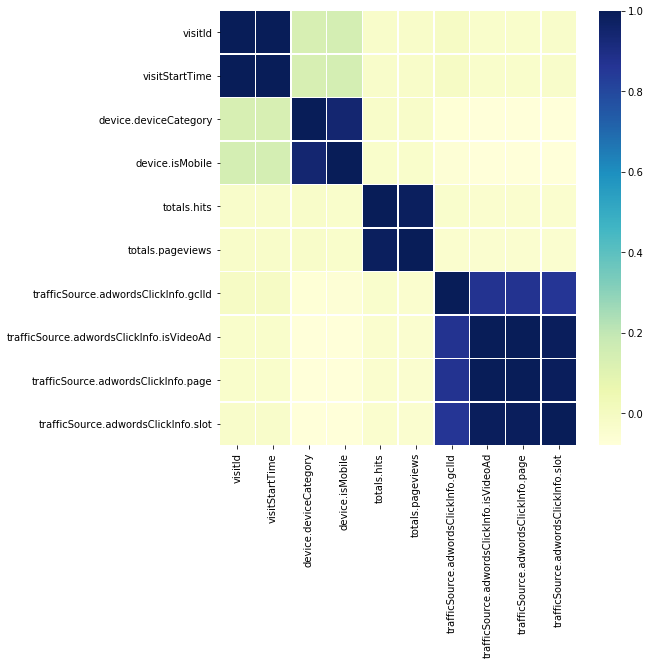
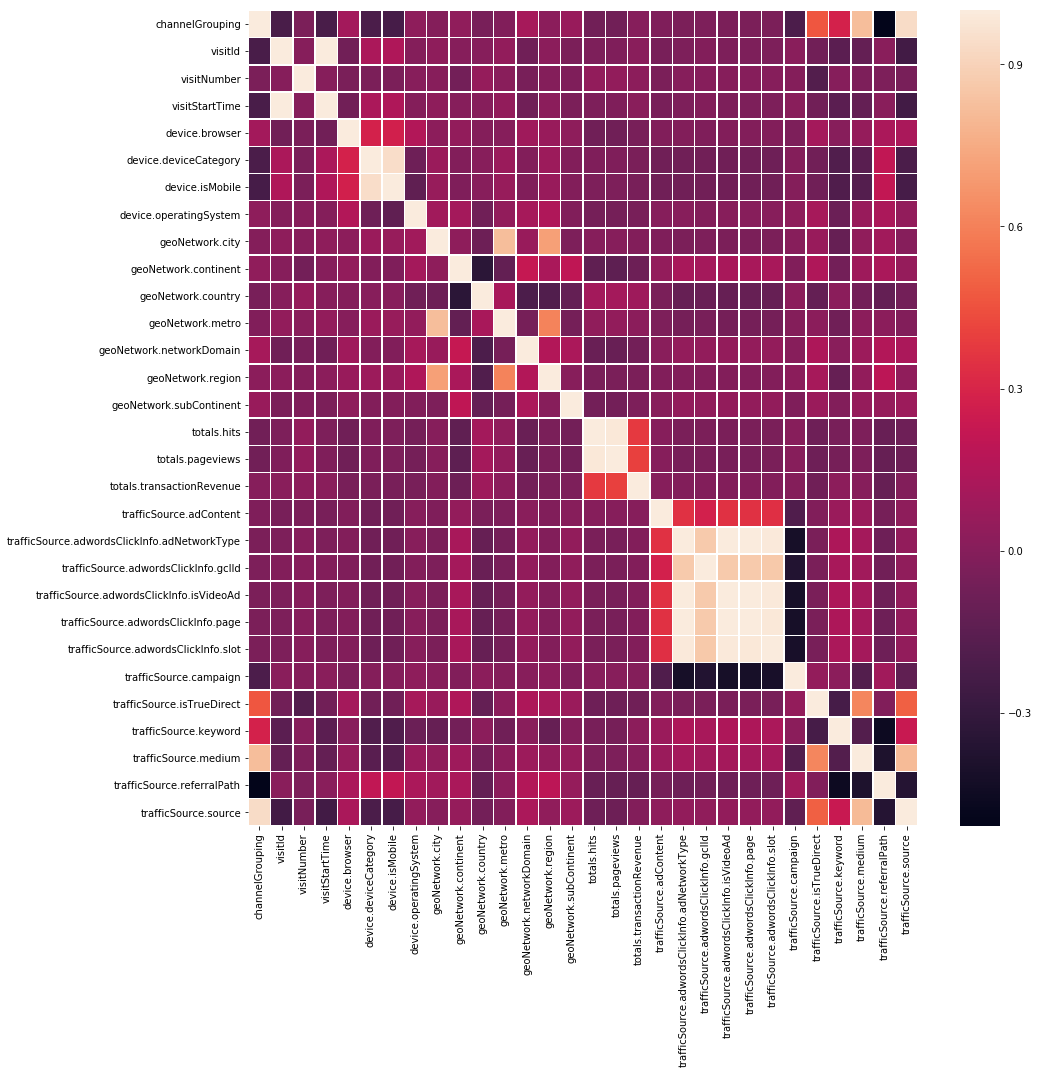
DSF Assignment 3 Questions

1. Explain what you did to clean the data. List each step/method as a separate item.

Ans: First I cleaned the dataset by getting the JSON values in the standard format. Then I replaced all the NaN values. Then I calculated the log value. Then I found out which all columns have constant values and removed those columns. Then I removed the columns that are not present in Train and Test

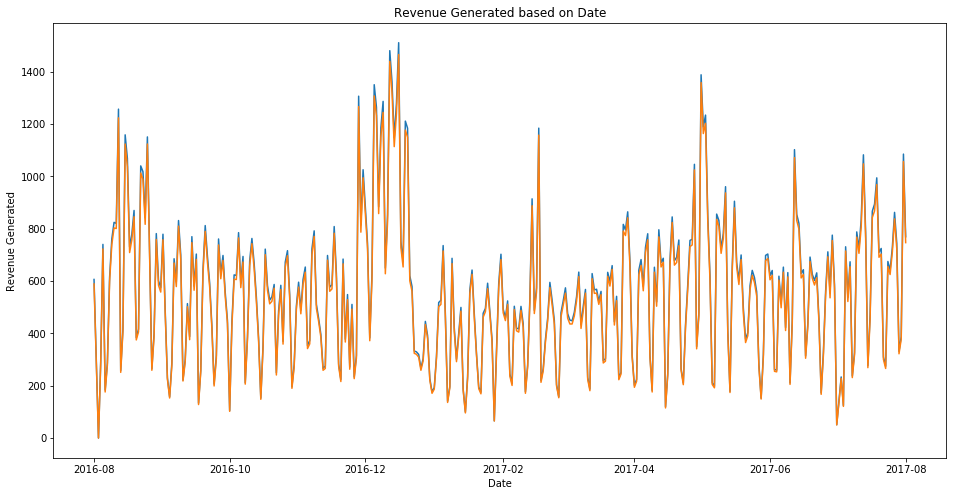
2.1.1) Upload your heatmap (or other exciting plot). 



2.1.2) Explain the reason for your choice for these variables and any interesting results associated with them. (based on 2.1.1)

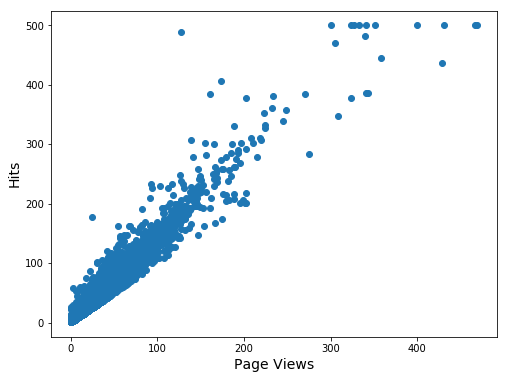
Ans: I made two heatmaps. In the first heatmap I considered all the variables. Then I took a heatmap which takes only the specific columns which are having high correlation value.

2.2.1) Upload a second plot of your choice



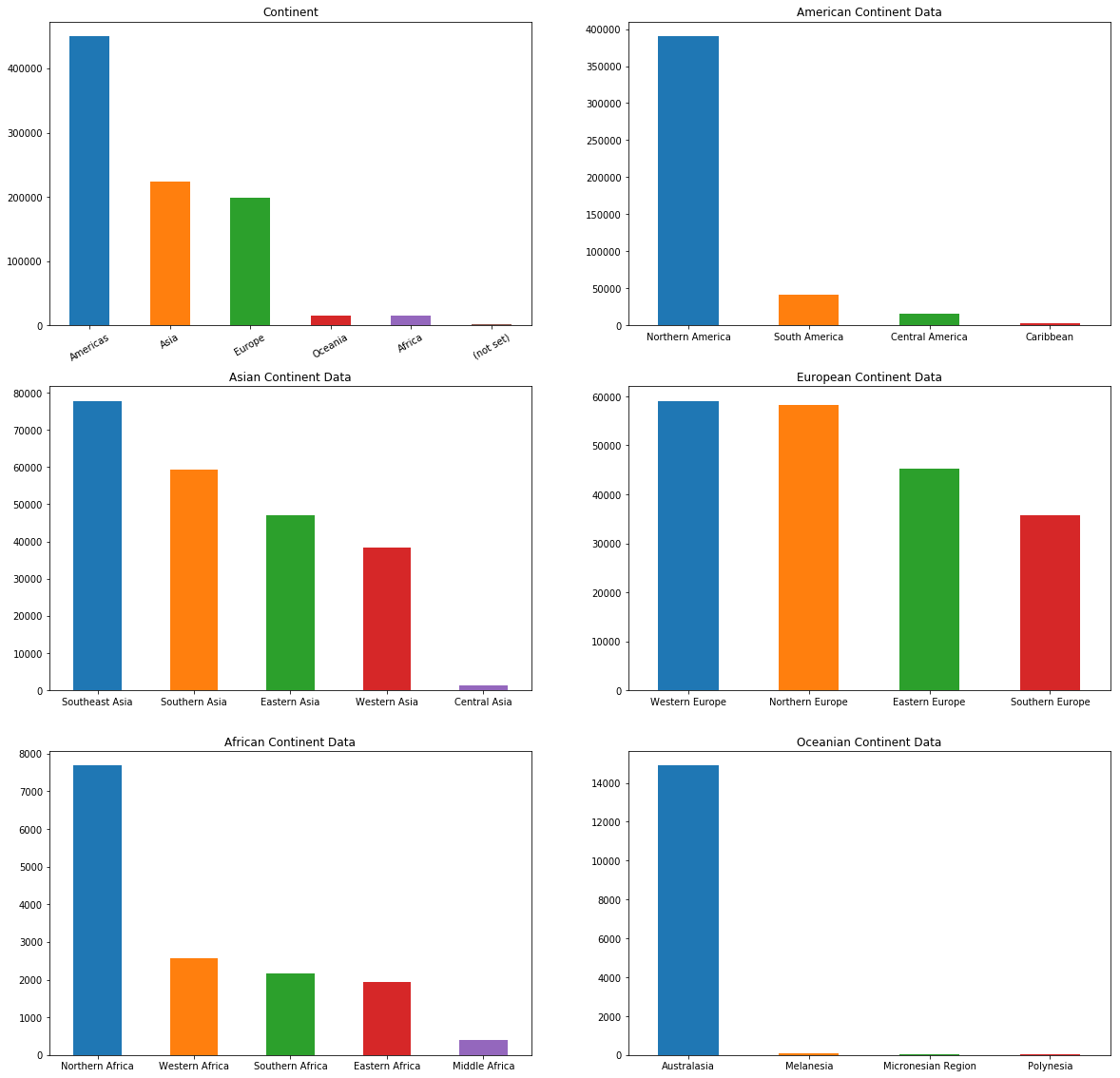
2.2.2) Explain the reason for your choice for these variables and any interesting results associated with them. (based on 2.2.1)

Ans: I generated a plot of Revenue generated with respect ot the time frame of the revenue generation. I considered this because I felt that this would show me a good relationship of how the revenue was improving or decreasing as time progressed

2.3.1) Upload a third plot of your choice.

2.3.2) Explain the reason for your choice for these variables and any interesting results associated with them. (based on 2.3.1)

Ans: I made scatter plot of Page Views v/s the number of hits for each page view and we can see that the relationship is linear amongst this data.

3.1) Upload a visualization of the clustering.

3.2) Describe your inferences from the clustering and discuss their significance. (based on 3.1)

Ans: I did a clustering of the continents and also on the basis of subcontinents, to generate a Bar graph which gives an relationship between various continents and subcontinents

4.1) Rank the ten most likely users who will buy a product from the store. List them here.

Channel Grouping: 4, 7, 4,4,1,4,2,7,6,2

4.2) Does your model produce good results? Why or why not? Explain the rationale behind your buying probability function

Ans: Yeah it produced good result for the given probability function because we used Logistic regression to generate the result after the data cleaning was done

5.1) Identify at least one external data set which you can integrate into your transaction prediction analysis to make it better.

I have taken the data from the source:  <https://www.kaggle.com/satian/exported-google-analytics-data>

5.2) Discuss/analyze the extent to which this data helps with the prediction task.

Ans: This data was merged with the existing data set on the given client ID of the external dataset and the visitor ID of the given dataset.

6.1) Report your best achieved rank

Ans: 2589

6.2) Report the score you received for your best rank

Ans: 1.75

6.3) Report the total number of entries you made during the course of this challenge

Ans: 1

7.1) For multiple (not necessarily all) relevant input variables, permute the value of each variable and see how it impact the accuracy of the results. Report the results below.

Ans: It shows how the pageviews contributes effectively the to dataset prediction than the other fields present in the dataset

7.2) Explain your process during the permutation test as well as your findings.

Ans: We permutated through each column of the given features of the given dataset. Then we find that pageviews generates the best result once the all the columns are shuffled. I made use of the eli5 library of sklearn. For high value of weight means that feature is the most important in the given list of features