**Diwali Sales Analysis Project Report**

**Task-**To analyse the Diwali Sales of a product based company and provide useful insights so that the company can figure out their prime customer base and target them better to achieve more sales in the next season

**About the data set !**

The data set which I used in this project has data of 11.25 k people , the data set has features like

['User\_ID', 'Cust\_name', 'Product\_ID', 'Gender', 'Age Group', 'Age',

'Marital\_Status', 'State', 'Zone', 'Occupation', 'Product\_Category',

'Orders', 'Amount', 'Status', 'unnamed1']

These features are describing the customer and also describing the products which he/she bought , it also has an in depth information about the region to which the customer belongs and also the occupational background of the customer

-The data set has 11.25k samples and 15 features

-There are significant amount of NaN entries in the data set

**Pandas Profiling**

Utilised pandas profiling modules’ profile report function to generate a html file which has detailed analysis of the above mentioned data set .

from ydata\_profiling import ProfileReport

**Preprocessing**

-The features ‘Status’, ‘unnamed1’ are having 11.25 k NaN entries each so I am dropping them

-The feature ‘Amount’ has about 12 NaN entries , and I am dropping these samples

-I am multiplying the ‘Orders’ feature with the ‘Amount’ feature to figure out the ‘total amount’ spent by a person , also I am retaining the ‘Orders’ feature and removing the ‘Amount’ feature

**Univariate and Multivariate Analysis**

**1-** Using the count plot of sns I figured out that there are about 8k females in the data set , which implies that product of this company are liked more by females

**2**-Using a countplot with hue-‘Marital status’ we can see that the products are slightly more preferred by unmarried males and females

**3**-Mean age of the people in the data set is 35 , where as more people are from the age group - 25-35 which I discovered using dist plot

**4**-Figured out that 25 percent people are under the age of 25 and middle 50 percent people are from the age group 35-45 , this I discovered using a box plot

**5**- sns.barplot(x='Gender',y='Total Amount',data=df2,hue='Marital\_Status') This helped me in figuring out that most amount is spent by married females

**6**-

a>df3=df.groupby(['State'],as\_index=False)['Orders'].count().sort\_values(by='Orders',ascending=False)

b>df3=df.groupby(['Gender','State'],as\_index=False)['Orders'].count().sort\_values(by='Orders',ascending=False

from this data frame it is quite visible that most of the customers are from

Uttar Pradesh ,Maharashtra ,Karnatka, and are females

**7**-

ax=sns.countplot(x=df['Occupation'],hue=df['Gender'])

most of the people are from IT Aviation and HealthCare sectors ,most of them are females

**8**- df4=df.groupby(['State','Gender','Occupation'],as\_index=False)['Orders'].count().sort\_values(by='Orders',ascending=False)

df4.head(10)

Further from this we can see that from the States UP, Karnatka, Maharashtra females from IT Health care Aviation and Banking are the prime customers

**9-**

**Now let us see which sector spent the most on the products of this company**

df5=df.groupby(['Occupation'],as\_index=False)['Total Amount'].sum().sort\_values(by='Total Amount',ascending=False)

df5.head(10)

well again we can see that IT ,Healthcare ,Aviation sectors spent the most amount

**10-**

df6=df.groupby(['Occupation','Gender'],as\_index=False)['Total Amount'].sum().sort\_values(by='Total Amount',ascending=False)

df6.head(10)

Agin we can see that its the females from the same sectors which are spending this much amount.

**11-**

**Let us see which product was the most sold product or the hot product**

df7=df.groupby(['Product\_Category','Gender'],as\_index=False)['Orders'].sum().sort\_values(by='Orders',ascending=False).head(8)

df7

ax=sns.barplot(x=df7['Product\_Category'],y=df7['Orders'],hue=df7['Gender'])

for i in ax.containers:

ax.bar\_label(i)

# food

# clothing

#electronic etc were the most popular categories with the most orders and amount spent , that too by females

**12-**

which product has been sold the most ?

df.groupby(['Product\_Category','Product\_ID'],as\_index=False)['Orders'].sum().sort\_values(by=['Orders'],ascending=False)

# the product POO265242 of FOOD CATEGORY was the hot product with the highest orders

**CONCLUSION**

**OVER ALL CONCLUSION IS THAT MARRIED/UNMARRIED FEMALES FROM UP KARNATKA MAHARASHTRA , FROM IT HEALTHCARE AVIATION**

**WERE THE MAJOR CUSTOMERS OF THE COMPANY AND THEY WERE INTERESTED IN THE PRODUCTS OF THE CATEGORY FOOD , CLOTHING AND ELECTRONICS**