Data Collection and Preprocessing Phase

Date	19 April 2024
Team ID	738220
Project Title	Walmart Sales Analysis for Retail Industry with Machine Learning
Maximum Marks	6 Marks

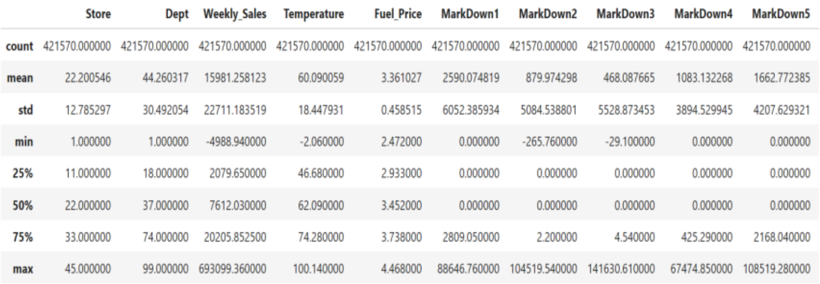
Data Exploration and Preprocessing

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>

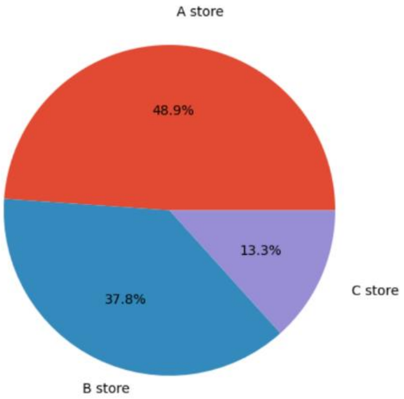
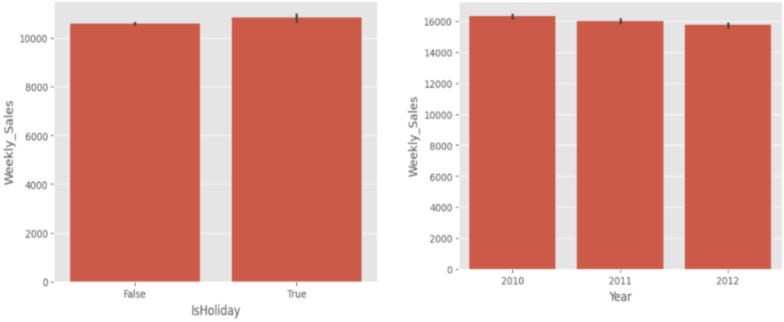
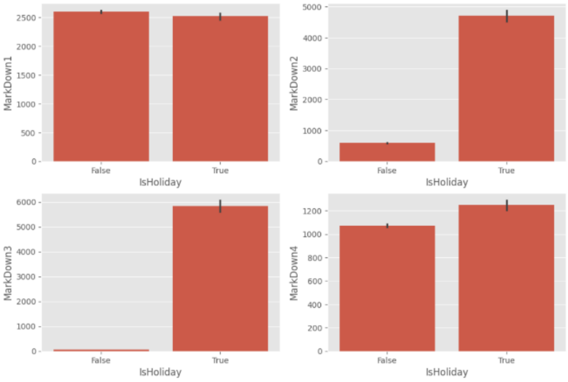
Dimension :
421570 rows × 17 columns

Descriptive Statistics:

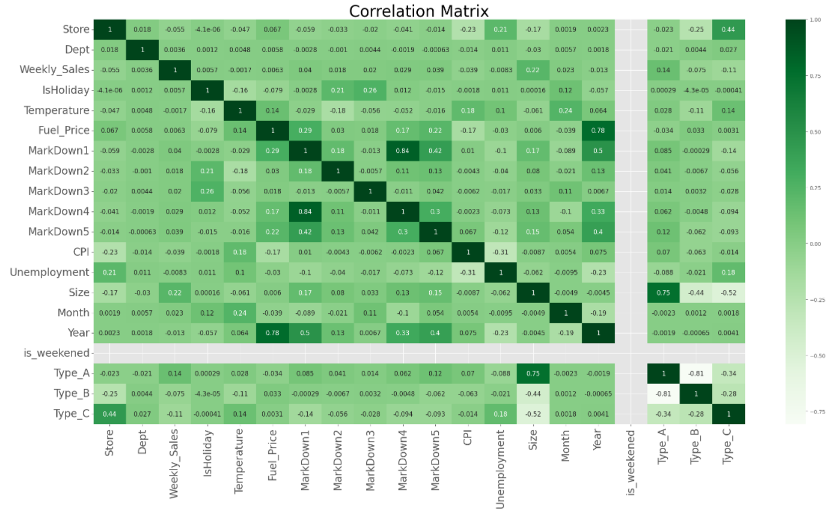


	Store	Dept	Weekly_Sales	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5
count	421570.000000	421570.000000	421570.000000	421570.000000	421570.000000	421570.000000	421570.000000	421570.000000	421570.000000	421570.000000
mean	22.200546	44.260317	15981.258123	60.090059	3.361027	2590.074819	879.974298	468.087665	1083.132268	1662.772385
std	12.785297	30.492054	22711.183519	18.447931	0.458515	6052.385934	5084.538801	5528.873453	3894.529945	4207.629321
min	1.000000	1.000000	-4988.940000	-2.060000	2.472000	0.000000	-265.760000	-29.100000	0.000000	0.000000
25%	11.000000	18.000000	2079.650000	46.680000	2.933000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	22.000000	37.000000	7612.030000	62.090000	3.452000	0.000000	0.000000	0.000000	0.000000	0.000000
75%	33.000000	74.000000	20205.852500	74.280000	3.738000	2809.050000	2.200000	4.540000	425.290000	2168.040000
max	45.000000	99.000000	693099.360000	100.140000	4.468000	88646.760000	104519.540000	141630.610000	67474.850000	108519.280000

</td></tr>
</table>

Univariate Analysis	
Bivariate Analysis	 

Multivariate Analysis



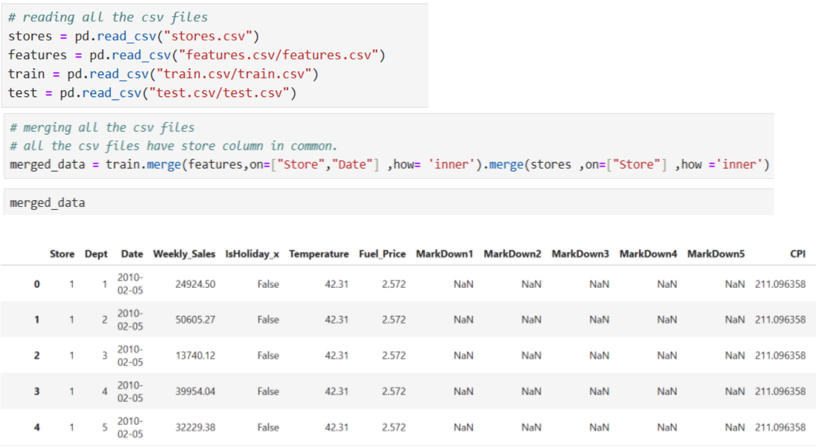
	Store	Dept	Weekly_Sales	IsHoliday	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	Size	Month	Year	is_weekened	Type_A	Type_B	Type_C
Store	1	0.018	-0.055	-4.1e-06	-0.047	0.067	-0.059	-0.033	-0.02	-0.041	-0.014	-0.23	0.21	-0.17	0.0019	0.0023		-0.023	-0.25	0.44
Dept	0.018	1	0.0036	0.0012	0.0048	0.0058	-0.0028	-0.001	0.0044	-0.0019	-0.00063	-0.014	0.011	-0.03	0.0057	0.0018		-0.021	0.0044	0.027
Weekly_Sales	-0.055	0.0036	1	0.0057	-0.0017	0.0063	0.04	0.018	0.02	0.029	0.039	-0.039	-0.0083	0.22	0.023	-0.013		0.14	-0.075	-0.11
IsHoliday	-4.1e-06	0.0012	0.0057	1	-0.16	-0.079	-0.0028	0.21	0.26	0.012	-0.015	-0.0018	0.011	0.00016	0.12	-0.057		0.00029	-4.3e-05	-0.00041
Temperature	-0.047	0.0048	-0.0017	-0.16	1	0.14	-0.029	-0.18	-0.056	-0.052	-0.016	0.18	0.1	-0.061	0.24	0.064		0.028	-0.11	0.14
Fuel_Price	0.067	0.0058	0.0063	-0.079	0.14	1	0.29	0.03	0.018	0.17	0.22	-0.17	-0.03	0.006	-0.039	0.78		-0.034	0.033	0.0031
MarkDown1	-0.059	-0.0028	0.04	-0.0028	-0.029	0.29	1	0.18	-0.013	0.84	0.42	0.01	-0.1	0.17	-0.089	0.5		0.085	-0.00029	-0.14
MarkDown2	-0.033	-0.001	0.018	0.21	-0.18	0.03	0.18	1	-0.0057	0.11	0.13	-0.0043	-0.04	0.08	-0.021	0.13		0.041	-0.0067	-0.056
MarkDown3	-0.02	0.0044	0.02	0.26	-0.056	0.018	-0.013	-0.0057	1	-0.011	0.042	-0.0062	-0.017	0.033	0.11	0.0067		0.014	0.0032	-0.028
MarkDown4	-0.041	-0.0019	0.029	0.012	-0.052	0.17	0.84	0.11	-0.011	1	0.3	-0.0023	-0.073	0.13	-0.1	0.33		0.062	-0.0048	-0.094
MarkDown5	-0.014	-0.00063	0.039	-0.015	-0.016	0.22	0.42	0.13	0.042	0.3	1	0.067	-0.12	0.15	0.054	0.4		0.12	-0.062	-0.093
CPI	-0.23	-0.014	-0.039	-0.0018	0.18	-0.17	0.01	-0.0043	-0.0062	-0.0023	0.067	1	-0.31	-0.0087	0.0054	0.075		0.07	-0.063	-0.014
Unemployment	0.21	0.011	-0.0083	0.011	0.1	-0.03	-0.1	-0.04	-0.017	-0.073	-0.12	-0.31	1	-0.062	-0.0095	-0.23		-0.088	-0.021	0.18
Size	-0.17	-0.03	0.22	0.00016	-0.061	0.006	0.17	0.08	0.033	0.13	0.15	-0.0087	-0.062	1	-0.0049	-0.0045		0.75	-0.44	-0.52
Month	0.0019	0.0057	0.023	0.12	0.24	-0.039	-0.089	-0.021	0.11	-0.1	0.054	0.0054	-0.0095	-0.0049	1	-0.19		-0.0023	0.0012	0.0018
Year	0.0023	0.0018	-0.013	-0.057	0.064	0.78	0.5	0.13	0.0067	0.33	0.4	0.075	-0.23	-0.0045	-0.19	1		-0.0019	-0.00065	0.0041
is_weekened																				
Type_A	-0.023	-0.021	0.14	0.00029	0.028	-0.034	0.085	0.041	0.014	0.062	0.12	0.07	-0.088	0.75	-0.0023	-0.0019		1	-0.81	-0.34
Type_B	-0.25	0.0044	-0.075	-4.3e-05	-0.11	0.033	-0.00029	-0.0067	0.0032	-0.0048	-0.062	-0.063	-0.021	-0.44	0.0012	-0.00065		-0.81	1	-0.28
Type_C	0.44	0.027	-0.11	-0.00041	0.14	0.0031	-0.14	-0.056	-0.028	-0.094	-0.093	-0.014	0.18	-0.52	0.0018	0.0041		-0.34	-0.28	1

Outliers and Anomalies

-

Data Preprocessing Code Screenshots

Loading Data



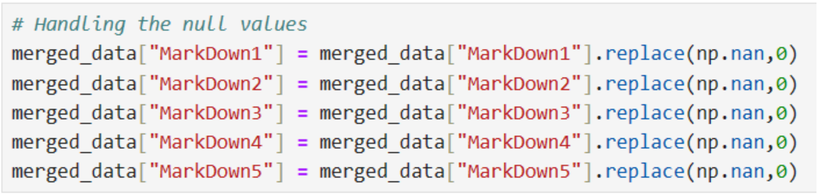
```python
# reading all the csv files
stores = pd.read_csv("stores.csv")
features = pd.read_csv("features.csv/features.csv")
train = pd.read_csv("train.csv/train.csv")
test = pd.read_csv("test.csv/test.csv")
```

```python
# merging all the csv files
# all the csv files have store column in common.
merged_data = train.merge(features,on=["Store","Date"] ,how= 'inner').merge(stores ,on=["Store"] ,how ='inner')
```

```python
merged_data
```

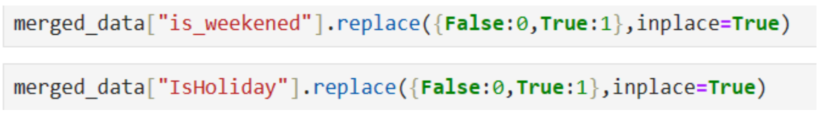
	Store	Dept	Date	Weekly_Sales	IsHoliday_x	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI
0	1	1	2010-02-05	24924.50	False	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358
1	1	2	2010-02-05	50605.27	False	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358
2	1	3	2010-02-05	13740.12	False	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358
3	1	4	2010-02-05	39954.04	False	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358
4	1	5	2010-02-05	32229.38	False	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358

Handling Missing Data



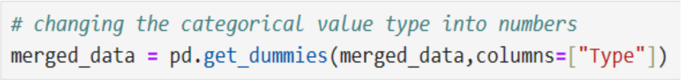
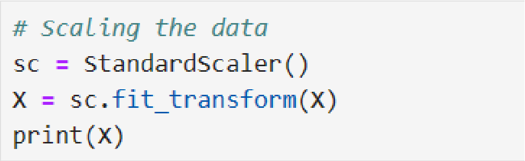
```python
# Handling the null values
merged_data["MarkDown1"] = merged_data["MarkDown1"].replace(np.nan,0)
merged_data["MarkDown2"] = merged_data["MarkDown2"].replace(np.nan,0)
merged_data["MarkDown3"] = merged_data["MarkDown3"].replace(np.nan,0)
merged_data["MarkDown4"] = merged_data["MarkDown4"].replace(np.nan,0)
merged_data["MarkDown5"] = merged_data["MarkDown5"].replace(np.nan,0)
```

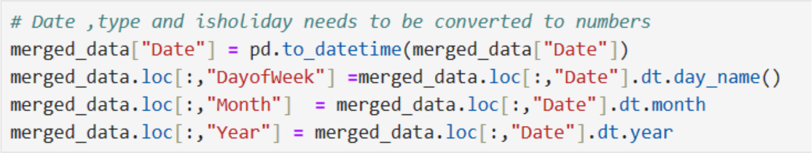
Data Transformation



```python
merged_data["is_weekened"].replace({False:0,True:1},inplace=True)
```

```python
merged_data["IsHoliday"].replace({False:0,True:1},inplace=True)
```

	
	# changing the categorical value type into numbers
	merged_data = pd.get_dummies(merged_data,columns=["Type"])
	
	# Scaling the data
	sc = StandardScaler()
	X = sc.fit_transform(X)
	print(X)

Feature Engineering	 # Date ,type and isholiday needs to be converted to numbers merged_data["Date"] = pd.to_datetime(merged_data["Date"]) merged_data.loc[:,"DayofWeek"] =merged_data.loc[:,"Date"].dt.day_name() merged_data.loc[:,"Month"] = merged_data.loc[:,"Date"].dt.month merged_data.loc[:,"Year"] = merged_data.loc[:,"Date"].dt.year
Save Processed Data	-