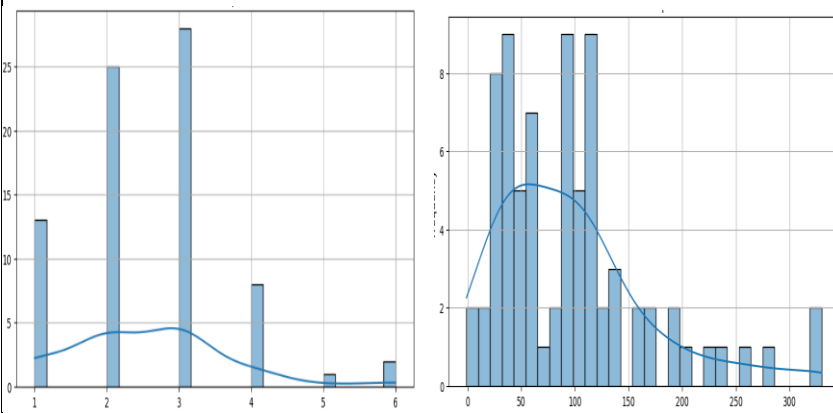


## Data Collection and Preprocessing Phase

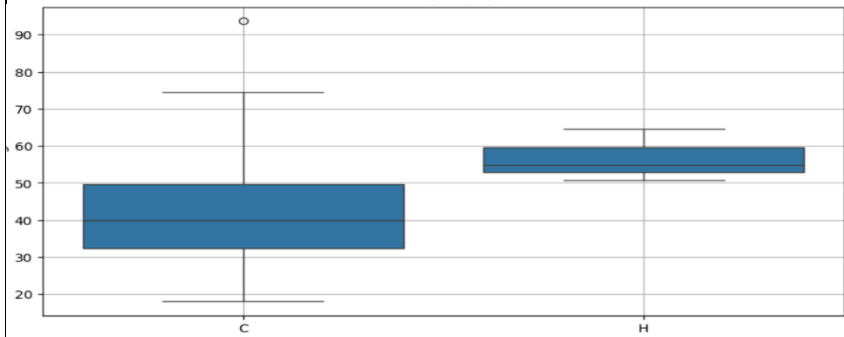
Date	15 March 2024
Team ID	LTVIP2024TMID24997
Project Title	Cereal Analysis Based on Ratings by using Machine Learning Techniques
Maximum Marks	6 Marks

## Data Exploration and Preprocessing Template

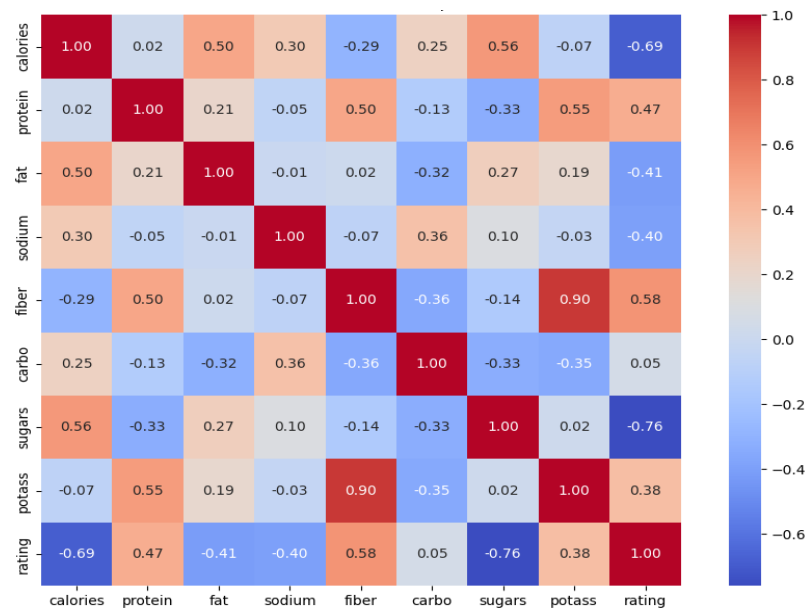
Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

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

## Bivariate Analysis



## Multivariate Analysis



## Outliers and Anomalies

—

## Data Preprocessing Code Screenshots

### Loading Data

```
# Load the dataset into the preferred environment
data = pd.read_csv('cereal.csv')
```

### Handling Missing Data

```
# Check for missing values
print(data.isnull().sum())

# Handle missing values (example: fill with mean)
data.fillna(data.mean(), inplace=True)
```

Data Transformation	<pre> from sklearn.preprocessing import StandardScaler  # Normalize numerical features scaler = StandardScaler() numerical_features = data.select_dtypes(include=['int64', 'float64']).columns data[numerical_features] = scaler.fit_transform(data[numerical_features]) </pre>
Feature Engineering	<pre> # One-hot encode categorical variables data = pd.get_dummies(data, columns=['mfr', 'type'], drop_first=True) </pre>
Save Processed Data	<pre> # Save the cleaned and processed data for future use data.to_csv('processed_cereal.csv', index=False) </pre>