# Data Collection and Preprocessing Phase

| | |
|---|---|
| Date | 23 September 2024 |
| Team ID | LTVIP2024TMID24997 |
| Project Title | Cereal Analysis Based on Ratings by using Machine Learning Techniques |
| Maximum Marks | 6 Marks |

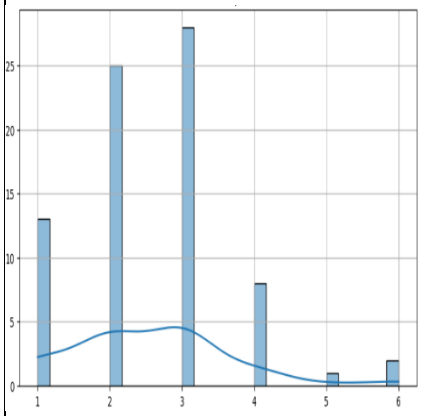
## Data Exploration and Preprocessing Template

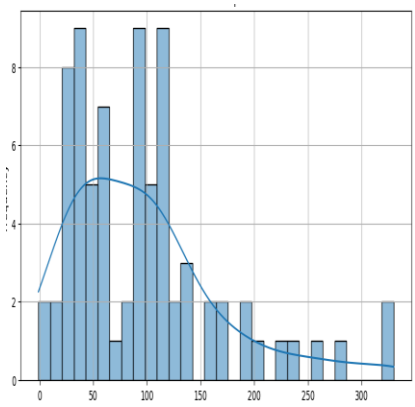
Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

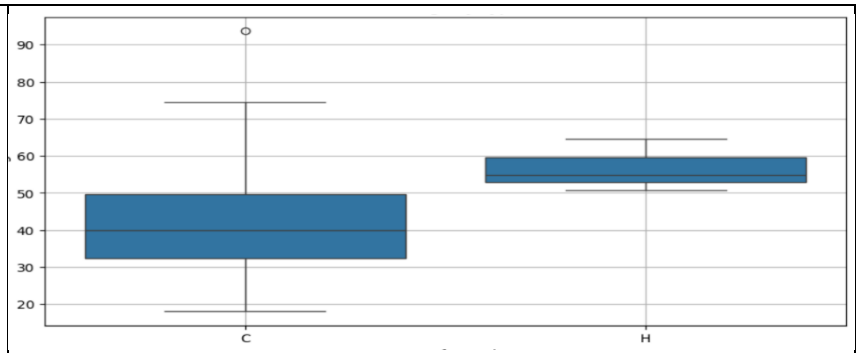
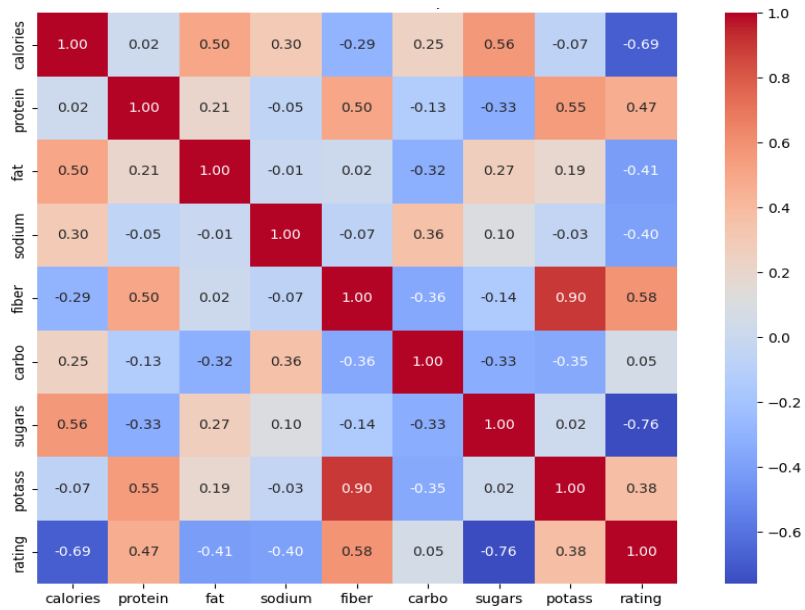
| Section | Description |
|---|---|
| Data Overview | |

| | 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



| | |
|---|---|
| Bivariate Analysis |  |
| Multivariate Analysis |  |
| Outliers and Anomalies | — |
| **Data Preprocessing Code Screenshots** | |
| Loading Data | ```python<br># Load the dataset into the preferred environment<br>data = pd.read_csv('cereal.csv')<br>```  |

| | |
|---|---|
| Handling Missing Data | ```python<br># Check for missing values<br>print(data.isnull().sum())<br><br># Handle missing values (example: fill with mean)<br>data.fillna(data.mean(), inplace=True)<br>```  |

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