Homework assignment 1

Mikayla Lalli

1. Install the babynames package

```
# Load required libraries
library(babynames)
library(tidyverse)
```

2. How many variables and observations does this package contain?

```
# Inspect the data
data("babynames")
glimpse(babynames)
```

```
Rows: 1,924,665
Columns: 5
$ year <dbl> 1880, 1880, 1880, 1880, 1880, 1880, 1880, 1880, 1880, 1880, 1880,~
$ sex  <chr> "F", "F", "F", "F", "F", "F", "F", "F", "F", "F", "F", "F", "F", ~
$ name <chr> "Mary", "Anna", "Emma", "Elizabeth", "Minnie", "Margaret", "Ida",~
$ n    <int> 7065, 2604, 2003, 1939, 1746, 1578, 1472, 1414, 1320, 1288, 1258,~
$ prop <dbl> 0.07238359, 0.02667896, 0.02052149, 0.01986579, 0.01788843, 0.016~
```

The babynames package contains 1,924,665 observations and 5 variables.

3. Create a data dictionary for each of the variables that includes the variable name, data type, and a description.

Variable name	Type	Description
name	Character	Name of person at birth
year	Numeric (double)	Birth year
sex	Character	Sex of person at birth
n	Numeric (integer)	Number of applicants of a given sex with a given name born in a given year
prop	Numeric (double)	Proportion of people of a given sex with a given name born in a given year

4. What is the range of years covered in babynames?

```
(range(babynames$year))
```

```
[1] 1880 2017
```

The range of years covered is 1880-2017

5. Create an object from the babynames package that does not include the variable n

```
babynames_new <- select(babynames, -n)
```

6. What is one reason for not including n, but keeping the variable prop?

Since the proportions are relative to the number of applicants born in the given year and the total number of people born each year is not readily available to us, the proportion data would be more meaningful/useful for making comparisons across birth years. Count data are absolute frequencies while proportions provide relative frequencies which can be compared across conditions with differing numbers of observations.

7. Using the object created in Question 5, what was the most popular name for both sexes in:

a) the 2nd millennium?

```r
babynames_mil_2 <- babynames_new |>
  filter(year >= 1900, year <= 1999) |>
  group_by(name, sex) |>
  summarize(prop = sum(prop)) |>
  group_by(sex) |>
  filter(prop == max(prop))
babynames_mil_2
```

```
# A tibble: 2 x 3
# Groups:   sex [2]
  name  sex    prop
  <chr> <chr> <dbl>
1 John  M      3.81
2 Mary  F      3.27
```

In the second millennium, the most popular name for males was John and for females was Mary.

b) the 3rd millennium?

```r
babynames_mil_3 <- babynames_new |>
  filter(year >= 2000, year <= 2017) |>
  group_by(name, sex) |>
  summarize(prop = sum(prop)) |>
  group_by(sex) |>
  filter(prop == max(prop))
babynames_mil_3
```

```
# A tibble: 2 x 3
# Groups:   sex [2]
  name  sex    prop
  <chr> <chr> <dbl>
1 Emma  F     0.171
2 Jacob M     0.198
```

In the third millennium, the most popular name for males was Jacob and for females was Emma.

8. What were the most popular names beginning with the letters Q, V, and X between 2000 and 2012?

```
names_2000_2012 <- babynames_new |>
  filter(year >= 2000, year <= 2012) |>
  group_by(name) |>
  summarize(prop = sum(prop)) |>
  filter(substr(name, 1, 1) %in% c("Q", "V", "X")) |>
  group_by(letter = substr(name, 1, 1)) |>
  filter(prop == max(prop))
names_2000_2012
```

```
# A tibble: 3 x 3
# Groups:   letter [3]
  name       prop letter
  <chr>     <dbl> <chr>
1 Quinn    0.0116 Q
2 Victoria 0.0523 V
3 Xavier   0.0328 X
```

Between 2000 and 2012, the most popular name beginning with Q was Quinn, V was Victoria, and X was Xavier.

9. Create a new object that retains all the variables of the babynames package, but create a new column that contains the decade each year is a part of named decade.

```
babynames_decades <- babynames |>
  mutate(decade = paste0(substr(babynames$year, 1, 3), "0"))
```

10. What is the mean and median number of female and male babies in each decade?

```
females <- babynames_decades |>
  filter(sex == "F") |>
  summarize(mean = mean(n), median = median(n), .by = decade)
```

```r
males <- babynames_decades |>
  filter(sex == "M") |>
  summarize(mean = mean(n), median = median(n), .by = decade)
```

```r
knitr::kable(females, col.names = c("Decade", "Mean", "Median"), caption = "Females")
knitr::kable(males, col.names = c("Decade", "Mean", "Median"), caption = "Males")
```

Table 2: Females

Decade	Mean	Median
1880	110.57017	13
1890	128.18406	13
1900	131.32904	12
1910	187.06284	12
1920	210.54574	12
1930	214.19867	12
1940	262.20824	12
1950	288.47692	13
1960	234.71960	12
1970	147.20851	11
1980	134.25355	11
1990	113.07160	11
2000	96.45799	11
2010	91.69925	11

Table 3: Males

Decade	Mean	Median
1880	100.76497	12
1890	93.59019	12
1900	94.38963	12
1910	180.83854	12
1920	226.78161	13
1930	253.28957	13
1940	368.40859	14
1950	460.86555	14
1960	415.51792	13
1970	265.55153	12
1980	236.98189	11
1990	187.35187	11
2000	149.06677	11
2010	133.67495	11

11. In which decade(s) and year(s), was:

```r
popular_names_year <- babynames_decades |>
  filter(name %in% c("Mikayla", "Michael", "Jack", "Scott")) |>
  group_by(name) |>
  filter(prop == max(prop))
popular_names_year
```

```
# A tibble: 4 x 6
# Groups:   name [4]
   year sex   name        n    prop decade
  <dbl> <chr> <chr>   <int>   <dbl> <chr>
```

```
1  1927 M     Jack    12795 0.0110  1920
2  1969 M     Michael 85208 0.0466  1960
3  1971 M     Scott   30918 0.0170  1970
4  1998 F     Mikayla  3858 0.00199 1990
```

```r
popular_names_decade <- babynames_decades |>
  group_by(decade, name) |>
  summarize(prop = sum(prop)) |>
  filter(name %in% c("Mikayla", "Michael", "Jack", "Scott")) |>
  group_by(name) |>
  filter(prop == max(prop))
popular_names_decade
```

```
# A tibble: 4 x 3
# Groups:   name [4]
  decade name      prop
  <chr>  <chr>    <dbl>
1 1920   Jack    0.103
2 1960   Michael 0.429
3 1960   Scott   0.137
4 2000   Mikayla 0.0117
```

a) your name the most popular?

Mikayla was most popular in the year 1998, and in the decade 2000

b) your supervisor's name the most popular?

Michael was most popular in the year 1969, and in the decade 1960.

c) Mike's kids' names, Jack and Scott, the most popular?

Jack was most popular in the year 1927, and in the decade 1920. Scott was most popular in the year 1971, and in the decade 1960.