ML/DL Track - Task 4

Created	@June 26, 2024 2:20 PM
Reviewed	☑

Nobel Prize Data Analysis Report

Introduction

This report analyzes the Nobel Prize data to identify various patterns and answers specific questions. The dataset includes information about Nobel Prize winners, including their gender, birth country, category of the prize, and other relevant details. The questions addressed in this report are:

1. What is the most commonly awarded gender and birth country?
2. Which decade had the highest ratio of US-born Nobel Prize winners to total winners in all categories?
3. Which decade and Nobel Prize category combination had the highest proportion of female laureates?
4. Who was the first woman to receive a Nobel Prize, and in what category?
5. Which individuals or organizations have won more than one Nobel Prize throughout the years?

Data Preparation

The dataset was loaded into a pandas DataFrame and the necessary columns were identified:

- `year`
- `category`
- `sex`
- `birth_country`
- `full_name`

Analysis and Results

1. Most Commonly Awarded Gender and Birth Country

The most frequently awarded gender and birth country were calculated using the `describe()` method on the respective columns.

```
x = df['sex'].describe().top
y = df['birth_country'].describe().top

print("Top Gender:", x)
print("Top Birth Country:", y)
```

Results:

- **Top Gender:** Male
- **Top Birth Country:** United States of America

2. Decade with the Highest Ratio of US-born Nobel Prize Winners to Total Winners

A new column `decade` was created by flooring the `year` divided by 10. Another column `usa_born_winner` indicated if the laureate was born in the USA. The mean proportion of USA-born winners by decade was calculated and the decade with the highest proportion identified.

```
df['decade'] = (np.floor(df['year'] / 10) * 10).astype(int)
df['usa_born_winner'] = df['birth_country'] == 'United States
of America'

df2 = df.groupby('decade', as_index=False)['usa_born_winne
r'].mean()
max_decade_usa = int(df2.loc[df2['usa_born_winner'].idxmax()]
['decade'])

print("Max Decade USA:", max_decade_usa)
```

Results:

- **Decade with Highest Ratio of US-born Winners:** 2000s

3. Decade and Nobel Prize Category Combination with the Highest Proportion of Female Laureates

The total and female laureates per decade and category were calculated. The proportion of female laureates for each combination was computed and the highest identified.

```
total_by_decade_category = df.groupby(['decade', 'categor
y']).size()
female_by_decade_category = df[df['sex'] == 'Female'].groupby
(['decade', 'category']).size()
proportion_female = (female_by_decade_category / total_by_dec
ade_category).reset_index(name='proportion_female')

max_female_row = proportion_female.loc[proportion_female['pro
portion_female'].idxmax()]
max_female_dict = {int(max_female_row['decade']): max_female_
row['category']}

print("Max Female Dict:", max_female_dict)
```

Results:

- **Decade and Category with Highest Proportion of Female Laureates:** {2000: 'Peace'}

4. First Woman to Receive a Nobel Prize and in What Category

The data was filtered to include only female laureates, sorted by year, and the first row identified.

```
female_laureates = df[df['sex'] == 'Female'].sort_values(by
='year')
first_female_laureate = female_laureates.iloc[0]
first_woman_name = first_female_laureate['full_name']
first_woman_category = first_female_laureate['category']

print("First Woman Name:", first_woman_name)
print("First Woman Category:", first_woman_category)
```

Results:

- **First Woman Name:** Marie Curie
- **First Woman Category:** Physics

5. Individuals or Organizations with More Than One Nobel Prize

The number of prizes for each laureate was counted and those with more than one prize were identified.

```
repeat_winners = df['full_name'].value_counts()
repeat_list = repeat_winners[repeat_winners > 1].index.tolist
()

print("Repeat Winners:", repeat_list)
```

Results:

- **Repeat Winners:**

- Marie Curie
- Linus Pauling
- John Bardeen
- Frederick Sanger
- International Committee of the Red Cross
- United Nations High Commissioner for Refugees (UNHCR)

Conclusion

This analysis of the Nobel Prize data provided insights into the most commonly awarded gender and birth country, identified the decade with the highest proportion of US-born winners, and found the decade and category combination with the highest proportion of female laureates. It also revealed the first woman to receive a Nobel Prize and listed individuals and organizations that have won more than one Nobel Prize.

This analysis demonstrates the value of historical data in identifying trends and patterns in prestigious awards like the Nobel Prize.