M. Shakaib Saleem

Case Study 1. The most Nobel of Prizes

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 nowadays it's not biased in any way whatsoever. Surely. Right?

Well, we're going to find out! The Nobel Foundation has made a dataset available of all prize winners from the start of the prize, in 1901, to 2016. Let's load it in and take a look.

In [2]:

```
# Loading in required libraries
# ... YOUR CODE FOR TASK 1 ...
import pandas as pd

# Reading in the Nobel Prize data
nobel = pd.read_csv('datasets/nobel.csv')

# Taking a look at the first several winners
# ... YOUR CODE FOR TASK 1 ...
display(nobel.head())
```

	year	category	prize	motivation	prize_share	laureate_id	laureate_type	full_nar
0	1901	Chemistry	The Nobel Prize in Chemistry 1901	"in recognition of the extraordinary services ...	1/1	160	Individual	Jacob Henric van 't H
1	1901	Literature	The Nobel Prize in Literature 1901	"in special recognition of his poetic composit...	1/1	569	Individual	Su Prudhom
2	1901	Medicine	The Nobel Prize in Physiology or Medicine 1901	"for his work on serum therapy, especially its...	1/1	293	Individual	Emil Ad von Behri
3	1901	Peace	The Nobel Peace Prize 1901	NaN	1/2	462	Individual	Jean Hei Duna
4	1901	Peace	The Nobel Peace Prize 1901	NaN	1/2	463	Individual	Frédé Pas

These are the first few records from our data set of nobel prize winners

2. So, who gets the Nobel Prize?

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 guys 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?

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

In [3]:

```python
# Display the number of (possibly shared) Nobel Prizes handed
# out between 1901 and 2016
# ... YOUR CODE FOR TASK 2 ...
print(len(nobel))

# Display the number of prizes won by male and female recipients.
# ... YOUR CODE FOR TASK 2 ...
display(nobel['sex'].value_counts())

# Display the number of prizes won by the top 10 nationalities.
# ... YOUR CODE FOR TASK 2 ...
display(nobel['birth_country'].value_counts().head(10))
```

```
911

Male      836
Female     49
Name: sex, dtype: int64

United States of America    259
United Kingdom               85
Germany                      61
France                       51
Sweden                       29
Japan                        24
Netherlands                  18
Canada                       18
Russia                       17
Italy                        17
Name: birth_country, dtype: int64
```

This tells us that there are 911 records in our database, which means the data consists of details on 911 prizes. Of these, 836 are won by Male whereas 49 are won by Female candidates. These are the only distributions for the 'Sex' category and other non-binary genders are not considered. Lastly, the descending list of countries tells us that "USA" has won by far the most number of Nobel prizes, followed by the UK, then Germany, France and Sweden.

3. USA dominance

Not so surprising perhaps: the most common Nobel laureate between 1901 and 2016 was a man born in the United States of America. But in 1901 all the winners were European. When did the USA start to dominate the Nobel Prize charts?

In [4]:

```
# Calculating the proportion of USA born winners per decade
nobel['usa_born_winner'] = nobel["birth_country"]=="United States of America"
nobel['decade'] = ((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
# ... YOUR CODE FOR TASK 3 ...
display(prop_usa_winners)
```

	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

This shows us the proportion of winners from USA for each decade. We can see that in the early 1900s, the ratio was less that 10% but since the 1930s, the proportion has been much higher, around 25-30%, and even 40% recently.

4. USA dominance, visualized

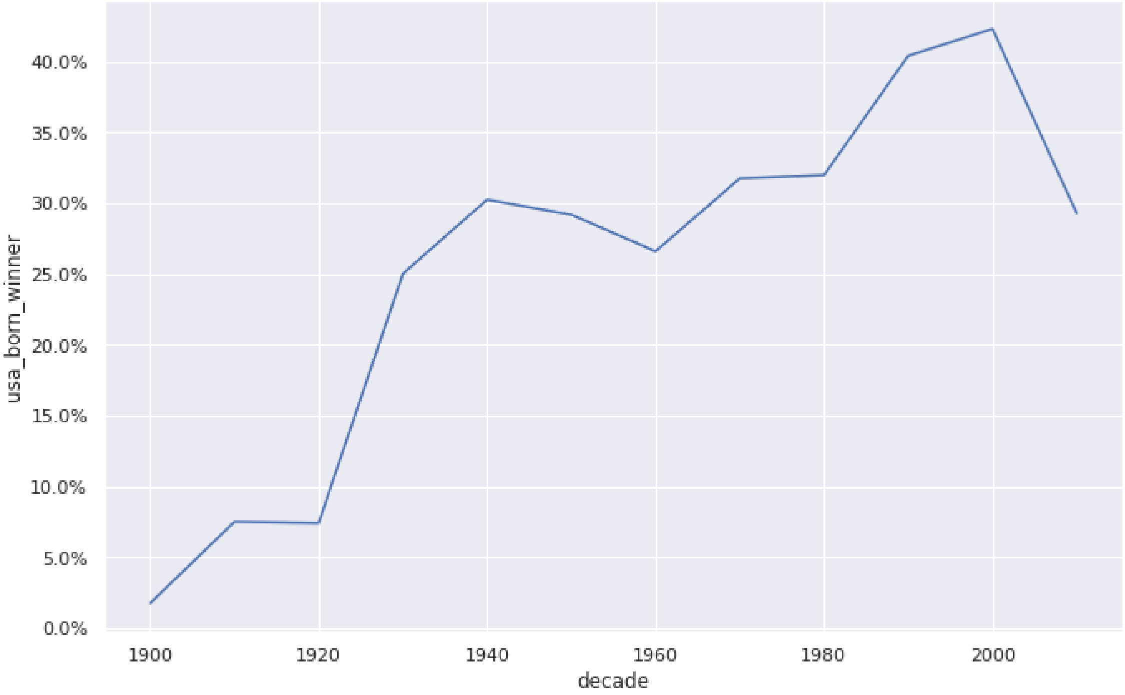
A table is OK, but to *see* when the USA started to dominate the Nobel charts we need a plot!

In [5]:

```
# Setting the plotting theme
import seaborn as sns
sns.set()
# and setting the size of all plots.
import matplotlib.pyplot as plt
plt.rcParams['figure.figsize'] = [11, 7]

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

# Adding %-formatting to the y-axis
from matplotlib.ticker import PercentFormatter
# ... YOUR CODE FOR TASK 4 ...
ax.yaxis.set_major_formatter(PercentFormatter(1))
```



We can again see the dominance of the USA, although this seems to have recently peaked and may now be on a decline.

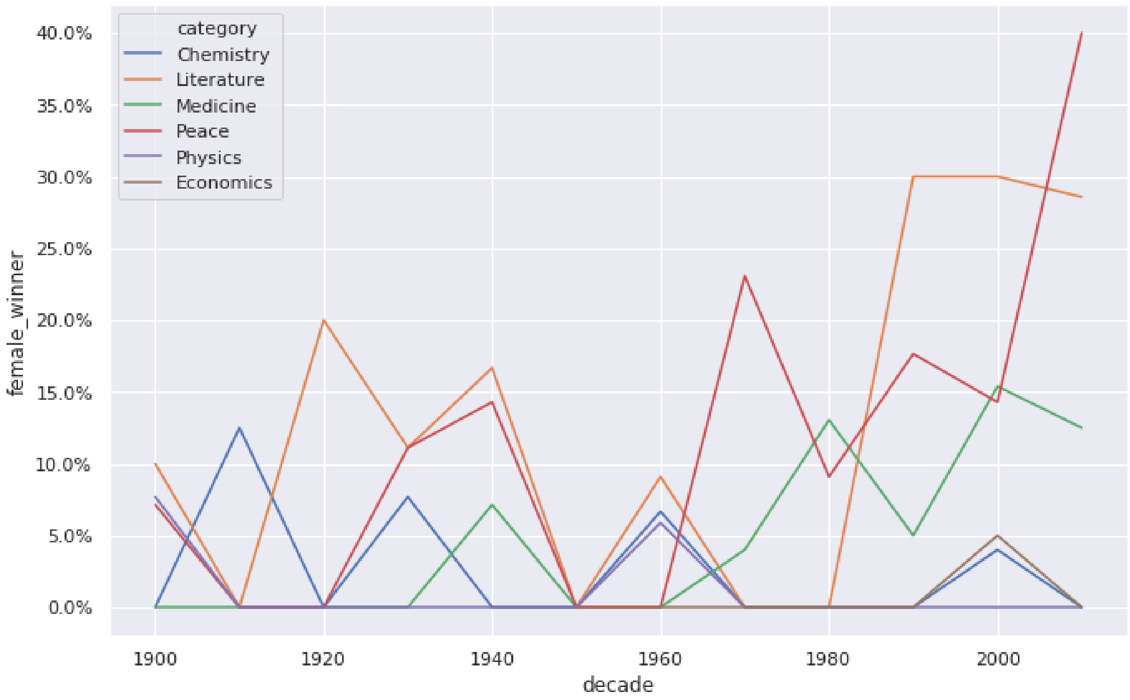
5. What is the gender of a typical Nobel Prize winner?

So the USA became the dominating winner of the Nobel Prize first in the 1930s and had kept the leading position ever since. But one group that was in the lead from the start, and never seems to let go, are *men*. Maybe it shouldn't come as a shock that there is some imbalance between how many male and female prize winners there are, but how significant is this imbalance? And is it better or worse within specific prize categories like physics, medicine, literature, etc.?

In [6]:

```
# 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
# ... YOUR CODE FOR TASK 5 ...
ax = sns.lineplot(x=prop_female_winners['decade'],y=prop_female_winners['female_winner'], hue=prop_female_winners['category'])
ax.yaxis.set_major_formatter(PercentFormatter(1))
```



This clearly shows the imbalance between the two sexes in winners. In all categories except Literature and Peace, men have clear dominance.

6. The first woman to win the Nobel Prize

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, but keep in mind that this just covers the years 2010 to 2016.

Given this imbalance, who was the first woman to receive a Nobel Prize? And in what category?

In [7]:

```
# Picking out the first woman to win a Nobel Prize
# ... YOUR CODE FOR TASK 5 ...
display(nobel.loc[nobel["sex"]=="Female"].nsmallest(1, 'year'))
```

	year	category	prize	motivation	prize_share	laureate_id	laureate_type	full_name
19	1903	Physics	The Nobel Prize in Physics 1903	"in recognition of the extraordinary services ...	1/4	6	Individual	Marie Curie, née Sklodowska

1 rows × 21 columns

This shows us that the first female winner was Marie Curie in 1903 who won the Physics Nobel Prize for her work on Radiology.

7. Repeat laureates

For most scientists/writers/activists a Nobel Prize would be the crowning achievement of a long career. But for some people, one is just not enough, and few have gotten it more than once. Who are these lucky few? (Having won no Nobel Prize myself, I'll assume it's just about luck.)

In [8]:

```
# Selecting the laureates that have received 2 or more prizes.
# ... YOUR CODE FOR TASK 5 ...
display(nobel.groupby('full_name').filter(lambda group: len(group) >= 2))
```

	year	category	prize	motivation	prize_share	laureate_id	laureate_type	
19	1903	Physics	The Nobel Prize in Physics 1903	"in recognition of the extraordinary services ...	1/4	6	Individual	
62	1911	Chemistry	The Nobel Prize in Chemistry 1911	"in recognition of her services to the advance...	1/1	6	Individual	
89	1917	Peace	The Nobel Peace Prize 1917	NaN	1/1	482	Organization	ir I
215	1944	Peace	The Nobel Peace Prize 1944	NaN	1/1	482	Organization	ir I
278	1954	Chemistry	The Nobel Prize in Chemistry 1954	"for his research into the nature of the chemi...	1/1	217	Individual	
283	1954	Peace	The Nobel Peace Prize 1954	NaN	1/1	515	Organization	U C
298	1956	Physics	The Nobel Prize in Physics 1956	"for their researches on semiconductors and th...	1/3	66	Individual	
306	1958	Chemistry	The Nobel Prize in Chemistry 1958	"for his work on the structure of proteins, es...	1/1	222	Individual	
340	1962	Peace	The Nobel Peace Prize 1962	NaN	1/1	217	Individual	
348	1963	Peace	The Nobel Peace Prize 1963	NaN	1/2	482	Organization	ir I
424	1972	Physics	The Nobel Prize in Physics 1972	"for their jointly developed theory of superco...	1/3	66	Individual	
505	1980	Chemistry	The Nobel Prize in Chemistry 1980	"for their contributions concerning the determ...	1/4	222	Individual	

	year	category	prize	motivation	prize_share	laureate_id	laureate_type
523	1981	Peace	The Nobel Peace Prize 1981	NaN	1/1	515	Organization

13 rows × 21 columns

This shows us that 13 records correspond to those who won multiple nobel prizes. This does not mean that there were 13 such laureates, but 13 prizes were given to such laurettes (which are 6 when we count unique ids from herein).

8. How old are you when you get the prize?

The list of repeat winners contains some illustrious names! We again meet Marie Curie, who got the prize in physics for discovering radiation and in chemistry for isolating radium and polonium. John Bardeen got it twice in physics for transistors and superconductivity, Frederick Sanger got it twice in chemistry, and Linus Carl Pauling got it first in chemistry and later in peace for his work in promoting nuclear disarmament. We also learn that organizations also get the prize as both the Red Cross and the UNHCR have gotten it twice.

But how old are you generally when you get the prize?

In [9]:

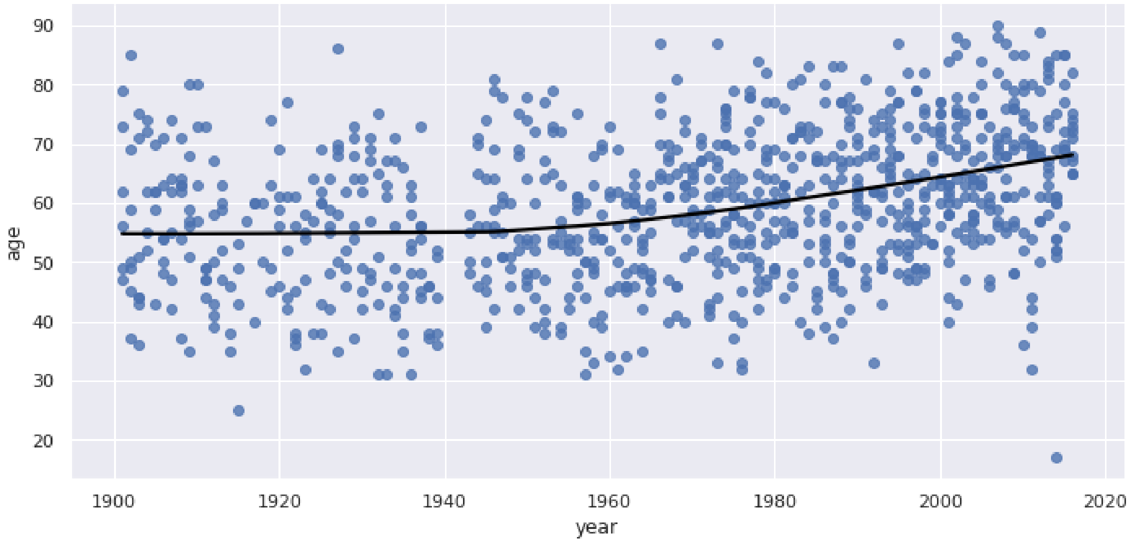
```python
# Converting birth_date from String to datetime
nobel['birth_date'] = pd.to_datetime(nobel['birth_date'])

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

# Plotting the age of Nobel Prize winners
# ... YOUR CODE FOR TASK 8 ...
sns.lmplot(data=nobel,x='year',y='age',lowess=True, aspect=2,line_kws={'color' : 'black'})
```

Out[9]:

```
<seaborn.axisgrid.FacetGrid at 0x7fe4385687b8>
```



We see the distribution of the ages of the winners in the form of the blue dots, which are getting denser with time, which means that now more and more number of people are winning nobel prizes. Also, the black line draws out the average age of the winner which is also on the rise as nowadays, mostly older people are winning nobel prizes as compared to before.

9. Age differences between prize categories

The plot above shows us a lot! We see that people use to be around 55 when they received the price, 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 high 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).

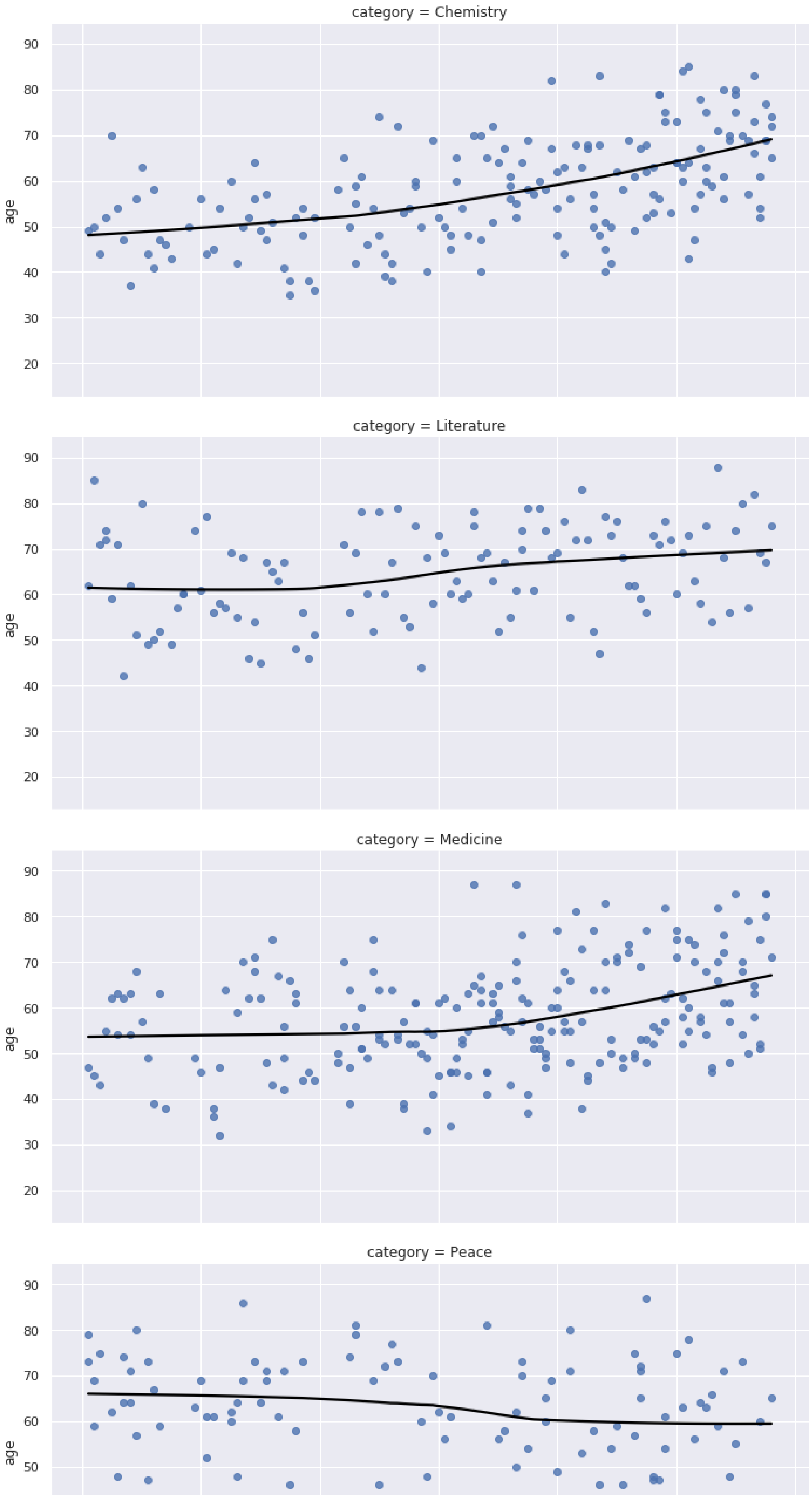
Let's look at age trends within different prize categories.

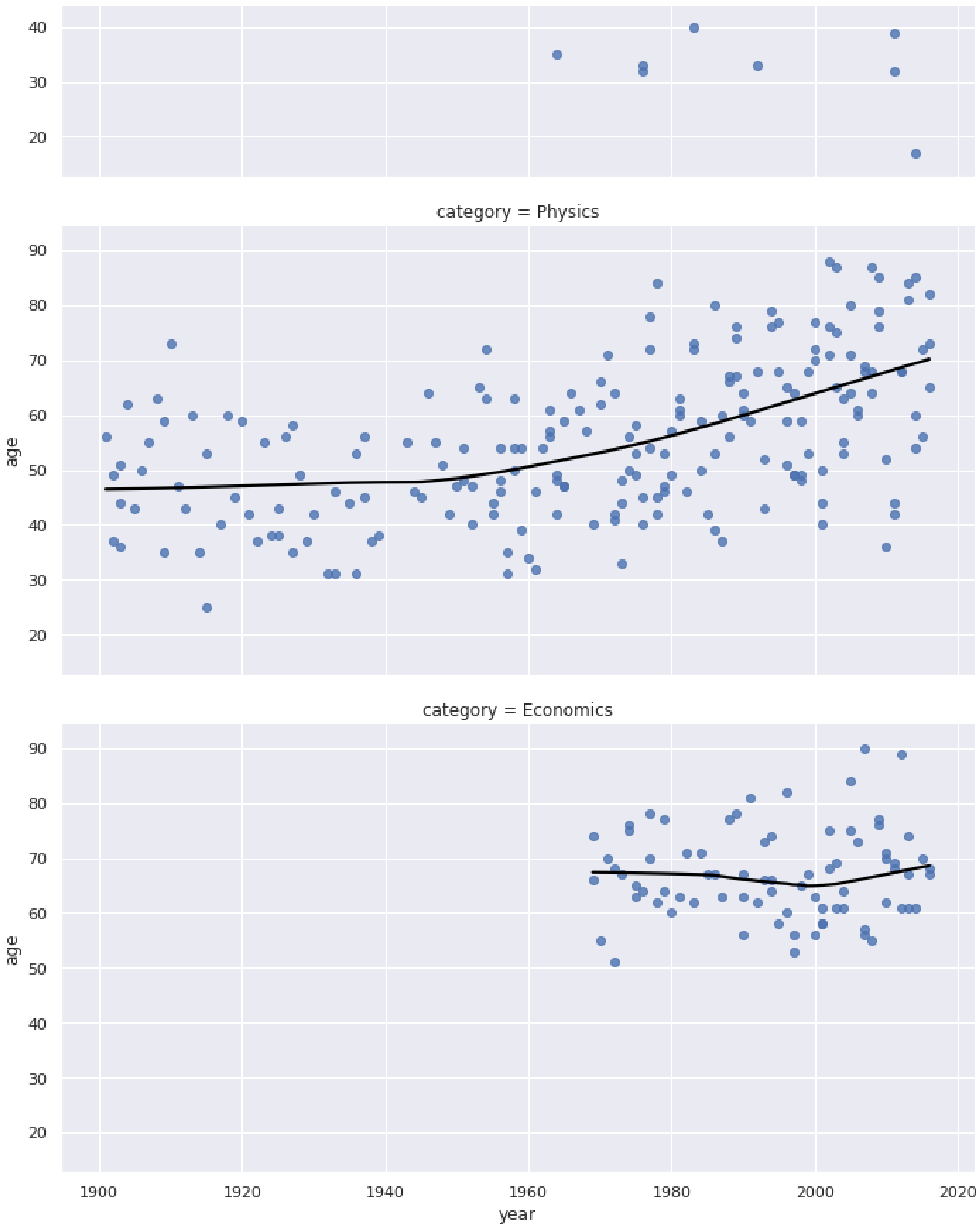
In [10]:

```
# Same plot as above, but separate plots for each type of Nobel Prize
# ... YOUR CODE FOR TASK 9 ...
sns.lmplot(data=nobel,x='year',y='age',lowess=True, aspect=2,line_kws={'color' : 'black'},row='category')
```

```
Out[10]:

<seaborn.axisgrid.FacetGrid at 0x7fe438500f98>
```


category = Chemistry
category = Literature
category = Medicine
category = Peace
age
90
80
70
60
50
40
30
20



We see that the average age for most categories is usually rising but for peace category, the average age is decreasing as the years pass by. One significant contributor to this our fellow Pakistani Malala Yusufzai who won the Nobel Peace prize in 2014.

10. Oldest and youngest winners

More plots with lots of exciting stuff going on! 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. This begs the questions, who are the oldest and youngest people ever to have won a Nobel Prize?

In [11]:

```
# The oldest winner of a Nobel Prize as of 2016
# ... YOUR CODE FOR TASK 10 ...
display(nobel.nlargest(1, 'age'))

# The youngest winner of a Nobel Prize as of 2016
# ... YOUR CODE FOR TASK 10 ..
display(nobel.nsmallest(1, 'age'))
```

	year	category	prize	motivation	prize_share	laureate_id	laureate_type	full_nam
793	2007	Economics	The Sveriges Riksbank Prize in Economic Scienc...	"for having laid the foundations of mechanism ...	1/3	820	Individual	Leon Hurwic

1 rows × 22 columns

	year	category	prize	motivation	prize_share	laureate_id	laureate_type	full_name	I
885	2014	Peace	The Nobel Peace Prize 2014	"for their struggle against the suppression of...	1/2	914	Individual	Malala Yousafzai	

1 rows × 22 columns

Here we can see that Malala was in fact the youngest laurette ever. On the other hand, the oldest winner was Leonid Hurwicz who won the Economics prize in 2007.

11. The category in which the prize is shared most often

In [12]:

```
nobel['is_shared'] = nobel['prize_share']!='1/1'
prop_shared_category = nobel.groupby('category', as_index=False)['is_shared'].mean()
#display(nobel['is_shared'].head())
display(prop_shared_category.sort_values('is_shared',ascending=False).head(1))
```

	category	is_shared
3	Medicine	0.815166

This result tells us that the prizes won in the category of Medicine are most often shared by more than one laurette

12. Countries with highest proportion of women laurettes

In [13]:

```
nobel['prop_women'] = nobel["sex"]=="Female"
prop_fem_country = nobel.groupby('birth_country', as_index=False)['prop_women'].mean()
display(prop_fem_country.sort_values('prop_women',ascending=False).head(25))
```

	birth_country	prop_women
120	Yemen	1.000000
111	Ukraine	1.000000
75	Ottoman Empire (Republic of Macedonia)	1.000000
63	Liberia	1.000000
77	Pakistan	1.000000
61	Kenya	1.000000
78	Persia (Iran)	1.000000
56	Iran	1.000000
24	Burma (Myanmar)	1.000000
12	Austrian Empire (Czech Republic)	1.000000
49	Guatemala	0.500000
26	Chile	0.500000
73	Northern Ireland	0.400000
95	Russian Empire (Poland)	0.400000
88	Romania	0.250000
6	Austria-Hungary (Czech Republic)	0.250000
20	British Mandate of Palestine (Israel)	0.200000
80	Poland	0.200000
36	Egypt	0.166667
45	Germany (Poland)	0.125000
58	Italy	0.117647
101	South Africa	0.111111
1	Australia	0.100000
33	Denmark	0.090909
27	China	0.090909

Here we can see the top countries in terms of proportion of women laurettes, however the highest score are perhaps misleading because they are caused by the low number winners from that country.

You get a prize!

Hey! You get a prize for making it to the very end of this notebook! It might not be a Nobel Prize, but I made it myself in paint so it should count for something. But don't despair, Leonid Hurwicz was 90 years old when he got his prize, so it might not be too late for you. Who knows.