Mini Project 2: Semi-structured Data Processing

IST652 - Scripting for Data Analysis

Khushi Shetty
SUID: 855125581

Exploratory Data Analysis of Nobel Prize Winners dataset using python.

Data and its Source

The datasets used in this analysis is the "prize.json" and “laureate.json”.

prize.json (Nobel Prize Information)
This dataset provides detailed information about Nobel Prize awards, including the year of the award, the category of the prize (e.g., physics, chemistry, peace), laureate details (such as their names and unique identifiers), the motivation behind the award, and the share of laureates if multiple individuals received the prize.

laureate.json (Nobel Laureate Information)
This dataset encompasses comprehensive data about Nobel laureates, covering details such as their unique identifiers, names, birth and death information, birthplaces (country and city), deathplaces, and gender. It serves as a fundamental resource for understanding the backgrounds and demographics of Nobel laureates across various fields.

These datasets together offer a comprehensive view of Nobel Prize history, providing insights into both the awards themselves and the individuals who have made significant contributions in their respective fields. Researchers and analysts can leverage this information to explore patterns, trends, and characteristics of Nobel laureates and their achievements over time.

Data Source: Github
Awesome-json-datasets(github.com)
Nobel Prize Information: api.nobelprize.org/v1/prize.json
Nobel Laureate Information: api.nobelprize.org/v1/laureate.json

Data Exploration and Data Cleaning

The data exploration and cleaning process consisted of several crucial steps to prepare the dataset for analysis:

Data Exploration:

1. **Loading Data:** The code begins by loading data from JSON files (`laureate.json` and `prize.json`), which contain information about Nobel laureates and prizes.
2. **Laureate Information Extraction:** The code then extracts relevant information about laureates, including their IDs, names, birth and death details, countries, and gender.
3. **Prize Information Extraction:** Similarly, information about Nobel Prizes is extracted, including the year, category, laureate details, motivation, and share.
4. **Dataframe Creation:** Two Pandas DataFrames (`df_laureates` and `df_prizes`) are created to store the extracted information.

Data Cleaning:

1. **Column Dropping:** Unnecessary columns, such as 'bornCountryCode' and 'diedCountryCode', are dropped from `df_laureates`.
2. **Birth and Death Year Extraction:** The 'born' and 'died' columns in `df_laureates` are used to create 'birth_year' and 'death_year' columns, respectively.
3. **Handling Non-Numeric Values:** Non-numeric values in 'birth_year' and 'death_year' are identified, replaced with NaN, and converted to integers.
4. **Handling Missing Values (NaT):** The code checks for NaT (Not a Time) values in 'born' and 'died', which represent missing or unknown date values.
5. **Age Calculation:** Age at death is calculated and added as a new column ('age_at_death') to `df_laureates`.
6. **Converting to Datetime:** The 'born' and 'died' columns are converted to datetime format for consistency.

These steps collectively prepare the data for analysis, addressing issues such as missing values, inconsistencies, and outliers.

Comparison Questions:

I addressed four key comparison questions in this analysis:

Analysis 1: Count of Entries by Born Country and Gender

Unit of Analysis: Born country of Nobel laureates
Comparison: To understand the distribution of Nobel laureates based on their born country and gender.

Computation:

- **Data Filtering:** Rows with 'bornCountry' labeled as 'Unknown' are excluded from the analysis to focus on known born countries.
- **Grouping:** The data is then grouped by 'bornCountry' and 'gender' to calculate counts for each combination.
- **Total Calculation:** The 'Total' column is added to represent the sum of female and male laureates for each born country.
- **Sorting:** The DataFrame is sorted based on the 'Total' column in descending order to highlight countries with the highest total number of laureates.

Summary of Findings:

1. **Top Countries:** The analysis reveals the distribution of Nobel laureates across various countries based on gender.
    - The United States (USA) has the highest total number of laureates (289), with 17 female and 272 male laureates.
    - The United Kingdom follows with 89 laureates, all of whom are male.
    - Germany, France, and Sweden also have notable counts of laureates.
2. **Gender Distribution:** The analysis provides insights into the gender distribution of laureates within each country.
    - Some countries, like the USA, show a more balanced distribution between male and female laureates, while others, like the United Kingdom, have a predominantly male representation.
    - There are instances of countries with only male laureates, indicating historical gender imbalances in Nobel laureates' representation from those regions.
3. **Total Laureates:** The 'Total' column showcases the overall contribution of each country to the pool of Nobel laureates. Sorting by this column highlights countries with the highest overall impact on Nobel recognition.

Analysis 2: Nobel Prize Distribution Trends Over 10-Year Intervals

Unit of Analysis: 10-year intervals based on the year of awarding the Nobel Prize.

Comparison: Understanding the distribution of Nobel Prizes across different 10-year intervals and exploring trends in the variety of recognized categories.

Computation:

- **DataFrame Preparation:** A copy of the original DataFrame (df_prize) is created to avoid modifying the original data.
- **Datetime Conversion:** The 'year' column is converted to datetime for better handling.

- **Year Interval Extraction:** Year intervals are created by binning the years into 10-year intervals, starting from 1900.
- **Grouping and Summary Statistics:** The data is grouped by 'year_interval,' and summary statistics are calculated, including the total number of prizes and the count of unique prize categories.
- **DataFrame Presentation:** The resulting DataFrame (prize_summary_statistics) is organized with 'year_interval' as the index for better clarity.

Summary of Findings:

1. **Total Prizes:** The analysis provides insights into the distribution of Nobel Prizes across different 10-year intervals, showing variations in the total number of prizes awarded over time.
2. **Unique Categories:** The count of unique prize categories is examined, reflecting the diversity and stability in the fields recognized by the Nobel Committee.

Key Observations:

- Across the analyzed intervals, the total number of prizes awarded varied, reaching a peak in the 2000s.
- The number of unique prize categories remained relatively consistent, suggesting a stable diversity in the fields recognized by the Nobel Committee over the years.
- This analysis contributes to understanding the temporal trends in Nobel Prize distribution, shedding light on the evolution of recognition in various fields over 10-year intervals.

Analysis 3: Frequent Laureates and Their Prize Information

Unit of Analysis: Individual laureates with multiple Nobel Prizes.
Comparison: Understanding the prize distribution and categories for laureates who have received the Nobel Prize more than once.
Computation:

- Frequencies Calculation: Counting the frequencies of each laureate's ID in the df_prize DataFrame.
- Filtering Frequent Laureates: Identifying laureates with frequencies (number of prizes won) greater than 1.
- Data Extraction and Formatting: For each frequent laureate, extracting unique prize categories, years of awards, and personal information such as first name, surname, and frequency of Nobel Prizes.
- DataFrame Creation: Organizing the extracted information into a new DataFrame (frequent_laureates_df).

- DataFrame Presentation: Displaying the new DataFrame with details about frequent laureates and their Nobel Prize information.

Summary of Findings:

The analysis delves into Nobel laureates who have received the prestigious prize on multiple occasions, shedding light on their diverse contributions across categories and years.

International Committee of the Red Cross (ID 482):

- Recognized for Peace in 1917, 1944, and 1963.
- Stands out with the highest frequency (3), symbolizing consistent impactful contributions.

Linus Pauling (ID 217):

- Awarded Peace in 1962 and Chemistry in 1954.
- Illustrates expertise spanning multiple disciplines.

Marie Curie (ID 6):

- Honored twice, showcasing excellence in Physics and Chemistry.
- Highlights the rare accomplishment of excelling in distinct scientific domains.

Office of the United Nations High Commissioner (ID 515):

- Peace Prize recipient in 1981 and 1954.
- Represents the diplomatic realm's enduring impact on global affairs.

John Bardeen (ID 66), Barry Sharpless (ID 743) and Frederick Sanger (ID 222):

- These laureates have been awarded twice in their fields.
- Exemplifies enduring excellence in a specific scientific discipline.

Key Observations:

- **Interdisciplinary Excellence:** Instances like Linus Pauling's recognition in both Peace and Chemistry, and Marie Curie’s recognition in Physics and Chemistry, underscore laureates' prowess across diverse domains.
- **Organizational Recognition:** The International Committee of the Red Cross and the Office of the United Nations High Commissioner showcase entities making impactful contributions.
- **Longitudinal Impact:** Frequent laureates' contributions span decades, emphasizing sustained excellence in their respective fields.

- This analysis offers a closer look at laureates who have made significant contributions, as evidenced by their multiple Nobel Prize recognitions, and provides a detailed overview of their achievements across different prize categories and years.

Analysis 4: Age and Gender Distribution of Nobel Laureates at the Time of Receiving the Prize

This analysis integrates two dataframes, df_prize and df_laureates, to uncover insights into the age and gender distribution of Nobel laureates at the time of being awarded the prestigious prize.

Unit of Analysis: Nobel laureates

Comparison: Understanding the age and gender distribution of Nobel laureates at the time of receiving the prize.

Computation:

- **Data Integration:**
  Merged two dataframes, df_prize and df_laureates, on the laureate_id column.
- **Age Calculation:**
  Derived the laureates' age at the time of receiving the Nobel Prize by subtracting their birth year from the prize year.
- **Age Grouping:**
  Binned laureates into age groups spanning 10-year intervals.
  Age categories include 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, and 90+.
- **Gender Differentiation:**
  Categorized laureates into 'Male' and 'Female,' excluding any organizational entries ('org').
- **Statistical Insights:**
  Provided a breakdown of female and male laureates within each age group.
  Calculated the total count and the percentage distribution of males and females in each age category.

Summary of Findings:

1. Age Group Patterns:
   - Age groups 50-69 exhibit the highest laureate counts, suggesting a period of heightened scientific and cultural contributions.
   - Notable peaks are observed in the 60-69 age group.
2. Gender Distribution:
   - Despite age variations, gender distribution is relatively balanced.
   - Older age groups show a historical gender disparity, with males dominating.
3. Impactful Insights:
   - The integrated dataset offers a comprehensive view, enabling nuanced observations about laureates' ages and gender dynamics.
   - Recognizing patterns in the age at which laureates receive Nobel Prizes contributes to understanding the evolution of achievements over time.

This analysis enhances our understanding of Nobel laureate demographics, providing valuable insights into the intersection of age and gender at the pinnacle of academic and scientific recognition.

Program Description

This Python-based analysis program encompasses data extraction, cleaning, and exploratory data analysis on Nobel laureates and prizes. Leveraging Pandas, NumPy, and Matplotlib, it dissects two datasets (df_laureates and df_prize), offering valuable insights into Nobel laureate demographics, prize categories, and temporal trends.

This analysis program provides a comprehensive understanding of Nobel laureate demographics, prize distribution, and trends. It equips users with actionable insights for historical context and future considerations in Nobel recognition.

Output Files

Four primary output files were generated as part of this analysis:

1. **"country_gender_distribution.csv":**
   This CSV file explores the distribution of Nobel laureates based on their birth countries and genders. It provides a comprehensive overview, highlighting the countries with the highest counts of laureates and the corresponding gender distribution. The file contains data for various countries, including the United States, the United Kingdom, Germany, and France.

2. **"prize_summary_statistics.csv":**
   This CSV file provides a summary of Nobel Prize statistics over the years, offering insights into the total number of prizes awarded and the diversity of prize categories. It spans different intervals of ten years, revealing trends in the total number of prizes awarded and the consistency of unique prize categories recognized by the Nobel Committee.

3. **"gender_age_counts.csv":**
   Focusing on Nobel Prize recipients, this CSV file explores the distribution of laureates across age groups and genders at the time of receiving the prize. It reveals that the majority of Nobel Prize recipients fall in the 50-69 age range, with a specific emphasis on male laureates, especially in older age groups. The file includes data on age groups such as 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, and 90+.

4. **"frequent_laureates_df.csv":**
   This CSV file compiles information about laureates who have received Nobel Prizes more than once. It includes details about laureates, such as Linus Pauling and John Bardeen, who have been recognized in different categories in different years. The presence of the International Committee of the Red Cross is also noted, reflecting collective recognition.