**A Choropleth Map and the Countries with the Most Prizes**

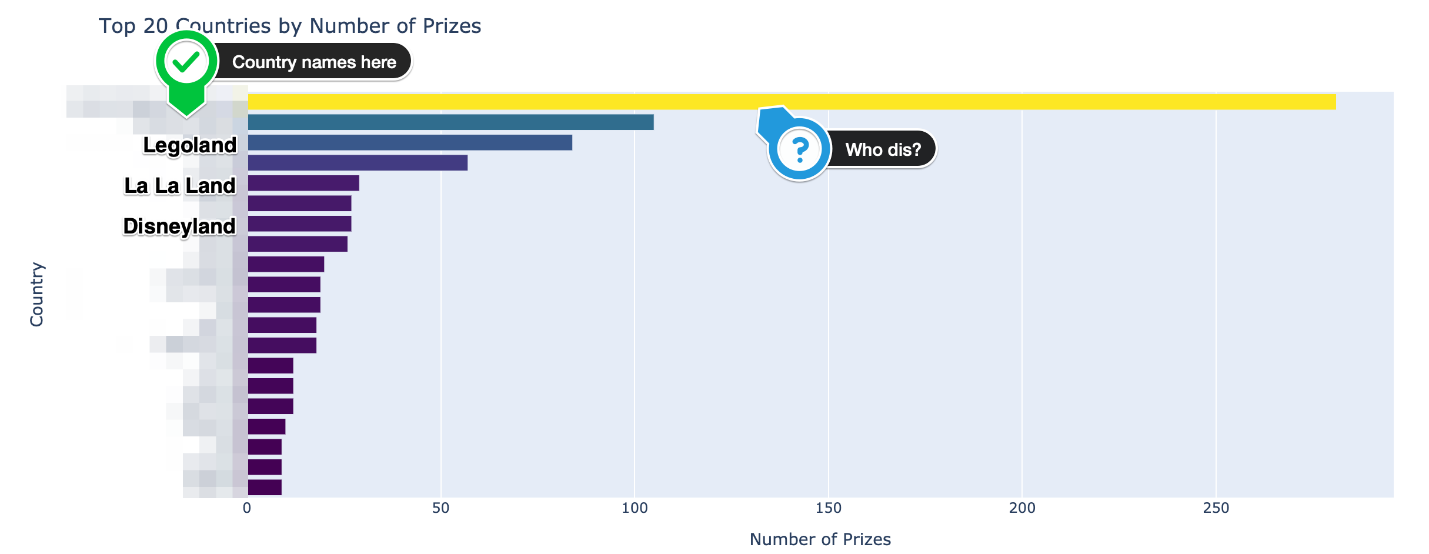
For this next bit, we're going to compare which countries actually get the most prizes. And we're also going to look at in which categories those prizes are awarded. This has me feeling a little like I'm at the Olympics 😊.

Challenge 1: Top 20 Country Ranking

* Create a Pandas DataFrame called top20\_countries that has the two columns. The prize column should contain the total number of prizes won.



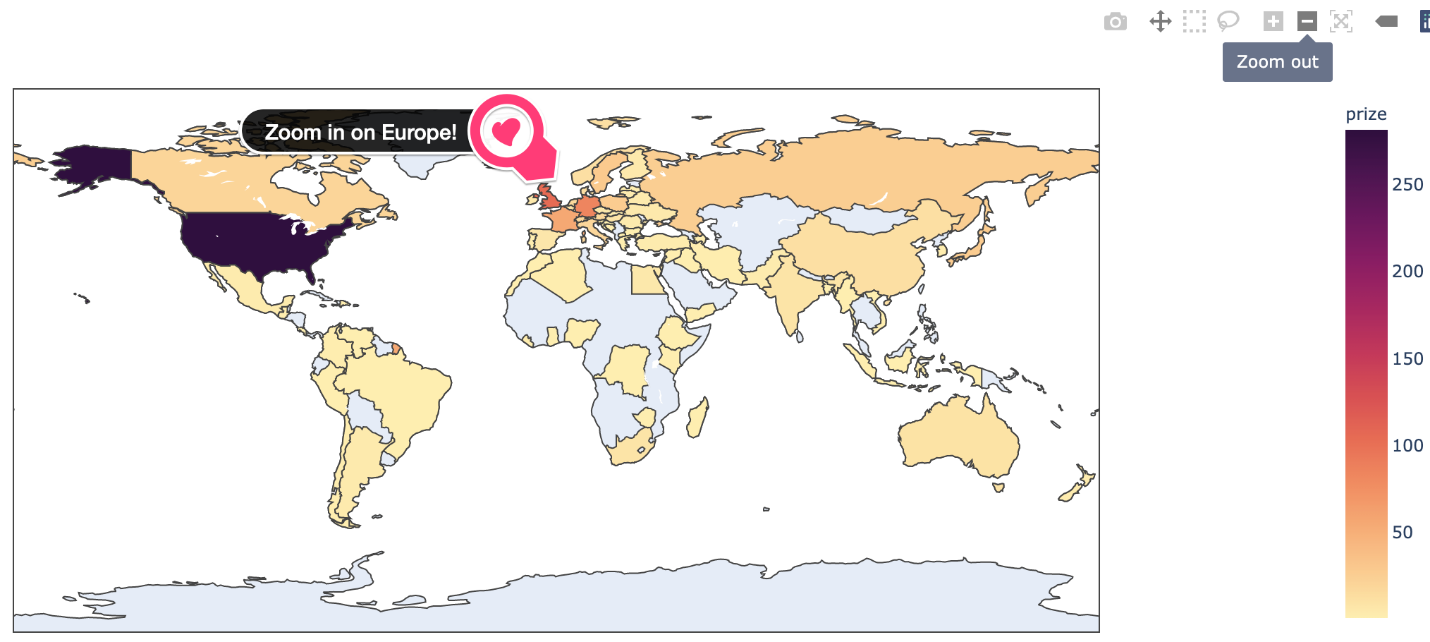
* Is it best to use birth\_country, birth\_country\_current or organization\_country?
* What are some potential problems when using birth\_country or any of the others? Which column is the least problematic?
* Then use plotly to create a horizontal bar chart showing the number of prizes won by each country. Here's what you're after:



* What is the ranking for the top 20 countries in terms of the number of prizes?

Challenge 2: Choropleth Map

* Create this choropleth map using [the plotly documentation](https://plotly.com/python/choropleth-maps/):

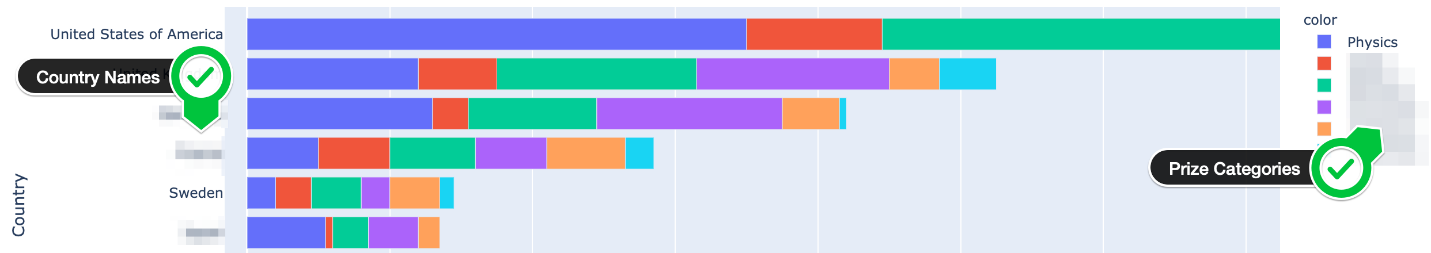


* Experiment with [plotly's available colours](https://plotly.com/python/builtin-colorscales/). I quite like the sequential colour matter on this map.

Hint: You'll need to use a 3 letter country code for each country.

Challenge 3: Country Bar Chart with Prize Category

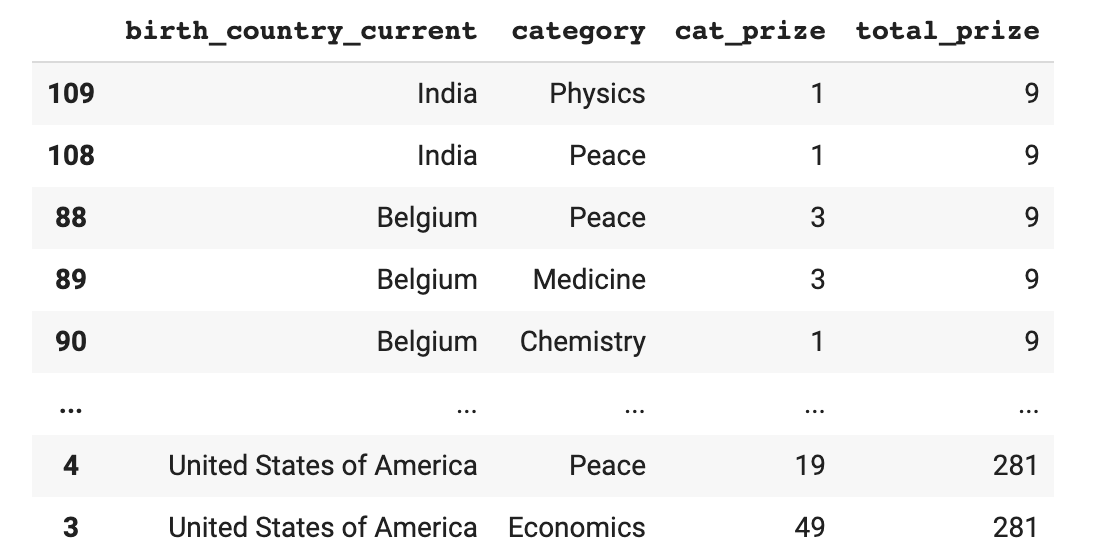
See if you can divide up the plotly bar chart you created above to show the which categories made up the total number of prizes. Here's what you're aiming for:



* In which category are Germany and Japan the weakest compared to the United States?
* In which category does Germany have more prizes than the UK?
* In which categories does France have more prizes than Germany?
* Which category makes up most of Australia's Nobel prizes?
* Which category makes up half of the prizes in the Netherlands?
* Does the United States have more prizes in Economics than all of France? What about in Physics or Medicine?

The hard part is preparing the data for this chart!

*Hint*: Take a two-step approach. The first step is grouping the data by country and category. Then you can create a DataFrame that looks something like this:



Challenge 4: Prizes by Country over Time

Every country's fortunes wax and wane over time. Investigate how the total number of prizes awarded changed over the years.

* When did the United States eclipse every other country in terms of the number of prizes won?
* Which country or countries were leading previously?
* Calculate the cumulative number of prizes won by each country in every year. Again, use the birth\_country\_current of the winner to calculate this.
* Create a [plotly line chart](https://plotly.com/python/line-charts/) where each country is a coloured line.

.

.

..

...

..

.

.

**Solution 1: Prize ranking by Country**

Looking at our DataFrame there are actually 3 different columns to choose from for creating this ranking: birth\_country, birth\_country\_current or organization\_country. However, they each have certain problems and limitations.

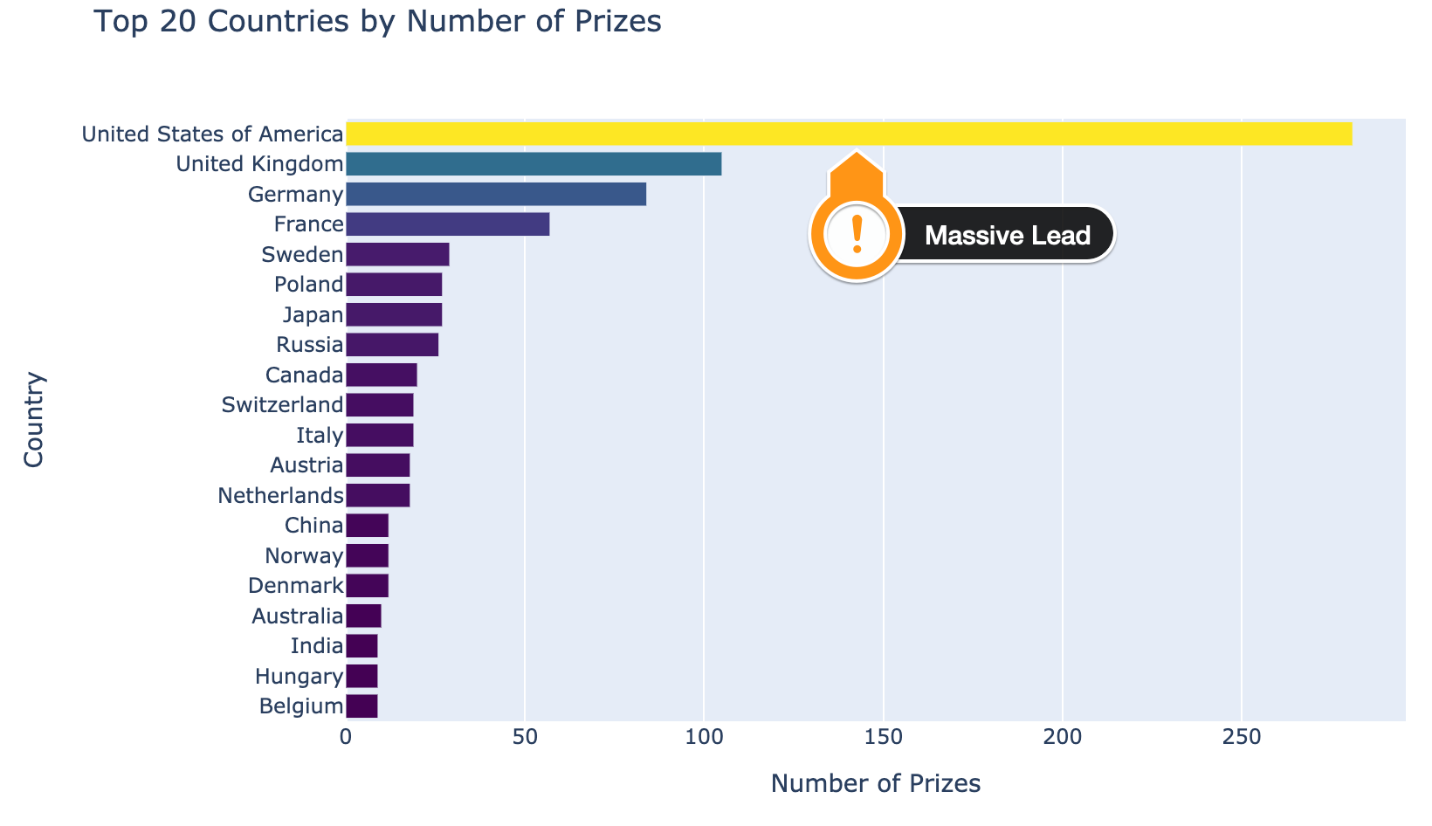
If you look at the entries in the birth country, you'll see that some countries no longer exist! These include the Soviet Union or Czechoslovakia for example. Hence, using birth\_country\_current is better, since it has the country name which controls the city where the laureate was born. Now, notice that this does not determine the laureates' nationality since some globetrotting folks gave birth to their future Nobel laureate children while abroad. Also, people's nationalities can change as they emigrate and acquire different citizenship or get married and change citizenship. What this boils down to is that we will have to be clear about the assumptions that we will make in the upcoming analysis.

We can create the list of the top 20 countries like this:

1. top\_countries = df\_data.groupby(['birth\_country\_current'],
2. as\_index=False).agg({'prize': pd.Series.count})
4. top\_countries.sort\_values(by='prize', inplace=True)
5. top20\_countries = top\_countries[-20:]

Note that the ranking here determines how our bar chart will be displayed.

1. h\_bar = px.bar(x=top20\_countries.prize,
2. y=top20\_countries.birth\_country\_current,
3. orientation='h',
4. color=top20\_countries.prize,
5. color\_continuous\_scale='Viridis',
6. title='Top 20 Countries by Number of Prizes')
8. h\_bar.update\_layout(xaxis\_title='Number of Prizes',
9. yaxis\_title='Country',
10. coloraxis\_showscale=False)
11. h\_bar.show()



The United States has a massive number of prizes by this measure. The UK and Germany are in second and third place respectively.

**Solution 2: Displaying the Data on a Map**

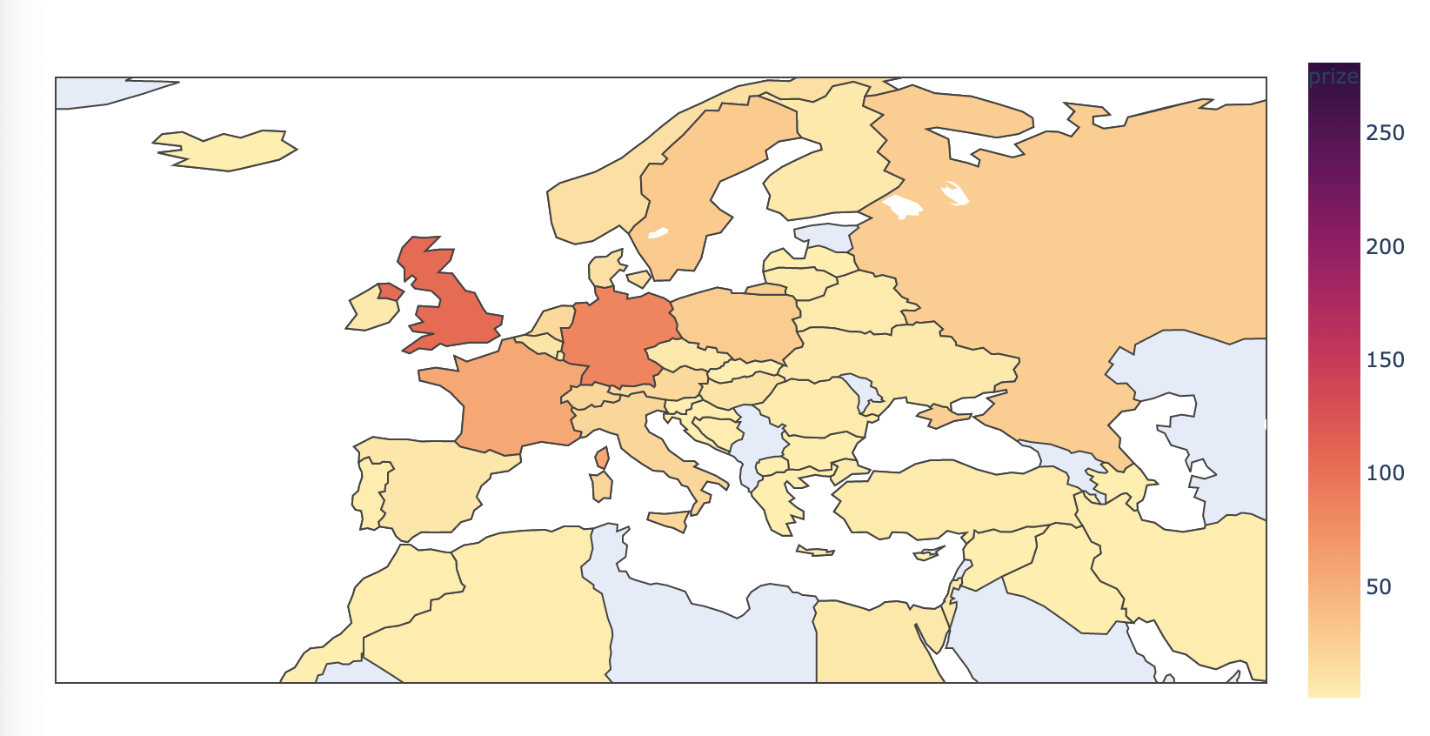
To show the above ranking on a colour coded map, we need to make use of the ISO codes.

1. df\_countries = df\_data.groupby(['birth\_country\_current', 'ISO'],
2. as\_index=False).agg({'prize': pd.Series.count})
3. df\_countries.sort\_values('prize', ascending=False)

This means we can use the ISO country codes for the locations parameter on the choropleth.

1. world\_map = px.choropleth(df\_countries,
2. locations='ISO',
3. color='prize',
4. hover\_name='birth\_country\_current',
5. color\_continuous\_scale=px.colors.sequential.matter)
7. world\_map.update\_layout(coloraxis\_showscale=True,)
9. world\_map.show()

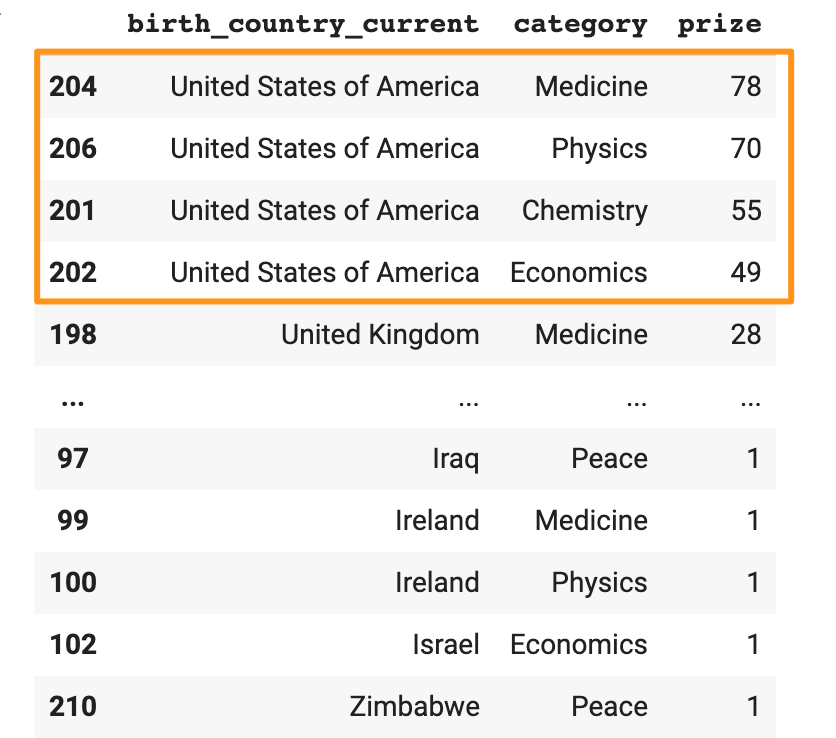
I love it how plotly allows you to zoom in and pan on the map it generates.



**Solution 3: The category breakdown by country**

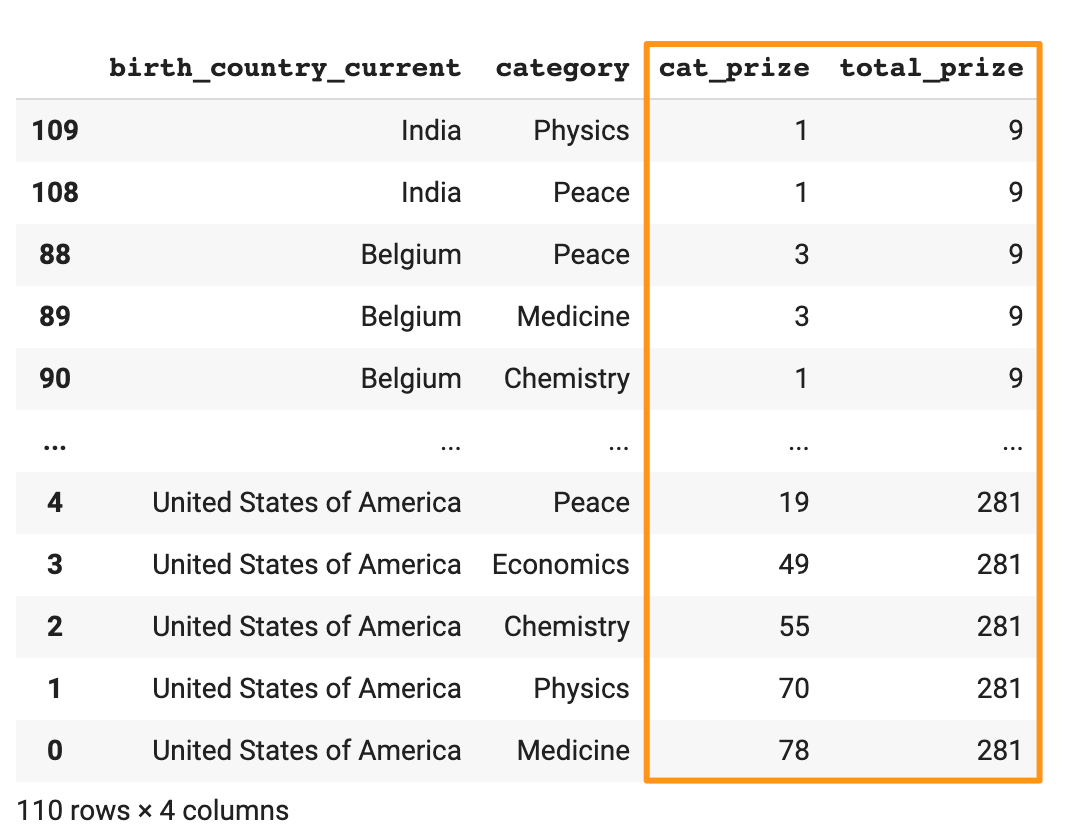
Preparing our data to show the breakdown by category and country is challenging. We'll take a two-step approach here. First we count the prizes by category in each country:

1. cat\_country = df\_data.groupby(['birth\_country\_current', 'category'],
2. as\_index=False).agg({'prize': pd.Series.count})
3. cat\_country.sort\_values(by='prize', ascending=False, inplace=True)



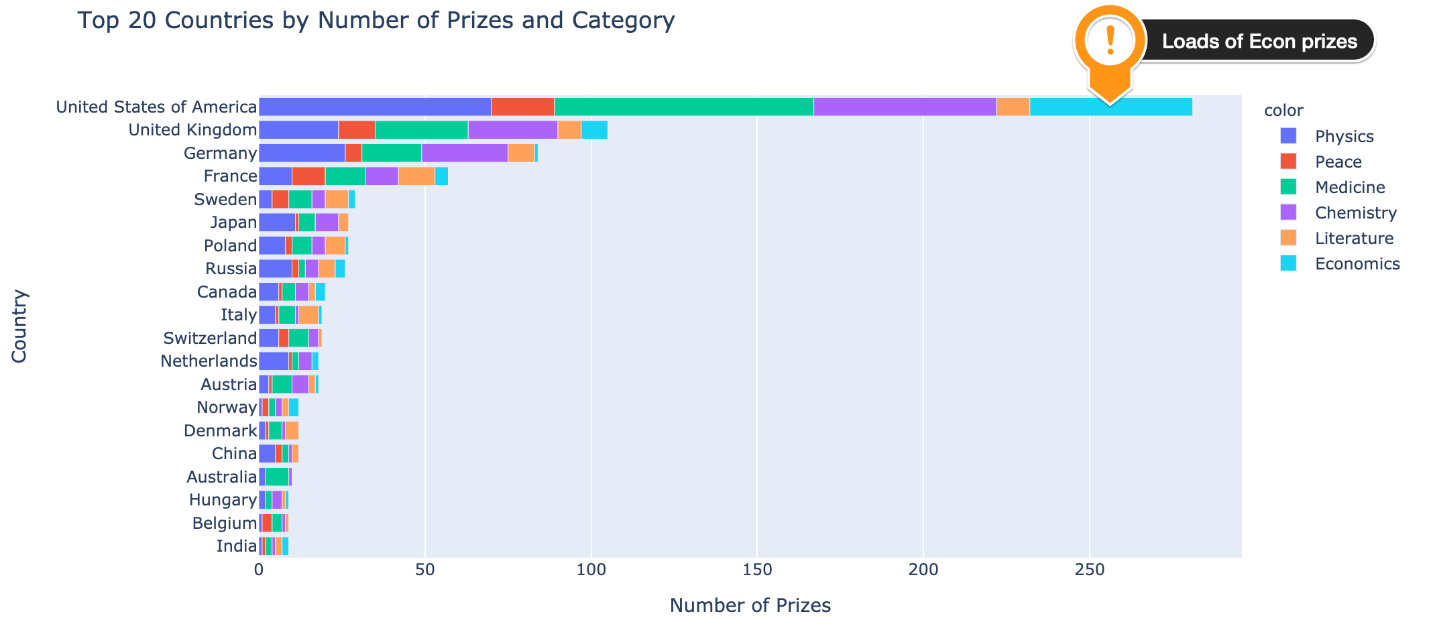
Next, we can merge the DataFrame above with the top20\_countries DataFrame that we created previously. That way we get the total number of prizes in a single column too. This is important since we want to control the order for our bar chart.

1. merged\_df = pd.merge(cat\_country, top20\_countries, on='birth\_country\_current')
2. # change column names
3. merged\_df.columns = ['birth\_country\_current', 'category', 'cat\_prize', 'total\_prize']
4. merged\_df.sort\_values(by='total\_prize', inplace=True)

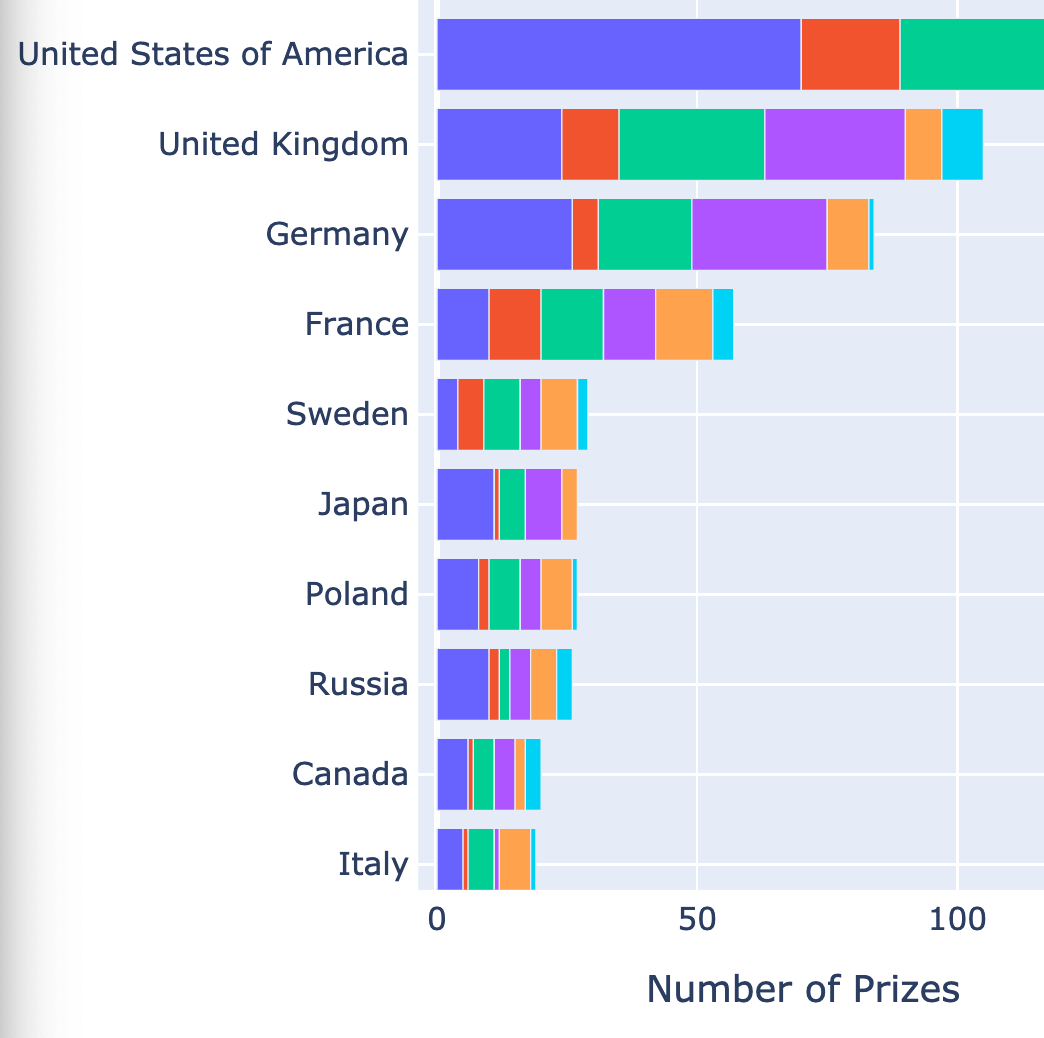


Now we can create our bar chart again. This time we use the color parameter based on the category.

1. cat\_cntry\_bar = px.bar(x=merged\_df.cat\_prize,
2. y=merged\_df.birth\_country\_current,
3. color=merged\_df.category,
4. orientation='h',
5. title='Top 20 Countries by Number of Prizes and Category')
7. cat\_cntry\_bar.update\_layout(xaxis\_title='Number of Prizes',
8. yaxis\_title='Country')
9. cat\_cntry\_bar.show()



Splitting the country bar chart by category allows us to get a very granular look at the data and answer a whole bunch of questions. For example, we see is that the US has won an incredible proportion of the prizes in the field of Economics. In comparison, Japan and Germany have won very few or no economics prize at all. Also, the US has more prizes in physics or medicine alone than all of France's prizes combined. On the chart, we also see that Germany won more prizes in physics than the UK and that France has won more prizes in peace and literature than Germany, even though Germany has been awarded a higher total number of prizes than France.



When did the United States become so dominant? Was it always this way? Has the prize become more global in scope?

**Solution 4: Country Prizes over Time**

To see how the prize was awarded over time. To do that, we can count the number of prizes by country by year.

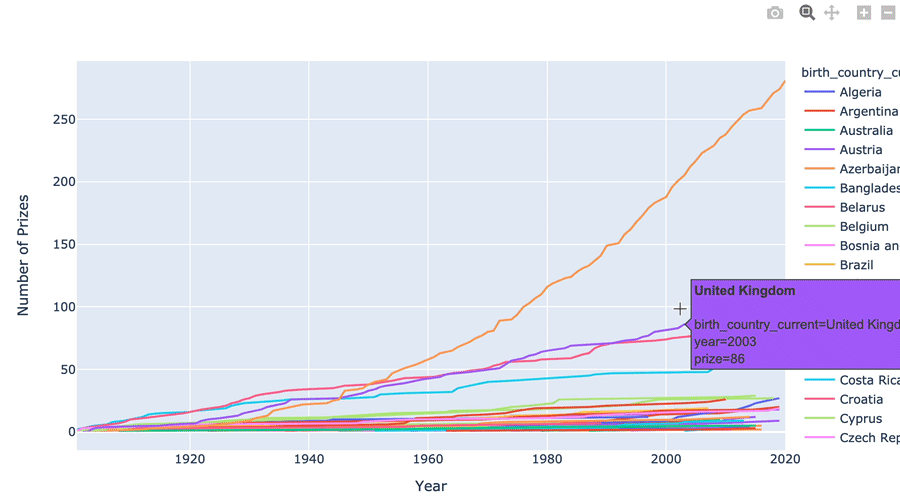
1. prize\_by\_year = df\_data.groupby(by=['birth\_country\_current', 'year'], as\_index=False).count()
2. prize\_by\_year = prize\_by\_year.sort\_values('year')[['year', 'birth\_country\_current', 'prize']]

Then we can create a series that has the cumulative sum for the number of prizes won.

1. cumulative\_prizes = prize\_by\_year.groupby(by=['birth\_country\_current',
2. 'year']).sum().groupby(level=[0]).cumsum()
3. cumulative\_prizes.reset\_index(inplace=True)

Using this, we can create a chart, using the current birth country as the color:

1. l\_chart = px.line(cumulative\_prizes,
2. x='year',
3. y='prize',
4. color='birth\_country\_current',
5. hover\_name='birth\_country\_current')
7. l\_chart.update\_layout(xaxis\_title='Year',
8. yaxis\_title='Number of Prizes')
10. l\_chart.show()



What we see is that the United States really started to take off after the Second World War which decimated Europe. Prior to that, the Nobel prize was pretty much a European affair. Very few laureates were chosen from other parts of the world. This has changed dramatically in the last 40 years or so. There are many more countries represented today than in the early days. Interestingly we also see that the UK and Germany traded places in the 70s and 90s on the total number of prizes won. Sweden being 5th place pretty consistently over many decades is quite interesting too. Perhaps this reflects a little bit of home bias? 😊

All this analysis of different countries makes me curious about where the actual research is happening. Where are the cities and organisations located where people actually make discoveries?