EXCEL Task 1

Dataset Overview: PersonalCare_Sales_Cost_Data

Dataset: Health_care_sales_data.xlsx

Description
This dataset contains sales and cost information for various personal care and wellness products from a brand. It provides insights into the financial performance of each product variant, enabling analysis of profitability, sales trends, and cost management across a wide range of categories such as body care, skincare, oral care, and hair care.

Data Structure

- **Columns:**
  - **Product:** The name of the personal care product variant (e.g., mouthwash, body wash, facial cream, shampoo, etc.).
  - **Sales:** The total revenue generated from the sales of each product (measured in units sold or monetary value).
  - **Costs:** The expenses associated with each product, which may include production costs, marketing expenses, and distribution costs.
- **Rows:**
  Each row represents a specific product, providing its sales and cost figures. The dataset covers 50 different personal care products from a variety of categories, providing a comprehensive view of the financial performance of each.

Task Requirements for Healthcare Product Sales and Cost Data

You are provided with a dataset that outlines the sales, costs, and profit margins of various healthcare products. The figures are in INR. Your task is to perform the following steps:

1. Calculate the Profit Margin for Each Product:

- Calculate the **Profit Margin** for each product using the following formula:
  Profit Margin = (Sales - Costs) / Sales
  Display the calculated margin as a percentage with two decimal places.

2. Set Up a Dropdown Menu:

- Create a **Dropdown Menu** in a designated cell that allows users to select from three options: **Sales**, **Costs**, or **Margin**.
  - The options must be directly embedded into the menu, **not** linked to any other cells in the worksheet.
  - Ensure that the dropdown menu is functional and accessible to users.

3. Use a Lookup Function to Display Values:

- In a column next to the **Profit Margin** column, use a **lookup function** to display the corresponding value (Sales, Costs, or Margin) for each product based on the selection made in the dropdown menu.
    - The displayed value should dynamically change depending on the option selected in the dropdown.

4. Create a Dynamic Chart:

- Create a **Dynamic Chart** that visualizes the values (Sales, Costs, or Margin) based on the selection made in the dropdown menu.
    - The chart should update automatically when the dropdown selection changes.
    - Make sure that the chart is clearly labeled and easily readable.

5. Organize and Format Data:

- Organize and format the worksheet to improve readability and ensure a professional appearance.
    - Use appropriate column widths and **bold headers**.
    - Format the **Sales** and **Costs** columns in **currency** (INR).
    - Format the **Profit Margin** column as a **percentage** with two decimal places.
    - Apply **conditional formatting** to highlight values that meet certain criteria (e.g., low profit margins).
    - Ensure that all data is aligned properly (numbers right-aligned, headers centered).
    - Add **borders** to distinguish between different sections of the worksheet.

Excel TASK - 2

Dataset Overview: Bollywood Box Office Collections by Indian States (2012-2019)

Dataset: BoxOfficeINDIA.xlsx

Explanation of the Data:

- **Columns**:
  - **Location**: Represents the state or region within India.
  - **2012 to 2019**: Box office collections by state for each year (in INR).
- **Rows**:
  - Each row represents a **state or region in India**, and the respective box office revenue from Bollywood and Tollywood films for the years 2012 to 2019.

Task Requirements for Bollywood Box Office Collections (2012-2019)

1. **Calculate the Total Box Office Collections for Each Location (Region)**
   Calculate the total box office collections for each location (Indian state/region) from 2012 to 2019. The total should sum the collections across all years for each state.
2. **Calculate the Total Box Office Collections for Each Year**
   Compute the total box office collections across all locations (Indian states) for each year from 2012 to 2019. This should give an overall picture of the box office performance for each year.
3. **Calculate the Average Box Office Collections Per Year**
   Determine the average box office collections for all locations combined per year over the specified period (2012-2019). This average should be calculated for each year.
4. **Count the Total Number of Locations (States/Regions) Included in the Analysis**
   Create a function to count the total number of locations (Indian states or regions) included in the dataset. This will help in understanding how many states contribute to the total collections.
5. **Create a Line Chart for a Specific Location (Region)**
   Create a line chart to illustrate the trend of box office collections for a specific region, such as **Maharashtra**, over the years (2012-2019). This chart will help visualize how the box office performance has changed in that region.
6. **Create a Column Chart for Comparative Analysis**
   Develop a column chart to compare the total box office collections of all locations for the year **2019**. The chart should display the contributions of each state to the total collections in that year.
   Additionally, create a **pie chart** to represent the percentage distribution of total collections recorded by each location over the entire period (2012-2019).

7. **Rename Sheets**
    - Change the name of **Sheet1** to **"Box Office Data"**.
    - Change the name of **Sheet2** to **"Annual Totals"**.
    - Change the name of **Sheet3** to **"Insights"**.
    - After renaming, copy the table headers (locations and years) from the **"Box Office Data"** worksheet and paste them into the **"Annual Totals"** worksheet for reference.
8. **Calculate Yearly Percentages**
    In the **"Annual Totals"** worksheet, compute the **percentage of box office collections** for each location per year in relation to the total box office collections for that year. This can be done using a formula that calculates the percentage for each cell and can be applied across the entire table.
9. **Implement a Rating System**
    Copy the table headers from the **"Box Office Data"** worksheet to the **"Insights"** worksheet.
    In the **"Insights"** worksheet, implement a function that outputs:
    - **"Above Average"** if the box office collections for a location in a given year are greater than the yearly average.
    - **"Below Average"** if they are not.
      This calculation should be done using a single formula that can be replicated across all applicable cells.
10. **Identify Trends in Box Office Collections**
    Identify and highlight any trends or significant changes in box office collections for specific locations between **2018 and 2019**.
    - Provide a **commentary** on the possible implications of these trends based on the data (e.g., which regions have seen growth or decline, and potential factors contributing to these trends)

SQL

Dataset Overview: Analyzing Employee Trends

Dataset:Employee_trends.csv

Problem Statement:

Analyzing Employee Retention and Satisfaction Trends

The goal of this project is to understand employee attrition, satisfaction, and demographics to inform decisions on improving retention and satisfaction within the organization. The dataset includes information on employees' demographics, job roles, education, business travel, and job satisfaction, enabling a comprehensive analysis of factors influencing employee turnover and engagement.

Columns in the dataset:

1. **emp_no**: Employee number (unique identifier for each employee).
2. **gender**: The gender of the employee (e.g., Male, Female).
3. **marital_status**: The marital status of the employee (e.g., Single, Married, Divorced).
4. **age_band**: Age group classification (e.g., 25-34, 35-44, 45-54, Over 55).
5. **age**: Actual age of the employee.
6. **department**: The department where the employee works (e.g., Sales, R&D).
7. **education**: The highest level of education attained by the employee (e.g., High School, Bachelor's Degree, Master's Degree).
8. **education_field**: The field of study the employee pursued (e.g., Life Sciences, Medical, Other).
9. **job_role**: The role or position held by the employee (e.g., Sales Executive, Research Scientist, Laboratory Technician, Manufacturing Director).
10. **business_travel**: The frequency of business travel required by the employee's role (e.g., Travel_Rarely, Travel_Frequently).
11. **employee_count**: Number of employees in the same situation (usually set to 1 for each entry in this dataset).
12. **attrition**: Indicates whether the employee has left the company or still employed (e.g., Yes, No).
13. **attrition_label**: Categorization of the employee's status as either "Ex-Employees" or "Current Employees."
14. **job_satisfaction**: A rating of the employee's job satisfaction (e.g., 1-4 scale, with 1 being low satisfaction and 4 being high satisfaction).
15. **active_employee**: An identifier that shows if the employee is currently active (e.g., 1 for active employees, 0 for inactive).

Analysis questions

1. General Employee Overview

1. Count the total number of employees in the dataset.
2. List the unique job roles and their employee counts.
3. Find the percentage of male and female employees in the company.
4. Count the number of employees in each age_band.
5. Determine the youngest and oldest employees in the dataset.
6. Find the average age of employees.
7. Count the number of employees in each education category.
8. List all distinct education_field values and the corresponding employee counts.
9. Count the number of employees who travel frequently, rarely, or never for business.

---

2. Attrition Analysis

1. Calculate the overall attrition rate in the dataset.
2. Count the number of current employees and ex-employees.
3. Find the attrition rate for each department.
4. List the attrition rates for different job roles.
5. Identify the age_band with the highest attrition rate.
6. Find the attrition rates for male vs. female employees.
7. Compare attrition rates for employees with job satisfaction levels of 1, 2, 3, and 4.
8. Identify departments where employees with low job satisfaction levels are leaving.

---

3. Job Satisfaction Analysis

1. Find the average job satisfaction score for the entire company.
2. Calculate the average job satisfaction score for current employees and ex-employees.
3. Determine which department has the highest average job satisfaction score.
4. Analyze the job satisfaction levels for employees with different education levels.
5. Identify job roles with the lowest average job satisfaction scores.
6. Compare job satisfaction scores for employees who travel frequently vs. rarely.
7. Analyze the average job satisfaction score by age_band.
8. Find the average job satisfaction score for each marital_status.
9. Identify the top 5 employees with the highest job satisfaction scores.
10. Compare the average job satisfaction scores of employees in R&D, Sales, a nd HR.

---

4. Department-Level Analysis

1. Count the number of employees in each department.

2. Find the average age of employees in each department.
3. Identify departments with the highest and lowest attrition rates.
4. Calculate the proportion of male and female employees in each department.
5. Determine the most common job_role in each department.
6. Find the average number of employees who travel frequently in each department.
7. List the number of active employees in each department.
8. Compare attrition rates for different departments by business_travel frequency.
9. Identify departments with a majority of employees in the "35 - 44" age range.

Tableau

Dataset: Walmart_sales_data.csv

About the dataset:

invoice_id: Unique identifier for each transaction.

branch: Branch code where the transaction occurred (e.g., "A", "B", "C").

city: City where the branch is located.

customer_type: Type of customer (e.g., "Normal", "Member").

gender: Gender of the customer.

product_line: Product category involved in the transaction.

unit_price: Price per unit of the product.

quantity: Number of units purchased.

tax_pct: Tax percentage applied.

total: Total amount including tax.

date: Date of the transaction.

time: Time of the transaction.

payment: Payment method used (e.g., "Credit card", "Ewallet", "Cash").

cogs: Cost of goods sold.

gross_margin_pct: Gross margin percentage.

gross_income: Gross income from the transaction.

rating: Customer rating of the service/product.

time_of_day: Time of day when the transaction occurred (e.g., "Morning", "Afternoon").

day_name: Day of the week.

month_name: Month of the transaction.

Project Title: Analyzing Walmart Sales Performance Across Branches

1. Problem Statement:

Walmart operates multiple branches across different cities, offering a diverse range of products. Understanding sales performance, customer behavior, and product trends is crucial for informed decision-making. The current dataset provides a detailed record of transactions, including product lines, customer demographics, and sales details. The challenge is to uncover actionable insights from this data to improve revenue, customer satisfaction, and operational efficiency.

2. Objectives:

1. **Sales Analysis**: Identify top-performing branches, cities, and product lines based on revenue and sales volume.
2. **Customer Behavior**: Analyze the customer demographics, including gender and customer type, and their purchasing patterns.
3. **Revenue Insights**: Evaluate the impact of payment methods on sales and identify high-gross-margin products.
4. **Time-Based Trends**: Examine sales trends by time of day, day of the week, and month to optimize staffing and promotions.
5. **Customer Feedback**: Analyze customer ratings to assess satisfaction levels across branches and product lines.

3. Tasks and Requirements:

Data Preparation

1. Load the dataset into Tableau.
2. Clean and preprocess the data, ensuring date and time fields are properly formatted.
3. Create calculated fields:
    - **Net Sales (Cogs + Gross Income)** for revenue analysis.
    - **Profit Margin Percentage** if additional insights on profitability are required.
    - Time-based segmentation fields (e.g., week number, hour buckets).

Visualizations

1. **Overview Dashboard**:
    - Total revenue, total sales, average rating.
    - Branch and city performance comparison.
2. **Product Line Analysis**:
    - Sales by product line.
    - Top 3 product categories based on gross income.
3. **Demographic Insights**:
    - Gender-wise and customer-type-wise sales distribution.
    - Rating comparison across demographics.
4. **Payment Method Analysis**:

   - Payment method popularity and its correlation with sales revenue.
5. **Time-Based Sales Trends**:
   - Hourly, daily, and monthly sales performance.
   - Heatmap showing peak sales periods across days and hours.

Advanced Features

1. Implement filters for branch, city, product line, and customer type to enable dynamic exploration.
2. Add KPIs (Key Performance Indicators) for revenue, average basket size, and customer satisfaction.
3. Use maps for geographic visualization of branch performance.

Insights and Recommendations

1. Provide a summary of findings based on dashboards.
2. Recommend strategies for underperforming branches.
3. Suggest high-performing product lines for focused promotions.
4. Highlight peak sales periods to optimize staffing and inventory.

Machine Learning

Air Quality and Pollution Assessment

Dataset link: pollution_dataset.csv

Problem Statement:

The goal of this analysis is to develop a predictive model that can assess air quality levels in various regions based on multiple environmental and demographic factors. The dataset contains 5000 samples, with features such as temperature, humidity, concentrations of particulate matter (PM2.5 and PM10), nitrogen dioxide (NO2), sulfur dioxide (SO2), carbon monoxide (CO), proximity to industrial areas, and population density. These factors collectively influence the level of pollution in a region.

The target variable, **Air Quality Levels**, is categorized into four classes:

- **Good**: Clean air with low pollution levels.
- **Moderate**: Acceptable air quality, but with some pollutants present.
- **Poor**: Noticeable pollution that may cause health issues for sensitive groups.
- **Hazardous**: Highly polluted air, posing serious health risks to the population.

The primary objective is to predict the **Air Quality Levels** (Good, Moderate, Poor, Hazardous) for a given region based on the above features, which can aid in environmental monitoring, policy-making, and public health planning.

About Dataset

This dataset focuses on air quality assessment across various regions. The dataset contains 5000 samples and captures critical environmental and demographic factors that influence pollution levels.

Key Features:

- Temperature (°C): Average temperature of the region.
- Humidity (%): Relative humidity recorded in the region.
- PM2.5 Concentration (µg/m³): Fine particulate matter levels.
- PM10 Concentration (µg/m³): Coarse particulate matter levels.
- NO2 Concentration (ppb): Nitrogen dioxide levels.
- SO2 Concentration (ppb): Sulfur dioxide levels.
- CO Concentration (ppm): Carbon monoxide levels.
- Proximity to Industrial Areas (km): Distance to the nearest industrial zone.
- Population Density (people/km²): Number of people per square kilometer in the region.

Target Variable: Air Quality Levels

- Good: Clean air with low pollution levels.
- Moderate: Acceptable air quality but with some pollutants present.
- Poor: Noticeable pollution that may cause health issues for sensitive groups.
- Hazardous: Highly polluted air posing serious health risks to the population.

Approach

To tackle the problem of Air quality prediction, we will employ a systematic machine learning workflow comprising several key tasks:

1. Data Exploration

- Investigate the dataset to understand its structure, identify data types, and uncover initial insights.
- Visualize distributions, relationships, and trends within the data using graphs and statistical measures.

2. Data Cleaning

- Handle missing values through imputation or removal based on their significance and impact.
- Detect and correct inconsistencies or anomalies within the dataset to ensure data integrity.

3. Feature Engineering

- Create new features that could enhance model performance, such as aggregating variables or encoding categorical variables.
- Select relevant features that contribute significantly to predicting air quality while eliminating redundant or irrelevant ones.

4. Model Building

- Experiment with various classical machine learning algorithms to determine which models best fit the data. Potential algorithms to consider include:
    - **Logistic Regression:** A statistical model suitable for binary classification.
    - **Decision Trees:** A non-linear model that makes decisions based on feature values.
    - **Random Forest:** An ensemble method that utilizes multiple decision trees to improve accuracy and robustness.

    - **Gradient Boosting Machines (GBM):** A powerful ensemble technique that builds models in a stage-wise manner to optimize prediction accuracy.
    - **Support Vector Machines (SVM):** A model that identifies the hyperplane that best separates classes in the feature space.

5. Model Testing

- Evaluate the performance of the models using appropriate metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC).
- Utilize cross-validation techniques to ensure the models generalize well to unseen data.