# Assignment #3 - TidyR

*Jacob Burke*

*02/10/2019*

## Problem 1

```
library(tidyr)
library(gapminder)
library(stringr)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```
library(dplyr)
```

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

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

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

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

```
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
library(knitr)
library(ggplot2)
## a)

unique(gapminder$continent)
```

```
## [1] Asia     Europe   Africa   Americas Oceania
## Levels: Africa Americas Asia Europe Oceania
```

```
## Therefore there are 5 continents in data set
```

```r
## b)
length(unique(gapminder$country))
```

```
## [1] 142
```

```r
## 142 countries

Africa <- filter(gapminder, gapminder$continent == "Africa")
length(unique(Africa$country))
```

```
## [1] 52
```

```r
## 52 countries for Africa

Asia <- filter(gapminder, gapminder$continent == "Asia")
length(unique(Asia$country))
```

```
## [1] 33
```

```r
## 33 countries for Asia

Americas <- filter(gapminder, gapminder$continent == "Americas")
length(unique(Americas$country))
```

```
## [1] 25
```

```r
## 25 countries for Americas

Europe <- filter(gapminder, gapminder$continent == "Europe")
length(unique(Europe$country))
```

```
## [1] 30
```

```r
## 30 countries for Europe

Oceania <- filter(gapminder, gapminder$continent == "Oceania")
length(unique(Oceania$country))
```

```
## [1] 2
```

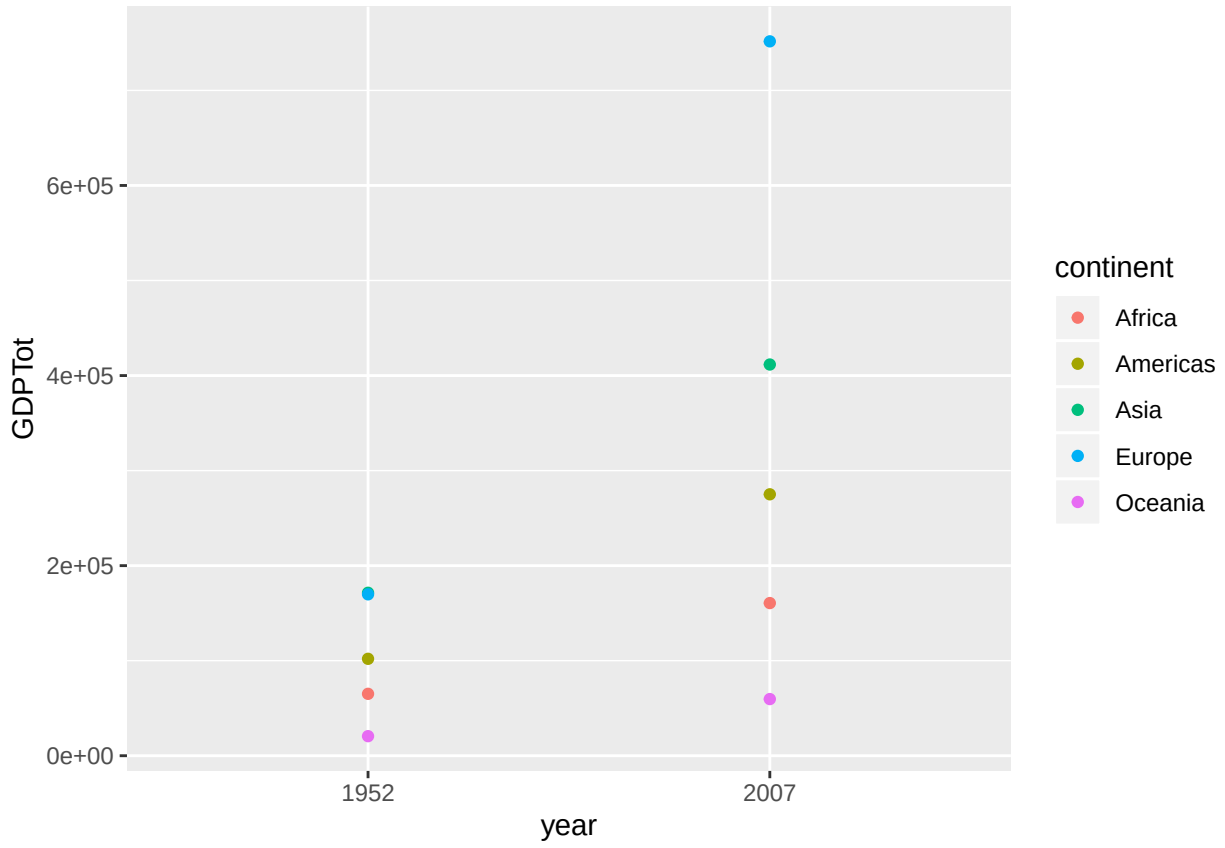
```r
## 2 countries for Oceania

## c)
Report <- gapminder

Report <- Report %>% group_by(continent) %>%
  summarise(PopTot = sum(as.numeric(pop)), GDPTot = sum(as.numeric(gdpPercap))) %>%
  arrange(continent, PopTot, GDPTot)

## d)

Report_1952 <- filter(gapminder, year == 1952)

Report_2007 <- filter(gapminder, year == 2007)

Report_1952 %<>% group_by(continent) %>%
  summarise(GDPTot = sum(as.numeric(gdpPercap))) %>%
  arrange(continent, GDPTot)
```

Table 1: GDP Totals: 1952, 2007

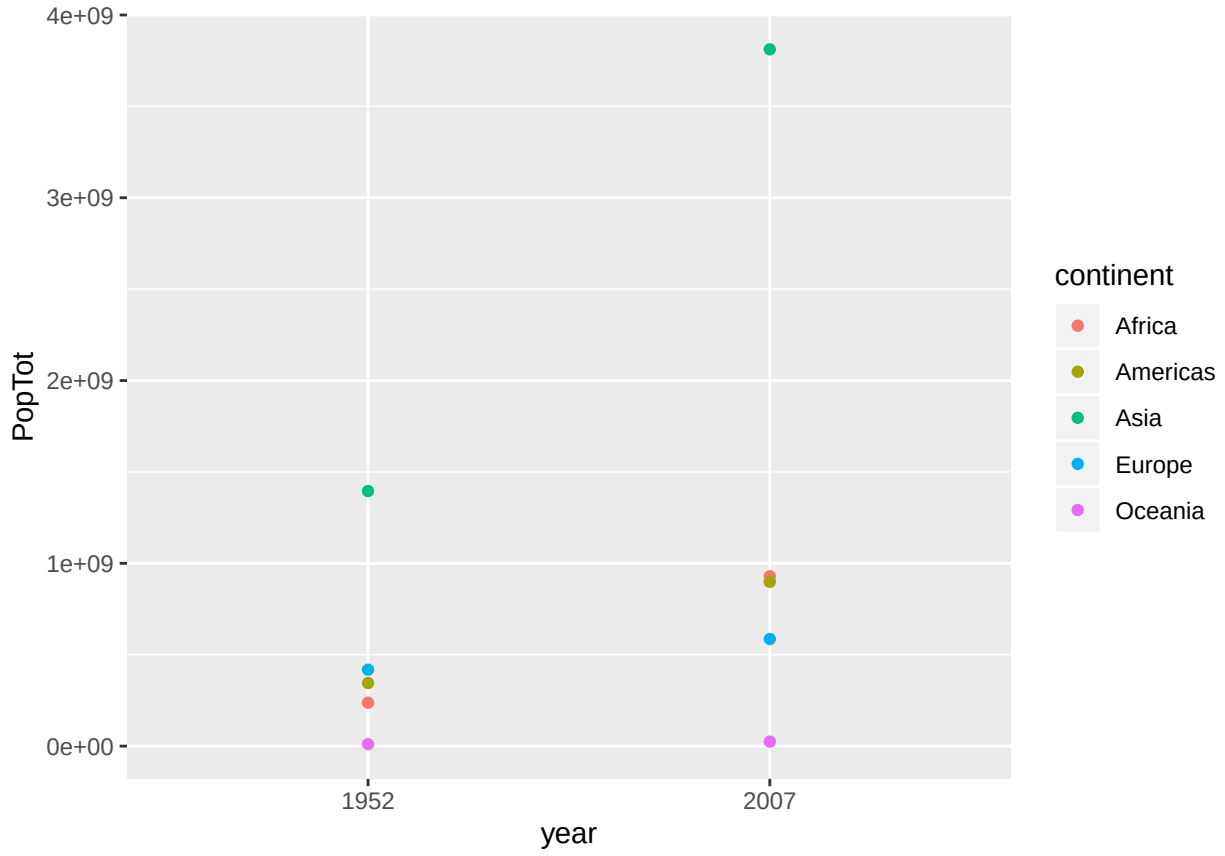
| 1952 | | 2007 | |
|---|---|---|---|
| continent | GDPTot | continent | GDPTot |
| Africa | 65133.77 | Africa | 160629.70 |
| Americas | 101976.56 | Americas | 275075.79 |
| Asia | 171450.97 | Asia | 411609.89 |
| Europe | 169831.72 | Europe | 751634.45 |
| Oceania | 20596.17 | Oceania | 59620.38 |

```
Report_2007 %<>% group_by(continent) %>%
  summarise(GDPTot = sum(as.numeric(gdpPercap))) %>%
  arrange(continent, GDPTot)

table <- cbind(Report_1952, Report_2007)

kable(table, digits = 2, format = "latex", booktabs=TRUE, caption = "GDP Totals: 1952, 2007") %>% kable_
  add_header_above(c("1952" = 2, "2007" = 2))
```

```
## e)

table2 <- rbind(Report_1952, Report_2007)
table2_Africa <- filter(table2, continent == "Africa")

table2_Africa <- mutate(table2_Africa, year = c("1952", "2007"))

table2_Asia <- filter(table2, continent == "Asia")

table2_Asia <- mutate(table2_Asia, year = c("1952", "2007"))

table2_Americas <- filter(table2, continent == "Americas")

table2_Americas <- mutate(table2_Americas, year = c("1952", "2007"))

table2_Europe <- filter(table2, continent == "Europe")

table2_Europe <- mutate(table2_Europe, year = c("1952", "2007"))

table2_Oceania <- filter(table2, continent == "Oceania")

table2_Oceania <- mutate(table2_Oceania, year = c("1952", "2007"))

plot_GDP <- rbind(table2_Africa, table2_Americas, table2_Asia, table2_Europe, table2_Oceania)


ggplot(plot_GDP) + geom_point(mapping = aes(x = year, y = GDPTot, colour = continent))
```




```
## We can see that there is no continent with decreased GDP

## f)

Report_1952 <- filter(gapminder, year == 1952)

Report_2007 <- filter(gapminder, year == 2007)

Report_1952 %<>% group_by(continent) %>%
  summarise(PopTot = sum(as.numeric(pop))) %>%
  arrange(continent, PopTot)

Report_2007 %<>% group_by(continent) %>%
  summarise(PopTot = sum(as.numeric(pop))) %>%
  arrange(continent, PopTot)

Report_2007 <- mutate(Report_2007, year = rep("2007", 5))
Report_1952 <- mutate(Report_1952, year = rep("1952", 5))

plot_pop <- rbind(Report_2007, Report_1952)

ggplot(plot_pop) + geom_point(mapping = aes(x = year, y = PopTot, colour = continent))
```




```
## again we can see no negative pop growth for each continent from 1952 to 2007

## g)

rate <- NULL

## getting Rate for each continent
for(i in seq(1,10, 2)){

  r <- abs(plot_GDP$GDPTot[i] - plot_GDP$GDPTot[i+1])/(55)
  rate <- c(rate, r)

}

GDP_Rate <- cbind(rate, c("Africa", "Asia", "America", "Europe", "Oceania"))


colnames(GDP_Rate) <- c("rate", "continent")

## Therefore we can see that Europe had the highest Rate of GDP growth over 55 years from 1952 to 2007
```

# Problem 2

```
library(AER)
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## Loading required package: lmtest
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## Loading required package: sandwich
```

```
## Loading required package: survival
```

```
data("Fertility")

Fertility <- as_tibble(Fertility)

?Fertility
```

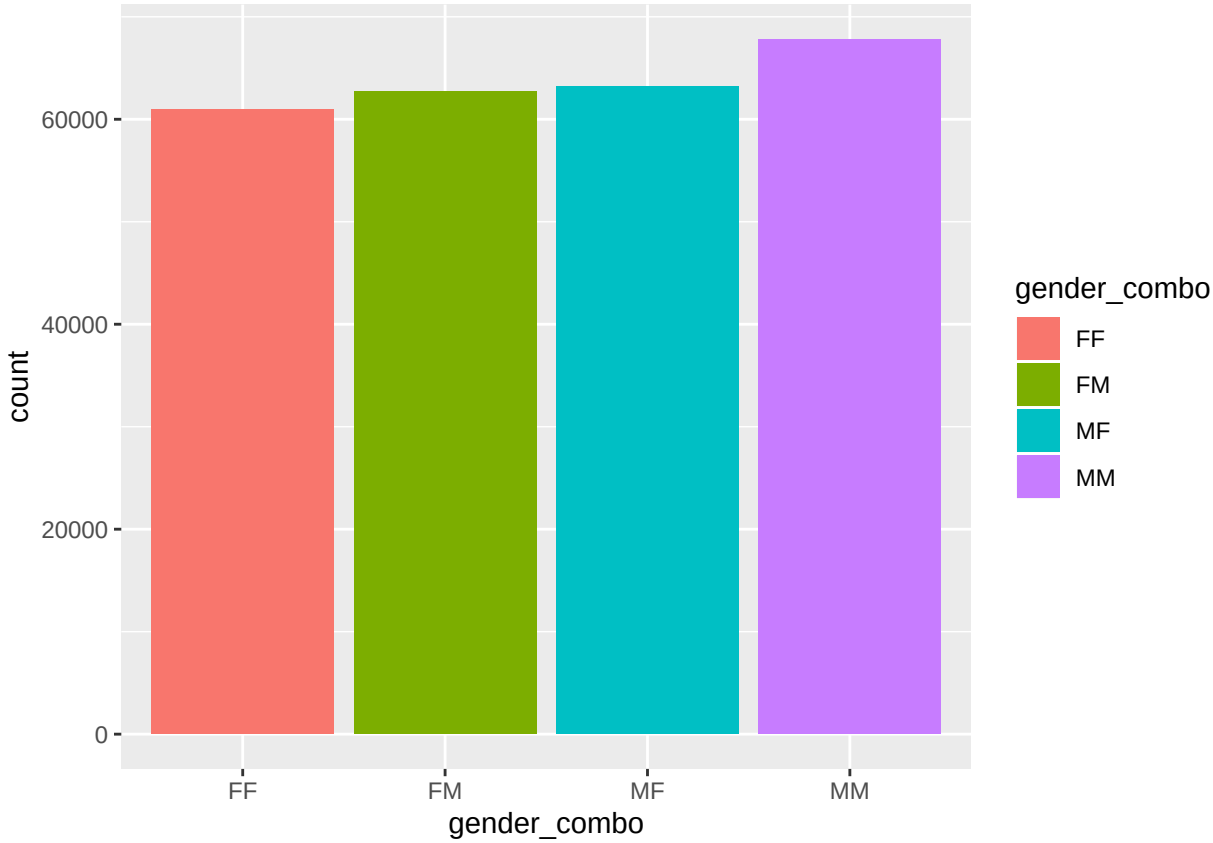
```
## starting httpd help server ...
```

```
##  done
```

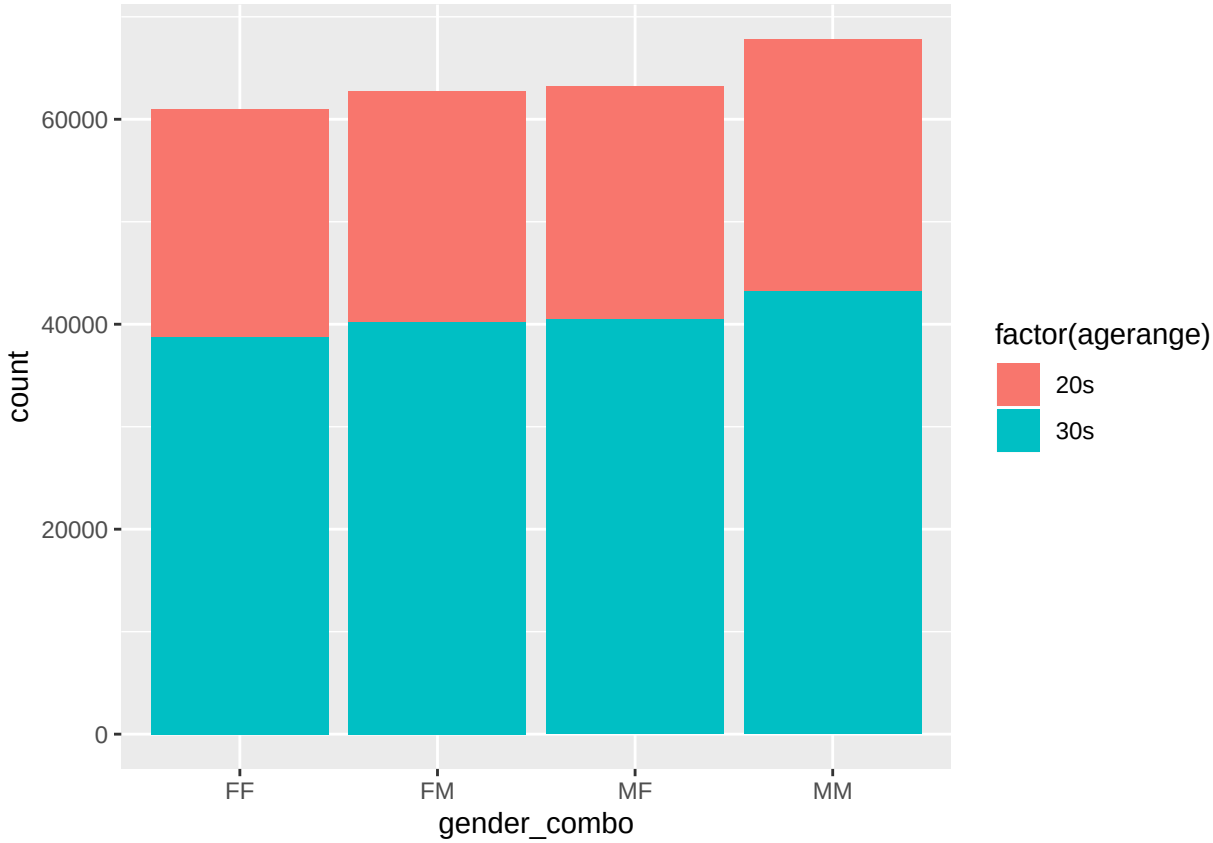
```
Fertility <- mutate(Fertility, gender_combo = NA)

Fertility$gender_combo[Fertility$gender1 == "male" & Fertility$gender2 == "male"] <- "MM"
Fertility$gender_combo[Fertility$gender1 == "male" & Fertility$gender2 == "female"] <- "MF"
Fertility$gender_combo[Fertility$gender1 == "female" & Fertility$gender2 == "male"] <- "FM"
Fertility$gender_combo[Fertility$gender1 == "female" & Fertility$gender2 == "female"] <- "FF"

table(Fertility$gender_combo)
```

```
##
##    FF    FM    MF    MM
## 60946 62724 63185 67799
```

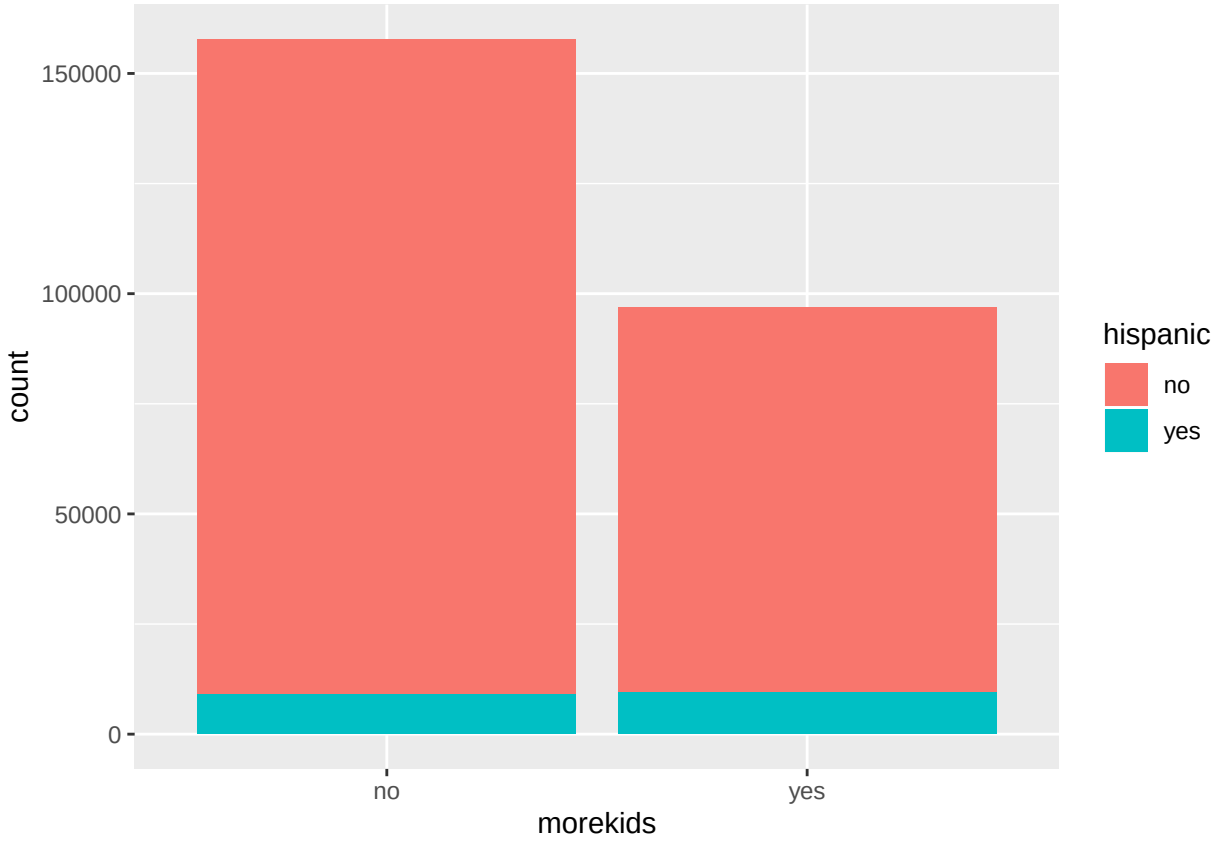
```
genderc <- table(Fertility$gender_combo)

ggplot(Fertility) + geom_bar(mapping = aes(x = gender_combo, fill = gender_combo))
```




```
## contrasting years for women in 20s versus women older than 29

Fertility <- mutate(Fertility, agerange = NA)

Fertility$agerange[Fertility$age >= 20 & Fertility$age <=29] <- "20s"

Fertility$agerange[Fertility$age > 29] <- "30s"

ggplot(Fertility) + geom_bar(mapping = aes(x = gender_combo, fill = factor(agerange)))
```




```
## as we can see there's fairly even frequencie distribution for women in their 20s and 30s

##b)

Fertility <- mutate(Fertility, morethan2 = NA)

Fertility$morethan2[Fertility$morekids == 'yes'] <- 1
Fertility$morethan2[Fertility$morekids == 'no'] <- 0

## contrasting more than 2 kids based on race (hispanic)

ggplot(Fertility) + geom_bar(aes(x = morekids, fill = hispanic))
```




```
## as you can see, there is more of a difference with people not
## hispanic, while hispanics have more of a 50:50 split with number
## of kids higher and lower than 2
```

## Problem 3

```
library(stringr)
data(mpg)
data(mtcars)

## a)

number_e <- str_count(rownames(mtcars), "e")

sum(number_e)
```

```
## [1] 25
```

```
## therefore 25 times does an e show up in row names of mtcars

## b)

number_merc_mt <- str_count(row.names(mtcars), "Merc")
sum(number_merc_mt)
```

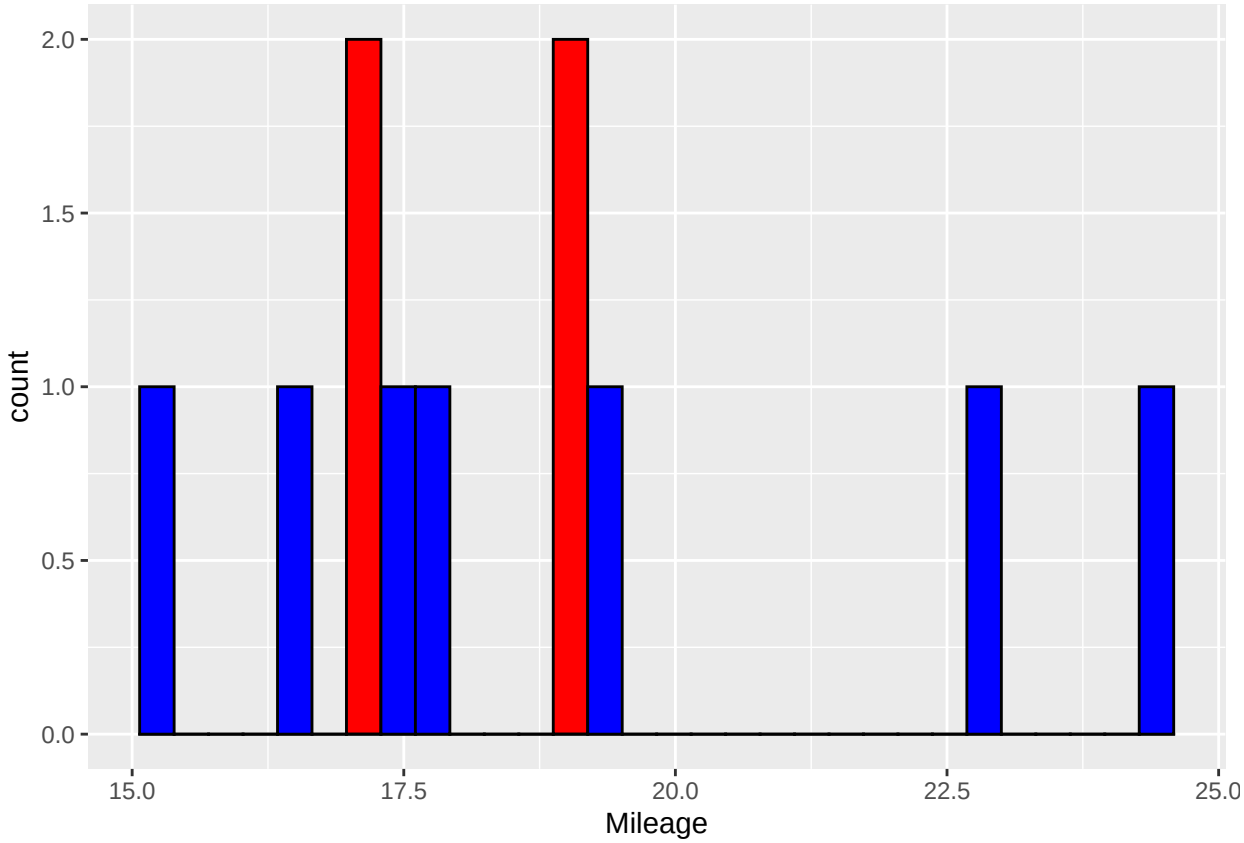
```
## [1] 7
```

```
## 7 times

## c)

number_merc <- str_count(mpg$manufacturer, "merc")

sum(number_merc)
```

```
## [1] 4
```

```
## 4 times

## d)

mpg <- mutate(mpg, merc = number_merc)

mtcars <- mutate(mtcars, merc = number_merc_mt)

mpg_merc <- filter(mpg, merc == 1)

mtcars_merc <- filter(mtcars, merc == 1)

## contrasting mpg mileage column of mtcars, and hwy mileage of mpg

ggplot() + geom_histogram(aes(x = mpg_merc$hwy), fill ='red',
                          color = 'black') +  geom_histogram(aes(x =
                          mtcars_merc$mpg), fill = 'blue', color =
                          'black') + xlab("Mileage")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## mpg in red, and mtcars in blue
```

## Problem 4

```
library(babynames)
data(babynames)

babynames<- as_tibble(babynames)

## a) (top baby names)

set.seed(2019)

baby_sample <- babynames[sample(nrow(babynames), 500000), ]

## 1880

baby1880 <- filter(baby_sample, year == 1880)
baby1880male <- filter(baby1880, sex == "M")

baby1880male <- baby1880male[
  with(baby1880male, order(-n)),]

baby1880male <- select(baby1880male, name)
```

```r
baby1880female <- filter(baby1880, sex == "F")

baby1880female <- baby1880female[
  with(baby1880female, order(-n)),]

baby1880female <- select(baby1880female, name)

baby_1880 <- cbind(baby1880male[1:3, ], baby1880female[1:3, ])

## 1920

baby1920 <- filter(baby_sample, year == 1920)
baby1920male <- filter(baby1920, sex == "M")

baby1920male <- baby1920male[
  with(baby1920male, order(-n)),]

baby1920male <- select(baby1920male, name)

baby1920female <- filter(baby1920, sex == "F")

baby1920female <- baby1920female[
  with(baby1920female, order(-n)),]

baby1920female <- select(baby1920female, name)


baby_1920 <- cbind(baby1920male[1:3, ], baby1920female[1:3, ])

## 1960

baby1960 <- filter(baby_sample, year == 1960)
baby1960male <- filter(baby1960, sex == "M")

baby1960male <- baby1960male[
  with(baby1960male, order(-n)),]

baby1960male <- select(baby1960male, name)

baby1960female <- filter(baby1960, sex == "F")

baby1960female <- baby1960female[
  with(baby1960female, order(-n)),]

baby1960female <- select(baby1960female, name)

baby_1960 <- cbind(baby1960male[1:3, ], baby1960female[1:3, ])

## 2000

baby2000 <- filter(baby_sample, year == 2000)
baby2000male <- filter(baby2000, sex == "M")
```

Table 2: Top Baby Boy and Girl Names By Year

| 1880 | | 1920 | | 1960 | | 2000 | |
|---|---|---|---|---|---|---|---|
| Boy | Girl | Boy | Girl | Boy | Girl | Boy | Girl |
| name | name | name | name | name | name | name | name |
| William | Mary | Harold | Ruth | James | Mary | Jacob | Alexis |
| Charles | Annie | Donald | Betty | William | Linda | Matthew | Jessica |
| Joe | Cora | Arthur | Edna | Thomas | Karen | Andrew | Lauren |

```
baby2000male <- baby2000male[
  with(baby2000male, order(-n)),]

baby2000male <- select(baby2000male, name)

baby2000female <- filter(baby2000, sex == "F")

baby2000female <- baby2000female[
  with(baby2000female, order(-n)),]

baby2000female <- select(baby2000female, name)


baby_2000 <- cbind(baby2000male[1:3, ], baby2000female[1:3, ])


## Bringing each year tables together

babynames_rank <- cbind(
                  baby_1880, baby_1920, baby_1960,
                  baby_2000)
## table

kable(babynames_rank, digits = 2, format = "latex", booktabs=TRUE,
      caption = "Top Baby Boy and Girl Names By Year") %>% kable_styling() %>%
      add_header_above(c("Boy" = 1, "Girl" = 1, "Boy" = 1, "Girl" = 1, "Boy" = 1,
                         "Girl" = 1, "Boy" = 1, "Girl" =1)) %>%
      add_header_above(c("1880" = 2, "1920" = 2, "1960" = 2, "2000" = 2))
```

```
## b)

baby_male <- filter(baby_sample, sex == "M")
baby_female <- filter(baby_sample, sex == "F")

match <- intersect(baby_male$name, baby_female$name)

length(match)
```

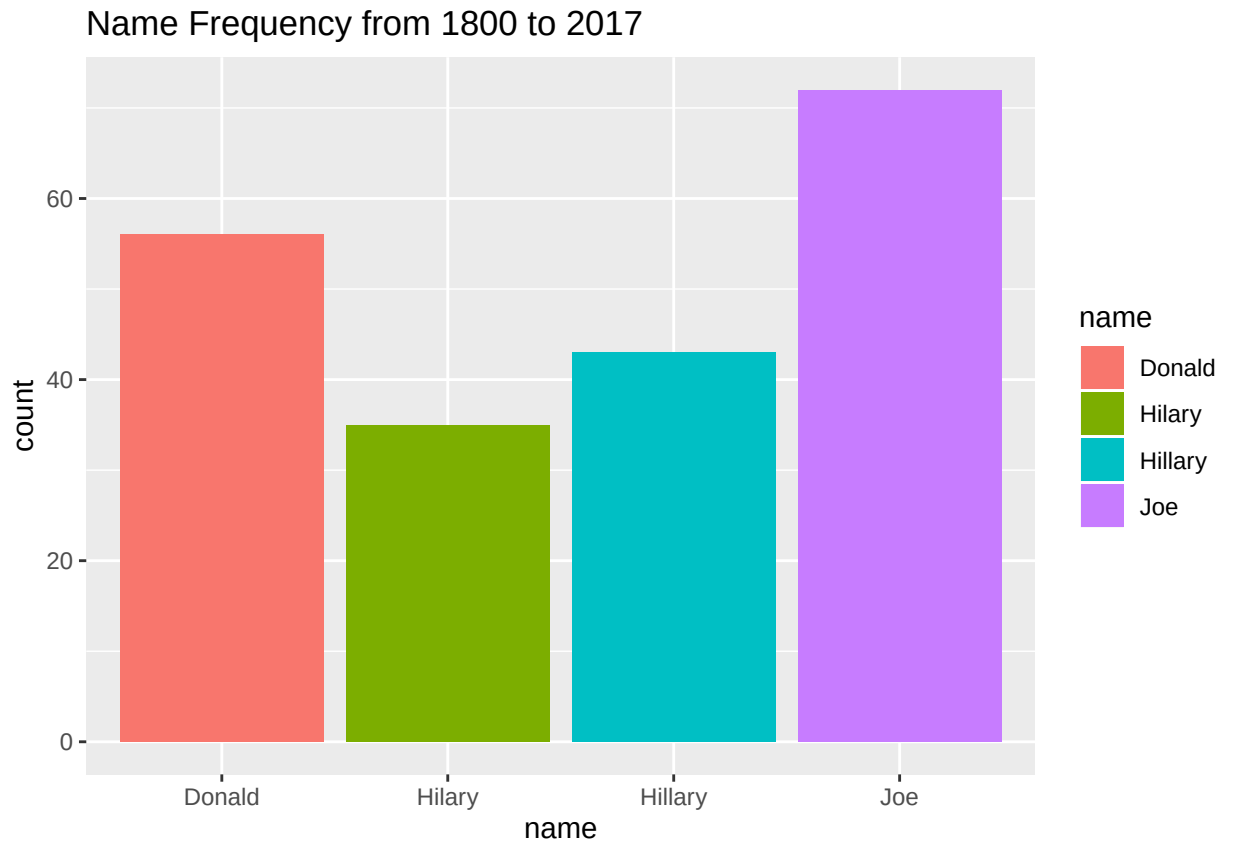
```
## [1] 7438
```

```
## Therefore there is overlap with 7464 names, all included in the "match" vector

## c)
```

```r
baby_1800s <- filter(baby_sample, 1800 <= year &
                       year < 1900)
baby_21st <- filter(baby_sample, 2000 <= year)

difference <- setdiff(baby_1800s$name, baby_21st$name)

length(difference)
```

```
## [1] 1015
```

```r
## Therefore there are 1032 names stored in 'difference" that
## were used in the 19th century but not in the 21st century


## d)

baby_hist_values <- filter(baby_sample, name == "Donald" | name == "Hilary" | name == "Hillary" | name =

## histogram

ggplot(baby_hist_values) + geom_bar(aes(x = name, fill = name)) + ggtitle("Name Frequency from 1800 to 1
```




```r
####
```