Final Evaluation Project

Excel: Top Global 2000 Largest Companies 2024 Dataset

About Dataset

This dataset contains financial information on the top 2000 global companies in 2024. It includes key metrics such as company name, country, sales, profit, assets, and market value. The data offers valuable insights into the financial performance and market presence of leading companies across various industries worldwide. It is ideal for financial analysis, visualization, and global market research.

Dataset Overview

- **Rows:** 2002
- **Columns:** 7
- **Column Description**
    - **Name:** The name of the company.
    - **Country:** The country where the company is based.
    - **Sales:** Total sales revenue of the company.
    - **Profit:** Net profit of the company.
    - **Assets:** Total assets held by the company.
    - **Market Value:** The market capitalization of the company.

Questions

- **Data Cleaning And Preprocessing**
    - **Convert Text To Numbers:** The Sales, Profit, Assets, and Market Value columns contain financial values with $ and B. Write a formula or use Power Query to convert them into numeric values.
    - **Handle Missing Data:** Identify and highlight missing or inconsistent values in the dataset. Use Excel functions like IFERROR, IFNA, and ISBLANK to handle them appropriately.
- **Data Analysis And Calculations**
    - **Top Performing Companies:** Using formulas, identify the top 10 companies with the highest profit.

- **Market Share Calculation:** Calculate the market share of each company by dividing its market value by the total market value of all companies.
- **Asset To Sales Ratio:** Compute the asset-to-sales ratio for each company and categorize companies as:
  - High Ratio (Ratio > 10)
  - Medium Ratio (5 < Ratio < 10)
  - Low Ratio (Ratio < 5)
- **Profitability Analysis:** Find companies with a Profit Margin above 20%. (Profit Margin = (Profit / Sales) * 100).
- **Conditional Formatting For Profit Growth:** Highlight companies with a Profit greater than $50B using conditional formatting.

- **Advanced Excel Concepts**
  - **Dynamic Data Extraction:** Create a dropdown filter that allows users to select a country and dynamically display the top 10 companies from that country.
  - **Lookup And Match Functions:** Implement a VLOOKUP or INDEX MATCH function to retrieve financial details of a company when its name is entered in a search cell.
  - **Company Segmentation:** Use pivot tables to classify companies into different tiers based on Market Value:
    - **Tier 1:** Market Value > $500B
    - **Tier 2:** $100B < Market Value < $500B
    - **Tier 3:** Market Value < $100B
- **Visualization And Dashboard**
  - **Create an Interactive Dashboard:** Build an Excel dashboard that includes:
    - A **Bar Chart** comparing Sales, Profit, and Assets for the top 10 companies.
    - A **Pie Chart** showing the distribution of companies by country.
    - A **Line Chart** visualizing total market value trends across different countries.
    - Interactive slicers for filtering by country and company size.

SQL: Olympics Dataset

About Dataset

This dataset contains historical records of Olympic athletes, events, and their performance across multiple editions of the Olympic Games.

athelete_events Dataset Overview

- **Rows:** 271117
- **Columns:** 15
- **Column Description**
    - **ID:** A unique identifier for each athlete entry.
    - **Name:** Name of the athlete.
    - **Sex:** Gender of the athlete where M represents male and F represents female.
    - **Age:** Age of the athlete during the event.
    - **Height:** Height of the athlete in centimeters.
    - **Weight:** Weight of the athlete in kilograms.
    - **Team:** The country the athlete represented.
    - **NOC:** The National Olympic Committee (NOC) code, linking to the noc_regions table.
    - **Games:** The specific Olympic Games the athlete participated in.
    - **Year:** The year of the Olympic Games.
    - **Season:** The season of the event.
    - **City:** The host city of the Olympic Games.
    - **Sport:** The sport the athlete competed in.
    - **Event:** The specific event the athlete participated in.
    - **Medal:** The medal won by the athlete.

noc_regions Dataset Overview

- **Rows:** 231
- **Columns:** 3
- **Column Description**

- **NOC:** A unique code representing the country.
- **Region**: The full name of the country or region associated with the NOC.
- **Notes**: Additional notes regarding the NOC, such as historical changes or special designations.

Questions

- **Data Cleaning And Preprocessing**
  - **Convert Data Types:** The age, height, and weight columns may be stored as text. If so, write a query to convert them into appropriate numeric data types.
  - **Handle Missing Data:** Identify and count missing values in the age, height, and weight columns.
  - **Standardize Medal Names:** The medal column may contain inconsistent values. If so, write a query to standardize all values to titlecase.
- **Data Analysis And Calculations**
  - **Top Medal-Winning Countries:** Find the top 10 countries with the highest number of total medals won.
  - **Medal Distribution By Season:** Count the number of medals won by the top 10 countries in the Summer Games versus Winter Games.
  - **Athlete Performance Analysis:** Identify the top 5 athletes with the highest number of Olympic medals.
  - **Country With The Highest Gold Medal Ratio:** Find the country with the highest percentage of Gold medals out of its total medals.
- **Advanced SQL Concepts**
  - **Dynamic Country-Based Medal Count:** Create a query that allows users to filter medals won by a selected country.
  - **Find Athletes Who Competed In Multiple Sports:** Retrieve a list of athletes who participated in more than one sport.
  - **Most Successful Athletes Per Country:** Find the most successful athlete (most medals won) from each country.
  - **Event Popularity Analysis:** Identify the top 5 most popular events based on the number of athletes participating.
- **Joins And Multi-Table Queries**
  - **Region-Based Medal Analysis:** Find the total number of medals won by each NOC.

- **Countries That Have Never Won a Medal:** List all countries that have participated in the Olympic Games but never won a medal.

Tableau: World Population Analysis Dataset

About Dataset

This dataset provides comprehensive insights into the population, geographical distribution, and growth dynamics of different countries worldwide. It includes key metrics such as land and water area, birth and death rates, migration trends, and population growth rates. This data is valuable for analyzing global demographic patterns, understanding population distribution, and forecasting future growth trends.

Dataset Overview

- **Rows:** 261
- **Columns:** 11
- **Column Description**
  - **Serial Number:** A unique identifier for each country in the dataset.
  - **Country Code:** A standardized country code representing each nation.
  - **Country Name:** The full name of the country.
  - **Area:** The total geographical area of the country, including both land and water, in square kilometers.
  - **Land Area:** The total landmass of the country, in square kilometers.
  - **Water Area:** The total water-covered area in the country, in square kilometers.
  - **Population:** The population of the country.
  - **Population Growth Rate:** The percentage increase or decrease in the country's population annually.
  - **Birth Rate:** The number of live births per 1,000 people per year.
  - **Death Rate:** The number of deaths per 1,000 people per year.
  - **Migration Rate:** The net migration rate, representing the number of people migrating into or out of a country per 1,000 people annually.

Questions

- **Data Cleaning And Preprocessing**
    - **Data Normalization:** The Population Growth Rate, Birth Rate, Death Rate, and Migration Rate columns contain percentages. Convert them into decimal format using calculated fields.
    - **Custom Grouping:** Group countries into three categories based on population size:
        - **Small Population:** Population < 10 million
        - **Medium Population:** 10 million < Population < 100 million
        - **Large Population:** 100 million > Population
    - Use a calculated field to dynamically assign each country to a category.
- **Complex Calculations and Advanced Functions**
    - **Running Total Analysis:** Create a running total chart to show cumulative population growth over time for the top 10 most populous countries.
    - **Weighted Growth Calculation:** Compute the weighted average population growth rate by considering both Population and Population Growth Rate.
    - **Dynamic Rank Calculation:** Use Tableau's RANK function to create a visualization that dynamically ranks countries by Population Growth Rate, updating automatically based on filters.
- **Advanced Filtering And Parameterized Visualizations**
    - **Dynamic Country-Based Medal Count:** Create a query that allows users to filter medals won by a selected country.
    - **Find Athletes Who Competed In Multiple Sports:** Retrieve a list of athletes who participated in more than one sport.
    - **Most Successful Athletes Per Country:** Find the most successful athlete (most medals won) from each country.
    - **Event Popularity Analysis:** Identify the top 5 most popular events based on the number of athletes participating.
- **Joins and Multi-Table Queries**
    - **Region-Based Medal Analysis:** Find the total number of medals won by each NOC.

- **Countries That Have Never Won A Medal:** List all countries that have participated in the Olympic Games but never won a medal.

Machine Learning: Hotel Booking Dataset

About Dataset

This dataset provides comprehensive insights into the population, geographical distribution, and growth dynamics of different countries worldwide. It includes key metrics such as land and water area, birth and death rates, migration trends, and population growth rates. This data is valuable for analyzing global demographic patterns, understanding population distribution, and forecasting future growth trends.

Dataset Overview

- **Rows:** 119391
- **Columns:** 32
- **Column Description**
  - **hotel:** Type of hotel.
  - **is_canceled:** Target variable, indicating whether the booking was canceled (1) or not (0).
  - **lead_time:** Number of days between the booking date and the arrival date.
  - **arrival_date_year:** Year of arrival.
  - **arrival_date_month:** Month of arrival.
  - **arrival_date_week_number:** Week number of the arrival date.
  - **arrival_date_day_of_month:** Day of the month of arrival.
  - **stays_in_weekend_nights:** Number of weekend nights (Saturday or Sunday) the guest stayed.
  - **stays_in_week_nights:** Number of weeknights (Monday to Friday) the guest stayed.
  - **adults:** Number of adults included in the booking.
  - **children:** Number of children included in the booking.
  - **babies:** Number of babies included in the booking.
  - **meal:** Type of meal booked.

- **country:** Country of origin of the guest.
- **market_segment:** Market segment designation.
- **distribution_channel:** Booking distribution channel.
- **is_repeated_guest:** Whether the guest is a repeated guest (1) or not (0).
- **previous_cancellations:** Number of previous cancellations by the guest.
- **previous_bookings_not_canceled:** Number of previous bookings not canceled.
- **reserved_room_type:** Code for the room type reserved.
- **assigned_room_type:** Code for the room type assigned.
- **booking_changes:** Number of changes made to the booking.
- **deposit_type:** Type of deposit made for the booking.
- **agent:** ID of the travel agency that made the booking.
- **company:** ID of the company that made the booking.
- **days_in_waiting_list:** Number of days the booking was on the waiting list.
- **customer_type:** Type of customer.
- **adr:** Average daily rate for the stay.
- **required_car_parking_spaces:** Number of car parking spaces required.
- **total_of_special_requests:** Number of special requests made.
- **reservation_status:** Final status of the reservation.
- **reservation_status_date:** Date of the last status change.

Questions

- **Data Cleaning And Preprocessing**
  - **Missing Data Handling:** Identify and handle missing values in columns like children, country, and agent. Decide whether to impute missing values or remove rows with missing data.
  - **Outlier Detection:** Check for anomalies in columns like adr (extreme values) and lead_time (very high values may need handling).
  - **Data Type Conversion:** Ensure all data columns have the correct data types.

- **Feature Engineering**
    - **Feature Scaling:** Apply scaling techniques to numerical features.
    - **Categorical Grouping:** Group bookings into three categories based on lead_time:
        - **Short:** lead_time < 30 days
        - **Medium:** 30 days < lead_time < 90 days
        - **Large:** 90 days > lead_time
- **Complex Calculations And Advanced Functions**
    - **Running Total Analysis:** Create a running total chart to show cumulative booking cancellations over time.
    - **Weighted Booking Score:** Compute a weighted cancellation rate by considering both previous_cancellations and previous_bookings_not_canceled.
    - **Dynamic Rank Calculation:** Use ranking techniques to dynamically rank hotels based on their cancellation rate.
- **Advanced Filtering And Parameterized Visualizations**
    - **Dynamic Market Segment Analysis:** Create a query that allows users to filter bookings by market segment dynamically.
    - **Find High-Risk Customers:** Retrieve a list of customers who had multiple cancellations in the past.
    - **Most Frequent Guests:** Identify the most frequent guests based on repeated bookings.
    - **Seasonal Booking Trends:** Identify peak booking months and trends across different hotels.
- **Joins And Multi-Table Queries**
    - **Hotel-Specific Cancellation Rate:** Find the total number of cancellations per hotel type.
    - **Countries With High Cancellation Rates:** List all countries from which guests have a high tendency to cancel bookings.
- **Model Building**
    - Split the Dataset: Split the dataset into training and testing sets (e.g., 80% training, 20% testing).
    - Experiment with models, mainly logistic regression, decision tree, random forest, gradient boosting machines, support vector machines.

    - **Hyperparameter Tuning:** Use Grid Search or Random Search to optimize hyperparameters.

- **Model Testing**
    - Evaluate performance using accuracy, precision, recall, F1-score and AUC-ROC.
    - Use cross-validation techniques.
    - Compare model performance and determine the best one based on evaluation metrics.
