# Inferential Statistics

## - ONLINE PURCHASE INTENTION -

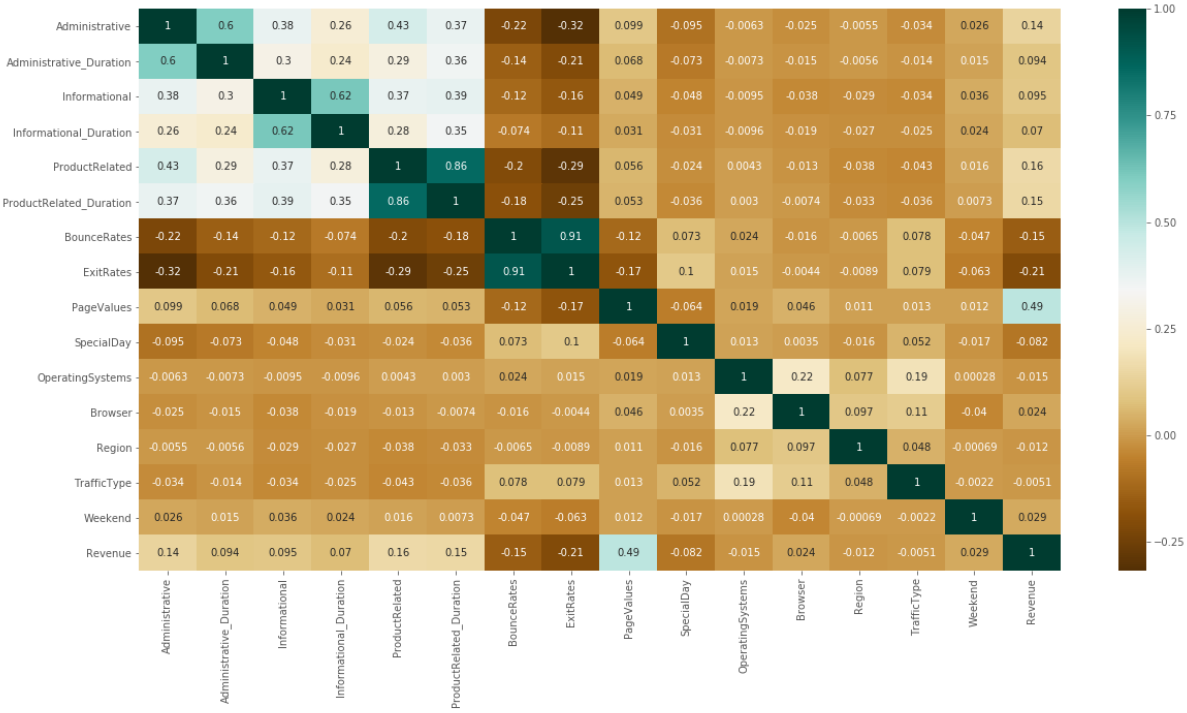
By: Tanvi Gandhi

## MOTIVATION

This document explores the possibility of identifying variables that are particularly significant to predict the Revenue target variable. It is also the aim to determine if there are strong correlations between pairs of independent variables or between an independent and a dependent variable.

## DATA INFORMATION

We analyzed the data for identifying any relationships between the different variables by constructing boxplots, scatter plots, pair plots etc. The heatmap, as shown below, was able to realize the correlation between the various variables.

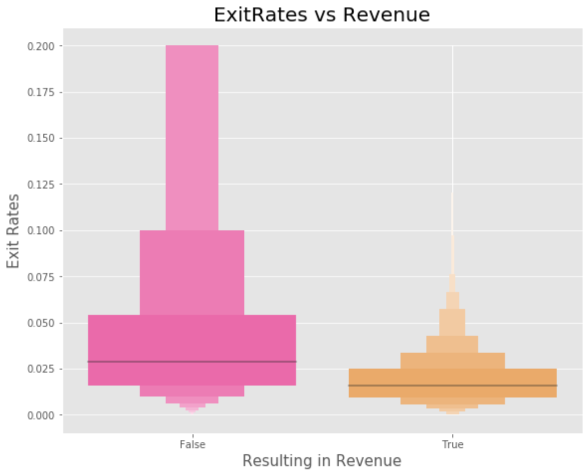


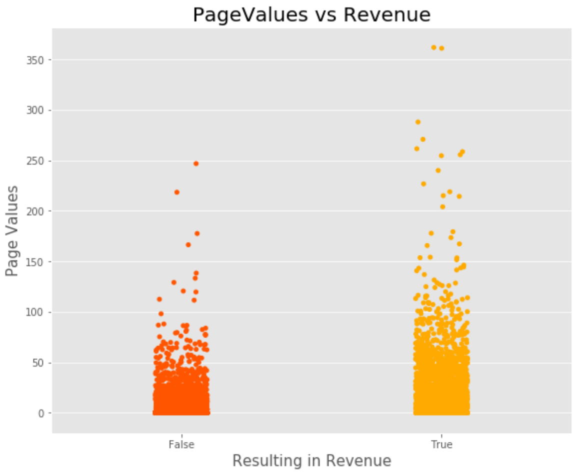
The significant observations are:

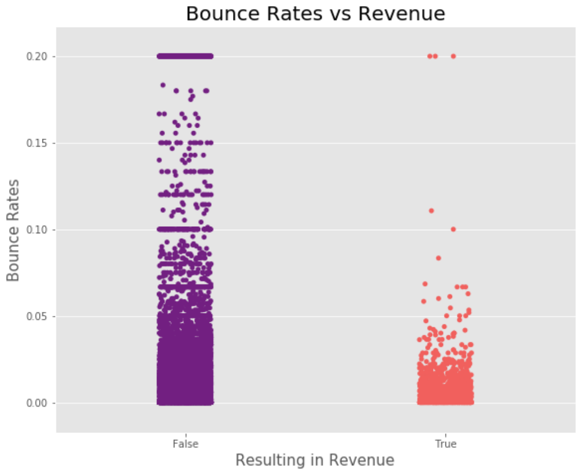
- ProductRelated and PoductRelated_Duration are highly correlated which may be an expected scenario since more the number of product pages the user will visit, the more will be the cumulative duration.
- From the above logic, we can also explain the significant relationship between Administrative & Administrative_Duration, and Informational & Informational_Duration.
- There is very high correlation between ExitRates and BounceRates since both terms intrinsically define the percentage of user visits which end up leaving a page, but in different ways.
- PageValues is a significant variable to predict the target Revenue.

## INTERESTING PLOTS

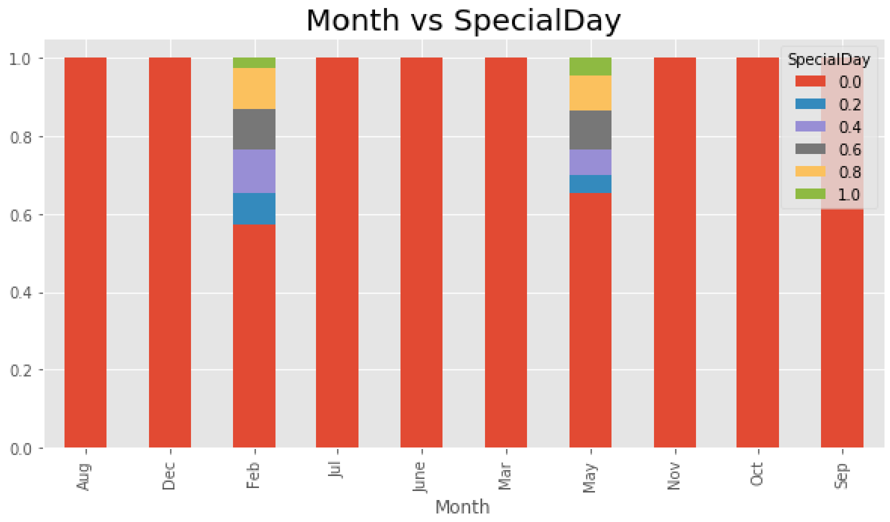
Below are some of the many interesting plots which allow us to visualize the relationships between the variables:


ExitRates vs Revenue

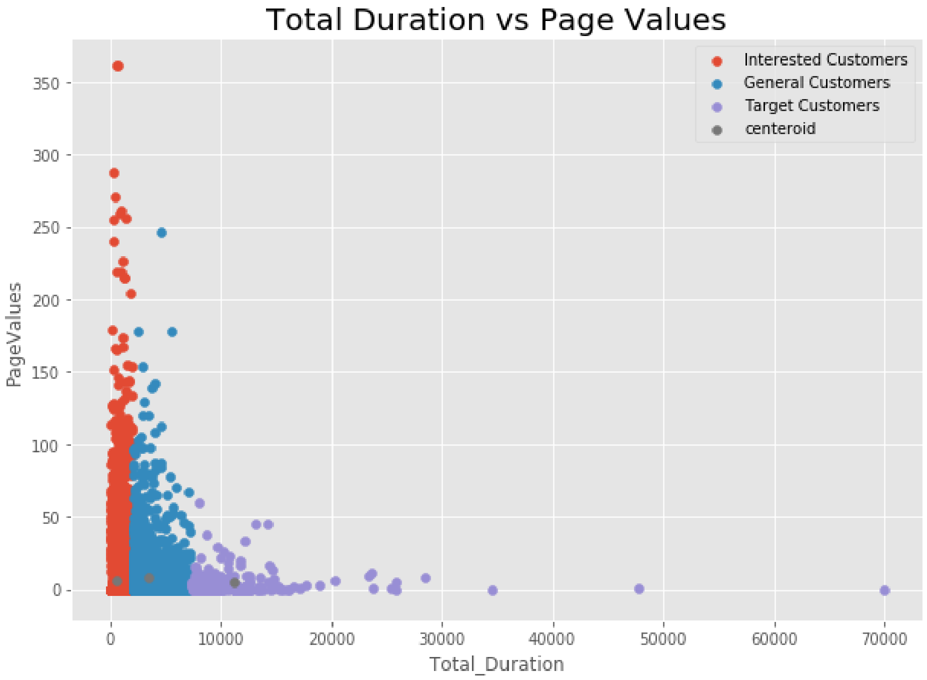

PageValues vs Revenue


Bounce Rates vs Revenue

The above three plots show the association of different page view analytics like ExitRates, BounceRates and PageValues with the target variable. Typically, the pages with average high BounceRates and/or ExitRates are corresponding to no revenue generation. The pages with average high PageValues are resulting in revenue generation.



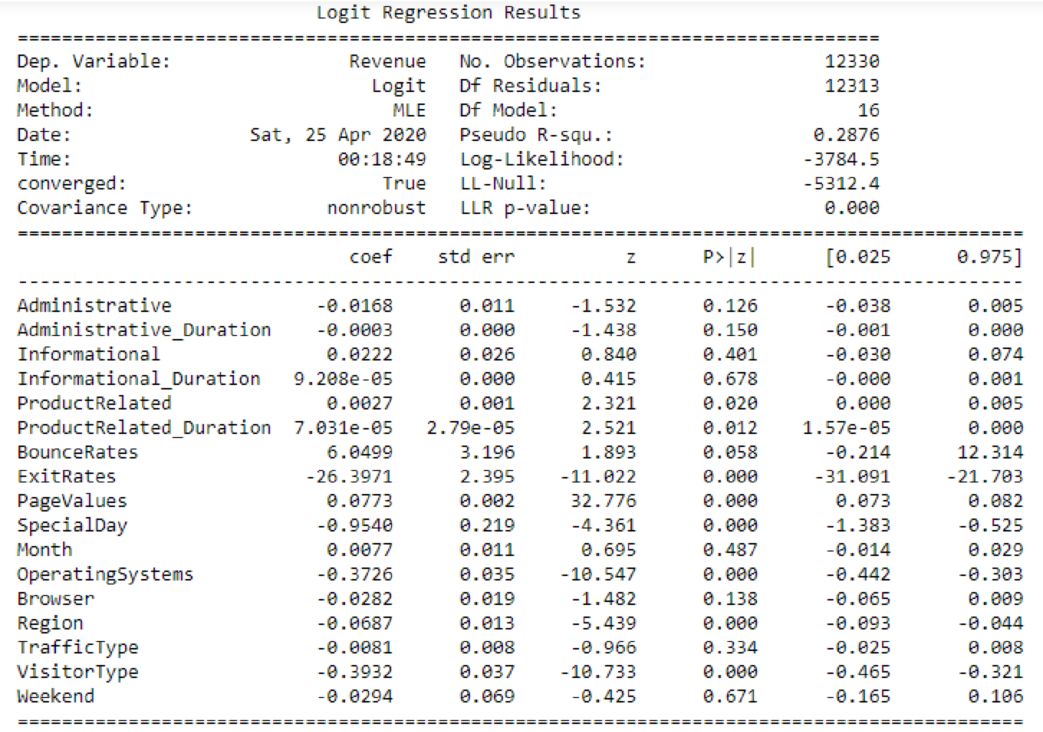
From the above graph, it is evident that the data is not available for all the twelve months. The data for January and April is not present. Further, the SpecialDay variable is not consistent for all the months as it is available for only two months. From other plots, it was observed that SpecialDay with value 0.6 was second most revenue generating after value 0.0 of SpecialDay.



In the above graph, the Total_Duration is calculated by the summation of Administrative_Duration, Informational_Duration and ProductRelated_Duration. The cluserting of different segments of the customers spending the specified total duration in accordance with the PageValues, shows the two extreme scenario. One in red reflect the customers who spend very less duration but visit pages with average high PageValues, these may be the interested customers who might visit with a specific product in mind. The other segment in the purple reflect the customers who spent a lot of time on the website but visit pages with low average PageValues, these might be the visitors who might randomly visit the website to search for a good product and end up spending a lot of time to explore the products but mostly visit pages with low average Page Values. These latter customers are our target to focus on and analyze the pages they visit.

## Logistic Regression Analysis

To analyze the P-values of the various variable with respect to the target variable, we performed Logistic Regression on the data. The resulting co-efficients were log odds (logarithm of odds of success/failure) and hence exponentiating these values would be more interpretable. The P-values are more than 0.05 for many variables, we can try to normalize the variables later on.



```
                           Logit Regression Results
==============================================================================
Dep. Variable:                Revenue   No. Observations:                12330
Model:                          Logit   Df Residuals:                    12313
Method:                           MLE   Df Model:                           16
Date:                Sat, 25 Apr 2020   Pseudo R-squ.:                  0.2876
Time:                        00:18:49   Log-Likelihood:                -3784.5
converged:                       True   LL-Null:                       -5312.4
Covariance Type:            nonrobust   LLR p-value:                     0.000
===========================================================================================
                              coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------
Administrative             -0.0168      0.011     -1.532      0.126      -0.038       0.005
Administrative_Duration    -0.0003      0.000     -1.438      0.150      -0.001       0.000
Informational               0.0222      0.026      0.840      0.401      -0.030       0.074
Informational_Duration   9.208e-05      0.000      0.415      0.678      -0.000       0.001
ProductRelated              0.0027      0.001      2.321      0.020       0.000       0.005
ProductRelated_Duration  7.031e-05   2.79e-05      2.521      0.012    1.57e-05       0.000
BounceRates                 6.0499      3.196      1.893      0.058      -0.214      12.314
ExitRates                 -26.3971      2.395    -11.022      0.000     -31.091     -21.703
PageValues                  0.0773      0.002     32.776      0.000       0.073       0.082
SpecialDay                 -0.9540      0.219     -4.361      0.000      -1.383      -0.525
Month                       0.0077      0.011      0.695      0.487      -0.014       0.029
OperatingSystems           -0.3726      0.035    -10.547      0.000      -0.442      -0.303
Browser                    -0.0282      0.019     -1.482      0.138      -0.065       0.009
Region                     -0.0687      0.013     -5.439      0.000      -0.093      -0.044
TrafficType                -0.0081      0.008     -0.966      0.334      -0.025       0.008
VisitorType                -0.3932      0.037    -10.733      0.000      -0.465      -0.321
Weekend                    -0.0294      0.069     -0.425      0.671      -0.165       0.106
===========================================================================================
```