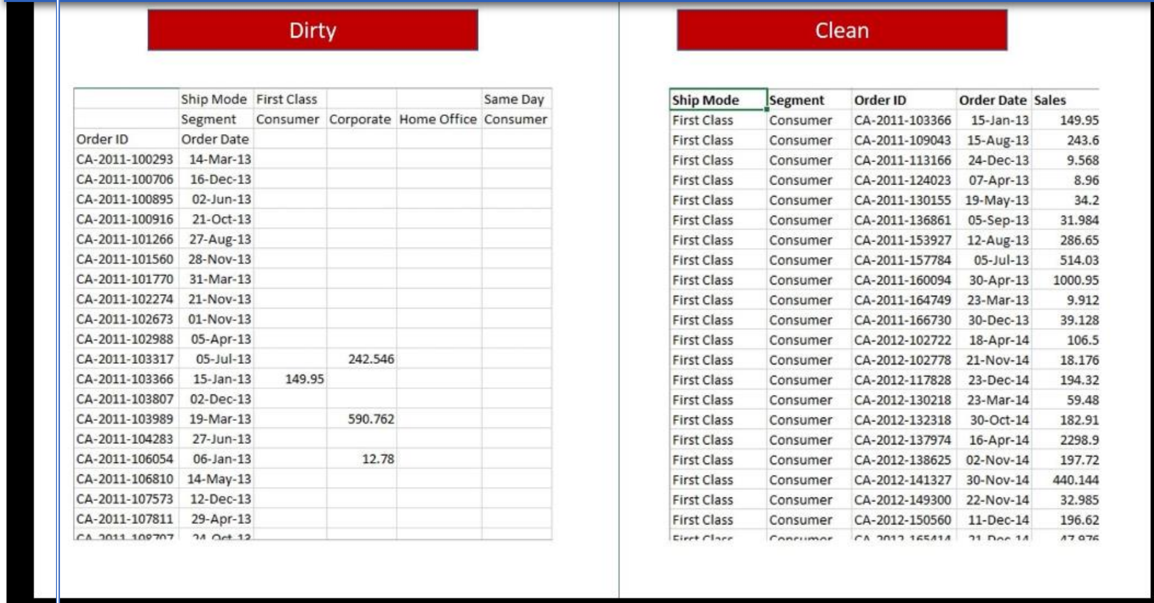
2025

SALES DATA CLEANING

For Exploratory Analysis



Dirty

	Ship Mode	First Class			Same Day
	Segment	Consumer	Corporate	Home Office	Consumer
Order ID	Order Date				
CA-2011-100293	14-Mar-13				
CA-2011-100706	16-Dec-13				
CA-2011-100895	02-Jun-13				
CA-2011-100916	21-Oct-13				
CA-2011-101266	27-Aug-13				
CA-2011-101560	28-Nov-13				
CA-2011-101770	31-Mar-13				
CA-2011-102274	21-Nov-13				
CA-2011-102673	01-Nov-13				
CA-2011-102988	05-Apr-13				
CA-2011-103317	05-Jul-13		242.546		
CA-2011-103366	15-Jan-13	149.95			
CA-2011-103807	02-Dec-13				
CA-2011-103989	19-Mar-13		590.762		
CA-2011-104283	27-Jun-13				
CA-2011-106054	06-Jan-13		12.78		
CA-2011-106810	14-May-13				
CA-2011-107573	12-Dec-13				
CA-2011-107811	29-Apr-13				
CA-2011-108707	24-Oct-13				

Clean

Ship Mode	Segment	Order ID	Order Date	Sales
First Class	Consumer	CA-2011-103366	15-Jan-13	149.95
First Class	Consumer	CA-2011-109043	15-Aug-13	243.6
First Class	Consumer	CA-2011-113166	24-Dec-13	9.568
First Class	Consumer	CA-2011-124023	07-Apr-13	8.96
First Class	Consumer	CA-2011-130155	19-May-13	34.2
First Class	Consumer	CA-2011-136861	05-Sep-13	31.984
First Class	Consumer	CA-2011-153927	12-Aug-13	286.65
First Class	Consumer	CA-2011-157784	05-Jul-13	514.03
First Class	Consumer	CA-2011-160094	30-Apr-13	1000.95
First Class	Consumer	CA-2011-164749	23-Mar-13	9.912
First Class	Consumer	CA-2011-166730	30-Dec-13	39.128
First Class	Consumer	CA-2012-102722	18-Apr-14	106.5
First Class	Consumer	CA-2012-102778	21-Nov-14	18.176
First Class	Consumer	CA-2012-117828	23-Dec-14	194.32
First Class	Consumer	CA-2012-130218	23-Mar-14	59.48
First Class	Consumer	CA-2012-132318	30-Oct-14	182.91
First Class	Consumer	CA-2012-137974	16-Apr-14	2298.9
First Class	Consumer	CA-2012-138625	02-Nov-14	197.72
First Class	Consumer	CA-2012-141327	30-Nov-14	440.144
First Class	Consumer	CA-2012-149300	22-Nov-14	32.985
First Class	Consumer	CA-2012-150560	11-Dec-14	196.62
First Class	Consumer	CA-2012-165414	21-Dec-14	47.976

Okello Raymond, 0778412136

Data Analyst

2/7/2025

SALES DATA CLEANING

1 INTRODUCTION

Data quality plays a crucial role in making accurate, data-driven decisions. Raw datasets often contain inconsistencies such as missing values, incorrect data types, and redundant information, which can lead to misleading insights.

This report outlines the process of cleaning and refining a dataset that initially contained 823 records and 14 columns. The dataset had:

- Missing values in several columns.
- Incorrect data types, with the Segment column stored as datetime instead of categorical.
- Redundant columns, including multiple segmented categories.

To ensure accuracy and efficiency, the dataset was transformed into a structured format, improving usability for analysis and visualization. This report details the exploration, cleaning steps, challenges faced, and the final cleaned dataset, providing insights into best practices for data preparation.

2 DATA EXPLORATION

Before cleaning the dataset, an initial exploration was conducted to understand its structure, data types, missing values, and potential inconsistencies.

The original dataset contained:

- 823 rows and 14 columns. These columns were not categorically correct for example, First Consumer as a column combines First Class Consumers which needs to be split as (Ship Mode and Segment) columns
- The column Segment lists the dates ordered, this needed to be renamed to its appropriate format and data values
- With this unnecessary column headers, sseveral columns with missing values were found

3 DATA CLEANING STEPS

To improve the dataset's quality and usability, several cleaning steps were performed. The goal was to remove inconsistencies, handle missing values, and restructure the data for better analysis.

- The dataset had two level headers, this was first reduced to one level in excel to combined headers such as First Consumer, Same Corporate, etc. This was only a step since there were still very many column headers which were unnecessary since the Segment, and Ship Mode columns could represent customer and shipping categorization.
- Creating a data frame for each of the four segments, this allowed me to create a column named "Ship Mode" and respective names were given according to the segment
- Null values were dropped from the respective data frame row meaning they don't belong in that segment
- The Ship Mode were picked from the column names contained orders, this meant that for first class, the order belonged to consumer if there is a sale in that column.
- All this was done for all the four segments which was then combined to form one data frame
- Column names were reviewed for clarity. Order Date, Ship Mode, and Sales were retained without changes, ensuring consistency.
- By reducing the number of columns from 14 to 5, memory usage dropped from 90.1 KB to 32.2 KB, making the dataset more efficient

4 CHALLENGES AND SOLUTIONS

4.1 HANDLING MISSING VALUES

Several columns, including First Consumer, First Corporate, and First Home Office, had significant missing values. These columns were deemed redundant and removed during the cleaning process, eliminating the issue of missing data in those columns.

4.2 INCORRECT DATA TYPES

The Segment column was incorrectly stored as a datetime64[ns] type, leading to potential misinterpretation. The data type of Segment was corrected to categorical (object) to better reflect its intended use.

4.3 Redundant Columns and Fragmented Data

The dataset contained multiple columns representing the same segmentation categories (e.g., First Consumer, Same Corporate), making the data unnecessarily complex. Redundant columns were removed, and the Segment column was retained to categorize the data, simplifying the structure.

4.4 Optimizing Memory Usage

The original dataset used 90.1 KB of memory due to numerous columns. Solution: By removing unnecessary columns and restructuring the data, the memory usage was reduced to 32.2 KB, improving efficiency. These challenges were addressed through strategic cleaning steps, making the dataset more manageable and suitable for analysis

5 EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) was performed on the cleaned dataset to uncover patterns, detect anomalies, and understand the relationships between variables. The key steps of the EDA process included summarizing statistics, visualizing distributions, and examining correlations.

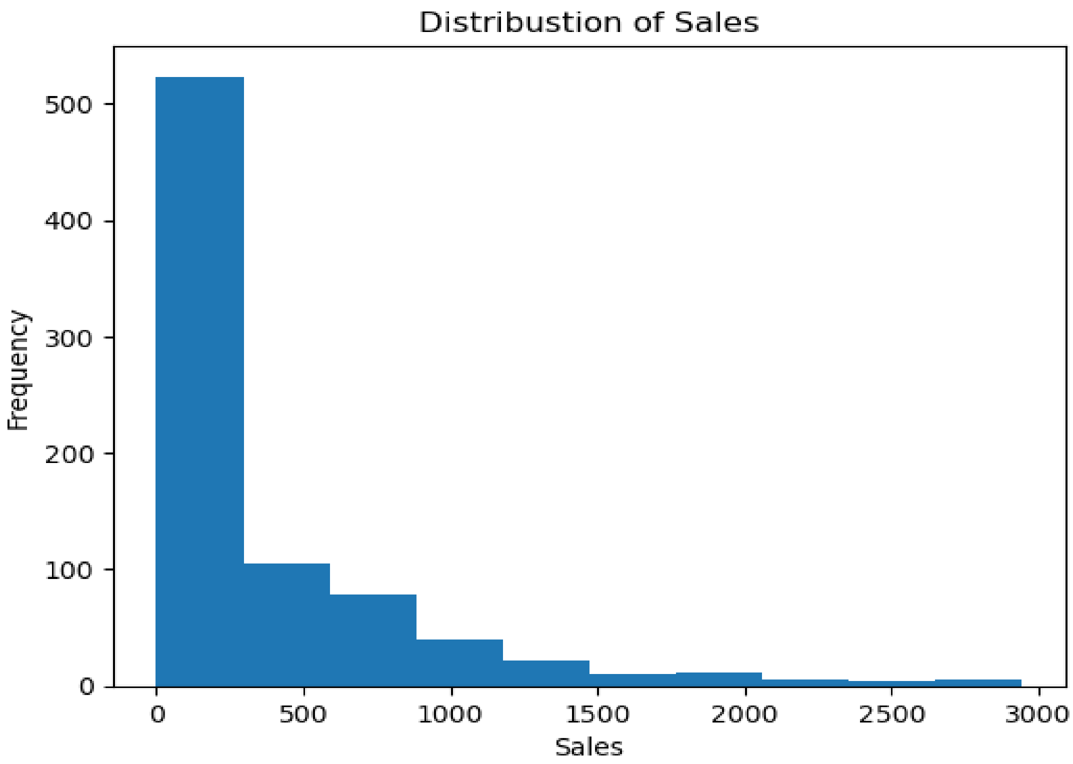
5.1 Summary Statistics

The cleaned dataset now contains 822 rows and 5 columns: Order ID, Order Date, Ship Mode, Segment, and Sales. Summary statistics for the Sales column revealed that the dataset was badly skewed to the left and sales above 3000 were filtered out. The result data set with 803 rows were used to describe the data set as follows:

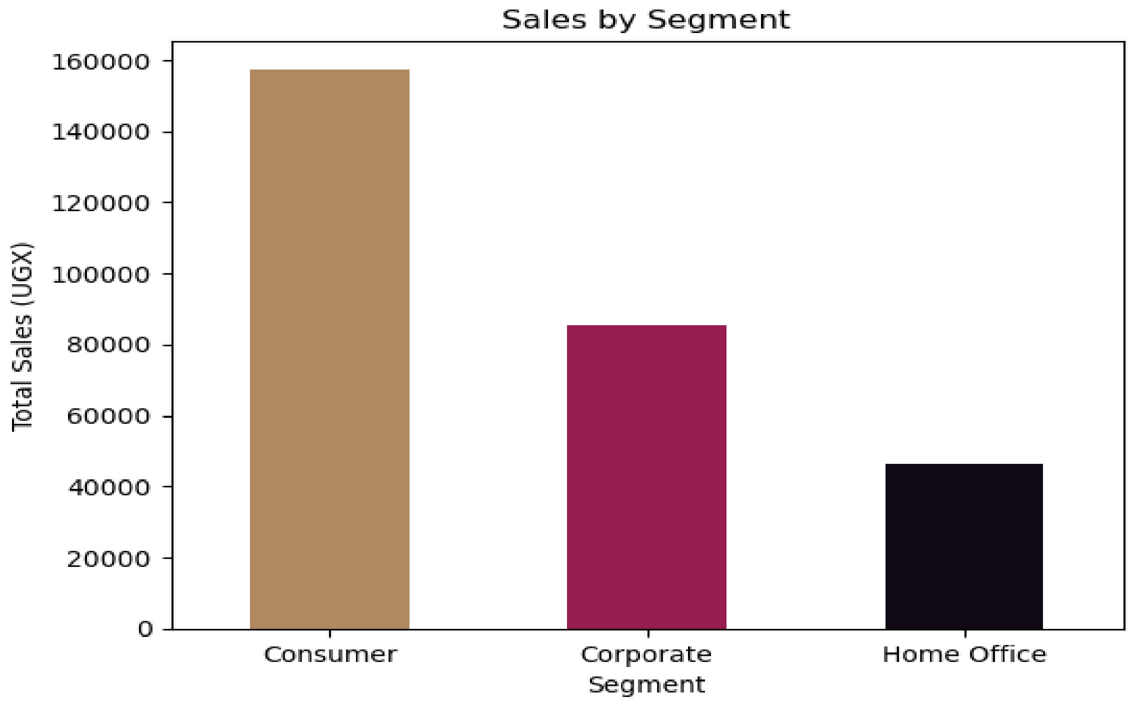
- Mean Sales: 359
- Median Sales: 148
- Sales Range: From 1 to 2942
- Standard Deviation: 500

5.2 Data Distribution and Visualizations

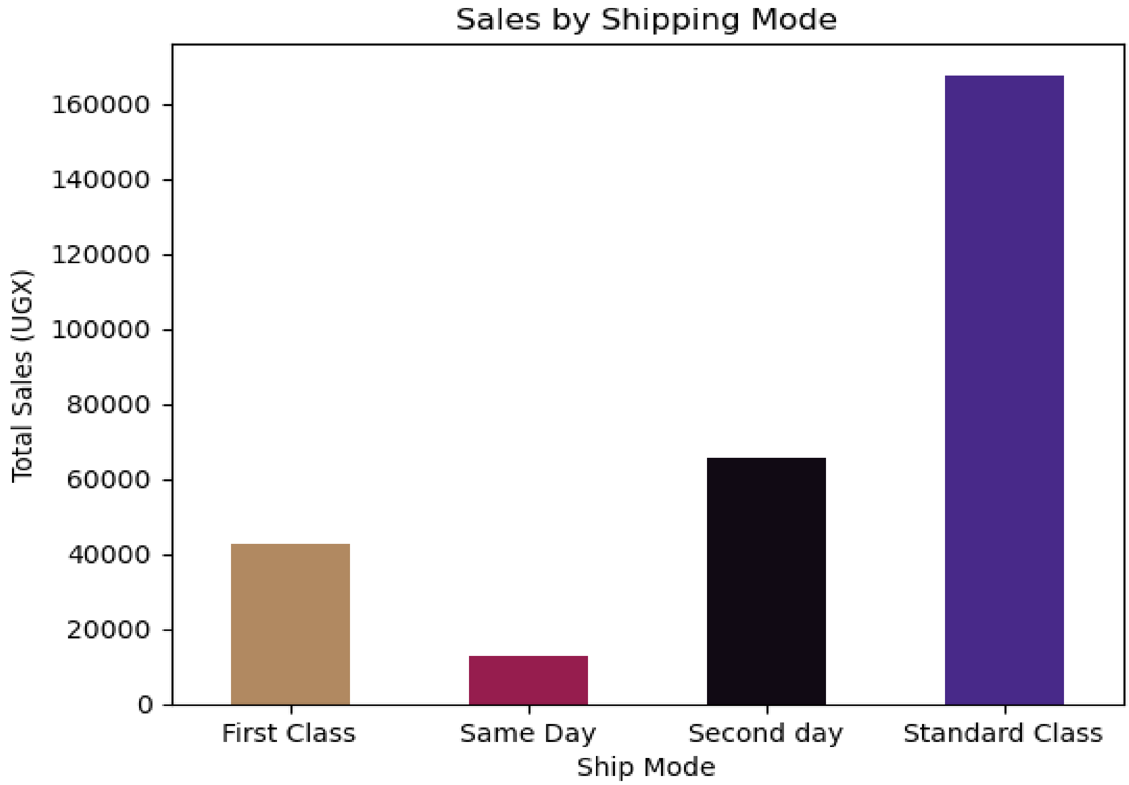
- **Sales Distribution**: A histogram was created to visualize the distribution of sales, which revealed a positively skewed distribution. This distribution indicates that most of the sales transactions are relatively low, with only a few of our products charged very highly.



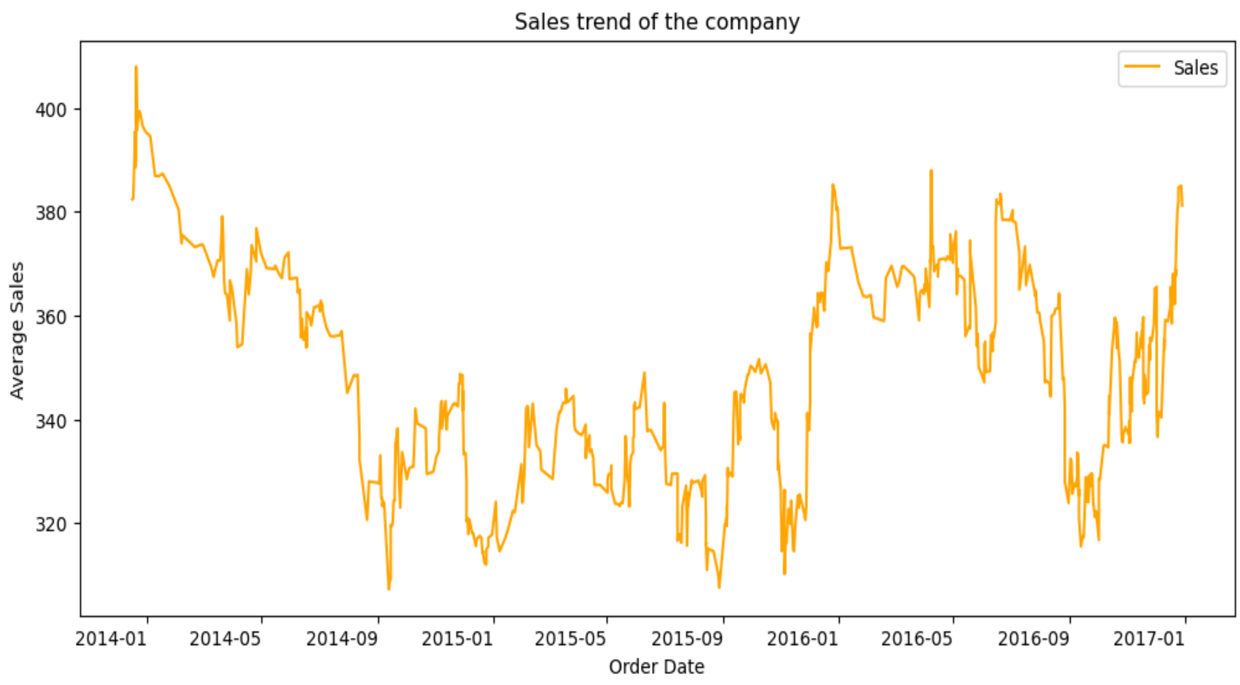
- Segment-wise Sales: A bar chart was used to compare sales across different segments, highlighting [key differences].



- Ship Mode Sales: A bar chart was used to compare sales across different sales, highlighting [key differences]



- Order Date Trends: A time series plot showed how sales varied over time, helping to identify seasonal patterns or trends.



6 CONCLUSION

- Summarize the impact of your cleaning process. How did it improve the dataset?
- Mention any next steps (e.g., analysis, modelling, or visualization)

7 TOOLS AND TECHNOLOGIES

- Python
- Pandas
- Matplotlib
- Excel

8 REFERENCES

To get a hold of the datasets used in the above process, and insights that might have gone unnoticed, go through the datasets below

Data Sources	
Unclean Data	Click Here
Clean Data	Click Here
Steps Taken	Click Here