EDA / Descriptive Statistics

Introduction:

Exploratory Data Analysis (EDA) and descriptive statistics are two powerful tools that data analysts use to understand a dataset before diving deeper into statistical modeling or machine learning. They work together to summarize the data, identify patterns and trends, and uncover potential issues.

Descriptive statistics provide a numerical summary of the data. It helps get a basic understanding of the data set by calculating measures of central tendency (like mean, median) and spread (like standard deviation, variance).

Overall design strategy:

- Conduct exploratory analysis to understand the distribution and relationships within the data.
- Visualize the distribution of Pharmacetucial Inventories using histograms or box plots to identify patterns and outliers.
- Explore relationships between Pharmacetucial Inventories and other variables such as ‘Qty_in_Un_of_Entry','Qty_in_OPUn', 'Qty_in_order_unit' and etc using scatter plots or correlation matrices.

Data Overview:

1. Dataset name: df

2. Total_number of rows before data cleaning 61567.

3. Total_number of columns before data cleaning is 25.

Total_column names:

['Material', 'Material Description', 'Plant', 'Storage Location', 'Movement Type', 'Posting Date', 'Qty in Un. of Entry', 'Unit of Entry', 'Movement Type Text', 'Document Date', 'Qty in OPUn', 'Order Price Unit', 'Order Unit', 'Qty in order unit', 'Entry Date', 'Amount in LC', 'Purchase Order, 'Movement indicator', 'Base Unit of Measure', 'Quantity', 'Material Doc. Year', 'Debit/Credit ind, 'Trans./Event Type', 'Material Type','Vendor Code']

4. In column ‘Trans_/Event_Type’ of pharmacetical data includes WA,WE,WL and WQ.

5. Material column has 2 different types of materials ZCON and ZCPC.

6. Plant has different types includes 8110, 8170, 8310, 8320, 8330, 8340.

7. Posting, Document and Entry date column have date on 2020 to 2024.

VARIABLES	DESCRIPTION
Material	Unique identifier for the material.
Material Description	Description of the material.
Plant	Identifier for the plant.
Storage Location	Specific storage location within the plant.
Movement Type	Type of inventory movement
Special Stock	Indicator for special stock situations.
Material Document	Document number associated with the material movement.
Material Doc.Item	Item number within the material document.
Posting Date	Date when the transaction was posted in the system.
Qty in Un. of Entry	Quantity of the material in the unit of entry.
Unit of Entry	Unit of measure for the quantity.
Batch	Batch number.
Order	Order number associated with the transaction.
Reference	Reference number for the transaction.
Movement Type Text	Text description of the movement type.
Asset	Asset number.
Subnumber	Subnumber of the asset.
Counter	Counter for operations.

Routing number for operations	**Routing number for production operations.**
Document Date	**Date when the document was created.**
Qty in OPUn	**Quantity in order price unit.**
Order Price Unit	**Unit of measure for the order price.**
Order Unit	**Unit of measure for the order.**
Qty in order unit	**Quantity in the order unit.**
Company Code	**Identifier for the company.**
Valuation Type	**Type of valuation.**
Entry Date	**Date when the entry was made in the system.**
Time of Entry	**Time when the entry was made.**
Amount in LC	**Amount in local currency.**
Purchase Order	**Purchase order number.**
Smart Number	**Smart number.**
Item	**Item number within the purchase order.**
Ext. Amount in Local Currency	**Extended amount in local currency.**
Sales Value	**Sales value.**
Reason for Movement	**Reason for the movement.**
Sales Order	**Sales order number.**
Sales Order Schedule	**Sales order schedule line.**
Sales Order Item	**Item number within the sales order.**
Cost Center	**Cost center associated with the movement.**
Customer	**Customer number.**

Movement indicator	**Indicator of the movement type.**
Consumption	**Indicator of consumption.**
Receipt Indicator	**Indicator of receipt.**
Supplier	**Supplier number.**
Base Unit of Measure	**Base unit of measure for the material.**
Quantity	**Quantity in the base unit of measure.**
Material Doc. Year	**Year of the material document.**
Network	**Network number.**
Activity	**Activity number.**
WBS Element	**Work Breakdown Structure element .**
Reservation	**Reservation number.**
Item No.Stock Transfer Reserv.	**Item number in stock transfer reservation.**
Debit/Credit ind	**Debit or credit indicator.**
User name	**Username of the person who made the entry.**
Trans./Event Type	**Transaction or event type.**
Sales Value inc. VAT	**Sales value including VAT.**
Currency	**Currency of the transaction.**
Goods Receipt/Issue Slip	**Slip number for goods receipt or issue.**
Item automatically created	**Indicator if the item was automatically created.**
Original Line Item	**Original line item.**
Multiple Account Assignment	**Indicator for multiple account assignment**
Product Code	**Product code.**

Product Description	**Description of the product.**
Material Type	**Type of material.**
Vendor Code	**Code of the vendor.**

3.Data cleaning:

Data Cleaning: Clean the data to handle missing values, outliers, and inconsistencies. This step involves tasks like imputation, removal of duplicates, and normalization.

4.Calculating measures of central tendency:

Variables	Mean values	Median values	mode
Material	3.102301e+07	3.102697e+07	31033204
Plant	8.126882e+03	8.110000e+03	8110
Movement Type	1.760058e+02	2.010000e+02	201
Vendor code	5.430134e+04	5.460500e+04	50115
Qty in Unit of entry	3.421237e+00	-1.000000e+00	-1
Qty in OPUn	3.111934e+01	0	0
Qty in order unit	3.111846e+01	0	0
Amount in LC	5.813507e+03	0	0
Purchase Order	4.421518e+09	4.400064e+09	4.7e+09
Quantity	3.423044e+00	-1.000000e+00	-1
Material document year	2.021919e+03	2.022000e+03	2023

5. Calculating measures of Dispersion:

Variables	Variance	Standard Deviation	Range
Material	9.031075e+09	9.503197e+04	1082091
Plant	2.927207e+03	5.410367e+01	230
Movement Type	5.126033e+03	7.159632e+01	540
Vendor code	6.430299e+06	2.535803e+03	10022
Qty in Unit of entry	4.586482e+05	6.772357e+02	68394
Qty in OPUn	2.236157e+05	4.728802e+02	26020
Qty in order unit	2.236157e+05	4.728802e+02	26020
Amount in LC	5.949191e+10	2.439096e+05	15443063
Purchase Order	6.271200e+15	7.919091e+07	799999345
Quantity	1.200206e+00	6.772356e+02	68394
Material document year	2.021919e+03	1.095539e+00	4

6. Calculating measures of ASSYMETRY:

Variables	Skewness
Material	-9.964307
Plant	3.088983
Movement Type	0.517433
Vendor code	-0.177573
Qty in Unit of entry	-1.245901
Qty in OPUn	36.168107
Qty in order unit	36.168095
Amount in LC	0.989753
Purchase Order	3.299860
Quantity	-1.245909
Material document year	-0.175085

7. Calculating measures of PEAKNESS:

Variables	Kurtosis
Material	104.539789
Plant	7.786756
Movement Type	-0.568004
Vendor code	-1.042748
Qty in Unit of entry	923.162268
Qty in OPUn	1456.782037
Qty in order unit	1456.781365
Amount in LC	109.009448
Purchase Order	9.665062
Quantity	923.162467
Material document year	-0.913627

Questions

1. Questions on Material Movements What is the total quantity of material received (goods receipt)?
2. Which plant has the highest number of materials?
3. Which document year has the highest materials and how many percentage materials they have?
4. What is the average quantity of material moved per transaction in 2022?
5. Questions on Financial Impact What is the total financial value of material received in 2024?

General Questions

1. How many different movement types are there in the dataset?
2. What are the different storage locations used in the dataset?
3. What is the earliest and latest posting date in the dataset?

Describe Visualization and how it answers the questions

Conclusion:

1 Question

201	101
24,817	27,967
40.3%	45.4%

2 Question

8110	55,298	90%

Plant 8110 has received the highest amount of material. Which is 55,298 that is 90% of total Material received.

3 Question

2023 – 31.58 %

2023 document year has the highest materials and they have 31.58% .

4 Question

5 Question

The total financial value of material received in 2024 is 2bn.

General Questions

1 Question

There are 11 movement types in this dataset.
(101,102,161,201,261,301,309,311,321,641)

2 Question

UTL, RD01, BOIS, HVAC, CCPJ, EQSP, 4017, BPRJ, CC01, CC03, CCP1, CCPJ, CN01, CN02, DEN1, DENG, DIN1, DINS, EEN1, ELS1, ELSP, ENO1, ENO2, EQS1, EQSP, GEN, GENE, HKS1, HKST, HVA1, HNAC, INS1, INSP, OIN1, OINS, PRCN, QCC1, QCCN, RD01, RDC1, RDCN, SHE1, SHEE, SS15, TTE1, UTL1, WHSL, WTR1.

3 Question

1 April 2020 is the earliest posting date in the dataset.

30 March 2024 is the latest posting date in the dataset.

Quantity Impact:

- There is a relationship between the quantity of materials requested in the tender and the corresponding tender prices, with larger quantities potentially leading to lower prices per unit.
- Suppliers may adjust their pricing based on the volume of materials being procured, offering discounts for larger quantities.

Comparison with Basic Price:

- There may be discrepancies between the basic prices listed in the tender documents and the actual awarded tender prices, indicating potential cost overruns or savings.
- Analyzing the differences between basic prices and awarded prices can help identify areas for cost optimization or negotiation.
- These conclusions provide insights into the dynamics of tender price analysis and can inform decision-making processes related to procurement, supplier management, and project planning.

Visualisation:

HISTOGRAM:

A histogram in Python is a graphical representation of the distribution of a dataset. It displays the frequency of different ranges (bins) of values in the data. Histograms are particularly useful for understanding a dataset's distribution, spread, and central tendency.

BOX PLOT:

A boxplot, also known as a box-and-whisker plot, is a standardized way of displaying the distribution of data based on a five-number summary: minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum. Boxplots can also highlight outliers and help compare distributions across different categories.

Q-Q PLOT:

A Quantile-Quantile (Q-Q) plot is a graphical tool to help you assess if a dataset follows a particular distribution, such as the normal distribution. It plots the quantiles of your data against the quantiles of a theoretical distribution (e.g., the normal distribution). The data will likely follow the specified distribution if the points form an approximately straight line.

DISTPLOT:

`Distplot` is a function from the Seaborn library used to plot the distribution of a univariate set of observations. It combines a histogram and a kernel density estimate (KDE) or fitted probability density function.

However, note that `distplot` has been deprecated in recent versions of Seaborn (since version 0.11.0). It is recommended to use `histplot` or `displot` instead. Below, I'll show how to use the deprecated `distplot` and the recommended `histplot`.

SCATTER PLOT:

A scatter plot is a type of data visualization displaying values for typically two variables for a data set. It is used to observe relationships between variables. Each point on the scatter plot represents an observation.

CONCLUSION:

The exploratory data analysis (EDA) and descriptive statistics of the pharmaceutical inventory dataset reveal critical insights into material movements and financial impacts. Plant 8110 emerged as the primary receiver of materials, accounting for 90% of the total inventory. The year 2023 saw the highest material activity, representing 31.58% of the total transactions. Additionally, the analysis highlighted 11 distinct movement types and various storage locations. Financially, the total value of materials received in 2024 amounted to 2 billion, underscoring the significant economic scale of operations. These findings provide a comprehensive understanding of inventory dynamics and financial performance, guiding strategic decision-making and resource allocation.