**Lab 1: Data Visualization, Data Preprocessing, and Statistical Analysis Using Python in Jupyter Notebook**

Romika Souda

Department of Computer Science, University of the Cumberlands

MSCS-634-M40: Advanced Big Data and Data Mining

Dr. Satish Penmatsa

May 21, 2025

**Step 1: Data Collection**

A screenshot of a computer

AI-generated content may be incorrect.

**Step 2: Data Visualization**

A screen shot of a computer screen

AI-generated content may be incorrect.Scatter Plot: Quantity Ordered vs Sales

We can see a noticeable positive link between how much is ordered and the sales revenue in the data. The more product ordered, the more likely sales will also go up which shows that large orders help boost total sales. A majority of data is found at low quantities and sales which means most orders are for small to medium amounts. Although the vast majority is small, a tiny part of the data reflects large-scale orders. It helps explain how sales increase with each order size and can be used to create better inventory and marketing approaches.

A screen shot of a computer

AI-generated content may be incorrect.Line Plot: Total Sales Over Time

The sales history is shown in the line plot which indicates a great deal of variation during the period observed. There were high levels of sales during late 2018 and late 2019 which may suggest that seasonal or promotional activities drove those results. The trend for resource usage appears stable overall, yet there are times when demand increases sharply. Businesses can use this kind of analysis to see trends, decide when sales may increase and manage their stock and resources more effectively.

A screenshot of a computer

AI-generated content may be incorrect.

Bar Chart: Number of Orders by Product Line

‘Classic Cars’ took the lead when it comes to the most orders, followed by ‘Vintage Cars’ and ‘Motorcycles’ in second and third place. It means that more people are buying toys in these lines than in ‘Trains’ or ‘Ships’ right now. The distribution process can guide how products are created, how advertising is launched and how stock is handled. It can be helpful for businesses to look into popular categories and search for ways to raise interest in less successful goodsA screenshot of a computer

AI-generated content may be incorrect.

Histogram: Distribution of Sales Amounts

We can see from the histogram that the majority of sales are between the lower and middle sales amounts. Most of the data shows a right-skewed distribution, suggesting that there are fewer very large sales. It appears that most sales are of average value, but high-value ones make a big difference to the company’s overall revenue. Knowing this distribution guides sales strategies to keep both quantity and revenue in line.A screenshot of a computer

AI-generated content may be incorrect. Box Plot: Sales by Deal Size

The data suggests that larger deals tend to lead to higher sales numbers and a few cases go well above the upper whisker. The sales levels and the spread of ‘Medium’ and ‘Small’ deals keep getting lower, suggesting that smaller deals are sold less consistently. There are occasional ‘Large’ deals with much higher amounts than the average. Through this visualization, we can determine the impact of the deal size on sales and detect unusual deals.A screenshot of a computer

AI-generated content may be incorrect. Pie Chart: Proportion of Orders by Deal Size

Nearly half of all orders are for ‘Medium’ deals and ‘Small’ deals come in closer second, according to the pie chart. Most deals are not large; a much smaller percentage is. Most of the transactions in this industry are either moderate or small, while the number of major large deals is smaller. Using these numbers helps businesses track how and what customers buy and develop sales methods to target them.

**Step 3: Data Preprocessing**

Perform data preprocessing steps following the provided examples:

Handling Missing Values:A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

Outlier Detection andRemoval:A screenshot of a computer

AI-generated content may be incorrect.

DataReduction: A screenshot of a computer

AI-generated content may be incorrect.

Data Scaling and Discretization:

A screenshot of a computer

AI-generated content may be incorrect.

**Step 4: Statistical Analysis**

General Overview of dataA screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.

CentralTendencyMeasures: A computer screen with a white background

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.

3.DispersionMeasures: A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.

4.CorrelationAnalysis: A screenshot of a computer

AI-generated content may be incorrect.