**Web and Social Analytics**

**Assignment 2**

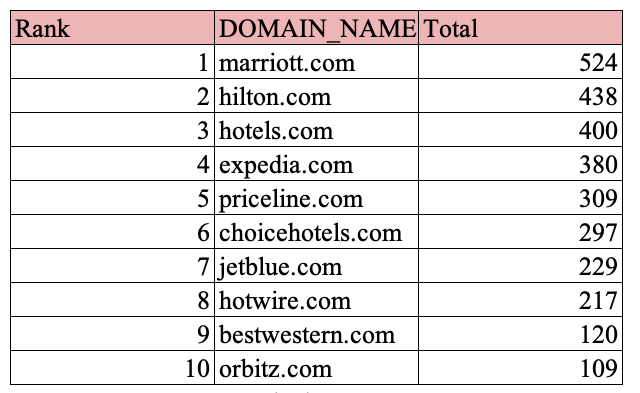
Mariam khan

Navya Manepalli

Meghana Rega

1. **(a) (4 points) Please provide a summary table showing the top 10 *domain names (DOMAIN\_NAME)* that generated the most volume of transactions the report should look like the following Table (Hint: one way to do this is to use the COUNTIF function in excel). Please summarize briefly your observations from the results**

A.



**1(a) Observations Made :**

By analyzing the clickstream data of online transactions for hotel booking ,this table shows the top 10 domain names with their respective ranks and the number of transactions.

The top 10 domain names include four official websites of the hotels mariott.com,Hilton.com,jetblue.com and bestwestern.com and the six others are travel agency or third party sites.

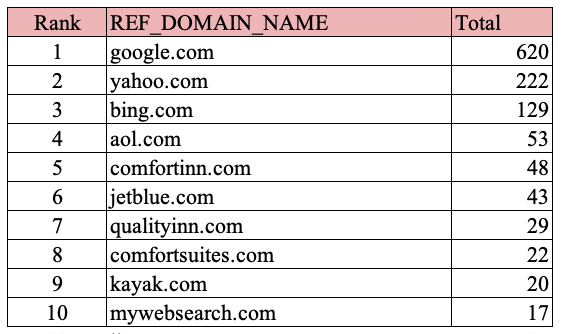
Marriot hotel’s website is found to be having the largest number of transactions **524** and Hilton.com with **438** number of transactions ranked second with a significant margin.Orbitz.com has the least number of transactions at 109.

From the table we can also infer that each domain name has achieved its market share due to the number of transactions.

With the major hotel domain names on the top we can infer that brand recognition plays a major role even though the travel agencies and the third-party have a large market share with lower prices.

**1.b)  (4 points) Please provide a summary table showing the top 10 reference domain names (REF\_DOMAIN\_NAME) that generated the most volume of transactions the report should look like the following Table. Please summarize briefly your observations from the results.**

1(b)



This table lists the top ten reference domain names that produced the most transaction volume.

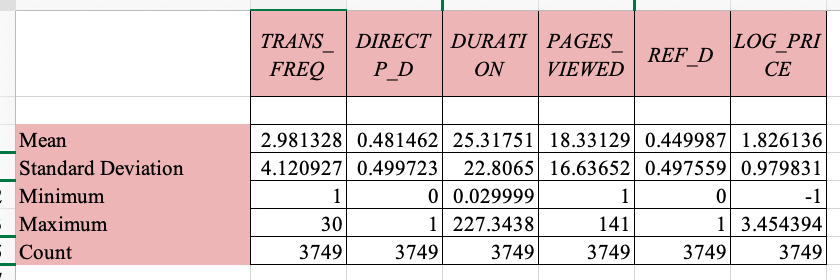
The table shows that Google ranks 1with 620 number of transactions comprising of 51.5% of the total number of transactions making it a market leader, followed by yahoo with only 18.4%.

80% of the total number of transactions comprise of googe.com,yahoo.com and bing.com.

Mywebsearch.com had the lowest number of transaction which is 17.

**1c) (4 points) Please provide summary statistics (N, Max, Min, Mean, and Std.) for variables: *DIRECTP\_D; REF\_D; DURATION; PAGES\_VIEWED; LOG\_PRICE*; and *TRANS\_FREQ*. Please report your summary statistics table and provide short descriptions (a few bullet points) of your observations.**

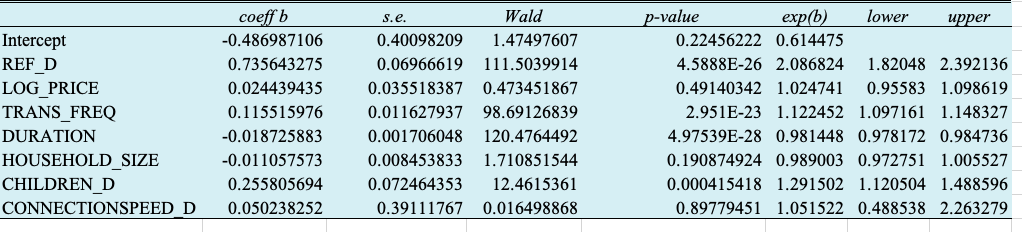
c)Summary Table

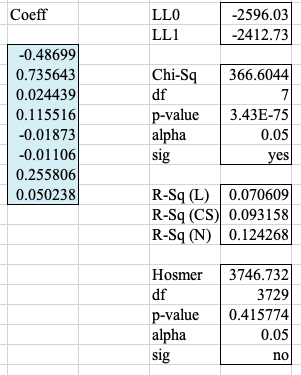
****

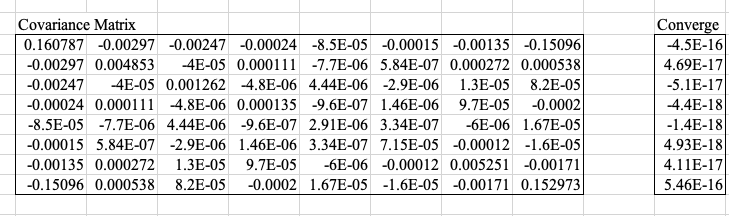
**Observations:**

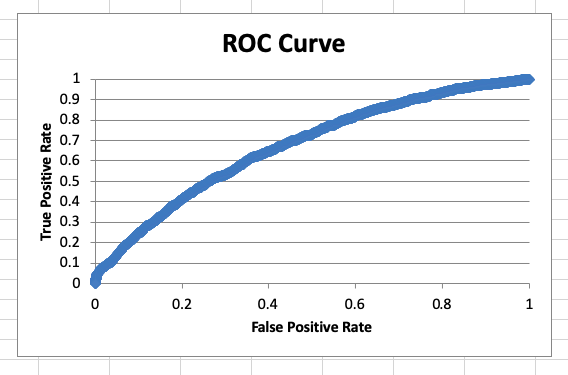
* The average transaction frequency made by customers was 2.98
* Considering the mean of DIRECTP\_D which is 0.481462 and REF\_D which is 0.449987 we can understand that less than half of the transaction volume is due to the hotel websites
* We can also infer that customers spent a maximum of 227.34 minutes and a minimum of 0.029 seconds
* The highest standard deviation is recorded by Duration which is the total time spent on a site.
* The data set has same number of values on each of the columns.
* We can infer that the maximum transaction frequency is 30 times the minimum transaction frequency.

1. **Please use the Binary Outcome (Logistic/Logit) regression technique to answer the question on “*what are the factors that influence people’s decision on whether to book directly on a hotel website or from other third party website?”* Please use *DIRECT\_D* as your Dependent Variable (DV); and *REF\_D, LOG\_PRICE, TRANS\_FREQ, DURATION, HOUSEHOLD\_SIZE, CHILDREN\_D*, and *CONNECTIONSPEED\_D* as your Independent Variables (IV). Please report and *interpret* your regression results, which should include the interpretation of the regression coefficients.**

**A.**

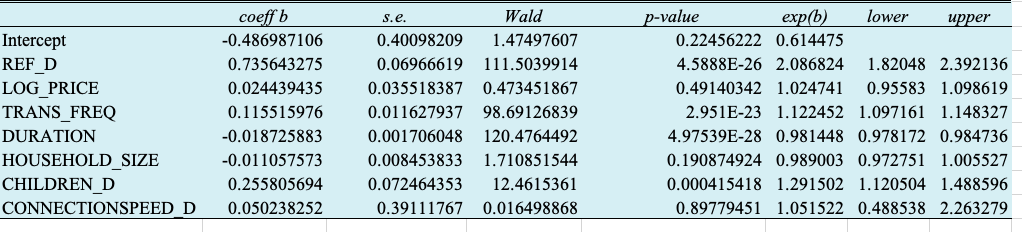






Logit regression is a type of regression analysis used for predicting the outcome of binary variables.

We employed Logit regression to understand the factors that influence peoples decision on whether to book directly on a Hotel website or from a third party website.

We used DIRECT\_D as the dependent variable. From this we tried to infer which independent variables affect the customer decisions.

Looking at the p value we can infer that there are four independent variables that have significant effect on the customer decision to book hotel from a hotel website or a third party. The p values of independent variables REFF-D, TRANS-FREQ, DURATION and CHILDREN\_D are less than 0.5 and the correlation with DIRECT\_D is statistically significant.

The following are our observations for the statistically significant variables.

REF\_D

The correlation between DIRECT\_D and REF\_D is significant as we can see that the p value is less than 0.5.The regression coefficient for REF\_D is 0.735 and increase in 1 unit of REF\_D results in log odds of the variable increased by 0.735. So, when RFD\_D is 1 then DIRECT\_D is increased by exponent of the coefficient which is e0.735 which is 2.09.This infers that there will be 2.09 or 109% higher probability that the customers will book from a hotel website rather than third party sites.

TRANS\_FREQ

Looking at this independent variable we can deduce that 1 unit increase in TRANS\_FREQ results in an increase of 0.115 in the log odds of the independent variable.e0.0115 gives 1.121 which means that for each customer transaction there will be a 1.121 or 12.1% increase in the likelihood that the transaction was from a hotel website.

Moreover ,households with a higher TRANS\_FREQ have a higher odd to book directly from the hotel’s website, since booking frequently from the same hotel can help the earn points and benefit from the loyalty programs.

DURATION

Considering the independent variable DURATION It can be observed first that the variable has a negative coefficient. So we can infer that 1 unit increase in DURATION will result in the independent variables odds falling off by -0.0187.Using it as an exponent in the equation e0.0187 which equals to 0.981 . this means one unit increase of the independent variable , there will be 0.981 lower probability that the transactions was taken from a hotel website.

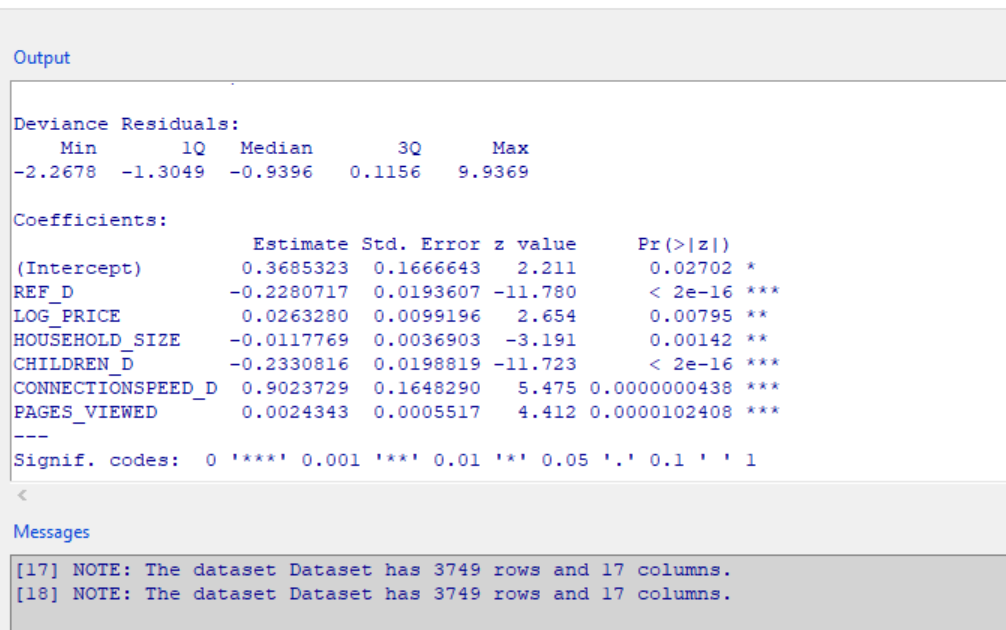
CHILDREN\_D

Considering this independent variable , it can inferred that an increase of one unit will result in an increase of 0.255 independent variable log odds. The exponent e0.255=1.29 .this 1.29 implies that there will be 1.29 or 29% of higher chances that the transaction occurred directly from a hotel website.

**3a) Please use the Count Data (Poisson) regression model to answer the question on “what are the factors that influence people’s booking frequencies?” Please use TRANS\_FREQ as your DV; and REF\_D, LOG\_PRICE, PAGES\_VIEWED, HOUSEHOLD\_SIZE, CHILDREN\_D, and CONNECTIONSPEED\_D as your IVs. Please report and interpret your regression results, which should include the interpretation of the regression coefficients.**

**b) Please repeat the analysis in question a) using the Negative Binomial Regression model. Please report and interpret your regression results and coefficients.**

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

3 a) 

After running the Count Data ( Poisson ) regression on the given data set, we observed that the p values of the IV ( Independent variable) s REF\_D,LOG\_PRICE,CHILDREN\_D,CONNECTIONSPEED\_D, PAGES\_VIEWED and HOUSEHOLD\_SIZE are relatively under 0.05. Which states that they were correlated with the DV (Dependent variable) TRANS\_FREQ mathematically. To answer to the question, yes the booking frequencies are influenced by the assumed/given factors.And, the 4 IVs(REF\_D,CHILDREN\_D,CONNECTIONSPEED\_D and PAGES\_VIEWED) are more significant. Also, among the 6 given IVs, 3 are negative and 3 are positive. More detailed observations are given below.

REF\_D : ( obtained by checking if referenced from other website, its value is 1, else 0)

The negative co-efficient value of REF\_D when correlated with the dependent variable TRANS\_FREQ is 0.228 . Which implies that when one unit of REF\_D is taken, it affects decrease in 0.228 of the log(TRANS\_FREQ) of the household. That is when REF\_D is equivalent to 1, it results in 1.26 declination of TRANS\_FREQ ( e0.23 = 1.26). For instance, when a household transaction is connected to another website other than direct booking site, 26% decrease in the total number of transaction frequency is noticed when destination booking site is directly linked.

This may be because other refencing websites have bad UI or not easy to navigate to main page. Also, It may be time taken

LOG\_PRICE: (obtained by applied log to total price paid for this transaction)

The co-efficient of LOG\_PRICE when correlated with TRANS\_FREQ is 0.026positive. Which says that when LOG\_PRICE is increased by 1 unit, it results in increase of 0.026 times of log (TRANS\_FREQ). For instance, the total transactions of household (TRANS\_FRQ) are increased by 1.026 ( e0.026 = 1.026)when the total price per transaction(LOG\_PRICE) increases by 1. Which means whenever there is one unit increase in LOG\_PRICE, TRANS\_FREQ increases by 1.026, that is 2.6% increase in transaction frequency.

This may be because, as the total price increases, there may be more savings deals offered or more points being scored which can be redeemed later so they come back and book again. Or bulk bookings may be more easier to book when planed in groups together.

PAGES\_VIEWED: (Total pages viewed at a site)

The coefficient of PAGES\_VIEWED is 0.0024 and positive. Which says it has a positive impact. That is, when the PAGES\_VIEWED is increased by 1 unit, the TRANS\_FREQ changes by 1.0024 (e0.0024 = 1.0024). It is inferred that households who are looking at more pages have 0.2% more frequency of transactions. Customers are more likely to make transactions if they visit more pages.

May be customers want to skim through more pages, and heck for better deals. As they keep going through more pages, they may find better fitting deals that they have and make more bookings. Also, they may be satisfied that this is all they have by looking at all websites and then they for sure make a decision among the options and confirm bookings.

HOUSEHOLD\_SIZE : ( Total number of people in the household)

The coefficient for HOUSEHOLD\_SIZE is 0.012 and negative. It is inferred from this that for every one unit increase in HOUSEHOLD\_SIZE, there is decrease in 1.012 units( e0.012 = 1.012) of TRANS\_FREQ. That is for every increase in the household size, there reduces the transaction rate by 1.2%. May be if there are more people living, they look for different deals or take them on spot or add extra beds in the room etc which needs in person assistance. Or they go with deals from other websites.

CHILDREN\_D: (Determines whether the household has any children or not)

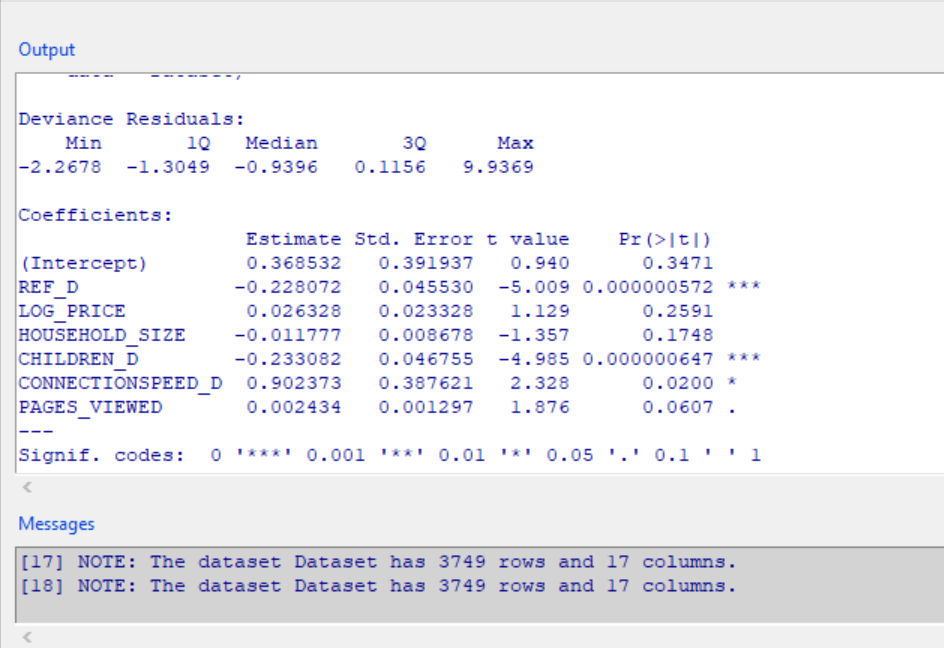
The coefficient for CHILDREN\_D is 0.233 and negative. Which means that for every one unit increase in CHILDREN\_D , there is a decrease in 1.262(e0.23 = 1.26) units of TRANS\_FREQ. It is inferred that when there are children present in the household, the transaction frequency decreases by 26%.

May be if there are children, number of vacations are less. There can be number of reasons for families with children preferring to be at home rather than handling children on trips. It also depends the age of children too. Therefore, the bookings of hotels decree with households having children.

CONNECTIONSPEED\_D: (Determines whether a household has highspeed internet)

The coefficient for CONNECTIONSPEED\_D is 0.902 and is positive. Which means that for everyone unit increase in CONNECTIONSPEED\_D, there is increase in 2.46 (e0.9 = 2.46) units of TRANS\_FREQ. That is whenever CONNECTIONSPEED\_D increase by 1 unit, the TRANS\_FREQ increase by 2.46 units. It is inferred that household with high connection speed have 146% increase in transaction frequency.

Because of high speed internet, it is easy to browse trough and complete transaction quickly. More speed more time and ease of transaction completion. Maybe they have time to skim through more websites and make decision and book the hotels from sites easily.

3b)

After running the Negative Binomial Regression on the given dataset, we obtained results as per the above figure. Our observation is that REF\_D and CHILDREN\_D are significant( since p values are less than 0.05) and are negatively related. Whereas CONNECTIONSPEED\_D is slightly significant with TRANS\_FREQ.

As per the significance, below is detailed interpretation for the IVs

REF\_D: (obtained by checking if referenced from other website, its value is 1, else 0)

As mentioned in the previous interpretation of 3a) , For every one unit change in REF\_D, there will a 1.256 e0.228=1.256), negative unit chnage in the TRANS\_FREQ. That is if the websites are referenced from other websites, the booking frequency falls down by 26%.

The reason could be the same as mentioned previously in 3a.

CHILDREN\_D:

Another significant IV. The coefficient is 0.23 and negative. Which is for every one unit increase in CHILDREN\_D, the TRANS\_FREQ decreases by 1.26(e0.23= 1.26) units. That is when there are children at household, the booking frequency decreases by 26%. Reasons can be as mentioned in 3a.

CONNECTIONSPEED\_D:

The last significant variable is CONNECTIONSPEED\_D. Very less significance, but still significant.

The coefficient is 0.9 and positive. That is when the connection speed is more in a household, they tend to book tickets more by around 146% statistically. The reasons are mentioned in 3a.

3c)

As per Poisson Regression Model, there were 4 significant variables (REF\_D, CHILDREN\_D, CONNECTIONSPEED\_D, and PAGES\_VIEWED). Among them, REF\_D and CHILDREN\_D had a negative correlation, CONNECTIONSPEED\_D and PAGES\_VIEWED had a positive correlation with the dependent variable TRANS\_FREQ.

Also, LOG\_PRICE and HOUSEHOLD\_SIZE, though the significance is not as related as the variables listed above, they still have a significant connection with the dependent variable

As per the Negative Binomial Regression Model, there are 2 very significant variables(REF\_D, CHILDREN\_D) and 1(CONNECTIONSPEED\_D) relatively reasonable significant variable. REF\_D and CHILDREN\_D had a negative correlation with TRANS\_FREQ while the independent variable CONNECTIONSPEED\_D had a positive correlation with the dependent variable.

Below is a comparison chart of the variables derived from the 2 models. There is difference in the Standard Deviation coefficient and also p-values are different for both models.

Considering the limitation of Poisson Distribution that Mean and Variance should be same and in the given dataset we see that there are a disparity and overdispersion for Variable TransFreq, Negative Binomial regression is reasonable. From looking at 3a and 3b, we see a distinction in the quantity of significant factors. (6 in Poisson and 2 in Negative Binomial Regression)

Also, there are similar coefficients for variables REF\_D and CHILDREN\_D in both regression models.

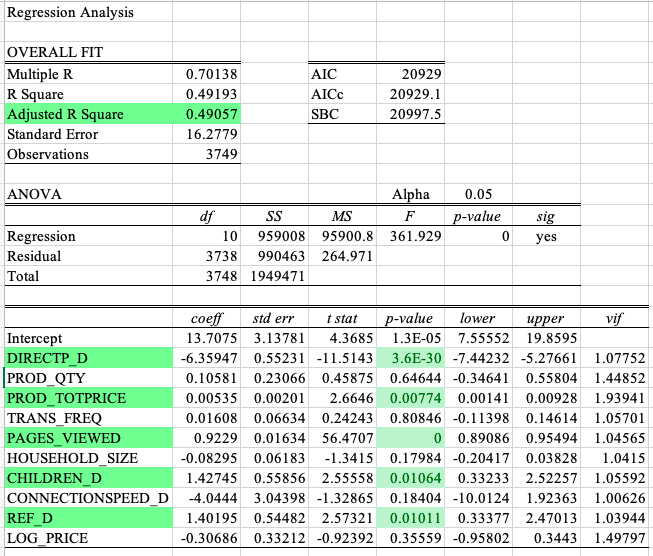
**Standard Deviation Error**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **REF\_D** | **LOG\_PRICE** | **PAGES\_VIEWED** | **HOUSEHOLD\_SIZE** | **CHILDREN\_D** | **CONNECTIONSPEED\_D** |
| Poisson Model | 0.019 | 0.0099 | 0.0005 | 0.003 | 0.019 | 0.164 |
| Negative Binomial Regression | 0.045 |  |  |  | 0.046 | 0.387 |

**Exponent of Coefficient**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **REF\_D** | **LOG\_PRICE** | **PAGES\_VIEWED** | **HOUSEHOLD\_SIZE** | **CHILDREN\_D** | **CONNECTIONSPEED\_D** |
| Poisson Model | 1.256 | 1.026 | 1.002 | 1.011 | 1.26 | 2.45 |
| Negative Binomial Regression | 1.256 |  |  |  | 1.26 | 2.45 |

**4a) Please use the linear regression technique to answer the question on “*what are the factors that influence how much time people spend on a website?”* Please use *DURATION* as your DV; and you may decide on the IVs by conducting the similar exercises in Assignment #1. Please ONLY report and *interpret* your final regression results.**



A single independent variable is used to predict the value of a dependent variable using the Linear regression technique.

We used DURATION as the dependent variable and considered DIRECTP\_D, PAGES\_VIEWED, PROD\_TOTPRICE to perform the linear regression.

1.CHILDREN\_D is statistically significant as the independent variable. CHILDREN\_D’s coefficient is 1.42 ,which means that if the household has at least 1 child, the total time spent on the website increases by 1.42 units.

2.DIRECTP\_D as an independent variable tells us that there is a 6.14 decrease in the DURATION which is the total time spent on the website when there is a one – unit increase in the transaction that occurs directly from the hotel website.

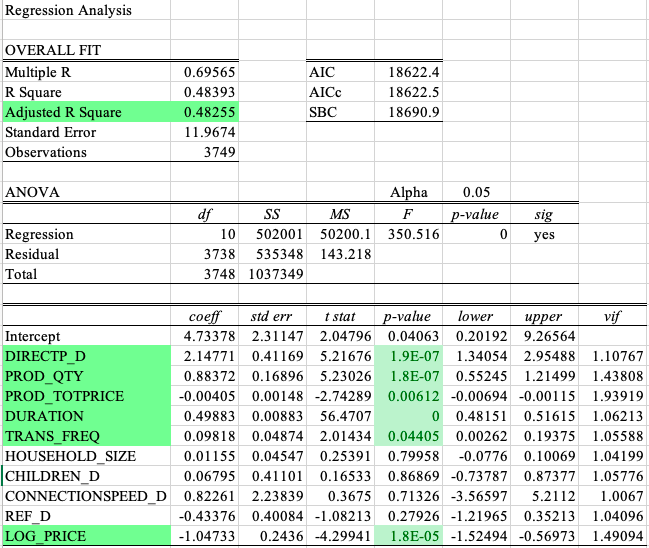
3.PAGES\_VIEWED is statistically significant as a dependent variable. We can deduct from the data that there is 0.9 unit increase in the total time spent by the customer on a website for every single unit increase in the pages viewed by the customer. The number of page views specifies the amount of time customer spends browsing and surfing the website.

4.For the dependent variable DURATION, PROD\_TOTPRICE is statistically significant independent variable. There is a 0.004 unit change in the total time the customer spends on the website for each increase in one unit of PROD\_TOTPRICE. The more time spent on the website ,the total price paid by the customers for the transactions is increased.

5. We can infer from the data that if REF\_D is increased by one unit i.e if the transaction is referred from other website , the total time spent on the website will increase by 1.40 units.

4(b)

**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?”* Please use *PAGES\_VIEWED* as your DV; and you may decide on the IVs by conducting the similar exercises in Assignment #1. Please ONLY report and *interpret* your final regression results.**

****

1. DIRECT\_D’s coefficient is 2.14 , this implies that when a unit transaction occurs directly form a hotel website the number of pages viewed increases by 2.14. At the hotel booking website, customers will be directed to multiple websites such as check out pages ,payment pages etc.
2. DURATION has a coefficient 0.49 which implies when the total time spent on a site increases by 1 unit ,the amount of pages viewed increases by 0.49 units. As the time spent researching a specific hotel reservation increases, so does the number of pages viewed by the customer.
3. LOG\_PRICE’s coefficient -1.04 which implies that when log price increases by 1 unit, the amount of pages viewed decreases by 1.04 units
4. For a unit increase in the PROD\_QTY there is a 0.88 unit increase in the number of pages viewed by the customer. If the customer makes multiple bookings , before completing the transaction, then they will visit multiple pages.
5. In case of PROD\_TOTPRICE for every one unit increase in the PROD\_TOTPRICE there is a 0.004 unit decrease in the number of pages viewed as the coefficient is -0.004
6. 0.098 is the coefficient of the TRANS\_FREQ ,This implies that for every unit increase in the transaction volume , there is 0.098 unit increase in total number of page views.

**4(c)**

**4(d)**