Analysing trends in Noble Prize winners using R

Summer Internship 2020

Submitted by:

Bhartendu Dubey (9917103102)

Under the supervision of:

Ms. Ambalika Sarkar

Department of CSE/IT

Jaypee Institute of Information Technology University, Noida

June 2020

Table of Contents

1. INTRODUCTON

The Nobel Prize is a set of annual international awards bestowed in several categories by Swedish and Norwegian institutions in recognition of academic, cultural, or scientific advances around the world. The will of the Swedish chemist, engineer and industrialist Alfred Nobel established the five Nobel prizes in 1895. The prizes in Chemistry, Literature, Peace, Physics, and Physiology or Medicine were first awarded in 1901. The prizes are widely regarded as the most prestigious awards available in their respective fields.

1.1 General Description

The Nobel Foundation, a private institution established in 1900, has ultimate responsibility for fulfilling the intentions in Alfred Nobel's will. The main mission of the Nobel Foundation is to manage Alfred Nobel's fortune in a manner that ensures a secure financial standing for the Nobel Prize over the long term and that the prize-awarding institutions are guaranteed independence in their work of selecting recipients.

The Nobel Foundation also strives to safeguard the prize-awarding institutions' common interests and to represent the Nobel organisation as a whole. In the past two decades a number of outreach activities have been developed with the aim of inspiring and disseminating knowledge about the Nobel Prize.

1.2 Objective and scope of the project

The prime objective of this project work is to analyse the rich history of over 100 years of the Nobel Prize winners. This Project will give useful insights using various data visualisations of R language. The Nobel laureate data was acquired from the Nobel Prize API which includes a record for every individual or organization that was awarded the Nobel Prize since 1901.

2. REQUIREMENT ANALYSIS

2.1 Hardware Requirements:

Processor	Intel® Core™ i5-8250U CPU@ 1.6GHz
RAM	8.00 GB
GPU	Nvidia GeForce MX150
GPU Size	4.00 GB
System type	64-bit operating system, x-64 based processor

2.2 Software requirements:

R lang.

RStudio

3. DETAILED ANALYSIS & IMPLEMENTATION

Following Libraries were used for working with the Nobel Data.

- **tidyverse**: It collects some of the most versatile R packages (ggplot2, dplyr, tidyr, readr, purrr, and tibble) which work in harmony to clean, process, model, and visualize data.

- **ggthemes**: It provides Extra Themes, Scales and Geoms for 'ggplot2'.

- **RColorBrewer**: It is an R package that contains a ready-to-use color palettes for creating beautiful graphics for different plots.

- **kableExtra**: It helps in building common complex tables and manipulate table styles.

- **knitr**: It is an engine for dynamic report generation with R. It is a package in the statistical programming language R that enables integration of R code into LaTeX, LyX, HTML, Markdown, AsciiDoc, and reStructuredText documents.

- **ggrepel**: It is an add-in for the ggplot2 package and it makes the text label show up nicely.

- **scales**: This package is to customise to control the appearance of axis and legend labels.

- **gridExtra**: It Provides a number of user-level functions to work with "grid" graphics, notably to arrange multiple grid-based plots on a page, and draw tables.

- **tidytext**: This package provide functionality to tokenize elements by commonly used units of text like these and convert to a one-term-per-row format.

- **wordcloud**: The word cloud generator package (wordcloud) are available in R to analyze texts and to quickly visualize the keywords as a word cloud.

- **lubridate**: It is an R package that makes it easier to work with dates and times.

```
> library(tidyverse)
-- Attaching packages --------------------------------------- tidyverse 1.3.0 --
v ggplot2 3.3.0     v purrr   0.3.4
v tibble  3.0.1     v dplyr   0.8.5
v tidyr   1.1.0     v stringr 1.4.0
v readr   1.3.1     v forcats 0.5.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
> library(ggthemes)
> library(RColorBrewer)
> library(kableExtra)

Attaching package: 'kableExtra'

The following object is masked from 'package:dplyr':

    group_rows

> library(knitr)
> library(ggrepel)
> library(scales)

Attaching package: 'scales'

The following object is masked from 'package:purrr':

    discard

The following object is masked from 'package:readr':

    col_factor

> library(gridExtra)

Attaching package: 'gridExtra'

The following object is masked from 'package:dplyr':

    combine

> library(tidytext)
> library(wordcloud)
> library(lubridate)

Attaching package: 'lubridate'

The following objects are masked from 'package:dplyr':

    intersect, setdiff, union

The following objects are masked from 'package:base':

    date, intersect, setdiff, union

>
```

Fig.1. Required Libraries

- **DataSet Properties:**

```
> nobel<-read.csv("D:/groot/Summer/archive.csv")
> str(nobel)
'data.frame':   969 obs. of  18 variables:
 $ Year                : int  1901 1901 1901 1901 1901 1901 1902 1902 1902 1902 ...
 $ Category            : chr  "Chemistry" "Literature" "Medicine" "Peace" ...
 $ Prize               : chr  "The Nobel Prize in Chemistry 1901" "The Nobel Prize in Literature 1901" "The Nobel Prize in
 Physiology or Medicine 1901" "The Nobel Peace Prize 1901" ...
 $ Motivation          : chr  "\"in recognition of the extraordinary services he has rendered by the discovery of the laws
 of chemical dynamic"| __truncated__ "\"in special recognition of his poetic composition, which gives evidence of lofty id
ealism, artistic perfection"| __truncated__ "\"for his work on serum therapy, especially its application against diphtheri
a, by which he has opened a new ro"| __truncated__ "" ...
 $ Prize.Share         : chr  "1/1" "1/1" "1/1" "1/2" ...
 $ Laureate.ID         : int  160 569 293 462 463 1 161 571 294 464 ...
 $ Laureate.Type       : chr  "Individual" "Individual" "Individual" "Individual" ...
 $ Full.Name           : chr  "Jacobus Henricus van 't Hoff" "Sully Prudhomme" "Emil Adolf von Behring" "Jean Henry Dunan
t" ...
 $ Birth.Date          : chr  "1852-08-30" "1839-03-16" "1854-03-15" "1828-05-08" ...
 $ Birth.City          : chr  "Rotterdam" "Paris" "Hansdorf (Lawice)" "Geneva" ...
 $ Birth.Country       : chr  "Netherlands" "France" "Prussia (Poland)" "Switzerland" ...
 $ Sex                 : chr  "Male" "Male" "Male" "Male" ...
 $ Organization.Name   : chr  "Berlin University" "" "Marburg University" "" ...
 $ Organization.City   : chr  "Berlin" "" "Marburg" "" ...
 $ Organization.Country: chr  "Germany" "" "Germany" "" ...
 $ Death.Date          : chr  "1911-03-01" "1907-09-07" "1917-03-31" "1910-10-30" ...
 $ Death.City          : chr  "Berlin" "ChÃ¢tenay" "Marburg" "Heiden" ...
 $ Death.Country       : chr  "Germany" "France" "Germany" "Switzerland" ...
>
```

Fig.2. Structure of Dataset

```
> head(nobel,3)
  Year   Category                                    Prize
1 1901  Chemistry             The Nobel Prize in Chemistry 1901
2 1901 Literature            The Nobel Prize in Literature 1901
3 1901   Medicine The Nobel Prize in Physiology or Medicine 1901

                                                                                                                Motivation
1                                                                                                  "in recognition of the e
xtraordinary services he has rendered by the discovery of the laws of chemical dynamics and osmotic pressure in solutions"
2                                                                  "in special recognition of his poetic composition, which
 gives evidence of lofty idealism, artistic perfection and a rare combination of the qualities of both heart and intellec
t"
3 "for his work on serum therapy, especially its application against diphtheria, by which he has opened a new road in the
 domain of medical science and thereby placed in the hands of the physician a victorious weapon against illness and death
s"
  Prize.Share Laureate.ID Laureate.Type                  Full.Name Birth.Date        Birth.City    Birth.Country
1         1/1         160    Individual Jacobus Henricus van 't Hoff 1852-08-30         Rotterdam      Netherlands
2         1/1         569    Individual             Sully Prudhomme 1839-03-16             Paris           France
3         1/1         293    Individual      Emil Adolf von Behring 1854-03-15 Hansdorf (Lawice) Prussia (Poland)
   Sex  Organization.Name Organization.City Organization.Country Death.Date Death.City Death.Country
1 Male  Berlin University           Berlin              Germany 1911-03-01     Berlin       Germany
2 Male                                                           1907-09-07  ChÃ¢tenay        France
3 Male Marburg University          Marburg              Germany 1917-03-31    Marburg       Germany
>
```

Fig.3.The Noble Dataset

- **Exploratory Data Analysis [EDA]:**

  - **#Gender**

    Finding gender ratio of Nobel prize winners.

Sex	n
	26
Female	50
Male	893

Fig.4. Gender Ratio

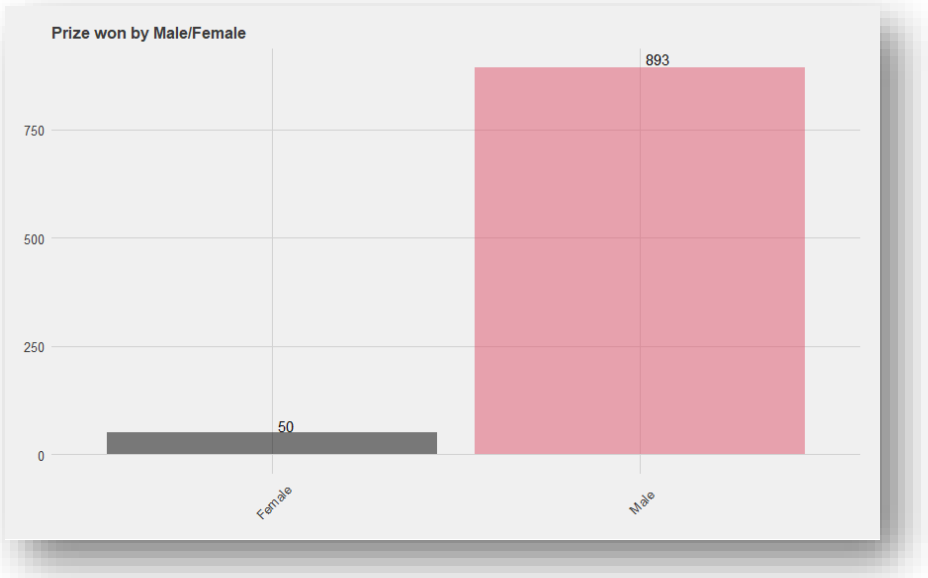

Fig.5. Gender Ratio Plot

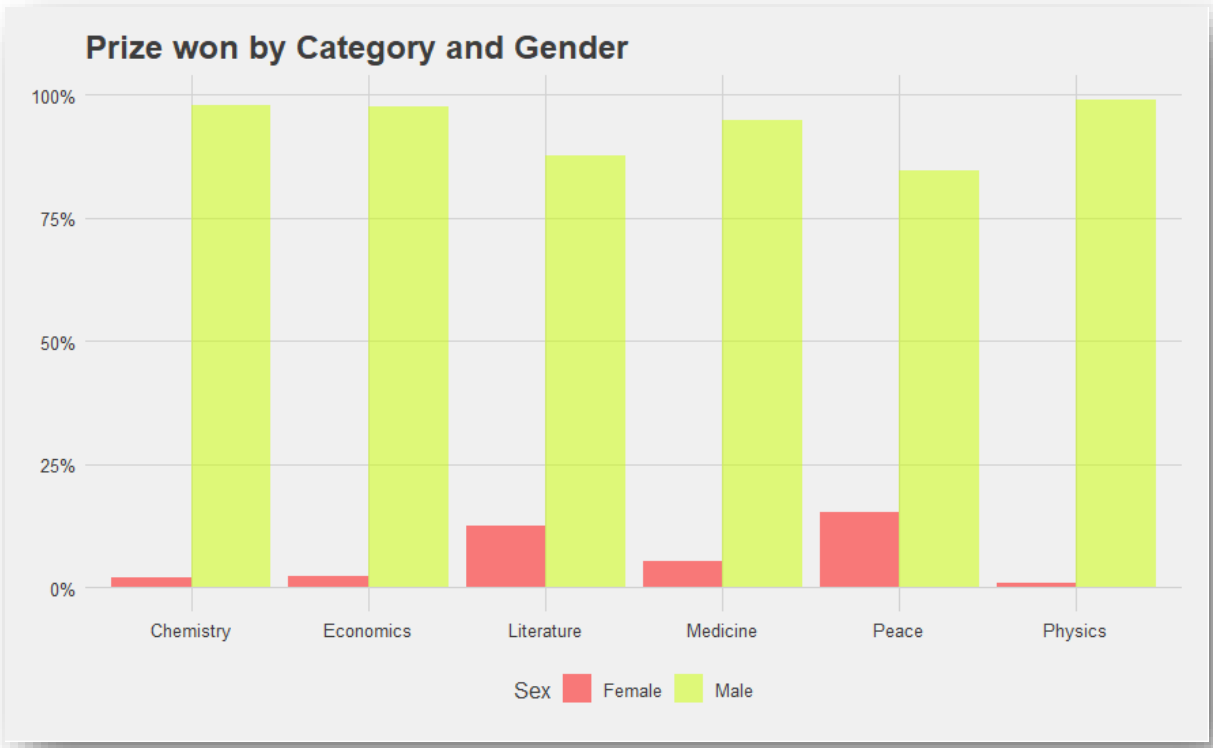

Fig.6. Category vs Gender plot

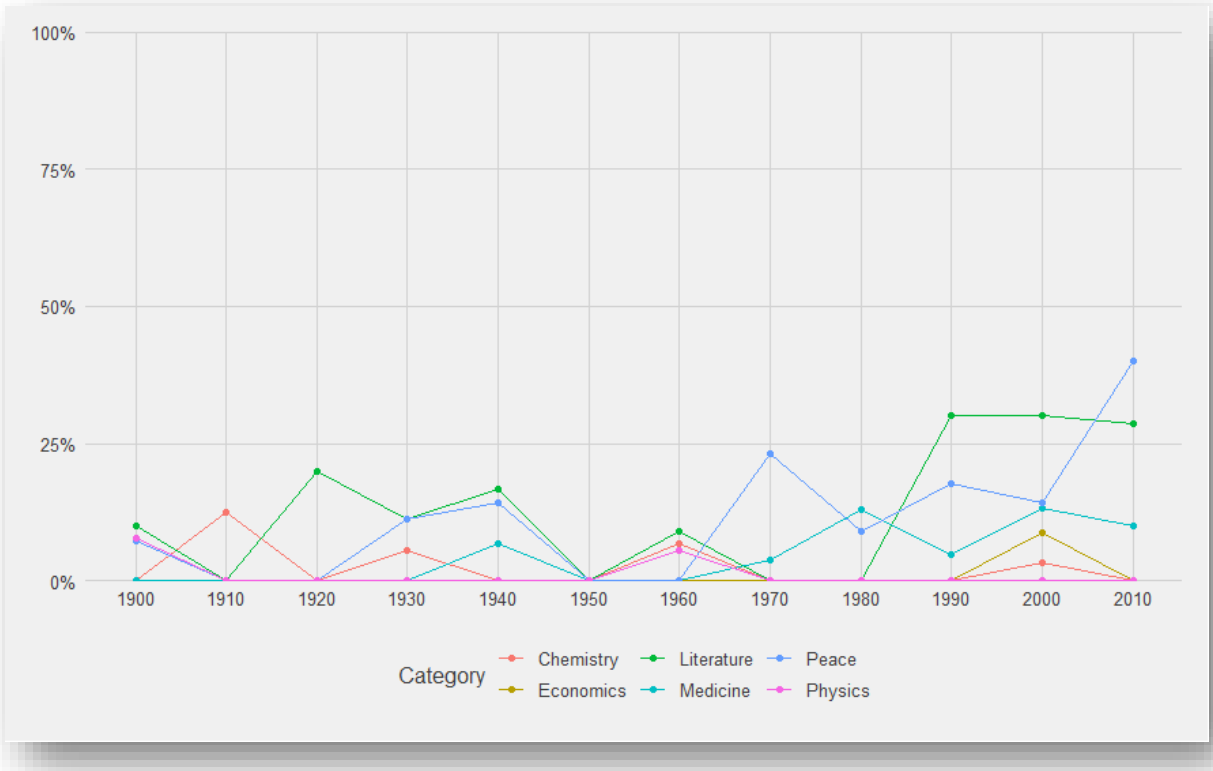

Fig.7. Female Laureates Proportion per decade

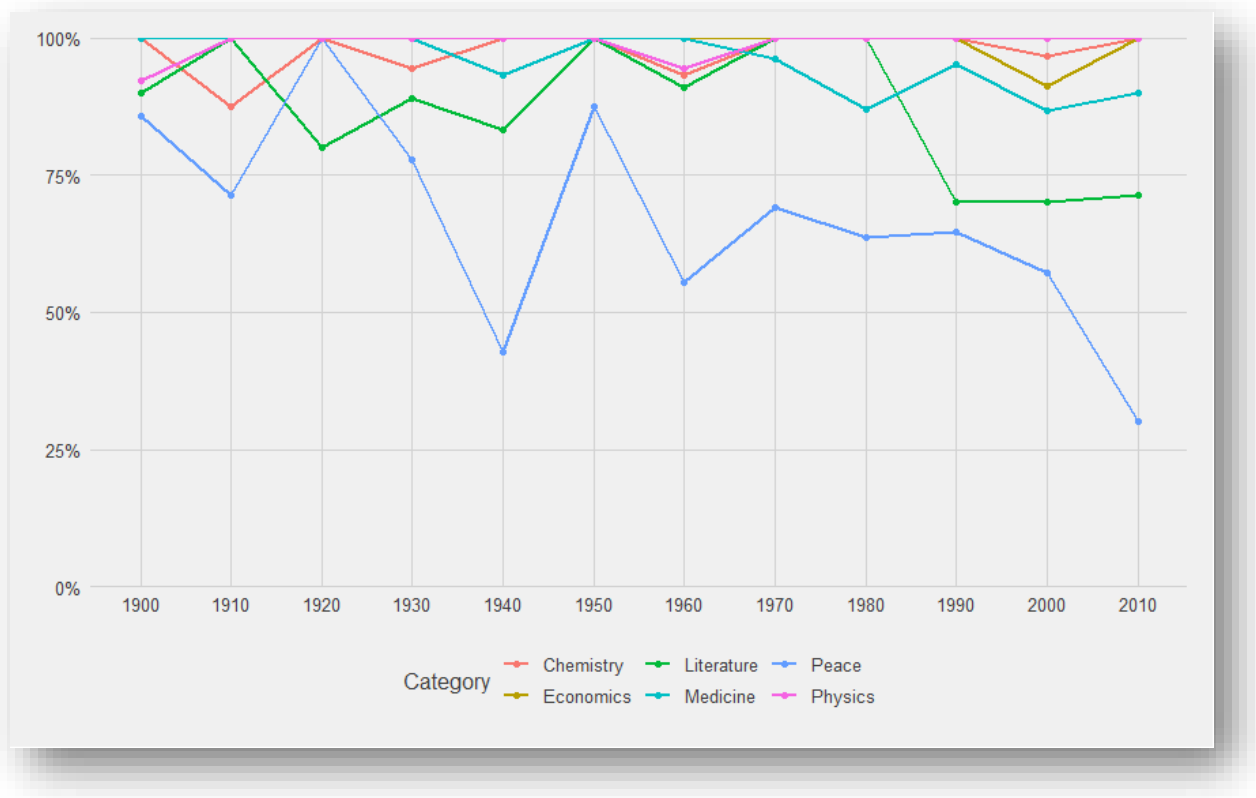

Fig.8. Male Laureates Proportion per decade

- **#Category:**

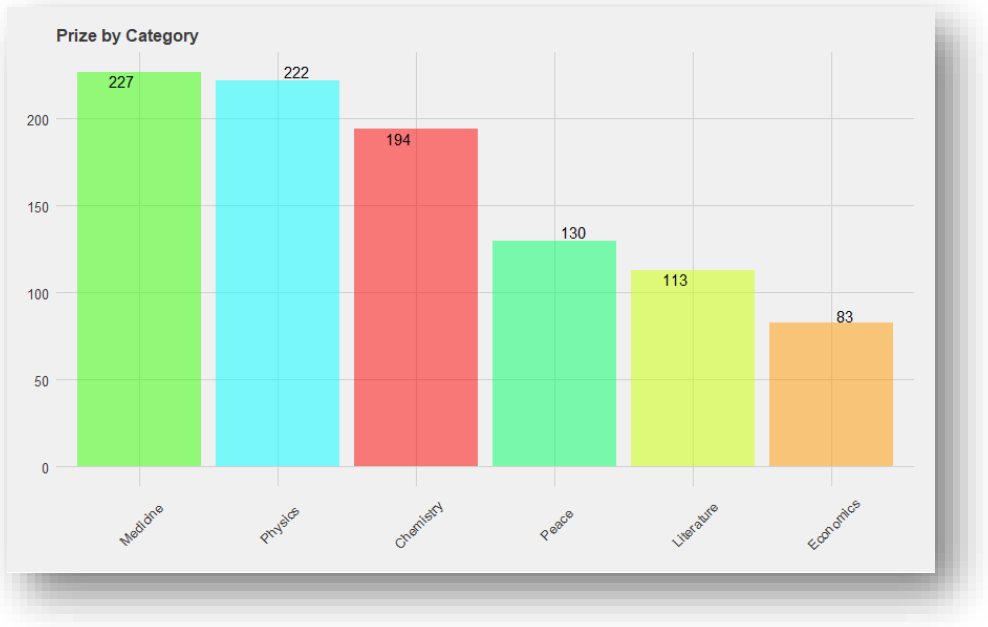

Fig.9. Prizes by category

- **#Sharing Prizes:**

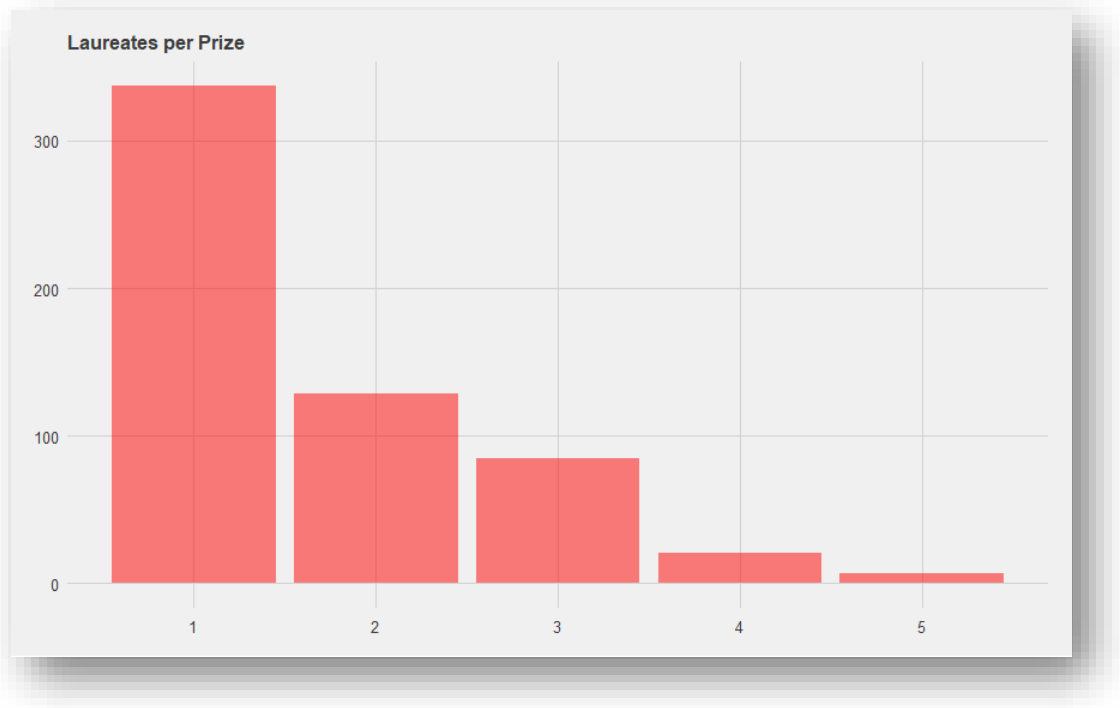

Fig.10. Laureates count Per Prize

- **#Multiple Nobel Laureates:**

Laureates/organizations may receive Nobel prize more than once, for their work in same or different field/topic.

Laureates won Prize more than once

Full.Name	Prize_Count
Comité international de la Croix Rouge (International Committee of the Red Cross)	3
Jack W. Szostak	3
Abdus Salam	2
Adam G. Riess	2
Adolf Friedrich Johann Butenandt	2

Fig.11. Laureates with multiple Nobel Prizes

➢ **#Age:**

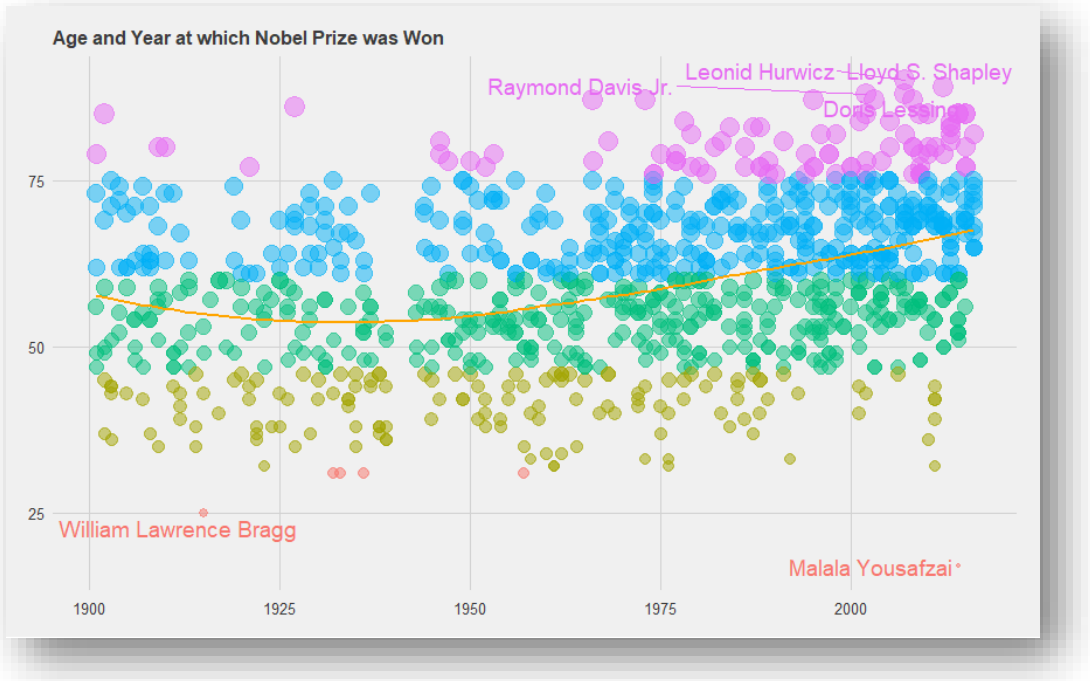

Fig.12. Age vs Year of Nobel Prize Win

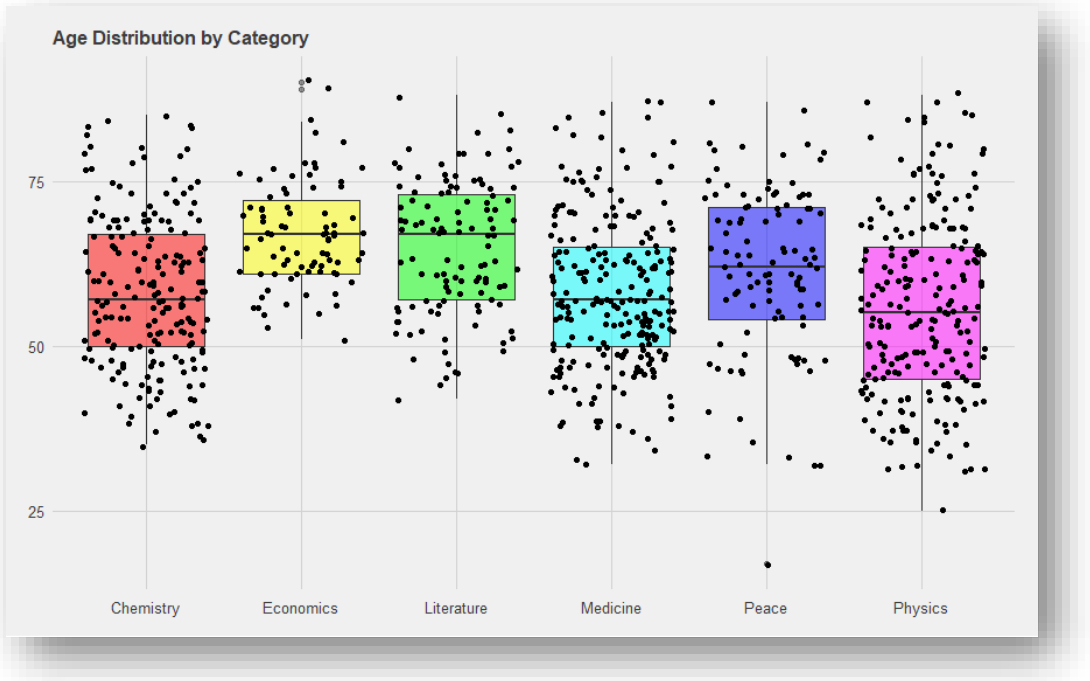

Fig.13. Age Distribution by Category

Fig.14. Age Trend for receiving Nobel price per Category

Age	Full.Name	Year	Category
17	Malala Yousafzai	2014	Peace
25	William Lawrence Bragg	1915	Physics
31	Werner Karl Heisenberg	1932	Physics
31	Paul Adrien Maurice Dirac	1933	Physics
31	Carl David Anderson	1936	Physics

Fig.15. Youngest Nobel laureates

Age	Full.Name	Year	Category
90	Leonid Hurwicz	2007	Economics
89	Lloyd S. Shapley	2012	Economics
88	Raymond Davis Jr.	2002	Physics
88	Doris Lessing	2007	Literature
87	Peyton Rous	1966	Medicine

Fig.16. Oldest Nobel laureates

Year	Full.Name	Category	Birth.Country
1903	Marie Curie, née Sklodowska	Physics	Russian Empire (Poland)
1905	Baroness Bertha Sophie Felicita von Suttner, née Countess Kinsky von Chinic und Tettau	Peace	Austrian Empire (Czech Republic)
1909	Selma Ottilia Lovisa Lagerlöf	Literature	Sweden
1911	Marie Curie, née Sklodowska	Chemistry	Russian Empire (Poland)
1947	Gerty Theresa Cori, née Radnitz	Medicine	Austria-Hungary (Czech Republic)
2009	Elinor Ostrom	Economics	United States of America
2009	Elinor Ostrom	Economics	United States of America

Fig.17. First Female Laureate in each category

➢ **#Life Span:**

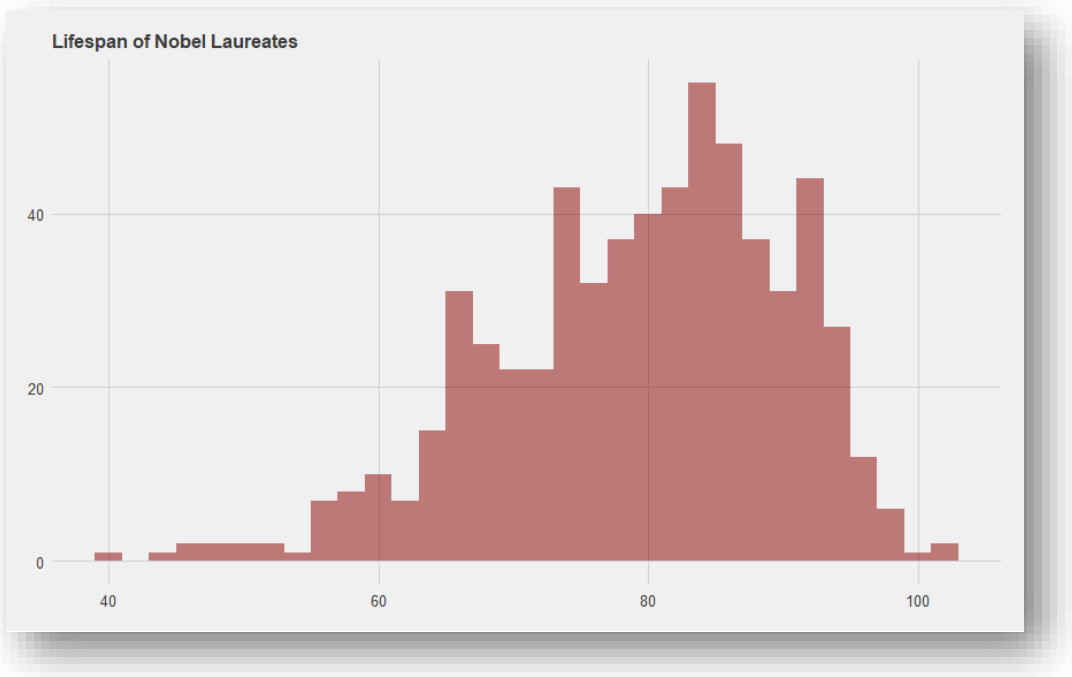

Fig.18. Life Span of Nobel Laureates

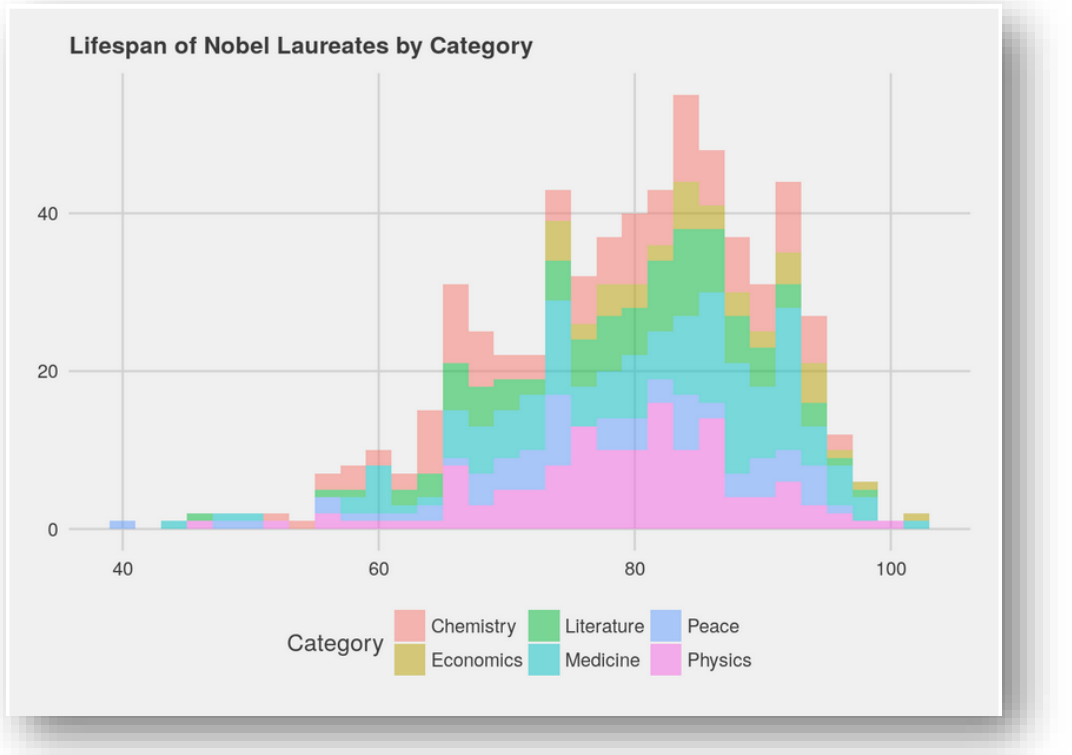

Fig.19. Life Span By category

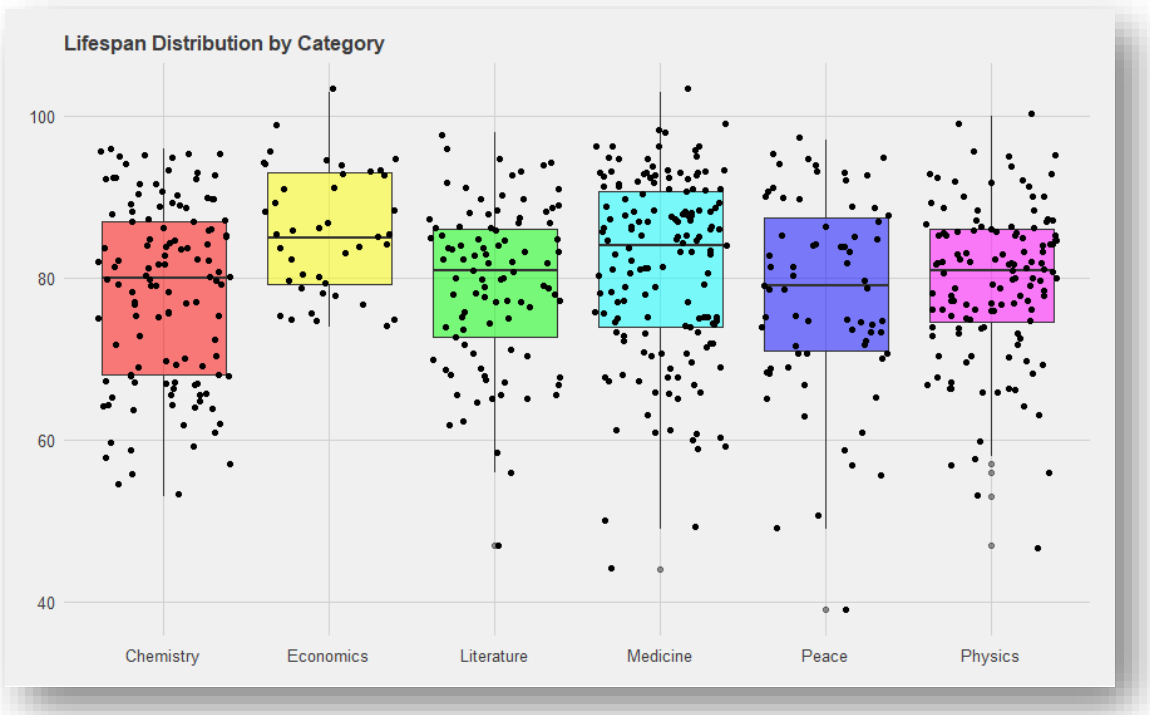

Fig.20. LifeSpan Distribution by Category

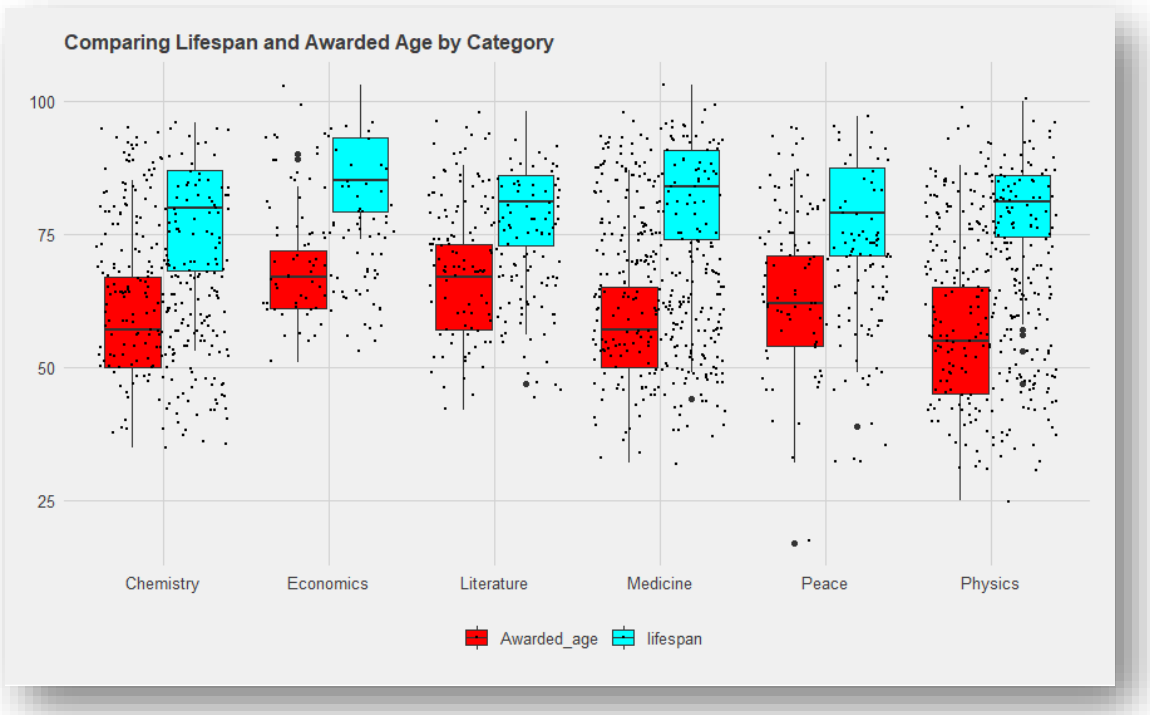

Fig.21. LifeSpan Vs Awarded age

- **#Country:**

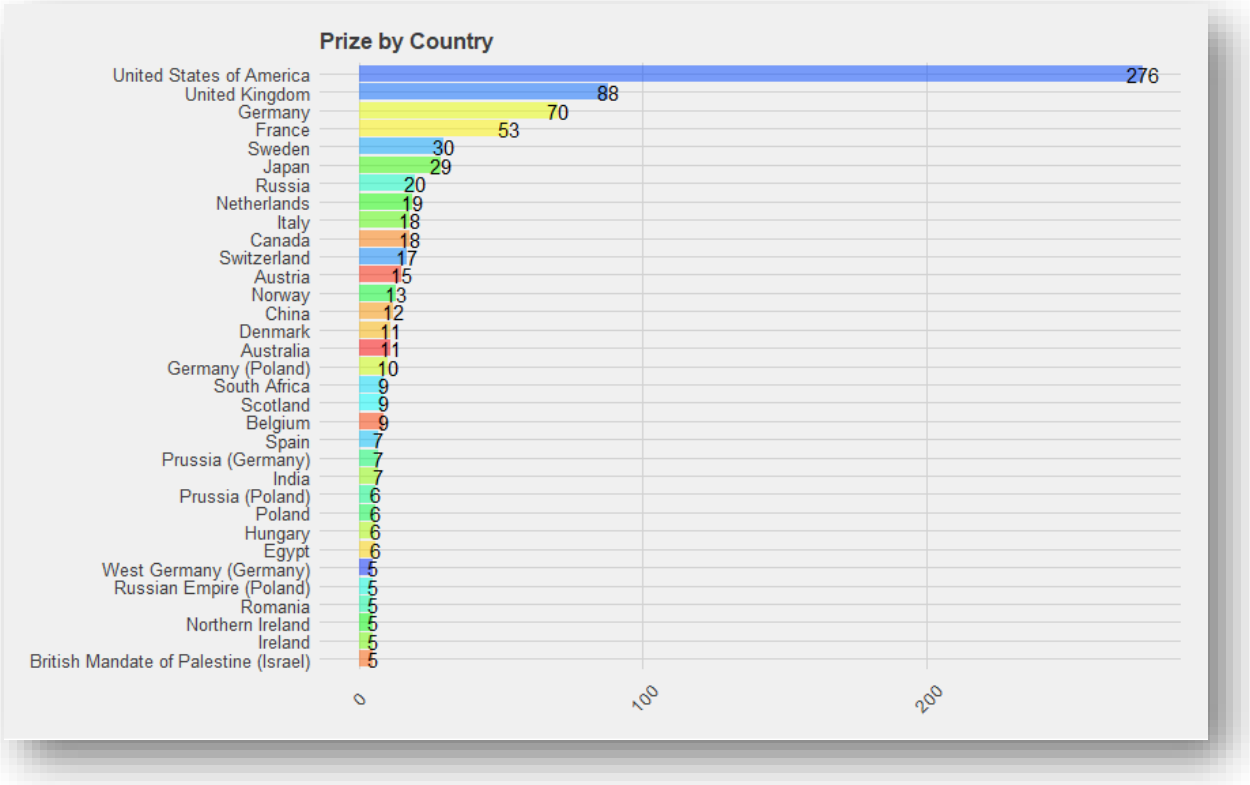

Fig.22. Prizes by country

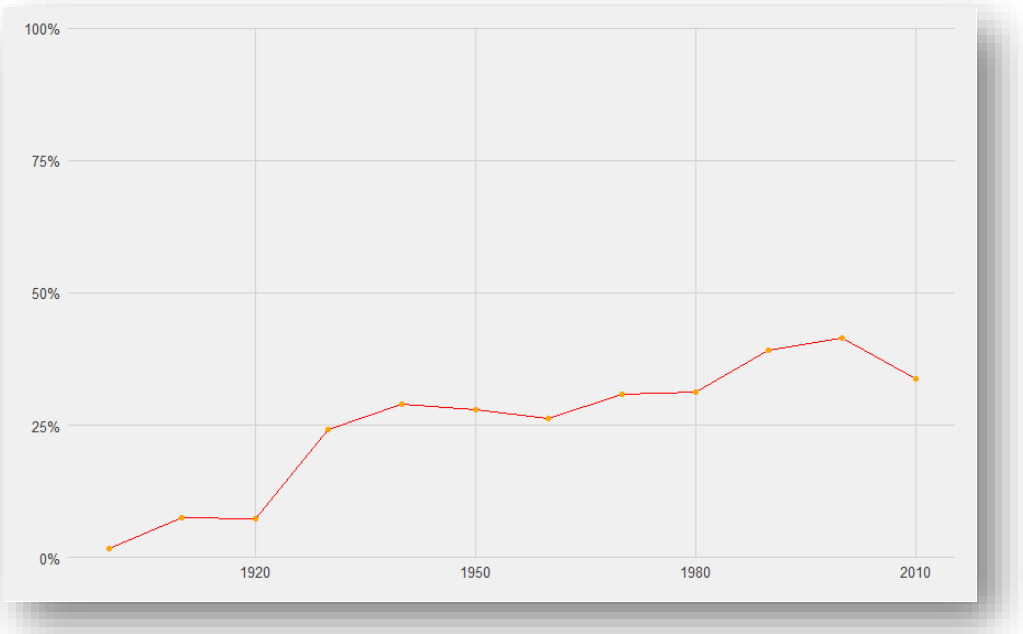

Fig.23. USA-Year wise proportion

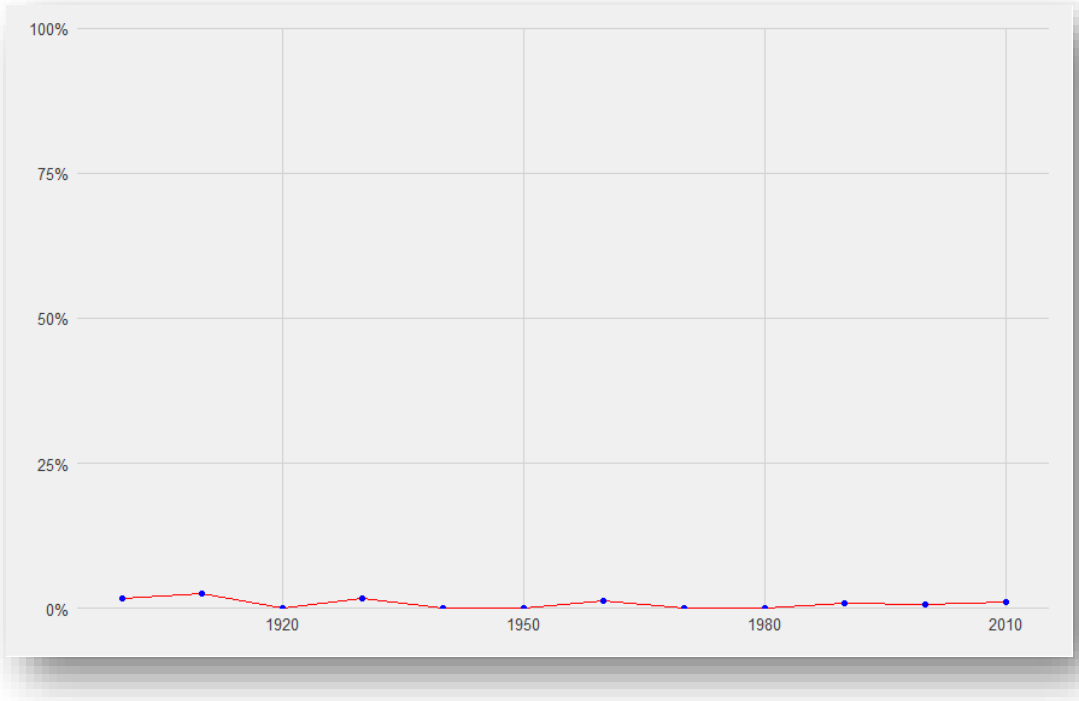

Fig.24. India-Year wise proportion

Top Countries who got nobel prize for Peace

region	value
United States of America	19
France	9
Germany	5
Sweden	5
United Kingdom	5
Northern Ireland	4
Belgium	3
Egypt	3
South Africa	3
Switzerland	3

Fig.25. Peace Laureates by Birth Country

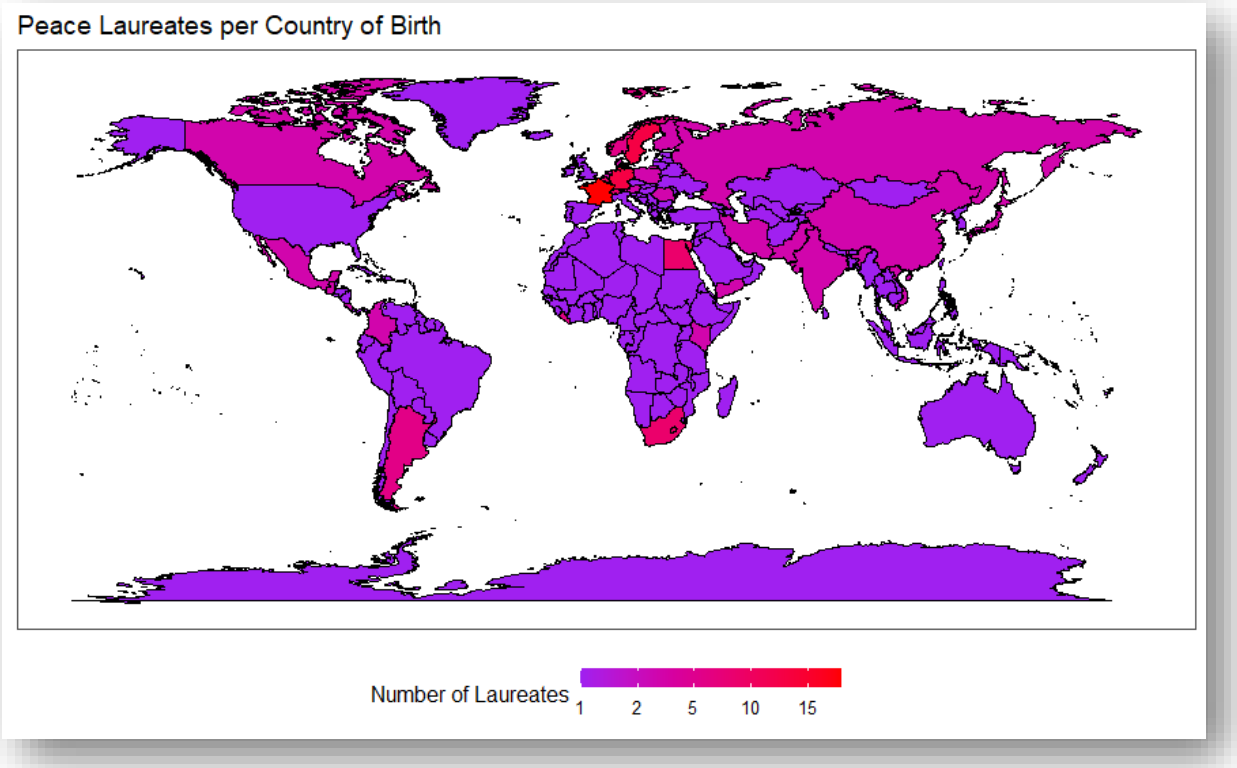

Fig.26. Peace Laureates country map

- **#Finding Prize Motivation:**

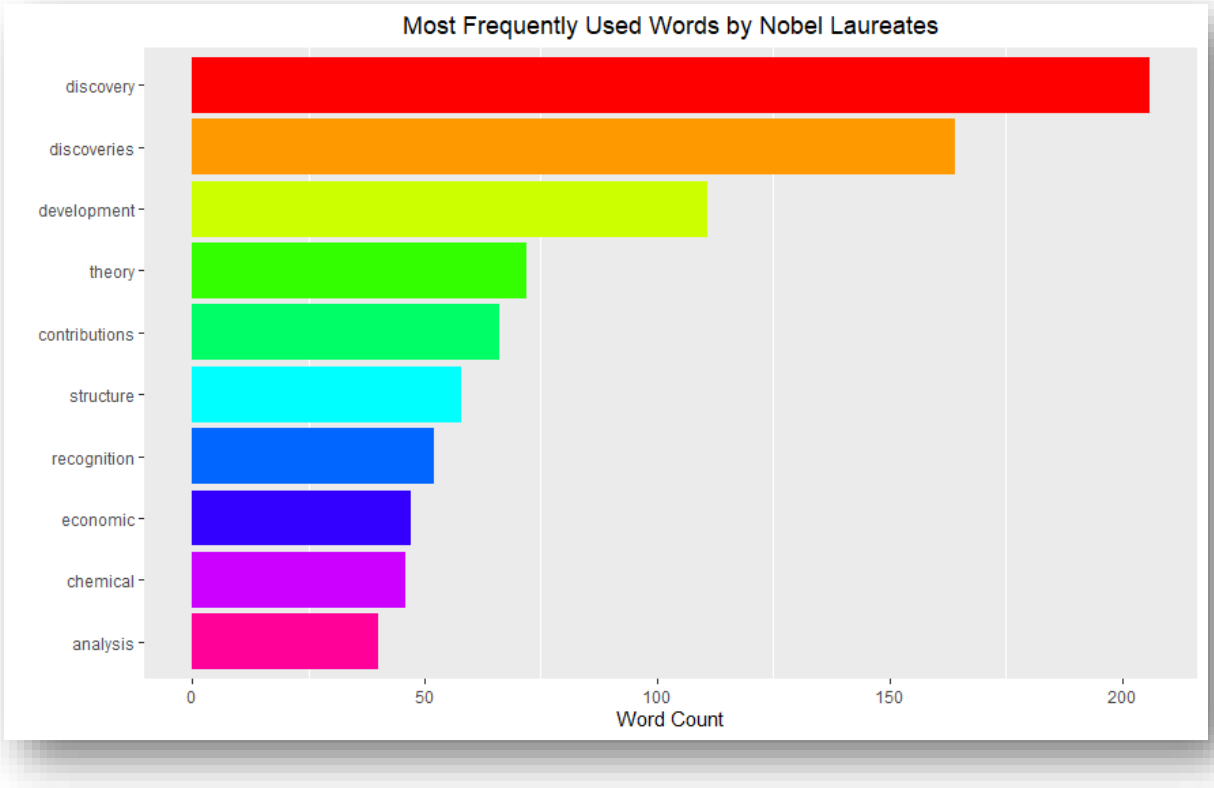

Fig.27. Most frequently used words by Nobel laureates

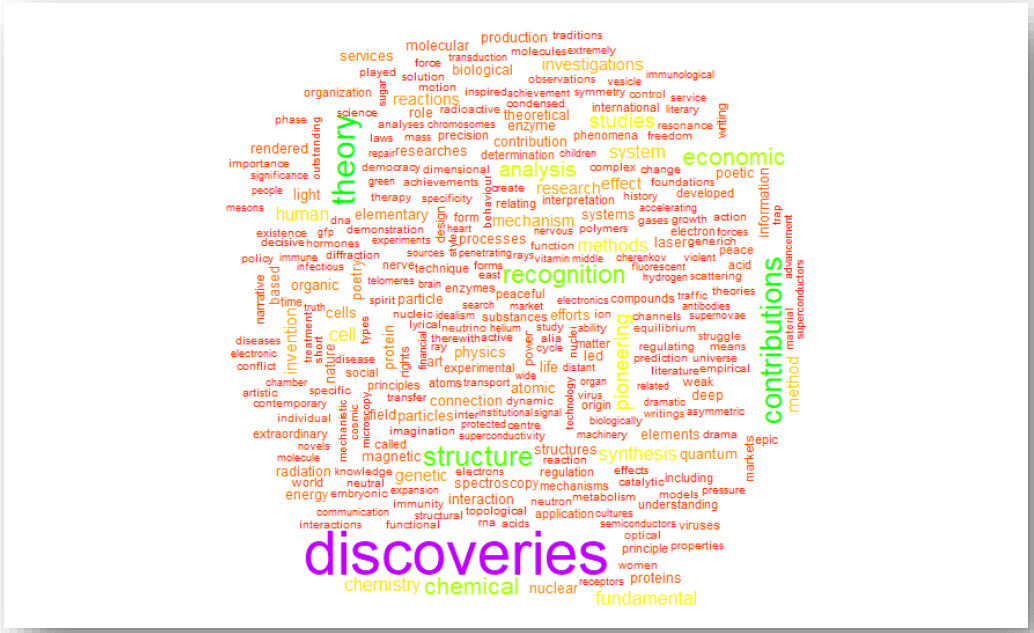

Fig.28. Word Cloud of motivation key-terms

4. CONCLUSION

The data was analysed and visualised to gain useful insights for a detailed study of over a century old rich history of The Nobel Prize winners. The insights such as country winning the most prizes in each category and most frequently written words in the prize motivation were analysed based on age, gender, and nationality of past Nobel laureates. This data can be used further to predict the next Nobel laureates based on the certain factors such as age, gender, nationality of nominees.