Capstone 1 Data Wrangling

Notebook Link:https://www.kaggle.com/henrysue/classifying-online-shopper-intention

The Data:

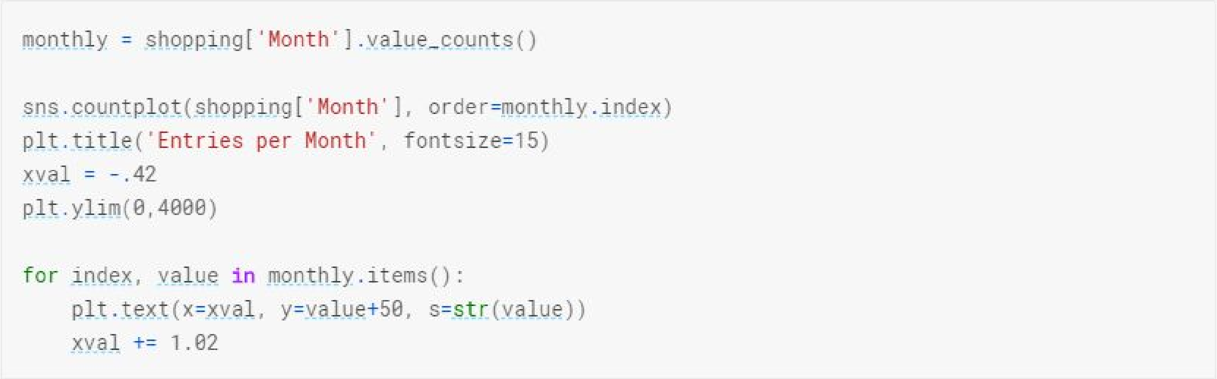
The data set is a set of 18 features: 10 numerical and 8 categorical. This dataset has 12330 entries, split into 10,422 entries where the shoppers did not purchase and 1908 entries where the shoppers did purchase. Each entry is based on unique users in a 1-year period to avoid any trends specific to a specific campaign.

Wrangling:

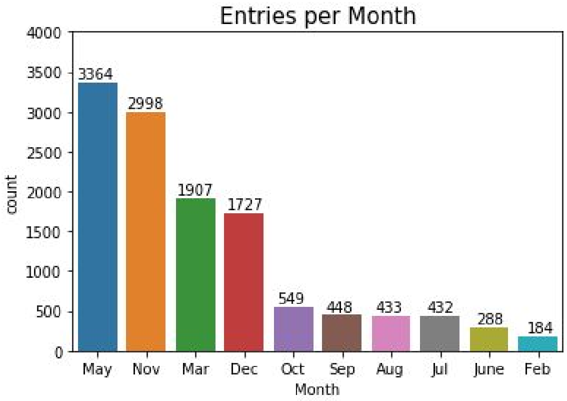
The dataset has no missing values, and therefore we do not need to replace any values.

The dataset has a few features that we do not want to observe.

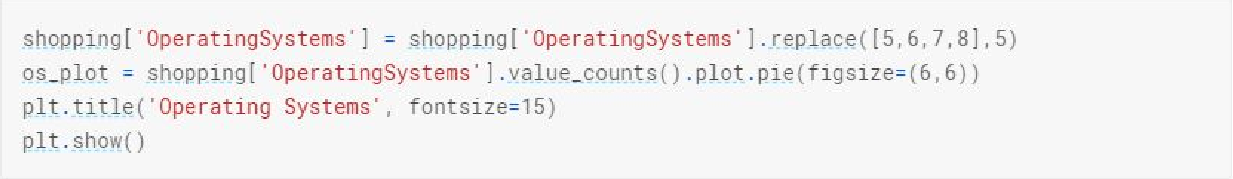
First, we drop the month column. The 'Month' column only has 10 unique types, which indicates that it is missing two months of data. Each month has varying numbers of entries, which could unfairly bias our data to prefer classification by month. We can see below the distribution of each month in the 'Month' column. Additionally, time-sensitivity is already contained in the 'SpecialDay' column, which influences buying decision, so the month column is slightly redundant.



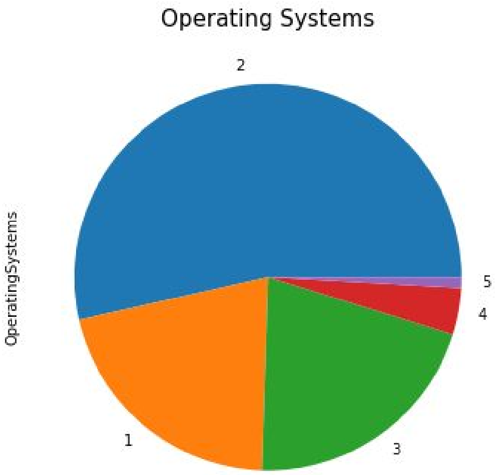
```
monthly = shopping['Month'].value_counts()

sns.countplot(shopping['Month'], order=monthly.index)
plt.title('Entries per Month', fontsize=15)
xval = -.42
plt.ylim(0,4000)

for index, value in monthly.items():
    plt.text(x=xval, y=value+50, s=str(value))
    xval += 1.02
```

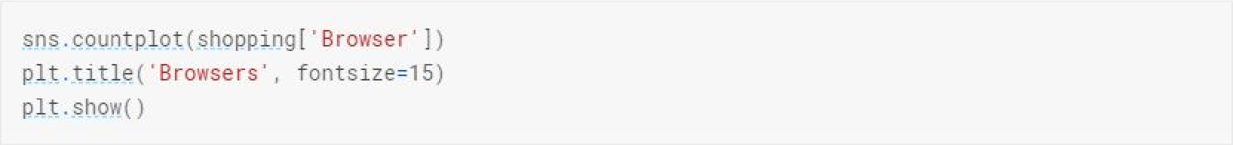


Here we have the Operating Systems, labeled by number. Low-usage browsers have been consolidated into label '5'. We can see that a majority of users use operating system #2. Operating systems can indicate users of a sepcifc type of computer and may portray certain user archetypes (Windows users, Mac users, Linux users). We do not use this for our analysis as they are an extra feature and we want to keep our model as simple as possible.

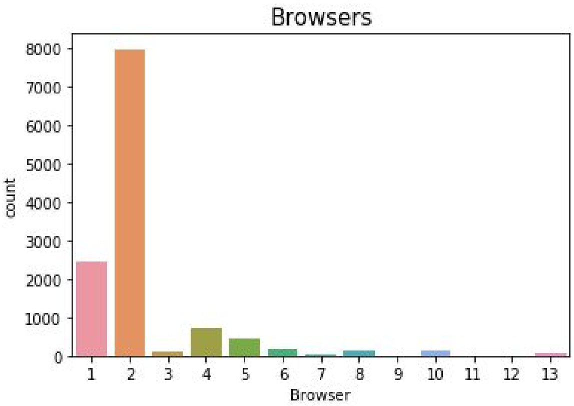


```
shopping['OperatingSystems'] = shopping['OperatingSystems'].replace([5,6,7,8],5)
os_plot = shopping['OperatingSystems'].value_counts().plot.pie(figsize=(6,6))
plt.title('Operating Systems', fontsize=15)
plt.show()
```





```
sns.countplot(shopping['Browser'])
plt.title('Browsers', fontsize=15)
plt.show()
```



Browser choice is even more polarizing than Operating System. Here we see that a large majority of users use browser 2, with a smaller number of users using browser 1. All other browsers represent a small subsection of online users. We will not use this as it does not contribute much to our model.

There are several other columns that we leave out:

'Region': We leave regionality out because the regionality may be slightly tied to purchase likelihood, but we want to train our model on a smaller set of features if possible.

'TrafficType': We leave this column out because Traffic sources are not quite useful for classifying if a user will make a purchase. It usually aids website owners in gauging traffic sources and can assist with determining where they should invest in advertisement.

'Weekend':There is a weak correlation between days of the week and online shopping. https://blog.workarea.com/trends-when-do-people-shop-online asserts that Sundays and Mondays have the highest traffic for eCommerce, but only by 16% of weekly revenue, and mostly on Monday, which is not classified as a weekend.