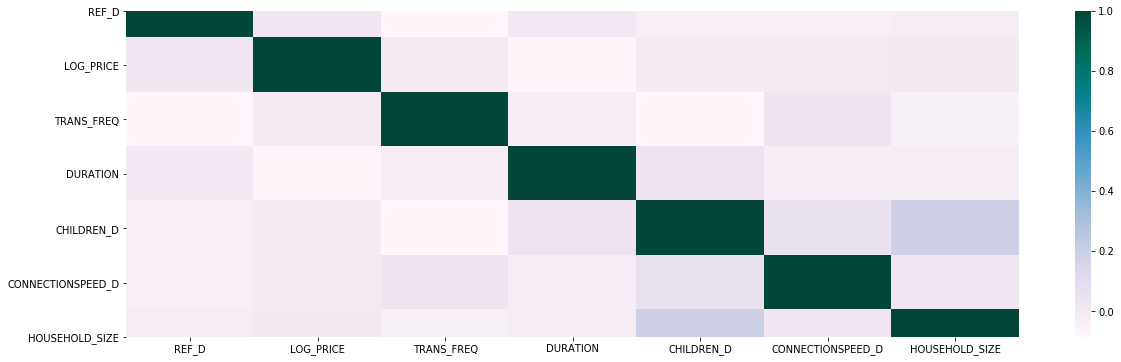
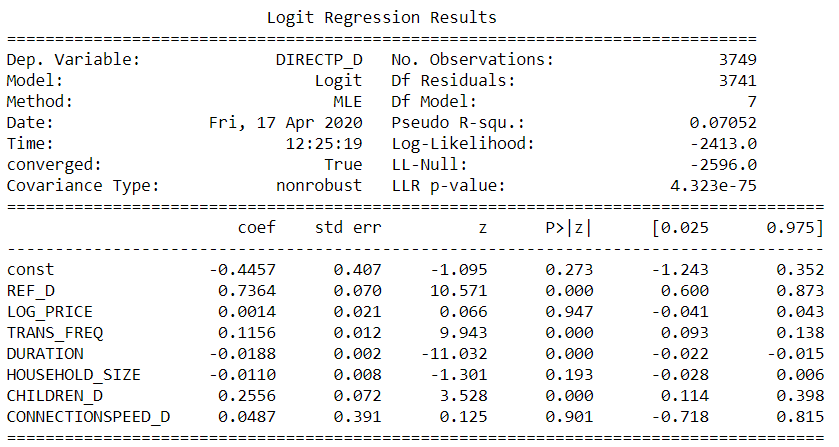
**CLICK STREAM ANALYTICS PROJECT PART 2 REPORT**

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?”*

a. Please use DIRECTP\_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).

We checked for Multicollinearity.



**b. Please report and interpret your regression results, which should include the interpretation of each of the regression coefficients.**

DIRECTP\_D = -0.4457 + (0.7364)REF\_D + (0.0014)LOG\_PRICE + (0.1156)TRANS\_FREQ + (-0.0188)DURATION + (-0.0110)HOUSEHOLD\_SIZE + (0.2556)CHILDREN\_D + (0.0487)CONNECTIONSPEED\_D

**INTERPRETATION:**

* HOUSESOLD\_SIZE. CONNECTIONSPEED\_D and LOG\_PRICE variables are excluded in the final model and equated to zero (0) since their P-values were not statistically significant
* **REF\_D:**

1. *It was directly accessed*
2. *It came from another website*

We noticed an incredibly significant relationship and we can see that the sign is positive as the coefficient is 0.7364. The odds of the transaction being incurred directly from a hotel website over the odds of the customers being referred from another website is exp (0.7364) = 2.0884. All things being equal, he odds of it coming from other website is **108%** higher than the odds of it being directly accessed.

**Insights:** From this, we endorse a marketing strategy that the hotel should pump more money into advertisement on other websites as this would be beneficial to attract more bookings. In addition, the company can also spread its reach with Search Engine Marketing (SEM) campaigns in order to optimize its performance and return on advertising ROA.

* **TRANS\_FREQ**

We noticed a significant relationship and we can see that the sign is positive as the coefficient is 0.1156. Based on our results, when there is a single increase in the number of transactions, we expect the log odds of DIRECTP\_D (whether the transaction is incurred directly from a hotel website or other third-party travel site) to increase by 0.1156. The change in odds is exp (0.1156) = 1.2225. All things being equal, this means that a customer who carries out one more transaction than another customer will have 12.25% higher odds of having the transaction being incurred directly from a hotel website.

**Insights:** From this, we suggest building several travel packages and incentives, which incorporate a greater number of transactions. For example, the company can include family packages for the summer that comprises of tours, theme parks, flights, and rentals as well as for business organizations in terms of conferences and team building. This would be beneficial to attract more potential customers to the hotel website.

* **DURATION**

We noticed a somewhat weak relationship and we can see that the sign is negative as the coefficient is -0.0188. Based on our results, where there is a **decrease** in the time spent on a website, we expect the log odds of DIRECTP\_D(whether the transaction is incurred directly from a hotel website or other third-party travel site) to decrease by 0.0188. The change in odds is exp (0.0188) = 1.0189. All things being equal, this means that a customer who spends one more minute on the site than another customer will have a 1.89% lower odds of having the transaction being incurred directly from a hotel website.

**Insights:** From this, duration can be said not to be one of the factors that influences people’s decision whether to book from a hotel website or the third-party website, as there is not a strong enough indication of the odds being swayed. We recommend that this not a good indication to use. However, we would advice that more analysis be carried out to determine the bounce rate.

* **CHILDREN\_D**

*Have Children - 1*

*No Children – 0*

We noticed a significant relationship and we can see that the sign is positive as the coefficient is 0.2356. The odds of the household having children over the odds of the household not having children is exp (0.2356) = 1.2656. All things being equal, the odds of the household having children is **26.57%** higher than the odds of the household not having children.

**Insights:** It shows when a household has children, it means that the transaction frequency is high and seeing from above analysis, that the higher the transaction frequency, the higher the odds of the transaction being incurred directly from a hotel website. We recommend that the hotel should offer more marketing incentives to cater towards children like day-care packages and theme parks packages.

**NOTE**

**Causality:** We noticed that REF\_D had a significant and strong influence on the dependent variable DIRECTP\_D (whether the transaction is incurred directly from a hotel website or other third-party travel site), this could also be as a result of the effect of **Security** and **Brand Awareness** which were not considered in the model.

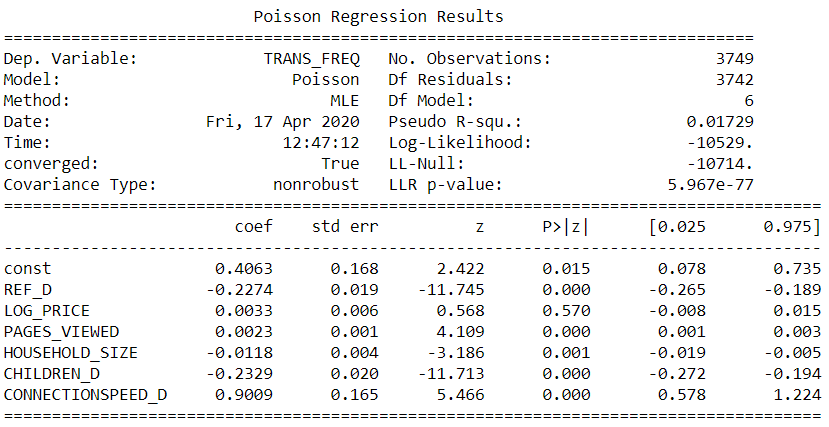
**c. (Bonus: 10 points)** **Given the regression results, your interpretation, and your experience/research on internet shopping, what kind of improvement would you make on this model? e.g. IVs to be removed or new IVs to be added? Or even other regression methodologies. This question is designed to motivate your self-exploration beyond lecture/tutorial coverage.**

Based on our results and knowledge, we believe that DURATION in this sense should not be included in the model and if there is a strong justification for it to be included in the model, then it means that data on other factors/variables such as click-through-rate, cost-per-click, bounce rate be made available for further analysis.

In addition, we also suggest that these variables LOGPRICE AND HOUSEHOLD\_SIZE, CONNECTION\_SPEED be removed in this analysis as they do not have a strong enough impact when looking at whether transactions are incurred on the website or not.

**2.** **Please use the Count Data (Poisson) regression model to answer the question on** *“what are the factors that influence people’s booking frequencies?”*

a. Please use TRANS\_FREQ as your DV; and REF\_D, LOG\_PRICE, PAGES\_VIEWED, HOUSEHOLD\_SIZE, CHILDREN\_D, and CONNECTIONSPEED\_D as your IVs.



**b. Please report and interpret your regression results, which should include the interpretation of the regression coefficients.**

TRANS\_FREQ = 0.4063 + (-0.2274)REF\_D + (0.0033)LOG\_PRICE + (0.0023) PAGES\_VIEWED + (-0.0118)HOUSEHOLD\_SIZE + (-0.2329)CHILDREN\_D + (0.9009)CONNECTIONSPEED\_D

* Mean = 2.98
* Standard Deviation = 4.12
* Variance = 16.97
* LOG\_PRICE variable has been excluded in the final model and equated to zero (0) since its P-value is not statistically significant and technically there is no enough justification to keep it.

**INTERPRETATION:**

* **REF\_D:**
  + It was directly accessed (0)
  + It came from another website (1)
* In the above model, REF\_D is statistically significant, and we can see that the sign is negative with a coefficient value of 0.2274.
* A change in REF\_D (whether the final purchase website was made through another website or directly accessed) is associated with a 0.2274 decrease in log (TRANS\_FREQ)
* TRANS\_FREQ (the number of transactions) as a result of a change in REF\_D (whether the final purchase website was made through another website or directly accessed) will decrease by 25.53%

**Insights:** In order to avoid reductions in the number of transactions, marketing campaigns should be carried out across multiple websites which enable customers view the different packages available and complete the booking or make payments accordingly. We also recommend that another marketing campaigns be carried out to drive visitors to the hotel website and converting them to customers while keeping click-costs minimized

* **PAGES\_VIEWED**
* In the above model, PAGES\_VIEWED is highly statistically significant, and we can see that the sign is positive with a coefficient value of 0.0023.
* A one-page increase in PAGES\_VIEWED (total number of pages viewed at a site) is associated with a 0.0023 increase in log (TRANS\_FREQ)
* As a result of a one-page increase in PAGES\_VIEWED (total number of pages viewed at a site), TRANS\_FREQ (the number of transactions) will increase slightly by 0.23%

**Insights:**

Although there is a positive impact of Pages Viewed on the number of transactions, it is seen that the change barely has a significant impact. So, we recommend that more analysis be conducted to determine more criteria that shows the actual engagement of customers on the website by measuring the duration, depth of visits, unique page views, relevancy and looking at the customer retention rate.

* **HOUSEHOLD\_SIZE**
* In the above model, HOUSEHOLD\_SIZE is statistically significant, and we can see that the sign is negative with a coefficient value of 0.0118.
* An increase by 1 person in the household (HOUSEHOLD\_SIZE) is associated with a 0.0118 decrease in log (TRANS\_FREQ)
* As a result of a 1 person increase in the household (HOUSEHOLD\_SIZE), TRANS\_FREQ (the number of transactions) will decrease slightly by 1.19%

**Insights:**

There is a negative impact of the size of the household on the number of transactions, however, it is seen that the change has a little impact. We endorse that more metrics be analyzed to investigate different household sizes and work towards making room sizes with different facilities available.

* **CHILDREN\_D**

Have Children - 1

No Children – 0

* In the above model, CHILDREN\_Dis statistically significant, and we can see that the sign is negative with a coefficient value of 0.2329.
* A change in CHILDREN\_D (whether the household has any children)is associated with a 0.2329 decrease in log (TRANS\_FREQ)
* The TRANS\_FREQ (the number of transactions) as a result of a change in whether the household has any children or not (CHILDREN\_D) will decrease by 26.23%

**Insights:**

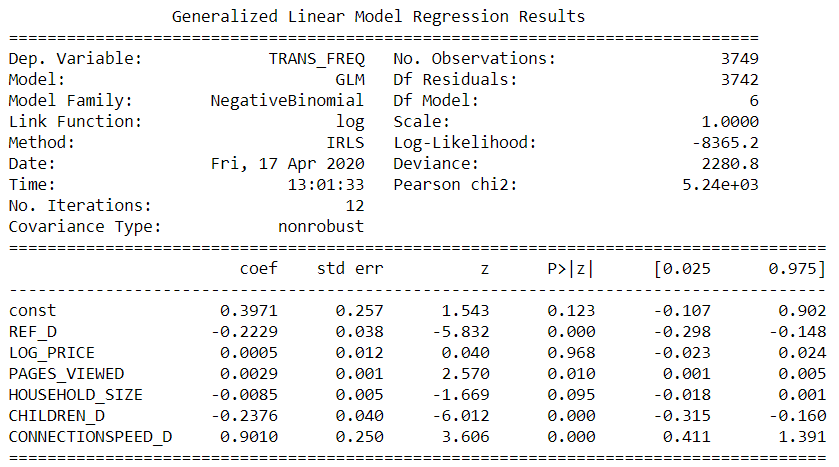
There is a negative impact of whether the household has any children on the number of transactions, it is seen that the change has a significant impact. We therefore endorse that more metrics be analyzed to investigate the actual impact on revenue that an additional child has on the hotel.

* **CONNECTIONSPEED\_D**
* In the above model, CONNECTIONSPEED\_D is statistically significant, and we can see that the sign is positive with a coefficient value of 0.9009
* A change in CONNECTIONSPEED\_D (whether the household has high speed internet or not)is associated with a 0.9009 increase in log (TRANS\_FREQ)
* The TRANS\_FREQ (the number of transactions) as a result of a change in whether the household has high speed internet or not will increase by 146.18%

**Insights:**

There is a positive impact of whether the household has high speed internet on the number of transactions, it is seen that the change has an incredibly significant impact. We suggest that a survey be carried to determine the household experience on the website and see if there are better alternatives to make sure their transactions are carried out efficiently.

**3. Please repeat the analysis in question 2. using the Negative Binomial Regression model.**



**a. Please report and interpret your regression results and coefficients.**

TRANS\_FREQ = 0.3971 + (-0.2229)REF\_D + (0.0005)LOG\_PRICE + (0.0029) PAGES\_VIEWED + (-0.0085)HOUSEHOLD\_SIZE + (-0.2376)CHILDREN\_D + (0.9010)CONNECTIONSPEED\_D

**Please note that the Insights are same as question 2b above.**

* Mean = 2.98
* Variance = 16.974
* LOG\_PRICE and HOUSEHOLD\_SIZE variables have been excluded in the final model and equated to zero (0) since its P-value is not statistically significant

**INTERPRETATION:**

* **REF\_D:**
  + It was directly accessed (0)
  + It came from another website (1)
* In the above model, REF\_D is statistically significant, and we can see that the sign is negative with a coefficient value of 0.2229.
* A change in REF\_D (whether the final purchase website was made through another website or directly accessed) is associated with a 0.2274 decrease in log (TRANS\_FREQ)
* TRANS\_FREQ (the number of transactions) as a result of a change in REF\_D (whether the final purchase website was made through another website or directly accessed) will decrease by 24.97%

**Insights:** In order to avoid reductions in the number of transactions, marketing campaigns should be carried out across multiple websites which enable customers view the different packages available and complete the booking or make payments accordingly. We also recommend that another marketing campaigns be carried out to drive visitors to the hotel website and converting them to customers while keeping click-costs minimized

* **PAGES\_VIEWED**
* In the above model, PAGES\_VIEWED is statistically significant, and we can see that the sign is positive with a coefficient value of 0.0029
* A one-page increase in PAGES\_VIEWED (total number of pages viewed at a site) is associated with a 0.0029 increase in log (TRANS\_FREQ)
* As a result of a one-page increase in PAGES\_VIEWED (total number of pages viewed at a site), TRANS\_FREQ (the number of transactions) will increase slightly by 0.29%

**Insights:**

Although there is a positive impact of Pages Viewed on the number of transactions, it is seen that the change barely has a significant impact. So, we recommend that more analysis be looked at to determine more criteria that shows the actual engagement of customers on the website by measuring the duration, depth of visits, unique page views, and looking at the customer retention rate.

* **CHILDREN\_D**
  + Have Children – 1
  + No Children – 0
* In the above model, CHILDREN\_Dis statistically significant, and we can see that the sign is negative with a coefficient value of 0.2376
* A change in CHILDREN\_D(whether the household has any children)is associated with a 0.2376 decrease in log(TRANS\_FREQ)
* The TRANS\_FREQ(the number of transactions) as a result of a change in whether the household has any children or not (CHILDREN\_D) will decrease by 26.82%

**Insights:**

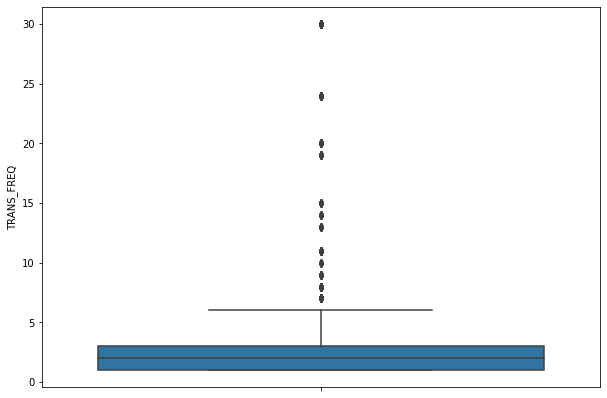
There is a negative impact of whether the household has any children on the number of transactions, it is seen that the change has a significant impact. We endorse that more metrics be analyzed to investigate the actual impact on revenue that an additional child has on the hotel.

* **CONNECTIONSPEED\_D**
* In the above model, CONNECTIONSPEED\_D is statistically significant, and we can see that the sign is positive with a coefficient value of 0.9010
* A change in CONNECTIONSPEED\_D (whether the household has high speed internet or not)is associated with a 0.9010 increase in log(TRANS\_FREQ)
* The TRANS\_FREQ(the number of transactions) as a result of a change in whether the household has high speed internet or not will increase by 146.21%

**Insights:**

There is a positive impact of whether the household has high speed internet on the number of transactions, it is seen that the change has an incredibly significant impact. We suggest that a survey be carried to determine the household experience on the website and see if there are better alternatives to make sure their transactions are carried out efficiently.

**4. Please summarize your observations by comparing the results from 2 and 3.**



The variance = 4.12^2 = 16.974

The Mean = 2.98

The above observations for the results from 2 and 3 are that (2) was analyzed using the Poisson Regression model and (3) was analyzed using the Negative Binomial Regression. Since Poisson assumes that the mean should be equals to the variance, while Negative Binomial allows that the mean be less than the variance. As a result, we adopt the Negative Binomial regression model for the TRANS\_FREQ.

We also observed that the coefficients of the two models are slightly the same however in the Negative Binomial regression model, LOG\_PRICE and HOUSEHOLD\_SIZE variables were excluded from the model since their p-values were not statistically significant while in the Poisson regression model, only the LOG\_PRICE variable was excluded.