasing product quantity will also increase the duration.

To increase PAGE\_VIWED, main variables we should focus on improving are DURATION, LOG\_PRICE. The LOG\_PRICE has the negative coefficient; websites should work on cost efficiency and lower the price of the products.