Thông tin nhóm

Nhóm 1

Nguyễn Hữu Kiên (Nhóm trưởng) - K224141670 Hồ Minh Nhí - K224141682 Phạm Ngọc Quang - K224141686 Ngô Nguyễn Đức Thắng - K224141694 Phan Nguyễn Xuân Toàn - K224141699

Import Dataset

In [27]:
```
import pandas as pd
data = pd.read_csv('Features_data_set.csv')
data
```

Out[27]:

	Store	Continent	Date	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	IsHoliday
0	1	Europe	5/4/2013	58.59	3.583	12872.34	5687.86	485.97	478.04	5092.33	225.086540	6.314	False
1	1	Europe	12/4/2013	62.72	3.529	3672.43	932.58	52.86	949.07	2836.64	225.170160	6.314	False
2	1	Australia	19/04/2013	67.10	3.451	3530.36	3.95	107.50	458.76	2062.05	225.170160	6.314	False
3	1	Africa	26/04/2013	59.23	3.417	2387.72	NaN	98.34	516.28	1421.63	225.170160	6.314	False
4	1	North America	4/1/2013	41.73	3.161	1214.08	25366.33	15.01	72.36	3940.02	224.080983	6.525	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...
8185	45	Europe	28/06/2013	76.05	3.639	4842.29	975.03	3.00	2449.97	3169.69	NaN	NaN	False
8186	45	Australia	5/7/2013	77.50	3.614	9090.48	2268.58	582.74	5797.47	1514.93	NaN	NaN	False
8187	45	South America	12/7/2013	79.37	3.614	3789.94	1827.31	85.72	744.84	2150.36	NaN	NaN	False
8188	45	South America	19/07/2013	82.84	3.737	2961.49	1047.07	204.19	363.00	1059.46	NaN	NaN	False
8189	45	South America	26/07/2013	76.06	3.804	212.02	851.73	2.06	10.88	1864.57	NaN	NaN	False

8190 rows × 13 columns

Filtering data to weed out the noise

In [28]:
```
data.filter(['Temperature', 'Fuel_Price']).loc[:343]
```

Out[28]:

	Temperature	Fuel_Price
0	58.59	3.583
1	62.72	3.529
2	67.10	3.451
3	59.23	3.417
4	41.73	3.161
...	...	...
339	81.81	2.705
340	86.26	2.653
341	85.81	2.637
342	83.40	2.668
343	51.26	2.732

344 rows × 2 columns

In [29]:
```
data['Temperature']
```

Out[29]:
```
0       58.59
1       62.72
2       67.10
3       59.23
4       41.73
        ...
8185    76.05
8186    77.50
8187    79.37
8188    82.84
8189    76.06
Name: Temperature, Length: 8190, dtype: float64
```

In [30]:
```
data[['Temperature']].loc[:343]
```

Out[30]:

	Temperature
0	58.59
1	62.72
2	67.10
3	59.23
4	41.73
...	...
339	81.81
340	86.26
341	85.81
342	83.40
343	51.26

344 rows × 1 columns

In [31]:
```
data.filter(['Temperature', 'Fuel_Price']).iloc[9:225]
```

Out[31]:

	Temperature	Fuel_Price
9	42.92	3.237
10	49.66	3.475
11	50.25	3.597
12	48.01	3.711
13	50.81	3.658
...	...	...
220	82.64	3.638
221	87.65	3.730
222	75.88	3.717
223	71.09	3.721
224	79.45	3.666

216 rows × 2 columns

In [32]:
```
data.filter([225,49,4,25,81,121],axis=0)
```

Out[32]:

	Store	Continent	Date	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	IsHoliday
225	2	South America	6/4/2012	68.43	3.891	12132.59	1.30	32.58	4874.69	5535.13	221.073764	6.891	False
49	1	Africa	18/05/2012	70.33	3.630	6154.14	NaN	45.11	1675.49	5508.18	221.742674	7.143	False
4	1	North America	4/1/2013	41.73	3.161	1214.08	25366.33	15.01	72.36	3940.02	224.080983	6.525	False
25	1	North America	30/11/2012	52.34	3.207	2460.03	NaN	3838.35	150.57	6966.34	223.610984	6.573	False
81	1	North America	13/05/2011	75.64	3.899	NaN	NaN	NaN	NaN	NaN	215.964053	7.682	False
121	1	Europe	24/12/2010	52.33	2.886	NaN	NaN	NaN	NaN	NaN	211.405122	7.838	False

In [33]:
```
data[4:225:4]
```

Out[33]:

	Store	Continent	Date	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	IsHoliday
4	1	North America	4/1/2013	41.73	3.161	1214.08	25366.33	15.01	72.36	3940.02	224.080983	6.525	False
8	1	Australia	25/01/2013	53.37	3.227	965.89	1097.91	0.10	225.36	1831.88	224.235552	6.525	False
12	1	North America	1/3/2013	48.01	3.711	10610.74	261.46	2.80	25.54	2747.59	224.564526	6.525	False
16	1	Asia	29/03/2013	51.00	3.606	13067.46	NaN	384.90	122.93	3903.80	225.002920	6.525	False
20	1	Australia	26/10/2012	69.16	3.506	2585.85	31.75	6.00	1057.16	1305.01	223.444251	6.573	False
24	1	Europe	23/11/2012	56.23	3.211	883.59	4.17	74910.32	209.91	303.32	223.561947	6.573	True
28	1	Australia	21/12/2012	56.02	3.098	8231.71	NaN	274.00	358.15	2834.02	223.839845	6.573	False
32	1	Asia	20/07/2012	80.42	3.311	3213.00	313.72	9.53	2262.02	3228.19	221.932727	6.908	False
36	1	South America	17/08/2012	84.85	3.571	3662.06	137.86	4.84	2752.20	3446.15	222.038411	6.908	False
40	1	North America	14/09/2012	74.97	3.717	17212.52	7.00	18.79	1523.11	7992.72	222.582019	6.908	False
44	1	South America	13/04/2012	69.07	3.891	6186.19	3288.69	17.07	1822.55	1063.78	221.510210	7.143	False
48	1	Asia	11/5/2012	73.77	3.688	8351.40	NaN	10.52	2443.14	3127.88	221.725663	7.143	False
52	1	Europe	8/6/2012	78.30	3.452	8813.81	116.80	64.55	2652.04	7161.91	221.749484	7.143	False
56	1	Africa	6/1/2012	49.01	3.157	6277.39	21813.16	143.10	1450.13	8483.00	219.714258	7.348	False
60	1	Asia	3/2/2012	56.55	3.360	34577.06	3579.21	160.53	32403.87	5630.40	220.172015	7.348	False
64	1	Africa	2/3/2012	60.96	3.630	15441.40	1569.00	10.80	25390.88	8067.61	220.848045	7.348	False
68	1	Europe	30/03/2012	67.61	3.845	10309.58	0.50	10.25	1654.17	2642.78	221.361012	7.348	False
72	1	South America	8/4/2011	67.84	3.622	NaN	NaN	NaN	NaN	NaN	215.074394	7.682	False
76	1	Asia	22/04/2011	72.99	3.807	NaN	NaN	NaN	NaN	NaN	215.459905	7.682	False
80	1	Australia	6/5/2011	64.61	3.906	NaN	NaN	NaN	NaN	NaN	215.796004	7.682	False
84	1	Asia	21/01/2011	44.04	3.016	NaN	NaN	NaN	NaN	NaN	211.827234	7.742	False
88	1	Europe	18/02/2011	57.36	3.045	NaN	NaN	NaN	NaN	NaN	213.247885	7.742	False
92	1	Asia	18/03/2011	62.76	3.488	NaN	NaN	NaN	NaN	NaN	214.362711	7.742	False
96	1	Australia	2/7/2010	80.91	2.669	NaN	NaN	NaN	NaN	NaN	211.223533	7.787	False
100	1	Africa	6/8/2010	87.16	2.627	NaN	NaN	NaN	NaN	NaN	211.504662	7.787	False
104	1	Australia	24/09/2010	80.94	2.624	NaN	NaN	NaN	NaN	NaN	211.597225	7.787	False
108	1	Africa	7/5/2010	72.55	2.835	NaN	NaN	NaN	NaN	NaN	210.339968	7.808	False
112	1	Europe	21/05/2010	76.44	2.826	NaN	NaN	NaN	NaN	NaN	210.617093	7.808	False
116	1	Australia	4/6/2010	80.69	2.705	NaN	NaN	NaN	NaN	NaN	211.176428	7.808	False
120	1	South America	31/12/2010	48.43	2.943	NaN	NaN	NaN	NaN	NaN	211.404932	7.838	True
124	1	Europe	3/12/2010	49.27	2.708	NaN	NaN	NaN	NaN	NaN	211.607193	7.838	False
128	1	South America	15/10/2010	67.18	2.720	NaN	NaN	NaN	NaN	NaN	211.813744	7.838	False
132	1	Asia	5/11/2010	58.74	2.689	NaN	NaN	NaN	NaN	NaN	211.956394	7.838	False
136	1	Europe	21/10/2011	63.71	3.353	NaN	NaN	NaN	NaN	NaN	217.515976	7.866	False
140	1	South America	18/11/2011	62.25	3.308	6074.12	254.39	51.98	427.39	5988.57	218.220509	7.866	False
144	1	North America	16/12/2011	51.63	3.159	5011.32	67.00	347.37	225.79	4011.37	219.179453	7.866	False
148	1	Africa	8/7/2011	85.83	3.480	NaN	NaN	NaN	NaN	NaN	215.277175	7.962	False
152	1	Africa	5/8/2011	91.65	3.684	NaN	NaN	NaN	NaN	NaN	215.544618	7.962	False
156	1	Asia	2/9/2011	87.83	3.533	NaN	NaN	NaN	NaN	NaN	215.797141	7.962	False
160	1	South America	30/09/2011	79.69	3.355	NaN	NaN	NaN	NaN	NaN	216.710597	7.962	False
164	1	Africa	12/2/2010	38.51	2.548	NaN	NaN	NaN	NaN	NaN	211.242170	8.106	True
168	1	North America	12/3/2010	57.79	2.667	NaN	NaN	NaN	NaN	NaN	211.380643	8.106	False
172	1	Africa	24/05/2013	77.19	3.494	7959.89	178.00	1621.47	3152.57	2938.70	NaN	NaN	False
176	1	Europe	21/06/2013	81.35	3.479	8104.02	417.99	327.33	5182.25	3754.44	NaN	NaN	False
180	1	North America	19/07/2013	79.26	3.556	3117.04	1060.39	199.05	1012.30	5381.72	NaN	NaN	False
184	2	Europe	19/04/2013	67.05	3.451	6858.52	12.72	104.72	219.88	2640.86	224.802531	6.112	False
188	2	North America	19/10/2012	68.08	3.594	4461.89	NaN	1.14	1579.67	2642.29	223.059808	6.170	False
192	2	North America	16/11/2012	52.72	3.252	26925.38	285.03	101.44	2014.28	8732.45	223.146903	6.170	False

	Store	Continent	Date	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	IsHoliday
196	2	Asia	14/12/2012	47.69	3.168	5550.46	NaN	50.00	871.41	2419.54	223.352397	6.170	False
200	2	North America	11/1/2013	47.42	3.243	6722.95	12098.17	13.68	1067.36	5383.20	223.834266	6.237	False
204	2	North America	8/2/2013	56.08	3.417	63622.34	687.56	358.29	44824.98	8167.67	223.869385	6.237	True
208	2	Asia	8/3/2013	51.12	3.658	24134.43	54.43	136.70	7235.69	6441.67	224.342767	6.237	False
212	2	Australia	6/7/2012	84.20	3.227	12355.50	295.05	100.15	6720.40	5506.53	221.521506	6.565	False
216	2	Europe	3/8/2012	90.22	3.417	27650.68	164.58	43.02	21801.90	6652.98	221.586980	6.565	False
220	2	Asia	31/08/2012	82.64	3.638	17500.26	73.22	21.38	12878.62	4756.50	221.941558	6.565	False
224	2	Europe	28/09/2012	79.45	3.666	7106.05	1.91	1.65	1549.10	3946.03	222.616433	6.565	False

In [34]:
```
data[(data.Store==4) | (data.Store==8)]
```

Out[34]:

	Store	Continent	Date	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	IsHoliday
546	4	Europe	5/10/2012	63.07	3.620	5918.34	NaN	126.57	3674.49	6807.07	131.075667	3.879	False
547	4	North America	12/10/2012	57.11	3.603	4975.39	NaN	61.17	1513.17	5905.53	131.108333	3.879	False
548	4	Asia	19/10/2012	64.46	3.610	6313.84	NaN	15.05	2421.08	5885.12	131.149968	3.879	False
549	4	Europe	26/10/2012	63.64	3.514	1763.13	88.76	66.76	NaN	7577.14	131.193097	3.879	False
550	4	South America	2/11/2012	53.31	3.404	3717.80	7665.66	23.00	190.24	1586.46	131.236226	3.879	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...
1451	8	Australia	28/06/2013	85.64	3.495	7133.36	388.98	20.20	2483.25	3965.72	NaN	NaN	False
1452	8	Asia	5/7/2013	76.18	3.422	6801.88	1592.93	880.32	5097.59	1717.64	NaN	NaN	False
1453	8	South America	12/7/2013	83.16	3.400	4018.39	893.36	95.35	1586.56	3186.00	NaN	NaN	False
1454	8	South America	19/07/2013	72.86	3.556	1034.72	486.84	43.83	818.13	7118.46	NaN	NaN	False
1455	8	Europe	26/07/2013	78.70	3.620	675.56	740.04	55.38	73.42	2343.11	NaN	NaN	False

364 rows × 13 columns

In [35]:
```
data[data.Store.isin([2, 46])]
```

Out[35]:

	Store	Continent	Date	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	IsHoliday
182	2	Europe	5/4/2013	58.30	3.583	16944.64	4983.34	411.07	2378.92	3007.06	224.719258	6.112	False
183	2	Europe	12/4/2013	61.23	3.529	10165.99	395.13	66.86	1537.62	3122.19	224.802531	6.112	False
184	2	Europe	19/04/2013	67.05	3.451	6858.52	12.72	104.72	219.88	2640.86	224.802531	6.112	False
185	2	South America	26/04/2013	58.13	3.417	2782.18	11.92	146.45	461.83	2046.53	224.802531	6.112	False
186	2	Africa	5/10/2012	70.27	3.617	6037.76	NaN	10.04	3027.37	3853.40	222.815930	6.170	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...
359	2	North America	28/06/2013	85.37	3.495	8638.45	2457.32	9.00	4713.20	9079.05	NaN	NaN	False
360	2	Africa	5/7/2013	79.48	3.422	11651.46	4984.50	2024.67	15196.91	2862.06	NaN	NaN	False
361	2	Australia	12/7/2013	85.41	3.400	7527.10	1244.78	84.18	2626.70	3881.66	NaN	NaN	False
362	2	Europe	19/07/2013	79.16	3.556	3313.12	723.52	94.85	1224.91	2471.69	NaN	NaN	False
363	2	Europe	26/07/2013	83.17	3.620	1966.46	609.55	91.00	493.60	2416.20	NaN	NaN	False

182 rows × 13 columns

In [36]:
```
data[(data.CPI <= 150) & (data.IsHoliday == True)]
```

Out[36]:

	Store	Continent	Date	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	IsHoliday
553	4	Europe	23/11/2012	55.09	3.233	3365.61	6.93	93310.30	497.28	3938.54	131.376667	3.879	True
558	4	Asia	28/12/2012	37.93	3.125	7712.77	50636.71	347.00	6.24	862.39	131.747000	3.879	True
568	4	Europe	8/2/2013	49.86	3.401	56705.09	564.98	521.82	45331.30	7730.26	132.215129	3.921	True
585	4	North America	7/9/2012	82.09	3.709	9082.61	16.00	59.06	7217.88	8026.47	130.932548	4.077	True
607	4	Asia	10/2/2012	33.00	3.411	11374.63	7208.51	118.37	12869.78	10618.93	130.384903	4.607	True
...	...	...	...	...	...	...	...	...	...	...	...	...	...
7940	44	South America	11/2/2011	30.83	3.034	NaN	NaN	NaN	NaN	NaN	127.859129	7.224	True
7955	44	South America	26/11/2010	28.22	2.830	NaN	NaN	NaN	NaN	NaN	126.669267	7.610	True
7960	44	Africa	31/12/2010	26.79	2.868	NaN	NaN	NaN	NaN	NaN	127.087677	7.610	True
7967	44	Asia	10/9/2010	65.74	2.870	NaN	NaN	NaN	NaN	NaN	126.114581	7.804	True
7988	44	Asia	12/2/2010	33.16	2.671	NaN	NaN	NaN	NaN	NaN	126.496258	8.119	True

286 rows × 13 columns

In [85]:
```
data[( ((data.CPI >= 200) | (data.CPI <130)) & (data.Store.isin([4,7])) )]
```

Out[85]:

	Store	Continent	Date	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	IsHoliday
615	4	South America	7/10/2011	65.79	3.299	NaN	NaN	NaN	NaN	NaN	129.693800	5.143	False
616	4	Asia	14/10/2011	63.75	3.283	NaN	NaN	NaN	NaN	NaN	129.770645	5.143	False
617	4	Australia	21/10/2011	64.79	3.361	NaN	NaN	NaN	NaN	NaN	129.782161	5.143	False
618	4	North America	28/10/2011	55.31	3.362	NaN	NaN	NaN	NaN	NaN	129.793677	5.143	False
619	4	Europe	4/11/2011	49.86	3.322	NaN	NaN	NaN	NaN	NaN	129.805194	5.143	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...
1105	7	North America	29/03/2013	26.56	3.605	7886.02	NaN	420.10	69.41	2403.73	201.155809	7.107	False
1106	7	South America	22/03/2013	30.94	3.609	4514.13	NaN	677.30	133.33	983.16	201.198428	7.107	False
1107	7	North America	8/3/2013	24.24	3.624	9370.38	85.98	107.66	2891.20	1385.72	201.212230	7.107	False
1108	7	Africa	15/03/2013	30.09	3.607	2521.46	NaN	219.71	879.39	1203.71	201.241047	7.107	False
1121	7	Europe	28/12/2012	2.32	3.108	7153.00	7398.79	106.33	24.43	1286.48	200.169074	7.557	True

117 rows × 13 columns

In [37]:
```
data.query("CPI >= 220 and Store == 1")
```

Out[37]:

	Store	Continent	Date	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	IsHoliday
0	1	Europe	5/4/2013	58.59	3.583	12872.34	5687.86	485.97	478.04	5092.33	225.086540	6.314	False
1	1	Europe	12/4/2013	62.72	3.529	3672.43	932.58	52.86	949.07	2836.64	225.170160	6.314	False
2	1	Australia	19/04/2013	67.10	3.451	3530.36	3.95	107.50	458.76	2062.05	225.170160	6.314	False
3	1	Africa	26/04/2013	59.23	3.417	2387.72	NaN	98.34	516.28	1421.63	225.170160	6.314	False
4	1	North America	4/1/2013	41.73	3.161	1214.08	25366.33	15.01	72.36	3940.02	224.080983	6.525	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...
64	1	Africa	2/3/2012	60.96	3.630	15441.40	1569.00	10.80	25390.88	8067.61	220.848045	7.348	False
65	1	Australia	9/3/2012	58.76	3.669	10331.04	151.88	6.00	671.43	5509.84	221.059189	7.348	False
66	1	Europe	16/03/2012	64.74	3.734	4298.16	7.50	2.02	2724.65	2017.69	221.211813	7.348	False
67	1	South America	23/03/2012	65.93	3.787	6118.56	9.48	4.97	426.72	3657.22	221.286413	7.348	False
68	1	Europe	30/03/2012	67.61	3.845	10309.58	0.50	10.25	1654.17	2642.78	221.361012	7.348	False

66 rows × 13 columns

Handling missing values

In [38]:
```
data['MarkDown1']=data.MarkDown1.fillna(data.MarkDown1.median())
data['MarkDown2']=data.MarkDown2.fillna(data.MarkDown2.median())
data['MarkDown3']=data.MarkDown3.fillna(data.MarkDown3.median())
data['MarkDown4']=data.MarkDown4.fillna(data.MarkDown4.median())
data['MarkDown5']=data.MarkDown5.fillna(data.MarkDown5.median())
data
```

Out[38]:

	Store	Continent	Date	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	IsHoliday
0	1	Europe	5/4/2013	58.59	3.583	12872.34	5687.86	485.97	478.04	5092.33	225.086540	6.314	False
1	1	Europe	12/4/2013	62.72	3.529	3672.43	932.58	52.86	949.07	2836.64	225.170160	6.314	False
2	1	Australia	19/04/2013	67.10	3.451	3530.36	3.95	107.50	458.76	2062.05	225.170160	6.314	False
3	1	Africa	26/04/2013	59.23	3.417	2387.72	364.57	98.34	516.28	1421.63	225.170160	6.314	False
4	1	North America	4/1/2013	41.73	3.161	1214.08	25366.33	15.01	72.36	3940.02	224.080983	6.525	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...
8185	45	Europe	28/06/2013	76.05	3.639	4842.29	975.03	3.00	2449.97	3169.69	NaN	NaN	False
8186	45	Australia	5/7/2013	77.50	3.614	9090.48	2268.58	582.74	5797.47	1514.93	NaN	NaN	False
8187	45	South America	12/7/2013	79.37	3.614	3789.94	1827.31	85.72	744.84	2150.36	NaN	NaN	False
8188	45	South America	19/07/2013	82.84	3.737	2961.49	1047.07	204.19	363.00	1059.46	NaN	NaN	False
8189	45	South America	26/07/2013	76.06	3.804	212.02	851.73	2.06	10.88	1864.57	NaN	NaN	False

8190 rows × 13 columns

In [39]:
```
CPI_mean = data.groupby('Store')['CPI'].transform('mean').round(7)
data['CPI'] = data['CPI'].fillna(CPI_mean)
data
```

Out[39]:

	Store	Continent	Date	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	IsHoliday
0	1	Europe	5/4/2013	58.59	3.583	12872.34	5687.86	485.97	478.04	5092.33	225.086540	6.314	False
1	1	Europe	12/4/2013	62.72	3.529	3672.43	932.58	52.86	949.07	2836.64	225.170160	6.314	False
2	1	Australia	19/04/2013	67.10	3.451	3530.36	3.95	107.50	458.76	2062.05	225.170160	6.314	False
3	1	Africa	26/04/2013	59.23	3.417	2387.72	364.57	98.34	516.28	1421.63	225.170160	6.314	False
4	1	North America	4/1/2013	41.73	3.161	1214.08	25366.33	15.01	72.36	3940.02	224.080983	6.525	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...
8185	45	Europe	28/06/2013	76.05	3.639	4842.29	975.03	3.00	2449.97	3169.69	187.298763	NaN	False
8186	45	Australia	5/7/2013	77.50	3.614	9090.48	2268.58	582.74	5797.47	1514.93	187.298763	NaN	False
8187	45	South America	12/7/2013	79.37	3.614	3789.94	1827.31	85.72	744.84	2150.36	187.298763	NaN	False
8188	45	South America	19/07/2013	82.84	3.737	2961.49	1047.07	204.19	363.00	1059.46	187.298763	NaN	False
8189	45	South America	26/07/2013	76.06	3.804	212.02	851.73	2.06	10.88	1864.57	187.298763	NaN	False

8190 rows × 13 columns

In [40]:
```
store_mode = data.groupby('Store')['Unemployment'].transform(lambda x: x.fillna(x.mode().iloc[0]))
data['Unemployment'] = store_mode
data
```

Out[40]:

	Store	Continent	Date	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	IsHoliday
0	1	Europe	5/4/2013	58.59	3.583	12872.34	5687.86	485.97	478.04	5092.33	225.086540	6.314	False
1	1	Europe	12/4/2013	62.72	3.529	3672.43	932.58	52.86	949.07	2836.64	225.170160	6.314	False
2	1	Australia	19/04/2013	67.10	3.451	3530.36	3.95	107.50	458.76	2062.05	225.170160	6.314	False
3	1	Africa	26/04/2013	59.23	3.417	2387.72	364.57	98.34	516.28	1421.63	225.170160	6.314	False
4	1	North America	4/1/2013	41.73	3.161	1214.08	25366.33	15.01	72.36	3940.02	224.080983	6.525	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...
8185	45	Europe	28/06/2013	76.05	3.639	4842.29	975.03	3.00	2449.97	3169.69	187.298763	8.625	False
8186	45	Australia	5/7/2013	77.50	3.614	9090.48	2268.58	582.74	5797.47	1514.93	187.298763	8.625	False
8187	45	South America	12/7/2013	79.37	3.614	3789.94	1827.31	85.72	744.84	2150.36	187.298763	8.625	False
8188	45	South America	19/07/2013	82.84	3.737	2961.49	1047.07	204.19	363.00	1059.46	187.298763	8.625	False
8189	45	South America	26/07/2013	76.06	3.804	212.02	851.73	2.06	10.88	1864.57	187.298763	8.625	False

8190 rows × 13 columns

Handling outliers

In [41]:
```
upper_limit = data['MarkDown1'].quantile(.95)
lower_limit = data['MarkDown1'].quantile(.05)

data = data[(data['MarkDown1'] < upper_limit) &
```

```
(data['MarkDown1'] > lower_limit)]

upper_limit_2 = data['MarkDown2'].quantile(.95)
lower_limit_2 = data['MarkDown2'].quantile(.05)

data = data[(data['MarkDown2'] < upper_limit_2) &
(data['MarkDown2'] > lower_limit_2)]

upper_limit_3 = data['MarkDown3'].quantile(.95)
lower_limit_3 = data['MarkDown3'].quantile(.05)

data = data[(data['MarkDown3'] < upper_limit_3) &
(data['MarkDown3'] > lower_limit_3)]

upper_limit_4 = data['MarkDown4'].quantile(.95)
lower_limit_4 = data['MarkDown4'].quantile(.05)

data = data[(data['MarkDown4'] < upper_limit_4) &
(data['MarkDown4'] > lower_limit_4)]

upper_limit_5 = data['MarkDown5'].quantile(.95)
lower_limit_5 = data['MarkDown5'].quantile(.05)

data = data[(data['MarkDown5'] < upper_limit_5) &
(data['MarkDown5'] > lower_limit_5)]

data
```

Out[41]:

	Store	Continent	Date	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	IsHoliday
1	1	Europe	12/4/2013	62.72	3.529	3672.43	932.58	52.86	949.070	2836.640	225.170160	6.314	False
9	1	Africa	18/01/2013	42.92	3.237	3772.69	3559.46	3.88	246.620	1900.400	224.235813	6.525	False
13	1	Australia	8/3/2013	50.81	3.658	5000.58	290.46	78.77	606.150	3697.110	224.708763	6.525	False
14	1	North America	15/03/2013	55.33	3.622	3808.13	364.57	15.65	2616.600	1909.170	224.835681	6.525	False
27	1	Asia	14/12/2012	48.89	3.168	3504.83	364.57	73.26	1636.800	2779.600	223.719277	6.573	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...
8176	45	Europe	12/3/2010	45.80	2.818	4743.58	364.57	36.26	1176.425	2727.135	182.162844	8.992	False
8179	45	Africa	17/05/2013	60.59	3.614	4515.35	667.88	6.12	522.700	2541.620	187.298763	8.625	False
8183	45	Australia	14/06/2013	70.01	3.632	2471.44	517.87	348.54	2612.330	3459.390	187.298763	8.625	False
8184	45	North America	21/06/2013	70.13	3.626	4989.34	385.31	178.56	2463.420	3117.940	187.298763	8.625	False
8187	45	South America	12/7/2013	79.37	3.614	3789.94	1827.31	85.72	744.840	2150.360	187.298763	8.625	False

4824 rows × 13 columns

Feature encoding techniques

In [45]:
```
from sklearn.preprocessing import LabelEncoder
label_encoder = LabelEncoder()
encoded_data = label_encoder.fit_transform(data['Continent'])
print(encoded_data)
from sklearn.preprocessing import OrdinalEncoder
order_encoder=OrdinalEncoder(categories=['Africa', 'Asia', 'Australia', 'Europe', 'North America', 'South America'])
data['Continent_encoded'] = label_encoder.fit_transform(data['Continent'])
data
```

```
[3 0 2 ... 2 4 5]
```

Out[45]:

	Store	Continent	Date	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	IsHoliday	C
1	1	Europe	12/4/2013	62.72	3.529	3672.43	932.58	52.86	949.070	2836.640	225.170160	6.314	False	
9	1	Africa	18/01/2013	42.92	3.237	3772.69	3559.46	3.88	246.620	1900.400	224.235813	6.525	False	
13	1	Australia	8/3/2013	50.81	3.658	5000.58	290.46	78.77	606.150	3697.110	224.708763	6.525	False	
14	1	North America	15/03/2013	55.33	3.622	3808.13	364.57	15.65	2616.600	1909.170	224.835681	6.525	False	
27	1	Asia	14/12/2012	48.89	3.168	3504.83	364.57	73.26	1636.800	2779.600	223.719277	6.573	False	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
8176	45	Europe	12/3/2010	45.80	2.818	4743.58	364.57	36.26	1176.425	2727.135	182.162844	8.992	False	
8179	45	Africa	17/05/2013	60.59	3.614	4515.35	667.88	6.12	522.700	2541.620	187.298763	8.625	False	
8183	45	Australia	14/06/2013	70.01	3.632	2471.44	517.87	348.54	2612.330	3459.390	187.298763	8.625	False	
8184	45	North America	21/06/2013	70.13	3.626	4989.34	385.31	178.56	2463.420	3117.940	187.298763	8.625	False	
8187	45	South America	12/7/2013	79.37	3.614	3789.94	1827.31	85.72	744.840	2150.360	187.298763	8.625	False	

4824 rows × 16 columns

Feature scaling

In [44]:
```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()

scaler.fit(data['CPI'].values.reshape(-1,1))
data['CPI_scaler']=scaler.transform(data['CPI'].values.reshape(-1,1))

scaler.fit(data['Unemployment'].values.reshape(-1,1))
data['Unemployment_scaler']=scaler.transform(data['Unemployment'].values.reshape(-1,1))
data.iloc[:1000,10:]
```

Out[44]:

	CPI	Unemployment	IsHoliday	Continent_encoded	CPI_scaler	Unemployment_scaler
1	225.170160	6.314	False	3	1.422863	-0.991820
9	224.235813	6.525	False	0	1.398705	-0.880278
13	224.708763	6.525	False	2	1.410933	-0.880278
14	224.835681	6.525	False	4	1.414215	-0.880278
27	223.719277	6.573	False	1	1.385349	-0.854904
...	...	...	...	...	...	...
1728	129.283258	8.257	False	1	-1.056397	0.035315
1729	129.325936	8.257	False	4	-1.055293	0.035315
1730	129.368613	8.257	True	2	-1.054190	0.035315
1731	129.430600	8.257	False	2	-1.052587	0.035315
1732	129.518333	8.257	False	4	-1.050319	0.035315

1000 rows × 6 columns

Save Dataset

In [46]:
```
data.to_csv('Cleaned_Features_data_set.csv', index=False)
```

In [ ]:
```
```