**Module 4: Final Exam Report**

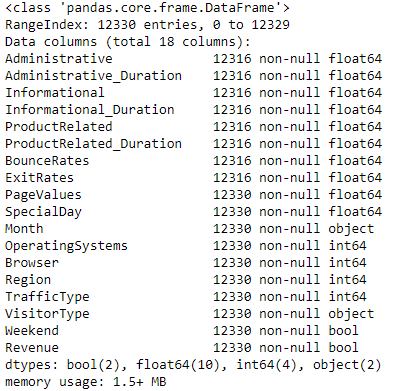
**Dataset:** ONLINE SHOPPER INTENTIONS.

**Description:**

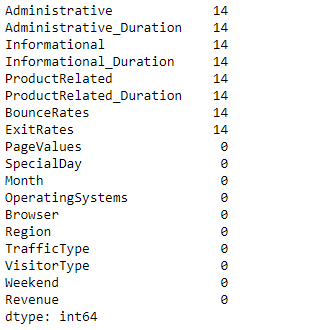
The dataset consists of feature vectors belonging to 12,330 sessions. The dataset was formed so that each session would belong to a different user in a 1-year period to avoid any tendency to a specific campaign, special day, user profile, or period.

1. **Data Cleaning and Data Munging:**

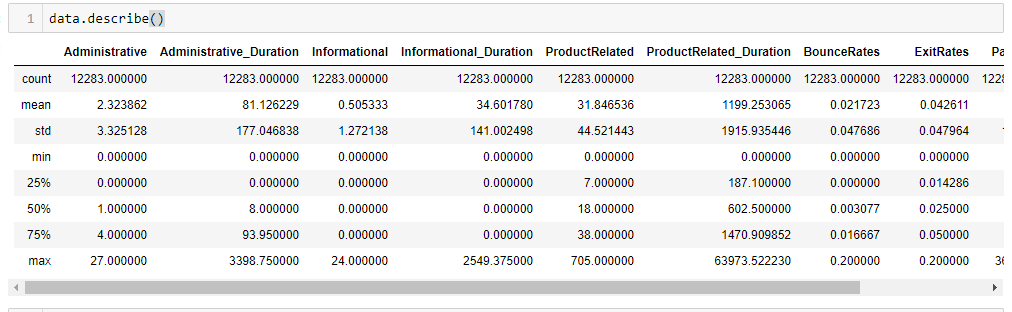
**Raw Data Info:**

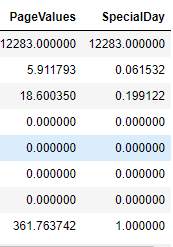


The above result shows the number of rows, number of columns, data types of columns and number of non-null rows.**Number of Null Values:**

****

There are 14 null rows in first 8 numerical features which should be removed.

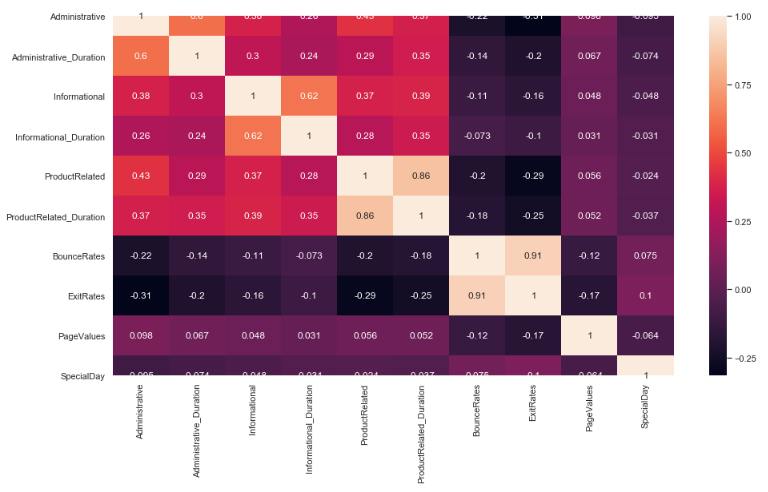
**Description of Data After Removing Null Values:**

****

The above result shows the different measures of all numerical attributes.

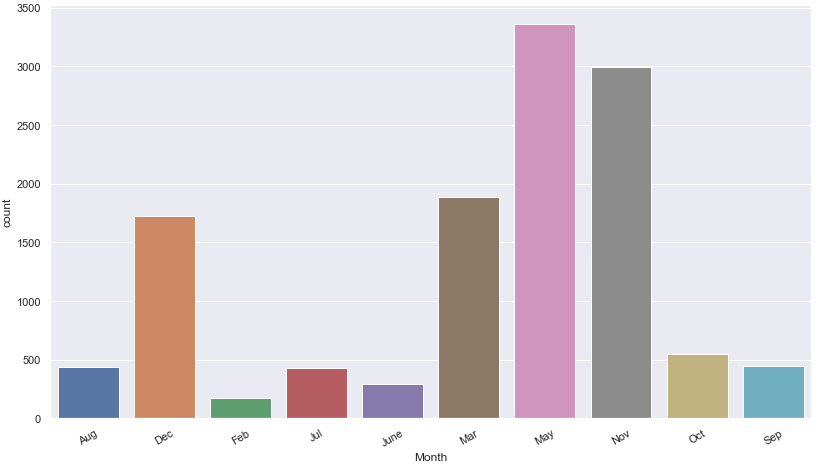
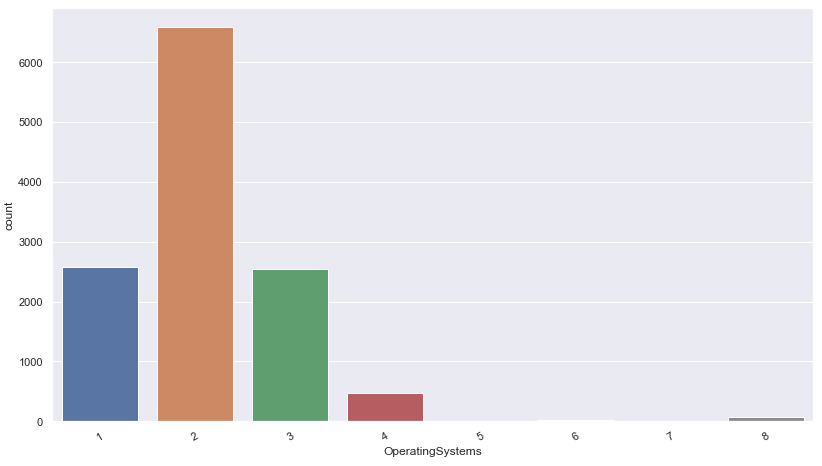
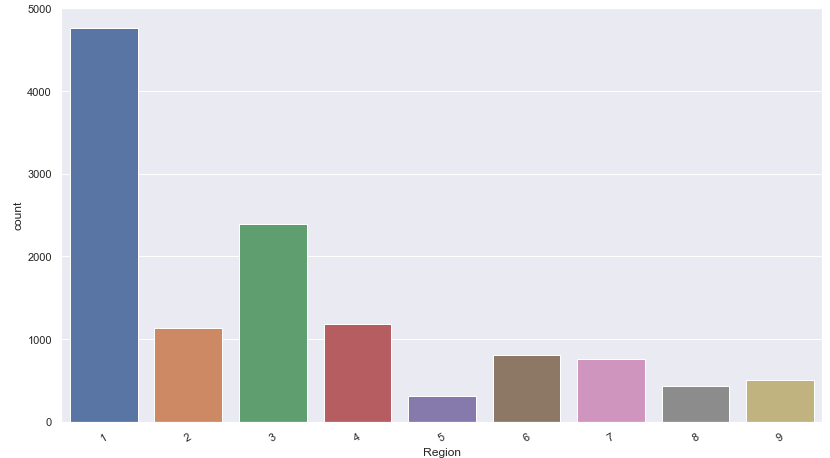
1. **EDA and Outliers Detection:**

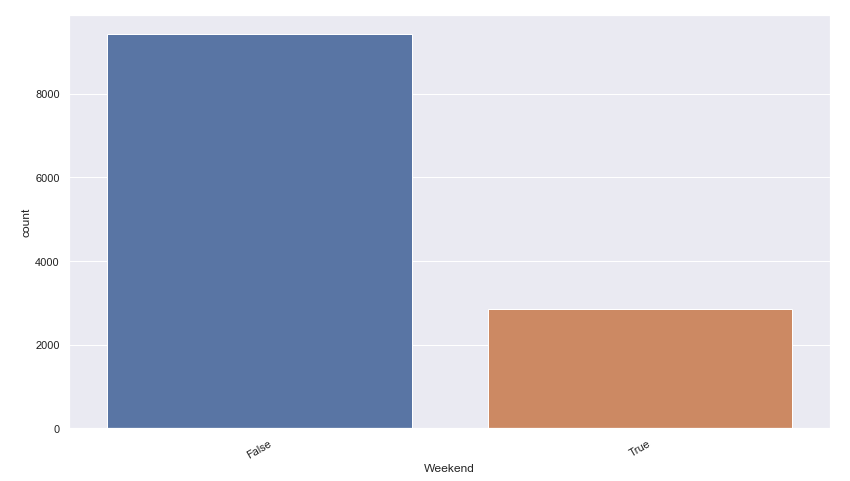
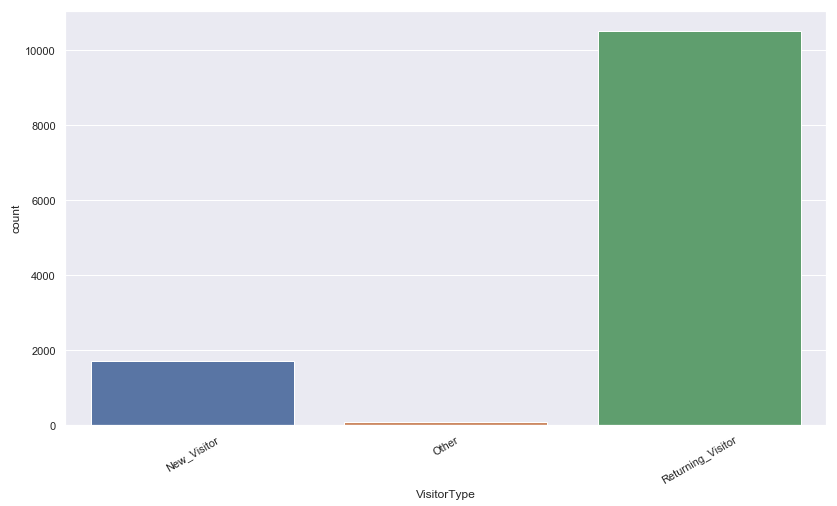
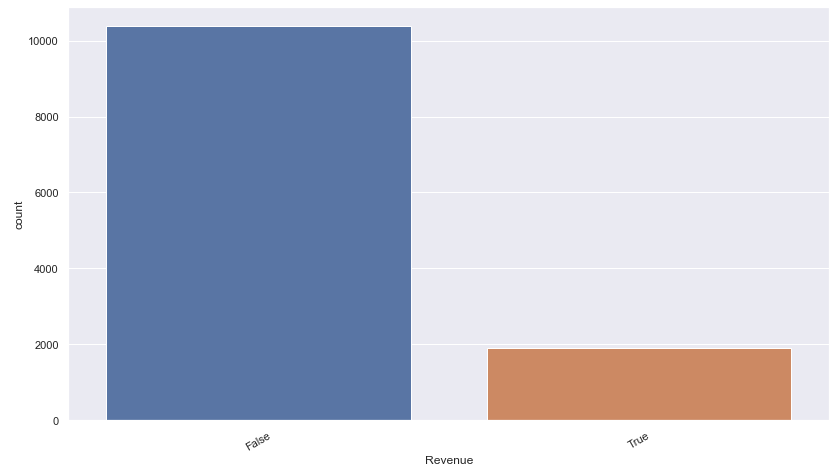
This phase will define all the analysis plots and the reason of plotting those charts which are used to analyze data.

1. **Correlation Heat Map:**

The above result shows the relationship between features. The orange color shows the positive relationship between features like revenue and page value have positive relationship. The dark purple color shows negative correlation like exit rates and product related duration have negative relationship.

1. **Univariate Analysis:**

****

****

The above plot satisfy the bellow statements

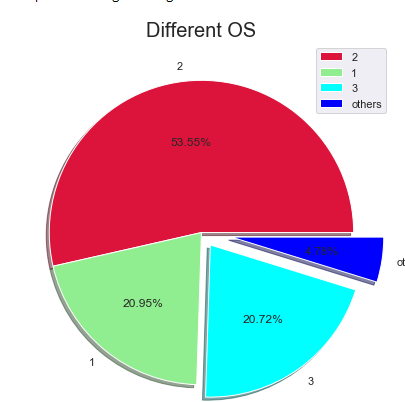
1) Most of the people visited the online shopping site in May.

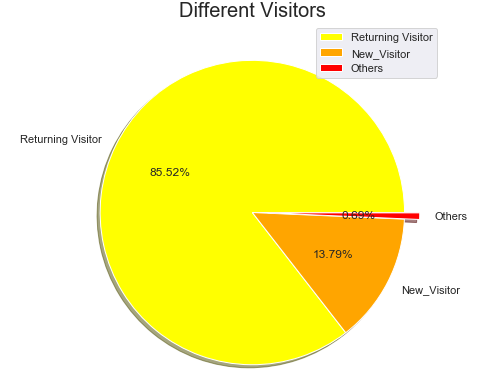
2) Shopping site is mostly visited from browser and operating system "2" in dataset.

3) People from region 1 mostly visited and made transactions

4) Mostly the visitors are returning visitors.

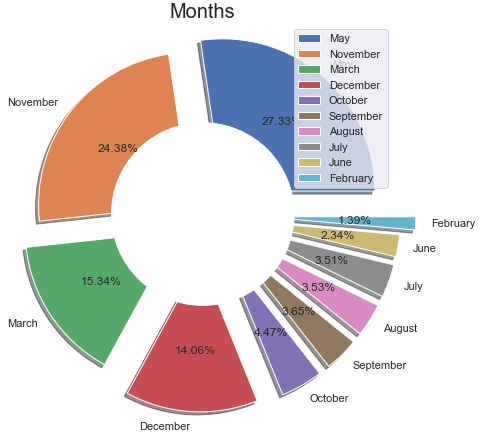
5) During weekdays, people visit the shopping sites mostly.

6) Overall revenue is very low throughout the dataset.

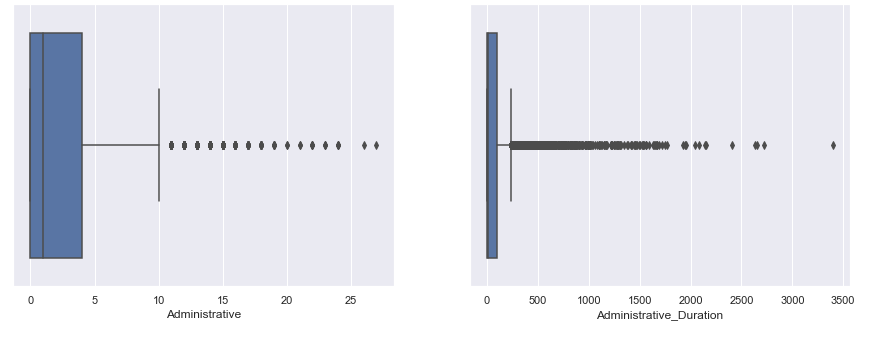


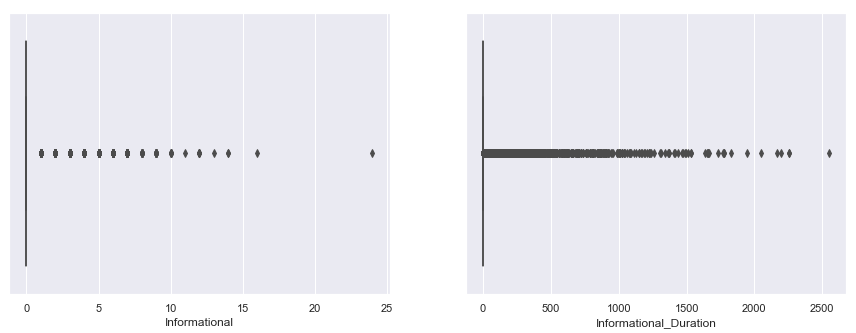
The above plots explain following things:

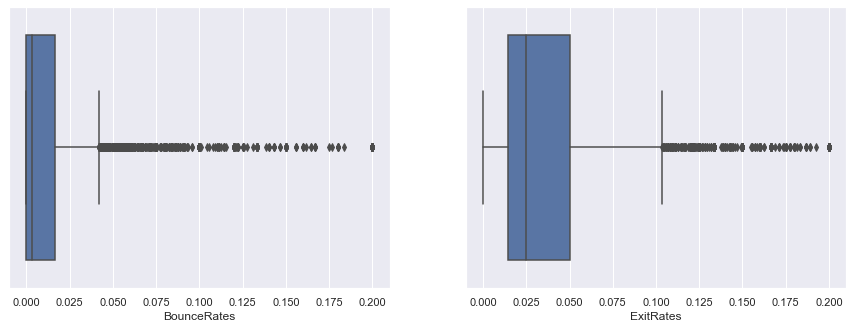
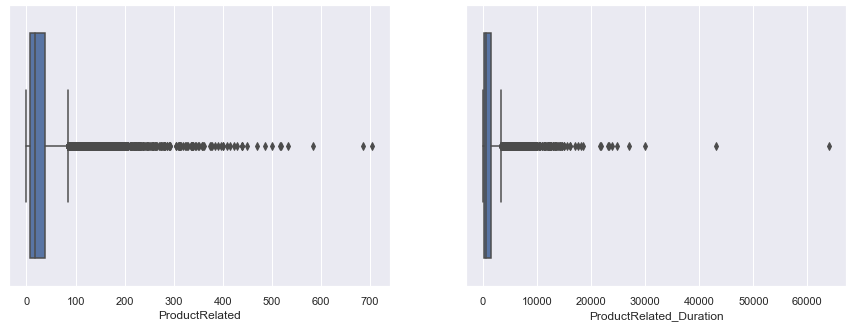
1. Ratio of Returning Visitor is 85.52%, New Visitor is 13.79% and 0.69% others.
2. Ratio of browser 1 is 20.95%, browser 2 is 53.55%, browser 3 is 20.72% and 4.78% others.

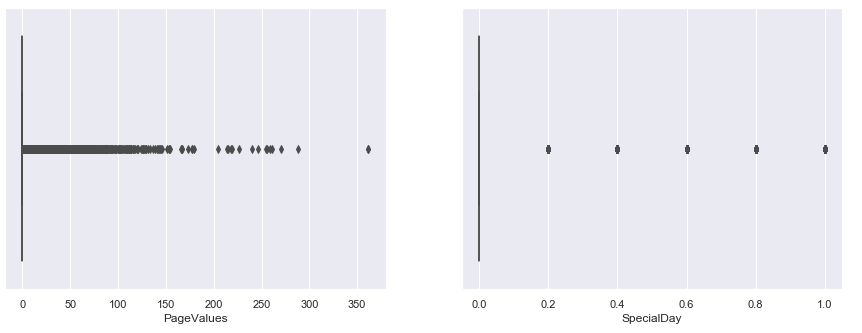


The above plot shows the distribution of months in the dataset in percentage. Most of the people visited the online shopping site in May, November, March and December.



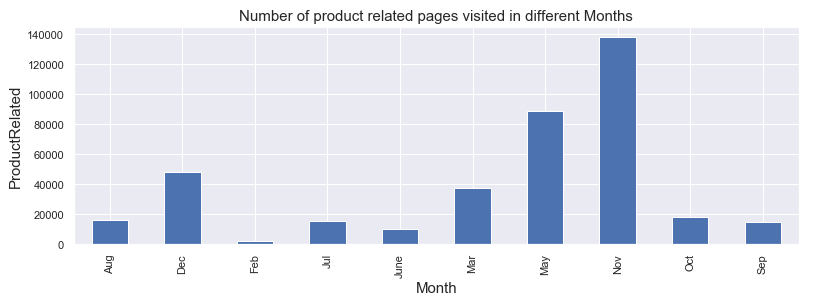




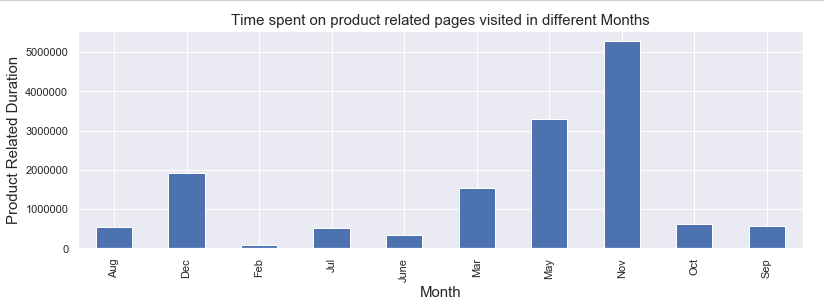


The above plots were drawn to find out the statistical measures like 1st quartile, mean and 3rd quartile. The main motive of these plots is to find out the outliers which might make hurdles in machine learning procedures and in cluster making. According to the above plots, “PageValues”, “specialDay”, “productRelated\_Duration”, “Informational” and “Informational\_Duration” have many outliers which will be handled at the time of preprocessing for applying clustering algorithms by using statistical procedures.

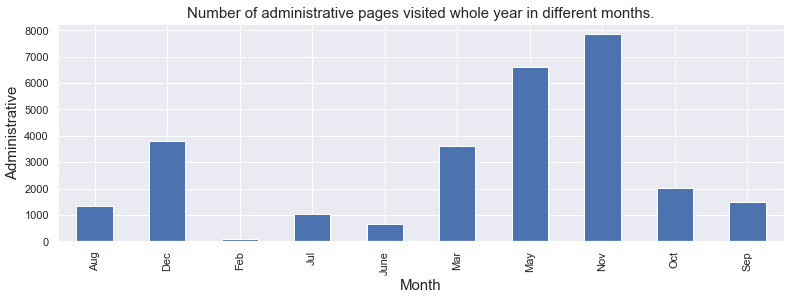
1. **Bivariate Analysis:**
2. **Categorical vs Numerical Plots:**

****

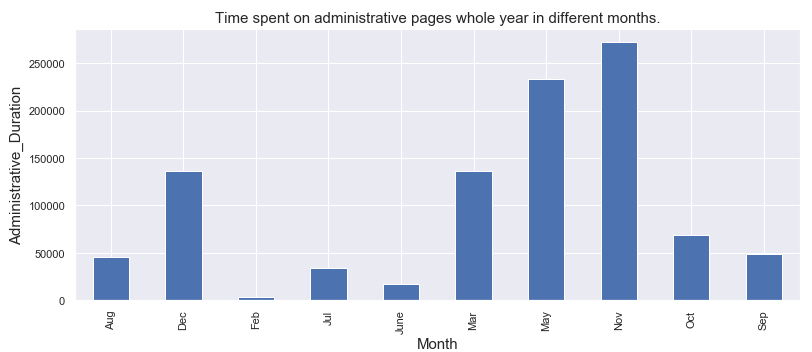
The above plot shows the sum of product related pages every. In November large number of pages related to products are visited which means people might doing preparations for Christmas, New year, weddings etc.

****

The above plot shows the duration spent on product related pages every month.

****

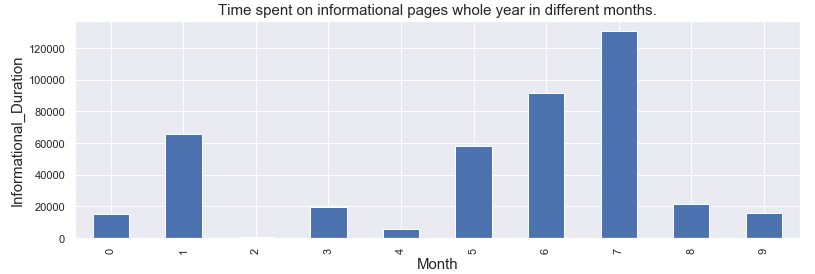
The above plot shows total number of times administrative pages visited whole year in different months.



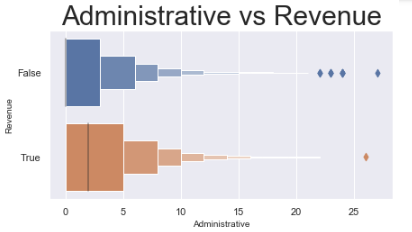
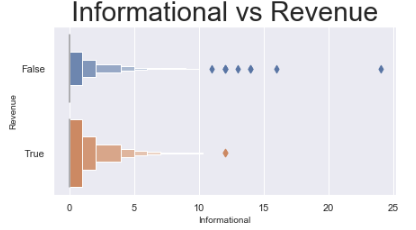
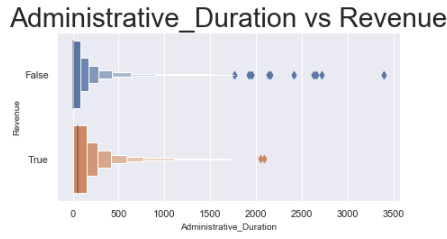
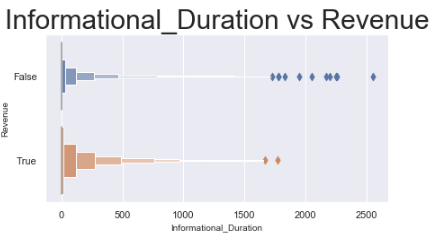
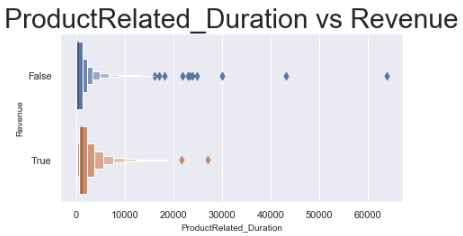
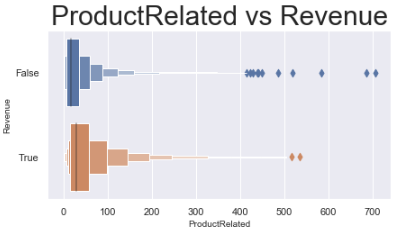
The above plot shows the duration spent on administrative related pages every month. People might have registered complains regarding delivery, product quantity or quality, delivery men's behavior etc.



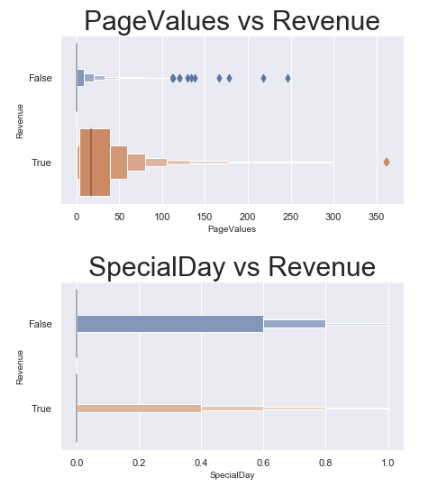
The above plot shows total number of times informational pages visited whole year in different months. Customers might have asked for the information about products quantity and gave their feedbacks on products and services.



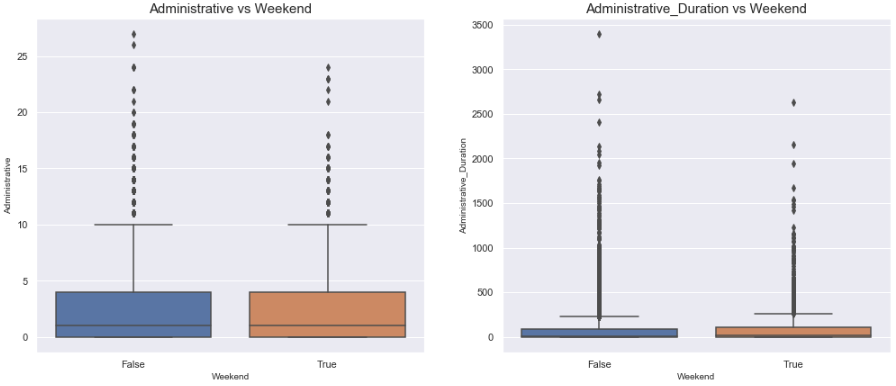
The above plot shows the duration spent on information related pages every month. People might encounter problems in making transactions or buying products, faqs about delivery and new offers.

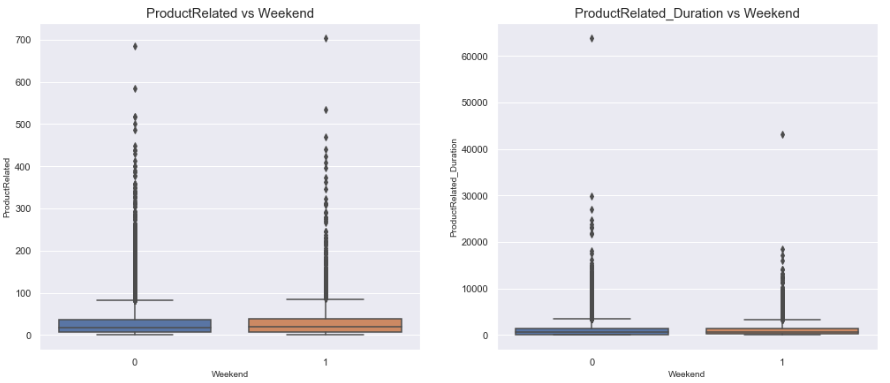


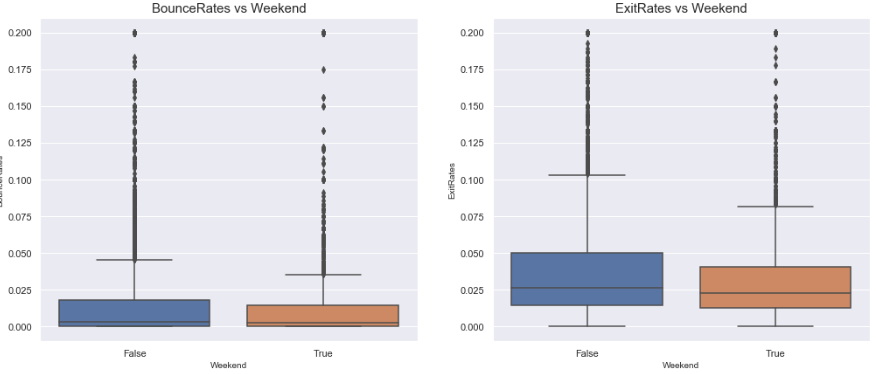
The above plots were drawn to find out the statistical measures like 1st quartile, mean and 3rd quartile according to the categorical feature i.e. “Revenue”. According to above plots, mostly outliers lie where there is no revenue or any purchase or transaction.



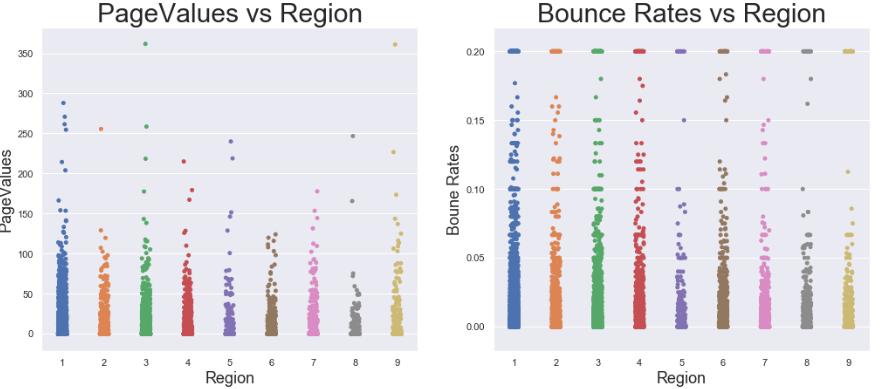
The above 4 plots were drawn to find out the statistical measures like 1st quartile, mean and 3rd quartile according to the categorical feature i.e. “Revenue”. According to the above plots there are no outliers except PageValues vs Revenue.

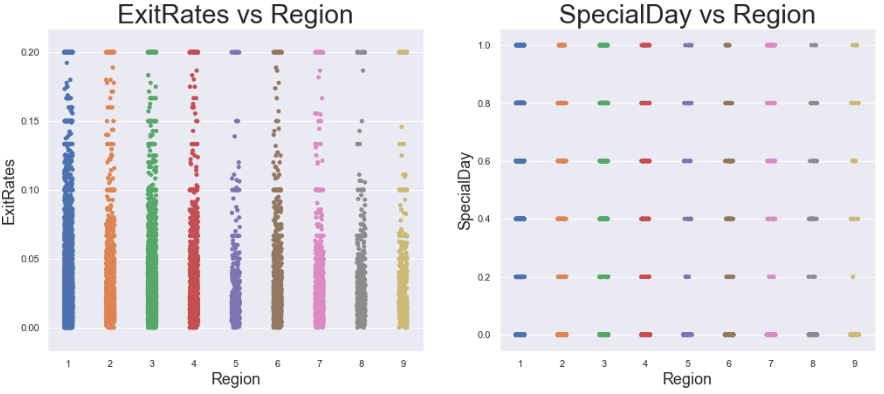


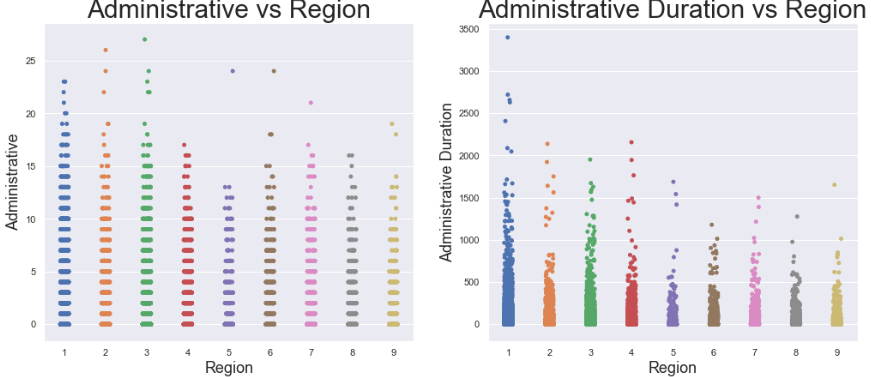


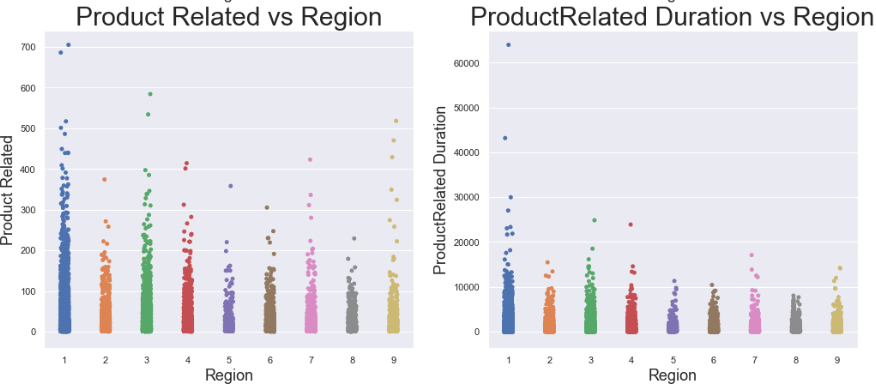


The above plots were drawn to find out the statistical measures like 1st quartile, mean and 3rd quartile according to the categorical feature i.e. “Weekend”. According to above plots, outliers according to weekend are present in every numerical feature but mostly outliers lie where there is not weekend or working day.



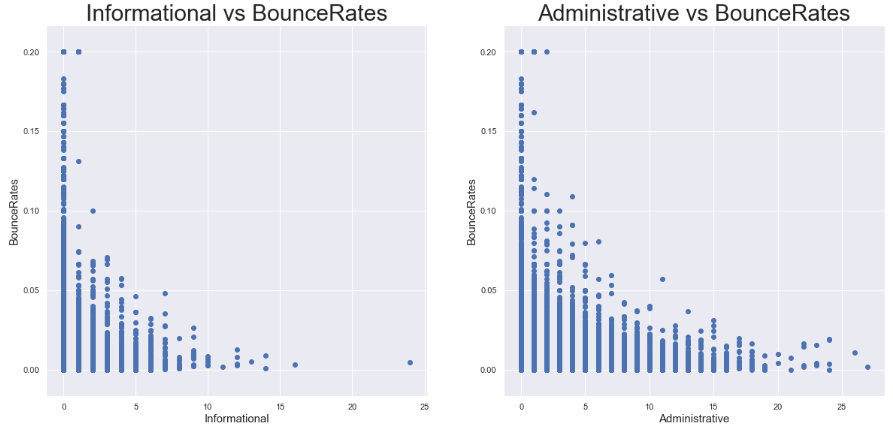


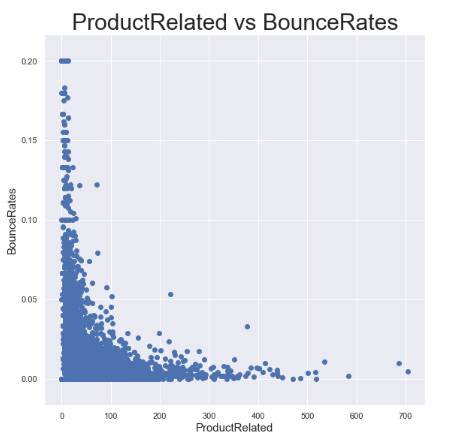




The above strip plots were drawn to identify the data points of numerical attributes distributed according to region. It is basically defining that how values of different numerical data points vary according region.

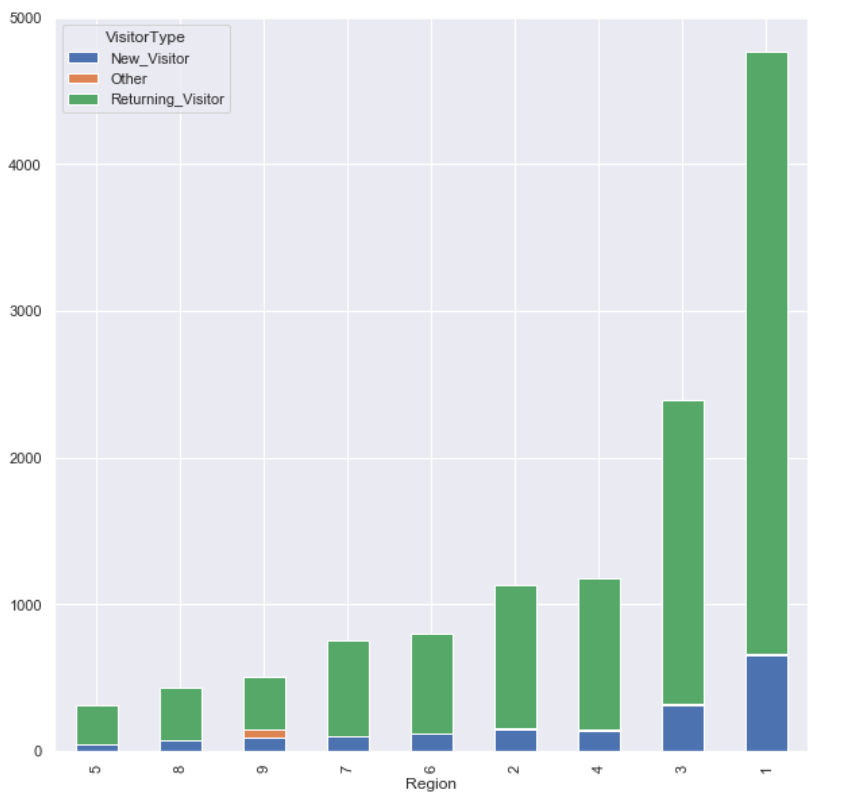
1. **Numerical vs Numerical Plots:**

****

****

The above scatter plot shows that how bounce rates vary according to the product, information and administration related pages visits. The above plots also show the negative correlation between bounce rates and other above-mentioned features.

1. **Categorical vs Categorical Plots:**

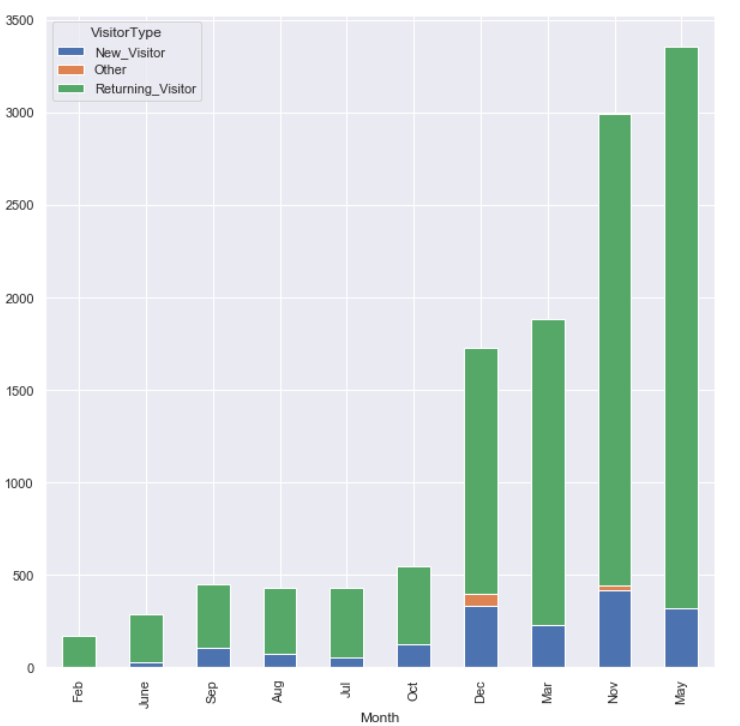
****

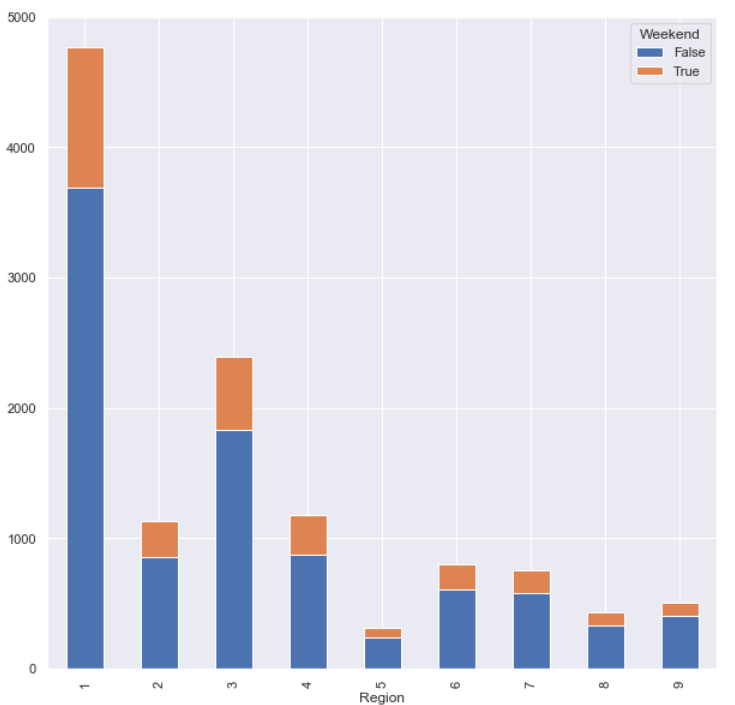
The above result shows how number of different types of visitors vary region wise.

1) Region 5,6 and 8 have lowest number of returning and new visitor which means online shopping is not so common in that region

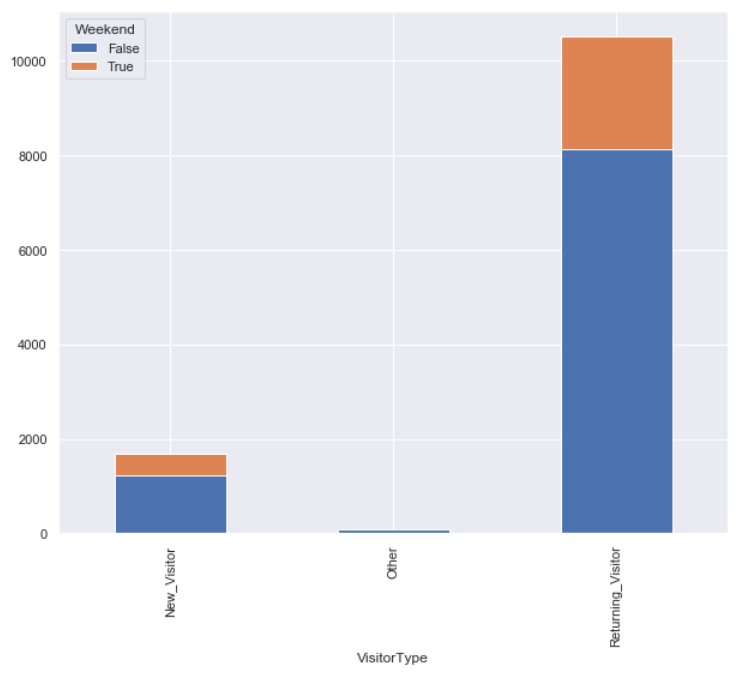
2) Region 7,9,2 and 4 have more returning visitors and less number of new visitors.

3) Region 3 and 1 have the highest numbers of new and returning visitors.

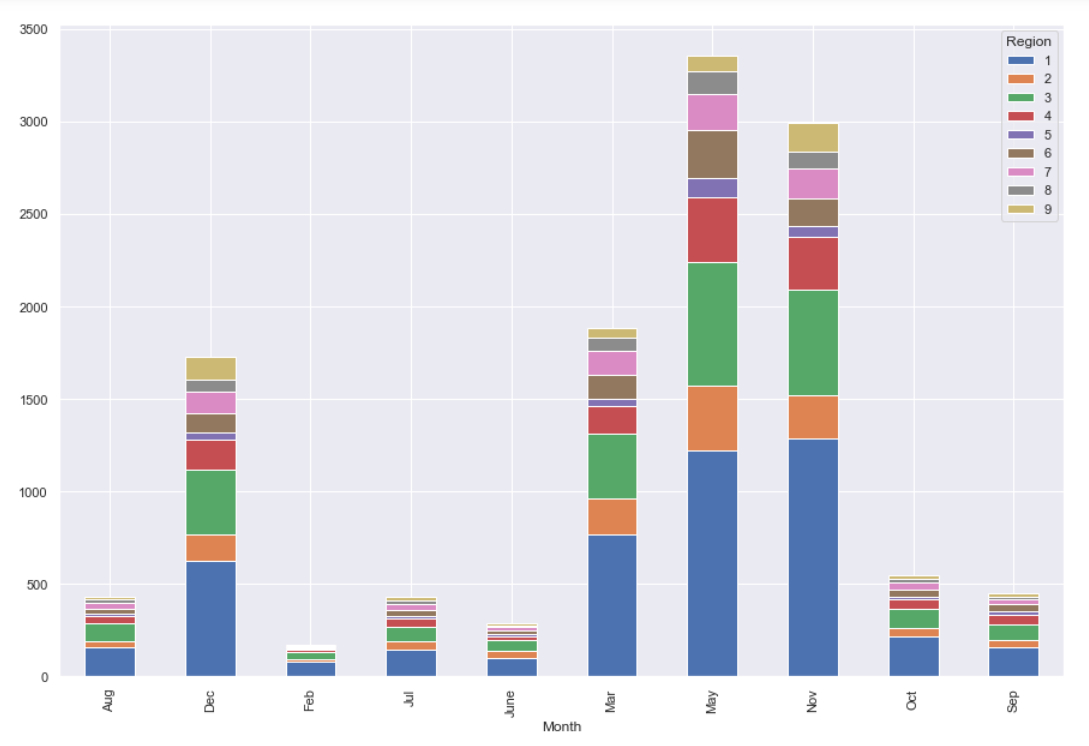
 The above result shows how number of different types of visitors vary month wise.



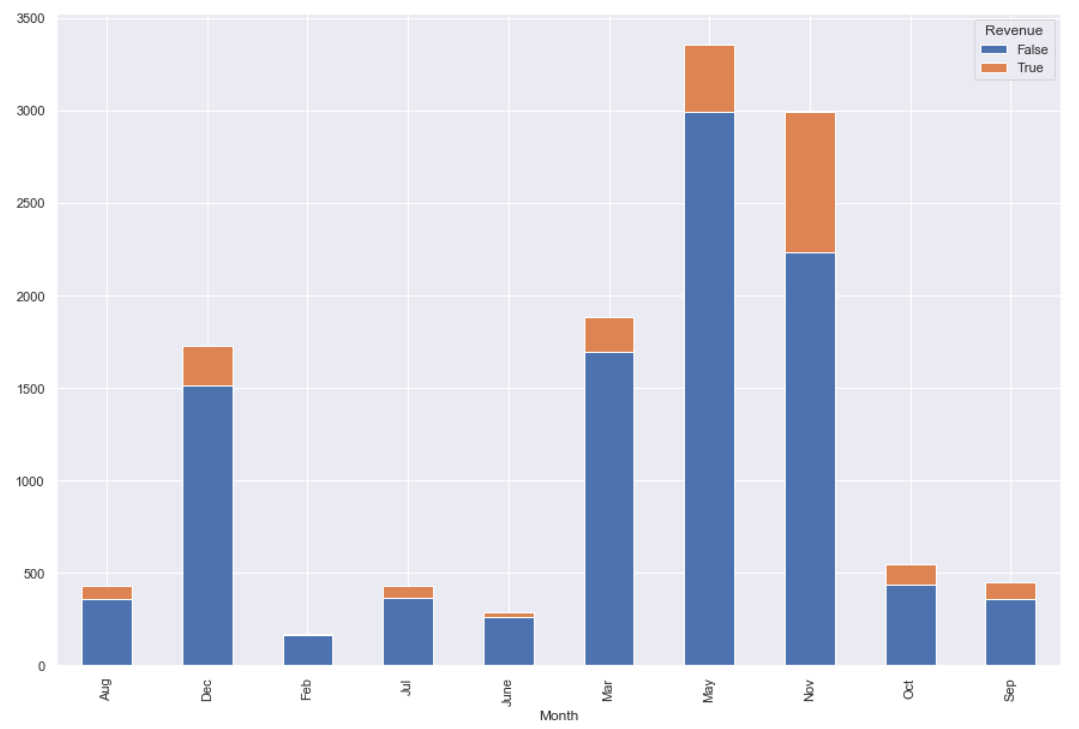
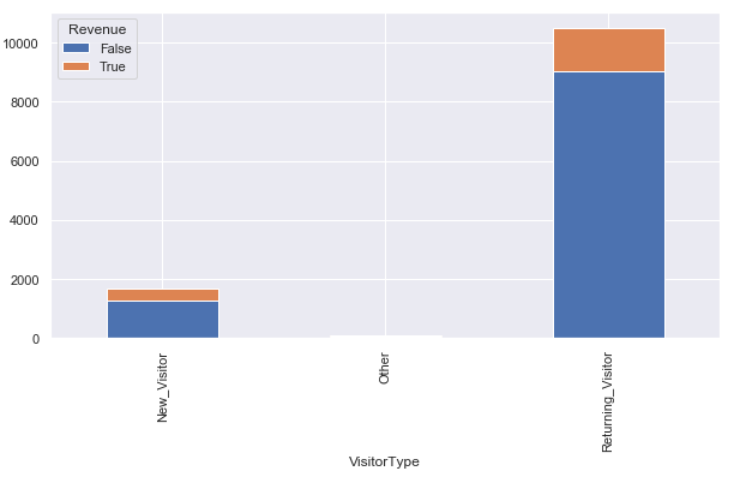
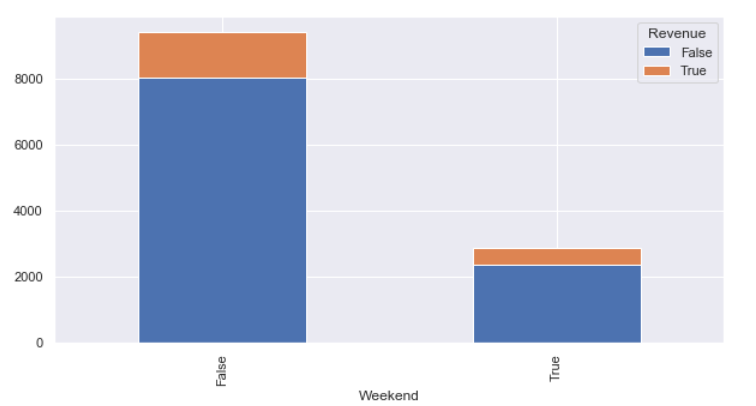
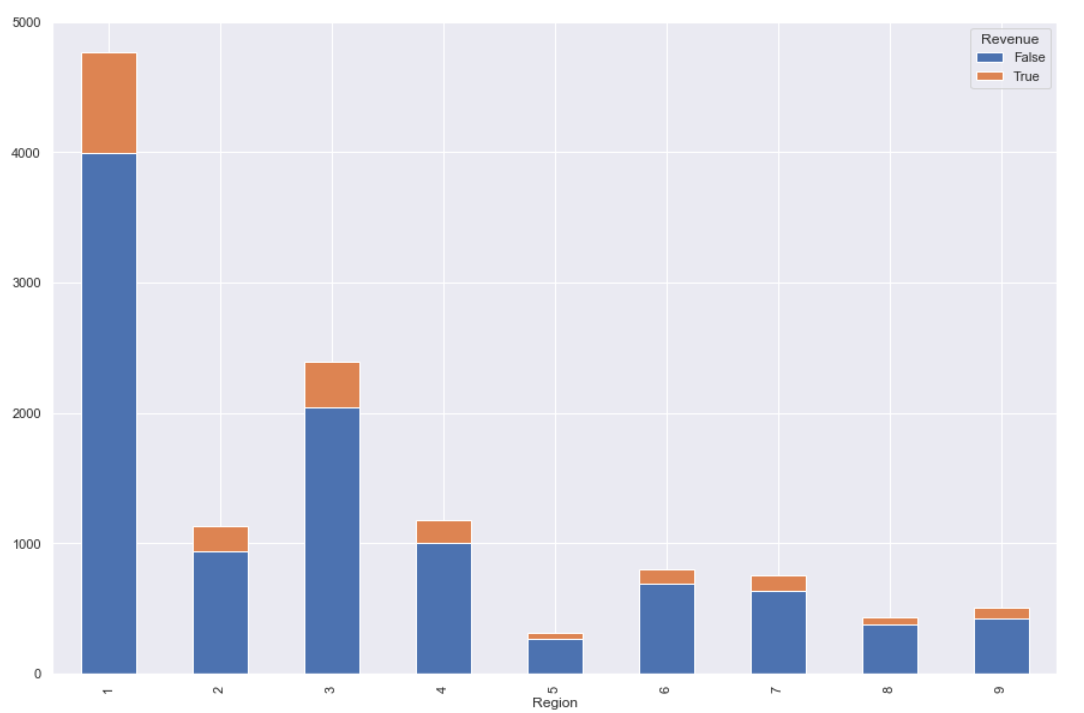
The above result shows how many sessions recorded at weekends or weekdays in different regions.

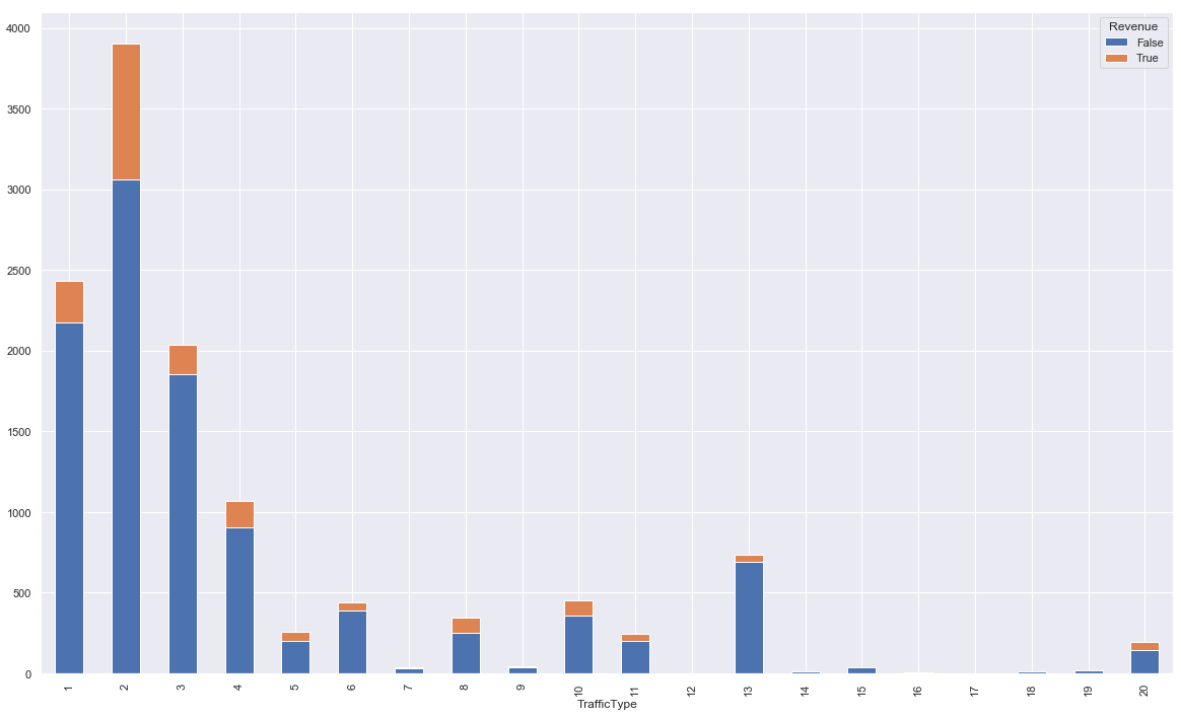


The above result shows how many sessions of different types of visitors recorded at weekends or weekdays.



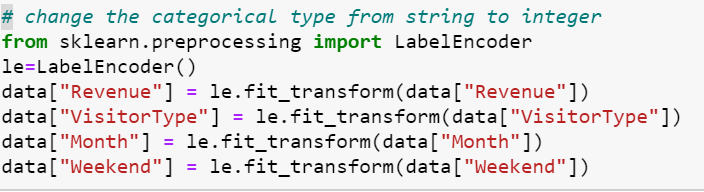
The above result shows how many sessions from different regions recorded in different months.

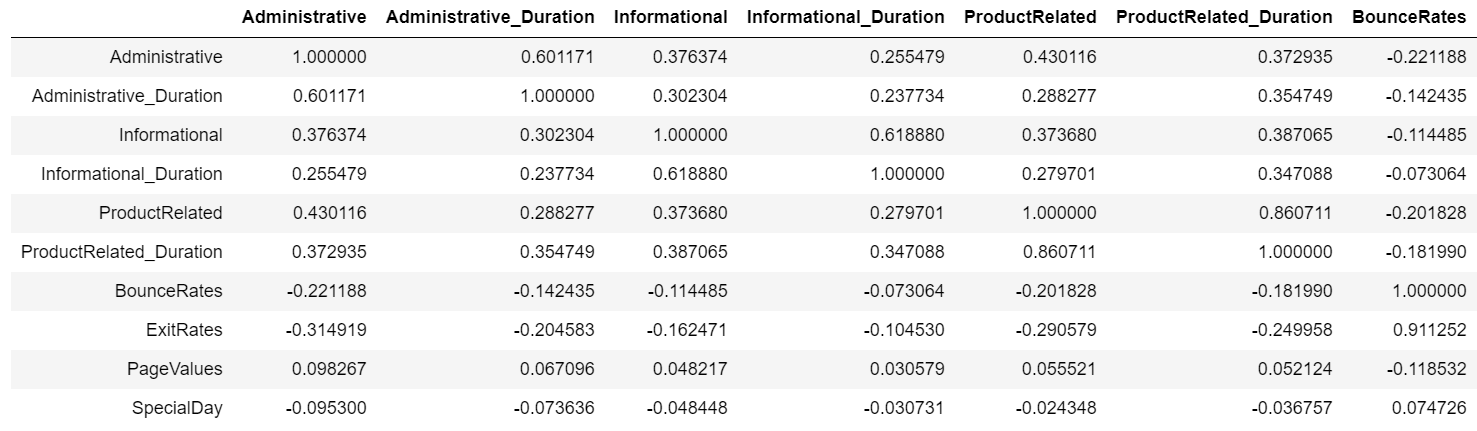
 

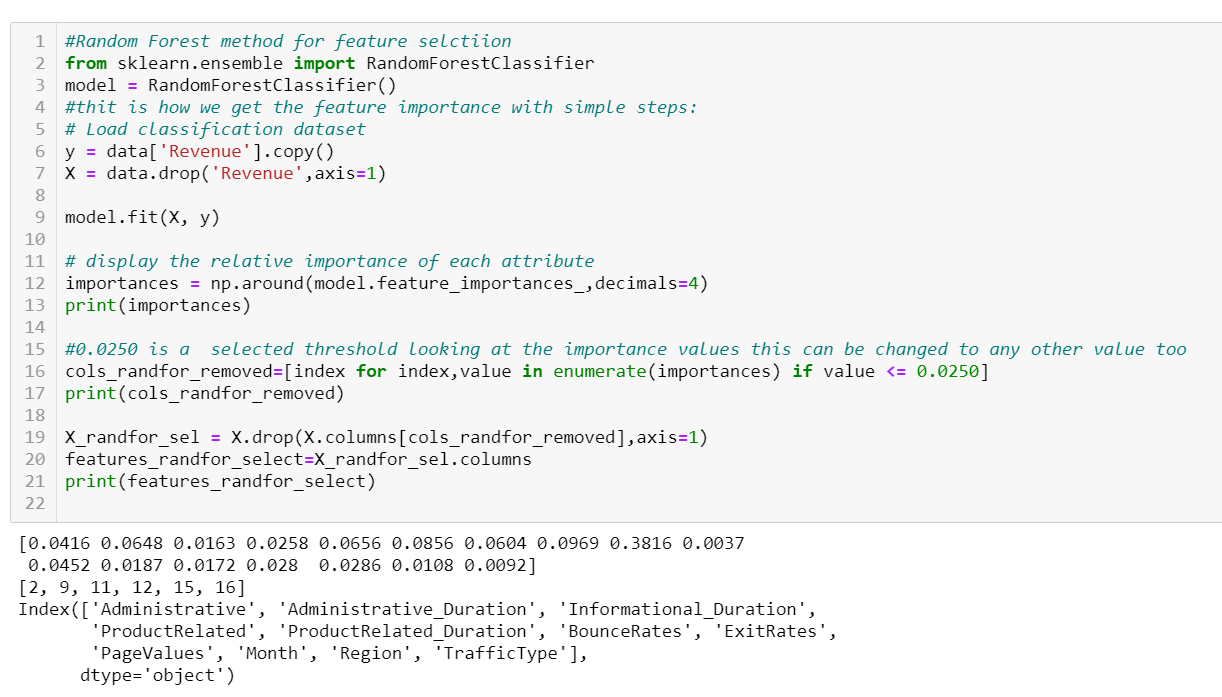


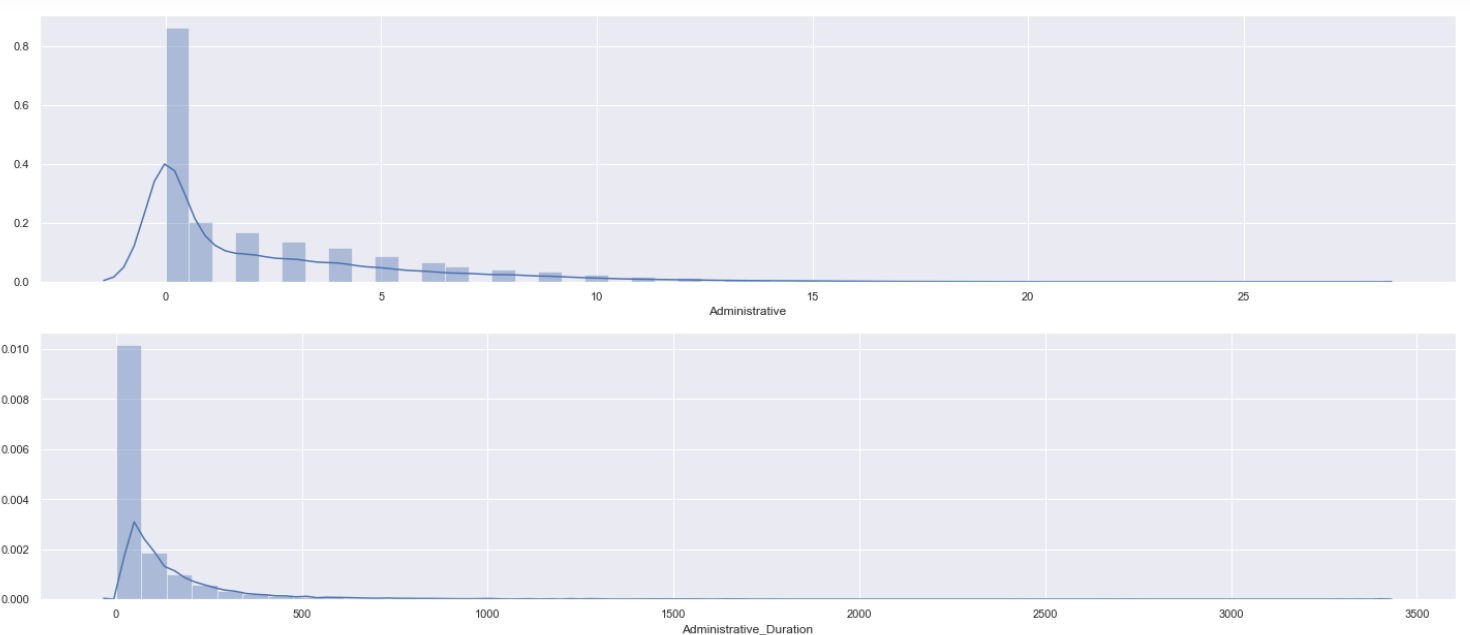
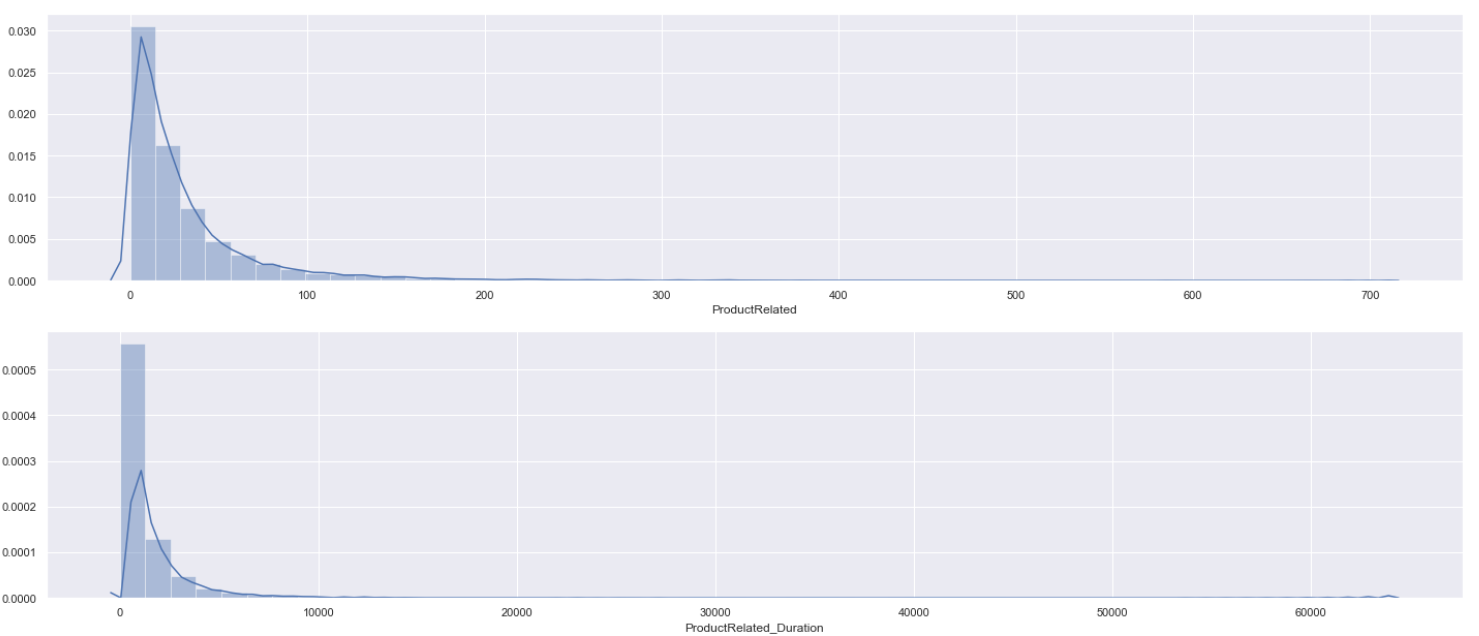
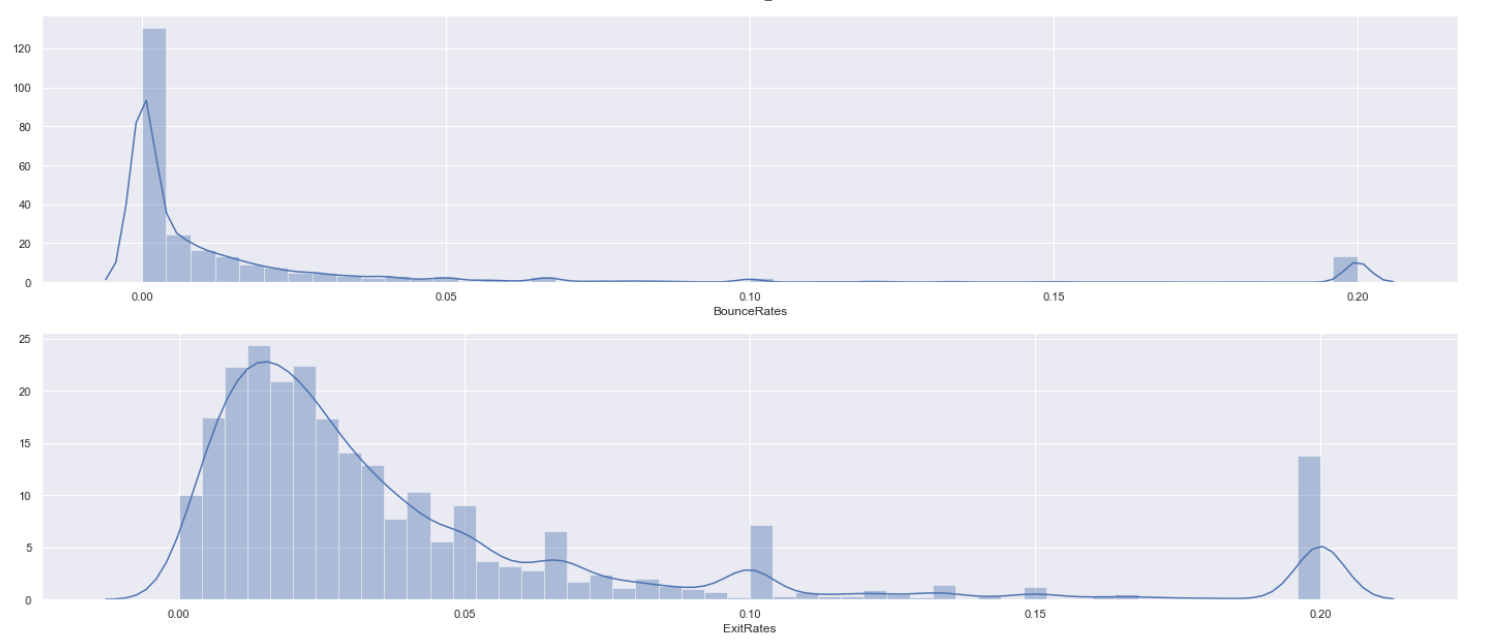
All the above graph shows the number of revenues generated and not according to the different categorical features:

1. In November March, May and December high number of revenues are generated.
2. Mostly revenues are generated from region 1.
3. Mostly revenues are generated from returning visitors.
4. Mostly revenues are generated in working days.
5. Mostly revenues are generated by 1, 2, 3, and 4 traffic type.
6. **Feature Engineering and Preprocessing:**



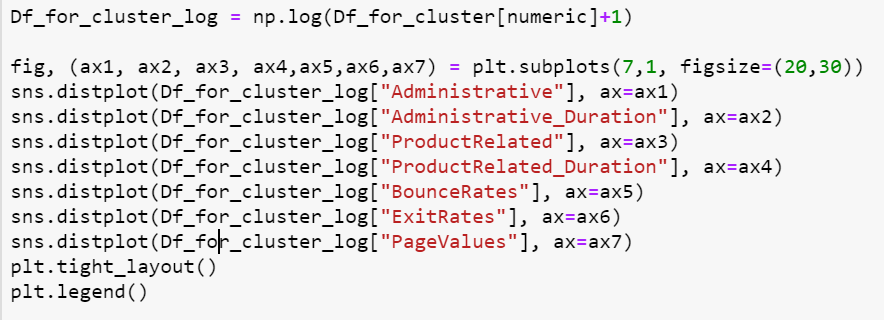
The above code shows the encoding of categorical features.

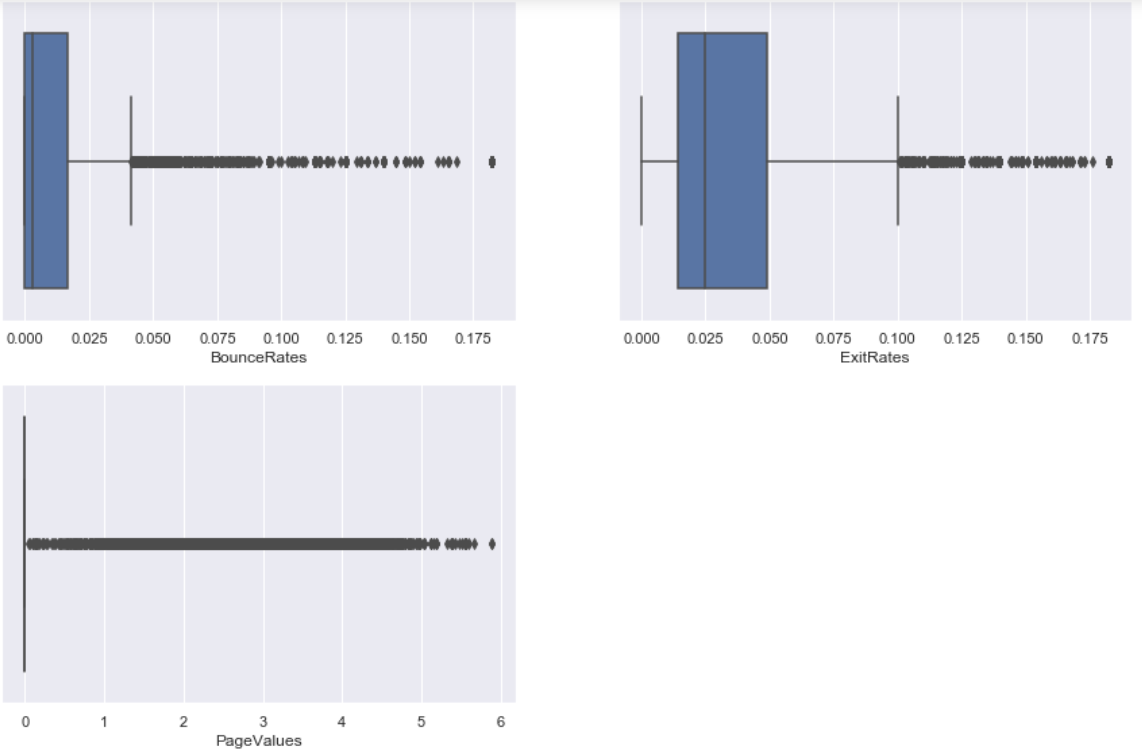
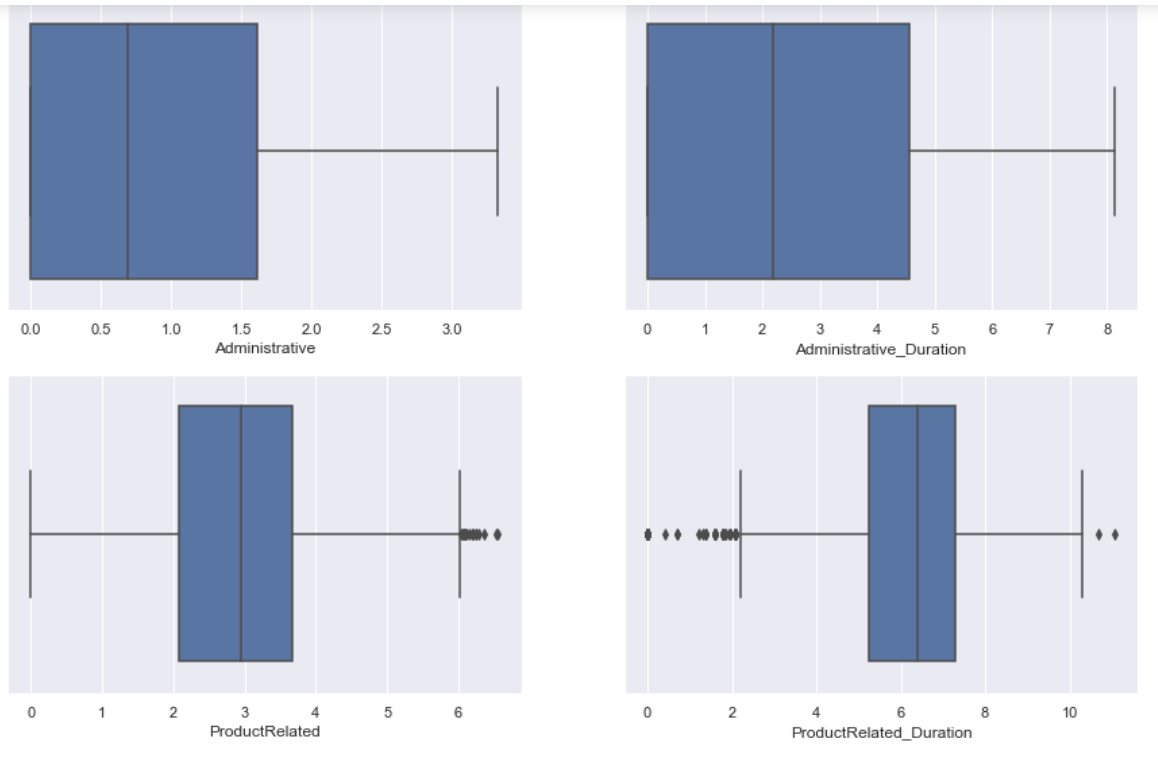


The above procedures are used to select important features. Random forest algorithm gives the weightage of every feature through which we can extract those which are greater than threshold value.  

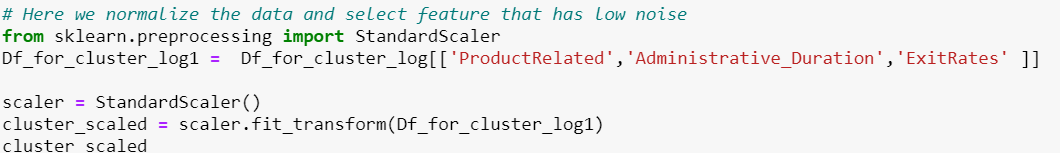
The above graphs show Positive Skewness which means when the tail on the right side of the distribution is longer or fatter. The mean and median will be greater than the mode.

We can use logarithmic transformation to reduce the outliers and distribution skewness. As we have 0 values which will result in inf value on log transform. To handle it we can add 1 in the values.





The above result shows that outliers from features administrative and administrative\_duration is removed where product related and product related duration and exit rates are better and remaining are in worst condition. That is why we can consider those features which have least number of outliers.



The above result shows the selection of important features and conversion of data frame to scaler.

1. **Clustering:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithm** | **Silhouette Index** | **Calinski Index** | **DB Index** |
| **K means** | 0.458 | 12486.85 | 0.803 |
| **MiniBatch Kmeans** | 0.295 | 5790.80 | 1.128 |
| **Mean Shift** | 0.4583 | 12260.136 | 0.791 |
| **Gaussian Mixture Models** | 0.421 | 10548.26 | 0.847 |
| **Affinity Propagation** |  |  |  |
| **DBScan** | 0.155 | 46.928 | 1.739 |
| **Optics** | 0.20 | 32.46 | 1.216 |

1. **Conclusion:**

According to the above clustering and the score generated from the different internal validation methods we can conclude that we should select K-means.

DB For K mean: 0.8039294201665187

Calinski Score For K mean: 12486.85824899695

Silhoutte ScoreK mean: 0.4696572914135361



