



Programming Languages for Data Engineering (AC50002)

R programming Assignment

Yashash Thittamaranahalli Venkatesh (2546871)

PART 1

Load the baby names Tibble. Use dplyr to find out which were the three most popular names given to babies of male sex during the 20th century (1900-1999) in the USA. Use ggplot2 to produce a single graph that shows how the popularity these three names changed over the course of the century. You should submit a one-page document containing your R code and your graph. Include comments in your code explaining what each line does.

Solution

I have done few steps which helps to execute part 1 of the assignment.

First, I installed the required packages and loaded the libraries to perform the operations which is required to filter the top three males name which were top between 1900-1999 and to plot a line graph to show the change of popularity of the top 3 names.

```
# imported libraries and packages -----
#to perform dplyr and ggplot2 we will import library tidyverse
library(tidyverse)
#installing babynames packages to perform tasks on dataset babynames
install.packages("babynames")
install.packages("plotly")
library(plotly)
library(babynames)

# Code to filter top 3 males -----
#Filtering the dataset to consider males of USA born in year between 1900 to 1999
males<-babynames%>%
  filter(sex=="M",year>=1900,year<=1999)
#It helps to displaying the filtered dataset for males
males
#Counts the total occurrences of each each names which is grouping by names
count_males<-males%>%
  group_by(name)%>%
```



```
summarise(total=sum(n))
#It helps to displaying the counts of males
count_males

#Selecting the top three names for males based on total occurrences and it is also arrange in descending
order
top_males<-count_males%>%
  top_n(3,total)%>%
  arrange(desc(total))
#It helps to display the top three names of males
top_males
#Filtering the males dataset to include only the top three names
filtering<-males%>%
  filter(name%in%top_males$name)
#It helps to display the filtered dataset of the top three names
filtering
# code to plot graph using ggplot -----
#Creates a line plot using ggplot to visualize the trends of the top three names over the year and how they
changed
ggplot(filtering, aes(x = year, y = n, color = name)) +
  geom_line() +
  labs(title = "Top three male names of USA and there popularity",
        x = "Year",
        y = "Numbers",
        color = "Names")+
  theme_minimal()
#ggplotly() converts the ggplot object to an interactive plot using plotly
ggplotly()
```



Top three male names of USA and there popularity

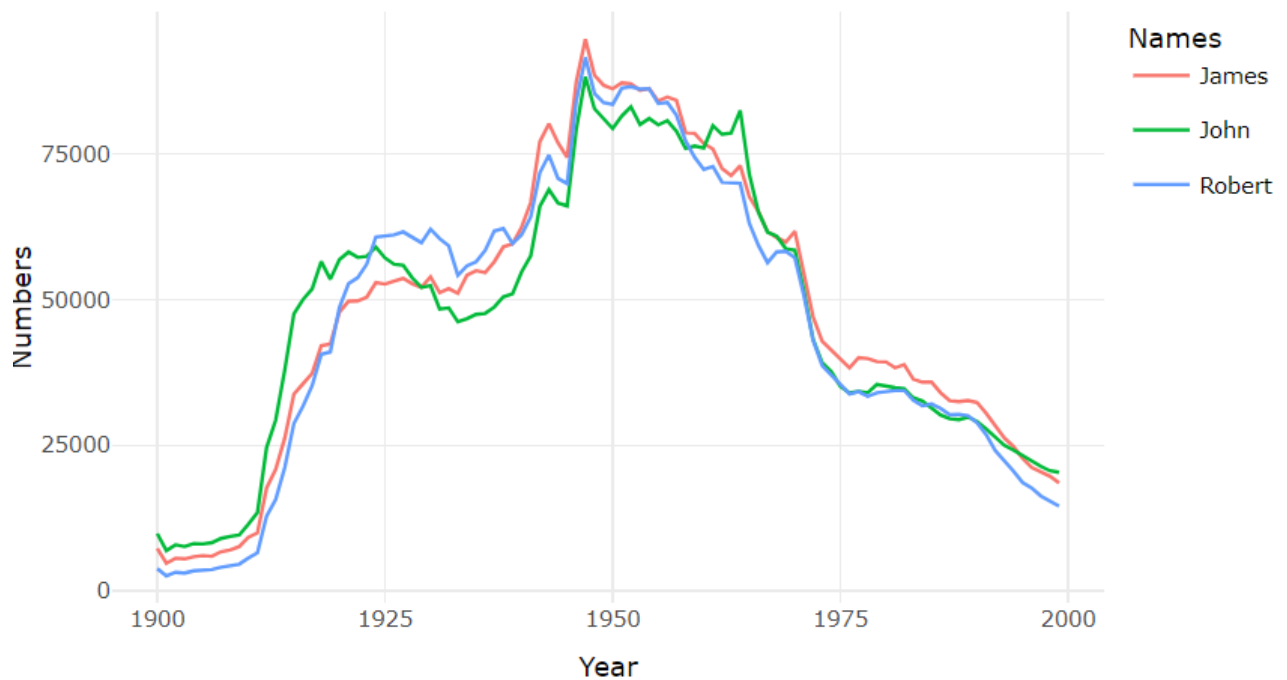


Figure 1 Graph shows how the popularity of these three names changed over the course of century.

The trends in the top three male names in the US from 1900 to 2000 are shown in the graph. Notably, "James" peaked in popularity in 1960 before declining and then rising once more. "John" had a similar trajectory, peaking in 1950, then falling off and then rising again in 1980. "Robert" came in third, declining until 1970 and then rising again. The graph shows how the popularity of names varies over time. For example, James was popular in 1900 and 2000 but not in the 1950s or 1960s. All in all, it provides information about the fluctuating popularity of masculine names in the United States during the previous century.