**1. Import required libraries and data inspection**

* Import all required libraries
* Import raw data
* Initial inspection of the data for duplicated rows, missing and data types

Mounted at /content/drive

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Show first five rows of the data:

Date Store ID Product ID Category Region Inventory Level \

0 1/1/2022 S001 P0001 Groceries North 231

1 1/1/2022 S001 P0002 Toys South 204

2 1/1/2022 S001 P0003 Toys West 102

3 1/1/2022 S001 P0004 Toys North 469

4 1/1/2022 S001 P0005 Electronics East 166

Units Sold Units Ordered Demand Forecast Price Discount \

0 127 55 135.47 33.50 20

1 150 66 144.04 63.01 20

2 65 51 74.02 27.99 10

3 61 164 62.18 32.72 10

4 14 135 9.26 73.64 0

Weather Condition Holiday/Promotion Competitor Pricing Seasonality

0 Rainy 0 29.69 Autumn

1 Sunny 0 66.16 Autumn

2 Sunny 1 31.32 Summer

3 Cloudy 1 34.74 Autumn

4 Sunny 0 68.95 Summer

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 73100 entries, 0 to 73099

Data columns (total 15 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Date 73100 non-null object

1 Store ID 73100 non-null object

2 Product ID 73100 non-null object

3 Category 73100 non-null object

4 Region 73100 non-null object

5 Inventory Level 73100 non-null int64

6 Units Sold 73100 non-null int64

7 Units Ordered 73100 non-null int64

8 Demand Forecast 73100 non-null float64

9 Price 73100 non-null float64

10 Discount 73100 non-null int64

11 Weather Condition 73100 non-null object

12 Holiday/Promotion 73100 non-null int64

13 Competitor Pricing 73100 non-null float64

14 Seasonality 73100 non-null object

dtypes: float64(3), int64(5), object(7)

memory usage: 8.4+ MB

None

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Date Store ID Product ID Category Region Inventory Level \

count 73100 73100 73100 73100 73100 73100.000000

unique 731 5 20 5 4 NaN

top 1/1/2024 S001 P0001 Furniture East NaN

freq 100 14620 3655 14699 18349 NaN

mean NaN NaN NaN NaN NaN 274.469877

std NaN NaN NaN NaN NaN 129.949514

min NaN NaN NaN NaN NaN 50.000000

25% NaN NaN NaN NaN NaN 162.000000

50% NaN NaN NaN NaN NaN 273.000000

75% NaN NaN NaN NaN NaN 387.000000

max NaN NaN NaN NaN NaN 500.000000

Units Sold Units Ordered Demand Forecast Price \

count 73100.000000 73100.000000 73100.000000 73100.000000

unique NaN NaN NaN NaN

top NaN NaN NaN NaN

freq NaN NaN NaN NaN

mean 136.464870 110.004473 141.494720 55.135108

std 108.919406 52.277448 109.254076 26.021945

min 0.000000 20.000000 -9.990000 10.000000

25% 49.000000 65.000000 53.670000 32.650000

50% 107.000000 110.000000 113.015000 55.050000

75% 203.000000 155.000000 208.052500 77.860000

max 499.000000 200.000000 518.550000 100.000000

Discount Weather Condition Holiday/Promotion Competitor Pricing \

count 73100.000000 73100 73100.000000 73100.000000

unique NaN 4 NaN NaN

top NaN Sunny NaN NaN

freq NaN 18290 NaN NaN

mean 10.009508 NaN 0.497305 55.146077

std 7.083746 NaN 0.499996 26.191408

min 0.000000 NaN 0.000000 5.030000

25% 5.000000 NaN 0.000000 32.680000

50% 10.000000 NaN 0.000000 55.010000

75% 15.000000 NaN 1.000000 77.820000

max 20.000000 NaN 1.000000 104.940000

Seasonality

count 73100

unique 4

top Spring

freq 18317

mean NaN

std NaN

min NaN

25% NaN

50% NaN

75% NaN

max NaN

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

number of missing values in each column:

Date 0

Store ID 0

Product ID 0

Category 0

Region 0

Inventory Level 0

Units Sold 0

Units Ordered 0

Demand Forecast 0

Price 0

Discount 0

Weather Condition 0

Holiday/Promotion 0

Competitor Pricing 0

Seasonality 0

dtype: int64

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

number of duplicated rows: 0

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

According to results of above cell, I conclude:

* No duplicated rows found in data
* Following columns have no missing values: 'Inventory Level', 'Units Sold', 'Units Ordered', 'Demand Forecast', 'Price', 'Discount'
* For other columns data type conversion need to be done first

# 2. Data type conversion for rest of columns

# 3. Data inspection after data type conversion

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 73100 entries, 0 to 73099

Data columns (total 15 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Date 73100 non-null datetime64[ns]

1 Store ID 73100 non-null string

2 Product ID 73100 non-null string

3 Category 73100 non-null string

4 Region 73100 non-null string

5 Inventory Level 73100 non-null int64

6 Units Sold 73100 non-null int64

7 Units Ordered 73100 non-null int64

8 Demand Forecast 73100 non-null float64

9 Price 73100 non-null float64

10 Discount 73100 non-null int64

11 Weather Condition 73100 non-null string

12 Holiday/Promotion 73100 non-null int64

13 Competitor Pricing 73100 non-null float64

14 Seasonality 73100 non-null string

dtypes: datetime64[ns](1), float64(3), int64(5), string(6)

memory usage: 8.4 MB

None

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Date Store ID Product ID Category Region \

count 73100 73100 73100 73100 73100

unique NaN 5 20 5 4

top NaN S001 P0001 Furniture East

freq NaN 14620 3655 14699 18349

mean 2022-12-31 23:59:59.999999744 NaN NaN NaN NaN

min 2022-01-01 00:00:00 NaN NaN NaN NaN

25% 2022-07-02 00:00:00 NaN NaN NaN NaN

50% 2023-01-01 00:00:00 NaN NaN NaN NaN

75% 2023-07-03 00:00:00 NaN NaN NaN NaN

max 2024-01-01 00:00:00 NaN NaN NaN NaN

std NaN NaN NaN NaN NaN

Inventory Level Units Sold Units Ordered Demand Forecast \

count 73100.000000 73100.000000 73100.000000 73100.000000

unique NaN NaN NaN NaN

top NaN NaN NaN NaN

freq NaN NaN NaN NaN

mean 274.469877 136.464870 110.004473 141.494720

min 50.000000 0.000000 20.000000 -9.990000

25% 162.000000 49.000000 65.000000 53.670000

50% 273.000000 107.000000 110.000000 113.015000

75% 387.000000 203.000000 155.000000 208.052500

max 500.000000 499.000000 200.000000 518.550000

std 129.949514 108.919406 52.277448 109.254076

Price Discount Weather Condition Holiday/Promotion \

count 73100.000000 73100.000000 73100 73100.000000

unique NaN NaN 4 NaN

top NaN NaN Sunny NaN

freq NaN NaN 18290 NaN

mean 55.135108 10.009508 NaN 0.497305

min 10.000000 0.000000 NaN 0.000000

25% 32.650000 5.000000 NaN 0.000000

50% 55.050000 10.000000 NaN 0.000000

75% 77.860000 15.000000 NaN 1.000000

max 100.000000 20.000000 NaN 1.000000

std 26.021945 7.083746 NaN 0.499996

Competitor Pricing Seasonality

count 73100.000000 73100

unique NaN 4

top NaN Spring

freq NaN 18317

mean 55.146077 NaN

min 5.030000 NaN

25% 32.680000 NaN

50% 55.010000 NaN

75% 77.820000 NaN

max 104.940000 NaN

std 26.191408 NaN

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

number of missing values in each column:

Date 0

Store ID 0

Product ID 0

Category 0

Region 0

Inventory Level 0

Units Sold 0

Units Ordered 0

Demand Forecast 0

Price 0

Discount 0

Weather Condition 0

Holiday/Promotion 0

Competitor Pricing 0

Seasonality 0

dtype: int64

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**According to the above results, no missing found in the columns and all data types for each column are same.**

# 4. Encode categorical variables using one-hot encoding.

# 5. Save cleaned data into a file