Premi Nobel: Esplorazione nello spazio e nel tempo

Mariachiara Manoccio, Giovanni Tripicchio

2024-02-20

Premi Nobel: Esplorazione nello spazio e nel tempo

In questo report abbiamo voluto effettuare un' analisi esplorativa sui vincitori dei premi Nobel nel corso degli anni. Il premio Nobel fu istituito dall'ingegnere svedese Alfred Nobel (1833-1896) inventore della dinamite che, amareggiato dagli effetti distruttivi della sua scoperta, decise che in futuro avrebbe voluto essere ricordato per azioni più nobili. Così decise di realizzare un premio, da assegnare "agli uomini capaci di creare cose belle e compiere azioni grandiose". Per questa ragione, in Svezia, a partire dal 1901, questo premio viene consegnato ogni anno il 10 Dicembre. Mostreremo come questo premio è cambiato negli anni.

About dataset

Il dataset è open access e scaricabile dal web all' indirizzo https://public.opendatasoft.com/explore/dataset/nobel-prize-laureates, le cui fonti provengono dal sito ufficiale della Fondazione Nobel http://www.nobelprize.org/nobel_organizations/nobelmedia/nobelprize_org/developer/. I dati sono stati analizzati tramite software R utilizzando l'interfaccia R-Studio.

```
# Importazione dataset
library(readxl)
load("~/Scrivania/Master_DataScience/R_giordano/Progetto_R/progetto.RData")
data_df <- as.data.frame(data)
names(data_df)
```

```
##  [1] "Id"                  "Firstname"           "Surname"
##  [4] "Born"                "Died"                "Born.country"
##  [7] "Born.country.code"   "Born.city"           "Died.country"
## [10] "Died.country.code"   "Died.city"           "Gender"
## [13] "Year"                "Category"            "Overall.motivation"
## [16] "Motivation"          "Organization.name"   "Organization.city"
## [19] "Organization.country" "Geo.Shape"          "Geo.Point.2D"
```

```
#str(data_df)
```

Il dataset è composto da 1000 righe e 21 colonne di tipo chr (ad eccezione dell' enumerazione di tipo int), contenenti informazioni anagrafiche, geografiche e sulla vittoria dei premi nobel dal 1901 al 2023. Siamo curiosi di conoscere l' età dei vincitori, quindi aggiungiamo una nuova colonna "Age" ottenuta dalla differenza tra la colonna "Year_Complete" (creata aggiungendo la stringa "-12-10"alla colonna "Year", anno di vincita del Nobel), e la colonna "Born" che contiene la data di nascita in formato YYYY-MM-DD, entrambe convertite in formato data. Successivamente trasformiamo i NULL in 0 per non incorrere in eventuali errori durante le elaborazioni successive. Infine applichiamo un filtro al dataset per filtrare tutti i valori della colonna "Age" diversi da zero derivanti da operazioni con colonne NULL da non tenere in considerazione. In questo modo riduciamo il dataset di partenza a 956 campi di cui conosciamo l' età. Infine, per completare il dataset aggiungiamo una nuova colonna "Name_surname" che unisce le due colonne "Firstname" e "Surname", che utilizzeremo per valutare il numero di vincitori distinti e per verificare se esistono vincitori pluripremiati.

```r
#trasformo in una nuova colonna la colonna "Year" per aggiungere il mese e l' anno
data_df$Year_Complete <- paste(data_df$Year, "-12-10", sep = "")
#per fare la differenza e ricavare l' età si converte prima in oggetti di tipo data.
library(lubridate)
data_df$Born <- ymd(data_df$Born)
data_df$Year_Complete <- ymd(data_df$Year_Complete)
# Calcola la differenza tra le due date
data_df$Age <- data_df$Year_Complete - data_df$Born
# Estrai l'età dall'intervallo di date ottenuto
data_df$Age <- as.integer(data_df$Age / dyears(1))
#conversione null
data_df[is.na(data_df)] <- 0
#rimuoviamo i valori nulli dell' età, ricavanti da informzioni NULL del dataset
data_df_age <- subset(data_df, Age != 0)
#Crea colonna Nome+Cognome
data_df$Name_surname <- paste(data_df$Firstname, data_df$Surname, sep = " ")
```

Esplorazione dei dati

Categorie, età, genere

In questa sezione osserviamo le numeriche relative all' anagrafica dei vincitori in relazione alle categorie del premio. Utilizzeremo le librerie "dplyr" per l' analisi dei dati e "ggplot2" per la rappresentazione grafica. Per prima cosa vediamo quante sono le categorie e il numero di vincitori che si sono susseguiti negli anni per categorie, come mostrato nella tabella 1. Il numero più alto di vincitori si osserva per le materie scientifiche, infatti medicina, chimica e fisica si aggiudicano complessivamente circa il 75% del totale dei vincitori). Il numero di vincitori più basso è stato osservato per il Nobel in Economia, che, come vedremo in seguito, è stato istituito solo in tempi più recenti.

```r
library(dplyr)
library(bookdown)
#vincitori per categorie
winnercat <- data_df %>%
  group_by(Category) %>%
  summarise(Count = n())%>%
  mutate(Percent_Categories = (Count / nrow(data_df)) * 100)
winnercat<-rename(winnercat, "%Categories"=Percent_Categories)
knitr::kable(winnercat, booktabs = TRUE,
  caption = 'Numero e percentuali di vincitori per categorie'
)
```

Table 1: Numero e percentuali di vincitori per categorie

Category	Count	%Categories
Chemistry	194	19.4
Economics	93	9.3
Literature	120	12.0
Medicine	227	22.7
Peace	141	14.1
Physics	225	22.5

Nella tabella 2 osserviamo il numero di vincitori, vincitrici e l' età media. La differenza di genere tra i vincitori è molto evidente, (6.4% per le femmine vs 90.1% per i maschi), mentre l'età media si differenzia di due anni. Si osserva anche una terza categoria che rappresenta importanti organizzazioni internazionali.

```r
#  valori distinti dalla colonna 'Name_surname'
elenco_distinti <- data_df[!duplicated(data_df$Name_surname), ]
#genere
genere <- elenco_distinti %>%
  group_by(Gender) %>%
  summarise(Count = n(),
            Mean_age = mean(Age)) %>%
  mutate(Percent_Gender = (Count / nrow(data_df)) * 100)
genere$Mean_age<-round(genere$Mean_age,digits = 1)
genere<-rename(genere, "%Gender"=Percent_Gender)
 knitr::kable(genere, booktabs = TRUE,
  caption = 'Numero di vincitori, vincitrici ed età media')
```

Table 2: Numero di vincitori, vincitrici ed età media

Gender	Count	Mean_age	%Gender
female	64	57.0	6.4
male	901	59.3	90.1
org	27	0.1	2.7

Il grafico in Figura 1 mostra la distribuzione del numero e genere di vincitori nel corso degli anni per categoria. Ogni categoria ha da un minimo di uno fino a un massimo di 3 vincitori per anno. Si osserva che la densità di punti è molto più alta oggi rispetto ai primi anni del 1900. Questo perchè oggi molti premi sono condivisi, ci sono molti più vincitori. La classe "organizzazioni" risulta vincitrice della categoria "Pace". Inoltre, si può vedere come i premi per la categoria Economia siano stati consegnati a partire dal 1969, anno di istituzione del premio. Possiamo anche constatare come il numero di vincitrici donne sia aumentato molto negli anni, che simboleggia che l'impatto scientifico e societario della donne è sempre più riconosciuto.

```r
# Creazione del grafico  del numero di maschi, femmine e altri per categoria e anno
gender_mean<-data_df%>%group_by(Category,Gender,Year_Complete)%>%summarize(Count=n())
gender_mean$Dim_gen <- ifelse(gender_mean$Gender == "male", 1,
                        ifelse(gender_mean$Gender == "org", 2, 3.5))

library(ggplot2)
grafico <- ggplot(gender_mean, aes(x = Year_Complete,
                           y =Count, color = Gender, size=Dim_gen, group = Gender)) +
  geom_point(alpha=0.8) +
  facet_wrap( ~ Category) +  # Suddivisione per categoria
  labs(title = "Genere dei vincitori negli anni per categorie",
       x = "Anno", y = "Numero Vincitori", color = "Genere",
       caption = 'Figura 1. Numero di vincitori per categoria negli anni, evidenziando
                il gender vincitore.
                 I punti a destra del grafico rappresentano le dimensioni dei vari
                 generi: dim_gem=1 e colore verde rappresenta il genere "uomo",
                 colore blue e dimensione 2.5 genere "organizzazioni" e
                dimensione 3.5 e colore rosa genere "donna"') +
  theme_minimal()
grafico <- grafico + scale_y_continuous(expand = c(0.99, 1.01))

#età massima e minima
max_min_age <- data_df_age %>%
    summarize(max_age = max(Age), min_age = min(Age) )
```

```
print(grafico)
```

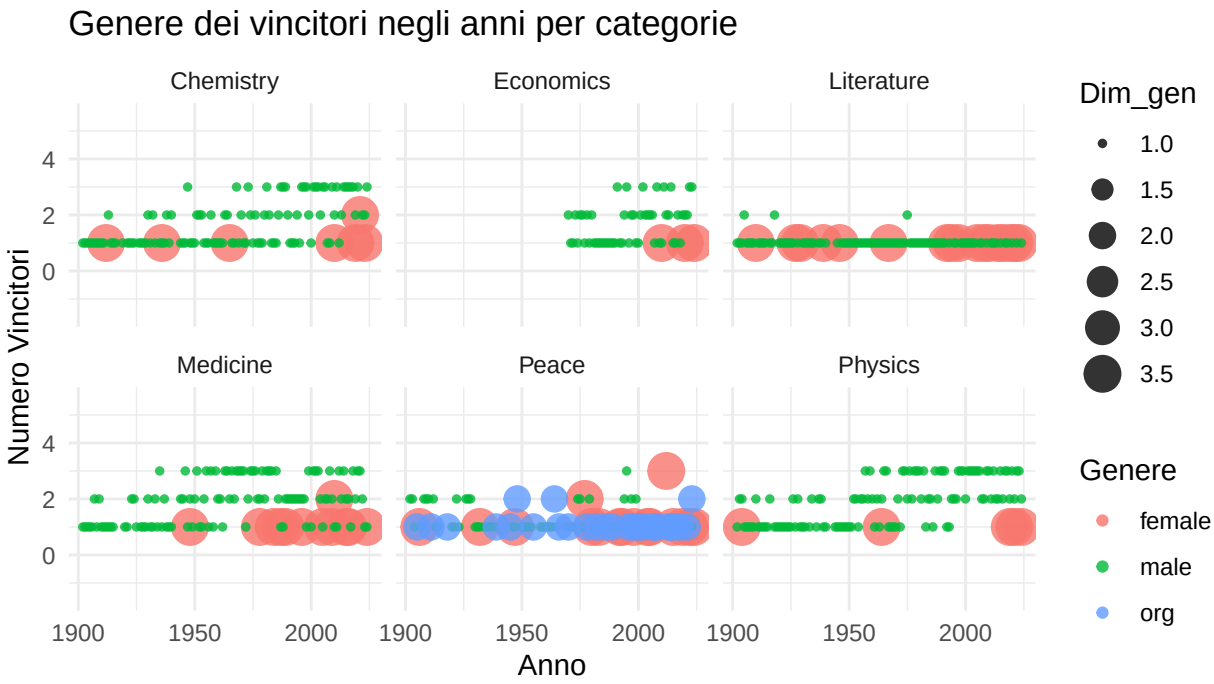

Figura 1. Numero di vincitori per categoria negli anni, evidenziando il gender vincitore.
I punti a destra del grafico rappresentano le dimensioni dei vari generi: dim_gem=1 e colore verde rappresenta il genere "uomo", colore blue e dimensione 2.5 genere "organizzazioni" e dimensione 3.5 e colore rosa genere "donna"

Il grafico in Figura 2 mostra un boxplot relativo alla distribuzione dell'età per categorie nel range interquartile. L'età media si colloca in un range tra i 55 e i 75 anni. Il boxplot mostra anche la presenza di outliers, che sono rappresentati dai punti più estremi. La più "giovane" è un' organizzazione internazionale (Nobel per la Pace a 3 anni dalla sua fondazione), l' Agenzia delle Nazioni Unite per i rifugiati (UNHCR). La persona più giovane ad aggiudicarsi il premio fu invece la 17enne Malala Yousafzai nel 2014 per la categoria Pace, per la lotta contro la sopraffazione dei bambini e il diritto di questi all'istruzione. Entrambi i giovani premiati, costituiscono due outliers del dataset esaminato. Il vincitore più anziano, ben 97 anni, fu il fisico Arthur Askin, che vinse il premio per la Fisica nel 2018 insieme ad altri due candidati, per le "invenzioni rivoluzionarie nel campo della fisica dei laser".

```
library(ggplot2)
#distribuzione età vincitori per categorie + boxplot eliminati i NA
attach(data_df_age)

get_box_stats <- function(y, upper_limit = max(data_df_age$Age) * 1.15) {
  return(data.frame(
    y = 0.95 * upper_limit,
    label = paste(
      "Count =", length(y), "\n",
      "Mean =", round(mean(y), 2), "\n",
      "Median =", round(median(y), 2), "\n",
      "IQR =", round(IQR(y), 2), "\n"
    )
  ))
}
```

```r
boxplot <- ggplot(data_df_age, aes(x = Category, y = Age, fill = Category)) +
  geom_boxplot(show.legend = F, alpha=0.5) +
  geom_point(show.legend = F, position = position_jitter(width = 0.2), color = "lightblue",
             size = 1.5) +
  labs(title = "Boxplot e distribuzione delle età per categoria", x = "Categoria",
       y = "Età",
        caption = 'Figura 2. Boxplot dell\'età dei vincitori dei premi Nobel per le sei
        categorie. La larghezza dei singoli boxplot indica il range interquartile, IQR
        (25 e 75 percentile). Le linee verticali catturano il 99% delle singole
        distribuzioni normali. I punti estremi delle linee verticali,
       rappresentano gli outliers.') +
   stat_summary(fun.data = get_box_stats, geom = "text", hjust = 0.4, vjust = 0.1,size =3) +
    ylim(0, 140)+
  theme_minimal()
print(boxplot)
```

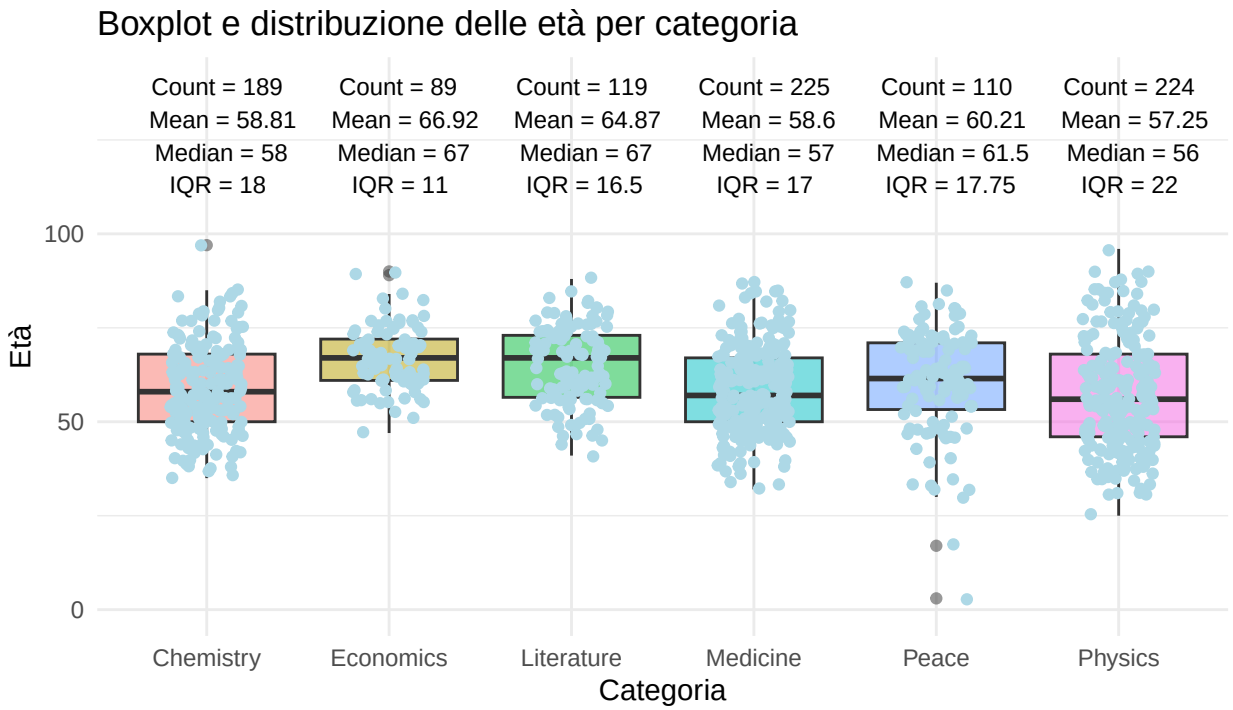

Figura 2. Boxplot dell'età dei vincitori dei premi Nobel per le sei categorie. La larghezza dei singoli boxplot indica il range interquartile, IQR (25 e 75 percentile). Le linee verticali catturano il 99% delle singole distribuzioni normali. I punti estremi delle linee verticali, rappresentano gli outliers.

E' inoltre accaduto ad oggi a ben sette fortunati vincitori di ricevere più volte il premio. Di questi ricordiamo: John Bardeen per la Fisica per lo studio sui transistor e per la superconduttività; Frederick Sanger, due volte per la chimica; Linus Carl Pauling prima la chimica e poi per la pace per il suo lavoro di promozione del disarmo nucleare. Anche organizzazioni come la Croce Rossa e l'UNHCR sono state pluripremiate negli anni. Anche la prima donna insignita del premio Nobel Marie Curie, si aggiudicò due premi: il primo per la Fisica nel 1903, per lo studio dei fenomeni di radiazione e nel 1911 per la Chimica, per la scoperta degli elementi radio e polonio. Inoltre, da un maggiore approfondimento sul dataset è risultato che oltre alla duplice vittoria della scienziata, anche suo marito Pierre Curie fu vincitore del premio del 1903 insieme a lei. Inoltre, anche la loro figlia Irène Joliot-Curie insieme al marito Frédéric Joliot-Curie nel 1935 ne vinsero un altro per la chimica. Insomma, una famiglia da Nobel!

```r
#plurivincitori
plurivincitori_ds <- data_df[duplicated(data_df$Name_surname) |
                               duplicated(data_df$Name_surname, fromLast = TRUE), ]
plurivincitori <- plurivincitori_ds %>%
  group_by(Name_surname) %>%
  summarise(Count = n())

#plurivincitrici
plurivincitrici_ds <- filter(plurivincitori_ds, Gender == "female")
plurivincitrici<- select(plurivincitrici_ds, Name_surname, Category, Year)

#curie_fam esplorazione
curie_fam <-data_df %>%
  filter(grepl("Curie", Name_surname, ignore.case = TRUE) |
           grepl("Joliot", Name_surname, ignore.case = TRUE))
```

Come è invece cambiata l' età negli anni? Il grafico seguente, Figura 3, mostra il trend dell'età dei vincitori nel corso degli anni. Vediamo che in passato le persone avevano circa 55 anni quando ricevevano il premio, ma oggi la media è più alta e più vicina ai 65 anni. Il caso opposto è osservato per i premi nobel per la Pace, per i quali è stato osservato un decremento dell'età media negli ultimi anni.

```r
#distribuzione età
grafico1 <- ggplot(data_df_age, aes(x = Year_Complete, y = Age)) +
  geom_point(size=2, color = "blue") +
  facet_wrap(~ Category) +
  labs(title = "Distribuzione età vincitori negli anni per categorie",
       x = "Anno", y = "Età Vincitori",
       caption = 'Figura 3. Trend dell\'età dei vincitori per ogni categoria negli anni.
       L\'andamento dell\'età risulta essere crescente per le categorie di chimica,
       fisica, letteratura e medicina, costante per economia e descrente per la
       categoria pace.') +
  theme_minimal() +
  stat_smooth(method = "lm", se = TRUE, color = "yellow") +
  coord_cartesian(ylim = c(2, 96))
print(grafico1)
```

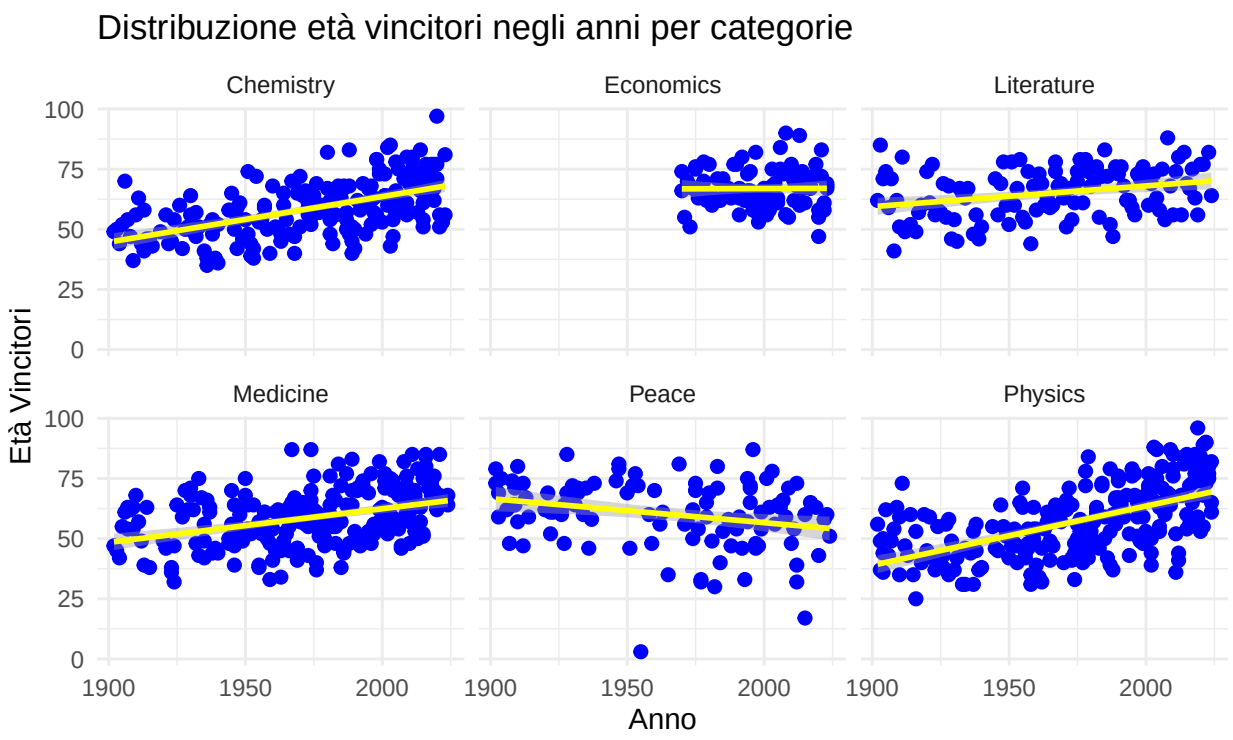

Figura 3. Trend dell'età dei vincitori per ogni categoria negli anni. L'andamento dell'età risulta essere crescente per le categorie di chimica, fisica, letteratura e medicina, costante per economia e descrente per la categoria pace.

Vincintori nel mondo

Per concludere vogliamo vedere come si distribuiscono geograficamente le provenienze dei vincitori nel panorama mondiale. Per questa analisi è stato utilizzato il pacchetto choroplethr. Il maggior numero di vincitori è distribuito nel Nord Europa, negli USA e in Russia (o meglio, nell'area ex URSS), seguito da Nord America, Europa Centro e Asia.

```
library(choroplethr)
library(countrycode)
data_1<-dplyr::select(data,Born.country.code, Category)
plotdata<-count(data_1,Born.country.code)
plotdata<-rename(plotdata,region=Born.country.code, value=n)
plotdata<-arrange(plotdata,desc(value))
plotdata$region<-countrycode(plotdata$region, "iso2c", "country.name")
plotdata<-mutate(plotdata, region = tolower(region))
# modifica di alcuni campi per problemi di formattazione
plotdata$region[1]="united states of america"
plotdata$region[24]="czech republic"
plotdata$region[39]="bosnia and herzegovina"
plotdata$region[54]="democratic republic of the congo"
plotdata$region[70]="macedonia"
plotdata$region[78]="trinidad and tobago"

country_choropleth(plotdata,
                   num_colors=9) +
  scale_fill_brewer(palette="YlOrRd")+
  labs(title = "Numero di vincitori nel Mondo",
```

```
        caption = "Figura 4. Distribuzione del numero di vincitori di Nobel per paese di nascita",
        fill = "# Vincitori")
```

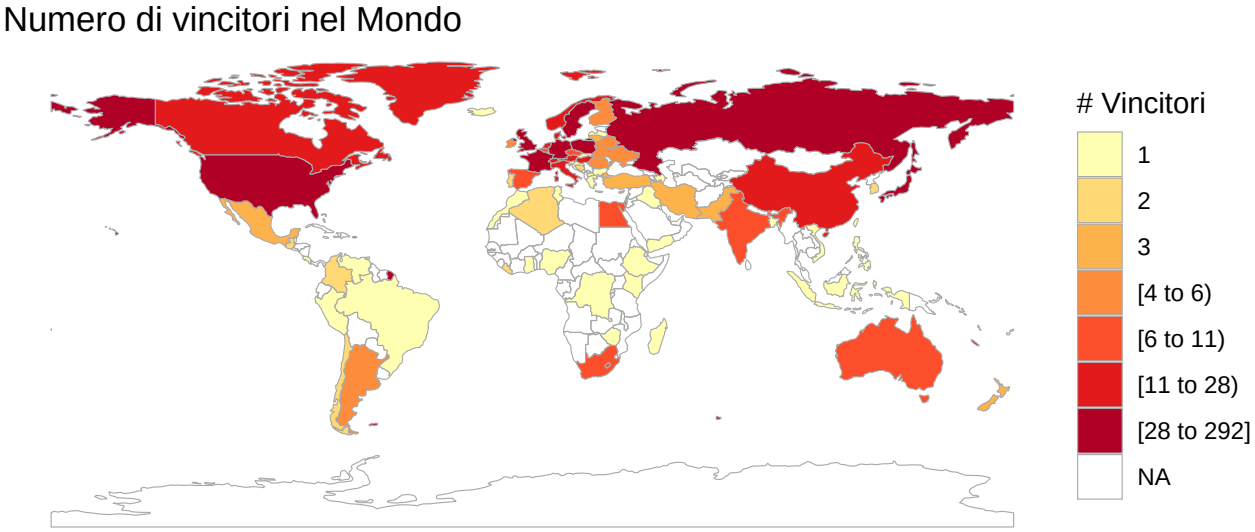

Figura 4. Distribuzione del numero di vincitori di Nobel per paese di nascita

Andando più nel dettaglio, sono state analizzate le differenze tra città di nascita e citta in cui è stato conferito il premio Nobel per la categoria Fisica, per osservare eventuali migrazioni per influenze storico-politiche e qualità dei centri di ricerca nel mondo.

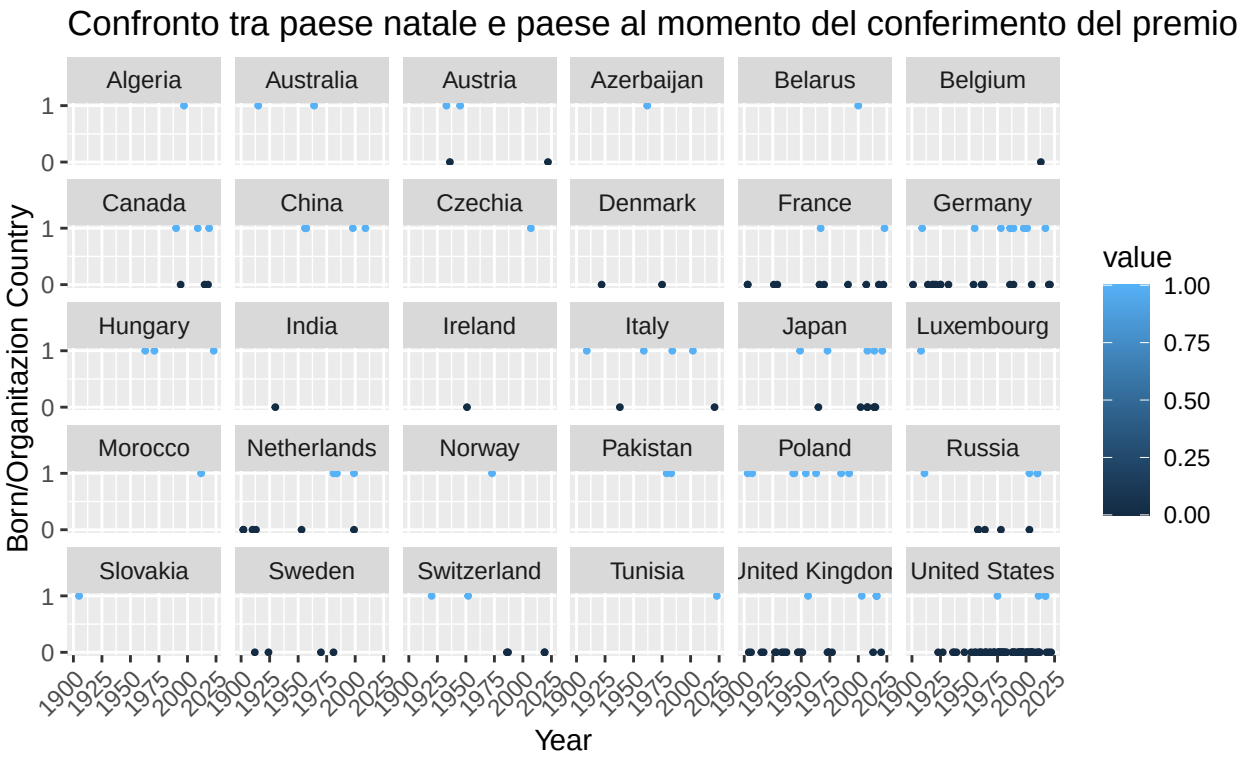

Figura 5. Confronto tra Paese di nascita e Paese al momento del conferimento del premio. La data indicata è reltiva all' anno di nascita del vincitore. Nel grafico e' stato assegnato il valore 1 nel caso in se i due Paesi sono diversi e valore 0 se i due Paesi coincidono

Nel grafico 5 si può osservare che che negli Stati Uniti e nel Regno Unito, per la Maggior parte dei vincitori per la categoria Fisica, i Paesi di nascita e Centro in cui il premio è stato vinto coincidono. In altre realtà

invece, i premi Nobel sono stati assegnati altrove. Questo può darci un'indicazione del fatto che, oltre ad aver un numero maggiore di premi Nobel vinti rispetto agli altri Paesi, la ricerca scientifica ad alti livelli è ben radicata, rispetto a Paesi più poveri o meno sviluppati. La forte ambiguità che si può osservare per i Paesi come la Polonia o la Germania, può essere attribuita invece al forte impatto che la comunità scientifica subì durante le guerre e le persecuzioni e che portarono alla migrazione di molte menti brillanti verso gli USA.

Conclusione

In questa analisi abbiamo esplorato l'universo dei Vincitori dei premi Nobel nella Storia. Abbiamo utilizzato un dataset open source scaricabile dal web e abbiamo effettuato un'analisi dati attraverso il software R-Studio. Nella prima parte, ci siamo concentrati sulle numeriche per categoria, per genere ed età, e su come queste caratteristiche siano cambiate negli anni. Abbiamo visto la forte differenza di genere tra maschi e femmine, anche se negli ultimi anni il numero di vincitrici è sempre più in aumento. Abbiamo inoltre visto che le organizzazioni mondiali sembrano dare un contributo alla Pace nel mondo, tanto da aggiudicarsi numerosi premi. Abbiamo anche osservato come cambia la distribuzione dell'età dei vincitori negli anni, che, ad eccezione della categoria Pace, sembra essere destinato a menti sempre più adulte. Infine abbiamo visto che il premio per l'Economia è stato istutito solo dopo, a partire dagli anni '70. Nella seconda parte abbiamo esplorato geograficamente la provenienza dei vincitori localizzata prevalentemente nel Nord Europa, negli Stati Uniti e nei Paesi ex URSS. Per finire volevamo verificare che il luogo di nascita dei vincitori fosse effettivamete quello di assegnazione del premio e per questa analisi abbiamo esaminato queste caratteristiche per la categoria Fisica, che ci ha condotti ad osservare che la distribuzione di vincitori nel mondo sia stata influenzata negli anni da vari fattori come le guerre e le persecuzioni. Infatti i premi Nobel tanto acclamati quanto spesso contestati, sono lo specchio del progresso scientifico e dell' innovazione culturale a passo con il contesto storico-politico.