The Spark Foundation Data Science and Business Analytics Intern Task 3 **Exploratory Data Analysis** • Objective : To find out the weak areas where you can work to make more profit. • Dataset : https://bit.ly/3i4rbWl • By: MAYURI ARUN PATHAK Setting Working Directory In [1]: import os os.chdir("H:\\Data Science\\Internship\\Spark") Importing Libraries import pandas as pd import numpy as np import seaborn as sns import matplotlib.pyplot as plt #Surpress warnings import warnings warnings.filterwarnings('ignore') Understanding Dataset df = pd.read_csv("SampleSuperstore.csv") print(df) Ship Mode Segment Country City State \ Second Class Kentucky Consumer United States Henderson Second Class Consumer United States Henderson Kentucky 1 Second Class Corporate United States Los Angeles California Standard Class Consumer United States Fort Lauderdale 3 Florida Standard Class Consumer United States Fort Lauderdale Florida 4 Second Class Florida Consumer United States Miami 9989 Costa Mesa California Costa Mesa California Costa Mesa California Standard Class Consumer United States 9990 9991 Standard Class Consumer United States 9992 Standard Class Consumer United States 9993 Second Class Consumer United States Westminster California Postal Code Region Category Sub-Category Sales Quantity \ 0 42420 South Furniture Bookcases 261.9600 2 Chairs 731.9400 42420 South Furniture 3 1 90036 West Office Supplies Labels 14.6200 2 2 33311 South Furniture Tables 957.5775 5 3 4 33311 South Office Supplies Storage 22.3680 ... Furniture Furnishings 25.2480 3 9989 33180 South 9990 Furniture Furnishings 91.9600 92627 West 9991 92627 West Technology Phones 258.5760 2 92627 West Office Supplies Paper 29.6000 9992 4 92683 West Office Supplies Appliances 243.1600 9993 Discount Profit 0.00 219.5820 0.00 6.8714 0.45 -383.0310 3 2.5164 4 0.20 9989 4.1028 0.20 9990 0.00 15.6332 9991 19.3932 0.20 13.3200 9992 0.00 72.9480 9993 0.00 [9994 rows x 13 columns] Basic Data Insights In [15]: df.sample(5) Out[15]: Ship Mode Segment City State Postal Code Region Category Sub-Category Sales Quantity Discount Profit Country 2587 Standard Class Consumer United States New York City New York 9.7812 10024 East Furniture Furnishings 22.230 0.0 1 7334 Standard Class Consumer United States Furnishings 54.528 0.2 14.3136 Florida 33614 South Furniture 3 1790 Standard Class Consumer United States New York City New York 10011 East Office Supplies Paper 48.940 1 0.0 24.4700 776 Standard Class Consumer United States 45231 East Office Supplies Art 32.760 0.2 3.6855 Cincinnati 82 Standard Class Consumer United States San Francisco California West Office Supplies 0.0 6.2031 94122 Storage 21.390 1 In [16]: df.head() Ship Mode Segment City State Postal Code Region **Category Sub-Category** Sales Quantity Discount Profit Out[16]: Country **0** Second Class Consumer United States Bookcases 261.9600 41.9136 Henderson Kentucky 42420 South Furniture 0.00 Second Class Consumer United States Henderson Kentucky 42420 South Furniture Chairs 731.9400 0.00 219.5820 2 **2** Second Class Corporate United States California 90036 West Office Supplies Labels 14.6200 0.00 6.8714 Los Angeles 3 Standard Class Consumer United States Fort Lauderdale 33311 South Furniture Tables 957.5775 0.45 -383.0310 4 Standard Class Consumer United States Fort Lauderdale Florida 33311 South Office Supplies Storage 22.3680 2 0.20 2.5164 df.tail() In [17]: Out[17]: Ship Mode Segment Country City State Postal Code Region **Category Sub-Category** Sales Quantity Discount **Profit 9989** Second Class Consumer United States 25.248 0.2 4.1028 Miami Florida 33180 South Furniture Furnishings Furnishings 9990 Standard Class Consumer United States Costa Mesa California 92627 West Furniture 91.960 0.0 15.6332 2 9991 Standard Class Consumer United States Costa Mesa Phones 258.576 0.2 19.3932 California 92627 West Technology 4 0.0 13.3200 9992 Standard Class Consumer United States Costa Mesa California 92627 West Office Supplies 29.600 Paper Appliances 243.160 2 9993 Second Class Consumer United States Westminster California 0.0 72.9480 92683 West Office Supplies In [18]: df.shape (9994, 13) Out[18]: In [19]: df.info() <class 'pandas.core.frame.DataFrame'> RangeIndex: 9994 entries, 0 to 9993 Data columns (total 13 columns): # Column Non-Null Count Dtype -----Ship Mode 0 9994 non-null object 1 Segment 9994 non-null object 2 9994 non-null Country object 3 9994 non-null City object 4 9994 non-null object State Postal Code 9994 non-null int64 9994 non-null object 9994 non-null Category Sub-Category 9994 non-null float64 Sales 9994 non-null 10 Quantity 9994 non-null float64 11 Discount 9994 non-null 12 Profit 9994 non-null float64 dtypes: float64(3), int64(2), object(8) memory usage: 1015.1+ KB df.describe() In [20]: Out[20]: **Postal Code** Discount Profit Sales Quantity 9994.000000 9994.000000 9994.000000 9994.000000 **count** 9994.000000 229.858001 28.656896 mean 55190.379428 3.789574 0.156203 std 32063.693350 623.245101 2.225110 0.206452 234.260108 1040.000000 0.444000 1.000000 0.000000 -6599.978000 **25%** 23223.000000 17.280000 2.000000 0.000000 1.728750 **50%** 56430.500000 54.490000 3.000000 0.200000 8.666500 **75%** 90008.000000 209.940000 5.000000 0.200000 29.364000 max 99301.000000 22638.480000 14.000000 0.800000 8399.976000 • Number of unique values in each column: In [21]: **for** i **in** df.columns : print(i,len(df[i].unique())) Ship Mode 4 Segment 3 Country 1 City 531 State 49 Postal Code 631 Region 4 Category 3 Sub-Category 17 Sales 5825 Quantity 14 Discount 12 Profit 7287 Check for null values df.isnull().sum() 0 Ship Mode Out[22] Segment 0 Country 0 City 0 State 0 Postal Code Region Category Sub-Category Sales Quantity 0 Discount 0 Profit dtype: int64 **Data Visualization** In [14]: sns.pairplot(df) <seaborn.axisgrid.PairGrid at 0x1ce69953fd0> 100000 80000 Postal Code 60000 40000 20000 20000 15000 5000 0.00000 00 000000 (0 0.0000 01((E 00)0 C(CDD C CE CD) 3((C)C) (C)C(C) (C(C)C) CO (E((O O EO O) COC 02 0 0 0 000 00 101 00 00 (0 (0 0)) **600333) GD 600** CO (O)000 O) CE (())))) 0.8 0.6 Discount P.0 . Since 0.2 (0) 0 (0) 0 0.0 7500 5000 2500 -2500-5000 50000 100000 20000 0.00 0.25 0.50 -5000 5000 10000 Ó Postal Code Profit fig, axes = plt.subplots(1,1,figsize = (12,7))sns.heatmap(df.corr()) plt.show() Postal Code - 0.8 - 0.6 0.2 Discount 0.0 Postal Code Sales Quantity Discount Profit In [26]: fig,axes = plt.subplots(1,2,figsize=(14,5)) fig.suptitle("Total profit Vs sales") sns.barplot(data = df.groupby('Sub-Category')['Sales', 'Profit'].agg(sum), x="Sales", y = "Profit", ax = axes[1]) df.groupby('Sub-Category')['Sales', 'Profit'].agg(sum).plot(kind='bar', ax=axes[0]) plt.xticks(rotation=90) plt.show() Total profit Vs sales Sales Profit 50000 300000 40000 250000 30000 200000 20000 150000 10000 100000 50000 -10000-20000 Paper 114879.9963 149528.03 107532.161000000001 167380.318 Sub-Category In [27]: fig, axes = plt.subplots(1,2,figsize =(14,5)) fig.suptitle("Total Sales Vs Quantity") sns.barplot(data=df.groupby('Sub-Category')['Sales', 'Quantity'].agg(sum), x='Sales', y='Quantity', ax=axes[1]) df.groupby('Sub-Category')['Sales', 'Quantity'].agg(sum).plot(kind='bar', ax=axes[0]) plt.xticks(rotation=90) plt.show() Total Sales Vs Quantity Sales 6000 Quantity 300000 5000 250000 4000 200000 3000 150000 2000 100000 1000 50000 91705.164 107532.161000000001 114879.9963 Sub-Category Sales fig, axes = plt.subplots(1,2,figsize =(14,5)) df.groupby('Sub-Category')['Discount', 'Profit'].agg(sum).plot(kind='bar',ax=axes[0]).set_title('Discount & Profit Relation based on Sub-Category') df.groupby('Sub-Category')['Profit','Quantity'].agg(sum).plot(kind='bar',ax=axes[1]).set_title('Quantity & Profit Relation based on Sub-Category') plt.xticks(rotation=90) plt.show() Discount & Profit Relation based on Sub-Category Quantity & Profit Relation based on Sub-Category Discount Quantity Profit 50000 50000 40000 40000 30000 30000 20000 20000 10000 10000 -10000-10000-20000 Phones Storage Supplies Copiers Chairs Chairs Sub-Category Sub-Category In [33]: fig, axes = plt.subplots(2,2,figsize= (16,8)) fig.suptitle("Distribution plots", fontsize = 16) sns.distplot(df['Sales'], ax=axes[0,0]) sns.distplot(df['Profit'], ax=axes[0,1]) sns.distplot(df['Discount'], ax=axes[1,0]) sns.distplot(df['Quantity'], ax=axes[1,1]) plt.show() Distribution plots 0.008 0.0025 0.007 0.006 0.0020 0.005 0.0015 0.0015 0.004 0.003 0.0010 0.002 0.0005 0.001 0.0000 0.000 5000 10000 15000 20000 -4000 -2000 4000 6000 8000 -6000 2000 25 20 0.6 £ 15 Del 10 0.2 Discount Quantity In [34]: fig, axes = plt.subplots(2,2,figsize=(16,8)) fig.suptitle("Sales with different shipping modes and Segments" , fontsize =16) sns.barplot(df['Ship Mode'], df['Sales'], ax=axes[0,0]) sns.lineplot(df['Ship Mode'], df['Sales'], ax=axes[0,1]) sns.barplot(df['Segment'], df['Sales'], ax=axes[1,0]) sns.lineplot(df['Segment'], df['Sales'], ax=axes[1,1]) plt.show() Sales with different shipping modes and Segments 300 280 250 260 200 S 150 <u>용</u> 240 100 220 50 200 Second Class Standard Class First Class Same Day Second Class Standard Class First Class Same Day Ship Mode Ship Mode 280 250 270 260 200 250 S 150 图 240 100 230 220 50 210 0 -Home Office Home Office Consumer Corporate Corporate Consumer Segment Segment fig, ax= plt.subplots(1,1,figsize=(12,7)) sns.countplot(df['Quantity'], hue=df['Region']) plt.show() 800 Region South West 700 Central East 600 500 400 400 300 200 Conclusions: • The features Profit and Discounts are highly related. Over less quantity of products also the sales were high. • The maximum quantity of product in demand was in range 2-4. • The mode of shipping doesn't affect much to the sales. • The Home Office provides highest sales followed by Corporate by slight variation. Suggestions are always Welcome! Thank You!