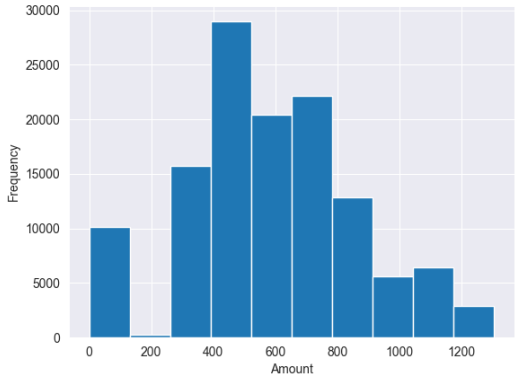
Step 1: Exploratory Data Analysis (EDA)

- After we loaded the dataset into data frame we inspect the key features ,we find that some columns not useful in analysis and modeling like (SKU,ASIN,Style,Order id,unnamed22,promotions ids) .
- Then we check the non-null records and types of columns using .info()
- We find multiple columns have null records.
- Some nulls are trivial like currency we can transform all records of currency to INR and fulfilled-by we can fill the nulls with another word opposing 'Easy Ship' like 'Non Easy Ship' to encode it later in modeling
- And the values that are null in Amount we find the value of the Qty in the same row are 0 then we assign value 0 to null values in amount as no items sold
- And the null values in 'Courier status' we find 0 in its rows in Qty so we can suppose it is cancelled
- And there are 33 records of null state and city and country ,we remove them

Then

- We generate statistics for numerical columns like Qty and Amount we find that the mean is 650 but the maximum is 5584 then there is outlier and for Qty we find the mean is 0.90 but the maximum is 15
- And when performing statistics on categorical data we find that
- In 'ship-service-level' : 'Expedited' is twice as 'Standard'
- And in 'Category': "Set" is the highest followed by 'Kurta'
- And in 'Size': 'M','L' sizes are the most common



2.Data Conversion

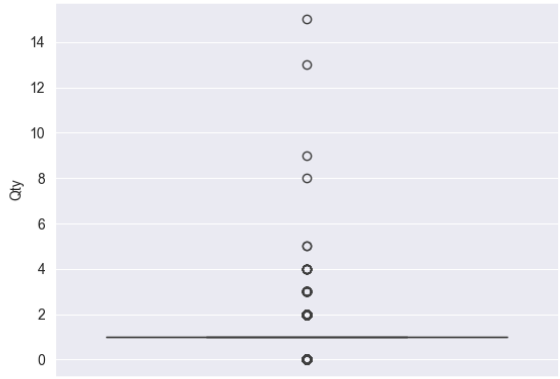
We find in the data frame that the ‘Date’ column has type object,so we transform its type to datetime to use it in analysis and visualization using

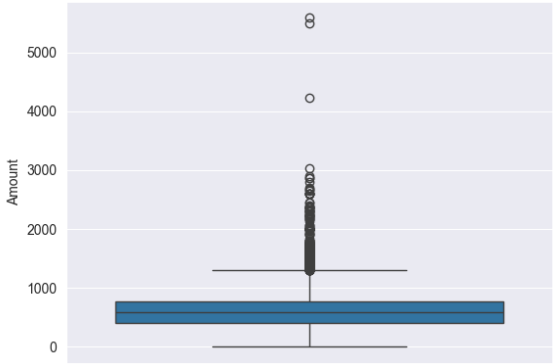
```
pd.to_datetime(data['Date'], format='%m-%d-%y', errors='coerce')
```

and transform the types from object to Category

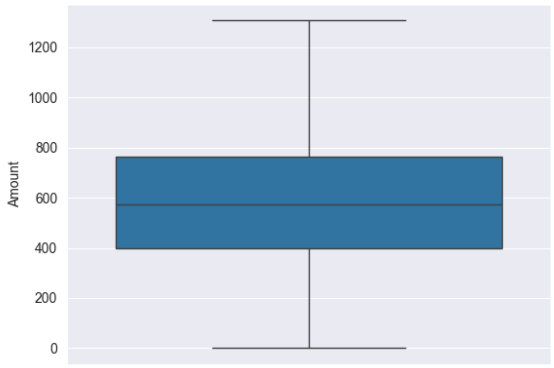
3.outliers

After we find the outliers in Qty and amount using describe() method ,we visualize it using box plot



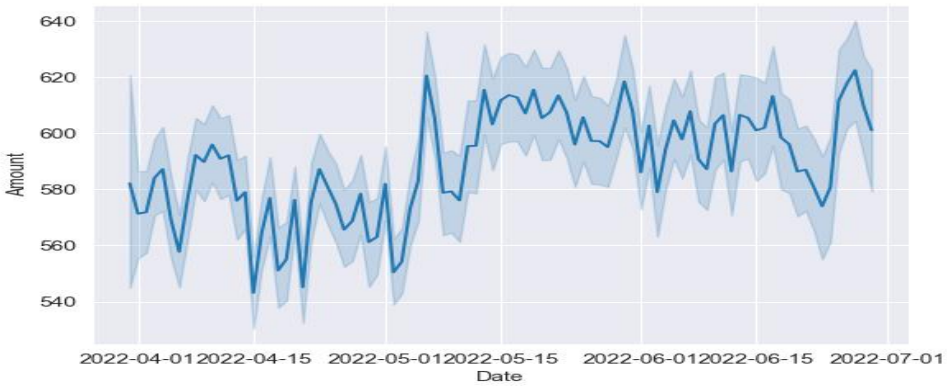


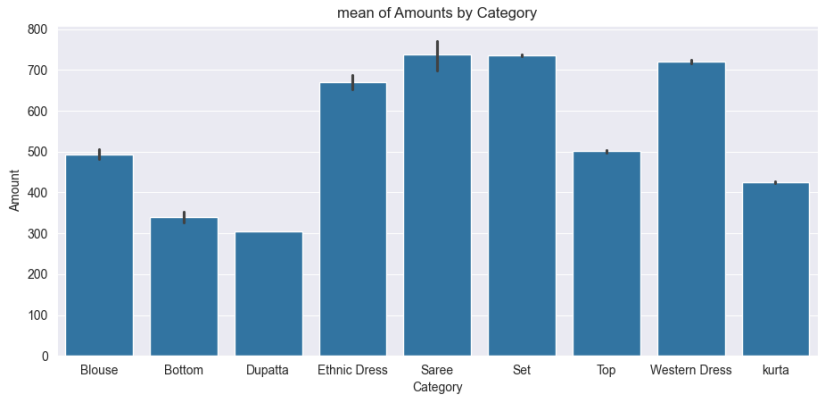
So we remove all values above 1 in Qty and above Q3+1.5*IQR in Amount

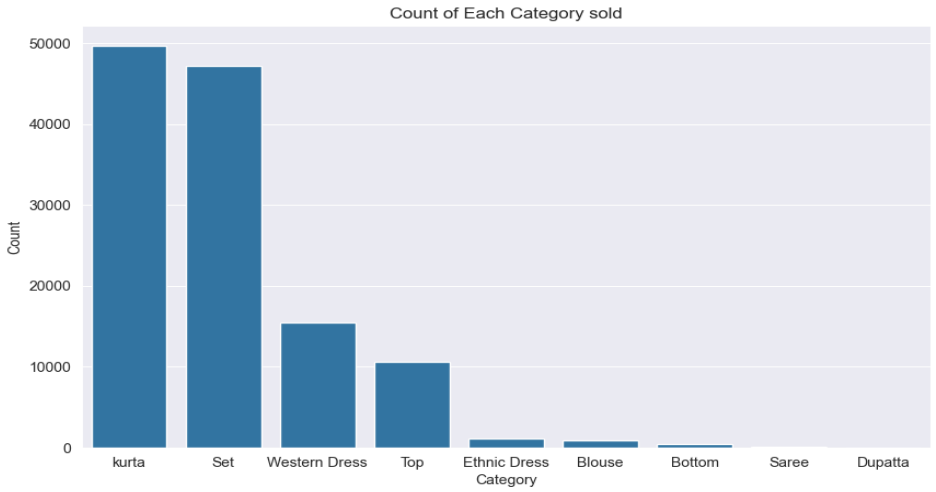


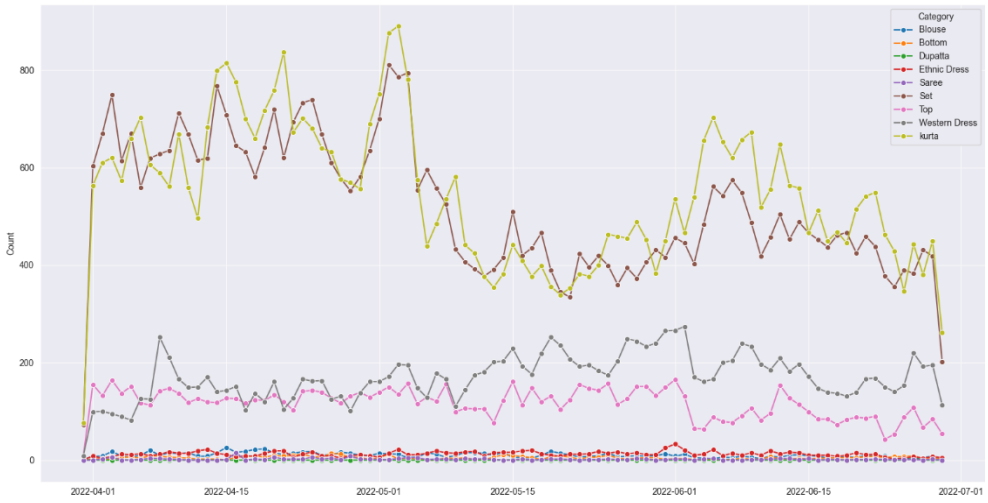
4.data visualization

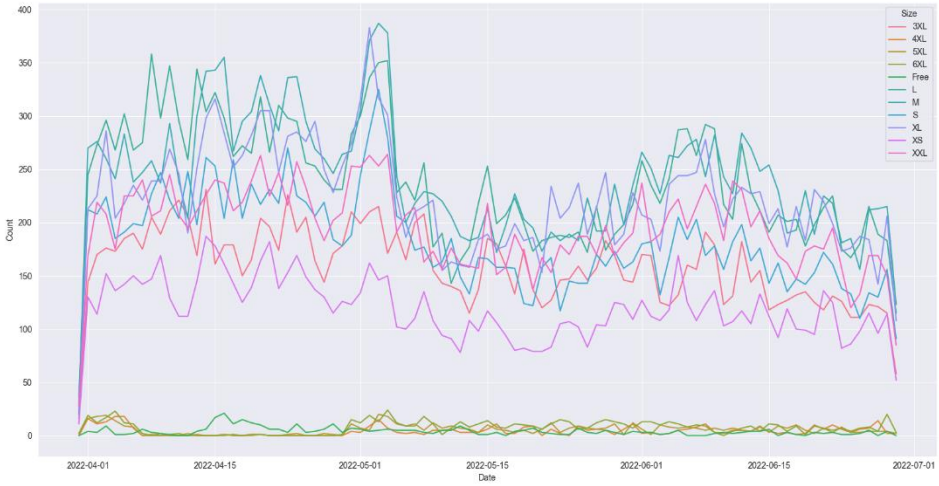
After the data cleaned from nulls and outliers we performed data visualizations using different visualization channels

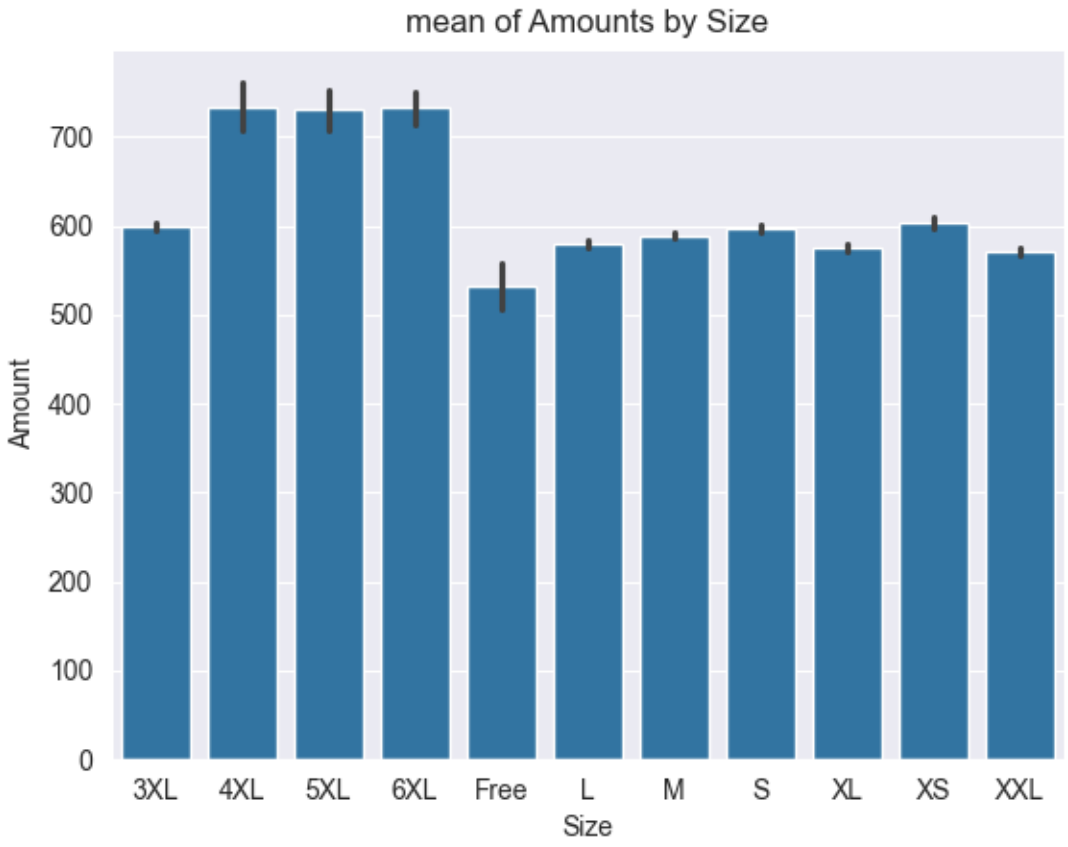





Count of Each Category sold
Count
50000
40000
30000
20000
10000
0
kurta
Set
Western Dress
Top
Ethnic Dress
Blouse
Bottom
Saree
Dupatta
Category

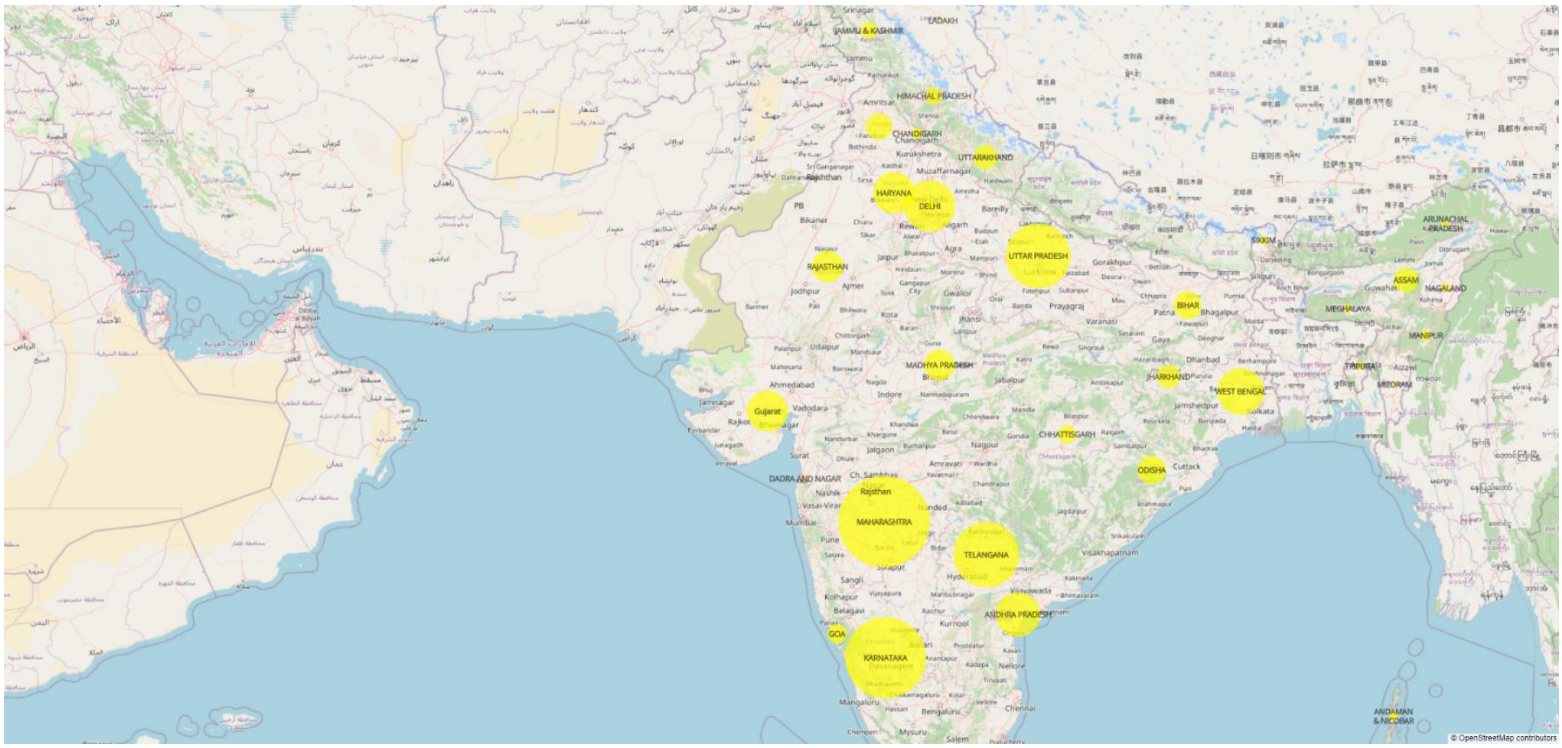

Category
Blouse
Bottom
Dupatta
Ethnic Dress
Saree
Set
Top
Western Dress
kurta
Count
800
600
400
200
0
2022-04-01
2022-04-15
2022-05-01
2022-05-15
2022-06-01
2022-06-15
2022-07-01


Size
3XL
4XL
5XL
6XL
Free
L
M
S
XL
XS
XXL
Count
400
350
300
250
200
150
100
50
0
2022-04-01
2022-04-15
2022-05-01
2022-05-15
2022-06-01
2022-06-15
2022-07-01
Date



And we performed geographical visualizations on amount but before performing that we need to provide important information to the dataframe to help it in visualization like ‘latitude ‘ and ‘longitude’ ,to get them we need

first to send the states names to the encoder to get their location then we aggregate the amounts for every state



5.data modeling

To model the data and make predictions for it we need first to formulate it to the proper format that the models understand, the models doesn't take any input except numeric so we need to transform the categorical data to numeric using proper encoding methods ,so the categorical columns that have many categories we use 'one hot encoding' with it because may large number affect on bias and the columns that have categorical data 3 or 2 categories ,the label encoder the most suitable encoder for them

So, after the data encoded we remove nulls if appears and split the dataset into training and testing and send the training data to the model to learn and then evaluate the model performance using testing data

5.1 logistic regression

The model process depends mainly on non linear function to determine the output

```
Accuracy: 0.9707506647576192
Precision: 0.9483662468611637
Recall: 0.9707506647576192
F1 Score: 0.9583863619364734

Classification Report:
              precision    recall  f1-score   support

         0.0       0.97      1.00      0.99      3518
         1.0       0.56      0.15      0.24       131
         2.0       0.49      0.65      0.56        46
         3.0       1.00      1.00      1.00     14702
         5.0       0.91      1.00      0.95      5480
         6.0       0.00      0.00      0.00         1
         7.0       0.00      0.00      0.00         2
         8.0       0.00      0.00      0.00       183
         9.0       0.00      0.00      0.00         2
        10.0       0.00      0.00      0.00       347
        11.0       0.00      0.00      0.00        32
        12.0       0.00      0.00      0.00         1

    accuracy                           0.97     24445
   macro avg       0.33      0.32      0.31     24445
weighted avg       0.95      0.97      0.96     24445
```

5.2. decision tree

Decision tree depends mainly on the probability,so it starts with the most important features and make branches and leafs of other features to predict the out put

```
Accuracy: 0.9706279402740847
Precision: 0.9484041535825082
Recall: 0.9706279402740847
F1 Score: 0.9581997417459145

Classification Report:
              precision    recall  f1-score   support

         0.0       0.97      1.00      0.99      3518
         1.0       0.55      0.12      0.20       131
         2.0       0.46      0.72      0.56        46
         3.0       1.00      1.00      1.00     14702
         5.0       0.91      1.00      0.95      5480
         6.0       0.00      0.00      0.00         1
         7.0       0.00      0.00      0.00         2
         8.0       0.00      0.00      0.00       183
         9.0       0.00      0.00      0.00         2
        10.0       0.00      0.00      0.00       347
        11.0       0.00      0.00      0.00        32
        12.0       0.00      0.00      0.00         1

    accuracy                           0.97     24445
   macro avg       0.32      0.32      0.31     24445
weighted avg       0.95      0.97      0.96     24445
```

5.3 Random Forest classifier

It contains of multiple decision trees that works at the same time and it evaluates the mean of them

```
Accuracy: 0.9570055226017591
Precision: 0.9485428062779476
Recall: 0.9570055226017591
F1 Score: 0.9525008841052426

Classification Report:
              precision    recall  f1-score   support

         0.0       0.97      0.99      0.98      3518
         1.0       0.35      0.18      0.24       131
         2.0       0.52      0.65      0.58        46
         3.0       1.00      1.00      1.00     14702
         5.0       0.91      0.94      0.92      5480
         6.0       0.00      0.00      0.00         1
         7.0       0.00      0.00      0.00         2
         8.0       0.04      0.02      0.03       183
         9.0       0.00      0.00      0.00         2
        10.0       0.06      0.04      0.05       347
        11.0       0.00      0.00      0.00        32
        12.0       0.00      0.00      0.00         1

    accuracy                           0.96     24445
   macro avg       0.32      0.32      0.32     24445
weighted avg       0.95      0.96      0.95     24445
```

5.4 cross-validation

We use it to check the performance of models doesn't affect by the difference of the data

We used here the K-folds which splits the dataset into k folds and get k-1 folds for training and the remaining 1 for testing, so it trained and tested all over the data

```
Cross Validation Scores:  [0.97807323 0.98343152 0.9779496  0.91531664 0.98093602]
Average CV Score:  0.9671414001536114
Number of CV Scores used in Average:  5
```