Capstone Project

Name: Ekta Rani
Bootcamp: Data Analyst Bootcamp
Workshop ID: DAT-V862-12023266/DA BATCH AUGUST
email:ektarani7909@gmail.com
Mobile: 7011563747

Project Title : **Boxify : Sales Analysis and Inventory Insights**

Summary:

As a data analyst tasked with delving into the sales Analysis and Inventory Insights, my objective is to conduct a comprehensive examination of a retail firm has many products in their inventory, and very few of them tend to sell (only about 10% sell each year) and many of the products only have a single sale in the course of a year

The sales and growth team of the retail firm wants to determine which products from their inventory should they retain to sell and the ones to discard.

To initiate this analysis, it is imperative to consider the variety of products from inventory.

Data Analysis Approach and Methodology:

Technical Aspects:

Language Used : Python is a popular programming language widely used in the field of data analysis. Its simplicity and versatility make it an excellent choice for handling and manipulating data .Python's ecosystem, combined with its readability and ease of learning, has contributed to its dominance in data analysis and has made it a preferred language for data scientists and analysts.

Libraries Used : **Pandas** is a powerful library for data manipulation and analysis. It provides data structures like DataFrames and Series, making it easy to handle and analyze structured data.

Matplotlib is a widely-used plotting library for creating static, interactive, and animated visualizations in Python.

Seaborn is a statistical data visualization library built on top of Matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

Plotly Express: Plotly Express is a high-level interface for creating interactive visualizations. It is built on top of the Plotly library and simplifies the creation of complex plots

Data Analysis Approach and Methodology:

- ✔ **Module imports :** Importing necessary imports to perform all activities for the analysis
- ✔ **Data Collection:** Acquiring the data given contains both historical sales data AND active inventory
- ✔ **Loading the Data:** Using pandas to read and load the dataset into a structured format for analysis.
- ✔ **Data Preprocessing:** Cleaning and organizing the dataset, addressing missing values and converting data types.
- ✔ **Inspecting the Data**: Examining the dataset's structure and initial content to identify patterns or outliers.
- ✔ **Finding Null Values:** Identifying and handling null values to ensure data completeness.
- ✔ **Finding Duplicates**: Detecting and managing duplicate entries to maintain data integrity**.**
- ✔ **Getting Descriptive Statistics:** Calculating descriptive statistics to gain insights into the numerical aspects of the dataset.

Data Analysis Approach and Methodology:

- ✔ **Figuring Out the Distribution:** Analyzing the distribution of variables within the dataset to understand the spread and central tendencies of the data.
- ✔ **Cleaning the Outliers:** Identifying and addressing outliers in the dataset to enhance the accuracy and reliability of subsequent analyses.
- ✔ **Plotting Correlation:** Creating correlation plots to visualize relationships between different nutritional components, providing insights into potential dependencies.
- ✔ **Data Visualization**: Employing data visualization techniques, such as histograms and box plots, to present a comprehensive overview of the content and identify patterns.
- ✔ **Nutrition-Based Insights:** Deriving meaningful insights from the data, including trends and patterns related to nutritional components, to inform decision-making.
- ✔ **Interactive Charts on salesInfo:** Utilizing interactive charting tools or libraries to create dynamic visualizations that allow users to explore product sales information interactively, enhancing engagement and understanding.

Exploratory data analysis findings and Insights

- **Size of the Dataset :** The dataset has 198918 entries (rows) and 14 columns.

- **Column Information :** There are three data types present in the DataFrame : float64, int64, and object (presumably strings or mixed types).

- **Numeric Columns Statistics :** Descriptive statistics for numeric columns (count, mean, std, min, 25%, 50%, 75%, max) are stating that there might be some severe outliers on all the nutritional columns

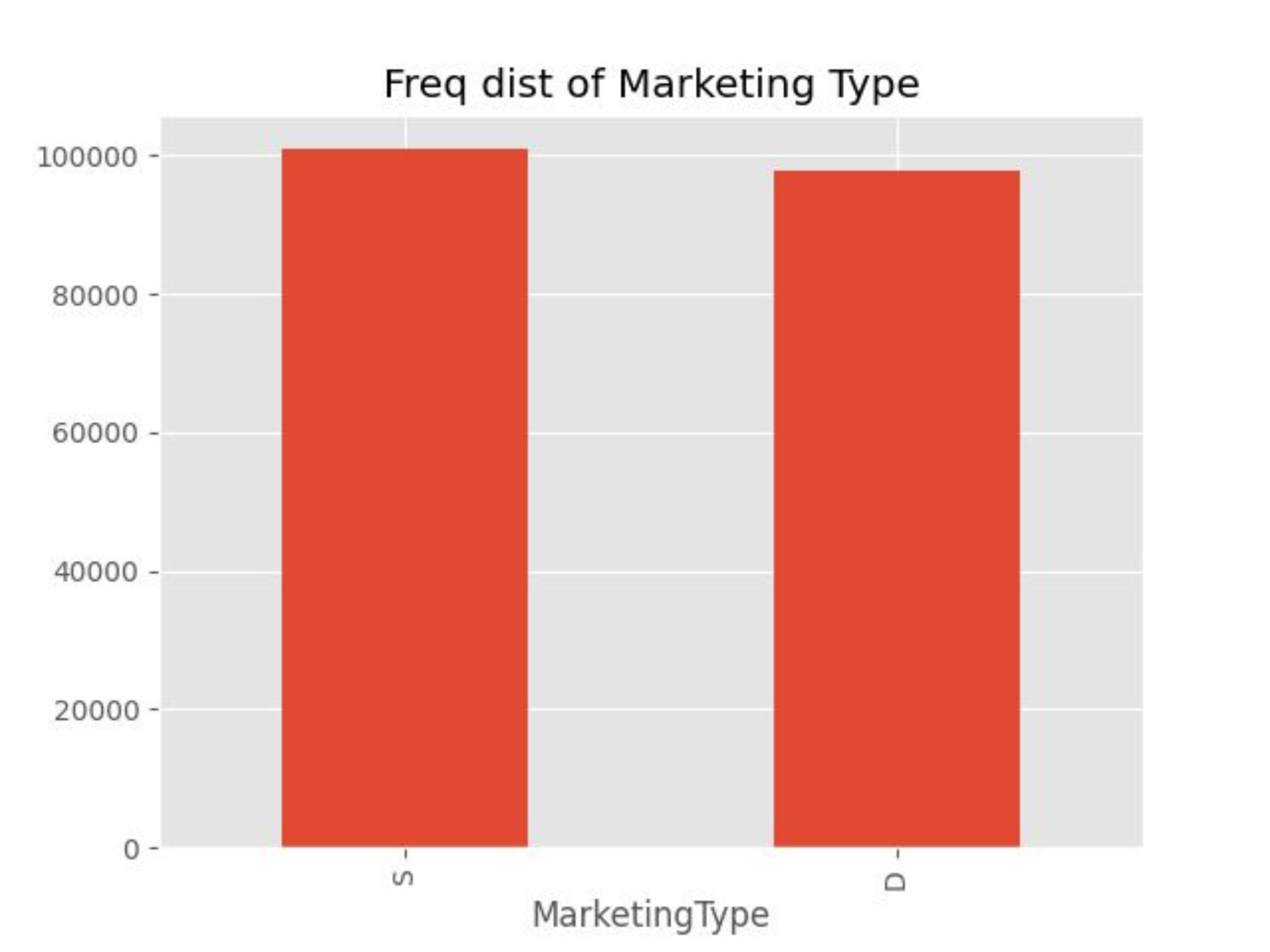
- **Categorical Columns :** The file type columns are of the object data type, indicating they are likely categorical or text data.

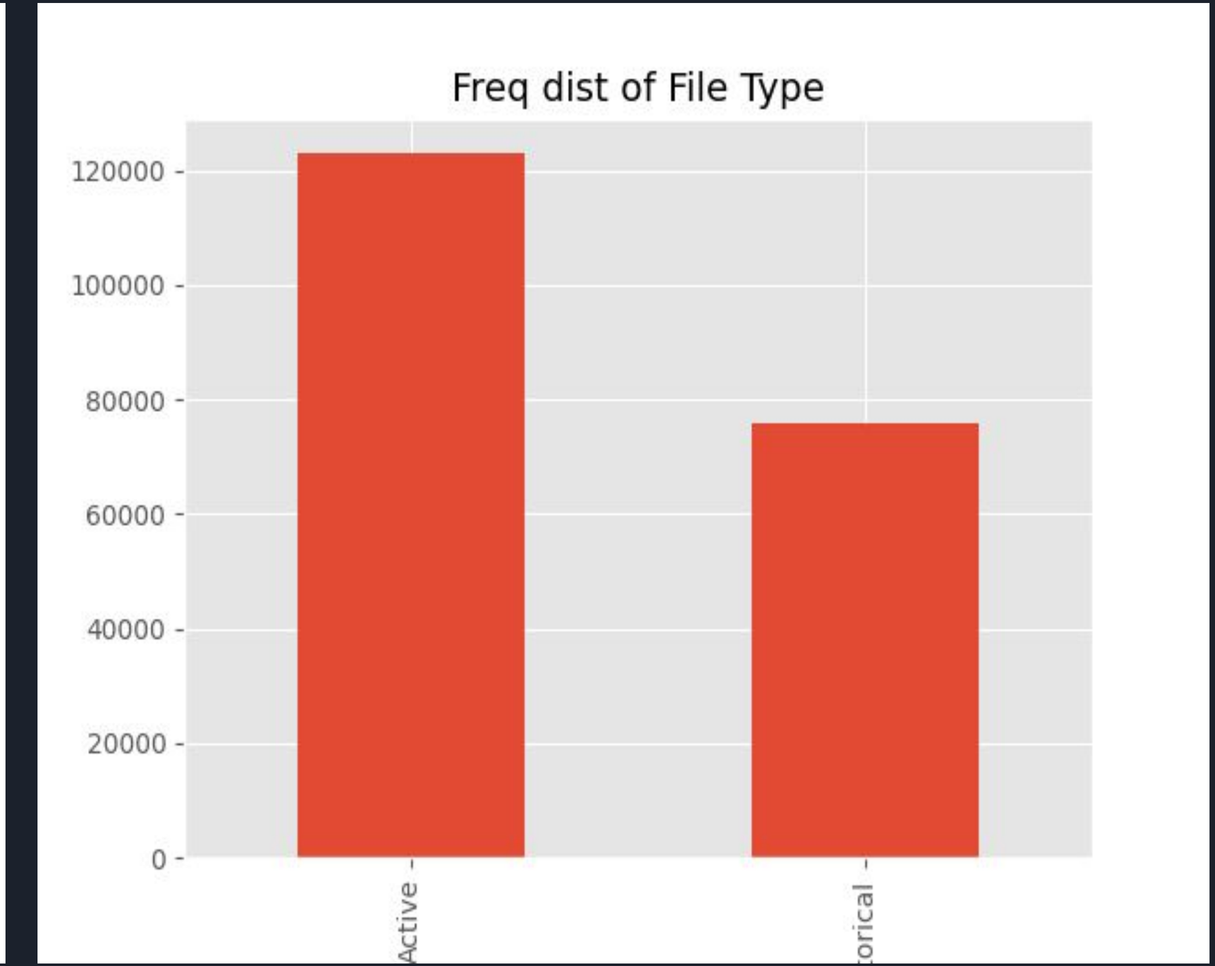
- **Inventory Information :** Columns like SKU_number,SoldFlag,SoldCount, Marketing type contains information about products sales and storage.

- **Data Completeness :** There are no missing values (Non-Null Count is 260 for all columns) and no duplicate values, indicating that the dataset is complete in terms of non-null entries.

Exploratory data analysis findings and Insights

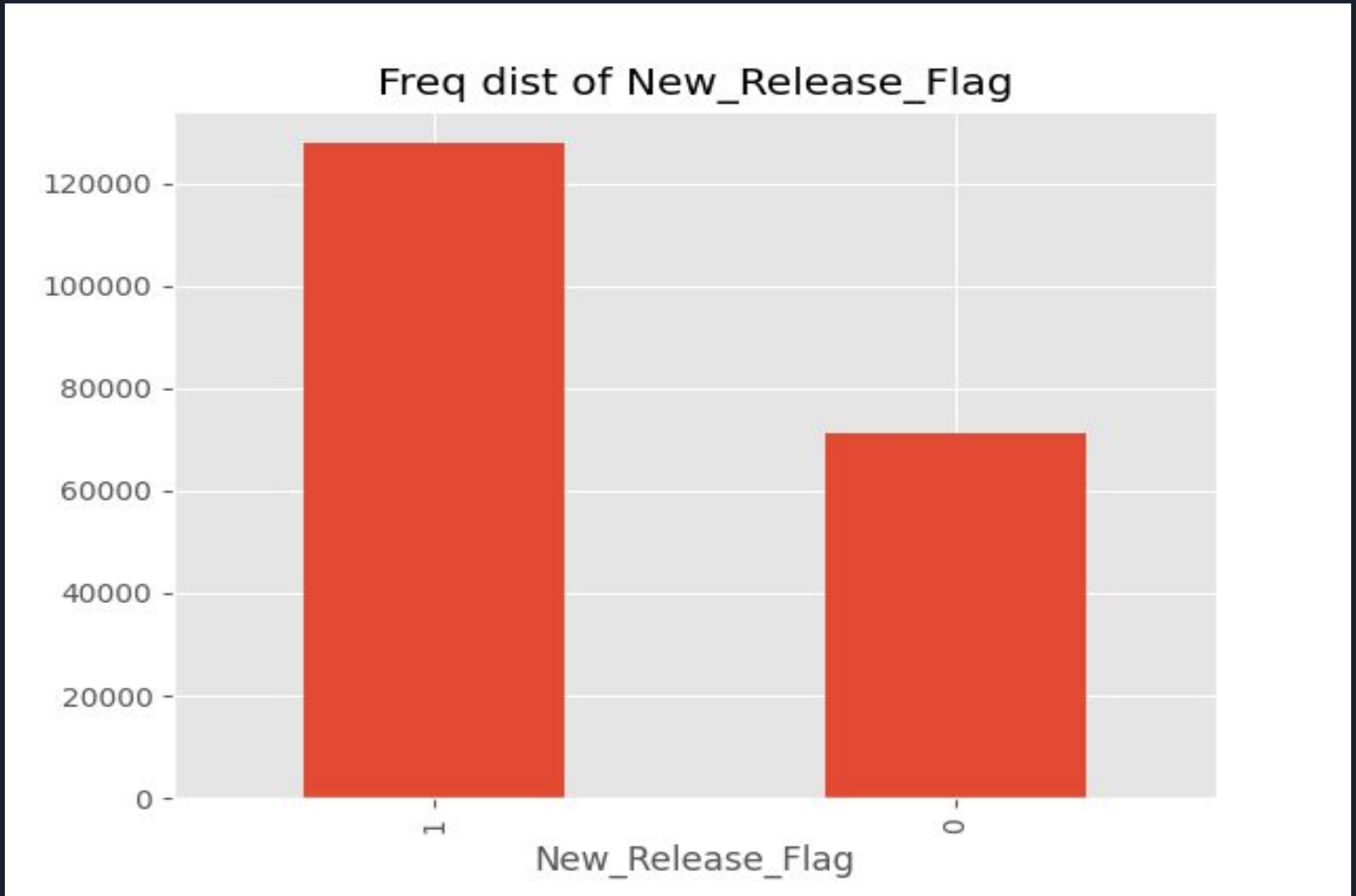
Figuring out the Shows the frequency distribution of the difference factors with a bar chart:





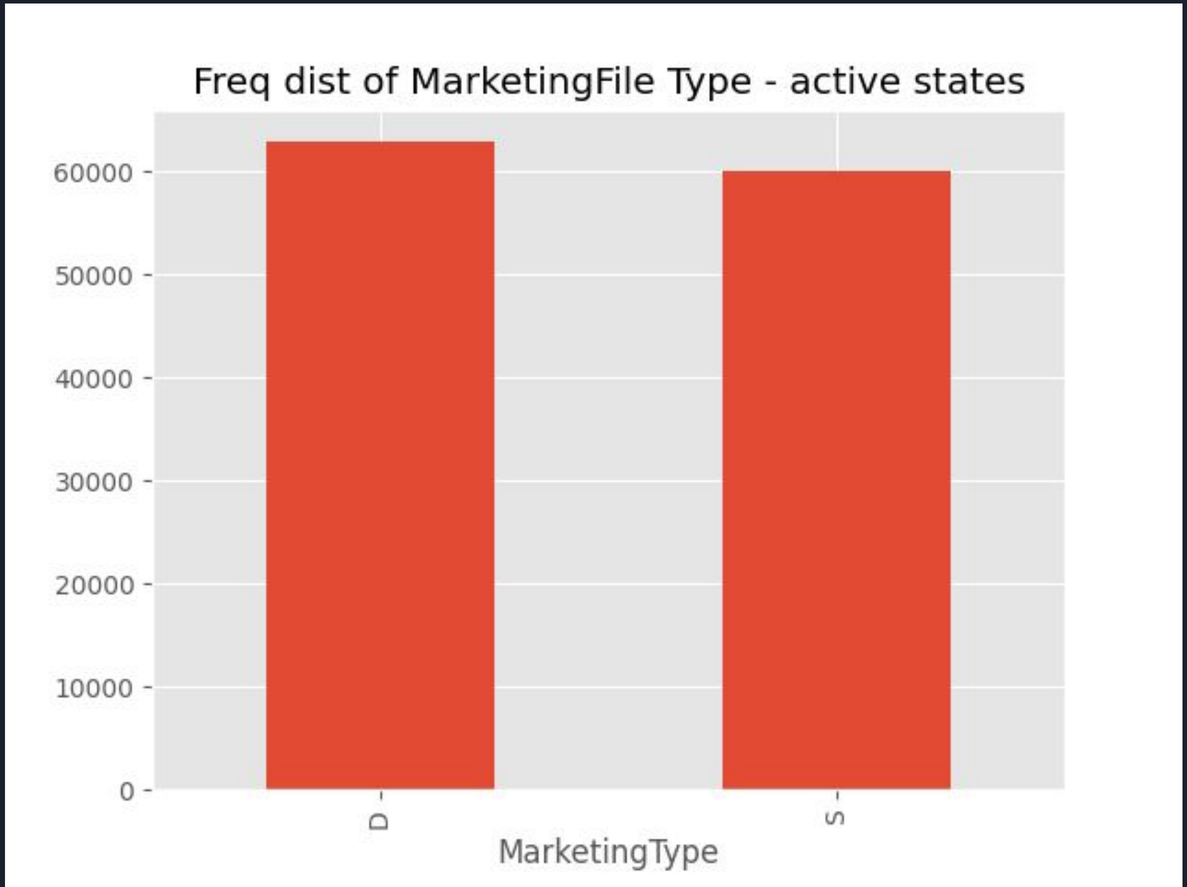
Exploratory data analysis findings and Insights

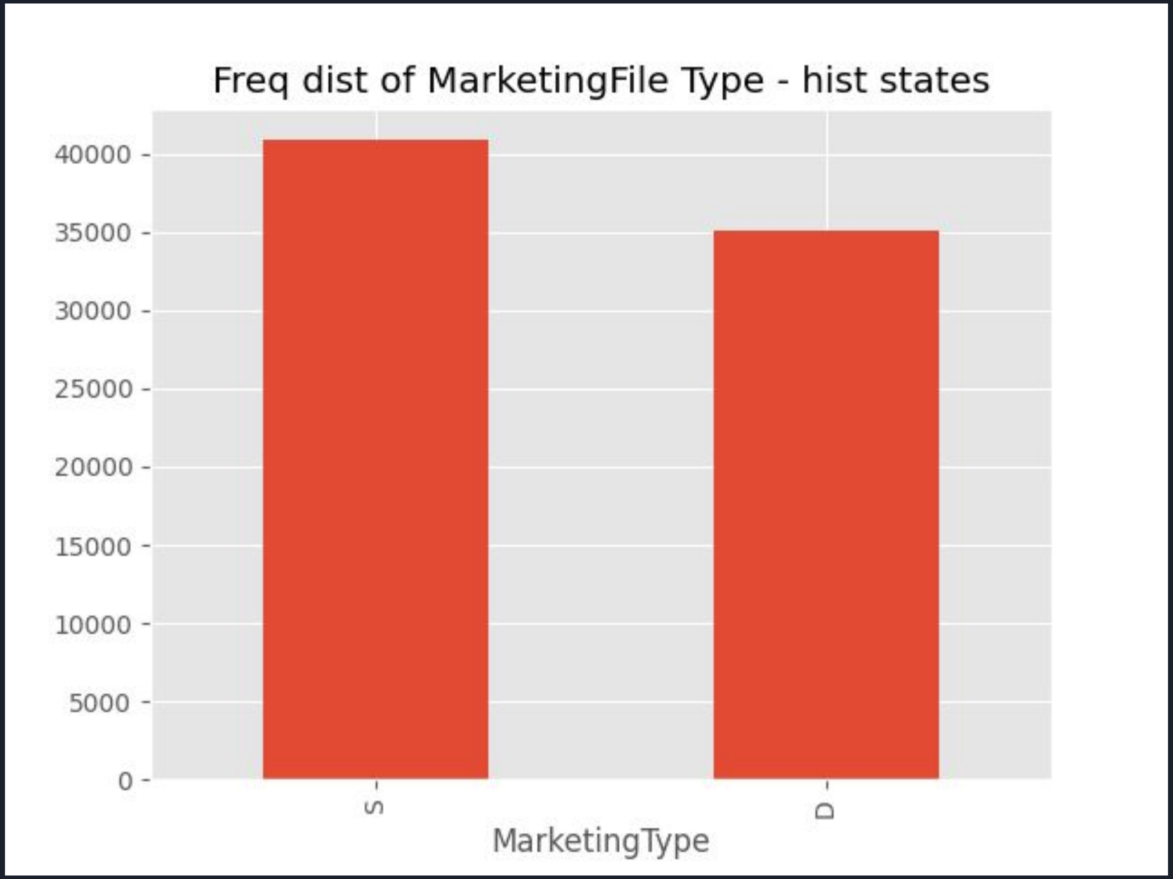
Finding new product inventorywith a Bar chart:



Exploratory data analysis findings and Insights

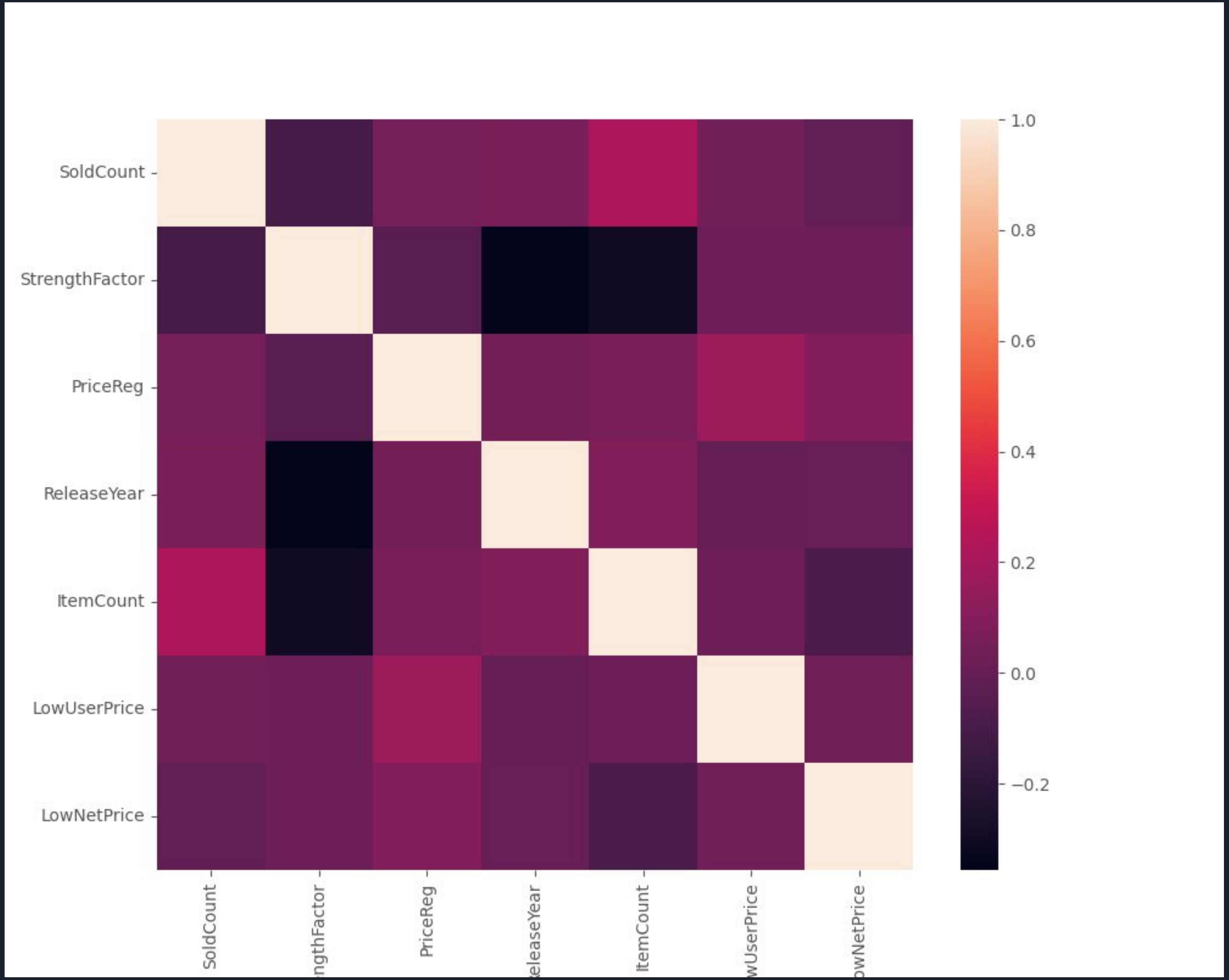
Figuring out the Shows the frequency distribution of the active states and hist states with a bar chart:





Exploratory data analysis findings and Insights

Correlation Plot:

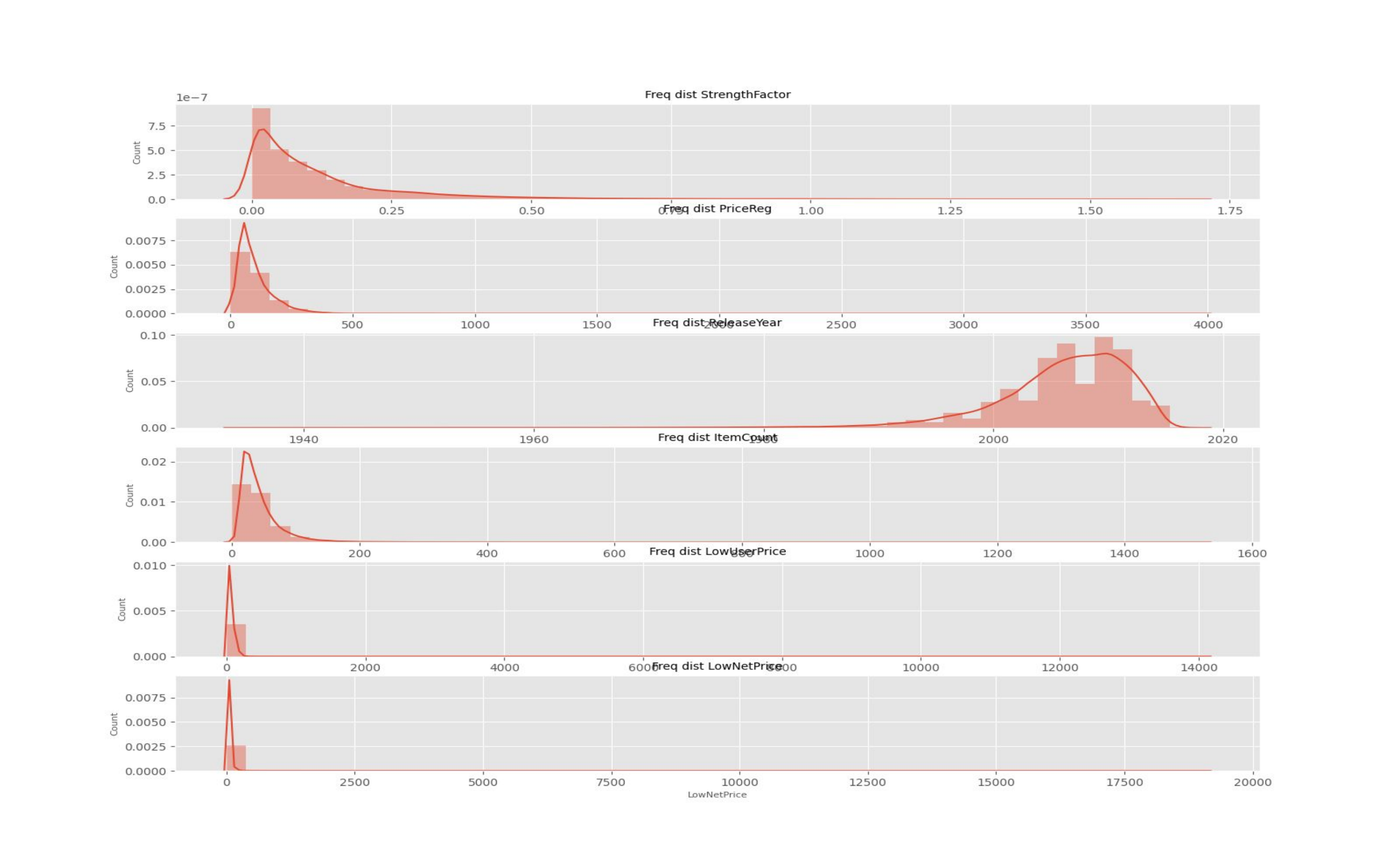


The correlation coefficients are proportionate to the box's size and color intensity.

Positively correlated variables will have correlation value close to +1 and negatively correlated variables will have correlation value close to -1.

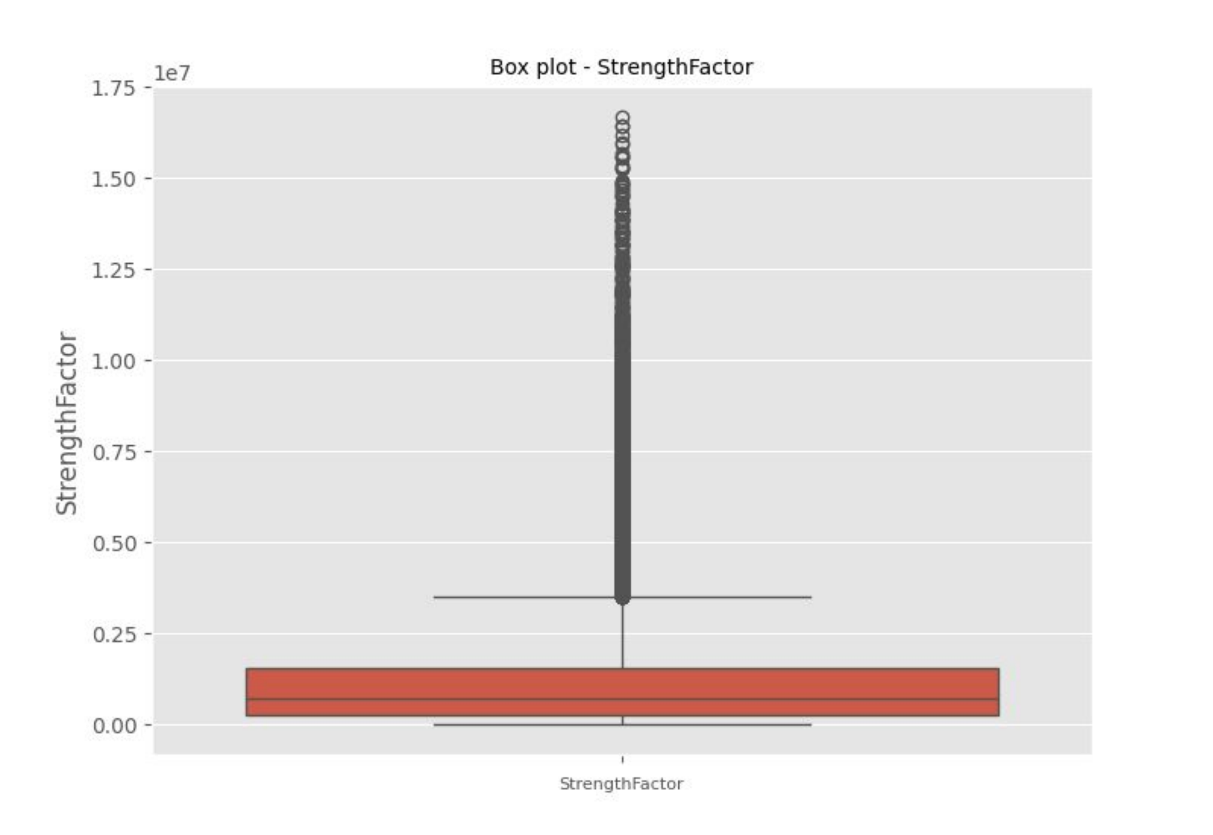
Data Visualizations :

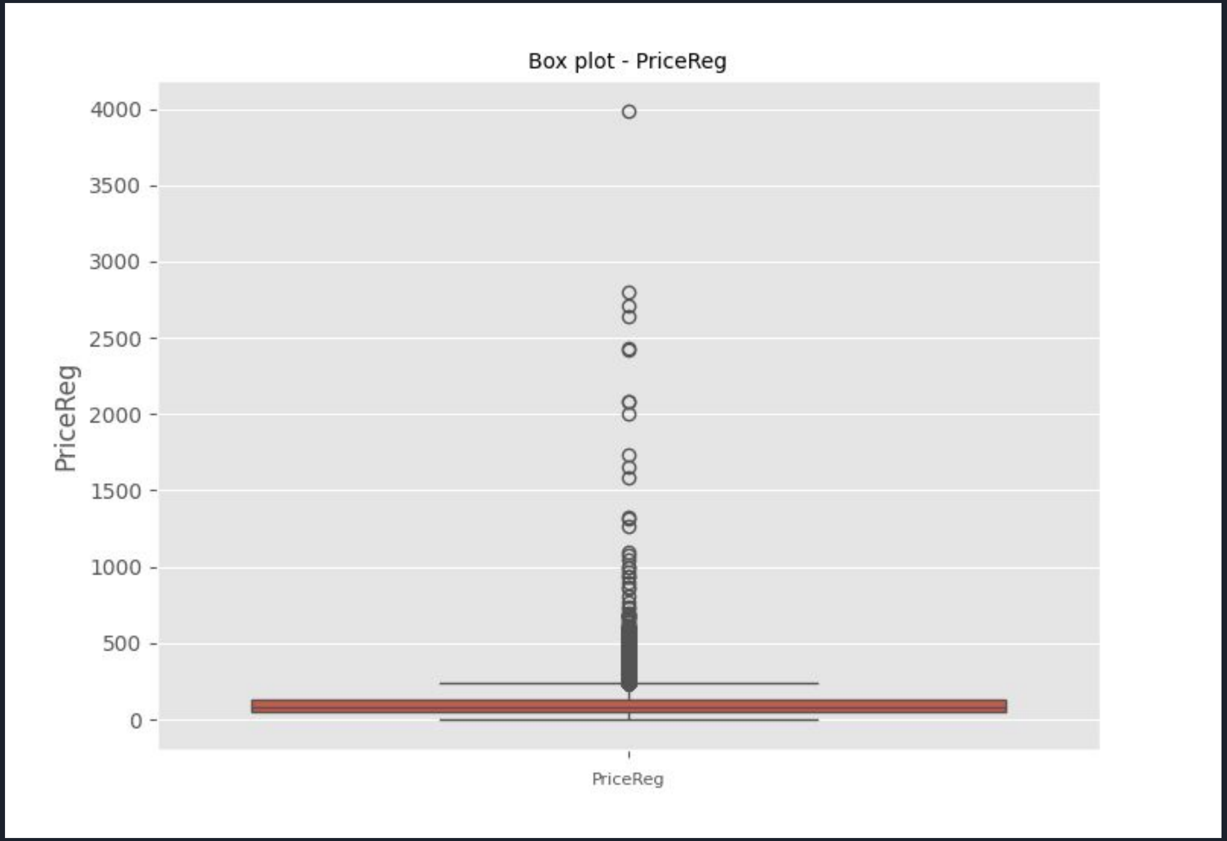
Plots with a kernel density estimate and histogram with bin size determined automatically



Data Visualizations :

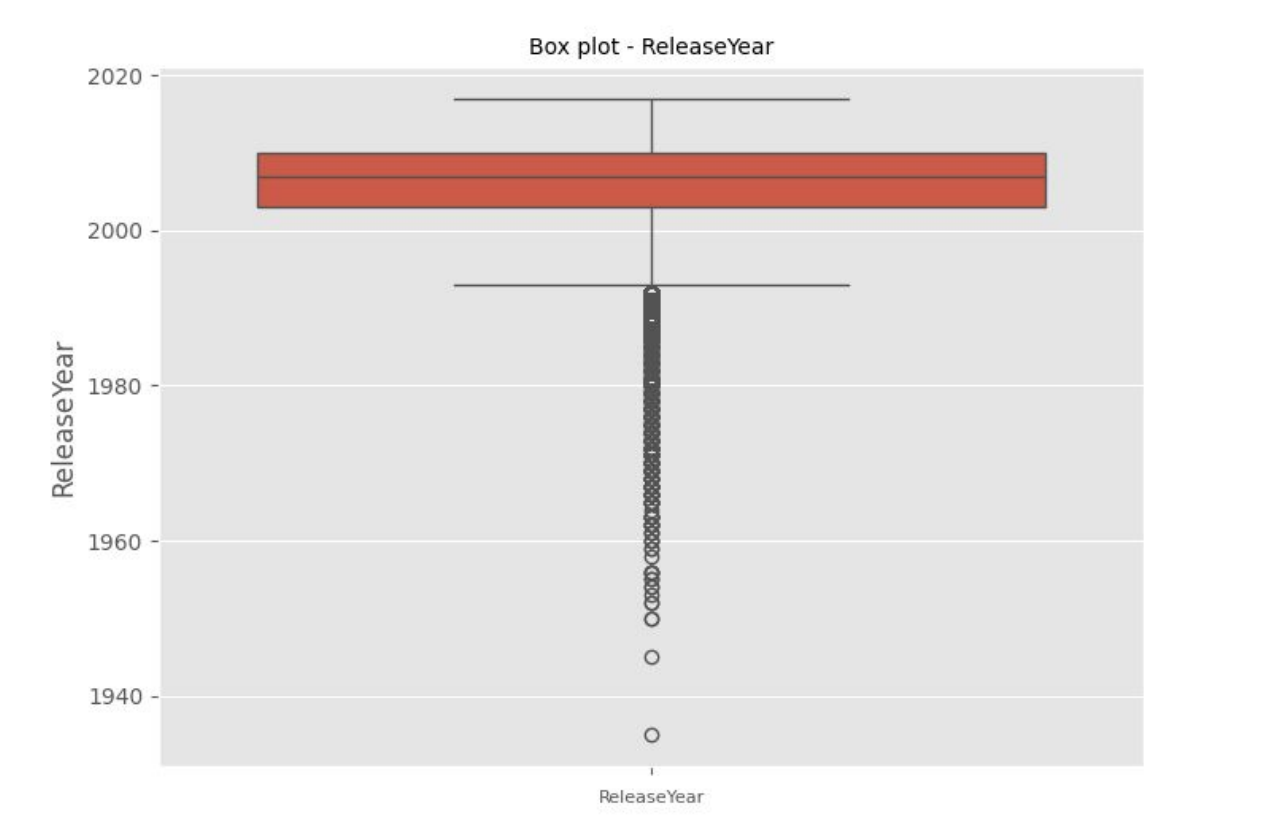
To analysis the outliers in the numeric features of the dataset:

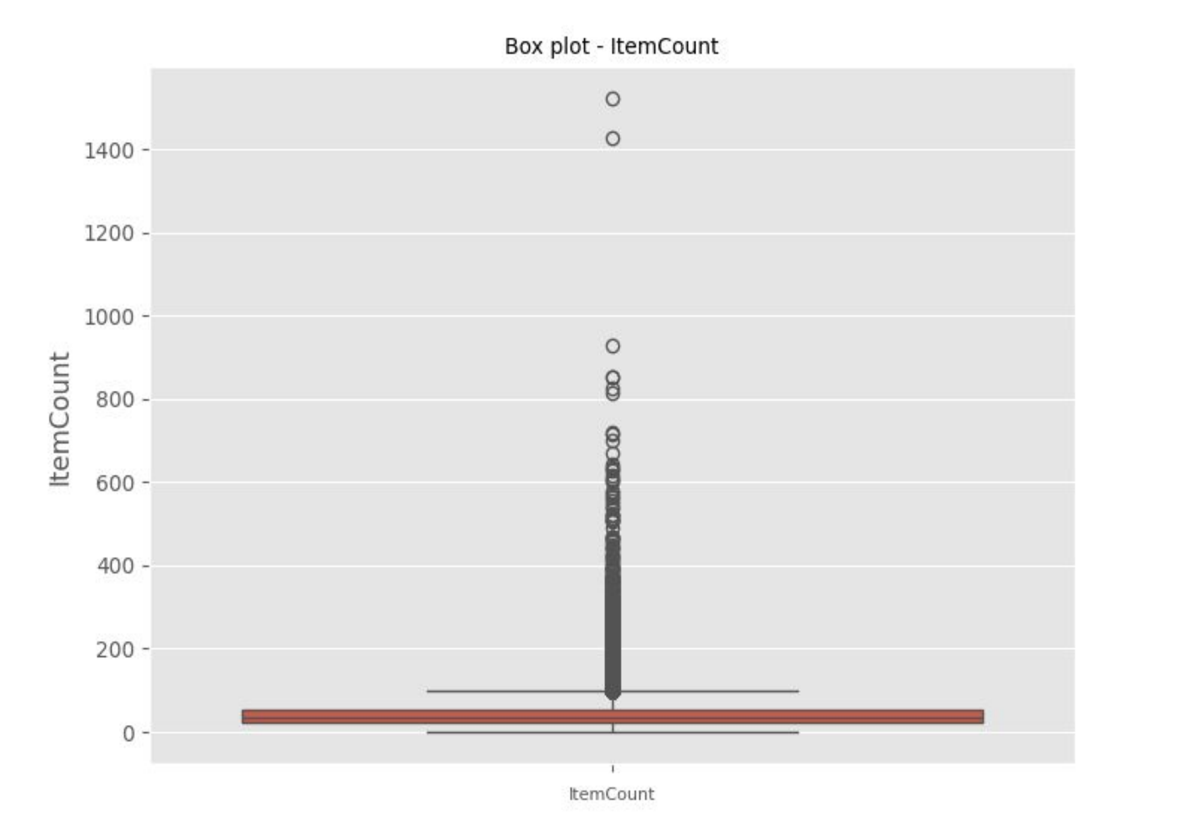




Data Visualizations :

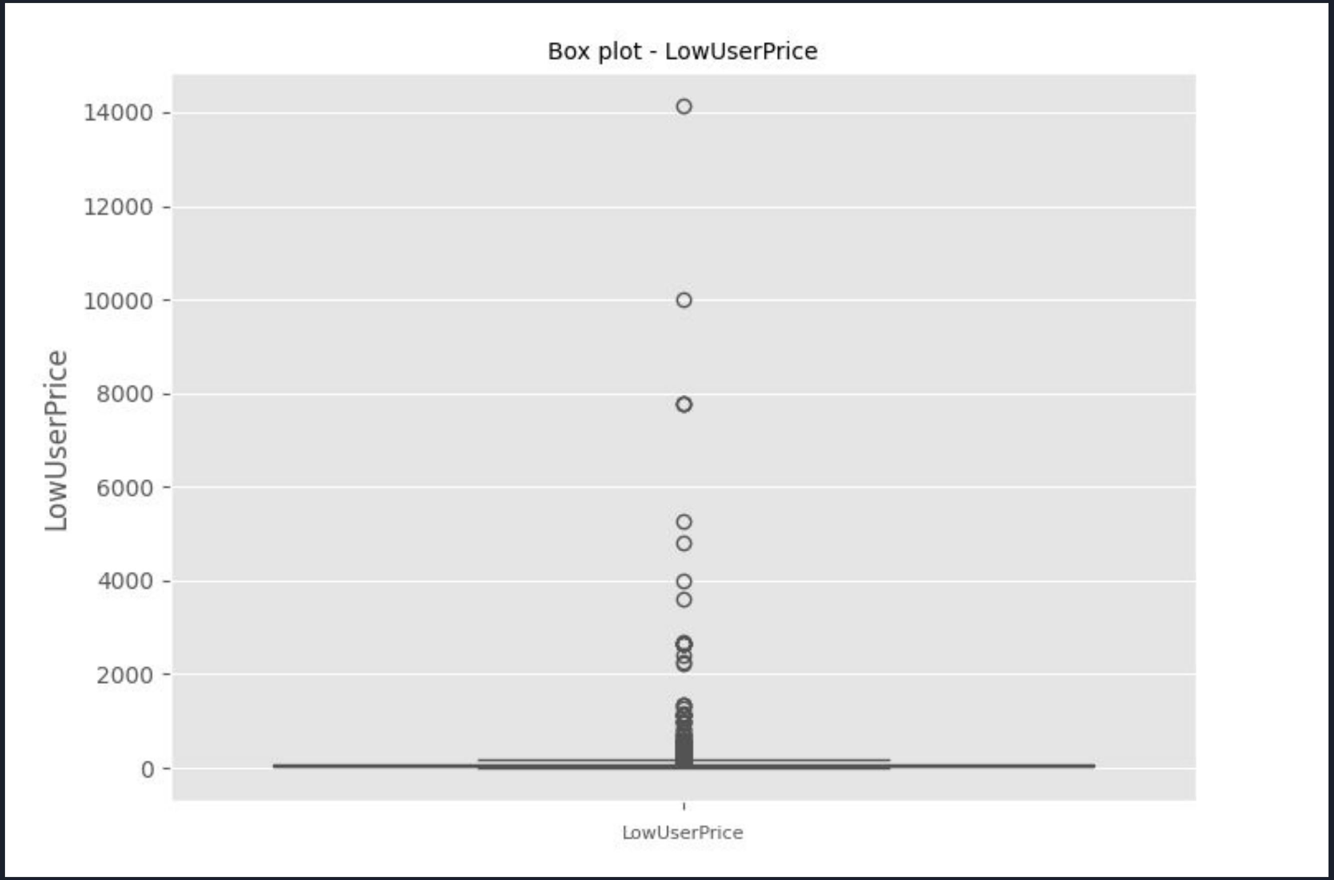
To analysis the outliers in the numeric features of the dataset :

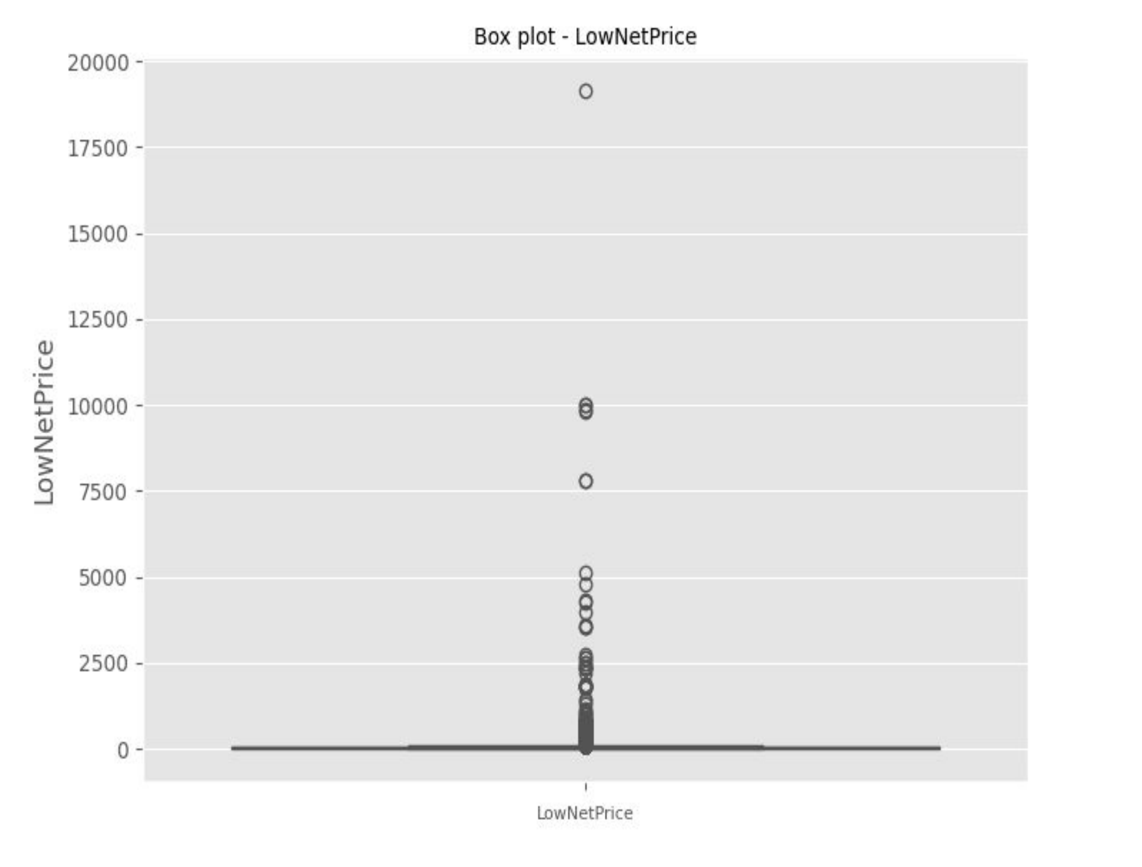




Data Visualizations :

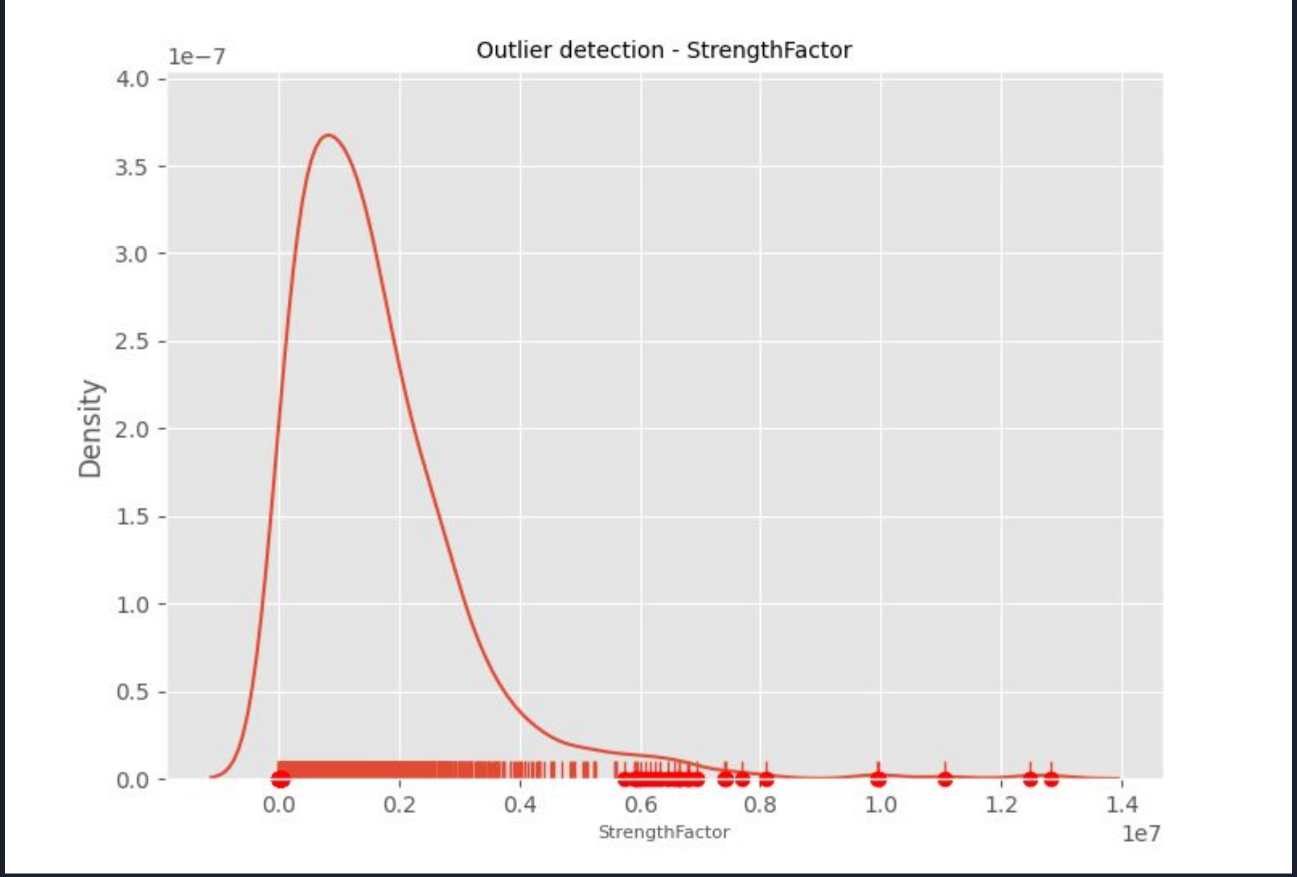
To analysis the outliers in the numeric features of the dataset :

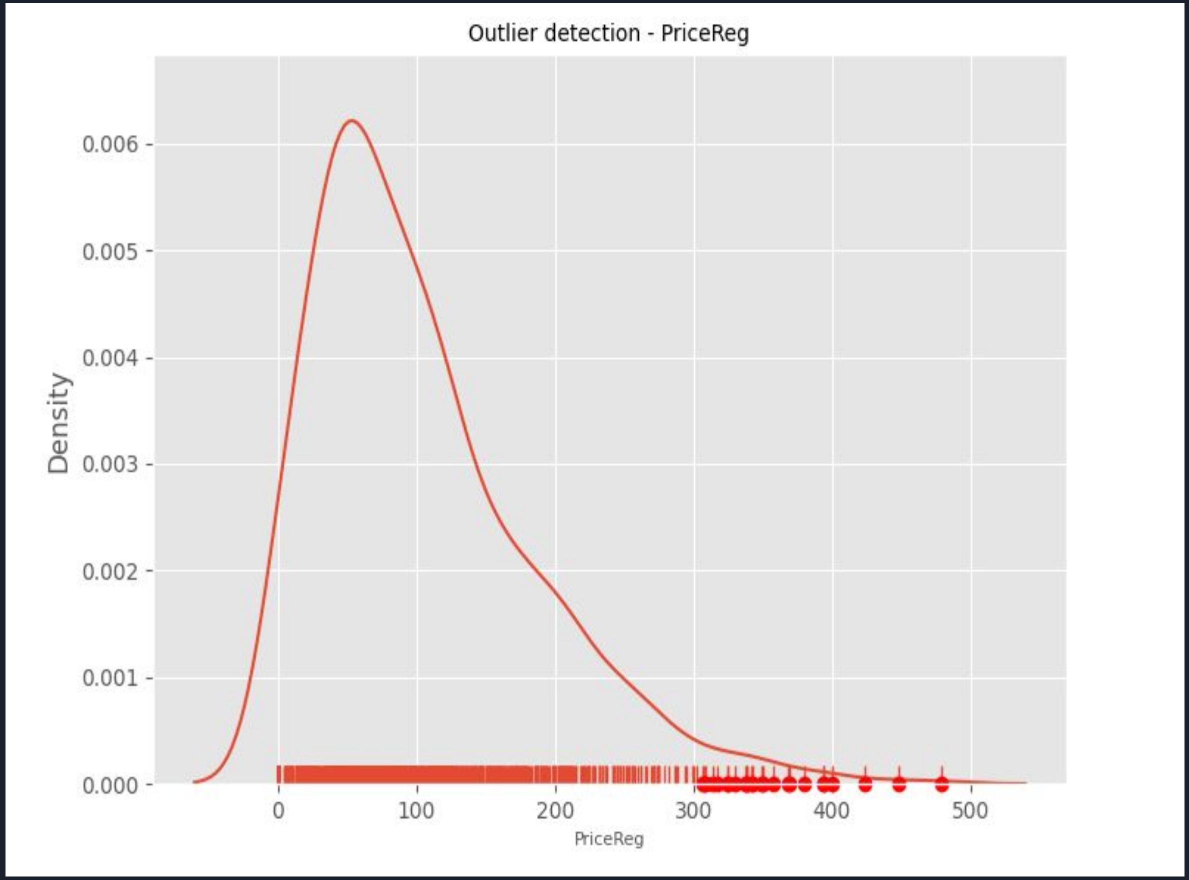




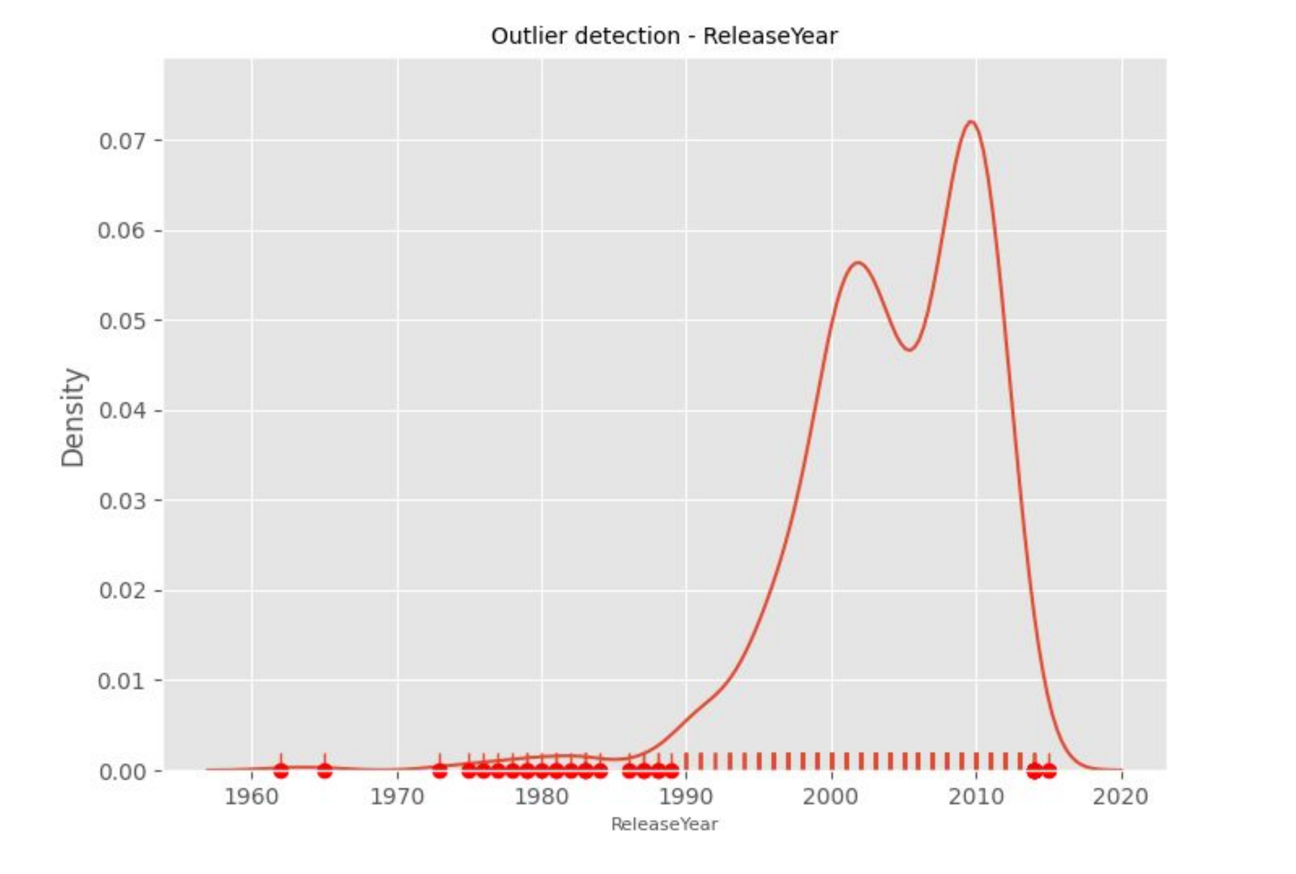
Data Visualizations :

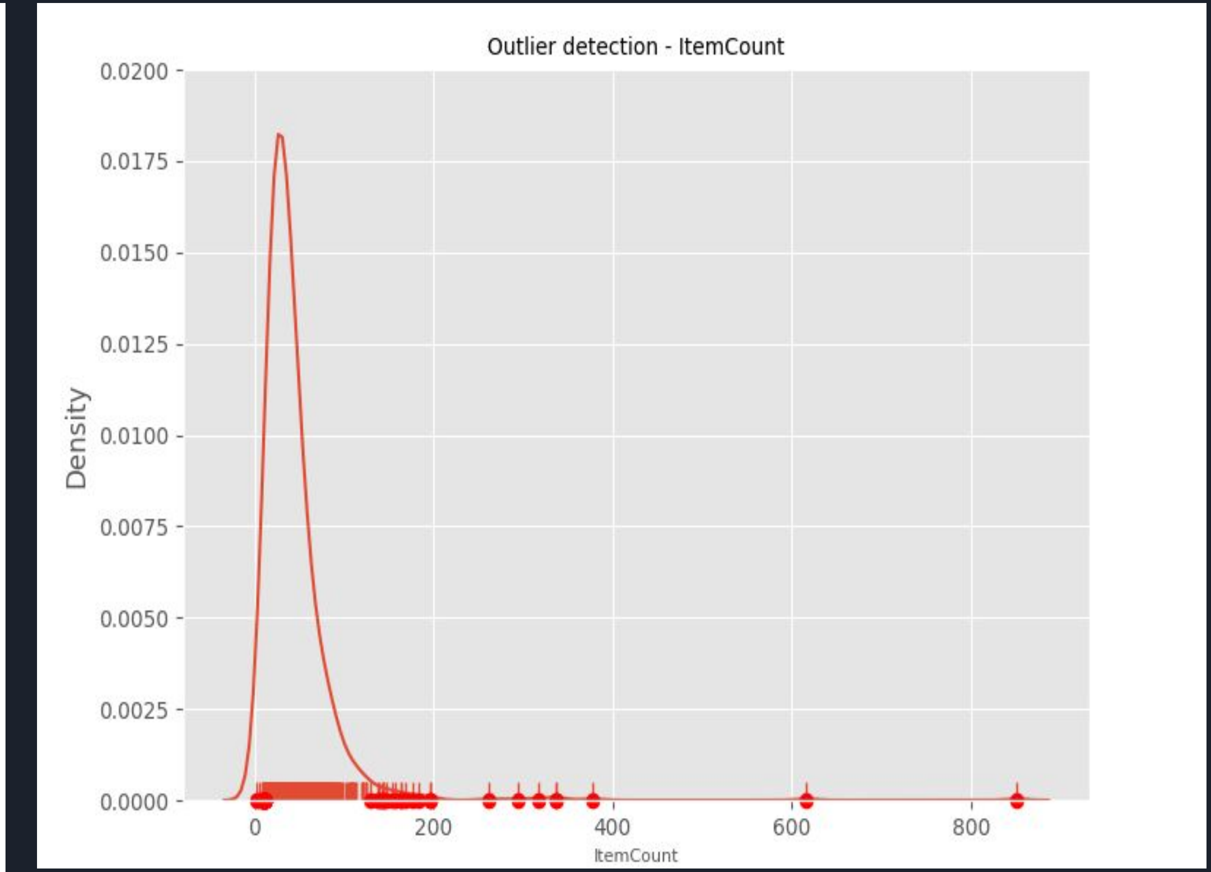
Outliers in input data can skew and mislead the results and make results less reliable, that's why we have to recognize all the outliers and treat them:



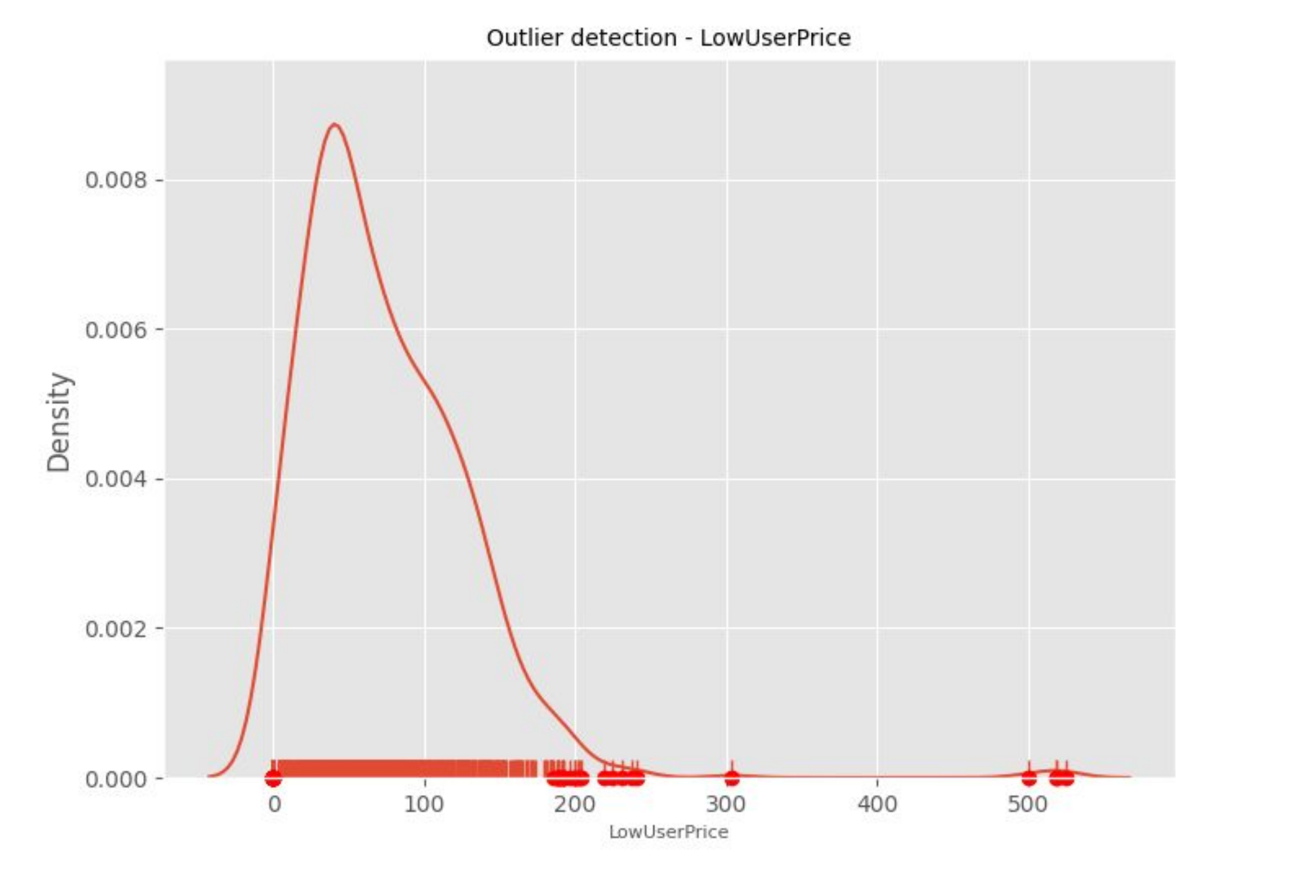


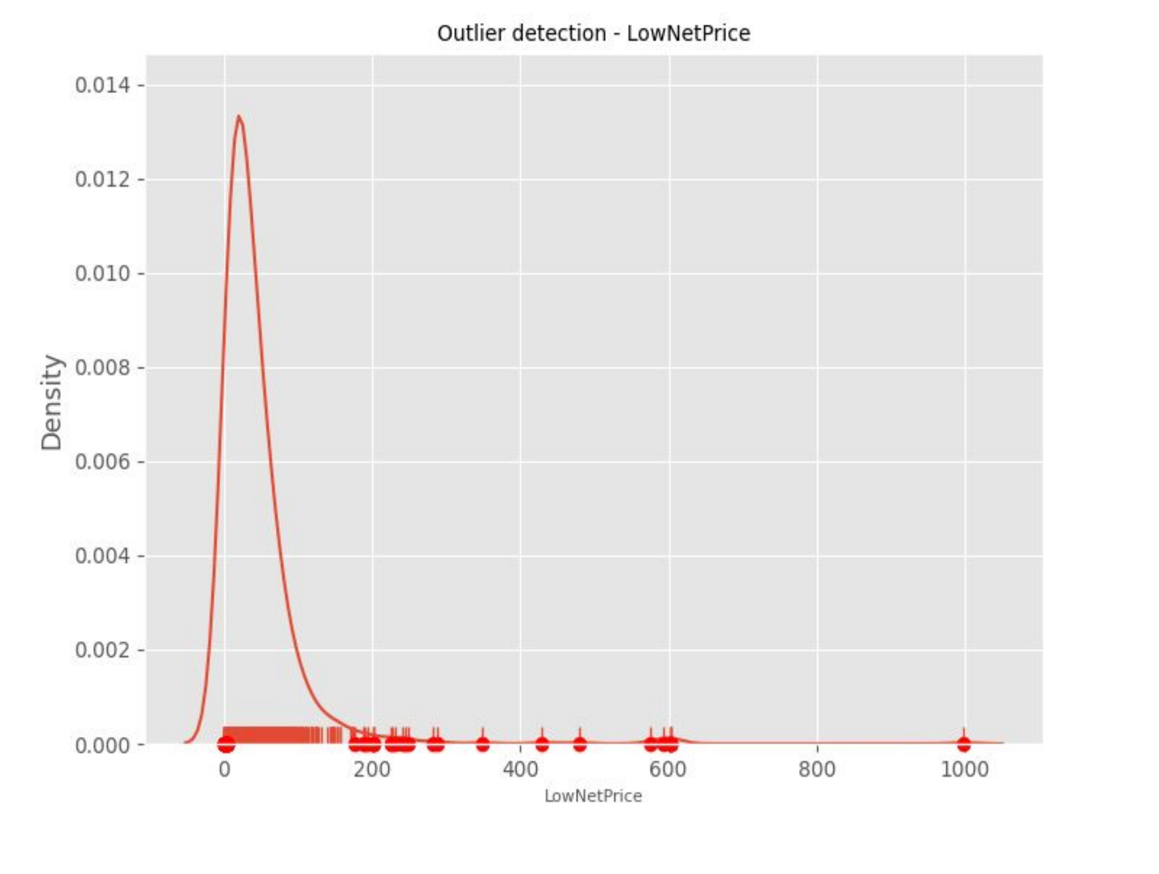
Data Visualizations :





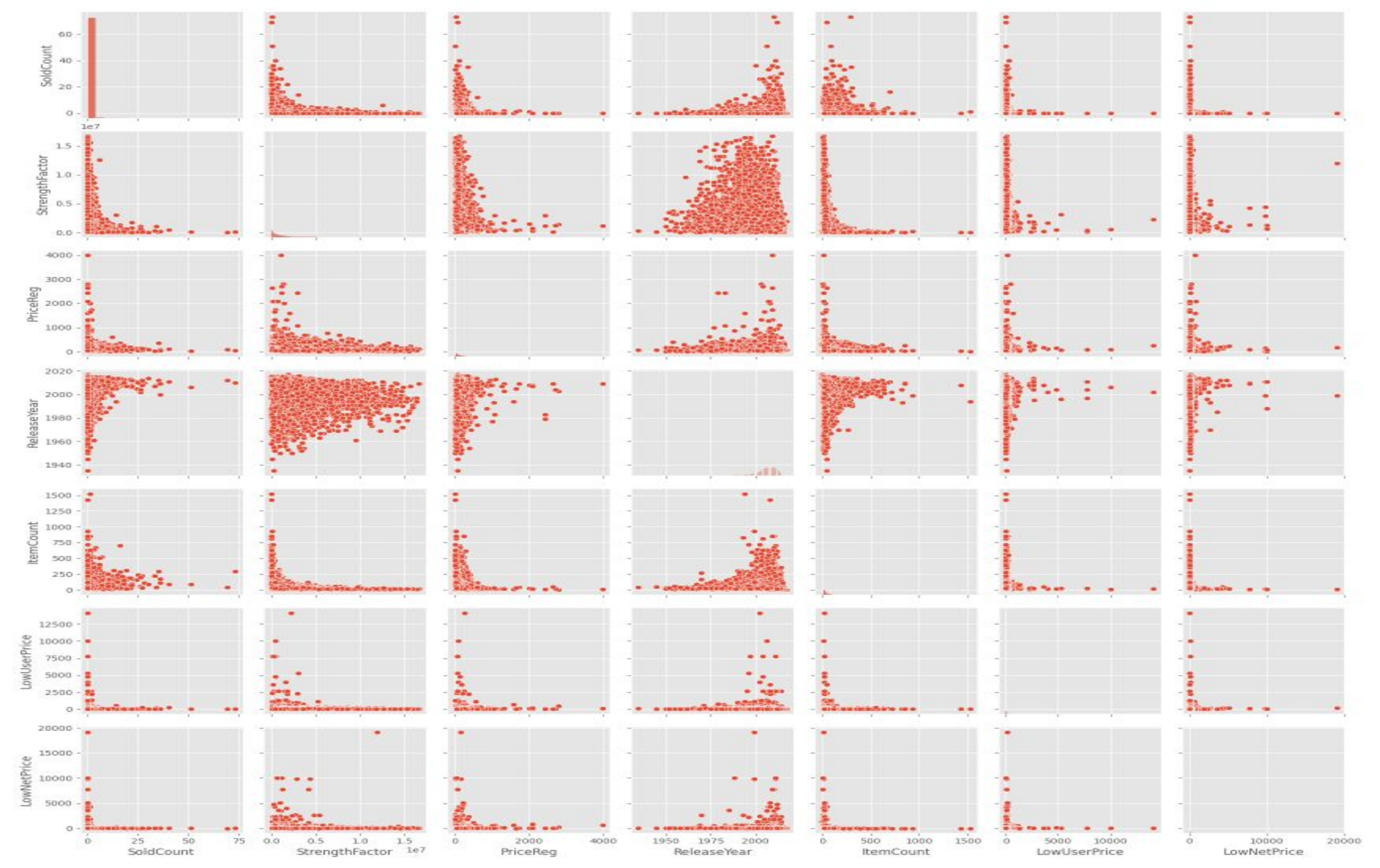
Data Visualizations :





Bivariate plots

To plot multiple pairwise bivariate distributions in a dataset:



Summary:

- The file contains historical sales data (identified with the column titled File_Type) along with current active inventory that is in need of evaluation (i.e., File Type = "Active"). The historical data shows sales for the past 6 months. The binary target (1 = sale, 0 = no sale in past six months) is likely the primary target that should drive the analysis.

- The other columns contain numeric and categorical attributes that we deem relevant to sales

- Some of the historical sales SKUs are also included in the active inventory.

- A few comments about the attributes included, as we realize we may have some attributes that are unnecessary or may need to be explained.

Summary:

- SKU_number: This is the unique identifier for each product.
- Order: Just a sequential counter. Can be ignored.
- SoldFlag: 1 = sold in past 6 mos. 0 = Not sold
- Marketing Type = Two categories of how we market the product. This should probably be ignored, or better yet, each type should be considered independently.
- New_Release_Flag = Any product that has had a future release (i.e., Release Number > 1)