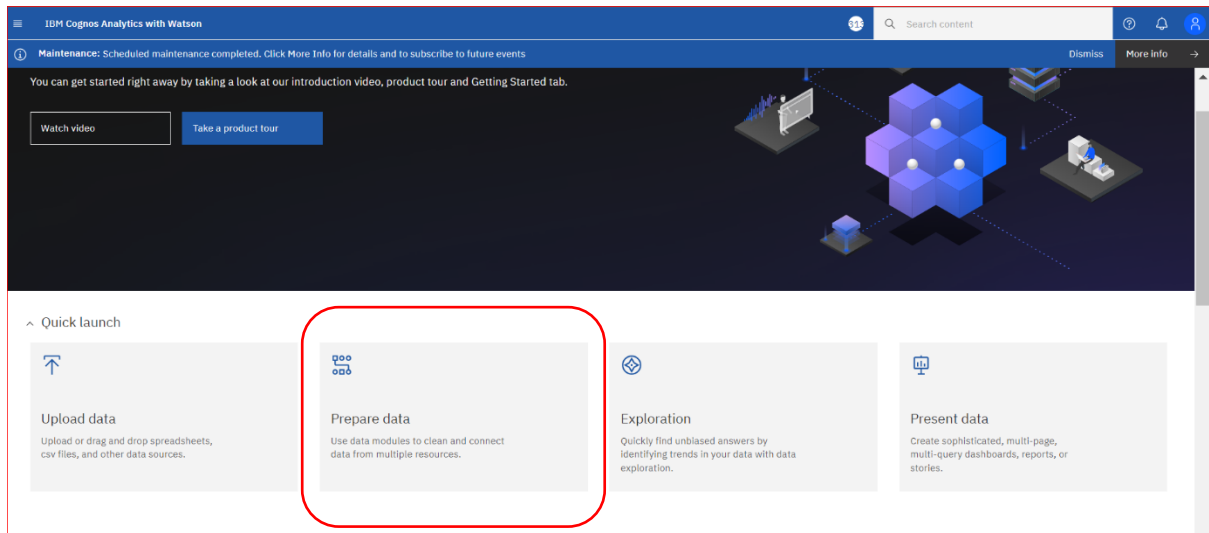
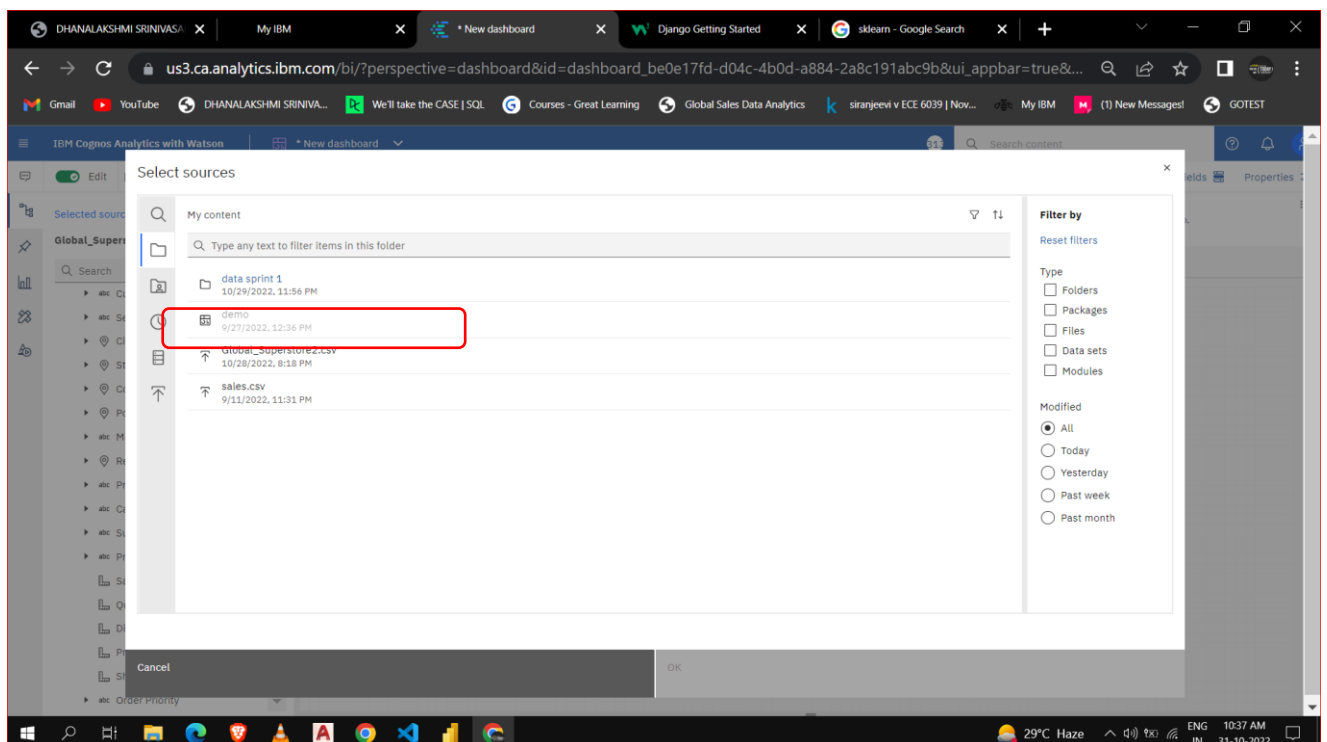


Project development sprint 2

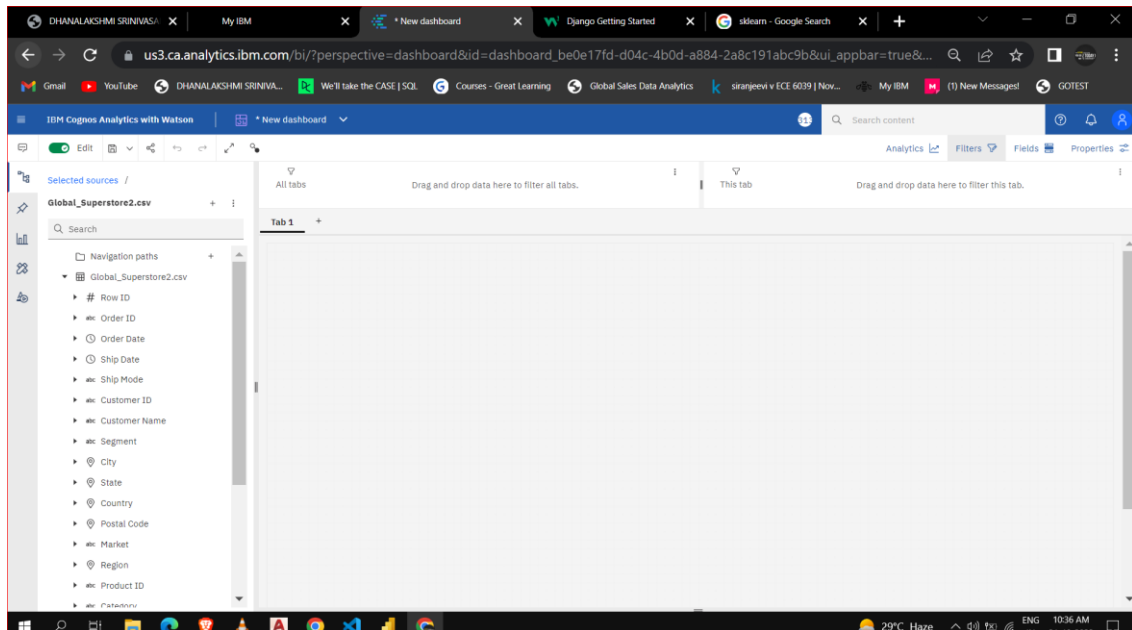
- Prepare Data in IBM COGNOS ANALYTICS



- Select sources file in IBM COGNOS



- New data module for prepare data.



- Loading dataset in Grid

The screenshot shows the IBM Cognos Analytics web interface with the 'Global_Superstore2.csv' dataset loaded into the 'Grid' view. The table displays the following data:

Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name
1	CA-2012-124891	2012-07-31	2012-07-31	Same Day	RH-19495	Rick Hansen
2	IN-2013-77878	2013-02-05	2013-02-07	Second Class	JR-16210	Justin Ritter
3	IN-2013-71249	2013-10-17	2013-10-18	First Class	CR-12730	Craig Reiter
4	ES-2013-1579342	2013-01-28	2013-01-30	First Class	KM-16375	Katherine Murray
5	SG-2013-4320	2013-11-05	2013-11-06	Same Day	RH-9495	Rick Hansen
6	IN-2013-42360	2013-06-28	2013-07-01	Second Class	JM-15655	Jim Mitchum
7	IN-2011-81826	2011-11-07	2011-11-09	First Class	TS-21340	Toby Swindell
8	IN-2012-86369	2012-04-14	2012-04-18	Standard Class	MB-18085	Mick Brown
9	CA-2014-135909	2014-10-14	2014-10-21	Standard Class	JW-15220	Jane Waco
10	CA-2012-116638	2012-01-28	2012-01-31	Second Class	JH-15985	Joseph Holt
11	CA-2011-102988	2011-04-05	2011-04-09	Second Class	GM-14695	Greg Maxwell

- Read the file data in Grid

The screenshot shows the IBM Cognos Analytics interface. The 'Data module' is open, and the 'Grid' view is selected. The data grid displays 11 rows of data. The 'Order ID' column is highlighted in the 'Properties' panel on the right, showing its label, usage as an identifier, and data type as text.

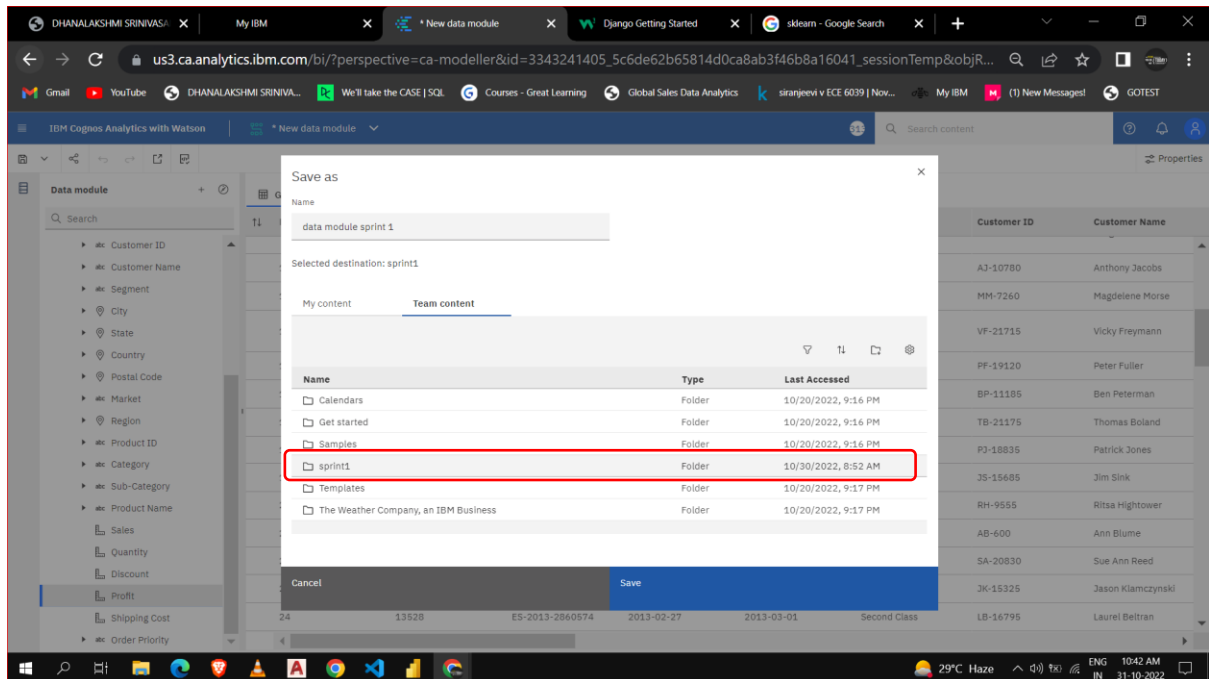
Row ID	Order ID	Order Date	Ship Date	Ship Mode
1	CA-2012-124891	2012-07-31	2012-07-31	Same Day
2	IN-2013-77878	2013-02-05	2013-02-07	Second Class
3	IN-2013-71249	2013-10-17	2013-10-18	First Class
4	ES-2013-1579342	2013-01-28	2013-01-30	First Class
5	SG-2013-4320	2013-11-05	2013-11-06	Same Day
6	IN-2013-42360	2013-06-28	2013-07-01	Second Class
7	IN-2011-81826	2011-11-07	2011-11-09	First Class
8	IN-2012-86369	2012-04-14	2012-04-18	Standard Class
9	CA-2014-135909	2014-10-14	2014-10-21	Standard Class
10	CA-2012-116638	2012-01-28	2012-01-31	Second Class
11	CA-2011-102988	2011-04-05	2011-04-09	Second Class

- Read the file data in Grid

The screenshot shows the IBM Cognos Analytics interface. The 'Data module' is open, and the 'Grid' view is selected. The data grid displays 11 rows of product data. The 'Sub-Category' column is highlighted in the 'Properties' panel on the right, showing its label, usage as an identifier, and data type as text.

Product ID	Category	Sub-Category	Product Name	Sales	Quantity	Discount	Profit
TEC-AC-10003033	Technology	Accessories	Plantronics CS510 - Over-the-Head monaural Wireless Headset System	2309.65	7	0	762.1845
FUR-CH-10003950	Furniture	Chairs	Novimex Executive Leather Armchair, Black	3709.395	9	0.1	-288.765
TEC-PH-10004664	Technology	Phones	Nokia Smart Phone, with Caller ID	5175.171	9	0.1	919.971
TEC-PH-10004583	Technology	Phones	Motorola Smart Phone, Cordless	2892.51	5	0.1	-96.54
TEC-SHA-10000501	Technology	Copiers	Sharp Wireless Fax, High-Speed	2832.96	8	0	311.52
TEC-PH-10000030	Technology	Phones	Samsung Smart Phone, with Caller ID	2862.675	5	0.1	763.275
FUR-CH-10004050	Furniture	Chairs	Novimex Executive Leather Armchair, Adjustable	1822.08	4	0	564.84
FUR-TA-10002958	Furniture	Tables	Chromcraft Conference Table, Fully Assembled	5244.84	6	0	996.48
OFF-BI-10003527	Office Supplies	Binders	Fellowes PB500 Electric Punch Plastic Comb Binding Machine with Manual Bind	5083.96	5	0.2	1906.485
FUR-TA-10000198	Furniture	Tables	Chromcraft Bull-Nose Wood Oval Conference Tables & Bases	4297.644	13	0.4	-1862.3124
OFF-SU-10002881	Office Supplies	Supplies	Martin Yale Chadless Opener Electric Letter	4164.05	5	0	83.281

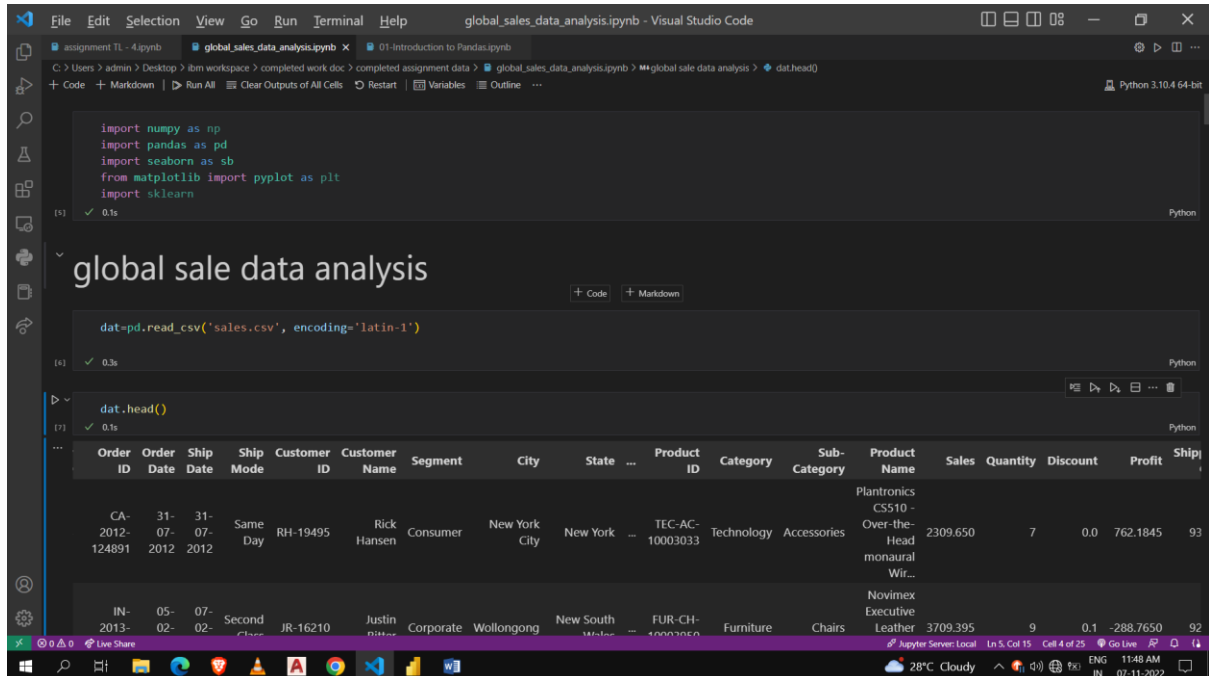
- Save as the preparing data



PROJECT PLANNING PHASE

SPRINT – 2

(DATA PREPRATION ,DATA CLEANING, DATA FORMAT)



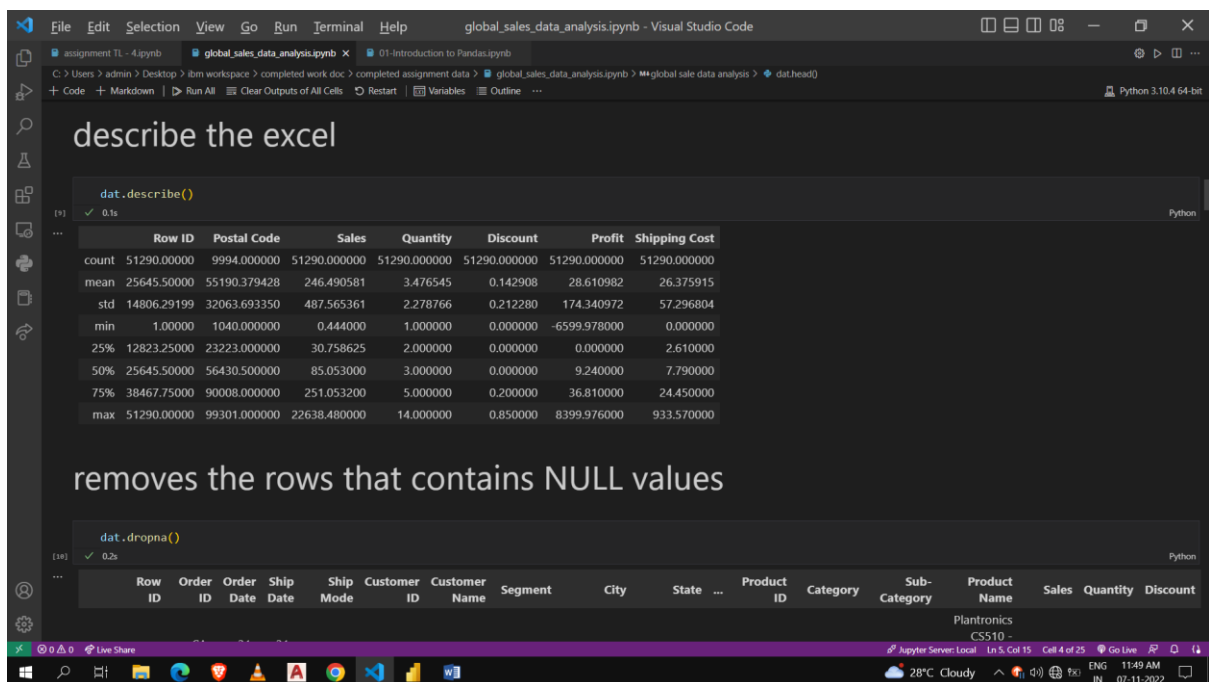
The screenshot shows a Jupyter notebook titled "global sale data analysis" in Visual Studio Code. The notebook has three cells. The first cell contains import statements for numpy, pandas, seaborn, matplotlib, and sklearn. The second cell contains the code to read a CSV file: `dat=pd.read_csv('sales.csv', encoding='latin-1')`. The third cell contains `dat.head()`, which displays the first five rows of the dataset. The output shows columns for Order ID, Order Date, Ship Date, Ship Mode, Customer ID, Customer Name, Segment, City, State, Product ID, Category, Sub-Category, Product Name, Sales, Quantity, Discount, Profit, and Ship Cost. The first row shows a Plantronics headset with a sales value of 2309.650.

```
import numpy as np
import pandas as pd
import seaborn as sb
from matplotlib import pyplot as plt
import sklearn
```

```
dat=pd.read_csv('sales.csv', encoding='latin-1')
```

```
dat.head()
```

Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name	Segment	City	State	Product ID	Category	Sub-Category	Product Name	Sales	Quantity	Discount	Profit	Ship Cost
CA-2012-124891	31-07-2012	31-07-2012	Same Day	RH-19495	Rick Hansen	Consumer	New York City	New York	TEC-AC-10003033	Technology	Accessories	Plantronics CS510 - Over-the-Head monaural Wir...	2309.650	7	0.0	762.1845	93
IN-2013-124891	05-02-2013	07-02-2013	Second Class	JR-16210	Justin Pitter	Corporate	Wollongong	New South Wales	FUR-CH-10003033	Furniture	Chairs	Novimex Executive Leather	3709.395	9	0.1	-288.7650	92



The screenshot shows the same Jupyter notebook with two additional cells. The fourth cell contains `dat.describe()`, which displays a summary of the data. The output shows statistics for Row ID, Postal Code, Sales, Quantity, Discount, Profit, and Shipping Cost. The fifth cell contains `dat.dropna()`, which removes rows containing NULL values. The output shows the first row of the dataset after cleaning, which is the same Plantronics headset as in the previous screenshot.

```
dat.describe()
```

	Row ID	Postal Code	Sales	Quantity	Discount	Profit	Shipping Cost
count	51290.000000	9994.000000	51290.000000	51290.000000	51290.000000	51290.000000	51290.000000
mean	25645.500000	55190.379428	246.490581	3.476545	0.142908	28.610982	26.375915
std	14806.29199	32063.693350	487.565361	2.278766	0.212280	174.340972	57.296804
min	1.000000	1040.000000	0.444000	1.000000	0.000000	-6599.978000	0.000000
25%	12823.25000	23223.000000	30.758625	2.000000	0.000000	0.000000	2.610000
50%	25645.50000	56430.500000	85.053000	3.000000	0.000000	9.240000	7.790000
75%	38467.75000	90008.000000	251.053200	5.000000	0.200000	36.810000	24.450000
max	51290.00000	99301.000000	22638.480000	14.000000	0.850000	8399.976000	933.570000

```
dat.dropna()
```

Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name	Segment	City	State	Product ID	Category	Sub-Category	Product Name	Sales	Quantity	Discount	Ship Cost
1	CA-2012-124891	31-07-2012	31-07-2012	Same Day	RH-19495	Rick Hansen	Consumer	New York City	New York	TEC-AC-10003033	Technology	Accessories	Plantronics CS510 -	2309.650	7	0.0	93

global_sales_data_analysis.ipynb - Visual Studio Code

assignment TL - 4.ipynb global_sales_data_analysis.ipynb x 01-Introduction to Pandas.ipynb

C:\Users> admin> Desktop> ibm workspace> completed work doc> completed assignment data> global_sales_data_analysis.ipynb> global sale data analysis> dat.head()

+ Code + Markdown | Run All | Clear Outputs of All Cells | Restart | Variables | Outline

Python 3.10.4 64-bit

removes the rows that contains NULL values

```
dat.dropna()
```

[14]: ✓ 0.2s Python

	Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name	Segment	City	State	Product ID	Category	Sub-Category	Product Name	Sales	Quantity	Discount
0	32298	CA-124891	2012-07-12	2012-07-12	Same Day	RH-19495	Rick Hansen	Consumer	New York City	New York	TEC-AC-10003033	Technology	Accessories	Plantronics CS510 - Over-the-Head monaural Wir...	2309.650	7	0.0
8	40155	CA-135909	2014-10-14	2014-10-14	Standard Class	JW-15220	Jane Waco	Corporate	Sacramento	California	OFF-BI-10003527	Office Supplies	Binders	Fellowes PB500 Electric Punch Plastic Comb Bin...	5083.960	5	0.2
9	40936	CA-116638	2012-01-28	2012-01-01	Second Class	JH-15985	Joseph Holt	Consumer	Concord	North Carolina	FUR-TA-10000198	Furniture	Tables	Chromcraft Bull-Nose Wood Oval Conference Tabl...	4297.644	13	0.4
10	34577	CA-102988	2011-04-05	2011-04-04	Second Class	GM-14695	Greg Maxwell	Corporate	Alexandria	Virginia	OFF-SU-10002881	Office Supplies	Supplies	Martin Yale Chadless Opener Electric Letter Op...	4164.050	5	0.0

28°C Cloudy 11:50 AM 07-11-2022

global_sales_data_analysis.ipynb - Visual Studio Code

assignment TL - 4.ipynb global_sales_data_analysis.ipynb x 01-Introduction to Pandas.ipynb

C:\Users> admin> Desktop> ibm workspace> completed work doc> completed assignment data> global_sales_data_analysis.ipynb> global sale data analysis> dat.head()

+ Code + Markdown | Run All | Clear Outputs of All Cells | Restart | Variables | Outline

Python 3.10.4 64-bit

4 columns

```
dat.tail()
```

[1]: ✓ 0.2s Python

	Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name	Segment	City	State	Product ID	Category	Sub-Category	Product Name	Sales	Quantity	Discount	Profit
51285	29002	IN-62366	2014-06-19	2014-06-06	Same Day	KE-16420	Katrina Edelman	Corporate	Kure	Hiroshima	OFF-FA-10000746	Office Supplies	Fasteners	Advantus Thumb Tacks, 12 Pack	65.100	5	0.0	4.5000
51286	35398	US-102288	2014-06-20	2014-06-06	Standard Class	ZC-21910	Zuschuss Carroll	Consumer	Houston	Texas	OFF-AP-10002906	Office Supplies	Appliances	Hoover Replacement Belt for Commercial Guardsm...	0.444	1	0.8	-1.1100
51287	40470	US-155768	2013-12-02	2013-12-12	Same Day	LB-16795	Laurel Beltran	Home Office	Oxnard	California	OFF-EN-10001219	Office Supplies	Envelopes	#10- 4 1/8" x 9 1/2" Security-Tint Envelopes	22.920	3	0.0	11.2308
51288	9596	MX-140767	2012-02-18	2012-02-02	Standard Class	RB-19795	Ross Baird	Home Office	Valinhos	São Paulo	OFF-BI-10000806	Office Supplies	Binders	Acco Index Tab, Economy	13.440	2	0.0	2.4000
51289	6147	MX-134460	2012-05-22	2012-05-05	Second Class	MC-18100	Mick Crebagga	Consumer	Tipitapa	Managua	OFF-PA-10004155	Office Supplies	Paper	Eaton Computer Printout Paper, 8.5 x 11	61.380	3	0.0	1.8000

5 rows x 19 columns 28°C Cloudy 11:49 AM 07-11-2022

global_sales_data_analysis.ipynb - Visual Studio Code

assignment1 - 4.ipynbglobal_sales_data_analysis.ipynb01-Introduction to Pandas.ipynb

C:\Users\> admin\> Desktop\> ibm workspace\> completed work doc\> completed assignment data\> global_sales_data_analysis.ipynb> global sale data analysis> dat.head()

+ Code + Markdown | Run All | Clear Outputs of All Cells | Restart | Variables | Outline

Python 3.10.4 64-bit

finding duplicates

```
print(df.duplicated())
```

0.7%

0 False
8 False
9 False
10 False
16 False
...
51270 False
51276 False
51277 False
51286 False
51287 False
Length: 9994, dtype: bool

Empty markdown cell, double click or press enter to edit.

+ Code + Markdown

Removing Duplicates

```
df.drop_duplicates(inplace = True)  
print(df.to_string())
```

2.5%

Output exceeds the size limit. Open the full output data in a text editor

Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name	Segment	City	State
Country	Postal Code	Market	Region	Product ID	Category	Sub-Category			
Product Name	Sales	Quantity	Discount	Profit	Shipping Cost	Order Priority			
0	32298	CA-2012-124891	31-07-2012	31-07-2012	Same Day	RH-19495			
States	10024.0	US	East	TEC-AC-10003033	Technology	Accessories			
Over-the-Head monaural Wireless Headset System	2309.6500	7	0.00	762.1845					
8	40155	CA-2014-135909	14-10-2014	21-10-2014	Standard Class	JW-15220			
States	95823.0	US	West	OFF-BI-10003527	Office Supplies	Binders			
Plastic Comb Binding Machine with Manual Bind	5083.9600	5	0.20	1906.4850					
9	40936	CA-2012-116638	28-01-2012	31-01-2012	Second Class	JH-15985			
States	28027.0	US	South	FUR-TA-10000190	Furniture	Tables			
Bull-Nose Wood Oval Conference Tables & Bases	4297.6440	13	0.40	1862.3124					
10	34577	CA-2011-102988	05-04-2011	09-04-2011	Second Class	GM-14695			
States	22304.0	US	South	OFF-SU-10002881	Office Supplies	Supplies			
Yale Chadless Opener Electric Letter Opener	4164.0500	5	0.00	83.2810					
16	36178	CA-2014-143567	03-11-2014	06-11-2014	Second Class	TB-21175			
States	42420.0	US	South	TEC-AC-10004145	Technology	Accessories			
Logitech diNovo Edge Keyboard	2249.9100	9	0.00	517.4793					
21	31784	CA-2011-154627	29-10-2011	31-10-2011	First Class	SA-20830			
States	60610.0	US	Central	TEC-PH-10001363	Technology	Phones			
Apple iPhone 5S	2735.9520	6	0.20	341.9940					
28	37311	CA-2013-159016	11-03-2013	12-03-2013	First Class	KF-16285			
States	90008.0	US	West	TEC-PH-10002885	Technology	Phones			
Apple iPhone 5	4158.9120	8	0.20	363.9048					
32	32735	CA-2012-139731	15-10-2012	15-10-2012	Same Day	JE-15745			
States	20100.0	US	Central	FUR-CU-10002034	Furniture	Chairs			

global_sales_data_analysis.ipynb - Visual Studio Code

assignment1 - 4.ipynbglobal_sales_data_analysis.ipynb01-Introduction to Pandas.ipynb

C:\Users\> admin\> Desktop\> ibm workspace\> completed work doc\> completed assignment data\> global_sales_data_analysis.ipynb> global sale data analysis> dat.head()

+ Code + Markdown | Run All | Clear Outputs of All Cells | Restart | Variables | Outline

Python 3.10.4 64-bit

Removing Duplicates

```
df.drop_duplicates(inplace = True)  
print(df.to_string())
```

2.5%

Output exceeds the size limit. Open the full output data in a text editor

Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name	Segment	City	State
Country	Postal Code	Market	Region	Product ID	Category	Sub-Category			
Product Name	Sales	Quantity	Discount	Profit	Shipping Cost	Order Priority			
0	32298	CA-2012-124891	31-07-2012	31-07-2012	Same Day	RH-19495			
States	10024.0	US	East	TEC-AC-10003033	Technology	Accessories			
Over-the-Head monaural Wireless Headset System	2309.6500	7	0.00	762.1845					
8	40155	CA-2014-135909	14-10-2014	21-10-2014	Standard Class	JW-15220			
States	95823.0	US	West	OFF-BI-10003527	Office Supplies	Binders			
Plastic Comb Binding Machine with Manual Bind	5083.9600	5	0.20	1906.4850					
9	40936	CA-2012-116638	28-01-2012	31-01-2012	Second Class	JH-15985			
States	28027.0	US	South	FUR-TA-10000190	Furniture	Tables			
Bull-Nose Wood Oval Conference Tables & Bases	4297.6440	13	0.40	1862.3124					
10	34577	CA-2011-102988	05-04-2011	09-04-2011	Second Class	GM-14695			
States	22304.0	US	South	OFF-SU-10002881	Office Supplies	Supplies			
Yale Chadless Opener Electric Letter Opener	4164.0500	5	0.00	83.2810					
16	36178	CA-2014-143567	03-11-2014	06-11-2014	Second Class	TB-21175			
States	42420.0	US	South	TEC-AC-10004145	Technology	Accessories			
Logitech diNovo Edge Keyboard	2249.9100	9	0.00	517.4793					
21	31784	CA-2011-154627	29-10-2011	31-10-2011	First Class	SA-20830			
States	60610.0	US	Central	TEC-PH-10001363	Technology	Phones			
Apple iPhone 5S	2735.9520	6	0.20	341.9940					
28	37311	CA-2013-159016	11-03-2013	12-03-2013	First Class	KF-16285			
States	90008.0	US	West	TEC-PH-10002885	Technology	Phones			
Apple iPhone 5	4158.9120	8	0.20	363.9048					
32	32735	CA-2012-139731	15-10-2012	15-10-2012	Same Day	JE-15745			
States	20100.0	US	Central	FUR-CU-10002034	Furniture	Chairs			

global_sales_data_analysis.ipynb - Visual Studio Code

assignment 11 - 4.ipynb global_sales_data_analysis.ipynb 01-Introduction to Pandas.ipynb

C:\Users> admin> Desktop> ibm workspace> completed work doc> completed assignment data> global_sales_data_analysis.ipynb> global sale data analysis> dat.head()

+ Code + Markdown | Run All | Clear Outputs of All Cells | Restart | Variables | Outline

Python 3.10.4 64-bit

These rows had cells with empty values removed

```
df = dat
emp=df.dropna(inplace = True)
print(df.to_string)
```

[11] ✓ 0.1s Python

Output exceeds the [size limit](#). Open the full output data [in a text editor](#).

<bound method DataFrame.to_string of

			Row ID	Order ID	Order Date	Ship Date	Ship Mode	\
0	32298	CA-2012-124891	31-07-2012	31-07-2012			Same Day	
8	40155	CA-2014-135909	14-10-2014	21-10-2014			Standard Class	
9	40936	CA-2012-116638	28-01-2012	31-01-2012			Second Class	
10	34577	CA-2011-102988	05-04-2011	09-04-2011			Second Class	
16	36178	CA-2014-143567	03-11-2014	06-11-2014			Second Class	
...
51270	38414	CA-2011-143168	18-10-2011	23-10-2011			Second Class	
51276	31558	US-2014-155299	09-06-2014	13-06-2014			Standard Class	
51277	37361	CA-2012-111780	25-12-2012	30-12-2012			Second Class	
51286	35398	US-2014-102288	20-06-2014	24-06-2014			Standard Class	
51287	40470	US-2013-155768	02-12-2013	02-12-2013			Same Day	

	Customer ID	Customer Name	Segment	City	\
0	RH-19495	Rick Hansen	Consumer	New York City	
8	JW-15220	Jane Waco	Corporate	Sacramento	
9	JH-15985	Joseph Holt	Consumer	Concord	
10	GM-14695	Greg Maxwell	Corporate	Alexandria	
16	TB-21175	Thomas Boland	Corporate	Henderson	
...
51270	IG-15085	Ivan Gibson	Consumer	Seattle	

28°C Cloudy 11:50 AM 07-11-2022