Project Nobel Prize Winner

Pilar Amat Rodrigo

9/7/2018

Contents

A VISUAL HISTORY OF NOBEL PRIZE WINNERS

1. Load the required libraries and the Nobel Prize dataset.

Data:https://ckan.oppnadata.se/dataset/nobel-prizes/resource/cafde48c-586d-4731-95f8-2e91091222d9

```
# Loading in required libraries
library(tidyverse)
library(gdata)
library(readxl)
```

2. Count up the Nobel Prizes. Also, split by sex and birth_country.

```
# Reading in the Nobel Prize data
  nobel <- read_csv(paste0("~/Documentos/DataCamp/RDataCamp/Proyectos DataCamp",
                           "/A Visual History of Nobel Prize Winners/datasets/nobel.csv"))
```

```
# Taking a look at the first couple of winners

head(nobel)
```

```
## # A tibble: 6 x 18
##    year category  prize motivation   prize_share laureate_id laureate_type
##   <int> <chr>     <chr> <chr>        <chr>             <int> <chr>
## 1  1901 Chemistry The ~ "\"in recog~ 1/1                 160 Individual
## 2  1901 Literatu~ The ~ "\"in speci~ 1/1                 569 Individual
## 3  1901 Medicine  The ~ "\"for his ~ 1/1                 293 Individual
## 4  1901 Peace     The ~ <NA>         1/2                 462 Individual
## 5  1901 Peace     The ~ <NA>         1/2                 463 Individual
## 6  1901 Physics   The ~ "\"in recog~ 1/1                   1 Individual
## # ... with 11 more variables: full_name <chr>, birth_date <date>,
## #   birth_city <chr>, birth_country <chr>, sex <chr>,
## #   organization_name <chr>, organization_city <chr>,
## #   organization_country <chr>, death_date <date>, death_city <chr>,
## #   death_country <chr>
```

```
tail(nobel)
```

```
## # A tibble: 6 x 18
##    year category  prize motivation   prize_share laureate_id laureate_type
##   <int> <chr>     <chr> <chr>        <chr>             <int> <chr>
## 1  2016 Literatu~ The ~ "\"for havi~ 1/1                 937 Individual
## 2  2016 Medicine  The ~ "\"for his ~ 1/1                 927 Individual
## 3  2016 Peace     The ~ "\"for his ~ 1/1                 934 Individual
## 4  2016 Physics   The ~ "\"for theo~ 1/2                 928 Individual
## 5  2016 Physics   The ~ "\"for theo~ 1/4                 929 Individual
## 6  2016 Physics   The ~ "\"for theo~ 1/4                 930 Individual
## # ... with 11 more variables: full_name <chr>, birth_date <date>,
## #   birth_city <chr>, birth_country <chr>, sex <chr>,
## #   organization_name <chr>, organization_city <chr>,
## #   organization_country <chr>, death_date <date>, death_city <chr>,
## #   death_country <chr>
```

```r
#this step is not necessary but it could have been interesting so I will keep it.
#nobel$year=as.integer(nobel$year)
```

```r
colnames(nobel)
```

```
##  [1] "year"                 "category"             "prize"
##  [4] "motivation"           "prize_share"          "laureate_id"
##  [7] "laureate_type"        "full_name"            "birth_date"
## [10] "birth_city"           "birth_country"        "sex"
## [13] "organization_name"    "organization_city"    "organization_country"
## [16] "death_date"           "death_city"           "death_country"
```

```r
#filter years from 1902-2016 and show per sex
nobel %>%
  filter(year>=1902 & year<=2016) %>%
  group_by(sex) %>%
  summarise(n=n())
```

```
## # A tibble: 3 x 2
##   sex        n
##   <chr>  <int>
## 1 Female    49
## 2 Male     830
## 3 <NA>      26
```

```r
# Counting the number of prizes won by different nationalities.

nobel %>%
  filter(year>=1902 & year<=2016) %>%
  group_by(birth_country) %>%
  summarise(count=n()) %>%
  arrange(desc(count))
```

```
## # A tibble: 122 x 2
##    birth_country            count
##    <chr>                    <int>
##  1 United States of America   259
##  2 United Kingdom              85
##  3 Germany                     61
##  4 France                      49
##  5 Sweden                      29
##  6 <NA>                        26
##  7 Japan                       24
##  8 Canada                      18
##  9 Italy                       17
## 10 Netherlands                 17
## # ... with 112 more rows
```

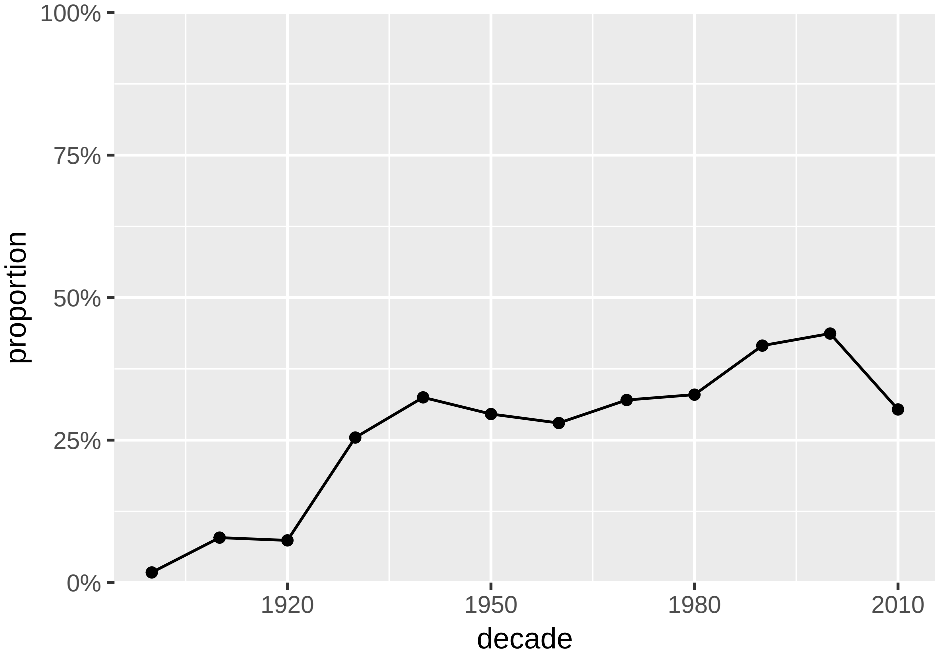
3. Calculate the proportion of USA born winners per decade starting from the nobel dataset and put the result into prop_usa_winners.

```
prop_usa_winners<- nobel %>%
                    mutate(usa_born_winners=(birth_country=="United States of America")) %>%
                    mutate(decade=(year-year%%10)) %>%
                    group_by(decade) %>%
                    summarise(proportion=mean(usa_born_winners, na.rm = TRUE))
prop_usa_winners
```

```
## # A tibble: 12 x 2
##    decade proportion
##     <dbl>      <dbl>
##  1   1900     0.0179
##  2   1910     0.0789
##  3   1920     0.0741
##  4   1930     0.255
##  5   1940     0.325
##  6   1950     0.296
##  7   1960     0.28
##  8   1970     0.320
##  9   1980     0.330
## 10   1990     0.416
## 11   2000     0.437
## 12   2010     0.304
```

4. Plot the proportion of USA born winners per decade.

```r
ggplot(prop_usa_winners, aes(x=decade, y=proportion))+
  geom_line()+
  geom_point()+
  scale_y_continuous(labels = scales::percent, limits = 0:1, expand = c(0,0))
```

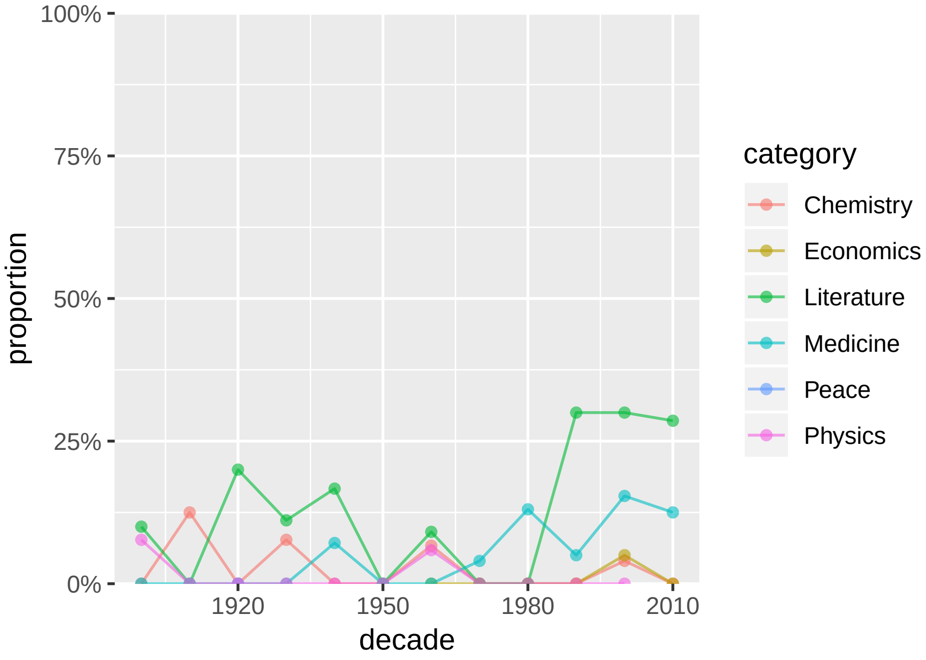


5. Plot the proportion of female laureates by decade split by prize category.

```r
prop_female_winners <- nobel %>%
                        mutate(female_winner=(sex=="Female")) %>%
                        mutate(decade=(year-year%%10)) %>%
                        group_by(decade, category) %>%
                        summarise(proportion=mean(female_winner))
prop_female_winners <- prop_female_winners[-nrow(prop_female_winners), ]
prop_female_winners
```

```
## # A tibble: 65 x 3
## # Groups:   decade [12]
##    decade category   proportion
##     <dbl> <chr>           <dbl>
##  1   1900 Chemistry      0
##  2   1900 Literature     0.1
##  3   1900 Medicine       0
##  4   1900 Peace         NA
##  5   1900 Physics        0.0769
##  6   1910 Chemistry      0.125
##  7   1910 Literature     0
##  8   1910 Medicine       0
##  9   1910 Peace         NA
## 10   1910 Physics        0
## # ... with 55 more rows
```

```
ggplot(prop_female_winners, aes(x=decade, y=proportion, color=category))+
  geom_line(alpha = 0.6)+
  geom_point(alpha = 0.6)+
  scale_y_continuous(labels=scales::percent, limits = 0:1, expand=c(0,0))
```



6. Extract and display the row showing the first woman to win a Nobel Prize.

```
nobel %>%
  filter(sex=="Female") %>%
  top_n(1,desc(year))
```

```
## # A tibble: 1 x 18
##    year category prize  motivation   prize_share laureate_id laureate_type
##   <int> <chr>    <chr>  <chr>        <chr>             <int> <chr>
## 1  1903 Physics  The N~ "\"in recog~ 1/4                   6 Individual
## # ... with 11 more variables: full_name <chr>, birth_date <date>,
## #   birth_city <chr>, birth_country <chr>, sex <chr>,
## #   organization_name <chr>, organization_city <chr>,
## #   organization_country <chr>, death_date <date>, death_city <chr>,
## #   death_country <chr>
```

7. Extract and display the names of repeat Nobel Prize winners.

```
nobel %>%
  #mutate(complete_name= paste(firstname, surname)) %>%
  group_by(full_name) %>%
  summarise(count=n()) %>%
  arrange(desc(count))
```

```
## # A tibble: 904 x 2
##    full_name                                                         count
##    <chr>                                                             <int>
##  1 Comité international de la Croix Rouge (International Committee ~     3
##  2 Frederick Sanger                                                      2
##  3 John Bardeen                                                          2
##  4 Linus Carl Pauling                                                    2
##  5 Marie Curie, née Sklodowska                                           2
##  6 Office of the United Nations High Commissioner for Refugees (UNH~     2
##  7 Aage Niels Bohr                                                       1
##  8 Aaron Ciechanover                                                     1
##  9 Aaron Klug                                                            1
## 10 Abdus Salam                                                           1
## # ... with 894 more rows
```

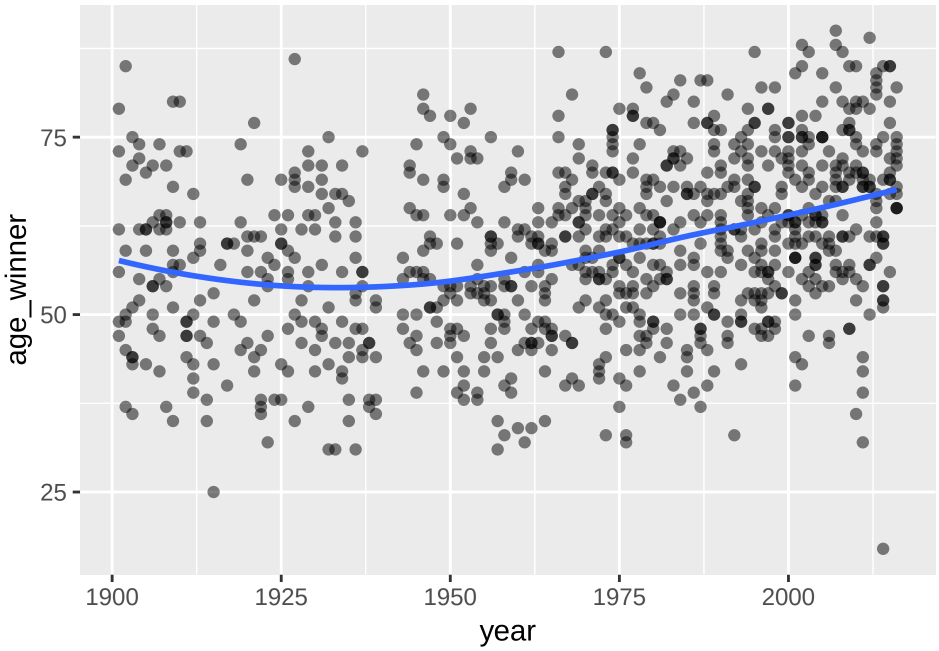
8. Calculate and plot the age of each winner when they won their Nobel Prize

```
library(lubridate)
nobel$born<-as.Date(nobel$birth_date)
head(nobel)
```

```
## # A tibble: 6 x 19
##    year category  prize motivation   prize_share laureate_id laureate_type
##   <int> <chr>     <chr> <chr>        <chr>             <int> <chr>
## 1  1901 Chemistry The ~ "\"in recog~ 1/1                 160 Individual
## 2  1901 Literatu~ The ~ "\"in speci~ 1/1                 569 Individual
## 3  1901 Medicine  The ~ "\"for his ~ 1/1                 293 Individual
## 4  1901 Peace     The ~ <NA>         1/2                 462 Individual
## 5  1901 Peace     The ~ <NA>         1/2                 463 Individual
## 6  1901 Physics   The ~ "\"in recog~ 1/1                   1 Individual
## # ... with 12 more variables: full_name <chr>, birth_date <date>,
## #   birth_city <chr>, birth_country <chr>, sex <chr>,
## #   organization_name <chr>, organization_city <chr>,
## #   organization_country <chr>, death_date <date>, death_city <chr>,
## #   death_country <chr>, born <date>
```

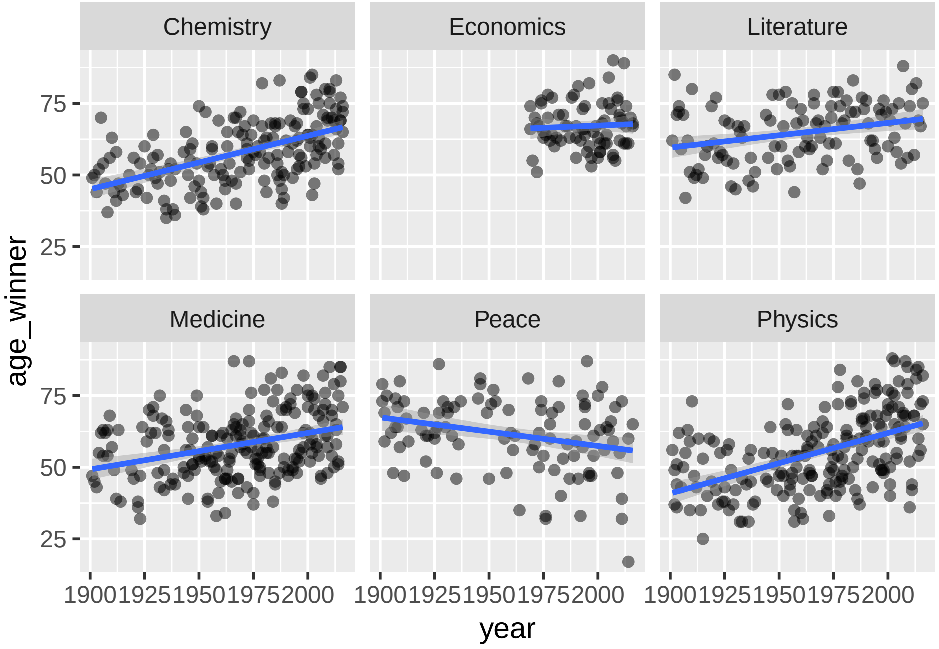
```
nobel_age<-nobel %>% mutate(age_winner=(year-year(birth_date)))

ggplot(nobel_age, aes(x= year, y= age_winner))+geom_point(alpha=0.5)+geom_smooth(se=FALSE)
```



9. Plot how old winners are within the different price categories.

```
ggplot(nobel_age, aes(x= year, y= age_winner))+geom_point(alpha=0.5)+geom_smooth(method=glm)+facet_wrap
```



10. Pick out the rows of the oldest and the youngest winner of a Nobel Prize.

```
# The oldest winner of a Nobel Prize as of 2016
nobel_age %>% top_n(1, age_winner)
```

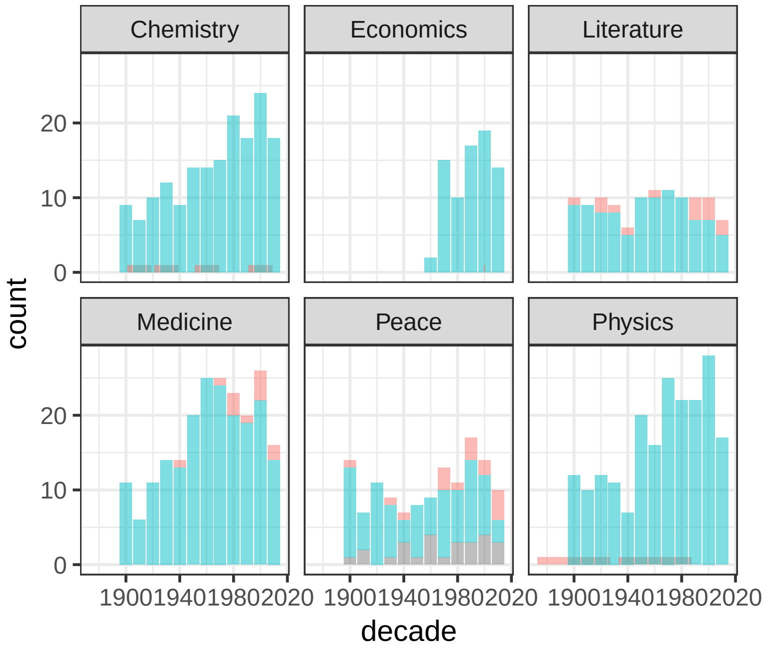
```
## # A tibble: 1 x 20
##    year category  prize  motivation  prize_share laureate_id laureate_type
##   <int> <chr>     <chr>  <chr>       <chr>             <int> <chr>
## 1  2007 Economics The S~ "\"for hav~ 1/3                 820 Individual
## # ... with 13 more variables: full_name <chr>, birth_date <date>,
## #   birth_city <chr>, birth_country <chr>, sex <chr>,
## #   organization_name <chr>, organization_city <chr>,
## #   organization_country <chr>, death_date <date>, death_city <chr>,
## #   death_country <chr>, born <date>, age_winner <dbl>
```

```
# The youngest winner of a Nobel Prize as of 2016
nobel_age %>% top_n(-1, age_winner)
```

```
## # A tibble: 1 x 20
##    year category prize  motivation   prize_share laureate_id laureate_type
##   <int> <chr>    <chr>  <chr>        <chr>             <int> <chr>
## 1  2014 Peace    The N~ "\"for thei~ 1/2                 914 Individual
## # ... with 13 more variables: full_name <chr>, birth_date <date>,
## #   birth_city <chr>, birth_country <chr>, sex <chr>,
## #   organization_name <chr>, organization_city <chr>,
## #   organization_country <chr>, death_date <date>, death_city <chr>,
## #   death_country <chr>, born <date>, age_winner <dbl>
```

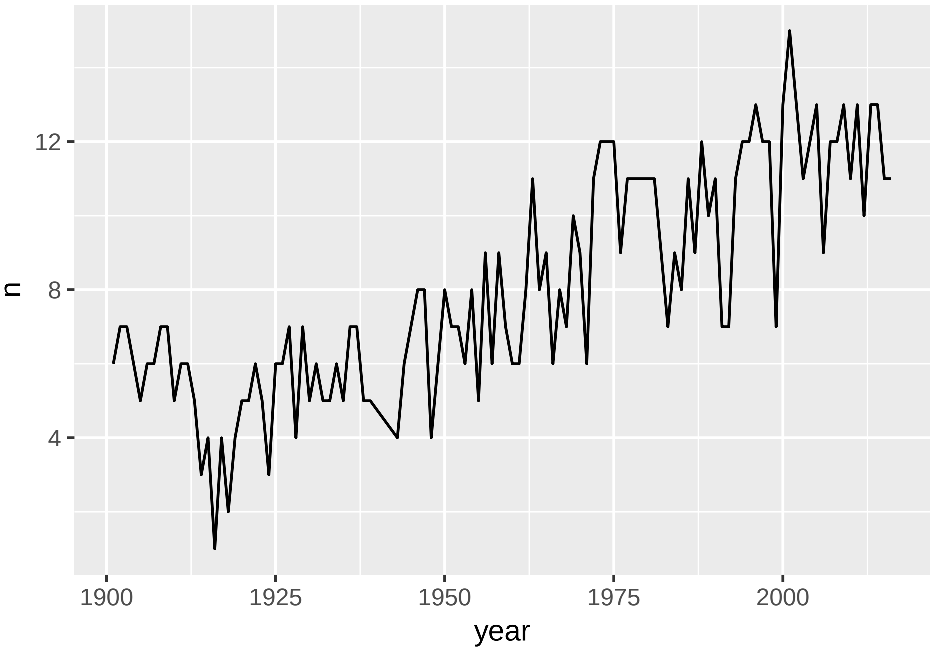
11. Take a look to the proportion of laureates by sex.

```
nobel2<-nobel %>% mutate(decade=(year-year%%10))
ggplot(nobel2, aes(x=decade, fill=sex))+ geom_bar(alpha=0.5)+ facet_wrap(~category)+ theme_bw()
```

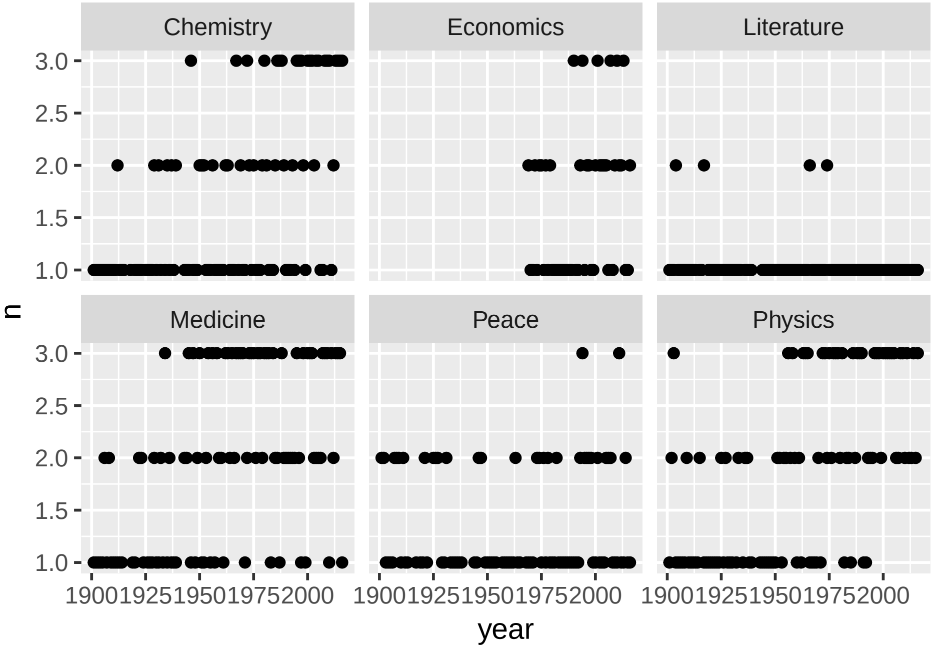


12. How did laureates changed over time? To me it seems that there is more and more cooperation, team of scientists.

```
nobelyearcount<- nobel %>% group_by(year) %>% summarise(n=n())
ggplot(nobelyearcount, aes(x=year,y=n))+geom_line()
```



```
nobelyearcount<- nobel %>% group_by(year, category) %>% summarise(n=n())
ggplot(nobelyearcount, aes(x=year,y=n))+geom_point()+facet_wrap(~category)
```


Chemistry
Economics
Literature
Medicine
Peace
Physics
n
year