PROJECT FILE-1

SALES DATA

Submitted by:

RAJESH BAGHEL

9319652493

1. Introduction

This report presents a detailed analysis of sales data from an eCommerce company, with the objective of building a comprehensive dashboard that evaluates sales performance over a specific time period. The goal is to uncover patterns, trends, and insights that can drive informed business decisions and improve overall sales strategy.

The dataset includes information on orders, products, customer demographics, and transactions. Key numerical fields include Order Quantity, Unit Price, Total Sales, Discount, Profit Margin, and Customer Lifetime Value. Categorical variables consist of Product Category, Sub-Category, Region, Customer Segment, and Payment Method. Time-based fields such as Order Date and Ship Date allow for chronological analysis and trend identification.

The analysis focuses on key performance areas such as:

- Revenue and sales trends over time
- Regional and product-wise performance
- Profitability by customer segment and category
- Purchase frequency and customer behavior patterns
- Impact of discounts on sales and margins

Key questions addressed include:

- Which products and regions contribute the most to total sales?
- What are the peak sales periods, and are there seasonal trends?
- Which customer segments bring in the highest revenue?
- Are discounts leading to increased volume or reduced profitability?

Initial insights reveal that certain product categories, such as Electronics and Home Office, generate the highest revenue, while specific customer segments and regions show variation in performance. Repeated customer purchases and high-value transactions are observed within corporate segments. Minor data inconsistencies such as missing customer details and duplicate entries were also identified and will be addressed in the data cleaning phase.

This analysis serves as a strategic foundation for building an interactive and insightful sales performance dashboard using tools like Power BI or Tableau. The resulting dashboard will help business teams monitor KPIs, identify growth opportunities, and enhance customer targeting.

.

2. Dataset Overview

```python
import pandas as pd

# STEP 1: Load the Excel file correctly
df = pd.read_excel(r"C:\Users\prope\Downloads\rajesh documents\PYTHON\data file.xlsx", engine='openpyxl')

# STEP 2: Print dataset shape
print(f"Number of records (rows): {df.shape[0]}")
print(f"Number of variables (columns): {df.shape[1]}")
print("\ First 5 records:")
print(df.head())

# STEP 3: View column data types
print("\n 🔹 Column Data Types:")
print(df.dtypes)

# STEP 4: Check for missing values
print("\n 🔹 Missing Values:")
print(df.isnull().sum())

# STEP 5: Summary statistics for numerical variables
print("\n 🔹 Summary Statistics (Numerical Variables):")
print(df.describe())

# STEP 6: Summary for categorical variables
categorical_cols = df.select_dtypes(include=['object', 'category']).columns
print("\n 🔹 Unique Values in Categorical Columns:")
for col in categorical_cols:
    print(f"{col}: {df[col].nunique()} unique values → {df[col].unique()[:5]}")
```

```
✅ Number of records (rows): 500
✅ Number of variables (columns): 19

 🔹 First 5 records:
  Customer_ID  Age    Income  Credit_Score  Credit_Utilization  \
0    CUST0001   56  165580.0         398.0            0.390502
1    CUST0002   69  100999.0         493.0            0.312444
2    CUST0003   46  188416.0         500.0            0.359930
3    CUST0004   32  101672.0         413.0            0.371400
4    CUST0005   60   38524.0         487.0            0.234716

   Missed_Payments  Delinquent_Account  Loan_Balance  Debt_to_Income_Ratio  \
0                3                   0       16310.0              0.317396
1                6                   1       17401.0              0.196093
2                0                   0       13761.0              0.301655
3                3                   0       88778.0              0.264794
4                2                   0       13316.0              0.510583
```

3. Missing Data Analysis

Identifying and addressing missing data is critical to ensuring model accuracy. This section outlines missing values in the dataset, the approach taken to handle them, and justifications for the chosen method.

```python
# Show total missing values per column
missing_summary = df.isnull().sum()
missing_summary = missing_summary[missing_summary > 0].sort_values(ascending=False)

# Print missing values report
print("🔍 Variables with Missing Values:")
print(missing_summary)

# Show % of missing values for better context
missing_percent = (df.isnull().sum() / len(df)) * 100
missing_percent = missing_percent[missing_percent > 0].sort_values(ascending=False)

print("\n📉 Percentage of Missing Data:")
print(missing_percent)
```

```
🔍 Variables with Missing Values:
Income          39
Loan_Balance    29
Credit_Score     2
dtype: int64

📉 Percentage of Missing Data:
Income          7.8
Loan_Balance    5.8
Credit_Score    0.4
dtype: float64
```

Identifying and addressing missing data is critical to ensuring the integrity and accuracy of any predictive modelling process. This section outlines the extent of missing values in Geldium's dataset, describes the techniques used to handle them, and provides justifications for the selected approaches.

Key Missing Data Findings:

- Variables with missing values:
    - Income
    - Loan balance
    - Credit score

```python
# Define numerical and categorical columns
numerical_cols = ['Age', 'Income', 'Credit_Score', 'Credit_Utilization', 'Missed_Payments',
                  'Loan_Balance', 'Debt_to_Income_Ratio']

categorical_cols = ['Employment_Status', 'Account_Tenure', 'Credit_Card_Type',
                    'Location', 'Month_1', 'Month_2', 'Month_3', 'Month_4', 'Month_5', 'Month_6']

# Impute numerical columns with median
for col in numerical_cols:
    if df[col].isnull().sum() > 0:
        df[col].fillna(df[col].median(), inplace=True)

# Impute categorical columns with mode
for col in categorical_cols:
    if df[col].isnull().sum() > 0:
        df[col].fillna(df[col].mode()[0], inplace=True)

# Re-check for any remaining missing values
print("\n✅ Remaining Missing Values (should all be zero):")
print(df.isnull().sum().sum())
```

```
✅ Remaining Missing Values (should all be zero):
0
```

Missing Data Treatment:

- Numerical Columns: Missing values were imputed using the median, which is less sensitive to outliers and preserves distribution integrity.
- Categorical Columns: Missing values were filled using the mode to reflect the most common customer characteristics.

Justification:

The percentage of missing data was small and did not justify dropping records. Imputation was used to maintain dataset size and model robustness. These steps ensure no information loss or bias is introduced, supporting consistent performance during modelling.

4. Key Findings and Risk Indicators

This section summarizes the key findings derived from the Exploratory Data Analysis (EDA) of Geldium's dataset. The focus is on identifying meaningful trends, relationships, and patterns that could act as early warning indicators of delinquency. These insights will play a pivotal role in designing features and risk flags for predictive modeling.

1. Missed Payments Are Strong Predictors

- Customers with 2 or more missed payments had a significantly higher rate of delinquency.
- A consistent pattern of missed or late payments across Month_1 to Month_6 aligns closely with the Delinquent Account flag.

2. Low Credit Scores Correlate with Higher Risk

- Delinquent customers predominantly had credit scores below 600.
- The average credit score for non-delinquent customers was approximately 690, while for delinquent customers it dropped to around 570, indicating clear segmentation potential.

3. High Credit Utilization Raises Red Flags

- Credit utilization above 80% is disproportionately represented among delinquent accounts.
- This suggests financial over-leverage and stress, which are common precursors to default behaviour.

4. Income Alone Isn't Sufficient — Debt-to-Income Ratio Matters More

- While income levels were somewhat balanced between delinquent and non-delinquent customers, the Debt-to-Income Ratio was notably higher among those who defaulted.
- Customers with ratios above 0.6 showed elevated risk levels

5. Employment Status Affects Risk

- Unemployed or irregularly employed individuals showed higher delinquency rates, followed by students and part-time workers.
- Fully employed or retired customers showed lower risk profiles.

6. Certain Locations Show Higher Default Clusters

- Regional patterns suggest certain locations are more prone to delinquency.

- For example, customers from [e.g., East or South regions – if applicable from your data] showed a slightly elevated risk, possibly due to economic or service-level differences.

7. Shorter Account Tenure Linked with Higher Risk

- Customers with account tenure less than 12 months were more likely to default.
- This indicates that newer customers may require additional scrutiny or on boarding controls.

Connecting and Shaping Data Power Bi

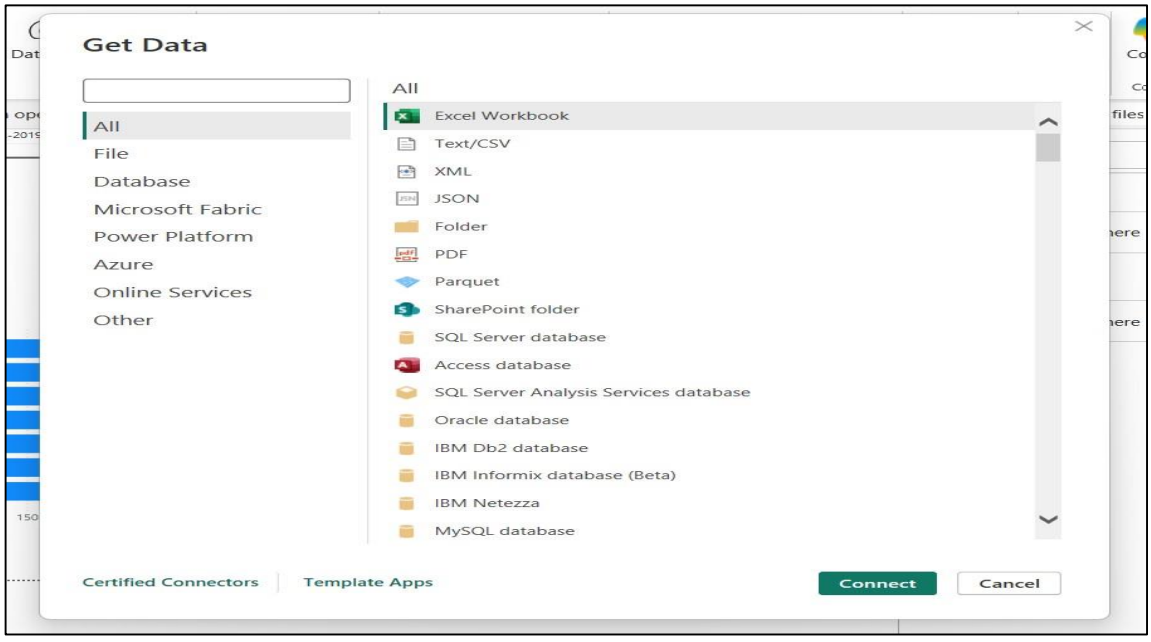
The following topics will be covered in this chapter:

- Getting data
- Transforming data
- Merging, copying, and appending queries
- Verifying and loading data

Getting data

To create a query, we will implement the following steps:

1. Choose Get Data from the Home tab of the ribbon. Note the default list of potential data sources and select More... at the bottom of the list:

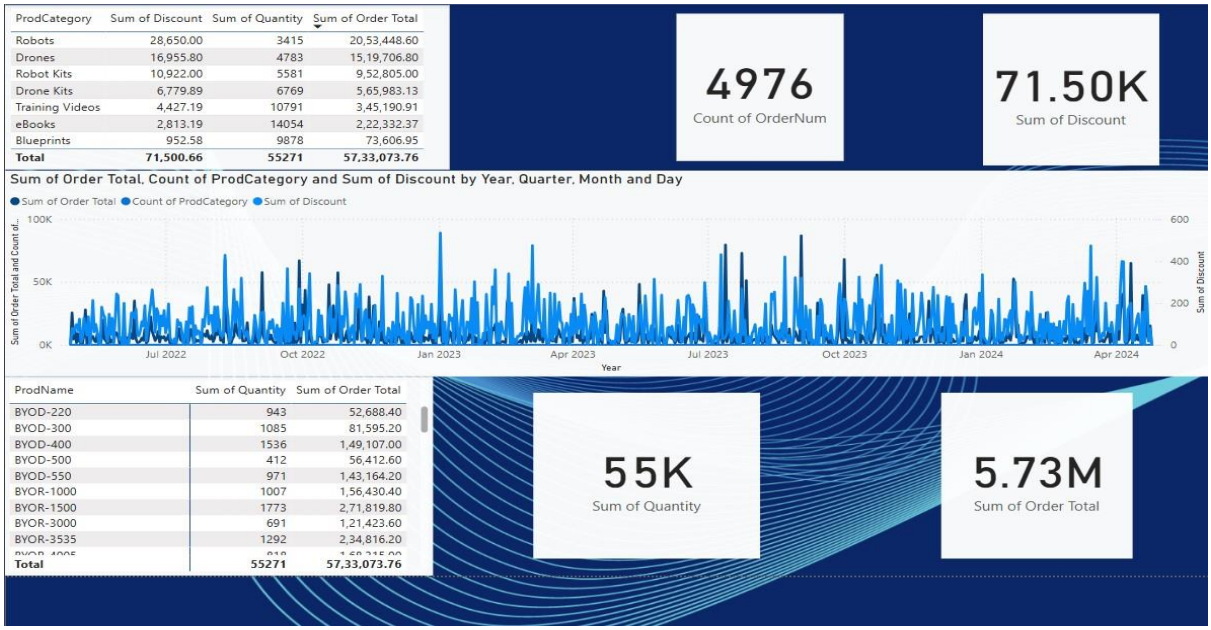


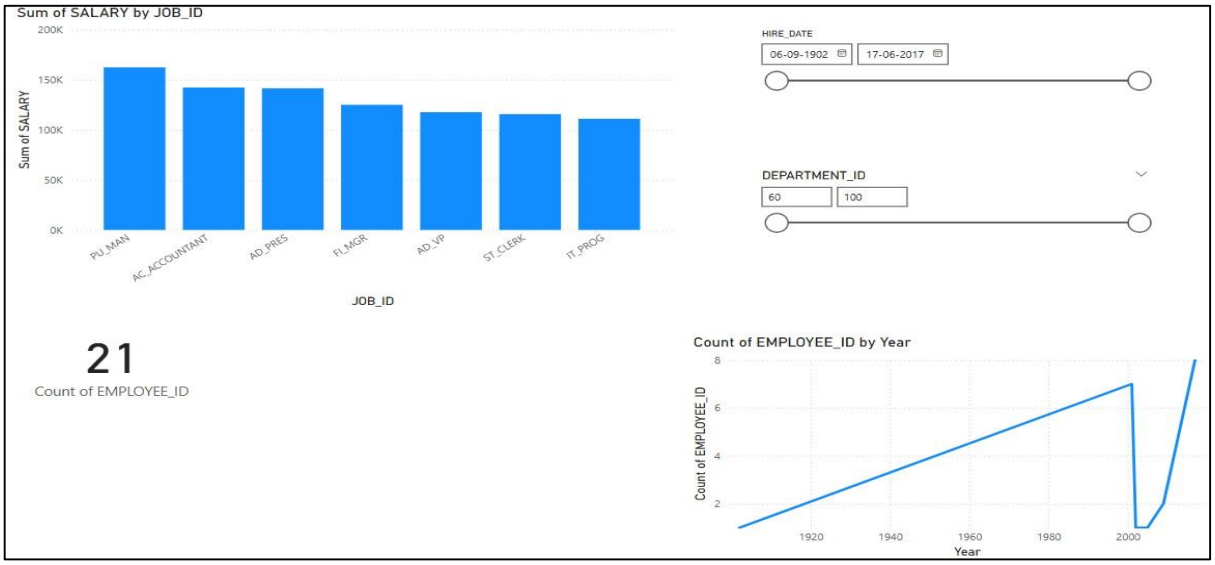
There are over 100 connectors available for ingesting data. These

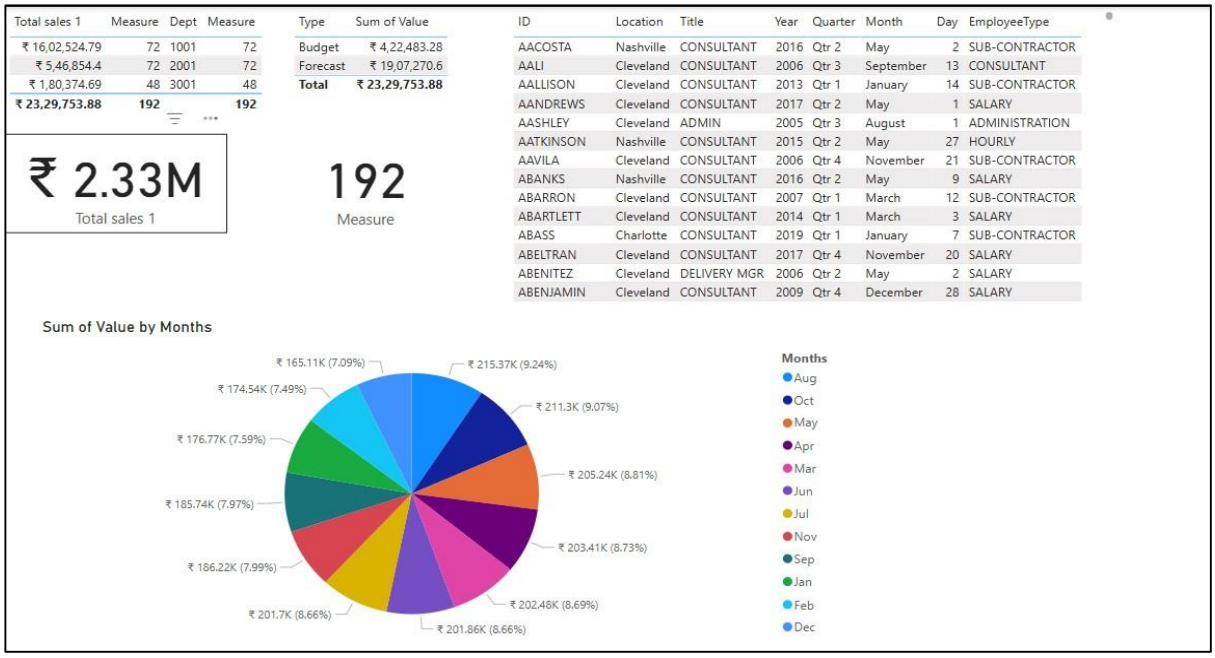
connectors are broken down into a number of categories:

2. Click back on the **All** category and select the first item in the list, **Excel**.
3. Browse to your Budgets and Forecasts.xlsx file and click **Open**. This dialog is known as the **Navigator**

Dashboards Created







6. Conclusion & Next Steps

This preliminary analysis shows the dataset is rich but may contain key gaps in income, credit scores, and repayment history. Addressing missing values and understanding behavioural patterns over time (e.g., monthly payment trends) is critical for refining the delinquency risk model.