BASIC STATS LEVEL 1

Descriptive Analytics for Numerical Columns

[data.head]

	Date	Day	SKU	City	Volume	BU	Brand	Model	Avg Price	Total Sales Value	Discount Rate (%)	Discount Amount	Net Sales Value
0	01-04-2021	Thursday	M01	C	15	Mobiles	RealU	RU-10	12100	181500	11.654820	21153.498820	160346.501180
1	01-04-2021	Thursday	M02	C	10	Mobiles	RealU	RU-9 Plus	10100	101000	11.560498	11676.102961	89323.897039
2	01-04-2021	Thursday	M03	C	7	Mobiles	YouM	YM-99	16100	112700	9.456886	10657.910157	102042.089843
3	01-04-2021	Thursday	M04	C	6	Mobiles	YouM	YM-99 Plus	20100	120600	6.935385	8364.074702	112235.925298
4	01-04-2021	Thursday	M05	C	3	Mobiles	YouM	YM-98	8100	24300	17.995663	4372.946230	19927.053770

- Dataset had 450 rows and 13 columns.
- Had no missing values.
- 6 are categorical columns
- 6 are numerical columns.

Date	object
Day	object
SKU	object
City	object
Volume	int64
BU	object
Brand	object
Model	object
Avg Price	int64
Total Sales Value	int64
Discount Rate (%)	float64
Discount Amount	float64
Net Sales Value	float64

dtype: object

STATISTICAL MEASURE

	Volume	Avg Price	Total Sales Value	Discount Rate (%)	Discount Amount	Net Sales Value
count	450.000000	450.000000	450.000000	450.000000	450.000000	450.000000
mean	5.066667	10453.433333	33812.835556	15.155242	3346.499424	30466.336131
std	4.231602	18079.904840	50535.074173	4.220602	4509.902963	46358.656624
min	1.000000	290.000000	400.000000	5.007822	69.177942	326.974801
25%	3.000000	465.000000	2700.000000	13.965063	460.459304	2202.208645
50%	4.000000	1450.000000	5700.000000	16.577766	988.933733	4677.788059
75%	6.000000	10100.000000	53200.000000	18.114718	5316.495427	47847.912852
max	31.000000	60100.000000	196400.000000	19.992407	25738.022194	179507.479049

- The columns with positive skewness are :
    1. Volume
    2. Avg Price
    3. Total Sales Value
    4. Discount Amount
    5. Net Sales Value

Skewness of data.

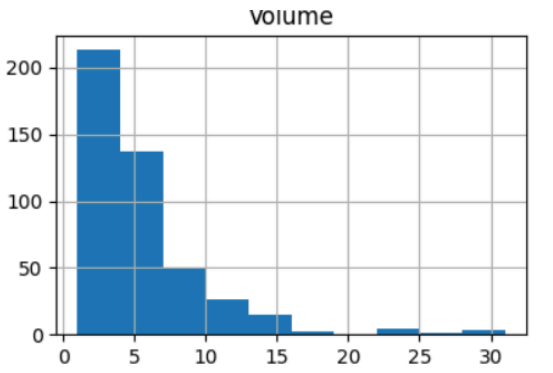
	0
Volume	2.731724
Discount Amount	1.913038
Net Sales Value	1.540822
Discount Rate (%)	-1.062294
Total Sales Value	1.534729
Avg Price	1.908873

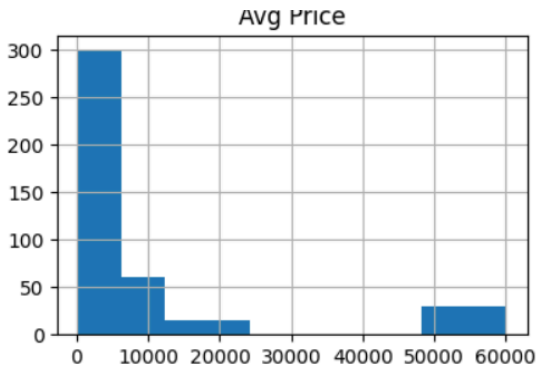
dtype: float64

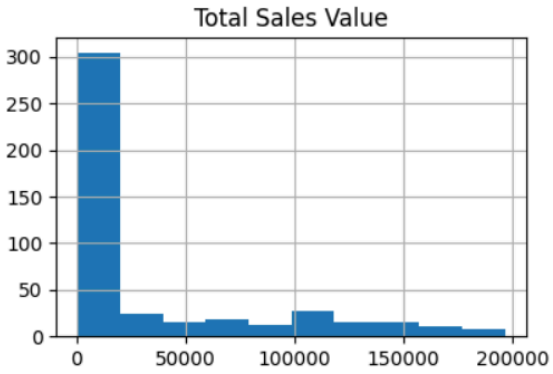
Data Visualisation

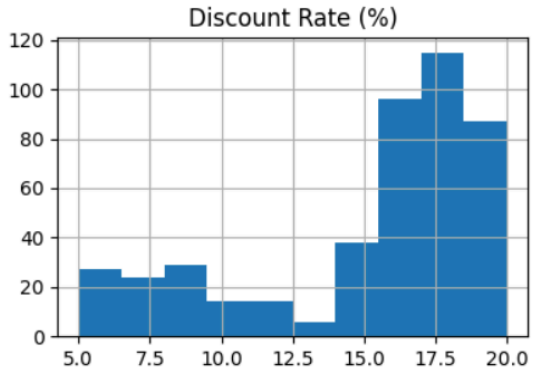
HISTOGRAM

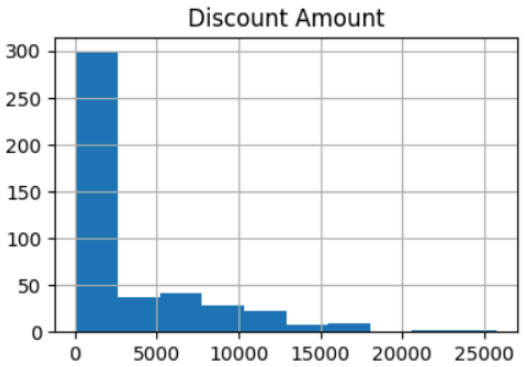
Now let's check the distribution of each numerical column:

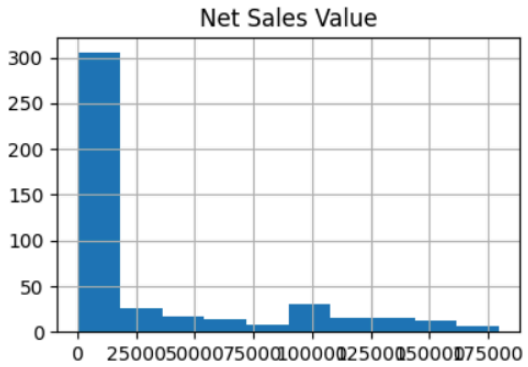












Columns which are normally distributed:

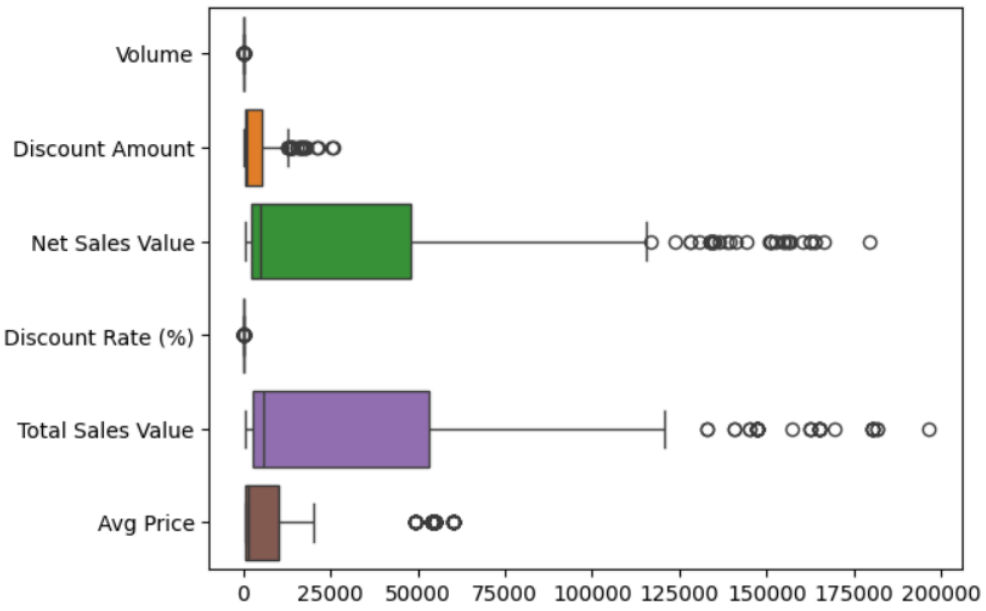
- Total sales value
- Net sales value

Columns which are negative skewed:

- Discount rate (%)

BOXPLOT

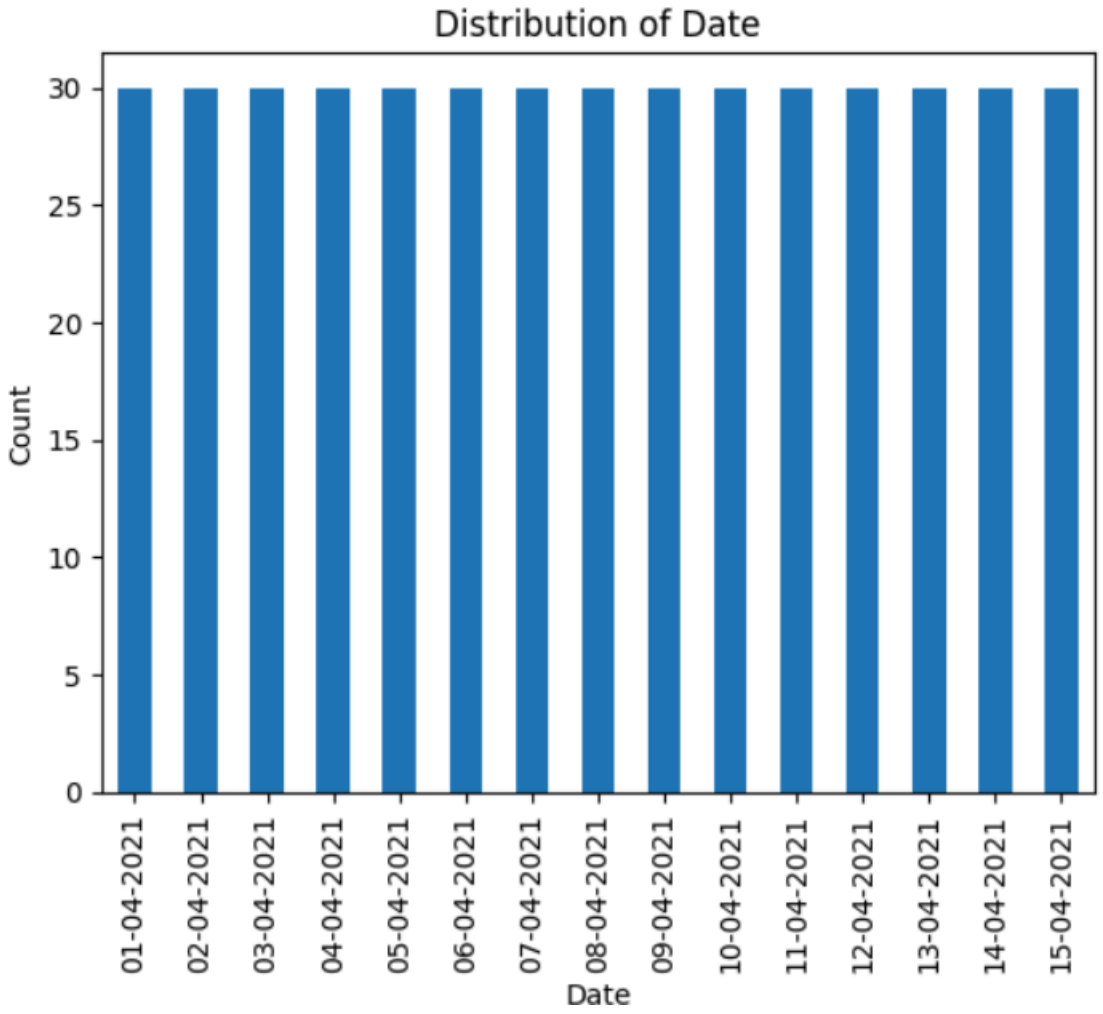
<Axes: >

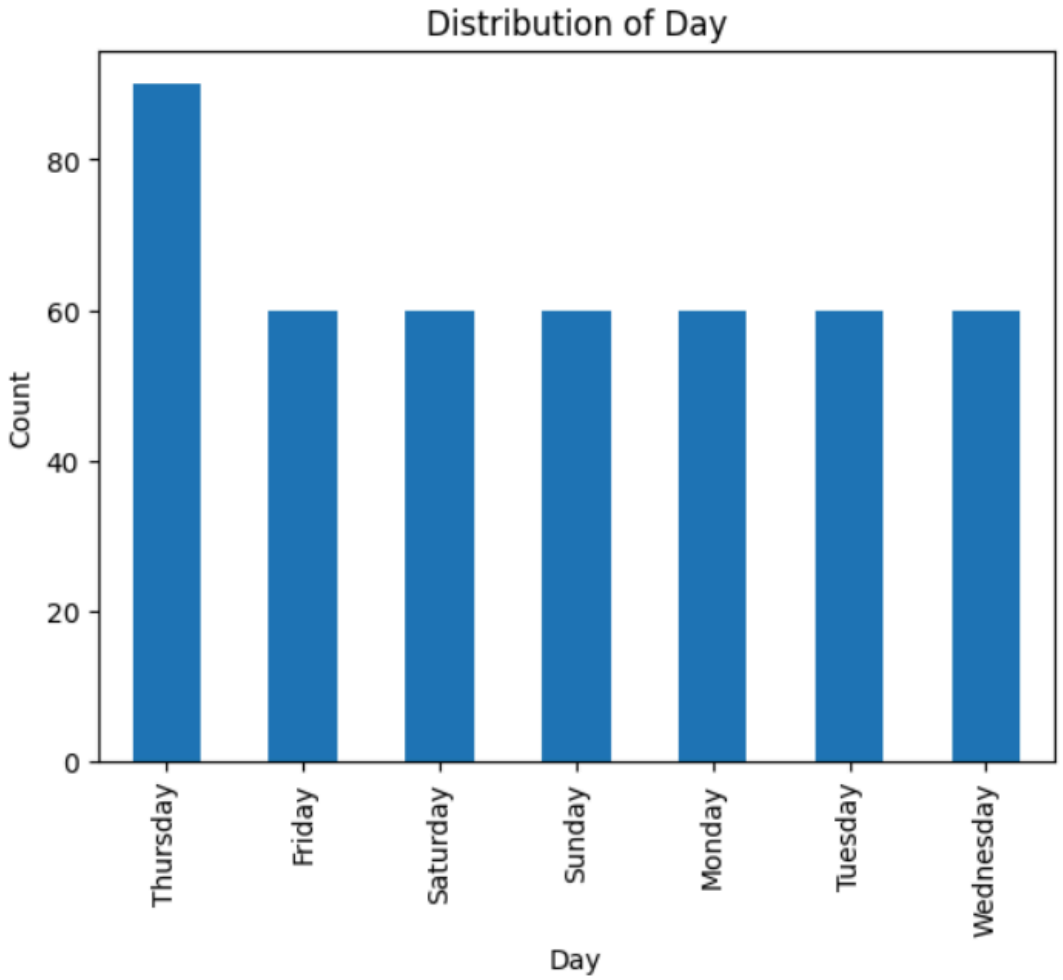


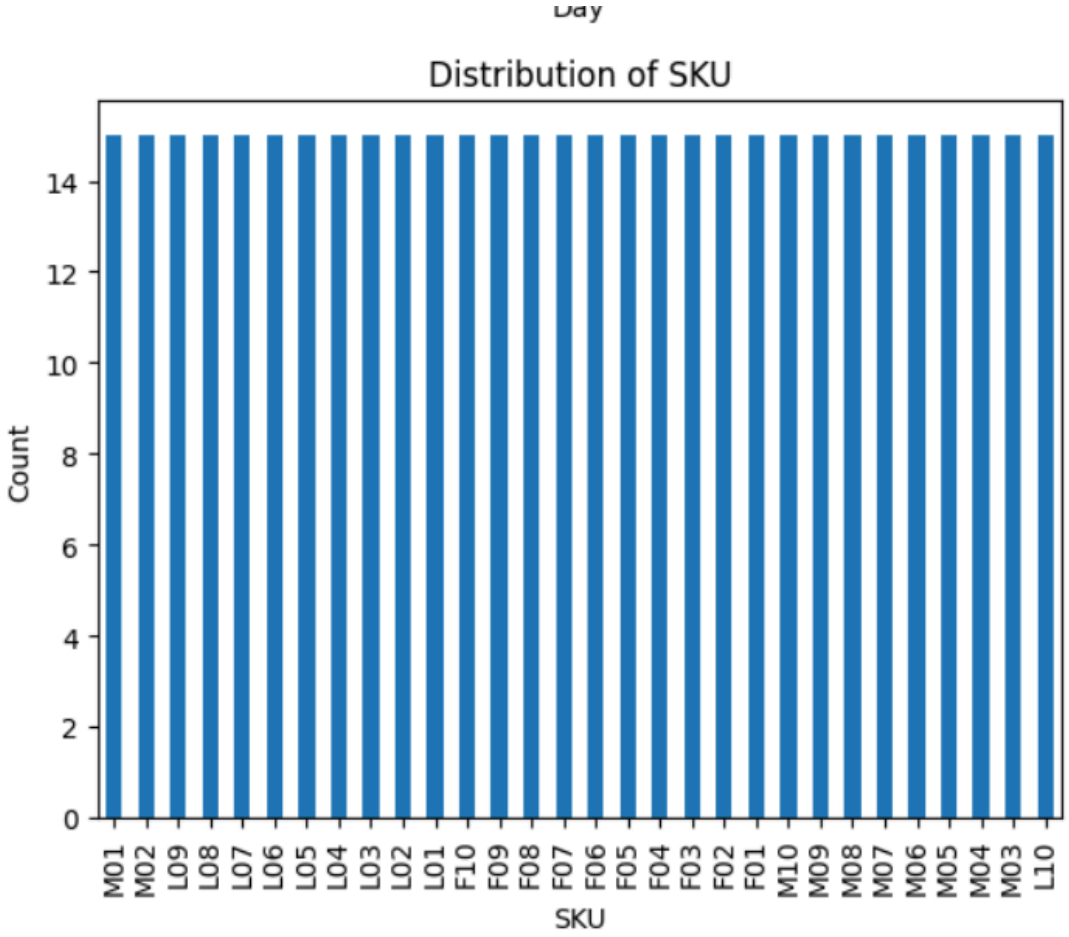
Columns with outliers:

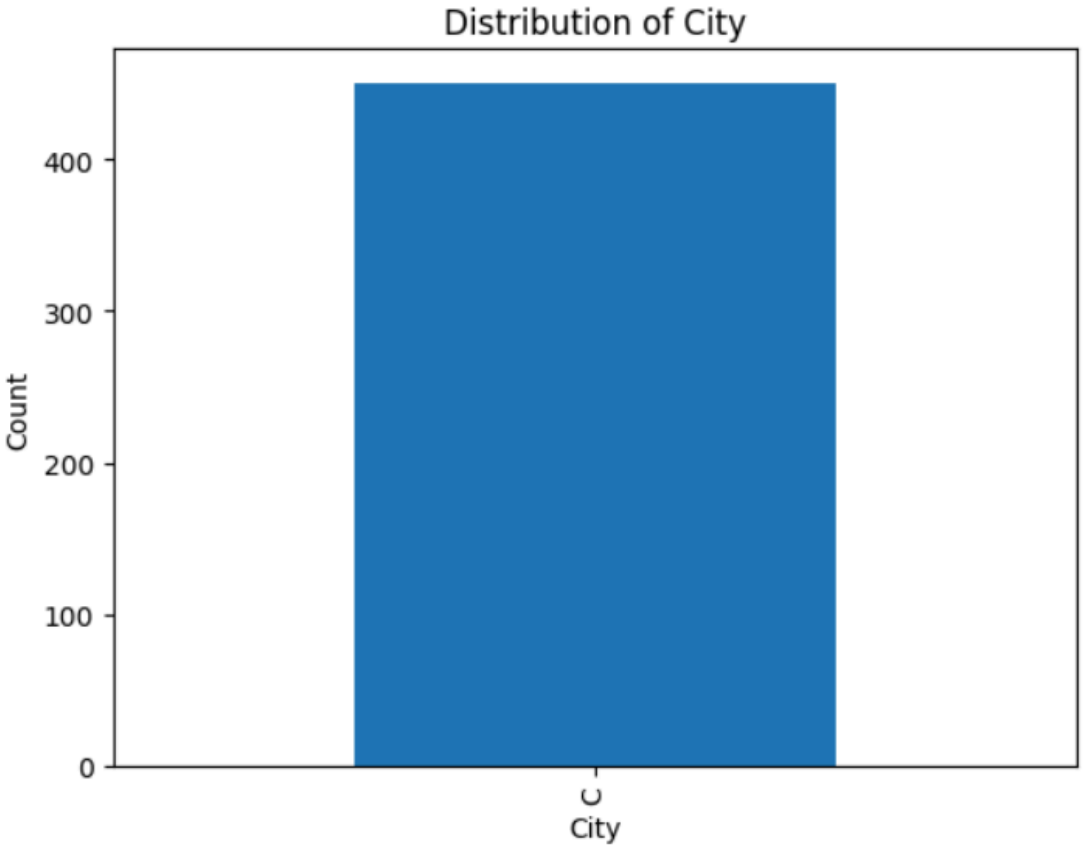
- Total sales value
- Net sales value
- Discount amount

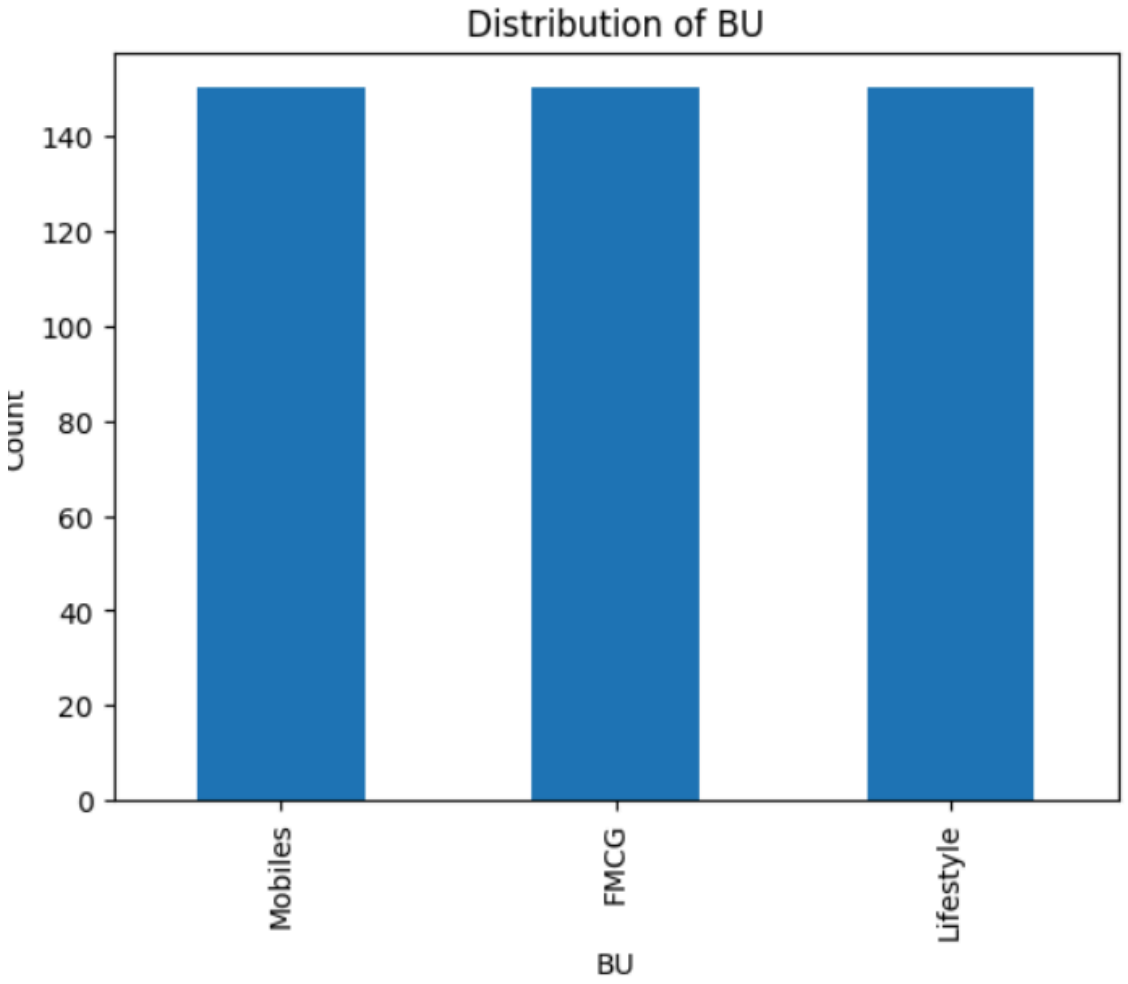
BARCHART FOR ANALYSIS OF CATEGORICAL COLUMNS

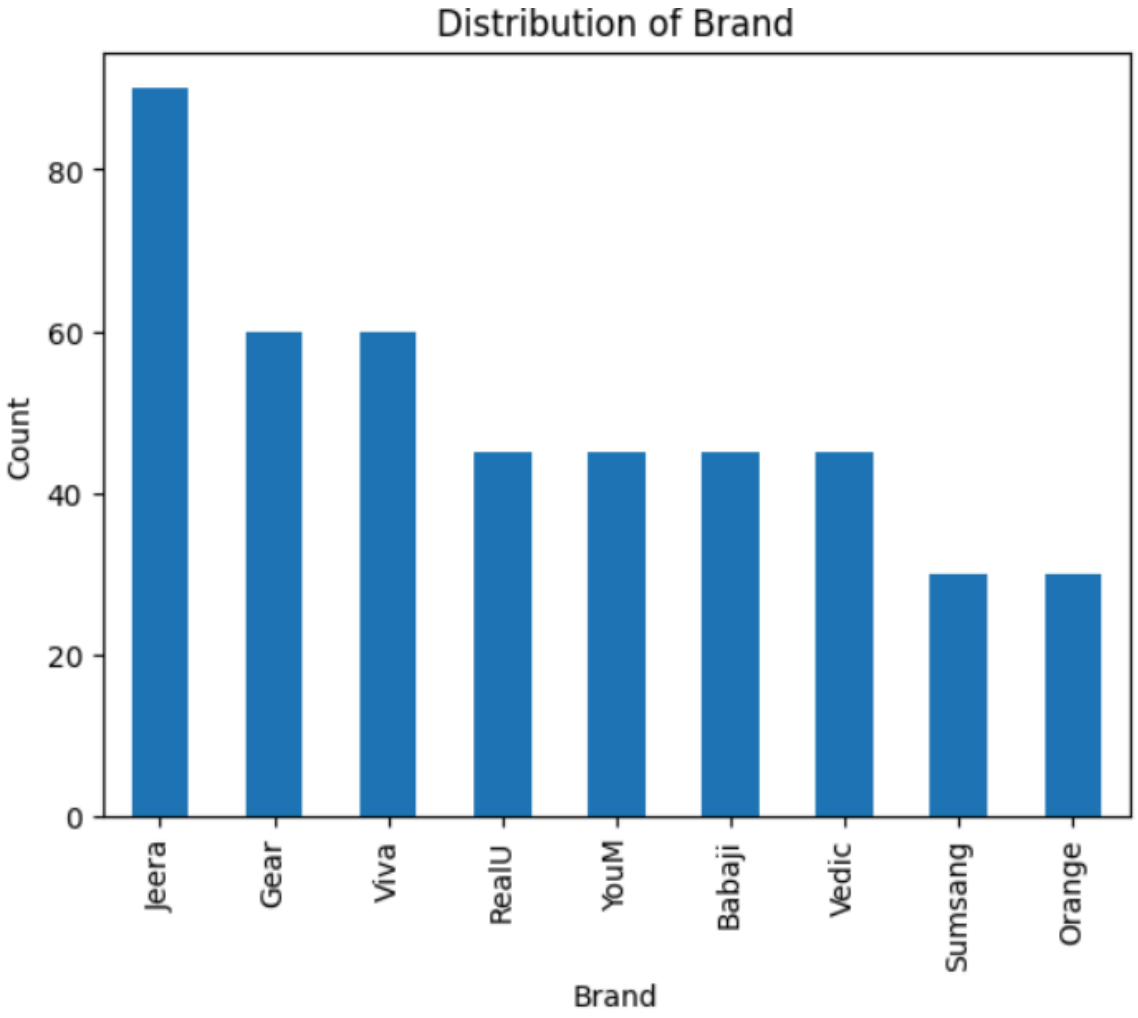

Distribution of Date
Count
30
25
20
15
10
5
0
01-04-2021
02-04-2021
03-04-2021
04-04-2021
05-04-2021
06-04-2021
07-04-2021
08-04-2021
09-04-2021
10-04-2021
11-04-2021
12-04-2021
13-04-2021
14-04-2021
15-04-2021
Date

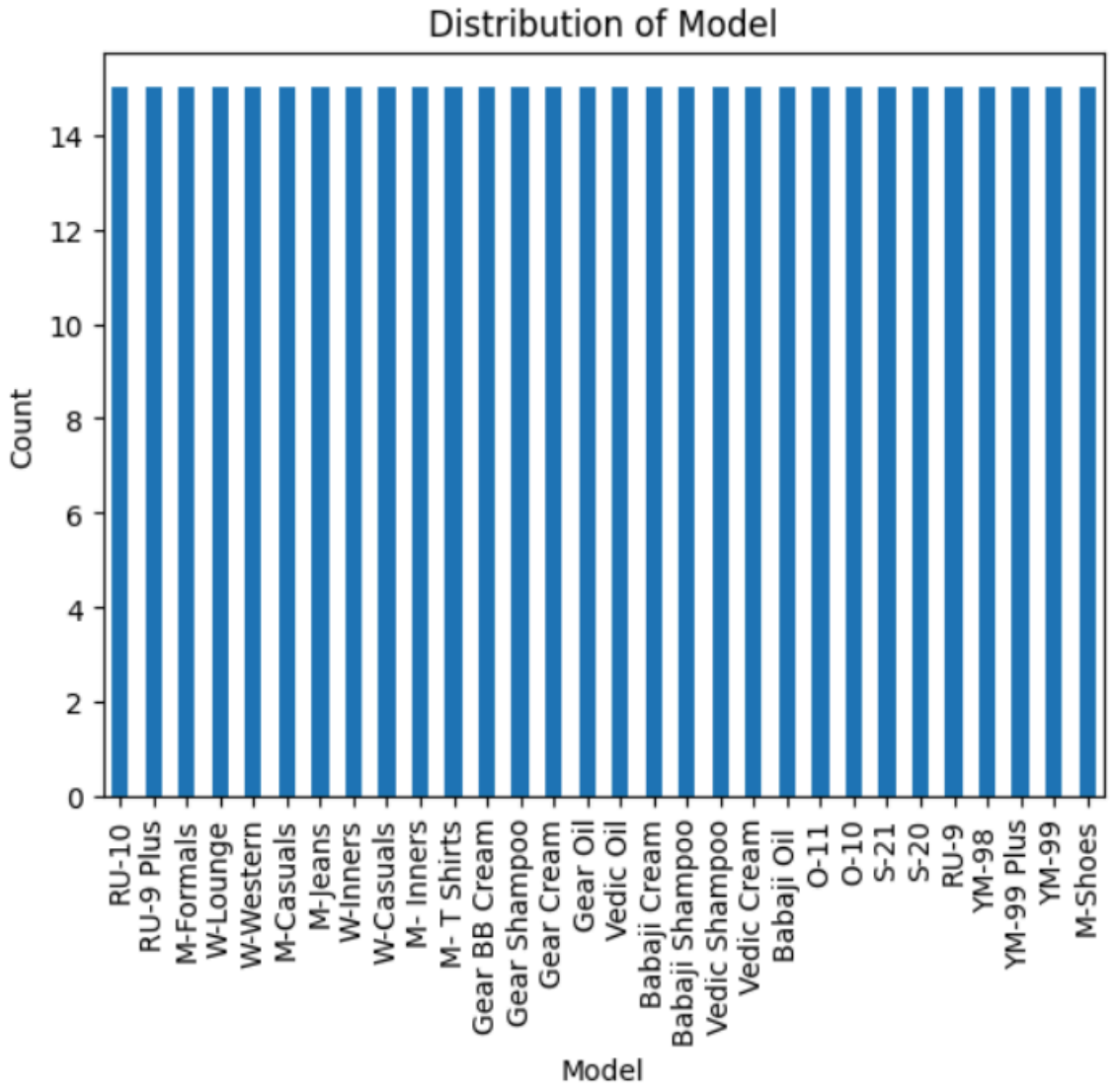

Distribution of Day
Count
0
20
40
60
80
Thursday
Friday
Saturday
Sunday
Monday
Tuesday
Wednesday
Day


Day
Distribution of SKU
Count
0
2
4
6
8
10
12
14
M01
M02
L09
L08
L07
L06
L05
L04
L03
L02
L01
F10
F09
F08
F07
F06
F05
F04
F03
F02
F01
M10
M09
M08
M07
M06
M05
M04
M03
L10
SKU


Distribution of City
Count
400
300
200
100
0
C
City


Distribution of BU
Count
140
120
100
80
60
40
20
0
Mobiles
FMCG
Lifestyle
BU


Distribution of Brand
Count
0
20
40
60
80
Jeera
Gear
Viva
RealU
YouM
Babaji
Vedic
Sumsang
Orange
Brand



Standardisation Of Numerical Variables

Standardisation is a technique used to transform data to have a mean of 0 and a standard deviation of 1.

After standardisation:

	0	1	2	3	4	5	6	7	8	9	10	11	12
0	-1.620185	0.47907	0.635438	0.0	2.350029	1.224745	0.076176	0.057767	0.091173	2.925721	-0.830289	3.952816	2.804756
1	-1.620185	0.47907	0.750973	0.0	1.167129	1.224745	0.076176	0.288836	-0.019570	1.330995	-0.852661	1.849014	1.271026
2	-1.620185	0.47907	0.866507	0.0	0.457388	1.224745	1.599686	1.559712	0.312659	1.562775	-1.351631	1.622995	1.545675
3	-1.620185	0.47907	0.982041	0.0	0.220808	1.224745	1.599686	1.675247	0.534146	1.719276	-1.949723	1.113807	1.765810
4	-1.620185	0.47907	1.097575	0.0	-0.488932	1.224745	1.599686	1.444178	-0.130313	-0.188452	0.673739	0.227852	-0.227595

Conversion Of Categorical Variables Into Dummies

One method for this is **one-hot encoding**, it transforms categorical variables into binary numbers(0 or 1)

One-hot encoding creates a sparse matrix where each category is represented by a binary vector, making it easy for algorithms to process the data.

After one-hot encoding:

Date_01-04-2021	Date_02-04-2021	Date_03-04-2021	Date_04-04-2021	...	Model_Vedic Cream	Model_Vedic Oil	Model_Vedic Shampoo	Model_W-Casuals	Model_W-Inners	Model_W-Lounge	Model_W-Western	Model_YM-98	Model_YM-99	Model_YM-99 Plus
1.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
1.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
1.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0

Conclusion

★ The shape of data has become (450, 85)

★ No missing values.

★ Outliers are present.

★ The data is not normally distributed.

★ Standardisation disturbed the data normally.

★ There is skewness in data.

★ One hot encoding increased the number of columns.