Classification on Flipkart-products

Prerequisites :

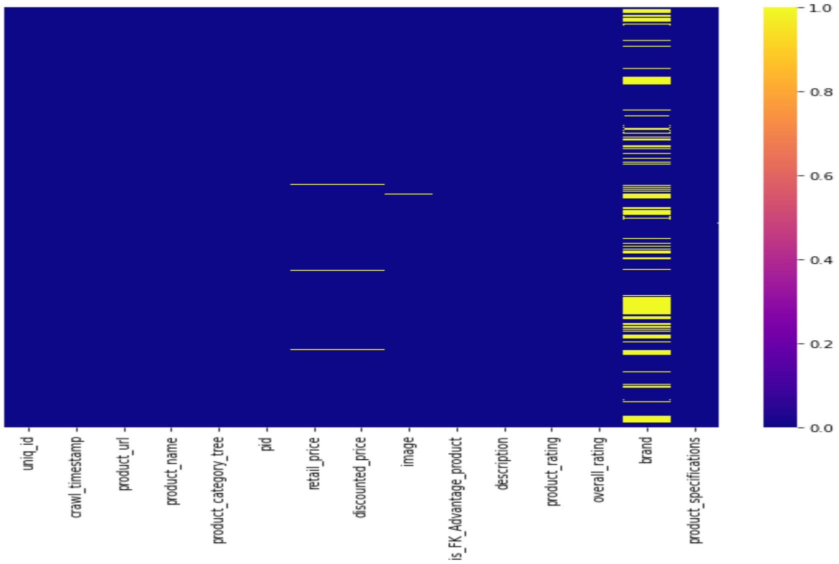
- To run the notebook file in colab, there is no need to install any libraries.
- Softwares and libraries in the local machine before running this project.
  - Anaconda : It will install an ipython notebook and most of the libraries which are needed like sklearn,pandas,seaborn,matplotlib,numpy, scipy.
  - Python 3
  - plotly
  - spacy
  - nltk
  - wordcloud

Dataset :

- The dataset has 20000 rows and 15 columns.
- All the columns are of type object except retail_price, discounted_price, is_FK_Advantage_product.
- retail_price, discounted_price are of data type float64.
- is_FK_Advantage_product is of type bool.
- Dataset link - [flipkart-products]()

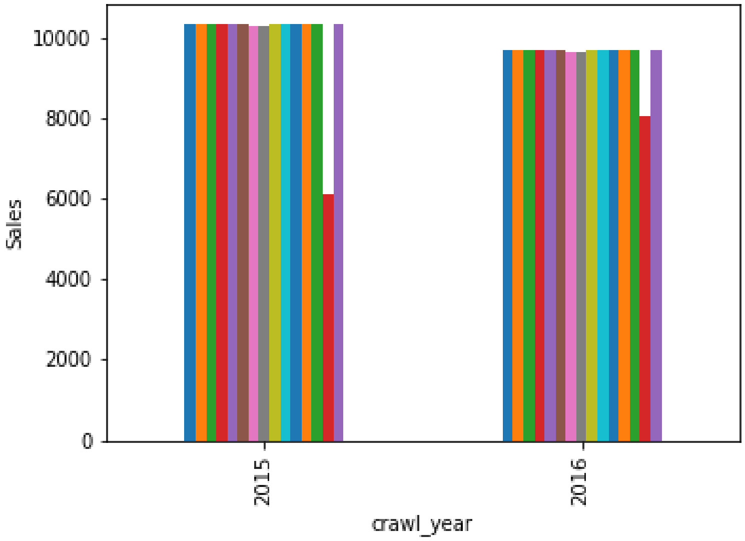
Data visualization :

- Column 'brand' has 5864 null values



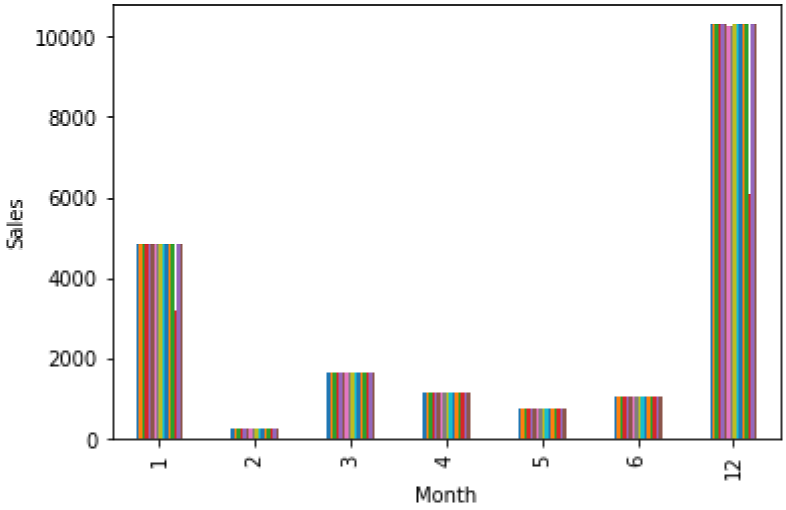
In the above heatmap, the yellow colour at brand shows the null values.

- Entry in ‘product_category_tree’ was a tree.
- To visualize the sales in a year we need to get the year for entries. To do this we use the lambda function on crawl_timestamp and get a crawl_year column.



2015 have more no. of sales

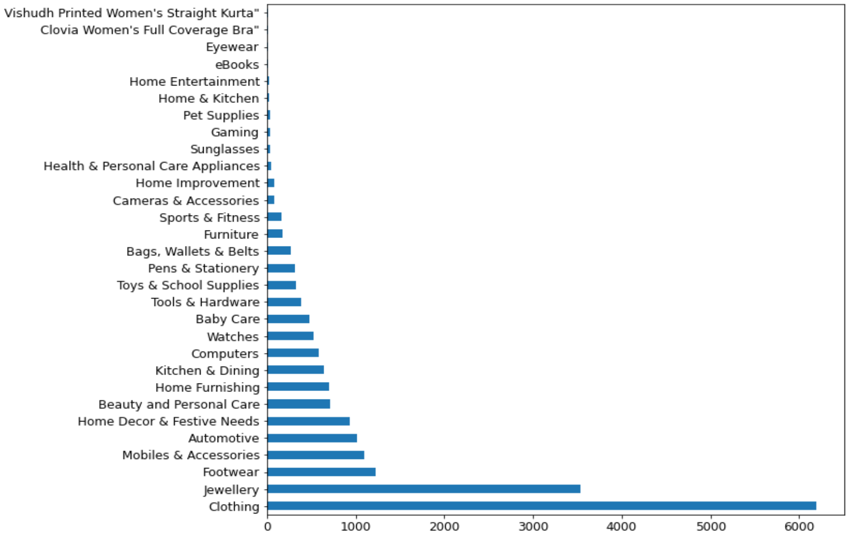
- To visualize the sales in a month we need to get the year for entries. To do this we use the lambda function on crawl_timestamp and get a Month column.



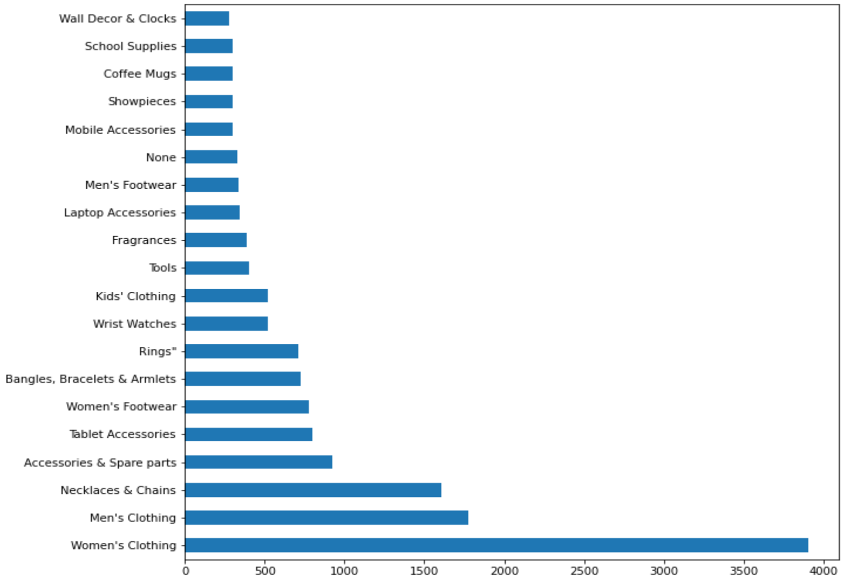
Month 12 have more no. of sales

- For instance - ["Clothing >> Women's Clothing >> Lingerie, Sleep & Swimwear >> Shorts >> Alisha Shorts >> Alisha Solid Women's Cycling Shorts"] says that user is more interested in clothing than women’s clothing etc.

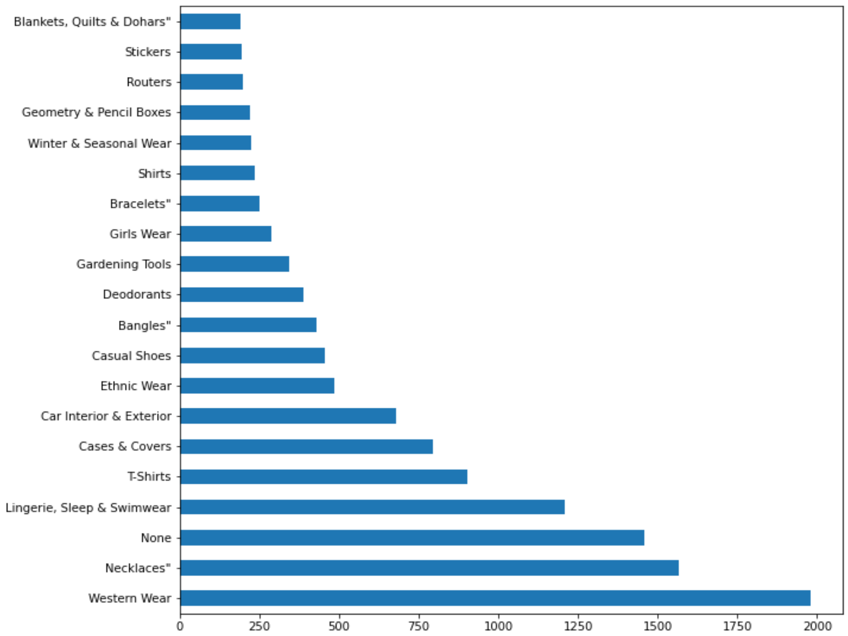
- Splitting this tree using lambda and split functions results in 6 new columns namely ‘primary_category’, ‘secondary’, ‘tertiary’, ‘quaternary’, ‘fifth’, ‘sixth’.
- Now the data frame has 20000 rows and 23 columns.
- If we visualize the ‘primary category’ using plt we say that most customers use flipkart for clothing.



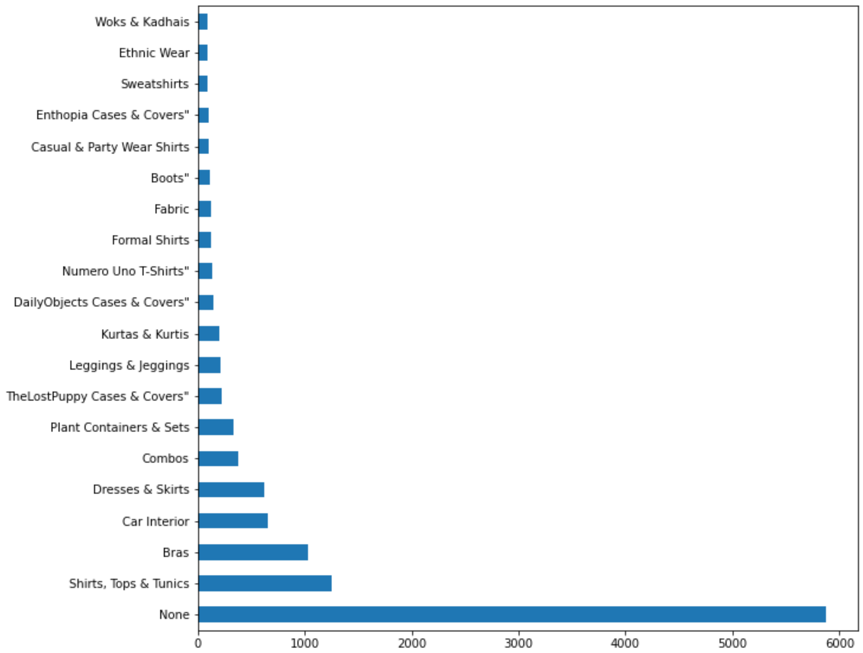
- If we visualize the column ‘secondary’ using plt we can say that womens use flipkart more for shopping than men.



- Visualizing the column ‘tertiary’ using plt we can say that there are customers who prefer western wear.



- If we visualize the column 'quaternary' using plt we can say that there are many null values.



Cleaning :

- For the prediction of category, description and main-category is sufficient. If the accuracy was not good, later other categories can be added.
- To find the main-category lets visualize all categories.
- Column 'secondary' has 328 null values which causes data loss.

- Column ‘tertiary’ has 1457 null values and ‘quaternary’ has 5876 null values which causes huge data loss.
- Column ‘primary_category’ has no null values.
- So ‘primary_category’ can be considered as main-category to predict the category using description.
- All the columns other than ‘primary_category’, ‘product_category_tree’ and ‘description’ can be omitted and these 3 columns are copied into a new data frame and saved as flipkart_com-ecommerce_cleaned_sample.csv

All the above visualizations and cleaning was done in notebook preprocessing.ipynb

Classifier :

- There are 266 unique values in ‘primary_category’ which means the problem is a multinomial classification problem with 266 classes.
- Considering description as X(feature vector) and primary_category as y(label), transform description using CountVectorizer.
- Split the dataset into test and train subsets using train_test_split by importing model_selection from sklearn.
- Here I considered three Machine learning methods
  - RandomForestClassifier
  - Naive Bayes classifier for multinomial models
  - DecisionTreeClassifier

Results :

RandomForestClassifier gave an accuracy of around 88%

- average precision = 0.87
- average recall = 0.87
- average F1 score = 0.86

Naive Bayes classifier for multinomial models gave an accuracy around 90%

- average precision = 0.89
- average recall = 0.90
- average F1 score = 0.89

DecisionTreeClassifier gave an accuracy of 93%

- average precision = 0.93
- average recall = 0.93
- average F1 score = 0.93

Improvements :

To improve the model, deep learning methods like CNN,RNN can be used for better accuracy.

References :

For cleaning and visualization : https://www.kaggle.com/learn/pandas

For models : https://scikit-learn.org/stable/modules/multiclass.html

Github repository link: Classification-on-Flipkart-products

Done by :

V . Sowmya

IIIT Sri City.