Nin Kongla
Doctor Zietz
INFO 2201
6 December 2022

Data Analysis of the Nobel Prize

Project Description:

For my project, I explored datasets of the Nobel prize by using two different data types with three data sources in total. I chose to use HTML web and CSV files as my data types because these are my top two favorite topics that we had covered in this course. The application that I have applied to my project using HTML is web scraping. I inspected the raw website and scraped through all the tags until I found the data I wanted to extract from that tag. After that, I display the information in an organized way to present a text description.

When I was done with the HTML, I moved on and worked on the CSV files. The first thing I did before anything with those two CSV files was that I turned them into readable dataframes that display rows and columns. After that step, requesting data from the table is easy to do. I added a visualization concept to help the audience better understand the dataframes.

What is the question I am trying to answer? How did I answer my question?:

In the project plan, I said the main question I wanted to answer is: which category has won the most award prize? I am able to answer this question by knowing how to access the dataframe. I found the column that has a category as a title, and then I used the value_counts method of pandas. It outputted the category name for me and added up each category's total occurrences. So, the answer is Medicine has won the most award prize of a total of 220 prizes given out to winners from 1901 to 2020.

There are also three sub-questions that I say I will try to answer in my project plan. The first question is, how many prizes were given in each category. The second question is, which country won the most? The third question is whether some people declined a Nobel prize. I answered the first question by using the same concept of how I was able to answer the main question. After I accessed the category column, I used the value_counts method. The answer to the first question was that there are six categories in the Nobel prize: medicine, chemistry, physics, literature, economics, and peace. From 1901 to 2020, awards given out to winners are 220 in medicine, 186 in chemistry, 216 in physics, 117 in literature, 86 in economics, and 135 in peace. I still use the value_counts method to answer question number two. I had to change from the ‘category’ column to the ‘birth_country’ column. The answer to the second question is that the United States of America won the most prize, a total of 281 prizes. Now, the way I answer the last question is depend on the information that I extract from the HTML website. There is a title and below description on the web that tells who had declined the Nobel prize, which I successfully extracted. The answer is that Jean-Paul Sartre declined the prize because he had consistently declined all official honors.

What were the results? Were they surprising or what you expected?

The fact that I found surprising while doing this project is that there are not many categories given out to Nobel laureates. I thought that they might value giving out the prize to those who did something technical, practical, or inventory, since the founder, Alfred Nobel, himself seems like the person who would hold that value. I expected that there would be more engineering categories.

Another interesting thing I found when I plotted bar charts at the very end of the project is that there seems to be a connection between why women were given more prize amounts than men even though women receive fewer Nobel prizes. My conclusion is that from when the Nobel prize was invented, in earlier decades, men were more likely to receive a Nobel prize than women. A lot of women laureates were given prizes when we moved toward modernity. As time goes on, money value has likely changed as well. With that being said, women got more award prizes because they got paid at a time when there was a big money value difference.

If you were to continue doing work on this topic, what would you do next?

I would find more data source that has more numeric values for me to do some interesting calculations and display them visually. And also ask more questions.