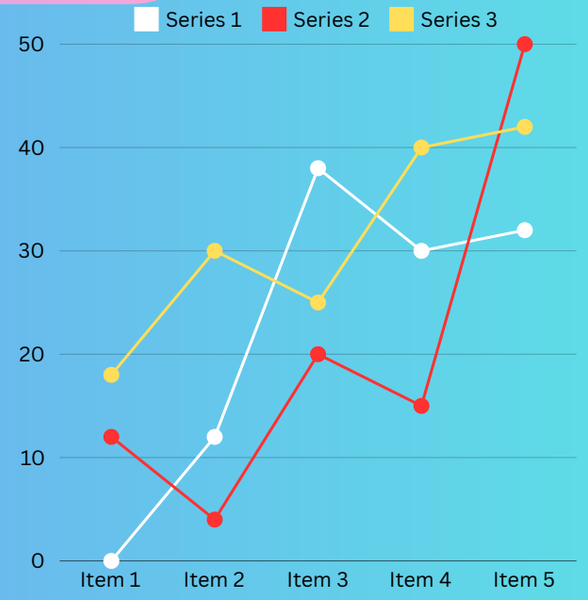
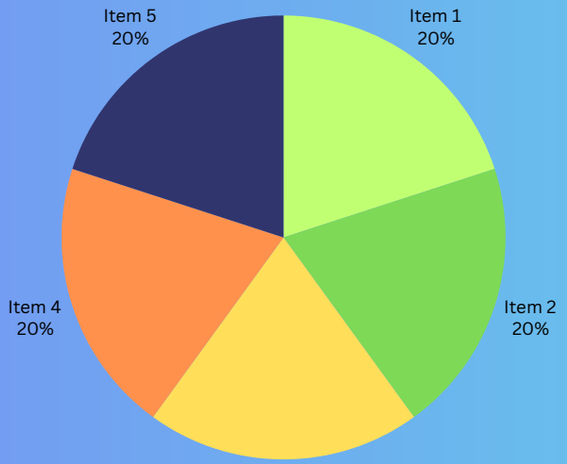
Business Analytics





Ankit Vishwakarma

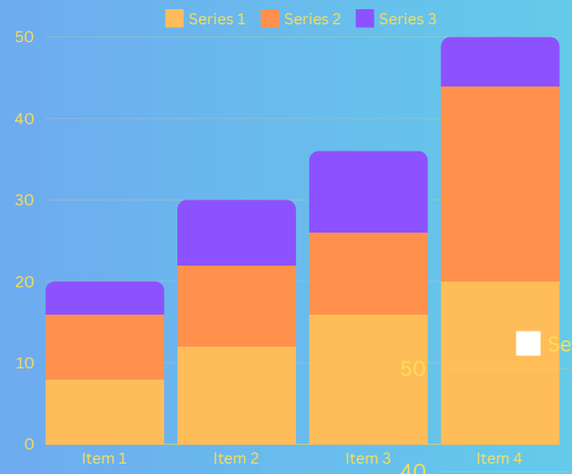
Sage University Bhopal

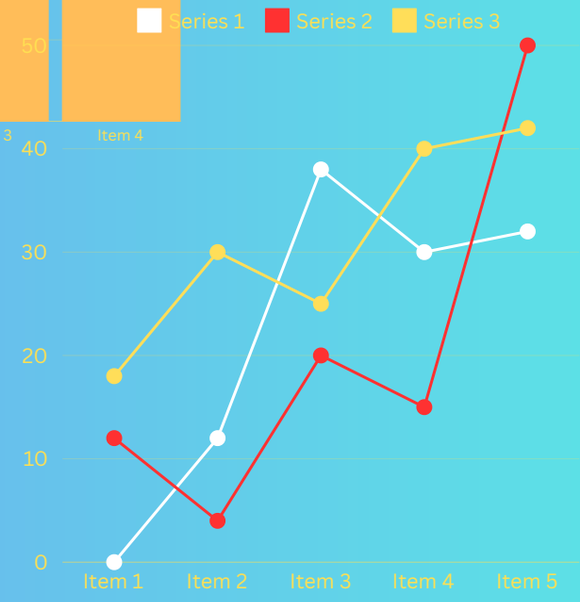
The task given to us is to perform data preprocessing and exploratory data analysis on our dataset.

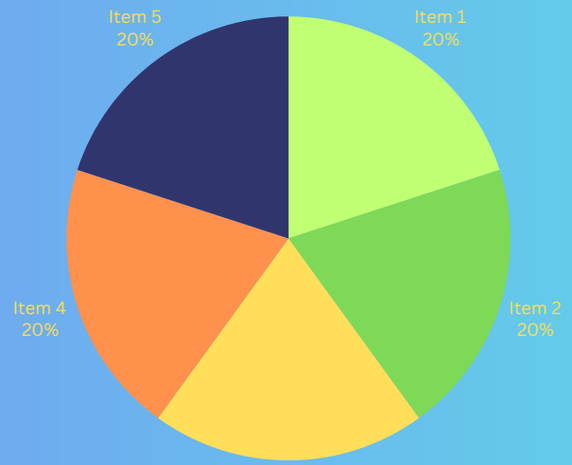
For this we have been given three datasets

1. Amazon - Review Sentiment
2. Customer Segmentation
3. Netflix Movies & TV Shows

I preferred to choose the first one which was Amazon .





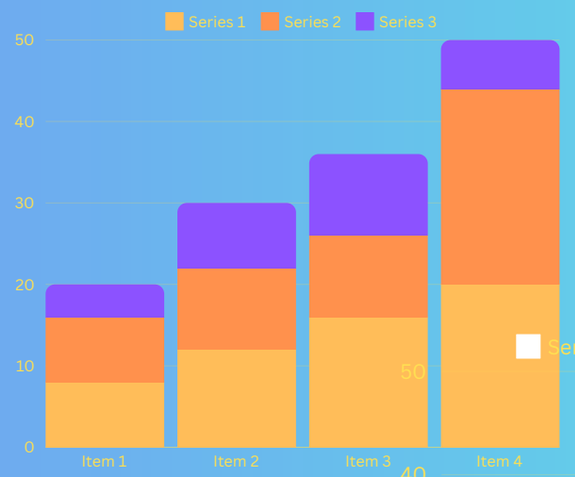


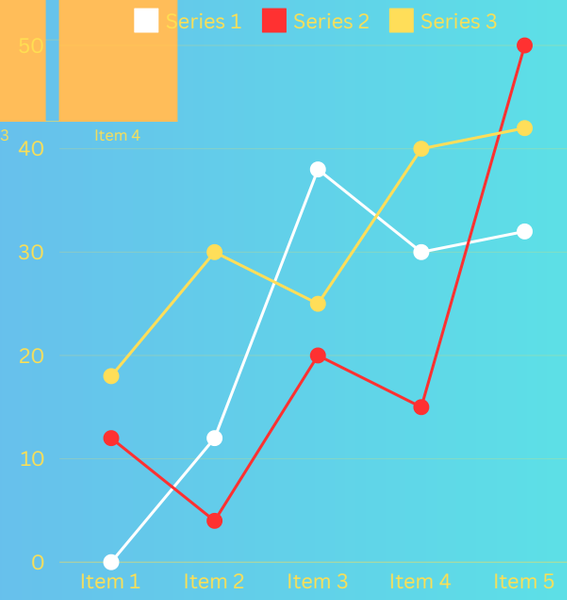
Amazon

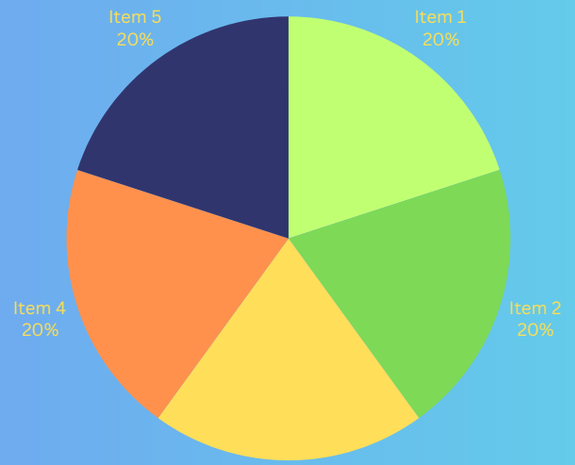
- **Why only Amazon?**

  Because,Amazon is a multinational technology company that specializes in e-commerce, cloud computing, digital streaming, and artificial intelligence. It is one of the world's largest online retailers and cloud service providers.

- I thought that if I work on Amazon dataset then I will get to know a lot.







1. **Importing the required libraries for EDA**
   - import pandas as pd
   - import numpy as np
   - import matplotlib.pyplot as plt
   - import seaborn as sns
2. **Loading the data into the data frame.**
   - df = pd.read_csv('amazon_reviews.csv')
3. **Checking the types of data**
   - df.dtypes
4. **We can change the data type of " reviewime" and ship data from object to date formate.**
   - df['reviewTime'] = pd.to_datetime(df['reviewTime'])
5. **Summary**
   - df.describe()
6. **Feature engineering,we can create more Columns,To know more information about dataset**
   - df['reviewYear'] = df['reviewTime'].dt.year
7. **Dropping irrelevant columns**
   - df.drop(['reviewText'],axis=1,inplace=True)
8. **Dropping the duplicate rows**
   - duplicate_rows_df = df[df.duplicated()]
9. **Dropping the missing or null values.**
   - df.dropna(inplace=True)

- Now i perform Exploratory Data Analysis,so that I can find the useful information from the given data,which will help the company to grow their business.

1. **Univeriate Analysis**

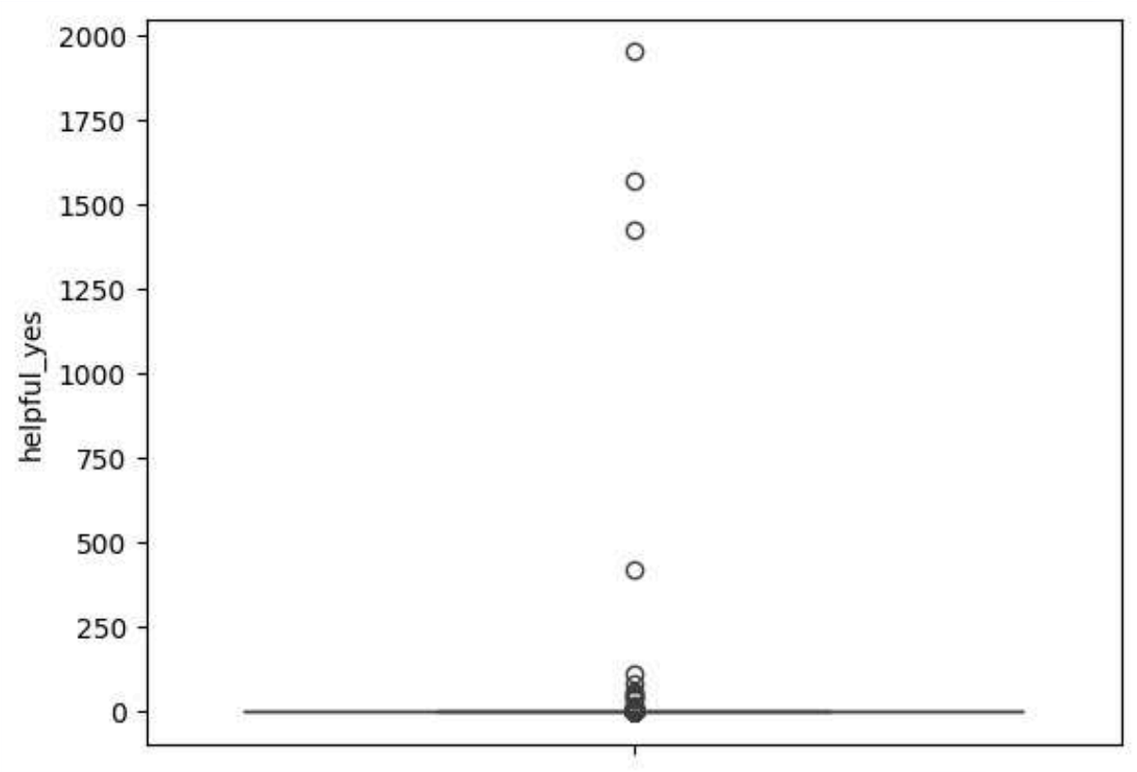
- **#**Detecting Outliers
- **sns.boxplot(df['helpful_yes'])**
- if we see in helpful_yes column of dataset
- then we find that their is more Outliers

2.**sns.countplot(x=df['reviewMonth'])**

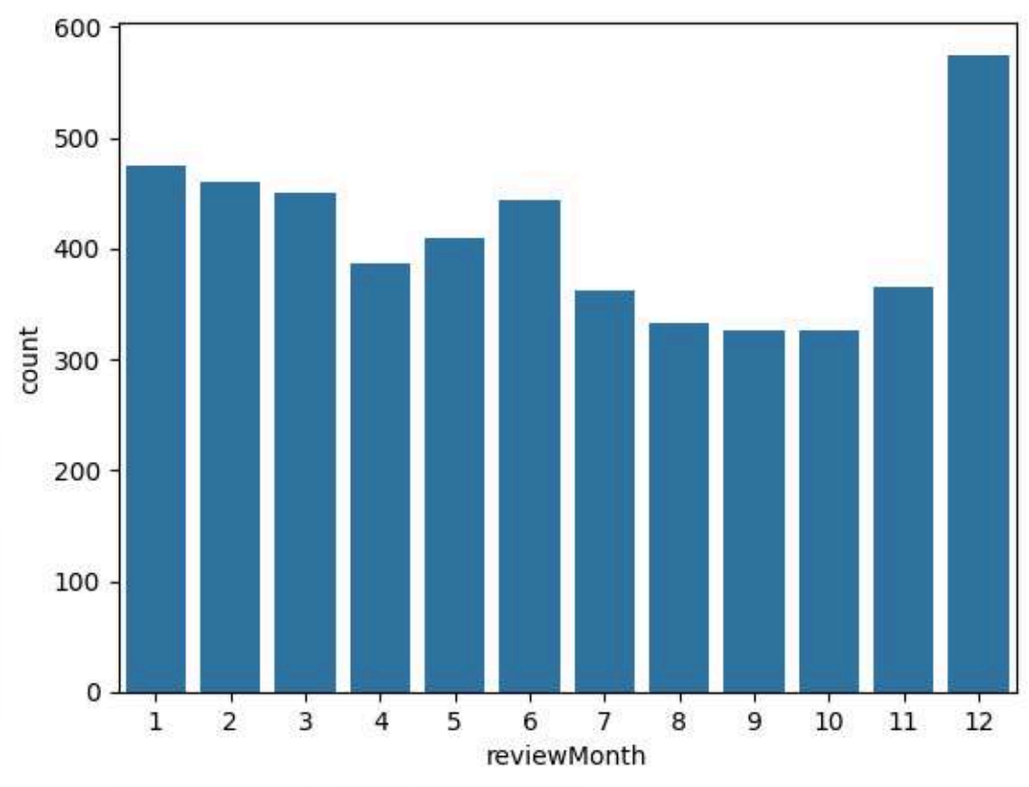
- After ploting review month of dataset
- we got to know that
- most of peoples give their review last
- month of the year

Graph

“hlepful_yes”



“reviewMonth”

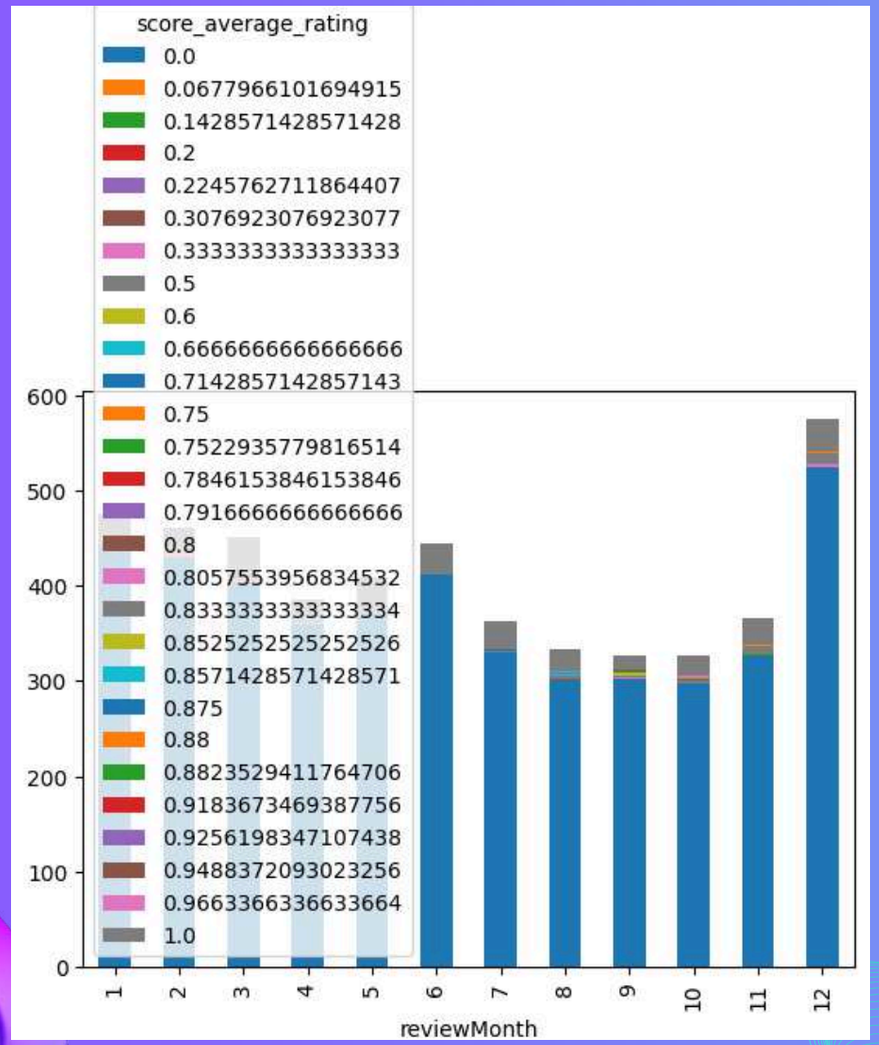


- **#Bivariate Analysis**
- da =df.groupby('reviewMonth')['score_average_rating'].value_counts()
- da.unstack().plot(kind="bar",stacked=True)
- #After ploting two columns together
- #reviewMonth and score_average_rating
- # here we got to know that which month get how much
- #scoreAverageRating

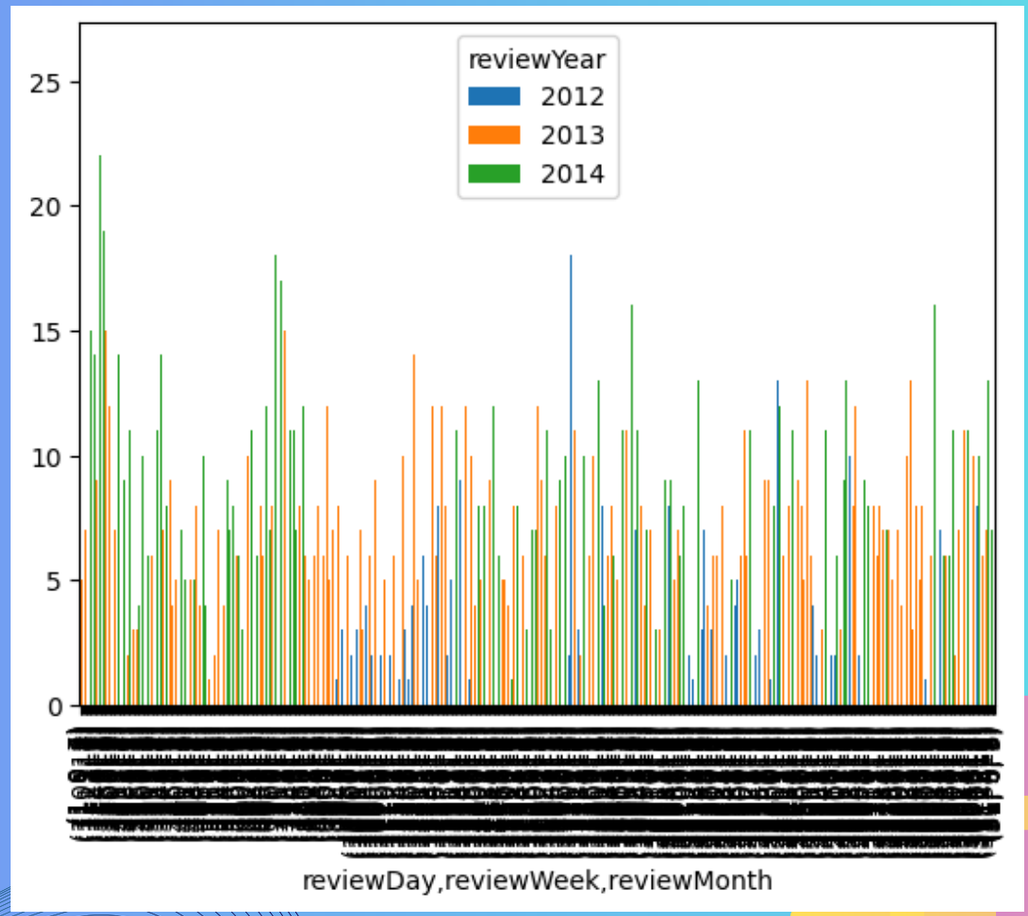
- **#Multivariate Analysis**
- dnew = df.groupby(['reviewDay','reviewWeek','reviewMonth'])['reviewYear'].value_counts()
- dnew.unstack().plot(kind="bar",stacked=True)
- #when i plot more then two columns together
- #reviewDay,reviewWeek,reviewMonth and reviewYear
- #then i find that i can see all the review according to Day,Week,Month,Year
- # at a single graph

Graph

“Bivariate Analysis”



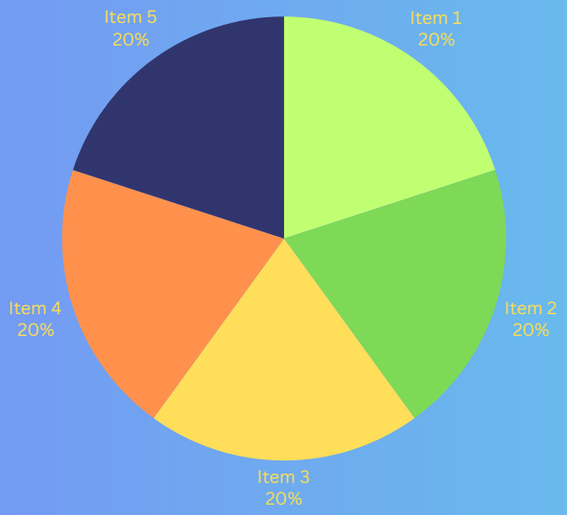
“Multivariate Analysis”

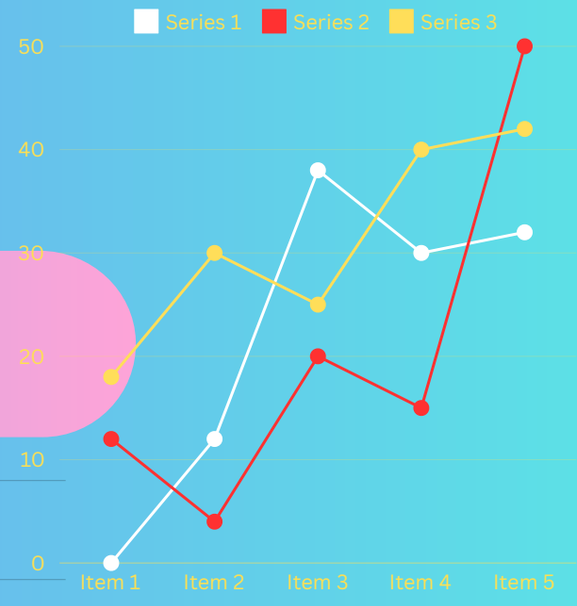


Conclusion

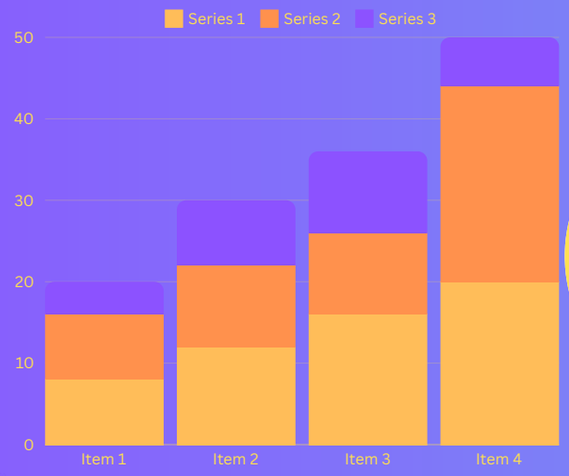
- Finally, after preprocessing and exploring the data, it turns out that the dataset's columns are missing a lot of data.
- There were many people who did not like to give feedback and reviews.
- We also found out from the dataset that most people like to give reviews in the last month of the year.
- At last we can say with the help of graphing aur ploting we can better understand the dataset ,because it not possible see whole Rows and Columns at time but we plot that data into a graph then it is possible to see.
- data preprocessing and exploring play a crucial role business analytics


Group 1
Subgroup 1
Subgroup 2
Subgroup 3
Group 4
Subgroup 8
Subgroup 9
Group 2
Subgroup 4
Subgroup 5
Group 3
Subgroup 6
Subgroup 7

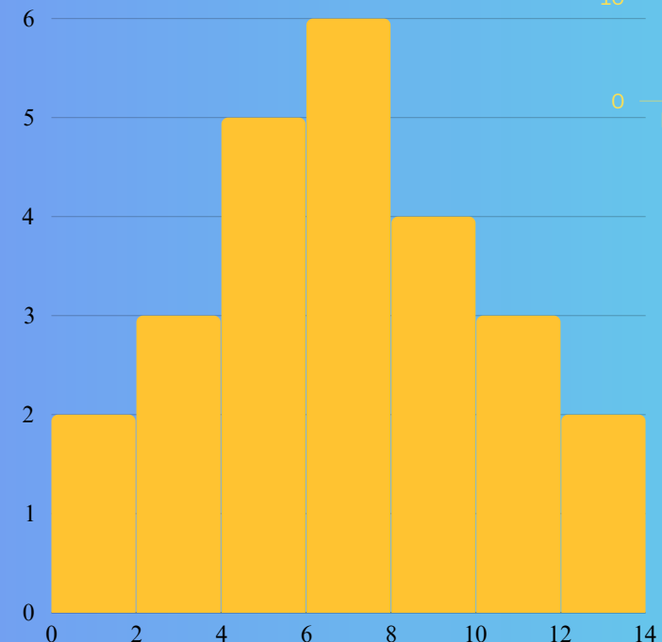

Item 5
20%
Item 1
20%
Item 4
20%
Item 2
20%
Item 3
20%


Series 1
Series 2
Series 3
50
40
30
20
10
0
Item 1
Item 2
Item 3
Item 4
Item 5

Thank you 🙏


Series 1
Series 2
Series 3
50
40
30
20
10
0
Item 1
Item 2
Item 3
Item 4


Group 1
Subgroup 1
Subgroup 2
Subgroup 3
Group 3
Subgr...
Subg...
Group 4
Subgroup 8
Subgroup 9
Group 2
Subgrou...
Subgro...


6
5
4
3
2
1
0
0
2
4
6
8
10
12
14