Flipkart E-Commerce Data Analysis Mohammad Areeb: Linkedin - www.linkedin.com/in/mohammadareeb2544: github - https://github.com/areeb399 Importing Libraries import numpy as np In [1]: import pandas as pd import matplotlib.pyplot as plt import seaborn as sns import plotly.express as px from plotly import graph\_objects as go Importing data In [2]: df = pd.read\_excel("flipkart.xlsx") In [3]: Out[3]: uniq\_id crawl\_timestamp product\_name product\_category\_tree pid retail\_price discounted\_price product\_url Alisha Solid ["Clothing >> Women's 2016-03-25 http://www.flipkart.com/alisha-solid-0 c2d766ca982eca8304150849735ffef9 Women's Clothing >> Lingerie, SRTEH2FF9KEDEFGF 999.0 379.0 ["http://img5a. 22:59:23 +0000 women-s-c... Cycling Shorts SI... FabHomeDecor ["Furniture >> Living http://www.flipkart.com/fabhomedecor-2016-03-25 22646.0 ["http://img6a Room Furniture >> Sofa SBEEH3QGU7MFYJFY **1** 7f7036a6d550aaa89d34c77bd39a5e48 Fabric Double 32157.0 22:59:23 +0000 fabric-do... Sofa Bed В... ["Footwear >> Women's 2016-03-25 http://www.flipkart.com/aw-**2** f449ec65dcbc041b6ae5e6a32717d01b AW Bellies Footwear >> Ballerinas SHOEH4GRSUBJGZXE 999.0 499.0 ["http://img5a. 22:59:23 +0000 bellies/p/itmeh4grg... Alisha Solid ["Clothing >> Women's 2016-03-25 http://www.flipkart.com/alisha-solid-Clothing >> Lingerie, 267.0 ["http://img5a. **3** 0973b37acd0c664e3de26e97e5571454 Women's SRTEH2F6HUZMQ6SJ 699.0 22:59:23 +0000 women-s-c... Cycling Shorts Sl... Sicons All ["Pet Supplies >> http://www.flipkart.com/sicons-all-2016-03-25 ["http Purpose Arnica Grooming >> Skin & PSOEH3ZYDMSYARJ5 210.0 bc940ea42ee6bef5ac7cea3fb5cfbee7 220.0 22:59:23 +0000 purpose-arn... Dog Shampoo Coat Care... WallDesign ["Baby Care >> Baby & 2015-12-01 http://www.flipkart.com/walldesign-["http 7179d2f6c4ad50a17d014ca1d2815156 STIE7KFJAKSTDY9G 730.0 19995 Small Vinyl Kids Gifts >> Stickers 1500.0 10:15:43 +0000 small-vinyl... Sticker Wallmantra ["Baby Care >> Baby & 2015-12-01 http://www.flipkart.com/wallmantra-19996 71ac419198359d37b8fe5e3fffdfee09 Large Vinyl Kids Gifts >> Stickers STIE9F5URNQGJCGH 1429.0 1143.0 ["http://img6a 10:15:43 +0000 large-vinyl... Stickers Sticker Elite Collection ["Baby Care >> Baby & 2015-12-01 http://www.flipkart.com/elite-collection-**19997** 93e9d343837400ce0d7980874ece471c Medium Acrylic Kids Gifts >> Stickers STIE7VAYDKQZEBSD 1299.0 999.0 ["http://img5a 10:15:43 +0000 mediu... Sticker Elite Collection ["Baby Care >> Baby & 2015-12-01 http://www.flipkart.com/elite-collection-669e79b8fa5d9ae020841c0c97d5e935 Medium Acrylic Kids Gifts >> Stickers STIE8YSVEPPCZ42Y 1499.0 1199.0 ["http://img5a 10:15:43 +0000 mediu... Sticker Elite Collection ["Baby Care >> Baby & 2015-12-01 http://www.flipkart.com/elite-collection-STIE88KN9ZDSGZKY ["http://img6a 19999 cb4fa87a874f715fff567f7b7b3be79c Medium Acrylic Kids Gifts >> Stickers 1499.0 999.0 10:15:43 +0000 mediu... Sticker >... 20000 rows × 15 columns Summary of the statistics for each numeric column df.describe() In [4]: retail\_price discounted\_price Out[4]: count 19922.000000 19922.000000 mean 2979.206104 1973.401767 9009.639341 7333.586040 std min 35.000000 35.000000 25% 666.000000 350.000000 50% 1040.000000 550.000000 75% 1999.000000 999.000000 max 571230.000000 571230.000000 Checking null(or missing) values df.isnull() In [5] pid retail\_price discounted\_price image is\_FK\_Advantage\_product description product\_rating overall\_rating be Out[5]: uniq\_id crawl\_timestamp product\_url product\_name product\_category\_tree 0 False F False False False False False False False False 1 False False False False False F 2 False F 3 False 4 False F 19995 False F 19996 False F False False False 19997 False F 19998 False F 19999 False F 20000 rows × 15 columns Total number of missing values in each columns df.isnull().sum() In [6]: uniq\_id 0 Out[6]: crawl\_timestamp 0 product\_url 0 0 product\_name product\_category\_tree 0 0 pid retail\_price 78 78 discounted\_price 3 image is\_FK\_Advantage\_product 0 2 description 0 product\_rating overall\_rating 0 brand 5864 product\_specifications 14 dtype: int64 Top 5 values In [7]: df.head() Out[7]: uniq\_id crawl\_timestamp product\_url product\_name product\_category\_tree pid retail\_price discounted\_price ["Clothing >> Women's Alisha Solid 2016-03-25 http://www.flipkart.com/alisha-solid-SRTEH2FF9KEDEFGF c2d766ca982eca8304150849735ffef9 379.0 ["http://img5a.flixca Women's 999.0 Clothing >> Lingerie, 22:59:23 +0000 women-s-c.. Cycling Shorts ["Furniture >> Living FabHomeDecor 2016-03-25 http://www.flipkart.com/fabhomedecor-7f7036a6d550aaa89d34c77bd39a5e48 Fabric Double Room Furniture >> Sofa SBEEH3QGU7MFYJFY 32157.0 22646.0 ["http://img6a.flixca 22:59:23 +0000 fabric-do... Sofa Bed В... ["Footwear >> Women's 2016-03-25 http://www.flipkart.com/aw-499.0 ["http://img5a.flixca **2** f449ec65dcbc041b6ae5e6a32717d01b AW Bellies Footwear >> Ballerinas SHOEH4GRSUBJGZXE 999.0 22:59:23 +0000 bellies/p/itmeh4grg... Alisha Solid ["Clothing >> Women's http://www.flipkart.com/alisha-solid-2016-03-25 699.0 **3** 0973b37acd0c664e3de26e97e5571454 Women's SRTEH2F6HUZMQ6SJ 267.0 ["http://img5a.flixca Clothing >> Lingerie, 22:59:23 +0000 women-s-c... Cycling Shorts Sl... Sicons All ["Pet Supplies >> 2016-03-25 http://www.flipkart.com/sicons-all-["http://im PSOEH3ZYDMSYARJ5 220.0 210.0 bc940ea42ee6bef5ac7cea3fb5cfbee7 Purpose Arnica Grooming >> Skin & 22:59:23 +0000 purpose-arn... Dog Shampoo Coat Care... Bottom 5 values df.tail() uniq\_id crawl\_timestamp product\_url product\_name product\_category\_tree pid retail\_price discounted\_price Out[8]: WallDesign ["Baby Care >> Baby & 2015-12-01 http://www.flipkart.com/walldesign-["http://img 19995 7179d2f6c4ad50a17d014ca1d2815156 Small Vinyl Kids Gifts >> Stickers STIE7KFJAKSTDY9G 1500.0 730.0 10:15:43 +0000 small-vinyl... Sticker >... Wallmantra ["Baby Care >> Baby & 2015-12-01 http://www.flipkart.com/wallmantra-19996 71ac419198359d37b8fe5e3fffdfee09 Kids Gifts >> Stickers STIE9F5URNQGJCGH 1429.0 1143.0 ["http://img6a.flixca Large Vinyl 10:15:43 +0000 large-vinyl... Stickers Sticker >... Elite Collection ["Baby Care >> Baby & 2015-12-01 http://www.flipkart.com/elite-**19997** 93e9d343837400ce0d7980874ece471c Medium Acrylic Kids Gifts >> Stickers STIE7VAYDKQZEBSD 1299.0 999.0 ["http://img5a.flixca 10:15:43 +0000 collection-mediu... Sticker >... ["Baby Care >> Baby & Elite Collection 2015-12-01 http://www.flipkart.com/elite-**19998** 669e79b8fa5d9ae020841c0c97d5e935 STIE8YSVEPPCZ42Y 1499.0 1199.0 ["http://img5a.flixca Medium Acrylic Kids Gifts >> Stickers 10:15:43 +0000 collection-mediu... Sticker Elite Collection ["Baby Care >> Baby & 2015-12-01 http://www.flipkart.com/elite-19999 cb4fa87a874f715fff567f7b7b3be79c Kids Gifts >> Stickers STIE88KN9ZDSGZKY 1499.0 999.0 ["http://img6a.flixca Medium Acrylic 10:15:43 +0000 collection-mediu... Sticker Filling missing values in a column 'retail\_price' with the median value of that column df['retail\_price'].fillna(df['retail\_price'].median(), inplace = True) In [9]: Filling missing values in a column 'discounted\_price' with the median value of that column In [10]: df['discounted\_price'].fillna(df['discounted\_price'].median(), inplace = True) Calculating and creating a new column named 'discount\_percentage' x = df['retail\_price'] - df['discounted\_price']  $y = (x / df['retail_price']) * 100$ df['discount\_percentage'] = y In [12]: df.head() uniq\_id crawl\_timestamp product\_url product\_name product\_category\_tree pid retail\_price discounted\_price Alisha Solid ["Clothing >> Women's 2016-03-25 http://www.flipkart.com/alisha-solidc2d766ca982eca8304150849735ffef9 SRTEH2FF9KEDEFGF 999.0 Women's Clothing >> Lingerie, 379.0 ["http://img5a.flixca 22:59:23 +0000 women-s-c... Cycling Shorts FabHomeDecor ["Furniture >> Living 2016-03-25 http://www.flipkart.com/fabhomedecor-22646.0 ["http://img6a.flixca **1** 7f7036a6d550aaa89d34c77bd39a5e48 Fabric Double SBEEH3QGU7MFYJFY 32157.0 Room Furniture >> Sofa 22:59:23 +0000 fabric-do... Sofa Bed В... ["Footwear >> Women's 2016-03-25 http://www.flipkart.com/aw-Footwear >> Ballerinas SHOEH4GRSUBJGZXE **2** f449ec65dcbc041b6ae5e6a32717d01b 999.0 499.0 ["http://img5a.flixca **AW Bellies** 22:59:23 +0000 bellies/p/itmeh4grg... Alisha Solid ["Clothing >> Women's 2016-03-25 http://www.flipkart.com/alisha-solid-**3** 0973b37acd0c664e3de26e97e5571454 SRTEH2F6HUZMQ6SJ 699.0 267.0 ["http://img5a.flixca Women's Clothing >> Lingerie, 22:59:23 +0000 women-s-c... Cycling Shorts Sicons All ["Pet Supplies >> 2016-03-25 http://www.flipkart.com/sicons-all-["http://im Purpose Arnica 220.0 210.0 Grooming >> Skin & PSOEH3ZYDMSYARJ5 bc940ea42ee6bef5ac7cea3fb5cfbee7 22:59:23 +0000 purpose-arn... Dog Shampoo Coat Care... df['discount\_percentage'] 62.062062 Out[13]: 29.576764 50.050050 61.802575 4.545455 19995 51.333333 19996 20.013996 19997 23.094688 19998 20.013342 19999 33.355570 Name: discount\_percentage, Length: 20000, dtype: float64 Converting the 'crawl' timestamp' column into datetime objects and storing them in a new column named 'timestamp' df['timestamp'] = pd.to\_datetime(df['crawl\_timestamp']) In [14]: Createing a new column named 'Time'. The lambda function lambda x : x.time extracts the time component from each datetime object and assigns it to the 'Time' column df['Time'] = df['timestamp'].apply(lambda x : x.time) In [15]: Createing a new column named 'Date'. The lambda function lambda x : x.date extracts the date component from each datetime object and assigns it to the 'Date' column df['Date'] = df['timestamp'].apply(lambda x : x.date) In [16]: Dropping a column named 'crawl\_timestamp' df.drop(['crawl\_timestamp'], axis = 1, inplace = True) In [17]: Creating a new column named 'main\_category' based on the values in the 'product\_category\_tree' column df['main\_category'] = df['product\_category\_tree'].apply(lambda x : x.split('>>')[0][2:len(x.split('>>')[0])-1]) df.head() In [19]: product\_url product\_name product\_category\_tree uniq\_id pid retail\_price discounted\_price Out[19]: Alisha Solid ["Clothing >> Women's http://www.flipkart.com/alisha-solidc2d766ca982eca8304150849735ffef9 SRTEH2FF9KEDEFGF Women's Clothing >> Lingerie, 999.0 379.0 ["http://img5a.flixcart.com/image/short/u women-s-c... Cycling Shorts Sl... FabHomeDecor ["Furniture >> Living http://www.flipkart.com/fabhomedecor-1 7f7036a6d550aaa89d34c77bd39a5e48 Room Furniture >> Sofa SBEEH3QGU7MFYJFY Fabric Double 32157.0 22646.0 ["http://img6a.flixcart.com/image/sofa-b fabric-do... Sofa Bed ["Footwear >> Women's http://www.flipkart.com/aw-Footwear >> Ballerinas SHOEH4GRSUBJGZXE f449ec65dcbc041b6ae5e6a32717d01b AW Bellies 999.0 499.0 ["http://img5a.flixcart.com/image/shoe/7 bellies/p/itmeh4grg... Alisha Solid ["Clothing >> Women's http://www.flipkart.com/alisha-solid-3 0973b37acd0c664e3de26e97e5571454 Women's Clothing >> Lingerie, SRTEH2F6HUZMQ6SJ 699.0 267.0 ["http://img5a.flixcart.com/image/short/6 women-s-c... Cycling Shorts Sl... ["Pet Supplies >> Sicons All http://www.flipkart.com/sicons-all-["http://img5a.flixcart.com/imag 220.0 210.0 bc940ea42ee6bef5ac7cea3fb5cfbee7 Purpose Arnica Grooming >> Skin & PSOEH3ZYDMSYARJ5 purpose-arn... Dog Shampoo Coat Care... Creating a DataFrame named 'top\_products' that contains information about the top 'n' main categories and their corresponding total counts from a DataFrame ('df'). In [20]: n = 10 top\_products=pd.DataFrame(df['main\_category'].value\_counts()[:n]).reset\_index() top\_products.rename(columns = {'index':'Top\_Products', 'main\_category':'Total\_Count'}, inplace = True) In [21]: top\_brands=pd.DataFrame(df['brand'].value\_counts()[:n]).reset\_index() top\_brands.rename(columns = {'index':'Top\_Brands','brand':'Total\_Count'}, inplace = True) df.head() uniq\_id product\_url product\_name product\_category\_tree pid retail\_price discounted\_price Out[22]: Alisha Solid ["Clothing >> Women's http://www.flipkart.com/alisha-solid-Clothing >> Lingerie, c2d766ca982eca8304150849735ffef9 Women's SRTEH2FF9KEDEFGF 999.0 379.0 ["http://img5a.flixcart.com/image/short/L women-s-c... Cycling Shorts SI... FabHomeDecor ["Furniture >> Living http://www.flipkart.com/fabhomedecor-7f7036a6d550aaa89d34c77bd39a5e48 Fabric Double Room Furniture >> Sofa SBEEH3QGU7MFYJFY 32157.0 22646.0 ["http://img6a.flixcart.com/image/sofa-b fabric-do... Sofa Bed В... ["Footwear >> Women's http://www.flipkart.com/aw-Footwear >> Ballerinas SHOEH4GRSUBJGZXE f449ec65dcbc041b6ae5e6a32717d01b AW Bellies 999.0 499.0 ["http://img5a.flixcart.com/image/shoe/7 bellies/p/itmeh4grg... ["Clothing >> Women's Alisha Solid http://www.flipkart.com/alisha-solid-3 0973b37acd0c664e3de26e97e5571454 Women's Clothing >> Lingerie, SRTEH2F6HUZMQ6SJ 699.0 267.0 ["http://img5a.flixcart.com/image/short/6 women-s-c... Cycling Shorts Sl... ["Pet Supplies >> Sicons All http://www.flipkart.com/sicons-all-["http://img5a.flixcart.com/imag 220.0 210.0 bc940ea42ee6bef5ac7cea3fb5cfbee7 Purpose Arnica Grooming >> Skin & PSOEH3ZYDMSYARJ5 purpose-arn... Dog Shampoo Coat Care... Creating a subplot containing two donut-like pie charts to visualize the distribution of top products and top brands from plotly.subplots import make\_subplots In [23]: label1 = top\_products['Top\_Products'] value1=top\_products['Total\_Count'] label2=top\_brands['Top\_Brands'] value2=top\_brands['Total\_Count'] fig\_both = make\_subplots(rows=1, cols=2, specs=[[{'type':'domain'}, {'type':'domain'}]]) fig\_both.add\_trace(go.Pie(labels=label1, values=value1, name="Top Products", pull=[0.3, 0, 0, 0]), fig\_both.add\_trace(go.Pie(labels=label2, values=value2, name="Top Brands",pull=[0.3, 0, 0, 0]), fig\_both.update\_traces(hole=.4, hoverinfo="label+percent+name") fig\_both.update\_layout( title\_text="Top products and brands distribution", annotations=[dict(text='Product', x=0.18, y=0.5, font\_size=20, showarrow=False), dict(text='Brand', x=0.82, y=0.5, font\_size=20, showarrow=False)]) 0 Top products and brands distribution Clothing Jewellery Footwear 19.4% Mobiles & Accessories 13% Automotive Home Decor & Festive Needs 37.3% Beauty and Personal Care 12.4% Home Furnishing 5.84% Kitchen & Dining **Product Brand** Computers 5.96% Allure Auto 11.9% 6.61% Regular 6.42% Voylla 6.09% Slim 9.48% 6.91% 8.73% TheLostPuppy Karatcraft Black White Filtering out rows where the 'discount percentage' is greater than 90 df\_discount = df.query('discount\_percentage > 90') In [24]: Dropping rows with null values df\_discount = df\_discount.dropna() In [25]: Handling spelling errors In [26]: df\_discount["brand"].replace('FashBlush', 'Fash Blush', inplace=True) Grouping data by the "brand" column and calculating the mean of the "discount" percentage" for each brand, sorting the results in descending order In [27]: max\_discount=pd.DataFrame(df\_discount\_groupby('brand')[['discount\_percentage']].mean().sort\_values(by=['discount\_percentage'], ascending=False).reset\_index()) Creating a bar chart from 'max\_discount' In [28]: fig = px.bar(max\_discount, x= 'brand', y='discount\_percentage',color='brand') fig.update\_xaxes(title\_text='Brands') fig.update\_yaxes(title\_text='Mean Discount Percentage') fig.update\_layout(title\_text='Mean Discount Percentages by Brand', title\_x = 0.5) Mean Discount Percentages by Brand 100 brand Rajcrafts Bling Fash Blush 80 Mydress Mystyle Mean Discount Percentage Soulful Threads Instella **Bond Beatz** 60 Fashblush Black KazamaKraft 40 Zaicus **CUBA** SDZ Gia 20 Mydress Mystyle Soulful Threads Fash Blush Kazamakraft Rajcrafts BONG BERK Fashblush SOZ Instella Black <a>e</a>icus CUBA Brands Grouping the data and aggregated by summing the "discounted price" for each unique "uniq id" value In [29]: df\_customer=df.groupby("uniq\_id")[["discounted\_price"]].sum().sort\_values(by=['discounted\_price'], ascending=[False]).reset\_index() #Top 20 customers spending the most In [30]: list1=df\_customer[:20] #plotting a bar graph fig = px.bar(list1, x= 'uniq\_id', y="discounted\_price", color="discounted\_price", color\_continuous\_scale='viridis') fig.update\_layout( title\_text='Top 20 Customers by Spending', title\_x = 0.5, xaxis\_title='Customer ID', yaxis\_title='Total Discounted Price' Top 20 Customers by Spending 600k discounted price 500k 500k Discounted Price 400k 400k 300k 300k 200k 200k 100k 100k 329c5f407ace063e1ce3e88f4105e7e6 agfa5b1a8917b841abaef2a1ce032114 07b0ar742cdcac28d09c29a1e246fft2 c4b0A528852Aa8770c760ed2bbca2ed5 710ed5f2393a4b9e8823aa0029f71f93 ad96000fald9e408a4fc47ea5c1123e5 30/324/95/b6/5c26284893000085/becf1 5f8e0c25e2915bc2383e60ad02b1e4aa e794bd226967af2298a2c7a1f8e2c9bb e3e08f67f1e0b5724e71c6845a7f0ffa b92dedcdcec57fe018148b9b2db237f7 069<sub>84</sub>9081<sub>4944</sub>92829<sub>bd914</sub>203<sub>88</sub>bd52 1b1c83326a749b99c9cafc740f9136fq e0de6958769d030f7e470e590f374d24 OSASZABOJAOBSABJAASCSZIC/b6 3<sub>82546675bc399953779e58d84d56650</sub> eb15c8c168e9ebb8d24deac65e0aec3> 44ac0b6808242ce5c71b971a7620c609 43e444307157aca6bb055c4bada67928 1923846a4f6ed68abe553a11d1d938d2 **Customer ID**