# Nobel Prize Data Visualization

Student ID: 20306228

Student Name: Chao Cui

## Description of data

Widely regarded as the most prestigious prize in the fields of physics, chemistry, medicine, literature, peace and economics, the Nobel Prize is awarded to scholars who have made outstanding contributions in these fields. Although the Nobel Prize is a household name, the public does not have intuitive access to complete information about the Nobel Prizes over the years. Therefore, it is especially important to present the data of the Nobel Prizes from 1901 to 2016 by integrating them and presenting them to the public through a visualization method. The data for this project are obtained from laureate.csv and prize.csv in the Nobel Prize Dataset in the Harvard Dataverse. The laureate.csv contains id, firstname, surname, born, died, bornCountry, bornCountryCode, bornCity, diedCountry, diedCountryCode, diedCity, gender, year, category, overallMotivation, share, motivation, name, city and country. The prize.csv contains year, category, overallMotivation, id, firstname, surname, motivation and share.

## Data pre-processing

Data pre-processing is an integral part of data visualization, which includes filtering, sorting, correlation, missing value cleaning, etc. First, use setwd() to set the R workspace to the path where laureate.csv and prize.csv are located, and use read.csv() to read the external data to respectively be stored in the Laureate and Prize variables in the R environment. Then use names(), and head() to see the contents of the data storage, and use mode() to detect the data as a list type. Finally, use sum(is.na()) to check if there are missing values in the data.

After the above operation, it is found that there are no NA values in Laureate and Prize, and the dataset is relatively clean. However, the string format of "Lennep (now Remscheid)" is present in bornCountry and diedCountry in Laureate. In addition, Northern Ireland and Scotland are not incorporated into the United Kingdom in Laureate. This causes the data to be less intuitive in subsequent visual analysis, so only the country name of the present international community in () needs to be read and Change Northern Ireland and Scotland to the United Kingdom. The stringr package is defined as a consistent, easy-to-use set of string tools. Its built-in functions greatly facilitate the manipulation of strings. The data processing for this project also use the stringr package. First use str\_detect(data,"\\(") to detect if the data contains a (, and return TRUE if it exists, and second use str\_extract(), combined with the regular expression (? <= \(). +? (? =\\))" to extract only the contents inside (). The extracted result is of type "now Remscheid". Nevertheless, there is still "now" in the data, so use str\_detect(data, "now") again to filter out the data which are not contain "now", and use word(data,-1) to extract the last word from the data. For issue that attribute Scotland and Northern Ireland to the United Kingdom, use str\_replace\_all() to replace "Scotland" and "Northern Ireland" in the data.

## Initial question

**Q1: How many Nobel winner have been given in each category?**

To visualize the number of Nobel Prize winners in different disciplines use tally() and group\_by() in the dplyr package to group the data in Prize by category. Meanwhile, use arrange() to sort the obtained data by number of people in ascending order and use mutate() and levels to customize the plotting order for the sorted category. After processing the data, the ggplot2 package was used to plot a bar chart with the category on the x-axis and the number of Nobel winners on the y-axis and using geom\_text() intuitive mark the number over the bar. In addition, stat =" identity" in geom\_bar() means that the sample points correspond to the vertical axis values without any statistical transformation.

According to the bar chart, medicine ranks first in the number of winners, followed by physics, while economics has the lowest number of winners in the Nobel Prize, however, this is since the Nobel Prize in economics was only established in 1968. At the same time, from an objective point of view, the large number of winners of medical and physics prizes shows the rapid progress of science and technology, but at the same time, the small number of winners of literary prizes also needs to make the public aware that literature and culture are also an integral part of the human development process.

**Q2: What is the gender ratio of Nobel Prize?**

From the Laureate data, the gender type of Nobel laureates is divided into "male", "female" and "org". A pie chart is considered a good choice for visualizing the respective percentages of Nobel Prize winners by gender type. First, the data in Laureate were counted in groups of gender using tally() and group\_by(). Subsequently, the percentage of each gender type was calculated and stored in the prop variable using mutate() and n/sum(n), where round() was also used to preserve the calculated values by three decimal places and percent() from the scales package to convert the preserved values into percentage form. Since ggplot() does not have a function specifically for drawing pie charts, the data is first plotted as a bar chart, where the x-axis is the empty string and the y-axis is the prop. Then use coord\_polar() to bend the Y-axis of the bar to form a pie chart. Finally, use geom\_text() and theme() to add prop data to the plot and hide the scale of the pie chart.

The pie chart shows that the percentage of male winners is significantly higher than the percentage of female winners and organizational winners.

**Q3: What is the average age of Nobel Prize winners by category?**

Nobel Prize winners are top people in various fields who have devoted their lives to scientific research, however, it is a topic of interest at what age they all made their field breakthroughs. First use filter() and substr(born,1,4) ! = "0000" to filter out the data in Laureate where the birth date is not recorded. Since there is no prize age, that uses the year of the prize to minus the year of the winner's birth. Therefore, the question uses mutate() and use as.numeric(year) - as.numeric(substr(born,1,4)) to get the prize age of the winner to store in the prize\_age variable. It is worth noting that both year and substr(born,1,4) are strings, so use as.numeric() to convert them to numeric types. Subsequently, the data were consolidated in groups by category using group\_by(), and the average age of each category winner was calculated using summarise() and mean(). Finally, a histogram is plotted using ggplot() + geom\_bar(), with the x-axis being the category, and the y-axis being the average age of Nobel Prize winners .

The bar chart shows that the average age of the winners in economics and literature is in the top two positions, while the average age of the winners in physics and medicine is in the last two positions. Combined with the bar chart in Q1, this may indicate that economics and literature are more difficult to win and require a long period of research and exploration, which leads to a lower number of winners. Physics and Medicine, on the other hand, are relatively easy to break through as times progress, resulting in a higher number of winners.

**Q4 & Q5: Which are the ten birth countries (or death countries) with the most Nobel Prize winners?**

The honor that comes with a Nobel Prize can greatly enhance the country's visibility. Therefore, the countries of birth and death of Nobel Prize laureates are often in the spotlight as well. Although the Laureate data has been preprocessed in this project, there are still null characters stored in the bornCountry(or diedCountry) in it, which is due to the unknown information of the country of birth and country of death for some winners. Therefore, the null characters in the data need to be filtered out using the filter(Laureate,bornCountry!="") in the dplyr package. The filtered information is then counted in groups of bornCountry (or diedCountry) using tally and group\_by(). Meanwhile, using arrange() to sort the bornCountry(or diedCounty) in ascending order, followed by tail(10) to select top 10 birth countries(or death countries). The problem still uses mutate() with levels to customize the order of the bar chart. Finally, ggplot() + geom\_bar() is used to draw the desired bar chart, where the x-axis is bornCountry (or diedCountry) and the y-axis is the number ot Nobel Prize winners.

As it can be seen from the bar chart of Q4 & Q5, the five countries of USA, United Kingdom, Gemany, France and Sweden are consistently in the top 5, both in terms of countries of birth and countries of death. In addition, these five countries are also internationally influential countries, which may attract elites in various fields, thus leading them to have more Nobel Prize winners.

## Advanced question

**Q6: What is ratio of male and female winners in each category?**

In questions Q1 and Q2, the project explored the number of winners in each category and the overall gender ratio of Nobel Prize winners. By combining questions Q1 and Q2, the proportion of male and female laureates in each category becomes further problem to explored the relationship between Q1 and Q2.

According to Q2, it is known that the gender types of Nobel laureates are "male", "female" and "org". Therefore, this question needs to filter out the data with "org" by using filter(gender == "male" | gender == "female"). When the data is filtered, tally() and group\_by() are used to count the data in the group of gender, and mutate() is used to add a new prop variable to the data according to the method of Q2, which stores the respective percentages of men and women in each category after being percentile by percent(). For visualization, the question plots the data as a bar chart using ggplot() + geom\_bar() and bends the bar chart to a pie chart with the y-axis via coord\_polar(). It is worth noting that facet\_wrap() is used here to divide the pie chart into multiple plots in the same panel conditional on category. Subsequently, to make the image more concise and clear, the question uses labs(), geom\_text() and theme() to add labels, text information and hide the pie chart scale.

I It can be seen here that female recipients are mainly concentrated in the fields of literature and peace, with only 0.9% of female recipients in the field of physics. This may imply that women are more talented at advocating for peace and creating literature than in other fields. On the contrary, the percentage of men in the field of physics is higher than other fields, and the percentage in the field of peace is the lowest, at 84.60%.

**Q7: What is the migration pattern between the birth country and death country of Nobel laureates?**

Q7 was developed as an extension of Q4 and Q5 to explore a migration pattern of Nobel laureates from country of birth to country of death. Since it is a migration model, it is necessary to filter out information about prize winners whose country of birth and country of death are the same. Use filter(bornCountry!="" & diedCountry ! = "" & bornCountry ! = diedCountry) to remove the data without country information and where the country of birth and the country of death are the same. The filtered data is used to customize the colors for subsequent plots using mutate(). The customization process uses case\_when() from the dplyr package, which assigns different colors to the top ten countries by combining the top ten bornCountries obtained from Q4. To clearly represent the migration model, the geom\_curve() from the ggplot2 package is used here to plot the curves, where the x-axis is birth country, and the y-axis is death country. Scale\_x\_discrete() and scale\_y\_discrete() are also added to set the x and y discrete scale aesthetics values. Finally, geom\_text(), labs() and theme() are still used to make visual images more intuitive and accurate.

The image of this question shows that a lot of Nobel Prize winners are born from one country and end up dying in another. This suggests part of the question of whether certain countries do not value their scientists, or whether certain countries do not have enough capacity to conduct scientific research. In addition, according to the image, there are far more migration curves in green than in other colors, which may imply that these countries with the top 10 Nobel Prize birth rates have their own unique education systems, or that these countries have sufficient funding to support their own scholars.

**Q8: What are the trends in the age of Nobel Prize winners in each category?**

The average age of Nobel Prize winners in each category can be known from Q3. As an extension of the question, the project needs to explore the trend of the age of Nobel Prize winners in each category. First use filter() and substr(born,1,4) ! = "0000" to filter out the data in Laureate where the birth date is not recorded. Since there is no prize age, that uses the year of the prize to minus the year of the winner's birth. Therefore, the question uses mutate() and use as.numeric(year) - as.numeric(substr(born,1,4)) to get the prize age of the winner to store in the prize\_age variable. It is worth noting that both year and substr(born,1,4) are strings, so use as.numeric() to convert them to numeric types. Subsequently, the data were consolidated in groups by category using group\_by(). Finally, a scatter chart is plotted using ggplot() + geom\_point(), with the x-axis being the category, and the y-axis being the age of Nobel Prize winners and use facet\_wrap() to draw scatter charts in one drawing board by each category. Meanwhile, use geom\_smooth(), method "lm", to explore the trend of Nobel Prize winners' age using linear equations.

As can be seen from the picture, the average age of Nobel laureates in the fields of chemistry, physics and medicine has shown a significant increasing trend over time, while the fields of economics and literature have shown a stable trend, in contrast, the average age of Peace Prize laureates has shown a significant decreasing trend.

## Reflection

The project visualized and analyzed a dataset of Nobel laureates, which allows the public to better understand the message behind the Nobel Prize. Although the project visualized the dataset in terms of the number of awards, age, gender ratio, and country context, there are still additional questions that need further analysis of other data in the dataset, such as a visual exploration of the motivations for the awards. Meanwhile, the visualized questions can be extended with other datasets, such as the migration pattern of countries of birth and death, by combining the datasets of education level and research funding in each country to explore whether education and research funding attract more Nobel laureates. In addition, the project had attempted to visualize the number of Nobel Prize winners in combination with the Remap package in the form of a map, while the gradient color intervals represented the number of winners in each country. However, the method was not successfully implemented due to the mismatch of country names, and the project will try to use the countrycode from the Laureate data and combine it with other datasets to solve this problem in the subsequent Nobel Prize visualization analysis.

## Appendix

图表, 条形图

描述已自动生成Figure 1: Q1

图表, 饼图

描述已自动生成

Figure 2: Q2

Figure 3: Q图表, 条形图

描述已自动生成3

图表, 条形图

描述已自动生成

Figure 4: Q4

图表, 条形图

描述已自动生成Figure 5: Q5

示意图

低可信度描述已自动生成

Figure 6: Q6

图表, 图示

描述已自动生成Figure 7: Q7

图表, 散点图

描述已自动生成

Figure 8: Q8