Data Collection and Preprocessing Phase

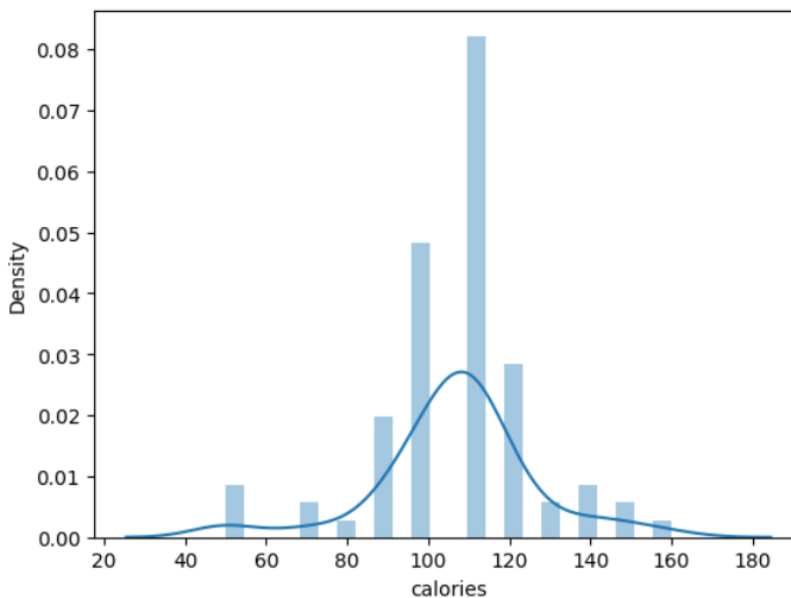|	
Date	7 July 2024
Team ID	SWTID1720104754
Project Title	Cereal Analysis Based On Rating By Using Machine Learning Techniques
Maximum Marks	6 Marks

Data Exploration and Preprocessing Template

Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.
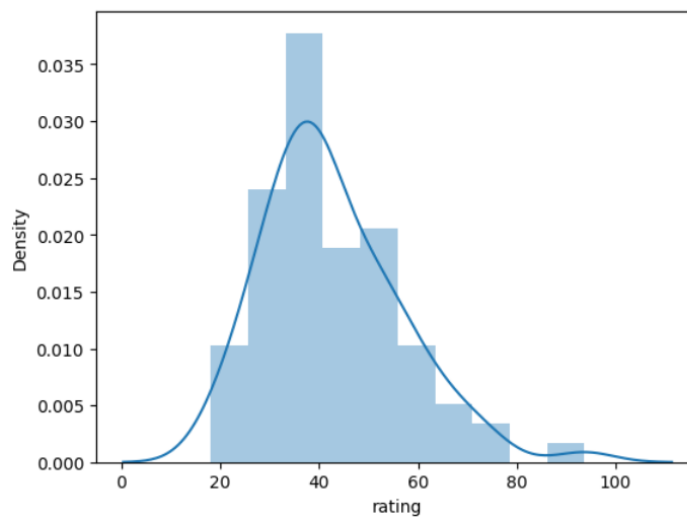
Section	Description
Data Overview	Dimension: 77 rows × 16 columns Descriptive statistics:

Descriptive statistics:

	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	shelf	weight	cups	rating
count	77.000000	77.000000	77.000000	77.000000	77.000000	77.000000	77.000000	77.000000	77.000000	77.000000	77.000000	77.000000	77.000000
mean	106.883117	2.545455	1.012987	159.675325	2.151948	14.597403	6.922078	96.077922	28.246753	2.207792	1.029610	0.821039	42.665705
std	19.484119	1.094790	1.006473	83.832295	2.383364	4.278956	4.444885	71.286813	22.342523	0.832524	0.150477	0.232716	14.047289
min	50.000000	1.000000	0.000000	0.000000	0.000000	-1.000000	-1.000000	-1.000000	0.000000	1.000000	0.500000	0.250000	18.042851
25%	100.000000	2.000000	0.000000	130.000000	1.000000	12.000000	3.000000	40.000000	25.000000	1.000000	1.000000	0.670000	33.174094
50%	110.000000	3.000000	1.000000	180.000000	2.000000	14.000000	7.000000	90.000000	25.000000	2.000000	1.000000	0.750000	40.400208
75%	110.000000	3.000000	2.000000	210.000000	3.000000	17.000000	11.000000	120.000000	25.000000	3.000000	1.000000	1.000000	50.828392
max	160.000000	6.000000	5.000000	320.000000	14.000000	23.000000	15.000000	330.000000	100.000000	3.000000	1.500000	1.500000	93.704912

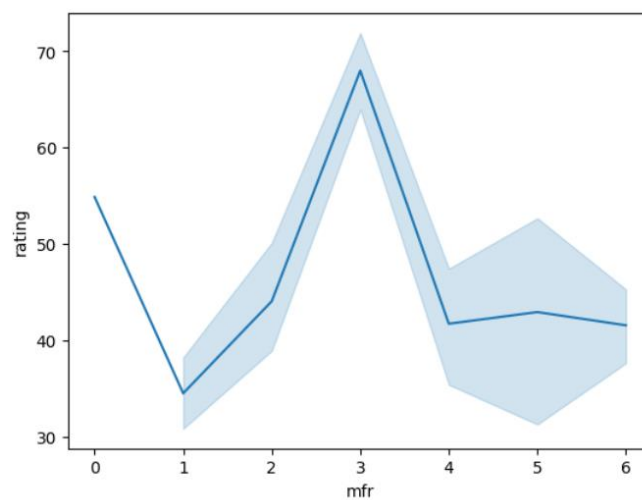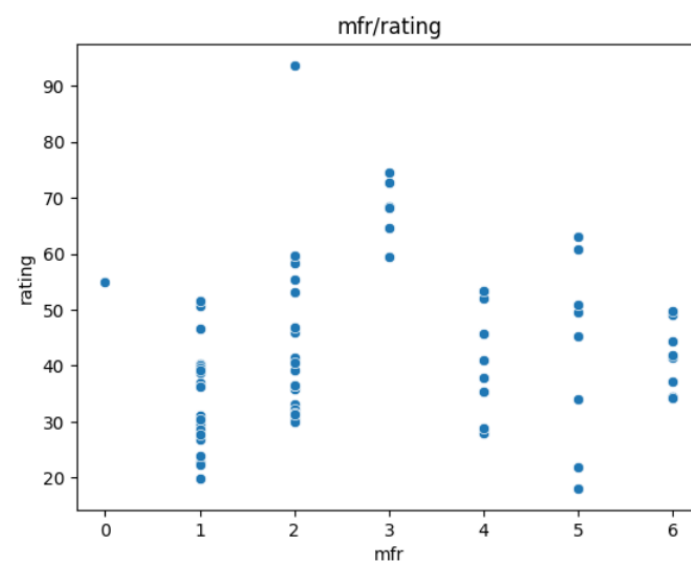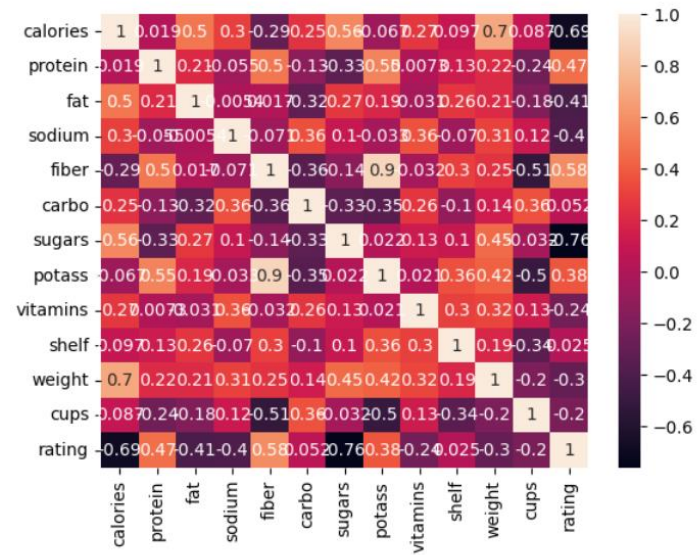| Univariate Analysis | <Axes: xlabel='calories', ylabel='Density'> <Axes: xlabel='rating', ylabel='Density'> |

Bivariate Analysis

```
<Axes: xlabel='mfr', ylabel='rating'>
```



```
Text(0.5, 1.0, 'mfr/rating')
```


.

Multivariate Analysis	 `<Axes: >`
Outliers and Anomalies	 `<Axes: xlabel='mfr', ylabel='rating'>`
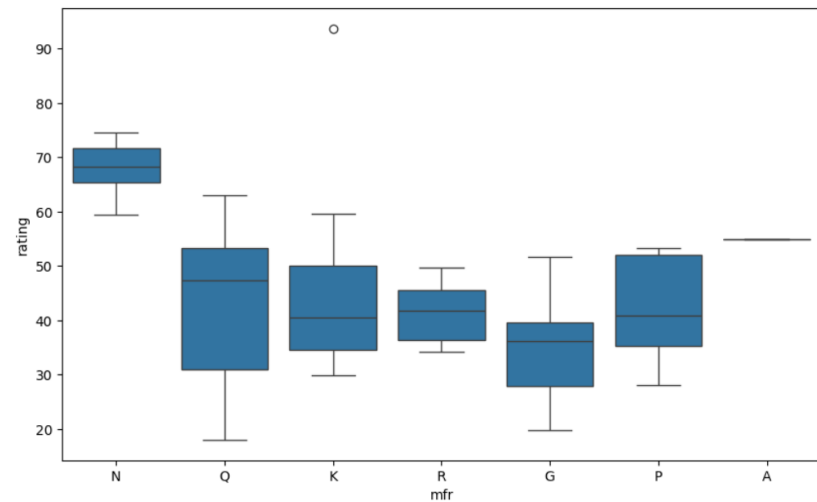
Data Preprocessing Code Screenshots

Loading Data	```python
data=pd.read_csv('cereal.csv')
```
data

| | name | mfr | type | calories | protein | fat | sodium | fiber | carbo | sugars | potass | vitamins | shelf | weight | cups | rating |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 100% Bran | N | C | 70 | 4 | 1 | 130 | 10.0 | 5.0 | 6 | 280 | 25 | 3 | 1.0 | 0.33 | 68.402973 |
| 1 | 100% Natural Bran | Q | C | 120 | 3 | 5 | 15 | 2.0 | 8.0 | 8 | 135 | 0 | 3 | 1.0 | 1.00 | 33.983679 |
| 2 | All-Bran | K | C | 70 | 4 | 1 | 260 | 9.0 | 7.0 | 5 | 320 | 25 | 3 | 1.0 | 0.33 | 59.425505 |
| 3 | All-Bran with Extra Fiber | K | C | 50 | 4 | 0 | 140 | 14.0 | 8.0 | 0 | 330 | 25 | 3 | 1.0 | 0.50 | 93.704912 |
| 4 | Almond Delight | R | C | 110 | 2 | 2 | 200 | 1.0 | 14.0 | 8 | -1 | 25 | 3 | 1.0 | 0.75 | 34.384843 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 72 | Triples | G | C | 110 | 2 | 1 | 250 | 0.0 | 21.0 | 3 | 60 | 25 | 3 | 1.0 | 0.75 | 39.106174 |
| 73 | Trix | G | C | 110 | 1 | 1 | 140 | 0.0 | 13.0 | 12 | 25 | 25 | 2 | 1.0 | 1.00 | 27.753301 |
| 74 | Wheat Chex | R | C | 100 | 3 | 1 | 230 | 3.0 | 17.0 | 3 | 115 | 25 | 1 | 1.0 | 0.67 | 49.787445 |
| 75 | Wheaties | G | C | 100 | 3 | 1 | 200 | 3.0 | 17.0 | 3 | 110 | 25 | 1 | 1.0 | 1.00 | 51.592193 |
| 76 | Wheaties Honey Gold | G | C | 110 | 2 | 1 | 200 | 1.0 | 16.0 | 8 | 60 | 25 | 1 | 1.0 | 0.75 | 36.187559 |

77 rows × 16 columns |
| **Handling Missing Data** | ▬▬▬▬▬▬ |
| **Data Transformation** | ```python
le = LabelEncoder()
data['mfr'] = le.fit_transform(data['mfr'])
data['type'] = le.fit_transform(data['type'])
```
data

	name	mfr	type	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	shelf	weight	cups	rating
0	100% Bran	3	0	70	4	1	130	10.0	5.0	6	280	25	3	1.0	0.33	68.402973
1	100% Natural Bran	5	0	120	3	5	15	2.0	8.0	8	135	0	3	1.0	1.00	33.983679
2	All-Bran	2	0	70	4	1	260	9.0	7.0	5	320	25	3	1.0	0.33	59.425505
3	All-Bran with Extra Fiber	2	0	50	4	0	140	14.0	8.0	0	330	25	3	1.0	0.50	93.704912
4	Almond Delight	6	0	110	2	2	200	1.0	14.0	8	-1	25	3	1.0	0.75	34.384843
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
72	Triples	1	0	110	2	1	250	0.0	21.0	3	60	25	3	1.0	0.75	39.106174
73	Trix	1	0	110	1	1	140	0.0	13.0	12	25	25	2	1.0	1.00	27.753301
74	Wheat Chex	6	0	100	3	1	230	3.0	17.0	3	115	25	1	1.0	0.67	49.787445
75	Wheaties	1	0	100	3	1	200	3.0	17.0	3	110	25	1	1.0	1.00	51.592193
76	Wheaties Honey Gold	1	0	110	2	1	200	1.0	16.0	8	60	25	1	1.0	0.75	36.187559	
Feature Engineering	▬▬▬▬▬▬▬
| **Save Processed Data** | ```python
import pickle
pickle.dump(lr,open('cerealanalysis.pkl','wb'))
model=pickle.load(open('cerealanalysis.pkl','rb'))
``` |