Batch: A3	**Roll No.: 16014022050**
Experiment 02	

Title: Dataset pre-processing

Objective:

1. **To learn how to prepare the dataset**
2. **To learn various steps in Data -Preprocessing**

Course Outcome:

CO1: Learn how to locate and download datasets, extract insights from that data and present their findings in a variety of different formats.

Books/ Journals/ Websites referred:

https://www.tableau.com/
https://www.kaggle.com/datasets

Resources used:

Google and you tube videos (specifically to find mean, using filters, deletion duplicate data and to find outfliers)

Theory (About Data Preprocessing):

Data preprocessing is a very crucial step in data analysis. In this process, the raw data is cleaned and transformed into a format. This format is used for data analysis and can be used by computers

Advantages:

- It helps in eliminating errors.
- It helps in handling the missing values.
- It helps in removing duplicates.
- It helps in standardizing formats.

Following points should be written by students:

Data processing is collecting and manipulating data into the usable and desired form. The data can be manipulated manually or automatically, depending on the predefined sequence of operations.

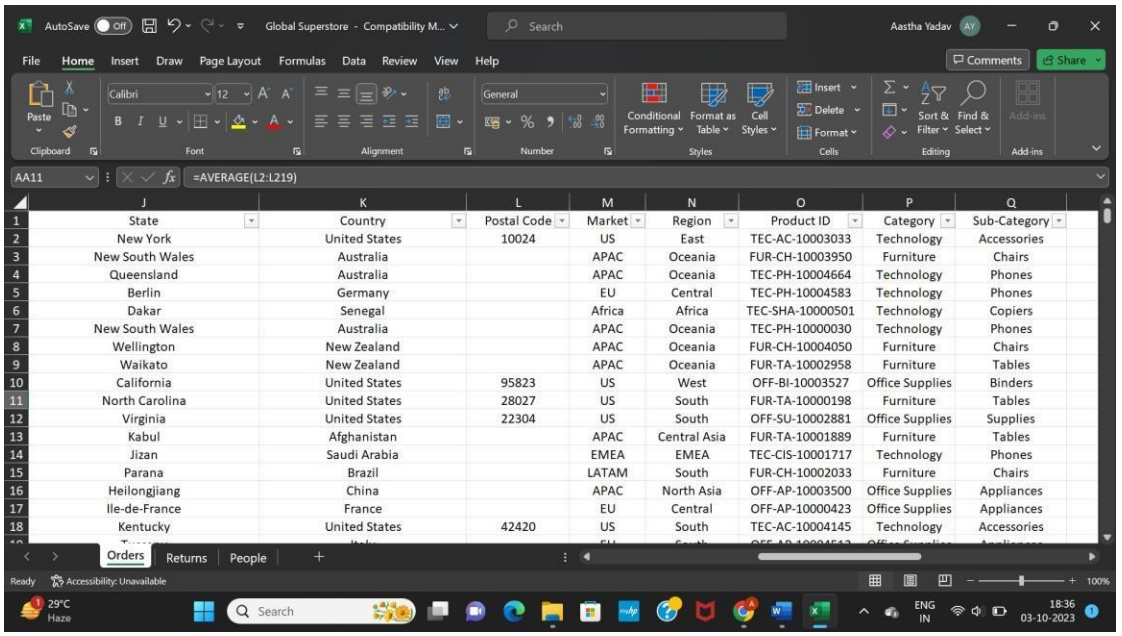
Different steps in Data Preprocessing:

- Finding missing, null values
- Replacing missing, null values with statistical parameters
- Encoding categorical data
- Normalization

Note: Student can use any technology like Tableau, Tableau-Prep, PowerBI, Google spreadsheet, excel, R programming, Python, Java any other technology for preprocessing.

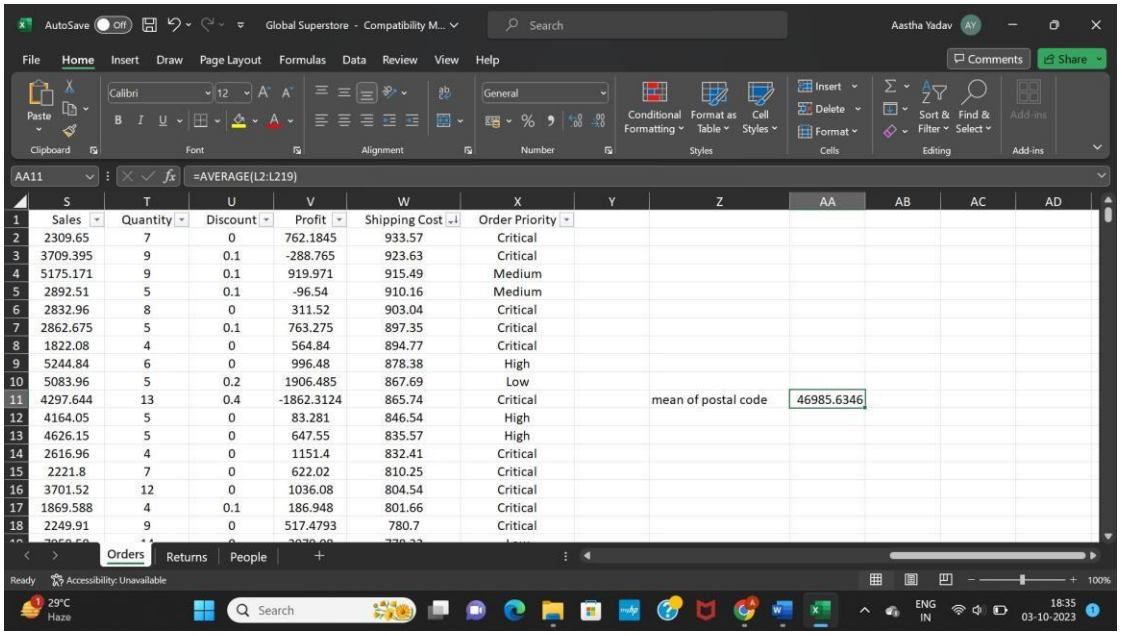
Platform used by the student:

Working (Paste the code and Output for each Data Preprocessing task):



	J	K	L	M	N	O	P	Q
1	State	Country	Postal Code	Market	Region	Product ID	Category	Sub-Category
2	New York	United States	10024	US	East	TEC-AC-10003033	Technology	Accessories
3	New South Wales	Australia		APAC	Oceania	FUR-CH-10003950	Furniture	Chairs
4	Queensland	Australia		APAC	Oceania	TEC-PH-10004664	Technology	Phones
5	Berlin	Germany		EU	Central	TEC-PH-10004583	Technology	Phones
6	Dakar	Senegal		Africa	Africa	TEC-SHA-10000501	Technology	Copiers
7	New South Wales	Australia		APAC	Oceania	TEC-PH-10000030	Technology	Phones
8	Wellington	New Zealand		APAC	Oceania	FUR-CH-10004050	Furniture	Chairs
9	Waikato	New Zealand		APAC	Oceania	FUR-TA-10002958	Furniture	Tables
10	California	United States	95823	US	West	OFF-BI-10003527	Office Supplies	Binders
11	North Carolina	United States	28027	US	South	FUR-TA-10000198	Furniture	Tables
12	Virginia	United States	22304	US	South	OFF-SU-10002881	Office Supplies	Supplies
13	Kabul	Afghanistan		APAC	Central Asia	FUR-TA-10001889	Furniture	Tables
14	Jizan	Saudi Arabia		EMEA	EMEA	TEC-CIS-10001717	Technology	Phones
15	Parana	Brazil		LATAM	South	FUR-CH-10002033	Furniture	Chairs
16	Heilongjiang	China		APAC	North Asia	OFF-AP-10003500	Office Supplies	Appliances
17	Ile-de-France	France		EU	Central	OFF-AP-10000423	Office Supplies	Appliances
18	Kentucky	United States	42420	US	South	TEC-AC-10004145	Technology	Accessories

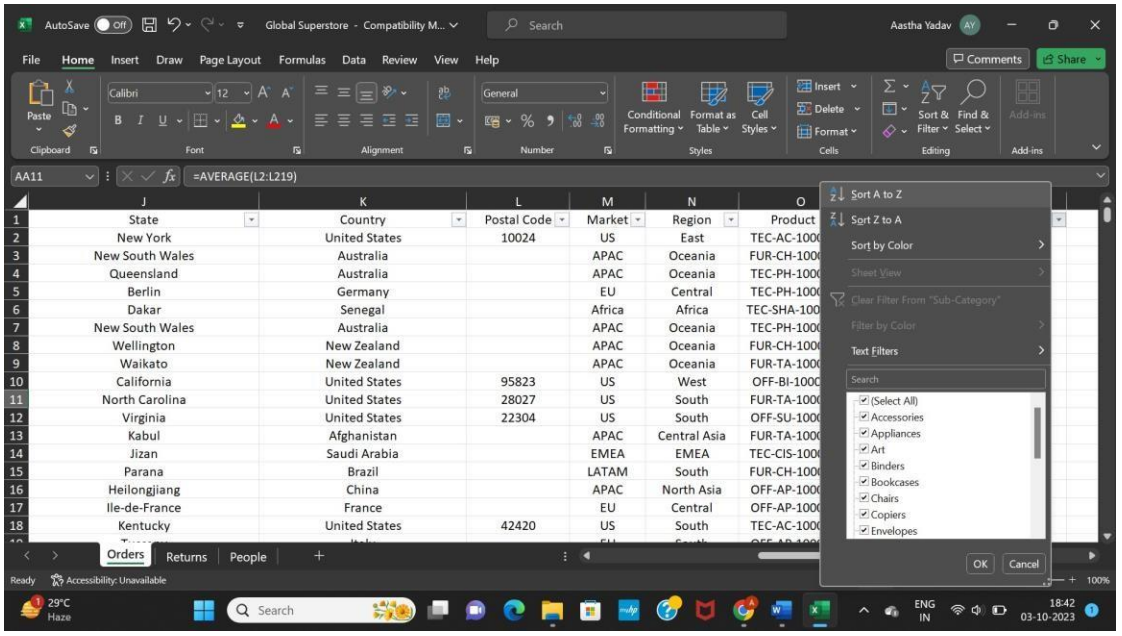
This screenshot shows empty spaces in dataset.



	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD
1	Sales	Quantity	Discount	Profit	Shipping Cost	Order Priority						
2	2309.65	7	0	762.1845	933.57	Critical						
3	3709.395	9	0.1	-288.765	923.63	Critical						
4	5175.171	9	0.1	919.971	915.49	Medium						
5	2892.51	5	0.1	-96.54	910.16	Medium						
6	2832.96	8	0	311.52	903.04	Critical						
7	2862.675	5	0.1	763.275	897.35	Critical						
8	1822.08	4	0	564.84	894.77	Critical						
9	5244.84	6	0	996.48	878.38	High						
10	5083.96	5	0.2	1906.485	867.69	Low						
11	4297.644	13	0.4	-1862.3124	865.74	Critical		mean of postal code	46985.6346			
12	4164.05	5	0	83.281	846.54	High						
13	4626.15	5	0	647.55	835.57	High						
14	2616.96	4	0	1151.4	832.41	Critical						
15	2221.8	7	0	622.02	810.25	Critical						
16	3701.52	12	0	1036.08	804.54	Critical						
17	1869.588	4	0.1	186.948	801.66	Critical						
18	2249.91	9	0	517.4793	780.7	Critical						

Now we can fill the empty boxes in postal column according to the average of postal column.

Use of filters:



	J	K	L	M	N	O
1	State	Country	Postal Code	Market	Region	Product
2	New York	United States	10024	US	East	TEC-AC-1000
3	New South Wales	Australia		APAC	Oceania	FUR-CH-1000
4	Queensland	Australia		APAC	Oceania	TEC-PH-1000
5	Berlin	Germany		EU	Central	TEC-PH-1000
6	Dakar	Senegal		Africa	Africa	TEC-SHA-100
7	New South Wales	Australia		APAC	Oceania	TEC-PH-1000
8	Wellington	New Zealand		APAC	Oceania	FUR-CH-1000
9	Waikato	New Zealand		APAC	Oceania	FUR-TA-1000
10	California	United States	95823	US	West	OFF-BI-1000
11	North Carolina	United States	28027	US	South	FUR-TA-1000
12	Virginia	United States	22304	US	South	OFF-SU-1000
13	Kabul	Afghanistan		APAC	Central Asia	FUR-TA-1000
14	Jizan	Saudi Arabia		EMEA	EMEA	TEC-CIS-1000
15	Parana	Brazil		LATAM	South	FUR-CH-1000
16	Heilongjiang	China		APAC	North Asia	OFF-AP-1000
17	Ile-de-France	France		EU	Central	OFF-AP-1000
18	Kentucky	United States	42420	US	South	TEC-AC-1000

Conclusion (Students should write in their own words):

By delving into the intricacies of dataset preparation, we acquired invaluable insights into streamlining data organization and formatting. Our exploration of the multifaceted stages of data preprocessing unveiled essential techniques for enhancing data quality and optimizing model performance.

Post Lab Question:

1. **Write the importance of Data Preprocessing**
    - **It improves accuracy and reliability.**
    - **It makes data consistent.**
    - **It increases the data's algorithm readability.**