Final Project - Analyzing Sales Data

Date: 23 Novamber 2022

Author: Morakot Teanchai

Course: `Pandas Foundation`

```
# import data
import pandas as pd
df = pd.read_csv("sample-store.csv")
```

```
# preview top 5 rows
df.head(5)
```

	Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name	Segment	Country/Region	City
0	1	CA-2019-152156	11/8/2019	11/11/2019	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderso
1	2	CA-2019-152156	11/8/2019	11/11/2019	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderso
2	3	CA-2019-138688	6/12/2019	6/16/2019	Second Class	DV-13045	Darrin Van Huff	Corporate	United States	Los Angeles
3	4	US-2018-108966	10/11/2018	10/18/2018	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdal
4	5	US-2018-108966	10/11/2018	10/18/2018	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdal

5 rows × 21 columns

```
# shape of dataframe
df.shape
```

```
(9994, 21)
```

```
# see data frame information using .info()
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9994 entries, 0 to 9993
Data columns (total 21 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Row ID          9994 non-null   int64
 1   Order ID        9994 non-null   object
 2   Order Date      9994 non-null   object
 3   Ship Date       9994 non-null   object
 4   Ship Mode       9994 non-null   object
 5   Customer ID     9994 non-null   object
```

```
 6   Customer Name   9994 non-null   object
 7   Segment         9994 non-null   object
 8   Country/Region  9994 non-null   object
 9   City            9994 non-null   object
 10  State           9994 non-null   object
 11  Postal Code     9983 non-null   float64
 12  Region          9994 non-null   object
 13  Product ID      9994 non-null   object
 14  Category        9994 non-null   object
```

We can use `pd.to_datetime()` function to convert columns 'Order Date' and 'Ship Date' to datetime.

```
# example of pd.to_datetime() function
pd.to_datetime(df['Order Date'].head(), format='%m/%d/%Y')
```

```
0   2019-11-08
1   2019-11-08
2   2019-06-12
3   2018-10-11
4   2018-10-11
Name: Order Date, dtype: datetime64[ns]
```

```
# TODO - convert order date and ship date to datetime in the original dataframe
df['Order Date'] = pd.to_datetime(df['Order Date'], format='%m/%d/%Y')
df['Ship Date'] = pd.to_datetime(df['Ship Date'], format='%m/%d/%Y')
```

```
# TODO - count nan in postal code column
df['Postal Code'].isna().sum()
```

```
11
```

```
# TODO - filter rows with missing values
df[df.isna().any(axis =1)]
```

	Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name	Segment	Country/Region	City	...
2234	2235	CA-2020-104066	2020-12-05	2020-12-10	Standard Class	QJ-19255	Quincy Jones	Corporate	United States	Burlington	...
5274	5275	CA-2018-162887	2018-11-07	2018-11-09	Second Class	SV-20785	Stewart Visinsky	Consumer	United States	Burlington	...
8798	8799	US-2019-150140	2019-04-06	2019-04-10	Standard Class	VM-21685	Valerie Mitchum	Home Office	United States	Burlington	...
9146	9147	US-2019-165505	2019-01-23	2019-01-27	Standard Class	CB-12535	Claudia Bergmann	Corporate	United States	Burlington	...
9147	9148	US-2019-165505	2019-01-23	2019-01-27	Standard Class	CB-12535	Claudia Bergmann	Corporate	United States	Burlington	...
9148	9149	US-2019-165505	2019-01-23	2019-01-27	Standard Class	CB-12535	Claudia Bergmann	Corporate	United States	Burlington	...
9386	9387	US-2020-127292	2020-01-19	2020-01-23	Standard Class	RM-19375	Raymond Messe	Consumer	United States	Burlington	...
9387	9388	US-2020-127292	2020-01-19	2020-01-23	Standard Class	RM-19375	Raymond Messe	Consumer	United States	Burlington	...
9388	9389	US-2020-127292	2020-01-19	2020-01-23	Standard Class	RM-19375	Raymond Messe	Consumer	United States	Burlington	...
9389	9390	US-2020-127292	2020-01-19	2020-01-23	Standard Class	RM-19375	Raymond Messe	Consumer	United States	Burlington	...
9741	9742	CA-2018-117086	2018-11-08	2018-11-12	Standard Class	QJ-19255	Quincy Jones	Corporate	United States	Burlington	...

11 rows × 21 columns

```
# TODO - Explore this dataset on your owns, ask your own questions
```

Explore 1: Profit by Region

```
# Explore 1: Profit by Region
df[['Profit','Region']].\
    groupby('Region').sum().\
    sort_values(by=['Profit'], ascending=False)
```

	Profit
Region	
West	108418.4489
East	91522.7800
South	46749.4303
Central	39706.3625

Explore 2: Profit by Segment

```
# Explore 2: Profit by Segment
df[['Profit','Segment']].\
    groupby('Segment').sum().\
    sort_values(by=['Profit'], ascending=False)
```

	Profit
Segment	
Consumer	134119.2092
Corporate	91979.1340
Home Office	60298.6785

Explore 3: Profit by Category

```
# Explore 3:  Profit by Category
df[['Category','Profit']].\
    groupby(['Category']).sum().\
    sort_values(by=['Profit'], ascending=False)
```

	Profit
Category	
Technology	145454.9481
Office Supplies	122490.8008
Furniture	18451.2728

Explore 4: Top 10 Profit of product of 'Technology' category

```
#Explore 4: Top 10 Profit of product of 'Technology' category

df[(df['Category'] == 'Technology')][['Category','Product ID','Product Name','Pro
    groupby(['Category','Product ID','Product Name'])['Profit'].sum().reset_index
    sort_values(by=['Profit'], ascending=False).head(10)
```

	Category	Product ID	Product Name	Profit
159	Technology	TEC-CO-10004722	Canon imageCLASS 2200 Advanced Copier	25199.9280
149	Technology	TEC-CO-10001449	Hewlett Packard LaserJet 3310 Copier	6983.8836
156	Technology	TEC-CO-10003763	Canon PC1060 Personal Laser Copier	4570.9347
176	Technology	TEC-MA-10001127	HP Designjet T520 Inkjet Large Format Printer ...	4094.9766
212	Technology	TEC-MA-10003979	Ativa V4110MDD Micro-Cut Shredder	3772.9461
175	Technology	TEC-MA-10001047	3D Systems Cube Printer, 2nd Generation, Magenta	3717.9714
60	Technology	TEC-AC-10002049	Plantronics Savi W720 Multi-Device Wireless He...	3696.2820
162	Technology	TEC-MA-10000045	Zebra ZM400 Thermal Label Printer	3343.5360
153	Technology	TEC-CO-10002095	Hewlett Packard 610 Color Digital Copier / Pri...	3124.9375
91	Technology	TEC-AC-10003033	Plantronics CS510 - Over-the-Head monaural Wir...	3085.0325

Data Analysis Part

Answer 10 below questions to get credit from this course. Write `pandas` code to find answers.

```
# TODO 01 - how many columns, rows in this dataset
col_num = df.shape[1]
row_num = df.shape[0]
print(f"Columns: {col_num}\nRows: {row_num}")
```

```
Columns: 21
Rows: 9994
```

```
# TODO 02 - is there any missing values?, if there is, which colunm? how many nan
df[df.columns[df.isna().sum() > 0]].isna().sum()
```

```
Postal Code    11
dtype: int64
```

```
# TODO 03 - your friend ask for `California` data, filter it and export csv for h
df_result = df[df['State'] == 'California'] #filter
df_result.to_csv('result_California.csv')   #Export csv
```

```
# TODO 04 - your friend ask for all order data in `California` and `Texas` in 201
df_result =  df[((df['State'] == 'California' )| (df['State'] == 'Texas') )\
           & ( df['Order Date'].dt.strftime('%Y') == '2017')] #filter
df_result.to_csv('result_2017.csv')    #Export csv
```

```
# TODO 05 - how much total sales, average sales, and standard deviation of sales
df_2017 =  df[df['Order Date'].dt.strftime('%Y') == '2017']
print(f"total sales: {df_2017['Sales'].sum()}")
print(f"average sales: {df_2017['Sales'].mean() }")
print(f"standard deviation of sales: {df_2017['Sales'].std()}")
```

```
total sales: 484247.4981
average sales: 242.97415860511794
standard deviation of sales: 754.0533572593683
```

```
# TODO 06 - which Segment has the highest profit in 2018
df[df['Order Date'].dt.strftime('%Y') == '2018'][['Profit','Segment']].\
    groupby('Segment').sum().\
    sort_values(by=['Profit'], ascending=False).head(1)
```

	Profit
Segment	
Consumer	28460.1665

```
# TODO 07 - which top 5 States have the least total sales between 15 April 2019 -
import datetime
df[(df['Order Date'] >= datetime.datetime(2019, 4, 15) ) & (df['Order Date'] <=
    [['State','Sales']].groupby('State').sum().sort_values(by=['Sales'], ascendin
```

	Sales
State	
New Hampshire	49.05
New Mexico	64.08
District of Columbia	117.07
Louisiana	249.80
South Carolina	502.48

```
# TODO 08 - what is the proportion of total sales (%) in West + Central in 2019 e

print("the proportion of total sales (%) in West + Central in 2019: ")
(df[(df['Order Date'].dt.strftime('%Y') == '2019') & ((df['Region'] == 'West') |
/ df[(df['Order Date'].dt.strftime('%Y') == '2019')]['Sales'].sum() ) * 100
```

```
the proportion of total sales (%) in West + Central in 2019:


54.97479891837763
```

```
# TODO 09 - find top 10 popular products in terms of number of orders vs. total s
df_toporder= df[(df['Order Date'].dt.strftime('%Y') == '2019') | (df['Order Date'
                groupby(['Product ID','Product Name'])['Order ID'].count().reset
                .sort_values(by=['Order ID'] , ascending=False).head(10)
df_topsale= df[(df['Order Date'].dt.strftime('%Y') == '2019') | (df['Order Date']
                groupby(['Product ID','Product Name'])['Sales'].sum().reset_inde
                .sort_values(by=['Sales'] , ascending=False).head(10)
```

```
#Print result
print("------------- top 10 popular products in terms of number of orders during
df_toporder
```

```
------------- top 10 popular products in terms of number of orders during 2019-
```

	Product ID	Product Name	Order ID
318	FUR-TA-10001095	Chromcraft Round Conference Tables	12
107	FUR-CH-10003774	Global Wood Trimmed Manager's Task Chair, Khaki	11
625	OFF-BI-10000301	GBC Instant Report Kit	10
1294	OFF-ST-10001325	Sterilite Officeware Hinged File Box	10
693	OFF-BI-10001989	Premium Transparent Presentation Covers by GBC	9
347	FUR-TA-10003473	Bretford Rectangular Conference Table Tops	9
795	OFF-BI-10004364	Storex Dura Pro Binders	9
1534	TEC-AC-10004510	Logitech Desktop MK120 Mouse and keyboard Combo	9
790	OFF-BI-10004236	XtraLife ClearVue Slant-D Ring Binder, White, 3"	9
53	FUR-CH-10000454	Hon Deluxe Fabric Upholstered Stacking Chairs,...	9

```
#Print result
print("------------- top 10 popular products in terms of total sales during 2019-
df_topsale
```

```
------------- top 10 popular products in terms of total sales during 2019-2020
```

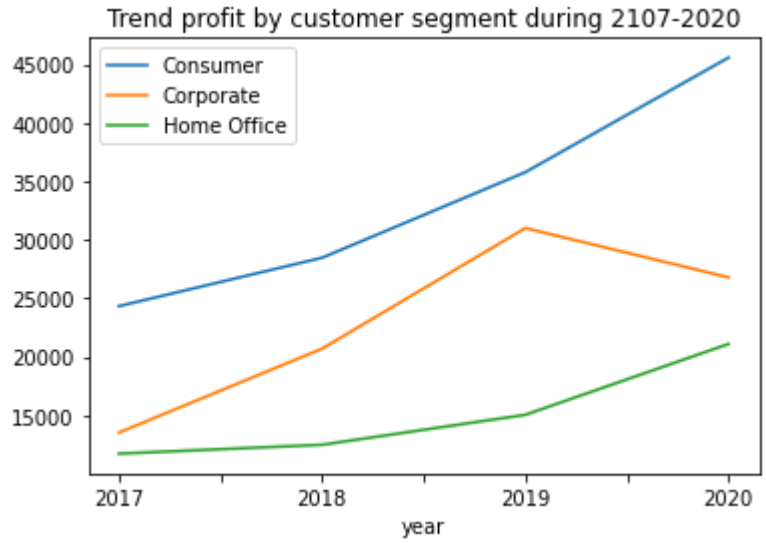
	Product ID	Product Name	Sales
1563	TEC-CO-10004722	Canon imageCLASS 2200 Advanced Copier	61599.824
1554	TEC-CO-10001449	Hewlett Packard LaserJet 3310 Copier	16079.732
1575	TEC-MA-10001047	3D Systems Cube Printer, 2nd Generation, Magenta	14299.890
633	OFF-BI-10000545	GBC Ibimaster 500 Manual ProClick Binding System	13621.542
667	OFF-BI-10001359	GBC DocuBind TL300 Electric Binding System	12737.258
822	OFF-BI-10004995	GBC DocuBind P400 Electric Binding System	12521.108
1653	TEC-PH-10001459	Samsung Galaxy Mega 6.3	12263.708
79	FUR-CH-10002024	HON 5400 Series Task Chairs for Big and Tall	11846.562
1402	OFF-SU-10002881	Martin Yale Chadless Opener Electric Letter Op...	11825.902
66	FUR-CH-10001215	Global Troy Executive Leather Low-Back Tilter	10169.894

```
# TODO 10 - plot at least 2 plots, any plot you think interesting :)
```

#1: Trend profit by customer segment during 2017-2020

```
df['year'] =  df['Order Date'].dt.strftime('%Y')  # Create year columns use to gr
df_result = df[['Profit','Segment','year']].groupby(['year','Segment']).sum().res
# Plot
df_result.set_index('year', inplace=True)
ax = df_result.groupby('Segment')['Profit'].plot(title='Trend profit by customer
```

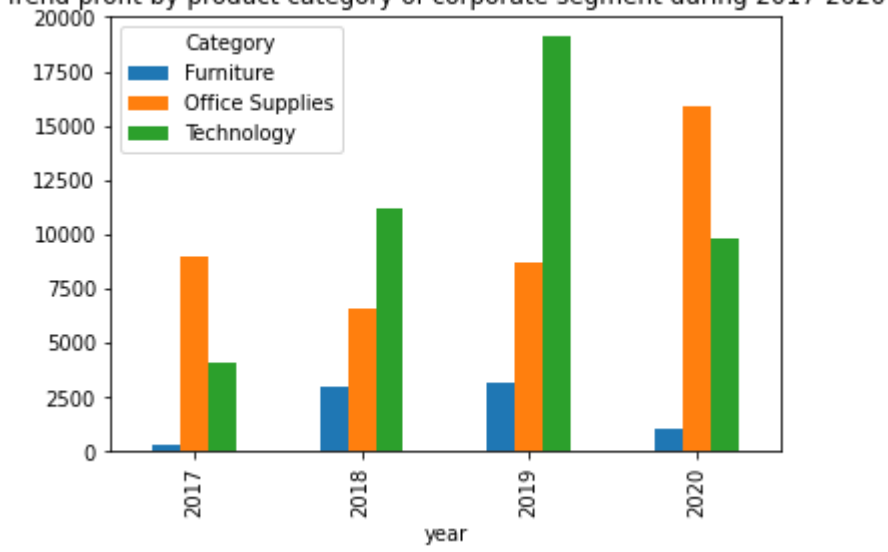
Download



#2: Trend profit by product category of corporate segment during 2017-2020

```
ax = df[(df['Segment'] == 'Corporate')] \
            [['Profit','year','Category']].groupby(['year','Category']).sum().uns
            plot(kind='bar',y='Profit',title = 'Trend profit by product category
```

Download



```
# TODO Bonus - use np.where() to create new column in dataframe to help you answe
```

Product that sale with big lot (big lot > dozen)

```
import numpy as np
```

```
df['Big Lot Sale'] = np.where(
    df['Quantity'] > 12,
    True,
    False
)
```

```
df[df['Big Lot Sale'] == True][['Category','Product Name','Quantity']].groupby(['
              .sort_values(by=['Category','Quantity'] , ascending=False)
```

	Category	Product Name	Quantity
45	Technology	Anker Ultra-Slim Mini Bluetooth 3.0 Wireless K...	14
46	Technology	Logitech Wireless Headset h800	14
50	Technology	Memorex Mini Travel Drive 16 GB USB 2.0 Flash ...	14
51	Technology	Plantronics Voyager Pro HD - Bluetooth Headset	14
52	Technology	PureGear Roll-On Screen Protector	14
53	Technology	Pyle PMP37LED	14
47	Technology	Logitech P710e Mobile Speakerphone	13
48	Technology	Macally Suction Cup Mount	13
49	Technology	Maxell 74 Minute CD-R Spindle, 50/Pack	13
54	Technology	Samsung Galaxy Mega 6.3	13
13	Office Supplies	Acco 7-Outlet Masterpiece Power Center, Wihtou...	14
14	Office Supplies	Avery 4027 File Folder Labels for Dot Matrix P...	14
16	Office Supplies	Cardinal EasyOpen D-Ring Binders	14
18	Office Supplies	Eldon ProFile File 'N Store Portable File Tub ...	14
22	Office Supplies	Ibico Recycled Linen-Style Covers	14
27	Office Supplies	OIC Colored Binder Clips, Assorted Sizes	14
28	Office Supplies	Personal Creations Ink Jet Cards and Labels	14
29	Office Supplies	Pressboard Covers with Storage Hooks, 9 1/2" x...	14
31	Office Supplies	Southworth 100% Résumé Paper, 24lb.	14
33	Office Supplies	Space Solutions Commercial Steel Shelving	14
35	Office Supplies	Staples in misc. colors	14
38	Office Supplies	Wilson Jones Clip & Carry Folder Binder Tool f...	14
40	Office Supplies	Wilson Jones Turn Tabs Binder Tool for Ring Bi...	14
41	Office Supplies	Xerox 1883	14
44	Office Supplies	Xerox 1964	14
15	Office Supplies	Avery 485	13
17	Office Supplies	DXL Angle-View Binders with Locking Rings, Black	13
19	Office Supplies	Fellowes Binding Cases	13
20	Office Supplies	GBC Ibimaster 500 Manual ProClick Binding System	13
21	Office Supplies	Great White Multi-Use Recycled Paper (20Lb. an...	13
23	Office Supplies	Iceberg Mobile Mega Data/Printer Cart	13
24	Office Supplies	Kensington 4 Outlet MasterPiece Compact Power ...	13
25	Office Supplies	Kensington 7 Outlet MasterPiece HOMEOFFICE Pow...	13
26	Office Supplies	Lock-Up Easel 'Spel-Binder'	13

30	Office Supplies	SlimView Poly Binder, 3/8"	13
32	Office Supplies	Southworth 25% Cotton Premium Laser Paper and ...	13
34	Office Supplies	Staples	13
36	Office Supplies	Storex Dura Pro Binders	13
37	Office Supplies	Trimflex Flexible Post Binders	13
39	Office Supplies	Wilson Jones Elliptical Ring 3 1/2" Capacity B...	13
42	Office Supplies	Xerox 1951	13
43	Office Supplies	Xerox 1952	13
2	Furniture	DAX Wood Document Frame	14
3	Furniture	Electrix Architect's Clamp-On Swing Arm Lamp, ...	14
5	Furniture	Global Push Button Manager's Chair, Indigo	14
6	Furniture	Global Stack Chair without Arms, Black	14
7	Furniture	High-Back Leather Manager's Chair	14
8	Furniture	Longer-Life Soft White Bulbs	14
9	Furniture	Metal Folding Chairs, Beige, 4/Carton	14
12	Furniture	Ultra Door Push Plate	14
0	Furniture	Bevis 36 x 72 Conference Tables	13
1	Furniture	Chromcraft Bull-Nose Wood Oval Conference Tabl...	13
4	Furniture	Executive Impressions Supervisor Wall Clock	13
10	Furniture	O'Sullivan 4-Shelf Bookcase in Odessa Pine	13
11	Furniture	Safco Chair Connectors, 6/Carton	13