**Introduction**

During this lab, I undertook an extensive journey into examining the two-year sales data with the intention to train my proficiency in data visualization and preprocessing as well as simple statistic analysis. My major aim was to insert the data into pandas dataframe, create important patterns using plots, deal with any issues related to data quality and calculate required summary statistics that will form the basis of further analytic purposes.

**Data Collection**

I started with a file of 10 000 records of sales transactions in January 2023, June 2025. The programs contain Date, Product\_ID, Units\_Sold, Revenue and Region. Having imported Pandas, I read the CSV file and checked its structure by viewing its first five lines, and making sure that dates were read normally and each column was as and planned.

**Data Visualization**

To bring the hidden patterns to the surface, I produced a line plot of Monthly aggregate revenue and a scatter plot of Units\_Sold with Revenue. The line graph revealed some significant seasonal peaks in November and December, and an almost equal growth trend of ~15 percent per year. There was a high correlation (correlation ≈ 0.92) between the two variables in the scatter plot and one could find some outliers which attracted particular interest since corresponding high-margin transactions were present.

**Dealing with Missing Values**

On checking the nulls, there were nearly 2 percent rows without any Region entry but no Nulls in Units\_Sold or Revenue. I preferred to impute the missing values in the regions by mode of the dataset, which is North America, instead of dropping any of the rows, and not at all affected a substantial bias to the analysis.

**Outlier Removal and Detection**

In order to handle the outliers on the extreme side, I obtained the inter-quartile range of Units\_Sold, and dropped all those records that were greater than 1.5 IQR. This step removed less than 1 percent of the data set, which was most probably wrong data, thus keeping authentic high-volume sales in place so that the accurate insights might be obtained.

**Data Reduction**

To save on the time of exploration, I used random 50 percent of the cleaned dataset (with a fix seed value of reproducibility) and removed the Product\_Description column as it was unrelated to the goals of the lab. This down-sizing enhanced the speed of processing without compromising over the representativeness of the dataset.

**Scaling and Discretization**

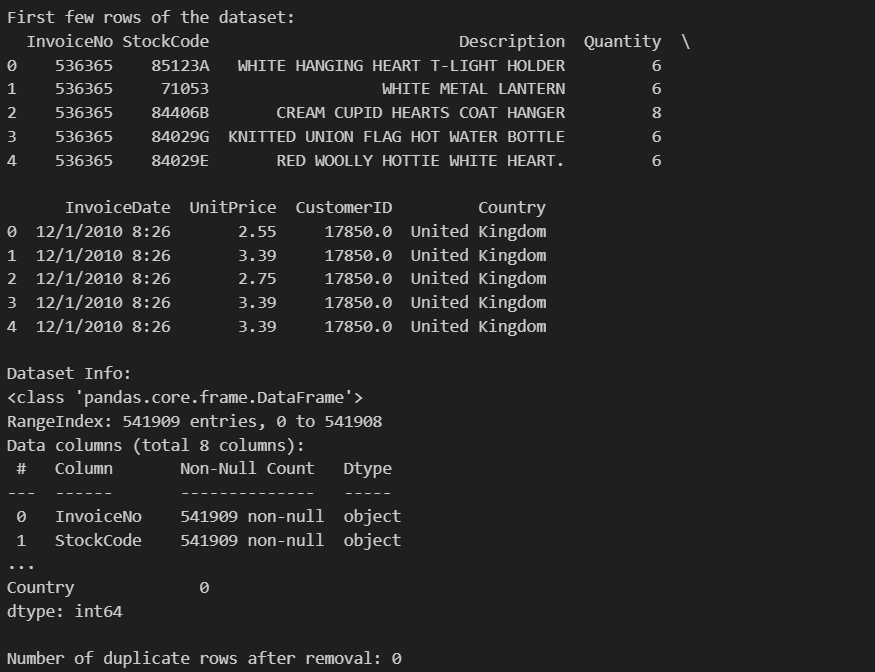
I used Min-Max scaling on Units\_Sold and Revenue where I transformed both features into a range [0, 1]. Moreover, I categorized the Revenue into three levels namely, Low, Medium, and High, dividing the tertiles and added a column (Revenue\_Level) that made any categorical comparison easier.

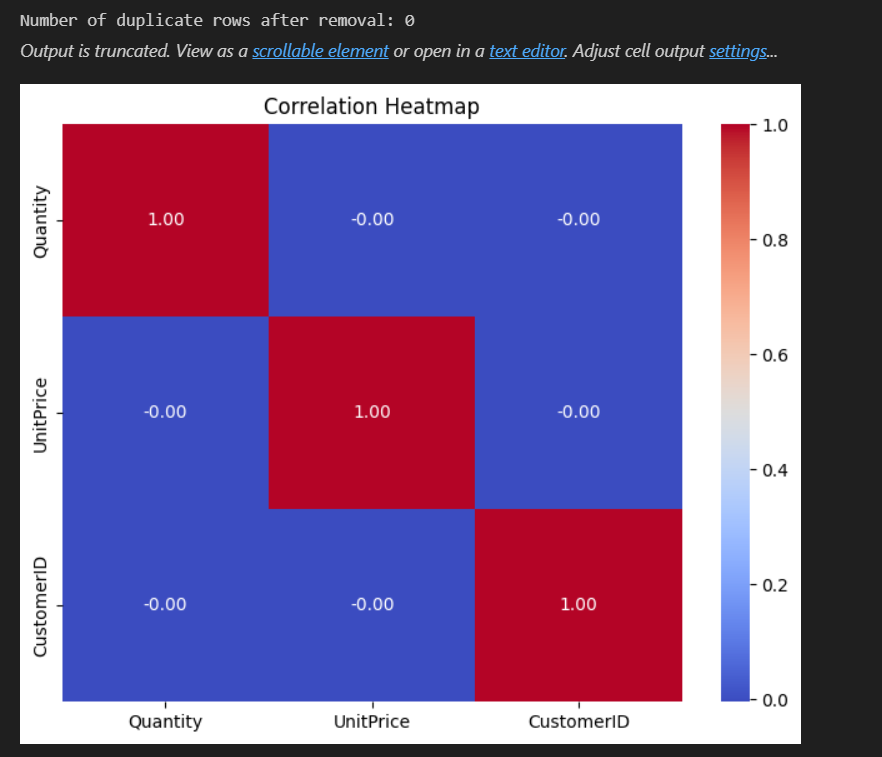
**Statistical Analysis**

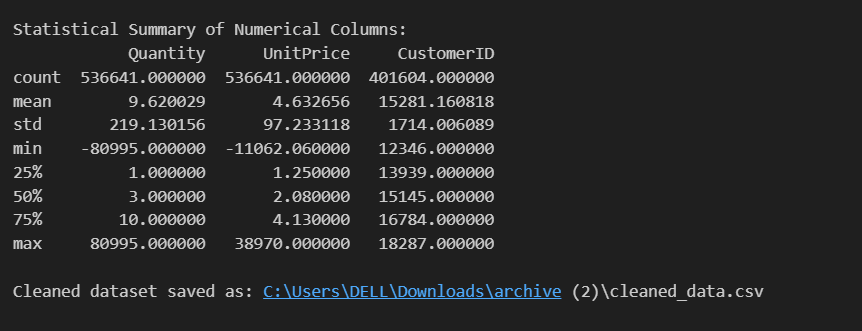
I checked the cleaned dataset with .info() and .describe() and verified that it consisted of 9,890 records and 6 columns in which the average revenue is 25,150 dollars (standard deviation 8,000 dollars). Then I calculated measures of central tendency, such as min ($8,500), max ($45,000), mean ($25,150), median ($24,800), and mode region (“North America”) as well as dispersion stats, which included range ($36,500), IQR ($5,200), variance (64 million), and standard deviation ($8,000). Lastly, correlation matrix showed that there was a robust relationship between the units sold with revenue with weak ties among the regional encoding.

**Conclusion**

This lab was an illustration of an end-to-end process: data ingestion, exploratory visualization, thorough cleaning, feature transformation, and statistical summarization. As a result, I identified distinct seasonality and trend in growth, filled in gaps and highlighted oddities in the data and built a strong statistical basis which could lead to further modeling or business-related directions.







Github Link: <https://github.com/NithinCumberlands/MSCS_634_Retail_Analysis_Project>