

EDS Theory Activity no. 1

Dataset :- Grocery

Name :- Swapnil Gajmol Ingole

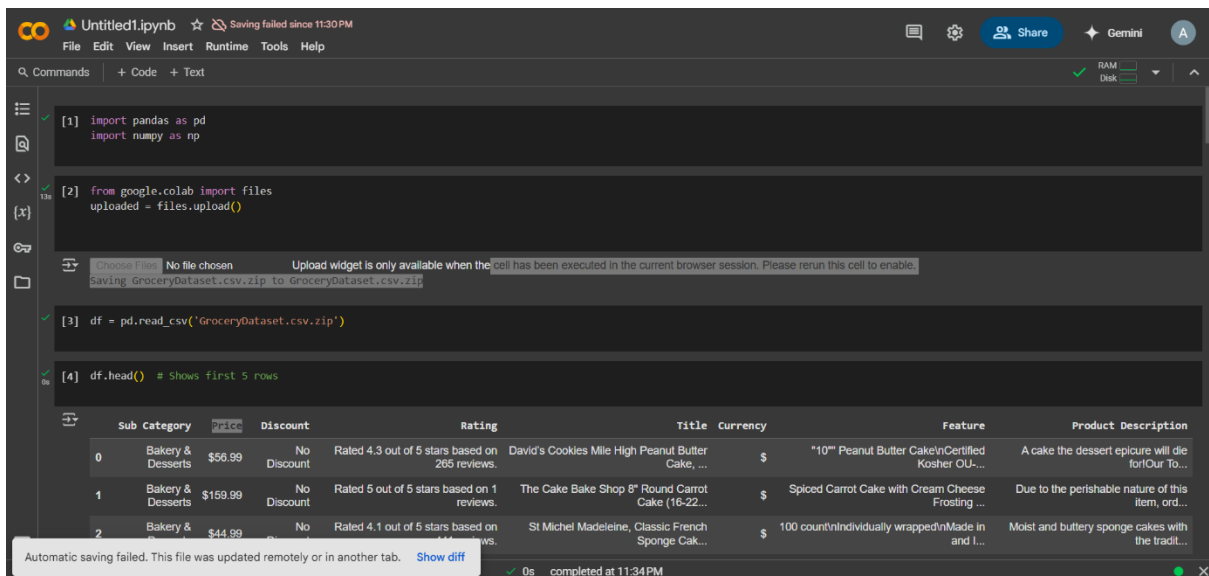
Div :- CS6

Batch :- C63

Roll No. :- CS6-66

PRN :- 202401080051

- Solution for formulated problem statement.
- Import dataset and all libraries.



The screenshot shows a Jupyter Notebook titled 'Untitled1.ipynb' with the following code cells:

```
[1] import pandas as pd
import numpy as np
```

```
[2] from google.colab import files
uploaded = files.upload()
```

A message indicates that no file was chosen for upload. The code then reads the dataset:

```
[3] df = pd.read_csv('GroceryDataset.csv.zip')
```

```
[4] df.head()
```

The output of the head() function shows the first 5 rows of the dataset:

| | Sub Category | Price | Discount | Rating | Title | Currency | Feature | Product Description |
|---|-------------------|----------|-------------|--|---|----------|--|--|
| 0 | Bakery & Desserts | \$56.99 | No Discount | Rated 4.3 out of 5 stars based on 265 reviews. | David's Cookies Mile High Peanut Butter Cake, ... | \$ | *10" Peanut Butter Cake in Certified Kosher OU... | A cake the dessert epicure will die for! Our To... |
| 1 | Bakery & Desserts | \$169.99 | No Discount | Rated 5 out of 5 stars based on 1 reviews. | The Cake Bake Shop 8" Round Carrot Cake (16-22... | \$ | Spiced Carrot Cake with Cream Cheese Frosting ... | Due to the perishable nature of this item, ord... |
| 2 | Bakery & Desserts | \$44.99 | No Discount | Rated 4.1 out of 5 stars based on 144 reviews. | St Michel Madeleine, Classic French Sponge Cak... | \$ | 100 count Individually wrapped in Made in and I... | Moist and buttery sponge cakes with the tradit... |

At the bottom, a status bar indicates 'Automatic saving failed. This file was updated remotely or in another tab. Show diff' and '0s completed at 11:34 PM'.

colab.research.google.com/drive/1y_vq1Rddc4j6BA-Hs_AMzvU5NqQjkVUM#scrollTo=IS-5uW6Mkrih

Untitled1.ipynb

File Edit View Insert Runtime Tools Help

Q Commands + Code + Text

RAM Disk

0s

[38] # 9. Products with No Discount

no_discount_products = df[df['Discount'] == 0]

print("Products with no discount:\n", no_discount_products[['Title', 'Discount']])

Products with no discount:

Empty DataFrame

Columns: [Title, Discount]

Index: []

0s

Products in the 'Electronics' category

electronics = df[df['Sub Category'] == 'Electronics']

print("Products in Electronics:\n", electronics)

Products in Electronics:

Empty DataFrame

Columns: [Sub Category, Price, Discount, Rating, Title, Currency, Feature, Product Description]

Index: []

colab.research.google.com/drive/1y_vq1Rddc4j6BA-Hs_AMzvU5NqQjkVUM#scrollTo=IS-5uW6Mkrih

Untitled1.ipynb

File Edit View Insert Runtime Tools Help

Q Commands + Code + Text

RAM Disk

1755

Fruit Roll-Ups, Variety Pack, 72-count

NaN

1756

Takis, Rolled Tortilla Chips, Intense Nacho, 1...

NaN

[1757 rows x 2 columns]

0s

[21] print("Bad/missing ratings count:", df['Rating'].isna().sum())

Bad/missing ratings count: 1757

0s

[22] print("Min rating:", df['Rating'].min())

Min rating: nan

0s

[24] print("Max rating:", df['Title'].max())

Max rating: uni-ball 207 Retractable Gel Pen, Medium Point 0.7mm, Assorted Ink Colors, 12-count

0s

[31] df['Rating'] = pd.to_numeric(df['Rating'], errors='coerce')

0s

[32] print("Unique sub categories:", df['Sub Category'].nunique())

Unique sub categories: 19

colab.research.google.com/drive/1y_vq1Rddc4j6BA-Hs_AMzvUSNqQjKVUM#scrollTo=IS-5uW6Mkrih

Untitled1.ipynb ☆ Saving failed since 11:30 PM

File Edit View Insert Runtime Tools Help

Commands + Code + Text

| 3 | Bakery & Desserts | \$39.99 | No Discount | Rated 4.7 out of 5 stars based on 9459 reviews. | David's Cookies Butter Pecan Meltaways 32 oz, ... | \$ | Butter Pecan Meltaways in 32 oz 2-Pack in No Prese... | These delectable butter pecan meltaways are th... | |
|---|-------------------|---------|-------------|---|---|----|---|--|--|
| 4 | Bakery & Desserts | \$59.99 | No Discount | Rated 4.5 out of 5 stars based on 758 reviews. | David's Cookies Premier Chocolate Cake, 7.2 lb... | \$ | "10" Four Layer Chocolate Cake in Certified Kosh... | A cake the dessert epicure will die for! To the... | |

```

[9] print(df.columns)

Index(['Sub Category', 'Price', 'Discount', 'Rating', 'Title', 'Currency',
      'feature', 'Product Description'],
      dtype='object')

print("Unique products:", df['Title'].nunique())
Unique products: 1484

[14] print("Unique sub-categories:", df['Sub Category'].nunique())
Unique sub-categories: 19

[19] # 1. Convert Rating to numeric, coercing non-numeric entries to NaN
df['Rating'] = pd.to_numeric(df['Rating'], errors='coerce')

# 2. Now you can safely compute min, max, mean, etc.
print("Minimum rating:", df['Rating'].min())
print("Maximum rating:", df['Rating'].max())

```

colab.research.google.com/drive/1y_vq1Rddc4j6BA-Hs_AMzvUSNqQjKVUM#scrollTo=IS-5uW6Mkrih

Untitled1.ipynb ☆ Saving failed since 11:30 PM

File Edit View Insert Runtime Tools Help

Commands + Code + Text

```

# 7. Average Rating per Sub Category
avg_rating_subcategory = df.groupby('Sub Category')['Rating'].mean()
print("Average rating per sub-category:\n", avg_rating_subcategory)

```

Average rating per sub-category:

| Sub Category | Rating |
|------------------------------|--------|
| Bakery & Desserts | NaN |
| Beverages & Water | NaN |
| Breakfast | NaN |
| Candy | NaN |
| Cleaning Supplies | NaN |
| Coffee | NaN |
| Deli | NaN |
| Floral | NaN |
| Gift Baskets | NaN |
| Household | NaN |
| Kirkland Signature Grocery | NaN |
| Laundry Detergent & Supplies | NaN |
| Meat & Seafood | NaN |
| Organic | NaN |
| Pantry & Dry Goods | NaN |
| Paper & Plastic Products | NaN |
| Poultry | NaN |
| Seafood | NaN |
| Snacks | NaN |

Name: Rating, dtype: float64

```
colab.research.google.com/drive/1y_vq1Rddc4j6BA-Hs_AMzvU5NqQjkVUM#scrollTo=IS-5uW6Mkrih

Untitled1.ipynb ☆ Saving failed since 11:30 PM
File Edit View Insert Runtime Tools Help

[19] dt['Rating'] = pd.to_numeric(df['Rating'], errors='coerce')
# 2. Now you can safely compute min, max, mean, etc.
print("Minimum rating:", df['Rating'].min())
print("Maximum rating:", df['Rating'].max())
print("Average rating:", df['Rating'].mean())

Minimum rating: nan
Maximum rating: nan
Average rating: nan

[20] bad_ratings = df[pd.isna(df['Rating'])]
print("Rows with non-numeric ratings:\n", bad_ratings[['Title', 'Rating']])

Rows with non-numeric ratings:
   Title Rating
0  David's Cookies Mile High Peanut Butter Cake, ...  NaN
1  The Cake Bake Shop 8" Round Carrot Cake (16-22...  NaN
2  St Michel Madeleine, Classic French Sponge Cak...  NaN
3  David's Cookies Butter Pecan Meltaways 32 oz, ...  NaN
4  David's Cookies Premier Chocolate Cake, 7.2 lb...  NaN
...  ...
1752 Oberto Thin Style Smoked Sausage Stick, Cockta...  NaN
1753 Cheetos Crunchy, Original, 2.1 oz, 64-count  NaN
1754 Sabritas Chile & Limon Mix, Variety Pack, 30-c...  NaN
1755 Fruit Roll-Ups, Variety Pack, 72-count  NaN
1756 Takis, Rolled Tortilla Chips, Intense Nacho, 1...  NaN

[1757 rows x 2 columns]
```

```
colab.research.google.com/drive/1y_vq1Rddc4j6BA-Hs_AMzvU5NqQjkVUM#scrollTo=IS-5uW6Mkrih

Untitled1.ipynb ☆ Saving failed since 11:30 PM
File Edit View Insert Runtime Tools Help

[32] print("Unique sub categories:", df['Sub Category'].nunique())

Unique sub categories: 19

[35] # 5. Top 5 Products by Discount
top_5_discount = df.sort_values(by='Discount', ascending=False).head(5)
print("Top 5 products by discount:\n", top_5_discount[['Title', 'Discount']])

Top 5 products by discount:
   Title \
39  Kirkland Signature Bottled Water 16.9 fl oz, 4...
0  David's Cookies Mile High Peanut Butter Cake, ...
1159 Namaste USDA Organic Gluten Free Perfect Sweet...
1157 Kirkland Signature, Organic Extra Virgin Olive...
1156 Kirkland Signature, Organic Chicken Stock, 32 ...

Discount
39  This item is not returnable.
0  No Discount
1159 No Discount
1157 No Discount
1156 No Discount
```