Question 1

Ramzan Kamoto

2024-06-18

Question 1

For the first part of this question I try to establish the most popular names nationally per decade. I start by creating a function that creates a tibble, which gives me the name with the highest count in every decade. I then use that table to construct a simple bar chart with names in the x axis and years on the y axis.

```
#load relevant data and packages
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
list.files('/Users/ramzankamoto/Documents/Masters/DS_EXAM/23550716/Question1/code', full.names = T, recu

custom_names <- c("Baby_Names", "charts", "population_data","HBO_credits","HBO_titles")

read_rds_files("/Users/ramzankamoto/Documents/Masters/DS_EXAM/23550716/Question1/data", custom_names)
```

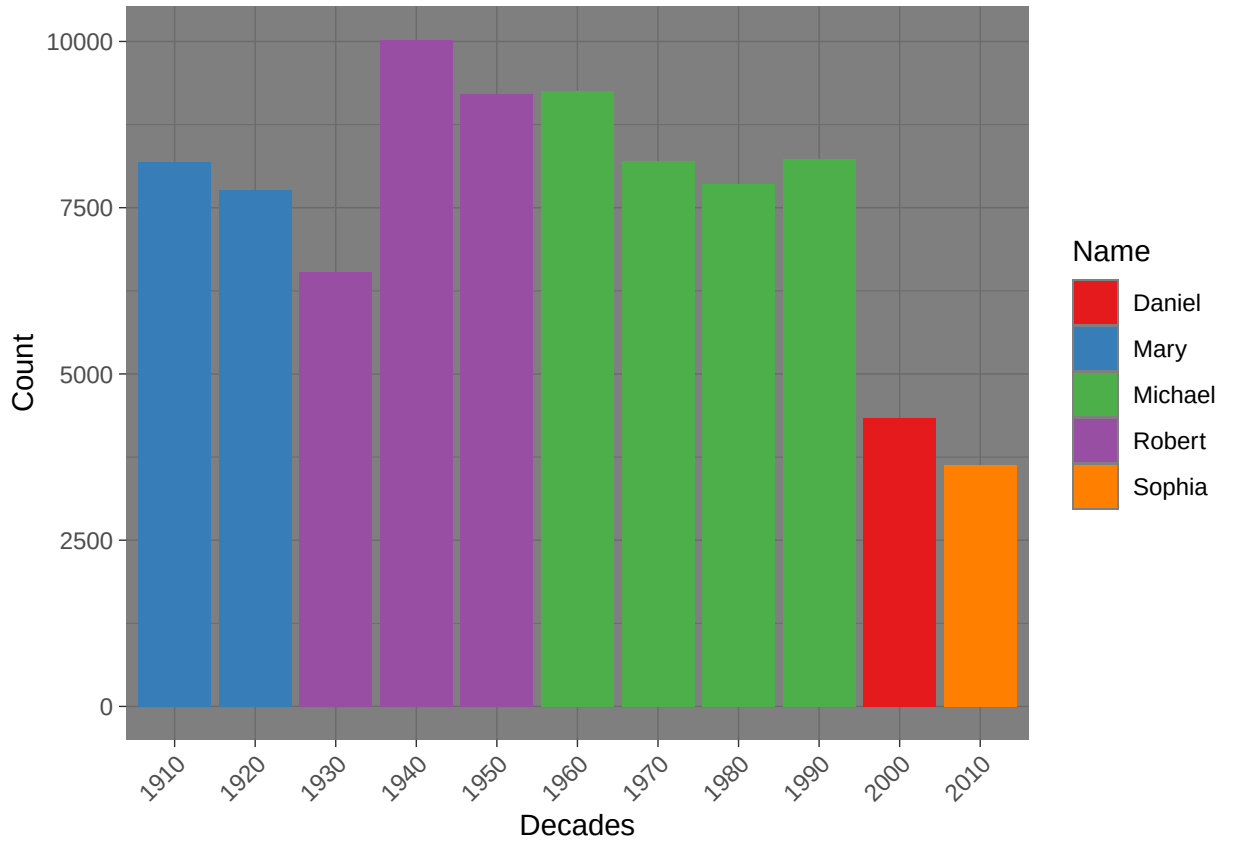
Plot

```
#use highest count function to find the name with the highest count per decade
highest_count_table <- highest_count_per_decade(data_frame = Baby_Names)
```

Admittedly, this is likely an over simplification. It might have been better to take the aggregate of all named Mary for example, rather than simply taking the highest figure.

```
Names_per_decade_plot <- plot_highest_names(highest_count_table)

print(Names_per_decade_plot)
```




```
##This plot still needs to be improved. Check Nicos slides on tidy visualizations.
```

Spearmans correlation

For the next part, we calculate the Spearmans rank correlation. It measures the magnitude and direction of of association between two variabales.

```
#creating a dataframes for top 25 names in 1995 and 1998
top_names <- create_top_names_datasets(Baby_Names)
top_25_males_1995 <- top_names$males_1995
top_25_females_1995 <- top_names$females_1995
top_25_males_1998 <- top_names$males_1998
top_25_females_1998 <- top_names$females_1998
```

Next we conduct the spearman correlation

```
cor.test(top_25_males_1995$Count, top_25_males_1998$Count, method = "spearman")
```

```
##
##  Spearman's rank correlation rho
##
## data:  top_25_males_1995$Count and top_25_males_1998$Count
## S = 0, p-value = 3.196e-07
## alternative hypothesis: true rho is not equal to 0
```

```
## sample estimates:
## rho
##   1
```

```r
cor.test(top_25_males_1998$Count, top_25_females_1998$Count, method = "spearman")
```

```
##
##  Spearman's rank correlation rho
##
## data:  top_25_males_1998$Count and top_25_females_1998$Count
## S = 0, p-value = 3.196e-07
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
## rho
##   1
```

The spearman correlation is suggesting, perfect positive correlation for any combination of top names in 1995 and 1998. There is likely an error in my calculation.