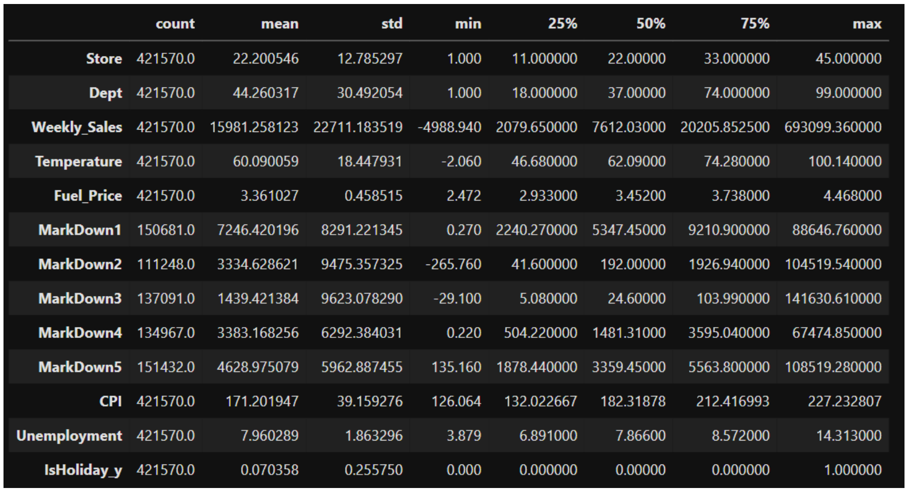
Data Collection and Preprocessing Phase

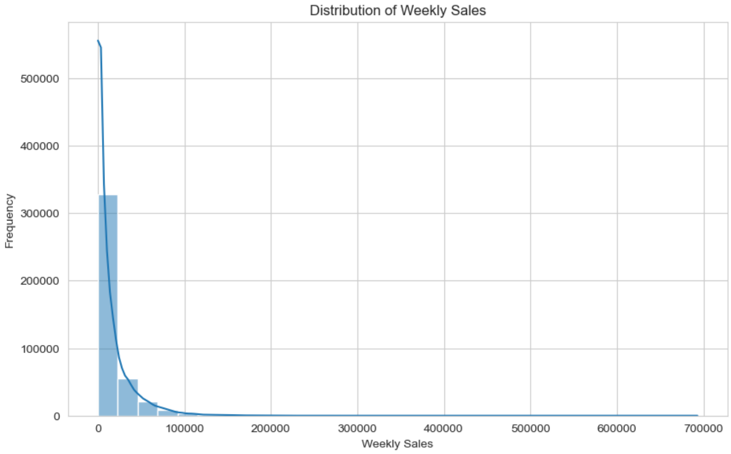
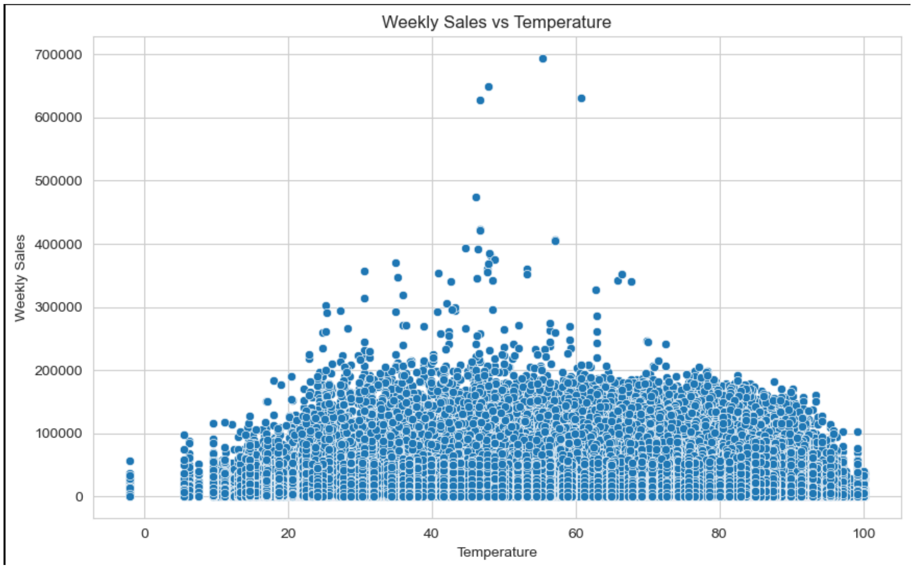
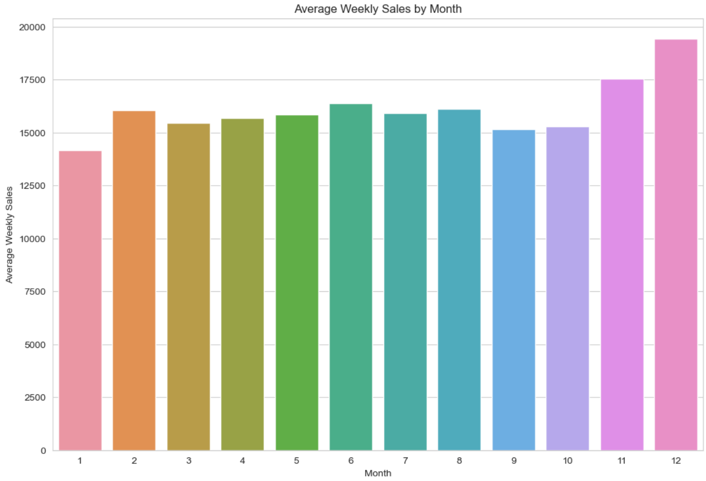
Date	9th July 2024
Team ID	SWTID1720435231
Project Title	Walmart Sales Analysis For Retail Industry With Machine Learning
Maximum Marks	6 Marks

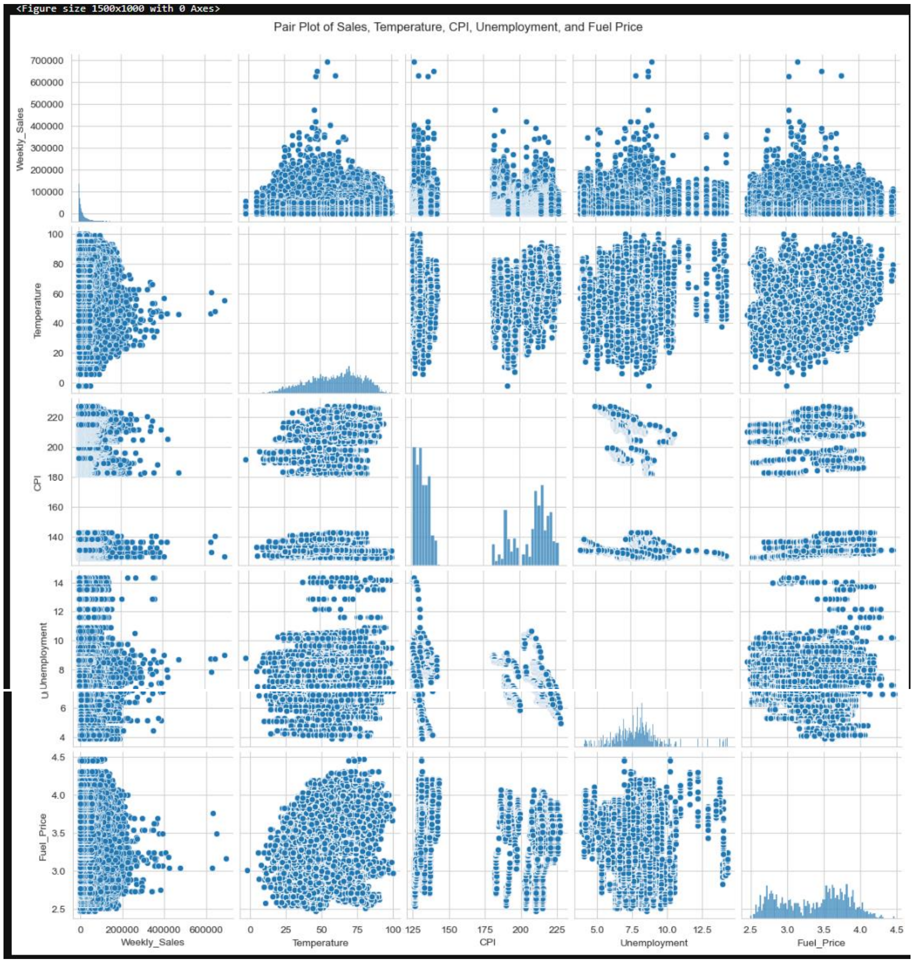
Data Exploration and Preprocessing Template

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

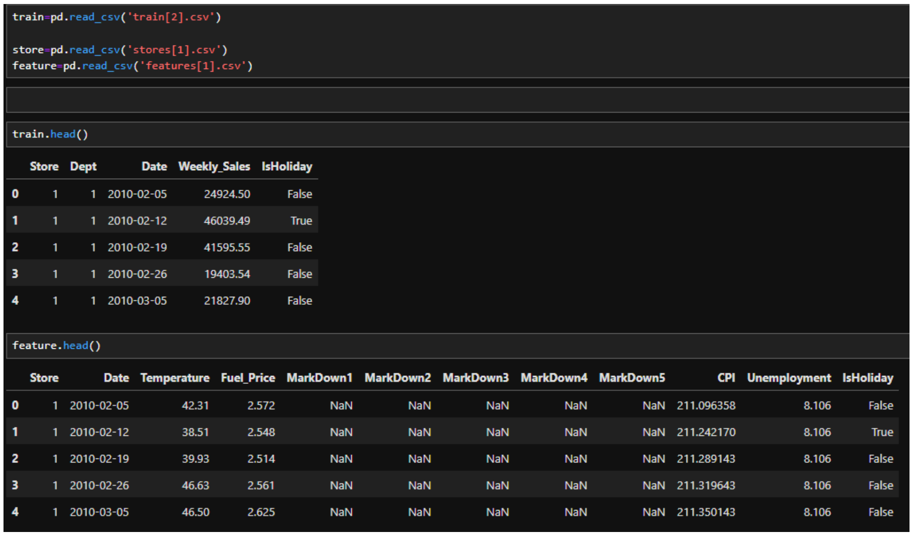
Section	**Description**
Data Overview	Dimensions: 421570 rows * 15 columns Descriptive statistics: 

	count	mean	std	min	25%	50%	75%	max
Store	421570.0	22.200546	12.785297	1.000	11.000000	22.00000	33.000000	45.000000
Dept	421570.0	44.260317	30.492054	1.000	18.000000	37.00000	74.000000	99.000000
Weekly_Sales	421570.0	15981.258123	22711.183519	-4988.940	2079.650000	7612.03000	20205.852500	693099.360000
Temperature	421570.0	60.090059	18.447931	-2.060	46.680000	62.09000	74.280000	100.140000
Fuel_Price	421570.0	3.361027	0.458515	2.472	2.933000	3.45200	3.738000	4.468000
MarkDown1	150681.0	7246.420196	8291.221345	0.270	2240.270000	5347.45000	9210.900000	88646.760000
MarkDown2	111248.0	3334.628621	9475.357325	-265.760	41.600000	192.00000	1926.940000	104519.540000
MarkDown3	137091.0	1439.421384	9623.078290	-29.100	5.080000	24.60000	103.990000	141630.610000
MarkDown4	134967.0	3383.168256	6292.384031	0.220	504.220000	1481.31000	3595.040000	67474.850000
MarkDown5	151432.0	4628.975079	5962.887455	135.160	1878.440000	3359.45000	5563.800000	108519.280000
CPI	421570.0	171.201947	39.159276	126.064	132.022667	182.31878	212.416993	227.232807
Unemployment	421570.0	7.960289	1.863296	3.879	6.891000	7.86600	8.572000	14.313000
IsHoliday_y	421570.0	0.070358	0.255750	0.000	0.000000	0.00000	0.000000	1.000000

Univariate Analysis	
Bivariate Analysis	 

Multivariate Analysis	 `<Figure size 1500x1000 with 0 Axes>`
Outliers and Anomalies	-
Data Preprocessing Code Screenshots	

Loading Data



```
train=pd.read_csv('train[2].csv')

store=pd.read_csv('stores[1].csv')
feature=pd.read_csv('features[1].csv')
```

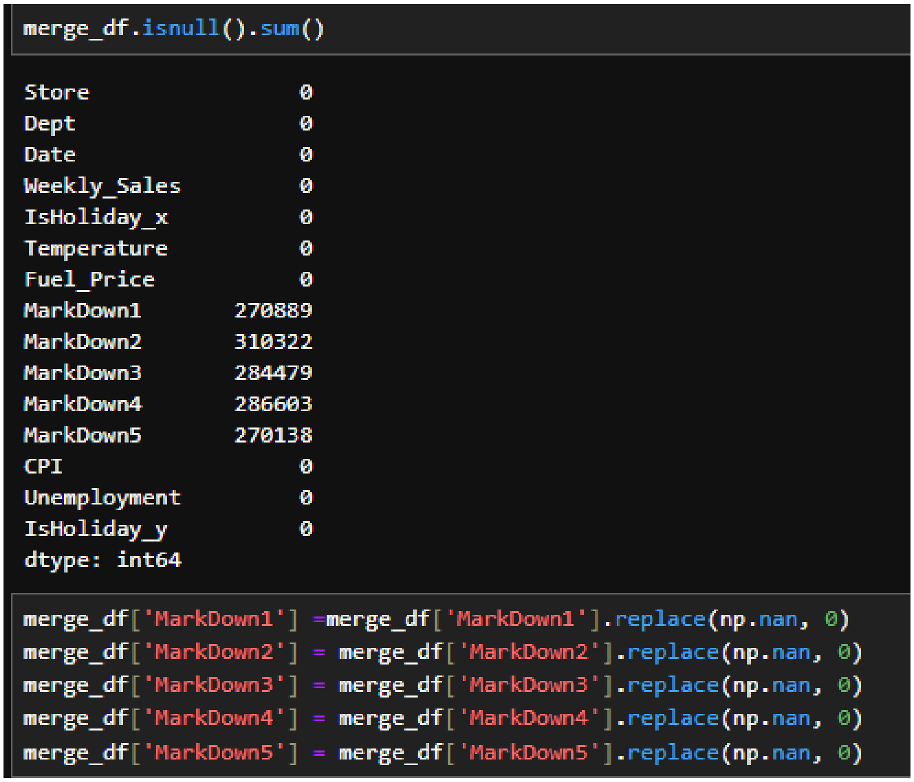
```
train.head()
```

	Store	Dept	Date	Weekly_Sales	IsHoliday
0	1	1	2010-02-05	24924.50	False
1	1	1	2010-02-12	46039.49	True
2	1	1	2010-02-19	41595.55	False
3	1	1	2010-02-26	19403.54	False
4	1	1	2010-03-05	21827.90	False

```
feature.head()
```

	Store	Date	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	IsHoliday
0	1	2010-02-05	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358	8.106	False
1	1	2010-02-12	38.51	2.548	NaN	NaN	NaN	NaN	NaN	211.242170	8.106	True
2	1	2010-02-19	39.93	2.514	NaN	NaN	NaN	NaN	NaN	211.289143	8.106	False
3	1	2010-02-26	46.63	2.561	NaN	NaN	NaN	NaN	NaN	211.319643	8.106	False
4	1	2010-03-05	46.50	2.625	NaN	NaN	NaN	NaN	NaN	211.350143	8.106	False

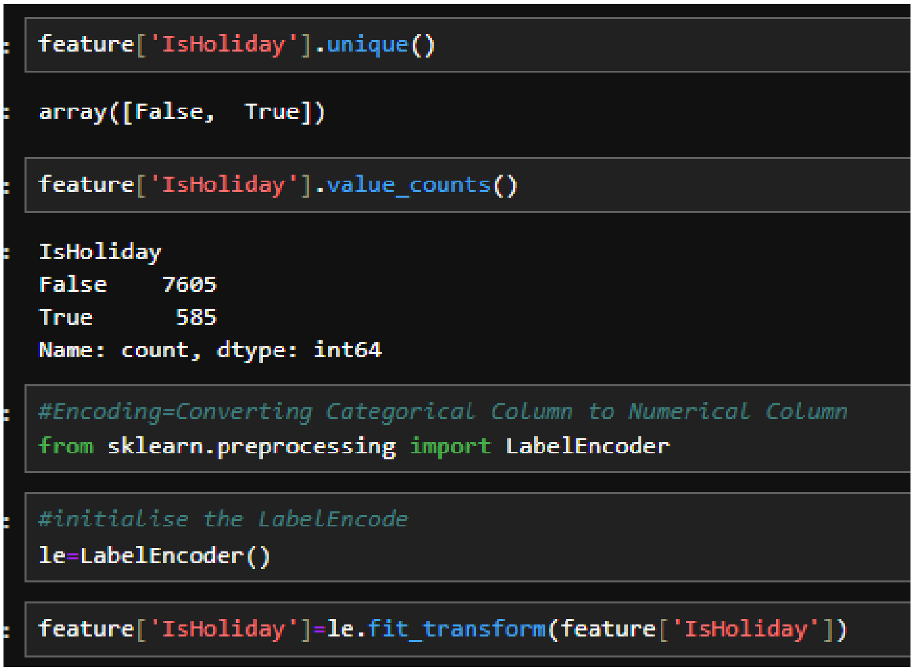
Handling Missing Data



```
merge_df.isnull().sum()
```

```
Store                0
Dept                 0
Date                 0
Weekly_Sales         0
IsHoliday_x          0
Temperature          0
Fuel_Price           0
MarkDown1       270889
MarkDown2       310322
MarkDown3       284479
MarkDown4       286603
MarkDown5       270138
CPI                  0
Unemployment         0
IsHoliday_y          0
dtype: int64
```

```
merge_df['MarkDown1'] =merge_df['MarkDown1'].replace(np.nan, 0)
merge_df['MarkDown2'] = merge_df['MarkDown2'].replace(np.nan, 0)
merge_df['MarkDown3'] = merge_df['MarkDown3'].replace(np.nan, 0)
merge_df['MarkDown4'] = merge_df['MarkDown4'].replace(np.nan, 0)
merge_df['MarkDown5'] = merge_df['MarkDown5'].replace(np.nan, 0)
```

Data Transformation	
	feature['IsHoliday'].unique()
	array([False, True])
	feature['IsHoliday'].value_counts()
	IsHoliday
	False 7605
	True 585
	Name: count, dtype: int64
	#Encoding=Converting Categorical Column to Numerical Column
	from sklearn.preprocessing import LabelEncoder
	#initialise the LabelEncode
	le=LabelEncoder()
	feature['IsHoliday']=le.fit_transform(feature['IsHoliday'])

Feature Engineering	Attached the code in the final submissions.
Save Processed Data	-