HNG stage Zero task

Objectives
1. Review any of the provided dataset.
2. Identify initial insights from the data at first glance.
3. Write a brief technical report detailing your observations.

Steps to Complete the Task
1. Dataset Familiarization:
    - Open one of the provided dataset to understand its structure and contents.

Samples sales data is a secondary data used for Pentaho DI Kettle for retail analytics. The data was Originally Written by María Carina Roldán, Pentaho Community Member, BI consultant (Assert Solutions), Argentina. This data can help with sales simulation training

    - Identify the key variables and data types (e.g., numerical, categorical).

Key Variables and Data Types
- Numerical: ORDER NUMBER, QUANTITY ORDERED, PRICEEACH, ORDERLINE NUMBER, SALES, QTR_ID, MONTH_ID, YEAR_ID, MSRP, POSTALCODE

- Categorical: ORDERDATE, STATUS, PRODUCT LINE, PRODUCTCODE, CUSTOMERNAME, PHONE, ADDRESSLINE1, ADDRESSLINE2, CITY, STATE, COUNTRY, TERRITORY, CONTACT LAST NAME, CONTACT FIRSTNAME, DEALSIZE

Observations
- The dataset contains a mix of numerical and categorical variables.

2. Initial Data Exploration:
    - Conduct a quick review of the dataset without deep analysis.

The dataset are not organised

    - Look for obvious patterns, trends, or anomalies in the data.

- Dataset are not formatted at first glance
- The `ORDERDATE` column is currently an object type and may need to be converted to a datetime type for time series analysis.
- There are some missing values in the `ADDRESSLINE2` and `STATE` columns.

3. Insight Identification:
    - Note any initial insights that can be immediately observed from the dataset.

Data not organised
The orderdate are not in the right format
The year_id and month_id are not well formatted and arranged
Have a lot of missing states
Data not formatted
4. Technical Report Writing:

- Write a brief technical report that includes:
    - Introduction: Briefly describe the dataset and the purpose of the review.
    - Observations: Present the initial insights you identified, supported by basic visualizations or summary statistics if necessary.
    - Conclusion: Summarize your observations and suggest potential areas for further analysis.

5. Review and Submission:
    - Proofread and edit your technical report for clarity and accuracy.
    - Publish your article on any online blog platform
    - Submit the final technical report article link for review.
    - Submission Link: https://forms.gle/xjqFgPJGQfLk728W9

This is my first task at HNG Tech internship

Introduction:

This report provides a preliminary analysis of the sample sales data used for Pentaho DI Kettle (for retail analytics). The data was Originally Written by María Carina Roldán, Pentaho Community Member, BI consultant (Assert Solutions), Argentina. This data can help with sales simulation training

The purpose of this review is to gain initial insights into the data's characteristics, identify potential areas for further exploration, and insight identification.

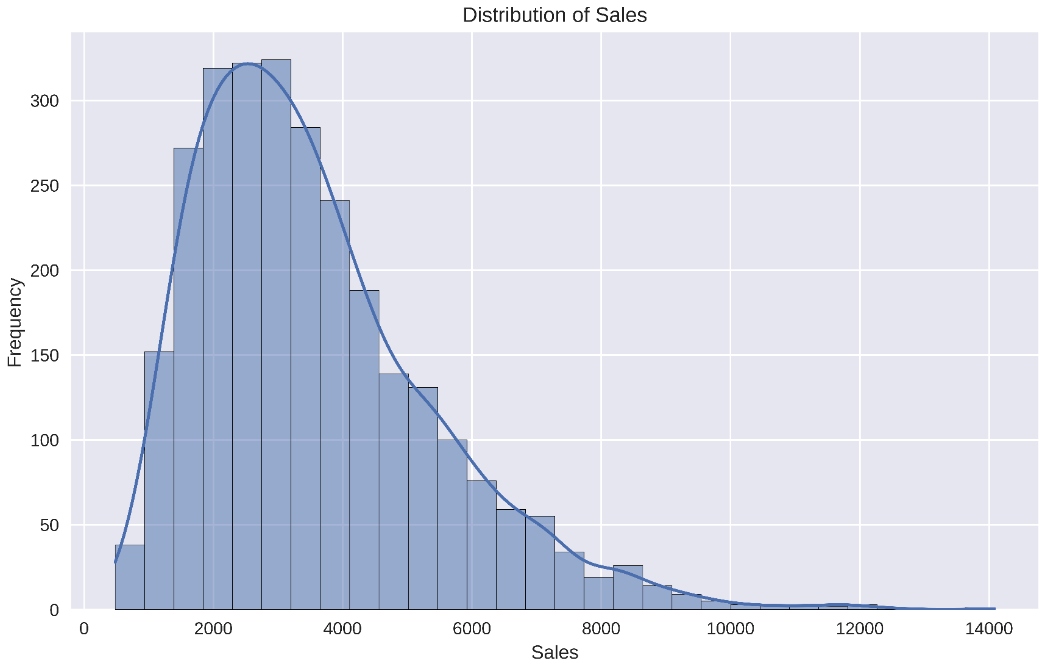
Observations:

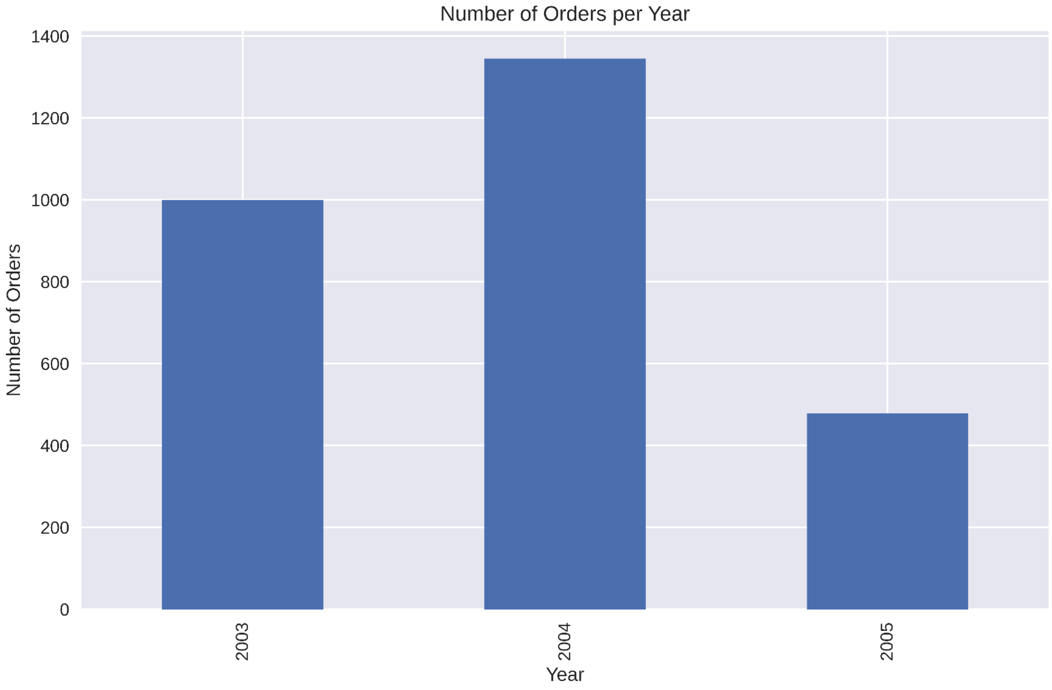
Data Overview:
The features comprise a mix of data variable and data types, including - Numerical: ORDER NUMBER, QUANTITY ORDERED, PRICEEACH, ORDERLINE NUMBER, SALES, QTR_ID, MONTH_ID, YEAR_ID, MSRP, POSTALCODE

- Categorical: ORDERDATE, STATUS, PRODUCT LINE, PRODUCTCODE, CUSTOMERNAME, PHONE, ADDRESSLINE1, ADDRESSLINE2, CITY, STATE, COUNTRY, TERRITORY, CONTACT LAST NAME, CONTACT FIRSTNAME, DEALSIZE

Initial Exploration:
Basic summary statistics reveal that Order Number: Ranges from 10100 to 10425.
Quantity Ordered: Ranges from 6 to 97, with a mean of approximately 35.
Price Each: Ranges from $26.88 to $100, with a mean of approximately $83.66.
Sales: Ranges from $482.13 to $14,082.80, with a mean of approximately $3,553.89.
Year: Data spans from 2003 to 2005.

Visualization of categorical features (e.g., histograms, bar charts) suggests





Data Quality:
The ORDERDATE column is not in datetime.
There are significant missing values in the ADDRESSLINE2 and STATE columns.

The distribution of sales shows a right-skewed pattern.
The number of orders per year is not consistent.
Sales over time show some fluctuations

This initial analysis provides a foundational understanding of the sample sales data. The data exhibits [summarize key observations, e.g., Order Number: Ranges from 10100 to 10425.
Quantity Ordered: Ranges from 6 to 97, with a mean of approximately 35.
Price Each: Ranges from $26.88 to $100, with a mean of approximately $83.66.
Sales: Ranges from $482.13 to $14,082.80, with a mean of approximately $3,553.89.
Year: Data spans from 2003 to 2005, with highest number of sales in 2004

The product lines are
Classic Cars: $3,919,615.66
Motorcycles: $1,166,388.34
Planes: $975,003.57
Ships: $714,437.13
Trains: $226,243.47
Trucks and Buses: $1,127,789.84
Vintage Cars: $1,903,150.84

Conclusion:

A more in-depth analysis of missing values and outliers is recommended to determine appropriate handling methods.

This report serves as a springboard for further investigation into the data. By delving deeper and employing more sophisticated techniques, we can extract valuable insights from the sample sales data.

To learn more about the program, click on any of these links https://hng.tech/internship, https://hng.tech/hire, or https://hng.tech/premium