DAT 301 Lab 2B

The purpose of this lab is to explore data using the `dplyr` package from `tidyr`. First, load the packages:

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
```

The first portion of the lab will consider the British babynames in years 2017 and 1996. The data below was obtained from the Office for National Statistics and compiled into a .CSV file in Excel. Load the data using the below code and name the dataset `uknames_df`:

```
uknames_df = read.csv("ukbabynames.csv", sep = ",", header = TRUE)
```

Question 1

Create a dataframe that displays only the top 10 names in 2017 and 1996 for male and female babies born in the UK (suggestion: create four columns with each combination gender and year, example Boys2017, Girls2017,...). Include commentary such as if you see any names appear on both lists or if the names are generally different, etc.

```
# Create slices of each group (Boys2017, Girls2017, Boys1996, Girls1996)
B2017 = uknames_df %>% filter(Gender == "M", Year == 2017) %>%
  slice_max(Count, n = 10)

G2017 = uknames_df %>% filter(Gender == "F", Year == 2017) %>%
  slice_max(Count, n = 10)

B1996 = uknames_df %>% filter(Gender == "M", Year == 1996) %>%
  slice_max(Count, n = 10)

G1996 = uknames_df %>% filter(Gender == "F", Year == 1996) %>%
  slice_max(Count, n = 10)
```

```
# extract the name colons for each group
Boys2017 = B2017[,"Name"]
Girls2017 = G2017[,"Name"]
Boys1996 = B1996[,"Name"]
Girls1996 = G1996[,"Name"]

Rank = c(1:10)

# create data.frame
TopTen = data.frame(Rank, Boys2017, Girls2017, Boys1996, Girls1996)
TopTen
```

```
##    Rank Boys2017 Girls2017 Boys1996 Girls1996
## 1     1   Oliver    Olivia     Jack    Sophie
## 2     2    Harry    Amelia   Daniel     Chloe
## 3     3   George      Isla   Thomas   Jessica
## 4     4     Noah       Ava    James     Emily
## 5     5     Jack     Emily   Joshua    Lauren
## 6     6    Jacob  Isabella  Matthew    Hannah
## 7     7      Leo       Mia     Ryan Charlotte
## 8     8    Oscar     Poppy   Joseph   Rebecca
## 9     9  Charlie      Ella   Samuel       Amy
## 10   10 Muhammad      Lily     Liam     Megan
```

Comparing the top ten names for girls in 1996 and 2017, the only name that repeats is “Emily”. For boys names in 1996 and 2017, the only name that occurs in both years is “Jack”. All other names in the top ten are distinct for both girls are boys.

Question 2

Part A

Create a dataframe with the counts of babies, separated by gender, and compare the results by using the top 10 names in the UK in the years 2017 and 1996.

```
Boys_2017 = sum(B2017[,"Count"])
Girls_2017 = sum(G2017[,"Count"])
Boys_1996 = sum(B1996[,"Count"])
Girls_1996 = sum(G1996[,"Count"])

TopTenCount = data.frame(Boys_2017, Girls_2017, Boys_1996, Girls_1996)
TopTenCount
```

```
##   Boys_2017 Girls_2017 Boys_1996 Girls_1996
## 1     43584      31946     80070      61100
```

Part B

Create a dataframe that compares the proportion of babies, separated by gender, named top 10 names in the UK in the years 2017 and 1996

```
boys2017 = Boys_2017/ sum(uknames_df %>% filter(Year == 2017, Gender == "M")%>%
                            select(Count))

girls2017 = Girls_2017/ sum(uknames_df%>%filter(Year == 2017, Gender == "F")%>%
                              select(Count))

boys1996 = Boys_1996/ sum(uknames_df%>%filter(Year == 1996, Gender == "M")%>%
                            select(Count))

girls1996 = Girls_1996/ sum(uknames_df%>%filter(Year == 1996, Gender == "F")%>%
                              select(Count))

TopTenProp = data.frame(boys2017, girls2017, boys1996, girls1996)
TopTenProp
```

```
##    boys2017 girls2017  boys1996 girls1996
## 1 0.1349759  0.106811 0.2513838 0.2064203
```

Part C

What trends do you see in parts A and B?

Part A shows the number of babies with top ten names in 1996 and 2017 by gender. The largest sum being the baby boys in 1996 and least sum being the baby girls in 2017. Part B shows the proportion of babies (by gender) in 1996 and 2017 that have a name from the top ten names. It appears that a larger proportion of boys in 1996 have a top ten name compared to boys in 2017. Similarly, a larger of girls in 1996 have a top ten name compared to girls in 2017.

Question 3

Regardless of gender, what are the top 20 baby names in the UK in 2017 and what are their counts?

```
UKTopTwenty2017 = uknames_df %>% filter(Year == 2017) %>% select(Name, Count)%>%
  slice_max(Count, n=20)
UKTopTwenty2017
```

```
##        Name Count
## 1    Oliver  6259
## 2    Olivia  5204
## 3     Harry  5031
## 4    George  4929
## 5    Amelia  4358
## 6      Noah  4273
## 7      Jack  4190
## 8     Jacob  3968
## 9       Leo  3781
## 10    Oscar  3738
## 11  Charlie  3724
## 12 Muhammad  3691
## 13  William  3437
## 14     Isla  3373
## 15      Ava  3289
## 16    Alfie  3287
## 17    Henry  3246
## 18   Thomas  3246
## 19   Joshua  3166
## 20  Freddie  3127
```

The next part of the analysis will cover US babynames. The data obtained below is from the `babynames` package. First load and save the dataset as `usnames_df`:

```
library(babynames)
usnames_df = babynames
```

Question 4

Create a column in the `usnames_df` dataset that displays the total number of babies named that specific baby name in the entire data set regardless of year. Display the top 20 girl and top 20 boy names.

```
# Top 20 Names for Girls
TopGirls = usnames_df %>% filter(sex == "F") %>% group_by(name) %>%
  summarise(count = sum(n))%>% slice_max(order_by = count, n = 20)
TopGirls
```

```
## # A tibble: 20 x 2
##    name        count
##    <chr>       <int>
##  1 Mary      4123200
##  2 Elizabeth 1629679
##  3 Patricia  1571692
##  4 Jennifer  1466281
##  5 Linda     1452249
##  6 Barbara   1434060
##  7 Margaret  1246649
##  8 Susan     1121440
##  9 Dorothy   1107096
## 10 Sarah     1073895
## 11 Jessica   1044939
## 12 Helen     1018290
## 13 Nancy     1002010
## 14 Betty      999474
## 15 Karen      985655
## 16 Lisa       964973
## 17 Anna       888505
## 18 Sandra     873512
## 19 Ashley     843819
## 20 Emily      841491
```

```
# Top 20 Names for Boys
TopBoys = usnames_df %>% filter(sex == "M") %>% group_by(name) %>%
  summarise(count = sum(n))%>% slice_max(order_by = count, n = 20)
TopBoys
```

```
## # A tibble: 20 x 2
##    name          count
##    <chr>         <int>
##  1 James       5150472
##  2 John        5115466
##  3 Robert      4814815
##  4 Michael     4350824
##  5 William     4102604
##  6 David       3611329
##  7 Joseph      2603445
##  8 Richard     2563082
##  9 Charles     2386048
## 10 Thomas      2304948
## 11 Christopher 2022164
## 12 Daniel      1907357
## 13 Matthew     1590440
## 14 George      1464186
## 15 Anthony     1432718
## 16 Donald      1410998
## 17 Paul        1386815
## 18 Mark        1349865
## 19 Edward      1288725
## 20 Andrew      1283910
```

Question 5

What percent of US girl and boy names were top 10 in 2017 vs. 1996? Display answer in a dataframe. Include commentary on anything you observe that you think is interesting.

```
G1996_US = usnames_df %>% filter(sex == "F", year == 1996)%>%
  slice_max(n, n = 10)
B1996_US = usnames_df %>% filter(sex == "M", year == 1996)%>%
  slice_max(n, n = 10)
G2017_US = usnames_df %>% filter(sex == "F", year == 2017)%>%
  slice_max(n, n = 10)
B2017_US = usnames_df %>% filter(sex == "M", year == 2017)%>%
  slice_max(n, n = 10)


girls1996_US = sum(G1996_US$n)/ sum(usnames_df %>% filter(sex == "F",
                                                     year == 1996)%>% select(n))
boys1996_US = sum(B1996_US$n)/ sum(usnames_df %>% filter(sex == "M",
                                                     year == 1996)%>% select(n))
girls2017_US = sum(G2017_US$n)/ sum(usnames_df %>% filter(sex == "F",
                                                     year == 2017)%>% select(n))
boys2017_US = sum(B2017_US$n)/ sum(usnames_df %>% filter(sex == "M",
                                                     year == 2017)%>% select(n))

TopTenProp_US = data.frame(girls1996_US, boys1996_US, girls2017_US, boys2017_US)
TopTenProp_US
```

```
##   girls1996_US boys1996_US girls2017_US boys2017_US
## 1    0.1147185   0.1551609   0.08386381  0.08008438
```

in 1996, a larger proportion of boys and girls were more likely to have a name in the top ten category compared to the boys and girls in 2017.

Question 6

Part A

Find the top 20 US names in 2017, regardless of gender. Display with their counts.

```
USTopTwenty2017 = usnames_df %>% filter(year == 2017) %>% select(name, n)%>%
  slice_max(n, n=20)
USTopTwenty2017
```

```
## # A tibble: 20 x 2
##    name          n
##    <chr>     <int>
##  1 Emma      19738
##  2 Liam      18728
##  3 Olivia    18632
##  4 Noah      18326
##  5 Ava       15902
##  6 Isabella  15100
##  7 William   14904
##  8 Sophia    14831
##  9 James     14232
## 10 Logan     13974
## 11 Benjamin  13733
## 12 Mason     13502
## 13 Mia       13437
## 14 Elijah    13268
## 15 Oliver    13141
## 16 Jacob     13106
## 17 Lucas     12951
## 18 Charlotte 12893
## 19 Michael   12579
## 20 Alexander 12467
```

Part B

Compare the top 20 names in the UK and the US in 2017. Which names were used in both the UK and the US?

```
intersect(UKTopTwenty2017$Name, USTopTwenty2017$name)
```

```
## [1] "Oliver"  "Olivia"  "Noah"    "Jacob"   "William" "Ava"
```

Question 7

Create a function that will look up a US babyname based on the name and gender and return the count of names of all babies in the dataset that have been named that name.

```r
# a_name - name to be searched
# a_sex - sex to be searched ("F" or "M")
countName = function(a_name, a_sex)
{
  df = usnames_df %>% group_by(name, sex) %>% summarise(count = sum(n))
  x = df$count[df$name == a_name & df$sex == a_sex]
  return(x)
}
```