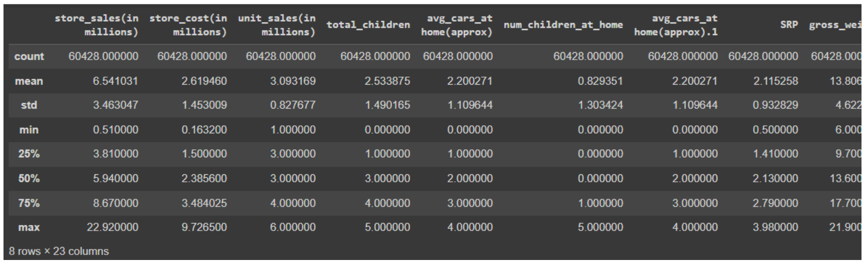
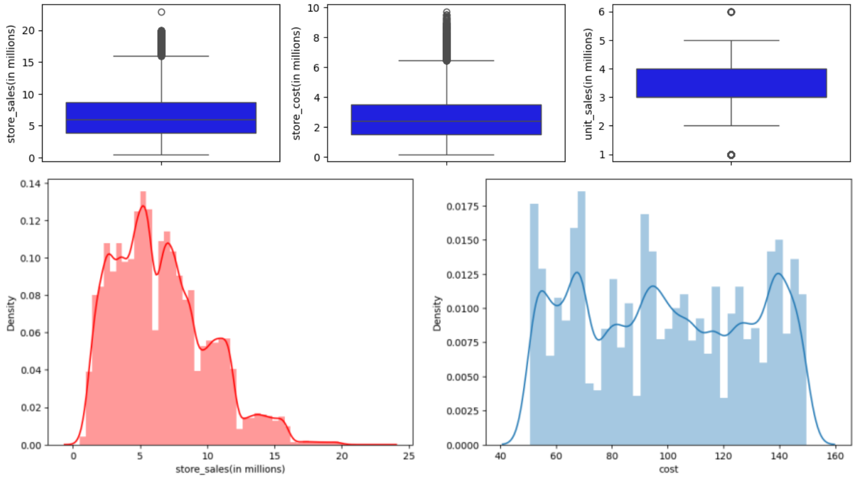
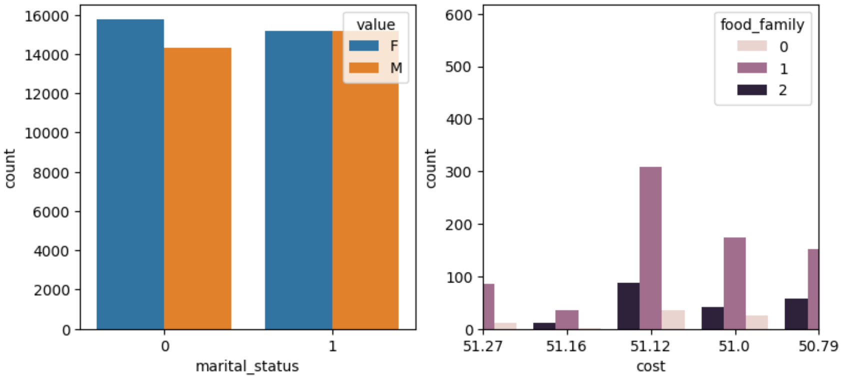
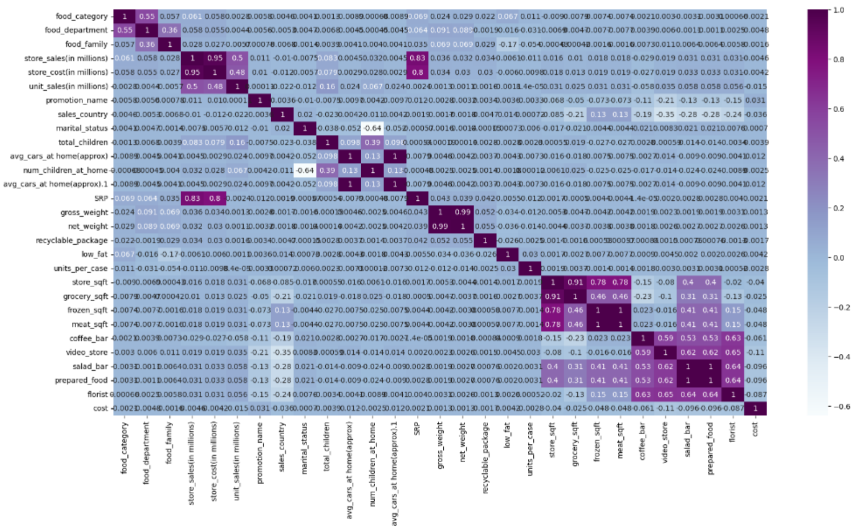
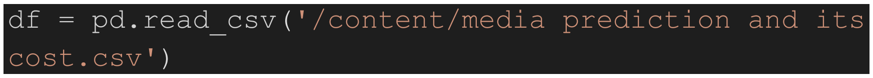
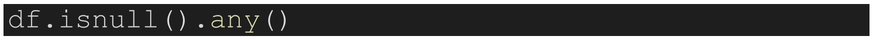
Data Collection and Preprocessing Phase

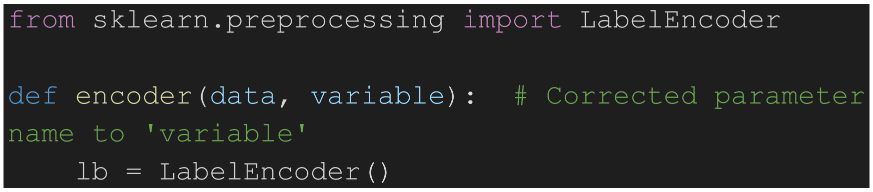
Date	4th July 2024
Team ID	739804
Project Title	Cost Prediction of Acquiring a Customer
Maximum Marks	6 Marks

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.

<table>
<tr><th>Section</th><th>Description</th></tr>
<tr><td>Data Overview</td><td>Dimensions: (60428, 40)</td></tr>
<tr><td>Univariate Analysis</td><td>
<table>
<tr><th></th><th>store_sales(in millions)</th><th>store_cost(in millions)</th><th>unit_sales(in millions)</th><th>total_children</th><th>avg_cars_at home(approx)</th><th>num_children_at_home</th><th>avg_cars_at home(approx).1</th><th>SRP</th><th>gross_wei</th></tr>
<tr><td>count</td><td>60428.000000</td><td>60428.000000</td><td>60428.000000</td><td>60428.000000</td><td>60428.000000</td><td>60428.000000</td><td>60428.000000</td><td>60428.000000</td><td>60428.000</td></tr>
<tr><td>mean</td><td>6.541031</td><td>2.619460</td><td>3.093169</td><td>2.533875</td><td>2.200271</td><td>0.829351</td><td>2.200271</td><td>2.115258</td><td>13.806</td></tr>
<tr><td>std</td><td>3.463047</td><td>1.453009</td><td>0.827677</td><td>1.490165</td><td>1.109644</td><td>1.303424</td><td>1.109644</td><td>0.932829</td><td>4.622</td></tr>
<tr><td>min</td><td>0.510000</td><td>0.163200</td><td>1.000000</td><td>0.000000</td><td>0.000000</td><td>0.000000</td><td>0.000000</td><td>0.500000</td><td>6.000</td></tr>
<tr><td>25%</td><td>3.810000</td><td>1.500000</td><td>3.000000</td><td>1.000000</td><td>1.000000</td><td>0.000000</td><td>1.000000</td><td>1.410000</td><td>9.700</td></tr>
<tr><td>50%</td><td>5.940000</td><td>2.385600</td><td>3.000000</td><td>3.000000</td><td>2.000000</td><td>0.000000</td><td>2.000000</td><td>2.130000</td><td>13.600</td></tr>
<tr><td>75%</td><td>8.670000</td><td>3.484025</td><td>4.000000</td><td>4.000000</td><td>3.000000</td><td>1.000000</td><td>3.000000</td><td>2.790000</td><td>17.700</td></tr>
<tr><td>max</td><td>22.920000</td><td>9.726500</td><td>6.000000</td><td>5.000000</td><td>4.000000</td><td>5.000000</td><td>4.000000</td><td>3.980000</td><td>21.900</td></tr>
</table>
8 rows × 23 columns</td></tr>
<tr><td>Bivariate Analysis</td><td>
</td></tr>
</table>

	
Multivariate Analysis	
Outliers and Anomalies	-
Data Preprocessing Code Screenshots	
Loading Data	 df = pd.read_csv('/content/media prediction and its cost.csv')
Handling Missing Data	 df.isnull().any()

Data Transformation	
	from sklearn.preprocessing import LabelEncoder
	
	def encoder(data, variable): # Corrected parameter name to 'variable'
	    lb = LabelEncoder()

	df[variable] = lb.fit_transform(df[variable])
	return lb

Feature Engineering	food_category_le = encoder(df,'food_category')
	brand_name_le = encoder(df,'brand_name')
	food_department_le = encoder(df,'food_department')
	food_family_le = encoder(df,'food_family')
	promotion_name_le = encoder(df,'promotion_name')
	store_city_le = encoder(df,'store_city')
	#unit_per_case_le = encoder(df,'unit_per_case')
	net_weight_le = encoder(df,'net_weight')
	sales_le = encoder(df,'sales_country')
	martial_le = encoder(df,'marital_status')

Save Processed Data	import pickle
	pickle.dump(rf,open('customers.pkl','wb'))
	pickle.dump(food_category_le,open('food_category_le.pkl','wb'))
	pickle.dump(brand_name_le,open('brand_name_le.pkl','wb'))
	pickle.dump(promotion_name_le,open('promotion_name_le.pkl','wb'))
	pickle.dump(store_city_le,open('store_city_le.pkl','wb'))