Exploratory Data Analysis:

Correlation:

After looking at the data and changing data types to relevant columns, a heat map was coded and can be seen as follows:

Chart

Description automatically generated

From the above heatmap, we observe the following:

In general, there is very little correlation among the different features in our dataset.

The very few cases of high correlation (|corr| >= 0.7) are:

BounceRates & ExitRates =0.9

ProductRelated & ProductRelated\_Duration =0.9

Moderate Correlations (0.3 < |corr| < 0.7):

Among the following features: Administrative, Administrative\_Duration, Informational, Informational\_Duration, ProductRelated, and ProductRelated\_Duration.

Also between PageValues and Revenue.

Web page analysis:

A [violint plot](https://www.data-to-viz.com/graph/violin.html) allow to visualize the distribution of a numeric variable for one or several groups.

Chart

Description automatically generated

Diagram

Description automatically generated with medium confidence

From the above boxplots, we can see that:

* In general, visitors tend to visit less pages, and spend less time, if they are not going to make a purchase.
* The number of product related pages, and the time spent on them, is way higher than that for account related or informational pages.
* The first 3 feature look like they follow a skewed normal distribution.

Page Metric Analysis: **-Google analytics values**

Chart, histogram

Description automatically generated

From the above visualizations of 3 google analytics metrics, we can conclude:

* *BounceRates* & *PageValues* do not follow a normal distribution.
* All 3 features have distributions that are skewed right.
* All 3 distributions have a lot of outliers.
* The average bounce and exit rates of most of the data points is low, which is good, since high rates identicate that visitors are not engaging with the website.
* Exit rate has more high values than bounce rate, which makes sense, where transaction confirmation pages for example will cause the average exit rate to increase.
* Bounce rate ==> the percentage where the first page visited was the only page visited in that session.
* Exit rate of a page ==> The percentage where that page was the last page visited in the session, out of all visits to that page.

Visitor Analysis:

Graphical user interface, chart, application, bar chart

Description automatically generated

* 1 Operating system is responsible for ~7000 of the examples in our dataset.
* 4 of the 8 operating systems used, are responsible of a very small number (<200) of the examples in our dataset.
* A similar story repeated with the browsers used by visitors, where there is 1 dominant browser, 3 with decent representation in the dataset, and the rest are rarey used.
* It looks like we have a very regionally diverse traffic in our dataset.
* Also Traffic sources are very diverse, with a few that did not contribute much to the dataset.

Visit Data Analysis:

Chart

Description automatically generated with medium confidence

Chart, pie chart

Description automatically generated

* On March and May, we have a lot of visits (May is the month with the highest number of visits), yet transactions made during those 2 months are not on the same level.
* We have no visits at all during Jan nor Apr.
* Most transactions happen during the end of the year, with Nov as the month with the highest number of confirmed transactions.
* The closer the visit date to a special day (like black Friday, new year's, ... etc) the more likely it will end up in a transaction.
* Most of transactions happen on special days (SpecialDay =0).
* It does not look like weekends affect the number of visits or transactions much, we can see only a slight increase in the number of transactions happening on weekends compared to those on weekdays.

A picture containing chart

Description automatically generatedChart, pie chart

Description automatically generated

* It's a good sign that 85% of visitors are returning to site. 15% of visitors are new. We can give offers to lure more new visitors.

data.groupby('Month')['Revenue'].value\_counts().unstack('Revenue').plot(kind='bar', stacked=True, figsize=(10, 5))

Chart, bar chart

Description automatically generated

data.groupby('Weekend')['Revenue'].value\_counts().unstack('Revenue').plot(kind='bar', stacked=True, figsize=(7, 7))

Chart, bar chart

Description automatically generated

Chart

Description automatically generated

# Bounce Rate vs Exit Rate

sns.lmplot(x="BounceRates", y="ExitRates", data=df,

           scatter\_kws={'alpha':0.3},

           line\_kws={'color':'black', 'ls':'--'})

plt.show()

Chart, scatter chart

Description automatically generated

google\_analytics\_features = ['BounceRates', 'ExitRates', 'PageValues']

corr\_mat = df[google\_analytics\_features].corr()

sns.heatmap(corr\_mat, annot=True,annot\_kws={'fontsize':16}, fmt='.3g', linewidths=2, cmap='Pastel1')

plt.show()

Table

Description automatically generated

# Impact of ProductRelated Page Views and ProductRelated Pageview Duration on Revenue

sns.lmplot(x="ProductRelated", y="ProductRelated\_Duration", hue='Revenue', data=df, scatter\_kws={'alpha':0.5})

plt.show()

Chart, scatter chart

Description automatically generated

Machine Learning:

Data Source:

Models:

KNN

Logistic Regression

Random Forest

Decision Tree Classifier

XGBoost

Adaboost

Naïve Bayes

Support Vector Machine