Data Collection and Preprocessing Phase

Date	20 June 2024
Team ID	739986
Project Title	Customer Acquisition cost estimation using Machine Learning.
Maximum Marks	6 Marks

Data Exploration and Preprocessing Report

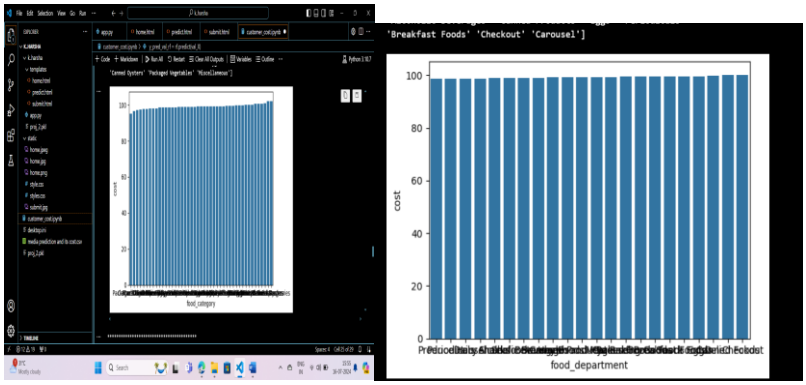
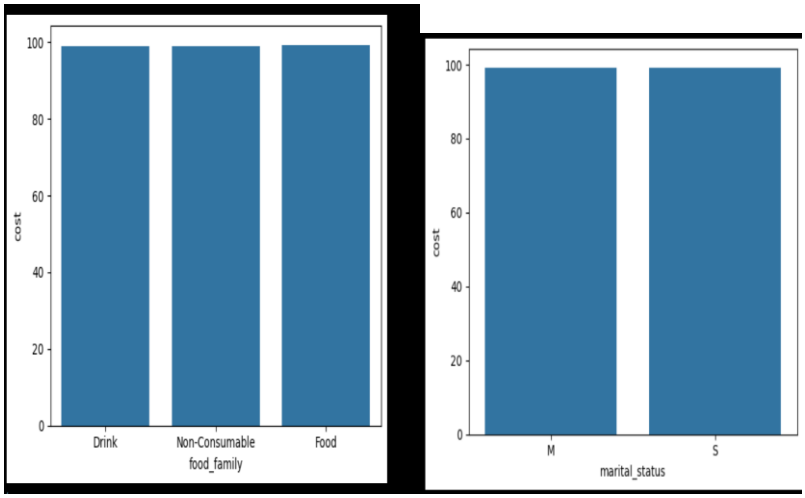
Dataset variables will be statistically analysed to identify patterns and outliers, with Python employed for preprocessing tasks like normalization and feature engineering. Data cleaning will address missing values and outliers, ensuring quality for subsequent analysis and modeling, and forming a strong foundation for insights and predictions.

Section	Description
Data Overview	

	store_sales(in millions)	store_cost(in millions)	unit_sales(in millions)	total_children	avg_cars_at home(approx)	num_children_at_home	avg_cars_at home(approx).1
count	60428.000000	60428.000000	60428.000000	60428.000000	60428.000000	60428.000000	60428.000000
mean	6.541031	2.619460	3.093169	2.533875	2.200271	0.829351	2.200271
std	3.463047	1.453009	0.827677	1.490165	1.109644	1.303424	1.109644
min	0.510000	0.163200	1.000000	0.000000	0.000000	0.000000	0.000000
25%	3.810000	1.500000	3.000000	1.000000	1.000000	0.000000	1.000000
50%	5.940000	2.385600	3.000000	3.000000	2.000000	0.000000	2.000000
75%	8.670000	3.484025	4.000000	4.000000	3.000000	1.000000	3.000000
max	22.920000	9.726500	6.000000	5.000000	4.000000	5.000000	4.000000

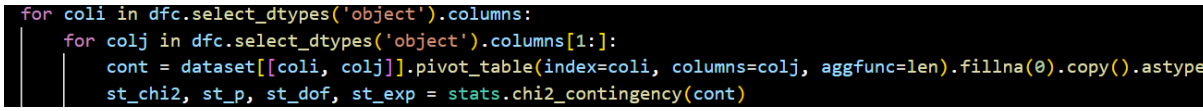
8 rows × 23 columns

SRP	gross_weight	net_weight	...	store_sqft	grocery_sqft	frozen_sqft	meat_sqft	coffee_bar	video_store
3.000000	60428.000000	60428.000000	...	60428.000000	60428.000000	60428.000000	60428.000000	60428.000000	60428.000000
2.115258	13.806433	11.796289	...	27988.477494	19133.799696	5312.852552	3541.846280	0.612646	0.354157
0.932829	4.622693	4.682986	...	5701.022090	3987.395735	1575.907263	1050.471635	0.487150	0.478261
0.500000	6.000000	3.050000	...	20319.000000	13305.000000	2452.000000	1635.000000	0.000000	0.000000
1.410000	9.700000	7.710000	...	23593.000000	16232.000000	4746.000000	3164.000000	0.000000	0.000000
2.130000	13.600000	11.600000	...	27694.000000	18670.000000	5062.000000	3375.000000	1.000000	0.000000
2.790000	17.700000	16.000000	...	30797.000000	22123.000000	5751.000000	3834.000000	1.000000	1.000000
3.980000	21.900000	20.800000	...	39696.000000	30351.000000	9184.000000	6122.000000	1.000000	1.000000

Univariate Analysis	
Bivariate Analysis	

Outliers and Anomalies	-

Data Preprocessing Code Screenshots

Loading Data	
Handling Missing Data	
Data Transformation	----------
Feature Engineering	Attached the codes in final submission.
Save Processed Data	-

```
dataset=pd.read_csv("media prediction and its cost.csv")
dataset.head()
```

Pytho

	food_category	food_department	food_family	store_sales(in millions)	store_cost(in millions)	unit_sales(in millions)	promotion_name	sales_country
0	Breakfast Foods	Frozen Foods	Food	7.36	2.7232	4.0	Bag Stuffers	USA
1	Breakfast Foods	Frozen Foods	Food	5.52	2.5944	3.0	Cash Register Lottery	USA
2	Breakfast Foods	Frozen Foods	Food	3.68	1.3616	2.0	High Roller Savings	USA
3	Breakfast Foods	Frozen Foods	Food	3.68	1.1776	2.0	Cash Register Lottery	USA
4	Breakfast Foods	Frozen Foods	Food	4.08	1.4280	3.0	Double Down Sale	USA

5 rows × 40 columns

```
for coli in dfc.select_dtypes('object').columns:
    for colj in dfc.select_dtypes('object').columns[1:]:
        cont = dataset[[coli, colj]].pivot_table(index=coli, columns=colj, aggfunc=len).fillna(0).copy().astype
        st_chi2, st_p, st_dof, st_exp = stats.chi2_contingency(cont)
```