

A Visual History of Nobel Prize Winners

November 25, 2019

1 Introduction - The Nobel Prize

The Nobel Prize is perhaps the world's most well known scientific award. Except for the honor, prestige and substantial prize money the recipient also gets a gold medal showing Alfred Nobel (1833 - 1896) who established the prize. Every year it's given to scientists and scholars in the categories chemistry, literature, physics, physiology or medicine, economics, and peace. The first Nobel Prize was handed out in 1901, and at that time the Prize was very Eurocentric and male-focused but it seems changed nowadays. And we are also curious about many other things about Nobel Prizers like the ages of people usually got Nobel Prize.

So, we're going to explore Nobel Prize dataset! The dataset made by Nobel Foundation has made available of all prize winners from the start of the prize, in 1901, to 2016. I download it from Kaggle.

Data Source: <https://www.kaggle.com/nobelfoundation/nobel-laureates>

Picture Credits: https://s3.amazonaws.com/assets.datacamp.com/production/project_441/img/Nobel_Prize.jpg

2 Data Importing

```
In [1]: # Loading in required libraries
import pandas as pd
import seaborn as sns
import numpy as np
import matplotlib.pyplot as plt

# Reading in the Nobel Prize data
nobel = pd.read_csv("nobel.csv")

# Taking a look at the first several winners
nobel.head(6)
```

```
Out[1]:
```

	year	category	prize \
0	1901	Chemistry	The Nobel Prize in Chemistry 1901
1	1901	Literature	The Nobel Prize in Literature 1901
2	1901	Medicine	The Nobel Prize in Physiology or Medicine 1901
3	1901	Peace	The Nobel Peace Prize 1901
4	1901	Peace	The Nobel Peace Prize 1901
5	1901	Physics	The Nobel Prize in Physics 1901

		motivation	prize_share	laureate_id	\
0	"in recognition of the extraordinary services ...		1/1	160	
1	"in special recognition of his poetic composit...		1/1	569	
2	"for his work on serum therapy, especially its...		1/1	293	
3		NaN	1/2	462	
4		NaN	1/2	463	
5	"in recognition of the extraordinary services ...		1/1	1	

	laureate_type	full_name	birth_date	birth_city	\
0	Individual	Jacobus Henricus van 't Hoff	1852-08-30	Rotterdam	
1	Individual	Sully Prudhomme	1839-03-16	Paris	
2	Individual	Emil Adolf von Behring	1854-03-15	Hansdorf (Lawice)	
3	Individual	Jean Henry Dunant	1828-05-08	Geneva	
4	Individual	Frédéric Passy	1822-05-20	Paris	
5	Individual	Wilhelm Conrad Röntgen	1845-03-27	Lennepe (Remscheid)	

	birth_country	sex	organization_name	organization_city	\
0	Netherlands	Male	Berlin University	Berlin	
1	France	Male	NaN	NaN	
2	Prussia (Poland)	Male	Marburg University	Marburg	
3	Switzerland	Male	NaN	NaN	
4	France	Male	NaN	NaN	
5	Prussia (Germany)	Male	Munich University	Munich	

	organization_country	death_date	death_city	death_country
0	Germany	1911-03-01	Berlin	Germany
1	NaN	1907-09-07	Châtenay	France
2	Germany	1917-03-31	Marburg	Germany
3	NaN	1910-10-30	Heiden	Switzerland
4	NaN	1912-06-12	Paris	France
5	Germany	1923-02-10	Munich	Germany

In [2]: nobel.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 911 entries, 0 to 910
Data columns (total 18 columns):
year                911 non-null int64
category            911 non-null object
prize               911 non-null object
motivation          823 non-null object
prize_share         911 non-null object
laureate_id         911 non-null int64
laureate_type       911 non-null object
full_name           911 non-null object
birth_date          883 non-null object
birth_city          883 non-null object
```

```

birth_country      885 non-null object
sex                885 non-null object
organization_name   665 non-null object
organization_city   667 non-null object
organization_country 667 non-null object
death_date         593 non-null object
death_city         576 non-null object
death_country      582 non-null object
dtypes: int64(2), object(16)
memory usage: 128.2+ KB

```

3 Who Gets the Nobel Prize? - Univariate Plots

Just looking at the first couple of prize winners, or Nobel laureates as they are also called, we already see a celebrity: Wilhelm Conrad Röntgen, the guy who discovered X-rays. And actually, we see that all of the winners in 1901 were males that came from Europe. But that was back in 1901, looking at all winners in the dataset, from 1901 to 2016, which sex and which country is the most commonly represented? And any other interesting thing we would find?

Let's look at several Univariate Plots of this dataset.

```

In [3]: # Display the number of (possibly shared) Nobel Prizes handed
        # out between 1901 and 2016
        display(len(nobel))

```

911

```

In [4]: sns.set()
        # Setting the size of all plots.
        plt.rcParams['figure.figsize'] = [13, 7]

```

3.1 Trend in Nobel Prize

```

In [5]: # Trend in Nobel Prize
        year=nobel['year'].value_counts()
        sns.lineplot(x=year.index,y=year.values,color='red')

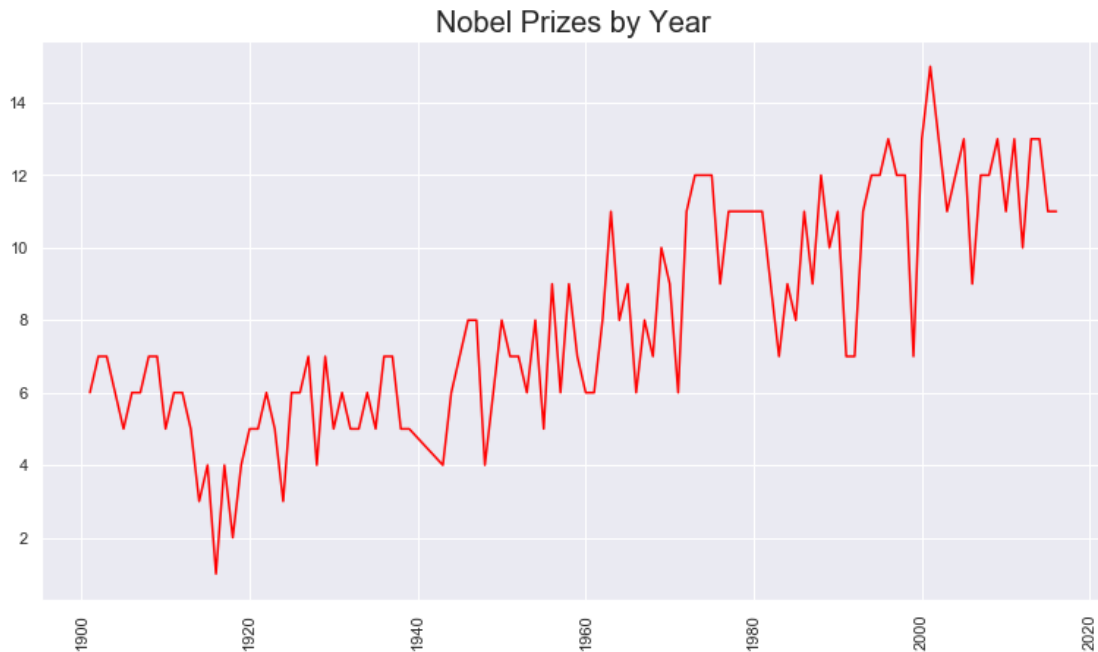
        plt.xticks(rotation=90)
        plt.title('Nobel Prizes by Year',fontsize=20)

```

```

Out[5]: Text(0.5, 1.0, 'Nobel Prizes by Year')

```



There is clear trend of ups and down of nobel prizes issued from 1901 to 2016. While overall, there is a growth trend. The prizes are increasing as time goes.

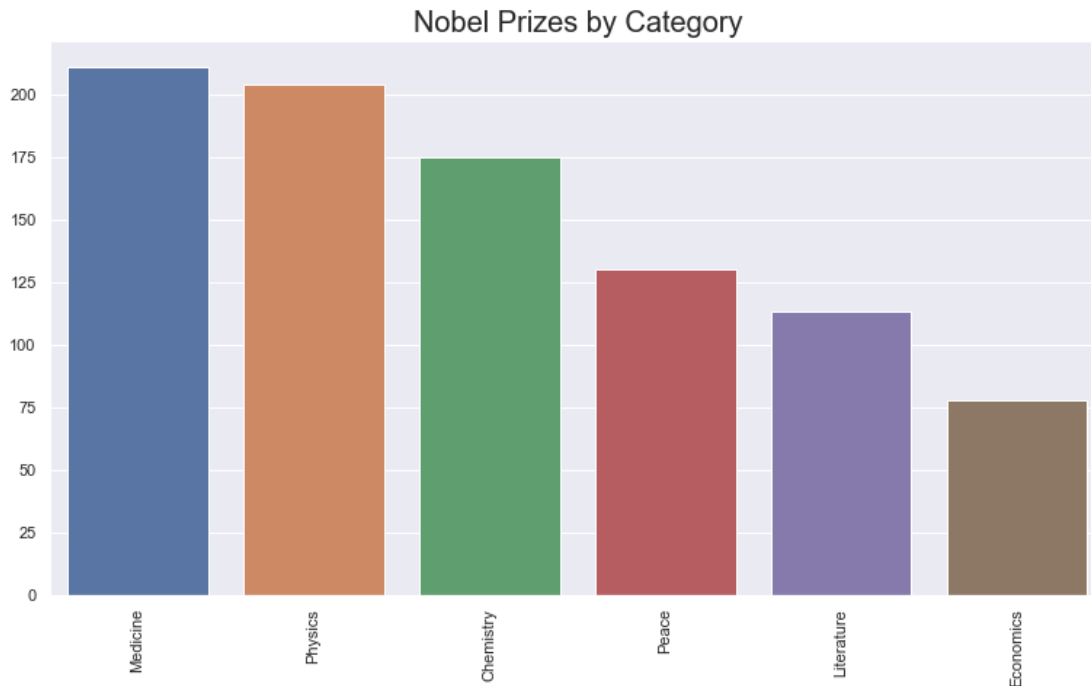
3.2 Nobel Prizes by Category

```
In [6]: # Display the number of prizes won by category.
        cat=nobel['category'].value_counts()
        display(cat)

        # Plot
        sns.barplot(x=cat.index,y=cat.values)
        plt.xticks(rotation=90)
        plt.title('Nobel Prizes by Category',fontsize=20)
```

```
Medicine      211
Physics       204
Chemistry     175
Peace         130
Literature    113
Economics     78
Name: category, dtype: int64
```

```
Out[6]: Text(0.5, 1.0, 'Nobel Prizes by Category')
```



The Nobel Prize is widely regarded as the most prestigious award available in the fields of literature, medicine, physics, chemistry, economics and activism for peace. Economics category there was only 78 laureates, because economics field was established since 1969. Medicine field got the highest number of laureates Medicine.

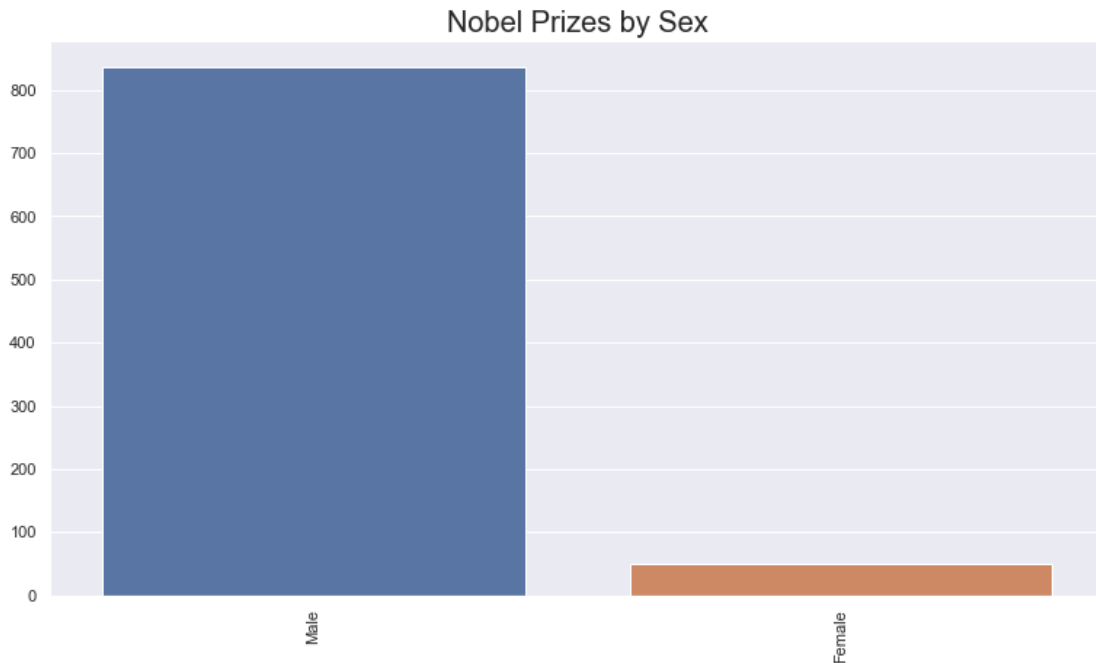
3.3 Nobel Prizes by Sex

```
In [7]: # Display the number of prizes won by sex.
sex=nobel['sex'].value_counts()
display(sex)

# Plot
sns.barplot(x=sex.index,y=sex.values)
plt.xticks(rotation=90)
plt.title('Nobel Prizes by Sex',fontsize=20)
```

```
Male      836
Female    49
Name: sex, dtype: int64
```

```
Out[7]: Text(0.5, 1.0, 'Nobel Prizes by Sex')
```



There is a huge gender gap between the male and female prize winners, more than 90% of prize winners are males.

3.4 Nobel Prizes by Birth Country

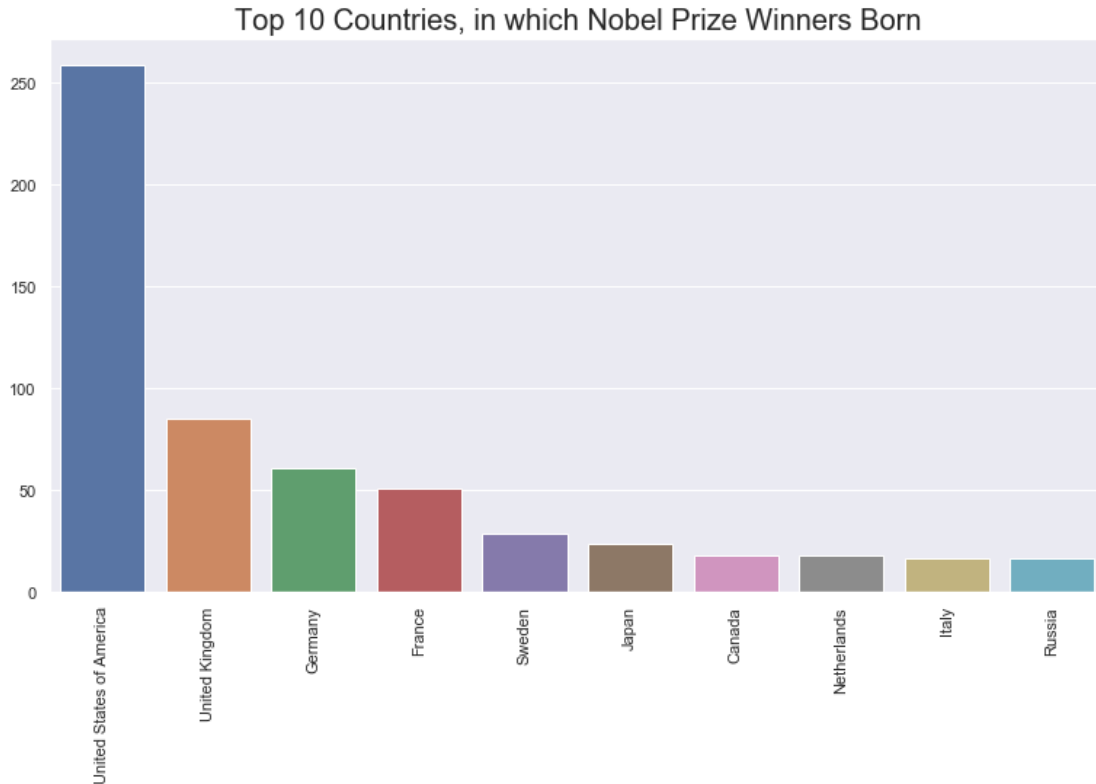
```
In [8]: # Display the number of prizes won by the top 10 Birth Country.
        ctry = nobel['birth_country'].value_counts().head(10)
        display(ctry)

        # Plot
        sns.barplot(x=ctry.index,y=ctry.values)
        plt.xticks(rotation=90)
        plt.title('Top 10 Countries, in which Nobel Prize Winners Born',fontsize=20)
```

United States of America	259
United Kingdom	85
Germany	61
France	51
Sweden	29
Japan	24
Canada	18
Netherlands	18
Italy	17
Russia	17

Name: birth_country, dtype: int64

Out[8]: Text(0.5, 1.0, 'Top 10 Countries, in which Nobel Prize Winners Born')



USA is the dominant country in receiving the prizes, Next comes United Kingdom, Germany, France, Sweden.

3.5 Nobel Prizes by Birth City

```
In [9]: # Display the number of prizes won by the top 10 Birth City.
city = nobel['birth_city'].value_counts().head(10)
display(city)
# Plot
sns.barplot(x=city.index,y=city.values)
plt.xticks(rotation=90)
plt.title('Top 10 Cities, in which Nobel Prize Winners Born',fontsize=20)
```

New York, NY	45
Paris	25
London	19
Vienna	14
Chicago, IL	12
Berlin	10
Budapest	8
Boston, MA	8

```

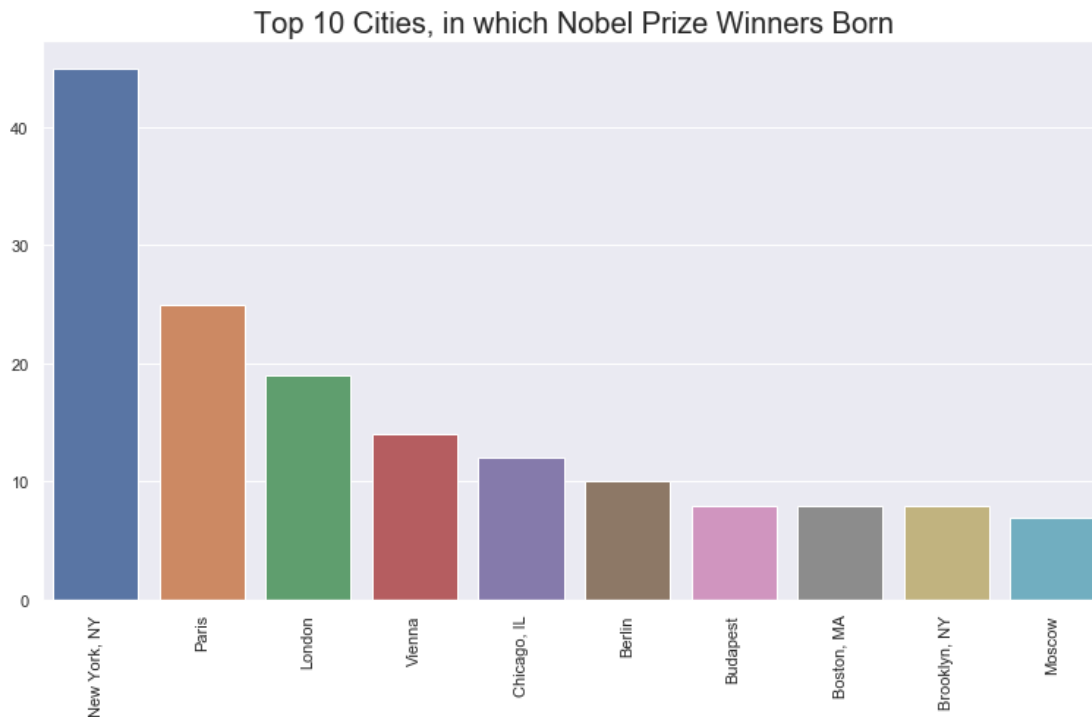
Brooklyn, NY      8
Moscow             7
Name: birth_city, dtype: int64

```

```

Out[9]: Text(0.5, 1.0, 'Top 10 Cities, in which Nobel Prize Winners Born')

```



45 nobel prize winners had born in New york city as on 2016. Next comes the cities Paris and London.

4 How It Changes as Time Goes? - Multivariate Plots

Let's look at several Bivariate Plots of this dataset.

To be clearer, I change the year to be decade.

4.1 USA Dominance - Prize Proportion per Decade

We already found that the most common Nobel laureate between 1901 and 2016 was born in the United States of America. But in 1901 all the winners were European. Then, when did the USA start to dominate the Nobel Prize charts?

(For country, we will use the birth_country of the winner, as the organization_country is NaN for all shared Nobel Prizes.)

```

In [10]: # Calculating the proportion of USA born winners per decade
nobel['usa_born_winner'] = nobel['birth_country'] == 'United States of America'

```



```
nobel['decade'] = (np.floor(nobel['year'] / 10) * 10).astype(int)
prop_usa_winners = nobel.groupby('decade', as_index=False)['usa_born_winner'].mean()

# Display the proportions of USA born winners per decade
prop_usa_winners
```

```
Out[10]:
```

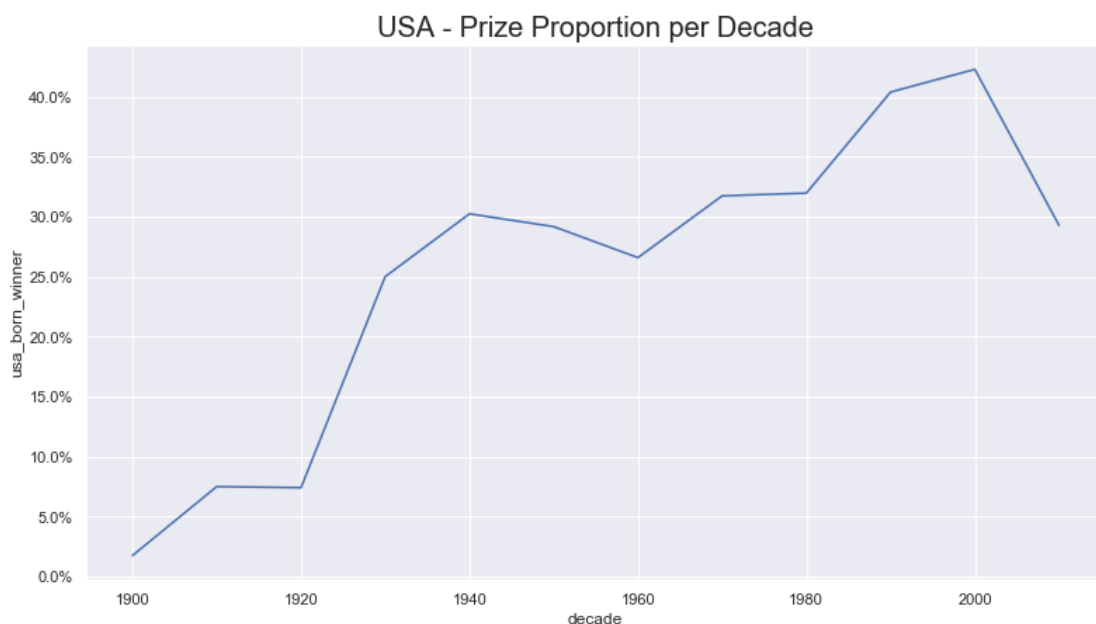
	decade	usa_born_winner
0	1900	0.017544
1	1910	0.075000
2	1920	0.074074
3	1930	0.250000
4	1940	0.302326
5	1950	0.291667
6	1960	0.265823
7	1970	0.317308
8	1980	0.319588
9	1990	0.403846
10	2000	0.422764
11	2010	0.292683

```
In [11]: # Plotting USA born winners
ax = sns.lineplot(x='decade', y='usa_born_winner', data=prop_usa_winners)

# Adding %-formatting to the y-axis
from matplotlib.ticker import PercentFormatter
ax.yaxis.set_major_formatter(PercentFormatter(1.0))

ax.set_title('USA - Prize Proportion per Decade', fontsize=20)
```

```
Out[11]: Text(0.5, 1.0, 'USA - Prize Proportion per Decade')
```



So the USA became the dominating winner of the Nobel Prize first in the 1930s and had kept the leading position ever since.

4.2 What is the gender of a typical Nobel Prize winner?

Well, when dominant birth country changed, while one group that was in the lead from the start, and never seems to let go, are *men*.

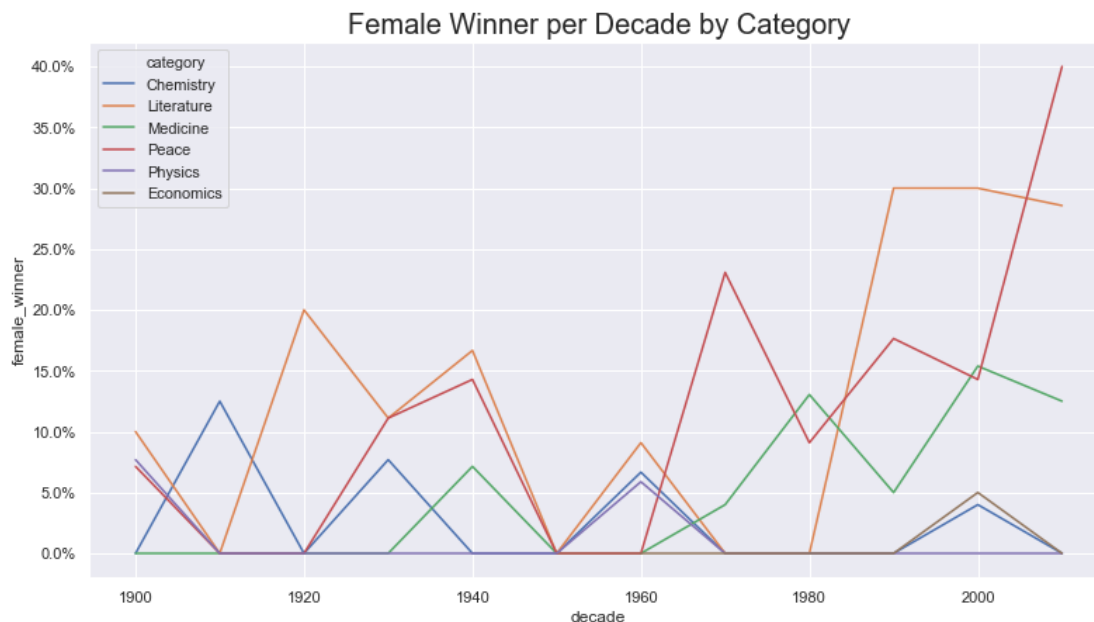
Maybe we weren't really shocked by this, but how significant is this imbalance? And is it better or worse within specific prize categories like physics, medicine, literature, etc.?

```
In [12]: # Calculating the proportion of female laureates per decade
nobel['female_winner'] = nobel['sex'] == 'Female'
prop_female_winners = nobel.groupby(['decade', 'category'],
                                   as_index=False)['female_winner'].mean()

# Plotting USA born winners with % winners on the y-axis
ax = sns.lineplot(x='decade', y='female_winner', hue='category',
                  data=prop_female_winners)

# Adding %-formatting to the y-axis
ax.yaxis.set_major_formatter(PercentFormatter(1.0))
ax.set_title('Female Winner per Decade by Category', fontsize=20)
```

```
Out[12]: Text(0.5, 1.0, 'Female Winner per Decade by Category')
```



The plot above is a bit messy as the lines are overplotting. But it does show some interesting trends and patterns. Overall the imbalance is pretty large with physics, economics, and chemistry having the largest imbalance. Medicine has a somewhat positive trend, and since the 1990s the literature prize is also now more balanced. The big outlier is the peace prize during the 2010s.

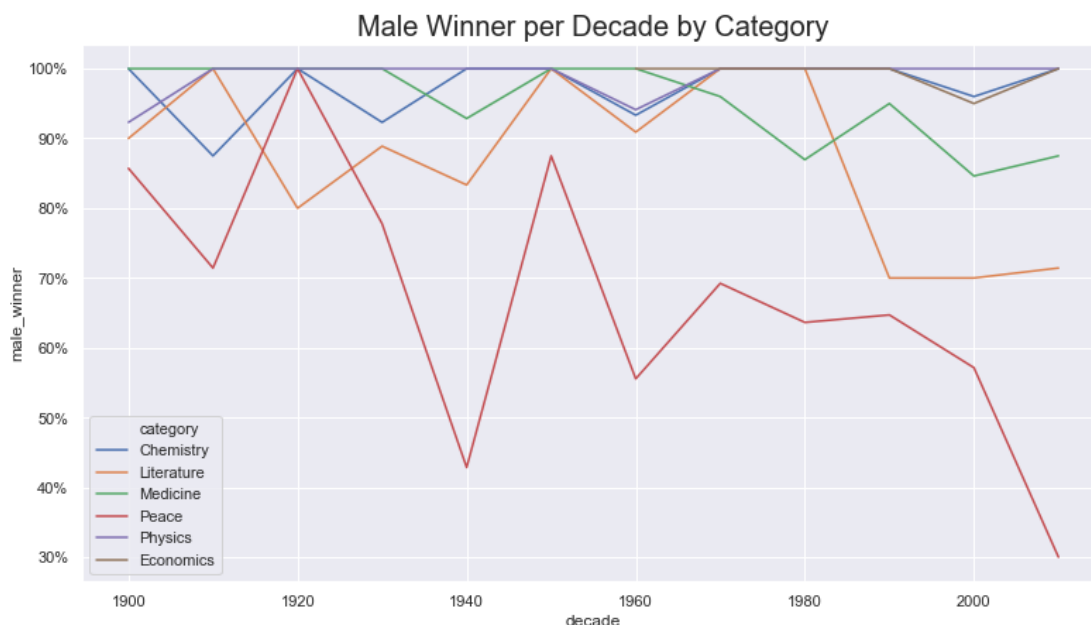
The plot below is Male Winner per Decade by Category, respectively.

```
In [13]: # Calculating the proportion of male laureates per decade
nobel['male_winner'] = nobel['sex'] == 'Male'
prop_male_winners = nobel.groupby(['decade', 'category'],
                                   as_index=False)['male_winner'].mean()

# Plotting USA born winners with % winners on the y-axis
ax = sns.lineplot(x='decade', y='male_winner', hue='category',
                  data=prop_male_winners)

# Adding %-formatting to the y-axis
ax.yaxis.set_major_formatter(PercentFormatter(1.0))
ax.set_title('Male Winner per Decade by Category', fontsize=20)
```

```
Out[13]: Text(0.5, 1.0, 'Male Winner per Decade by Category')
```



```
In [14]: ax = sns.lineplot(x='decade', y='male_winner', label="Male",
                             data=prop_male_winners)
ax = sns.lineplot(x='decade', y='female_winner', label="Female",
                  data=prop_female_winners)

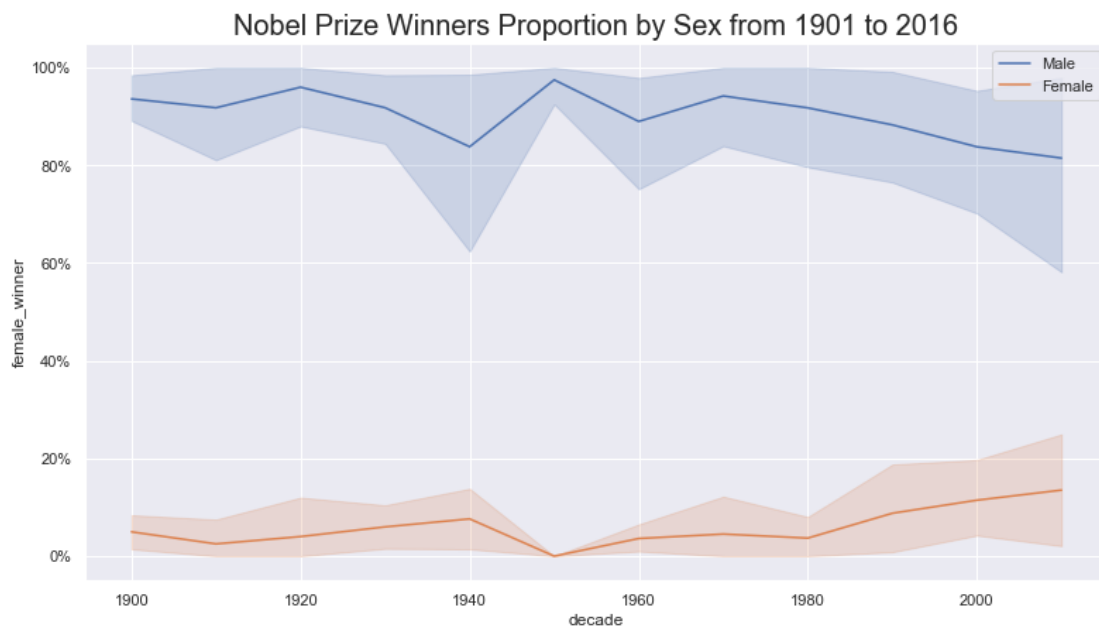
# Adding %-formatting to the y-axis
```

```
ax.yaxis.set_major_formatter(PercentFormatter(1.0))
ax.set_title('Nobel Prize Winners Proportion by Sex from 1901 to 2016', fontsize=20)
```

C:\Users\Zhoulz\Anaconda3\lib\site-packages\scipy\stats\stats.py:1713: FutureWarning: Using a non-integer for 'axis' is deprecated. Please use 'axis=axis' instead.

```
return np.add.reduce(sorted[indexer] * weights, axis=axis) / sumval
```

Out[14]: Text(0.5, 1.0, 'Nobel Prize Winners Proportion by Sex from 1901 to 2016')



4.3 How old are you when you get the prize?

Now, let's explore that how old when they get the prize.

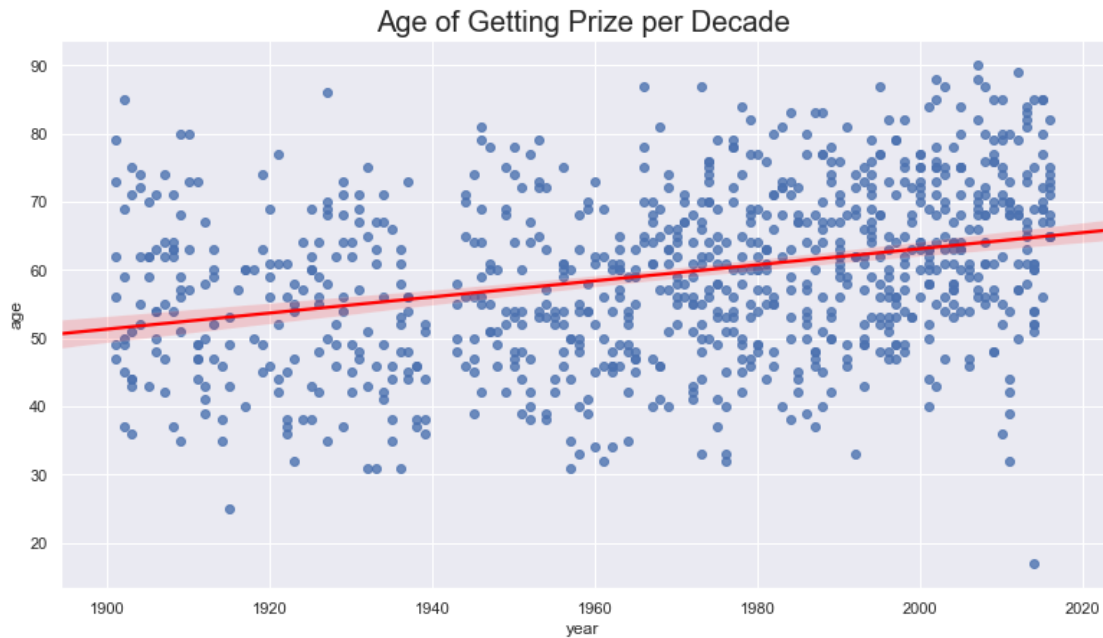
```
In [15]: # Converting birth_date from String to datetime
nobel['birth_date'] = pd.to_datetime(nobel['birth_date'])

# Calculating the age of Nobel Prize winners
nobel['age'] = nobel['year'] - nobel['birth_date'].dt.year
```

4.3.1 Plotting the age of Nobel Prize winners

```
In [16]: # Plotting the age of Nobel Prize winners
sns.regplot('year', 'age', nobel, line_kws={'color':'red'})
plt.title('Age of Getting Prize per Decade', fontsize=20)
```

Out[16]: Text(0.5, 1.0, 'Age of Getting Prize per Decade')



We see that people used to be around 55 when they received the prize, but nowadays the average is closer to 65. But there is a large spread in the laureates' ages, and while most are 50+, some are very young.

We also see that the density of points is much higher nowadays than in the early 1900s -- nowadays many more of the prizes are shared, and so there are many more winners. We also see that there was a disruption in awarded prizes around the Second World War (1939 - 1945).

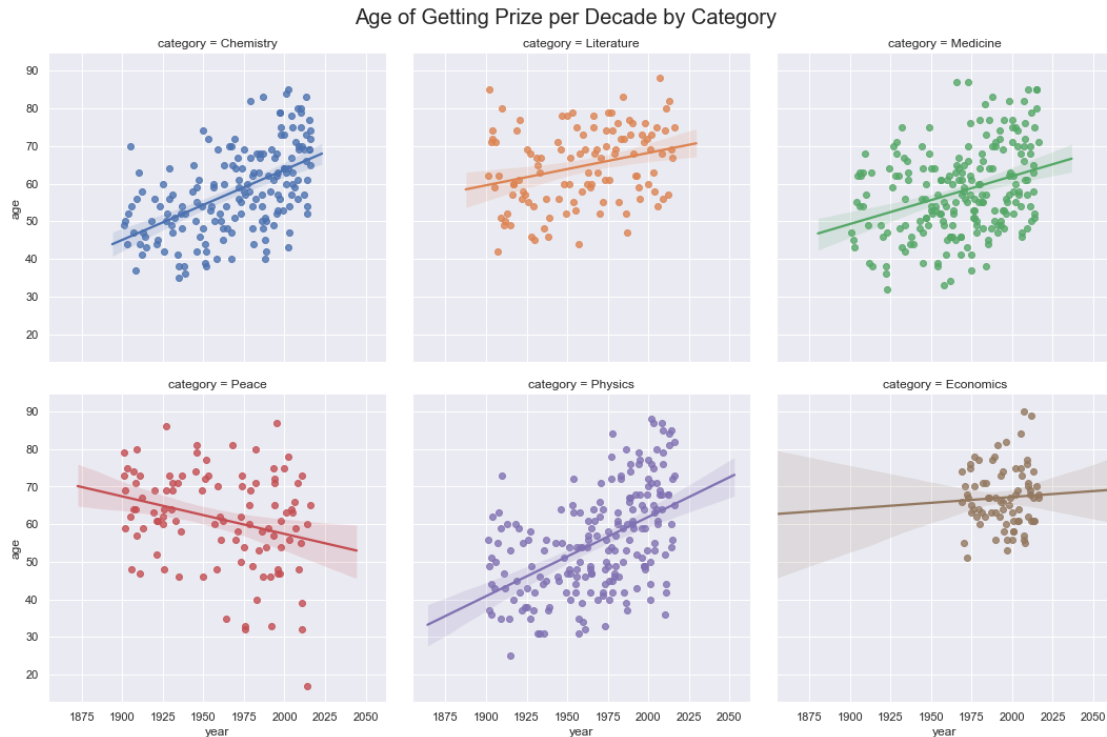
4.3.2 Age differences between prize categories

Let's look at age trends within different prize categories.

In [17]: *# Same plot as above, but separate plots for each type of Nobel Prize*

```
sns.lmplot('year', 'age', nobel, col='category', col_wrap=3, hue="category")
plt.subplots_adjust(top=0.92)
plt.suptitle('Age of Getting Prize per Decade by Category', fontsize=20)
```

Out[17]: Text(0.5, 0.98, 'Age of Getting Prize per Decade by Category')



We see that both winners of the chemistry, medicine, and physics prize have gotten older over time. The trend is strongest for physics: the average age used to be below 50, and now it's almost 70. Literature and economics are more stable. We also see that economics is a newer category. But peace shows an opposite trend where winners are getting younger!

In the peace category we also a winner around 2010 that seems exceptionally young.

5 There is a Nobel Prize for Everyone - Thank You!

Picture Credits: https://s3.amazonaws.com/assets.datacamp.com/production/project_441/img/paint_nobel_p