# Sample Sales Data Analysis

Name

Institution Affiliation

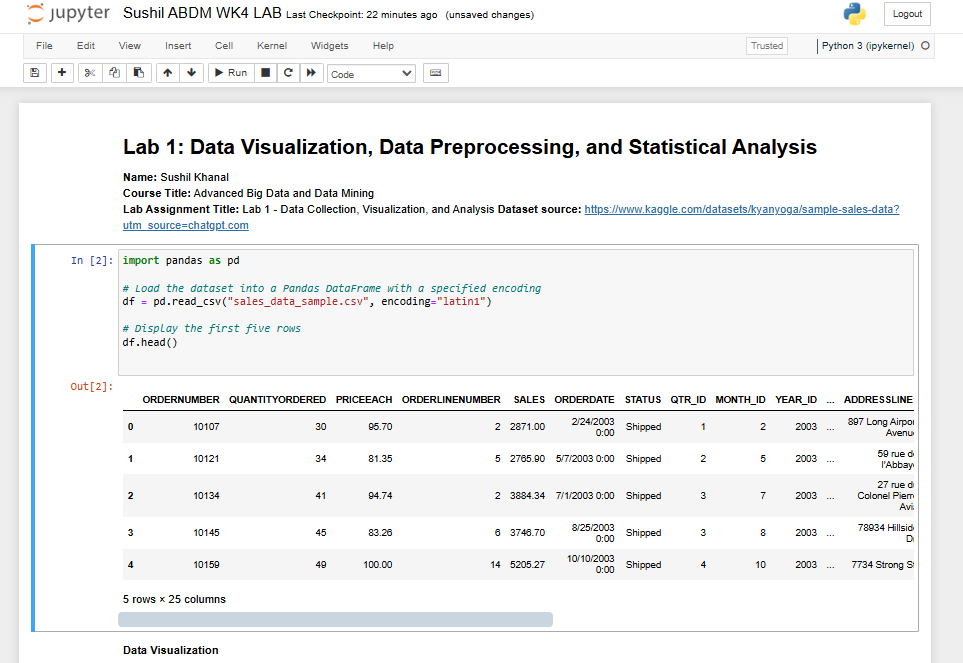
# Sample Sales Data Analysis

Introduction

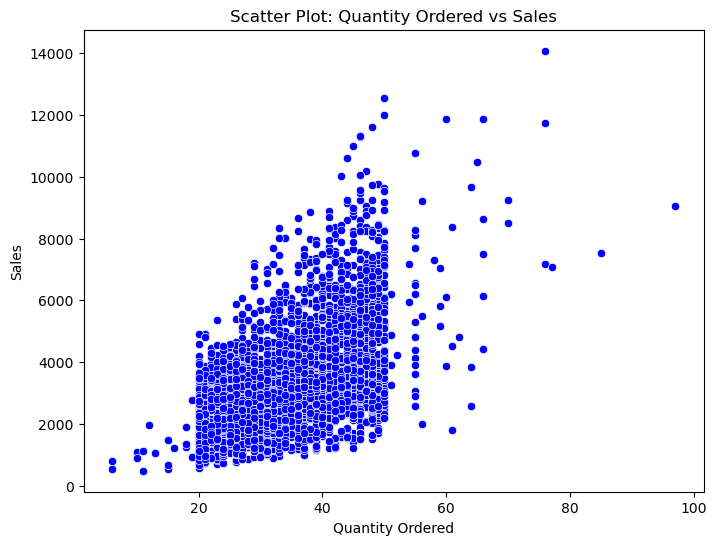
The purpose of this lab work is to explore, preprocess, and analyze data to extract meaningful insights through data analysis techniques such as visualization, statistical analysis, and data transformations. These steps ensure that the dataset is clean, structured, and ready for analysis. The dataset used is called sample sales dataset downloaded from the link below.

<https://www.kaggle.com/datasets/kyanyoga/sample-sales-data?utm_source=chatgpt.com>

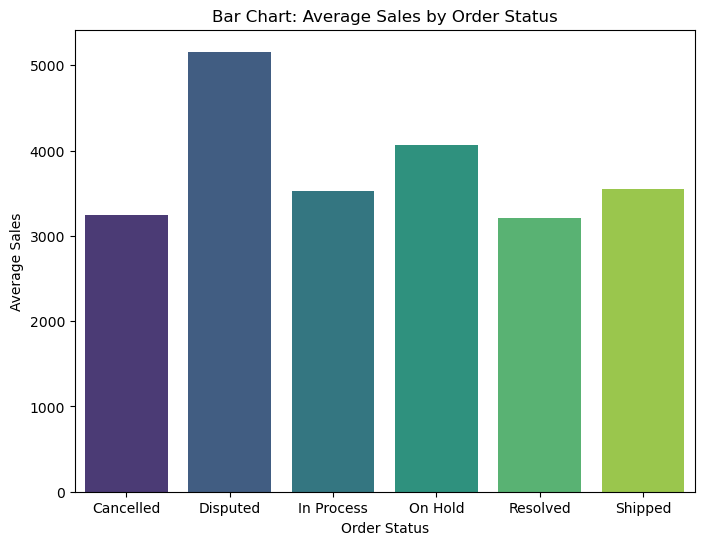
First few rows of the dataset



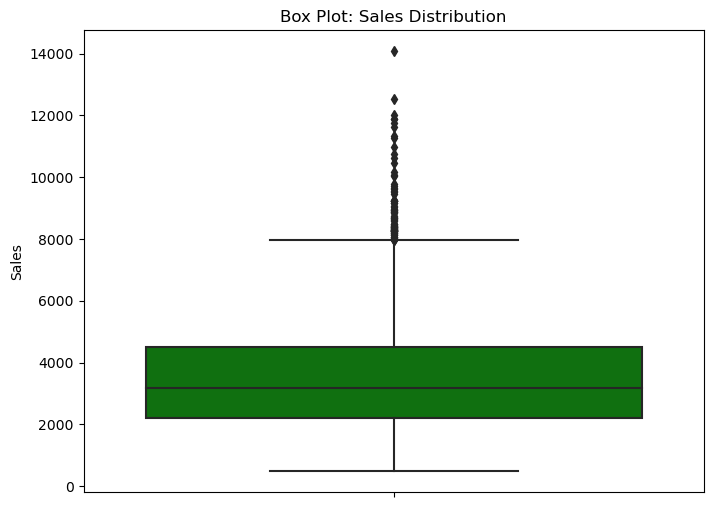
Data visualization



The scatter plot above showing the relationship between quantity ordered vs sales indicates that the two variables are positively correlated because an increase in quantity ordered leads to an increase of sales even though it is not a perfect correlation.



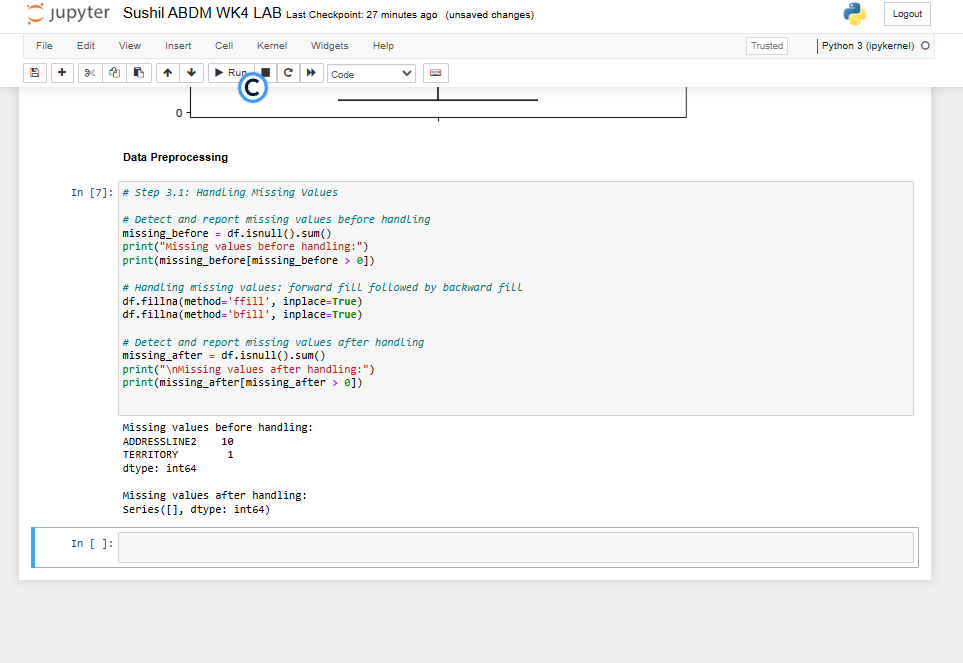
The bar chart above shows the average sales by order status. We can see that the Disputed order status has the highest average sales followed by on hold. The resolved and Cancelled status categories have the lowest average sales.



The box plot can help us understand the sales distribution of the dataset. The main insight we can get from the chart above is that most outliers are extremely high values as there are outliers noted on the upper end of the box plot.

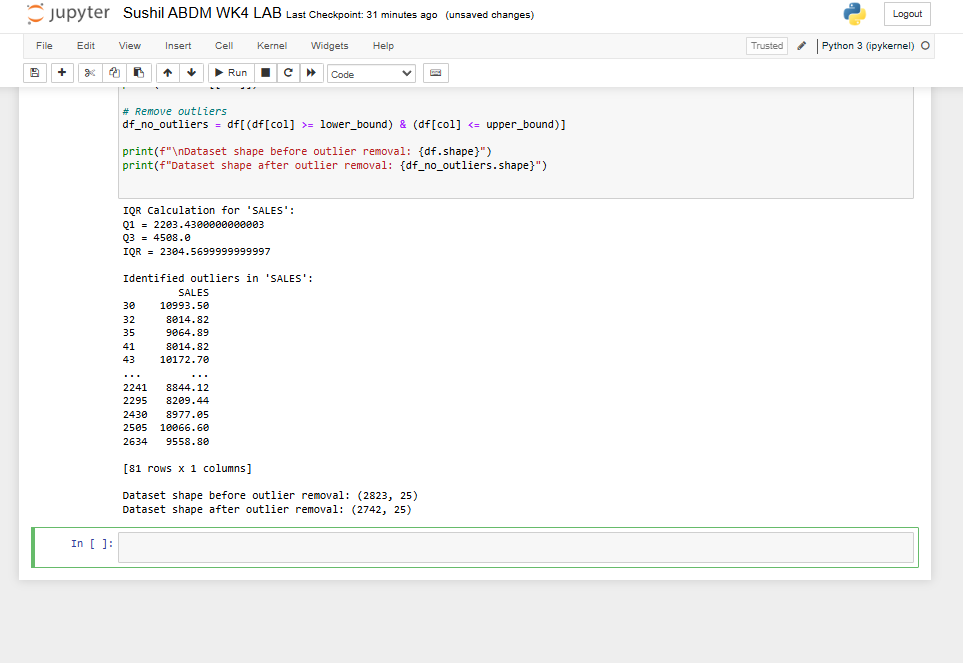
Data processing

Handling missing values



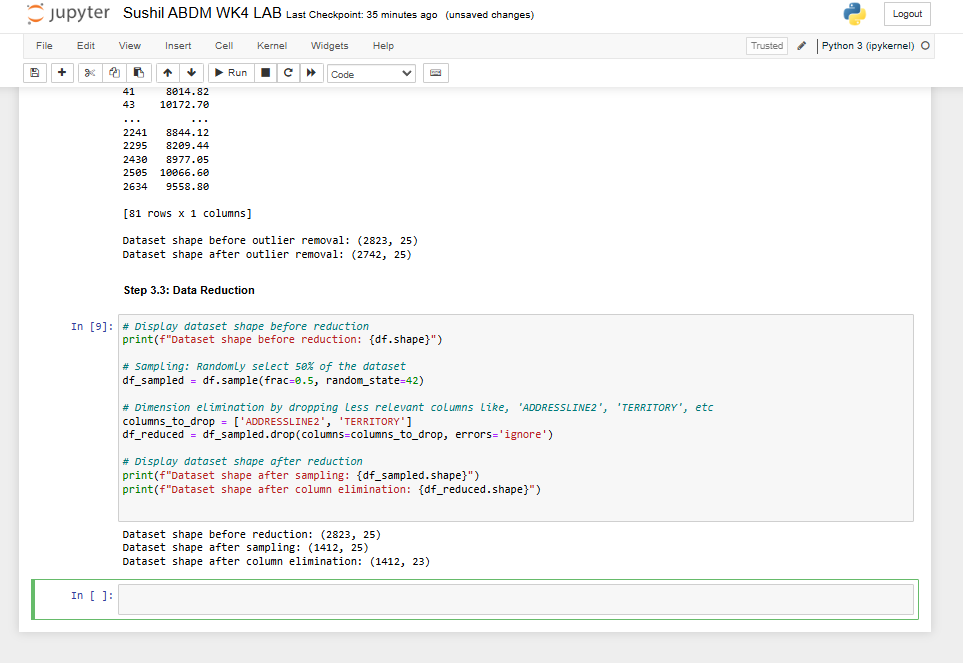
We had missing values in address line 2 and Territory which we handled successfully since the output after replacing the missing values shows that the dataset does not have any missing value.

Handling outliers using IQR



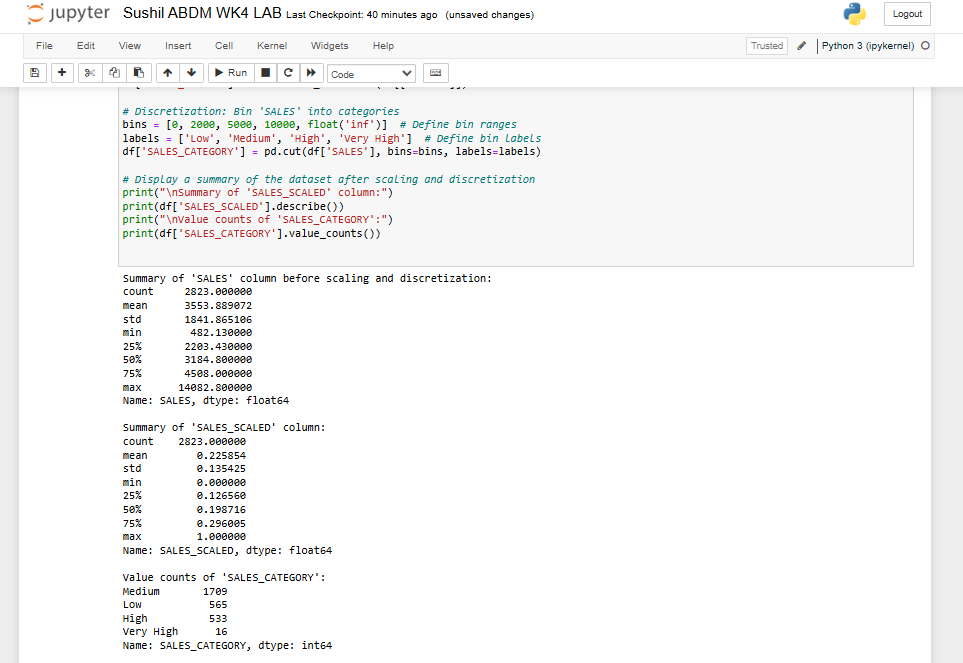
The dataset shape before and after handing outliers is different indicating that some rows identified as outliers were removed successfully.

Data reduction

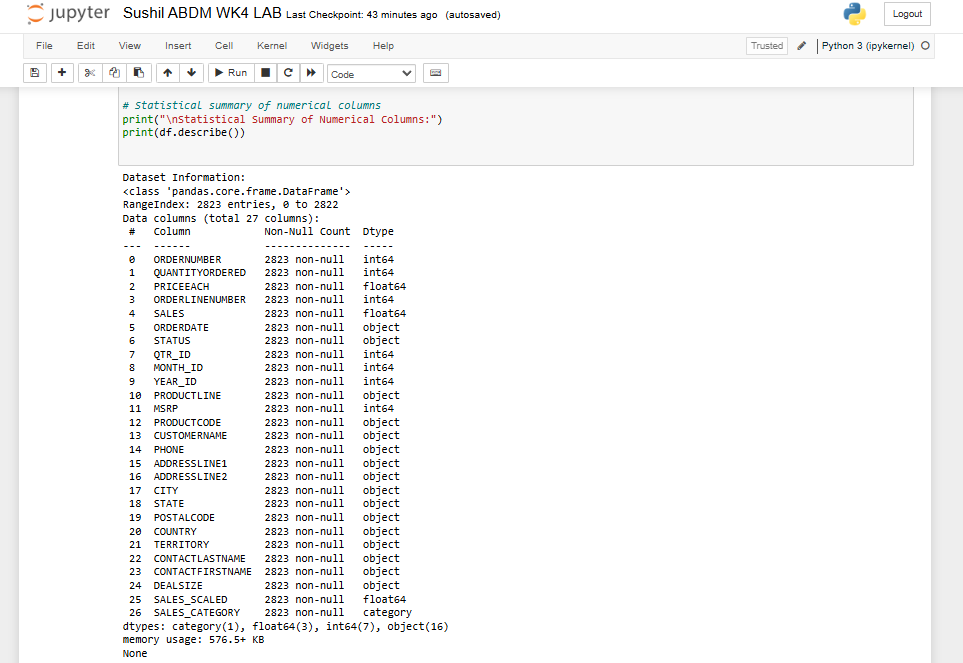


The dataset has been reduced from 2823 rows to 1412 rows and from 25 columns to 23 columns as shown in the screenshot above.

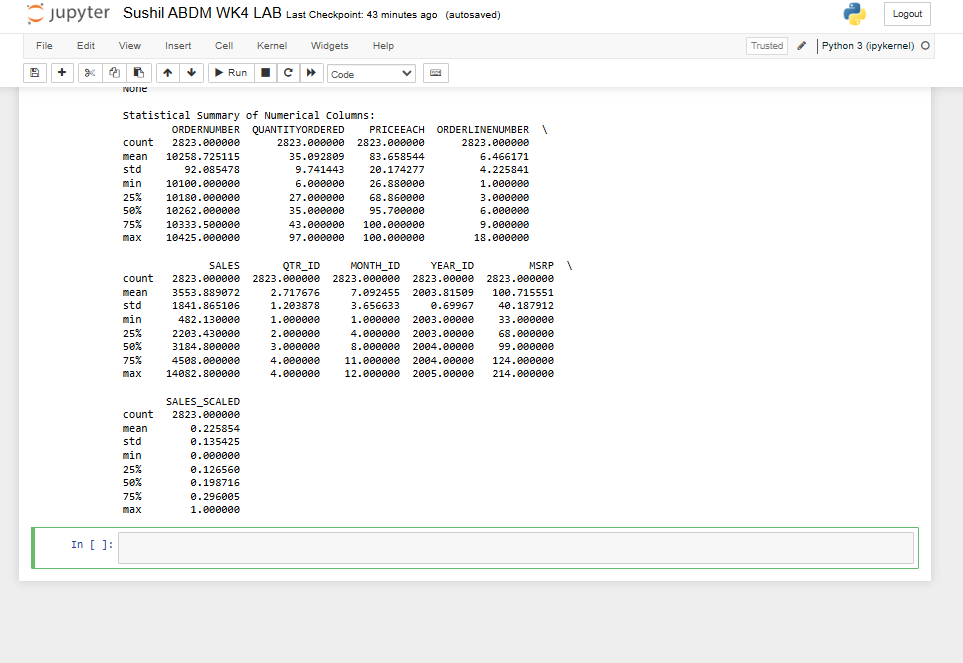
Data scaling



Data information

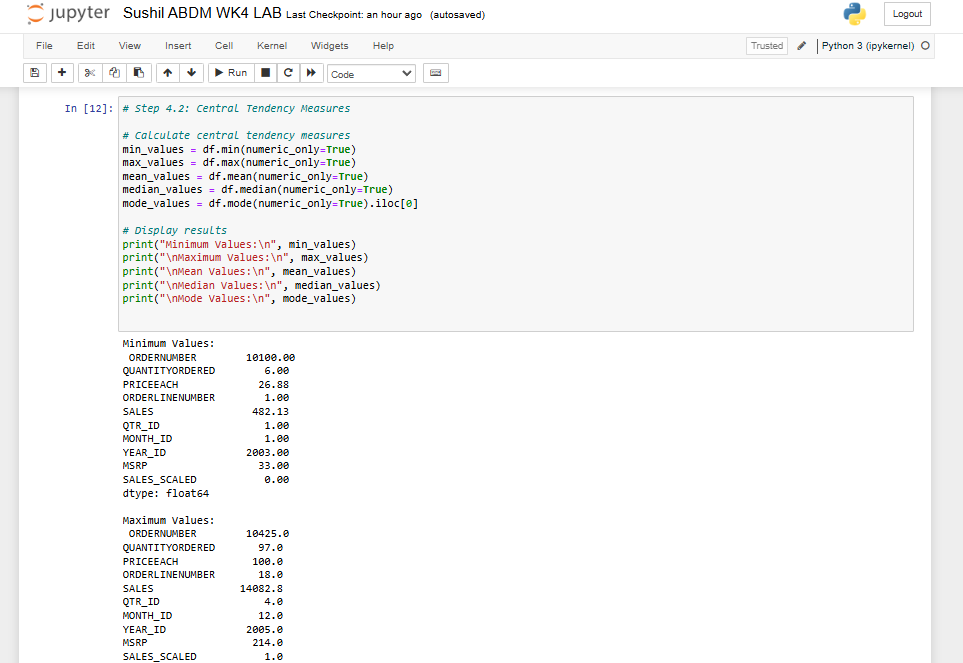


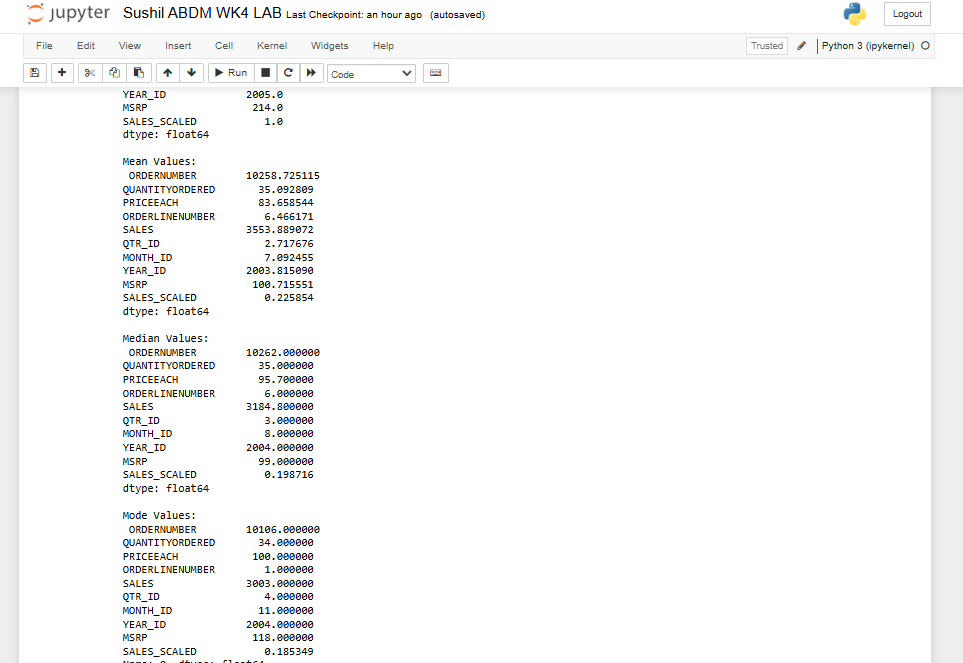
Statistical summary



Statistical summary of numerical columns provides summaries like std which represents the spread of the dataset and range measures like min and max. it also shows the quartiles and central tendency such as the mean of sales.

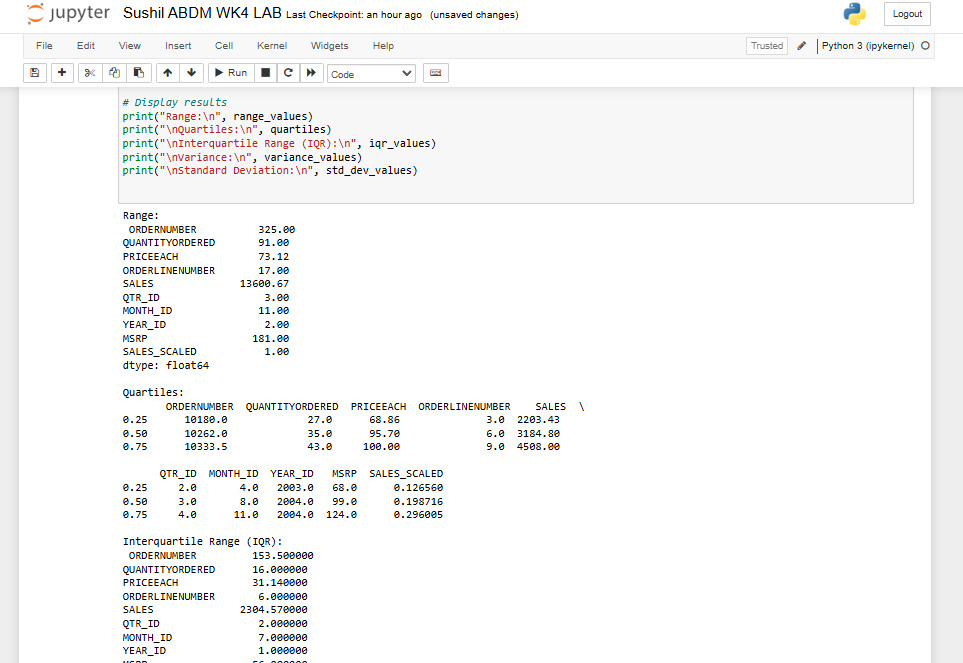
Central Tendency Measures

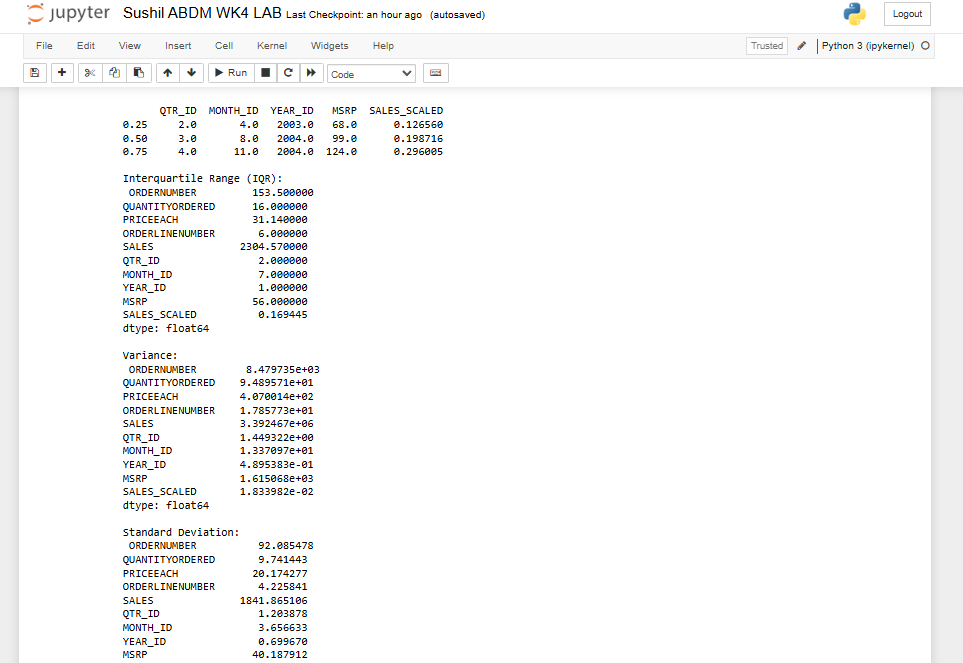




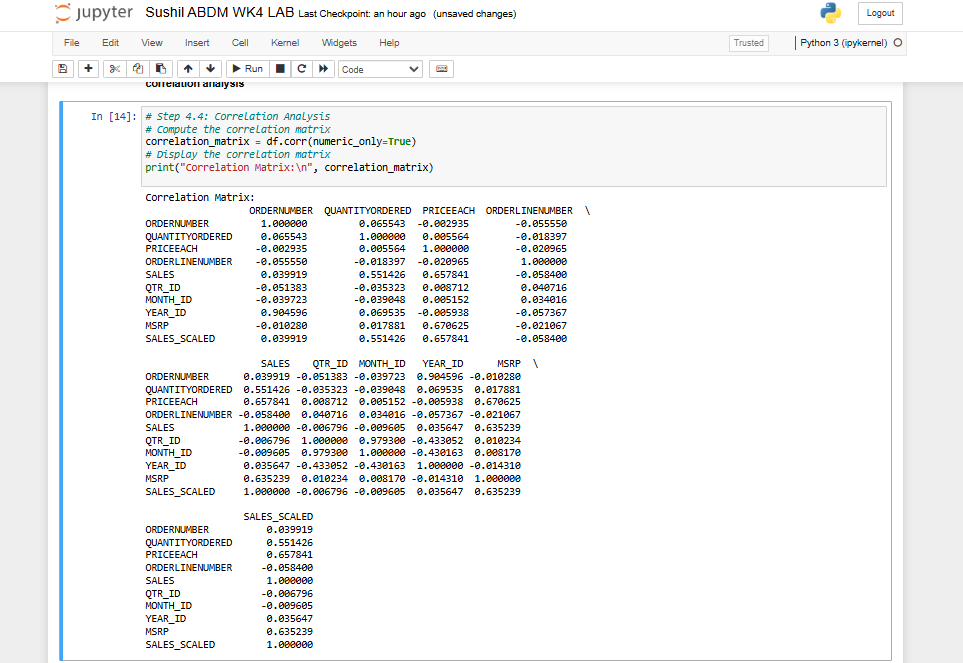
The central tendency measures analysis results above is similar to what we got from the data summary and info() functions we had used earlier.

Dispersion Measures





Correlation Analysis



Correlation analysis above shows that year ID and order number has a very strong correlation. Other variables with strong correlations are month ID vs order id, and price each vs MSRP and vs Sales. A positive correlation indicates that an increase in one leads to an increase in the other and vice versa.

Conclusion and challenges

The analysis process concluded successfully without any major challenges. All the issues I faced I could troubleshoot them using insights I got from a number of data analysis YouTube videos I had watched.