DATAWRANGLING_1

Vishwa Patel (vup4)

2023-01-30

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
install.packages("babynames", repos = "http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/ds/cdtrh2p16y94l1574pbn8q3w0000gn/T//RtmpmMQglX/downloaded_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(ggplot2)
library(magrittr)
library(babynames)
head(babynames)
```

```
## # A tibble: 6 x 5
##    year sex   name          n   prop
##   <dbl> <chr> <chr>     <int>  <dbl>
## 1  1880 F     Mary       7065 0.0724
## 2  1880 F     Anna       2604 0.0267
## 3  1880 F     Emma       2003 0.0205
## 4  1880 F     Elizabeth  1939 0.0199
## 5  1880 F     Minnie     1746 0.0179
## 6  1880 F     Margaret   1578 0.0162
```

```
babyname_taylor = filter(babynames, name=="Taylor") %>%
group_by(year,sex) %>% summarise(Total=sum(n))
```

```
## `summarise()` has grouped output by 'year'. You can override using the
## `.groups` argument.
```

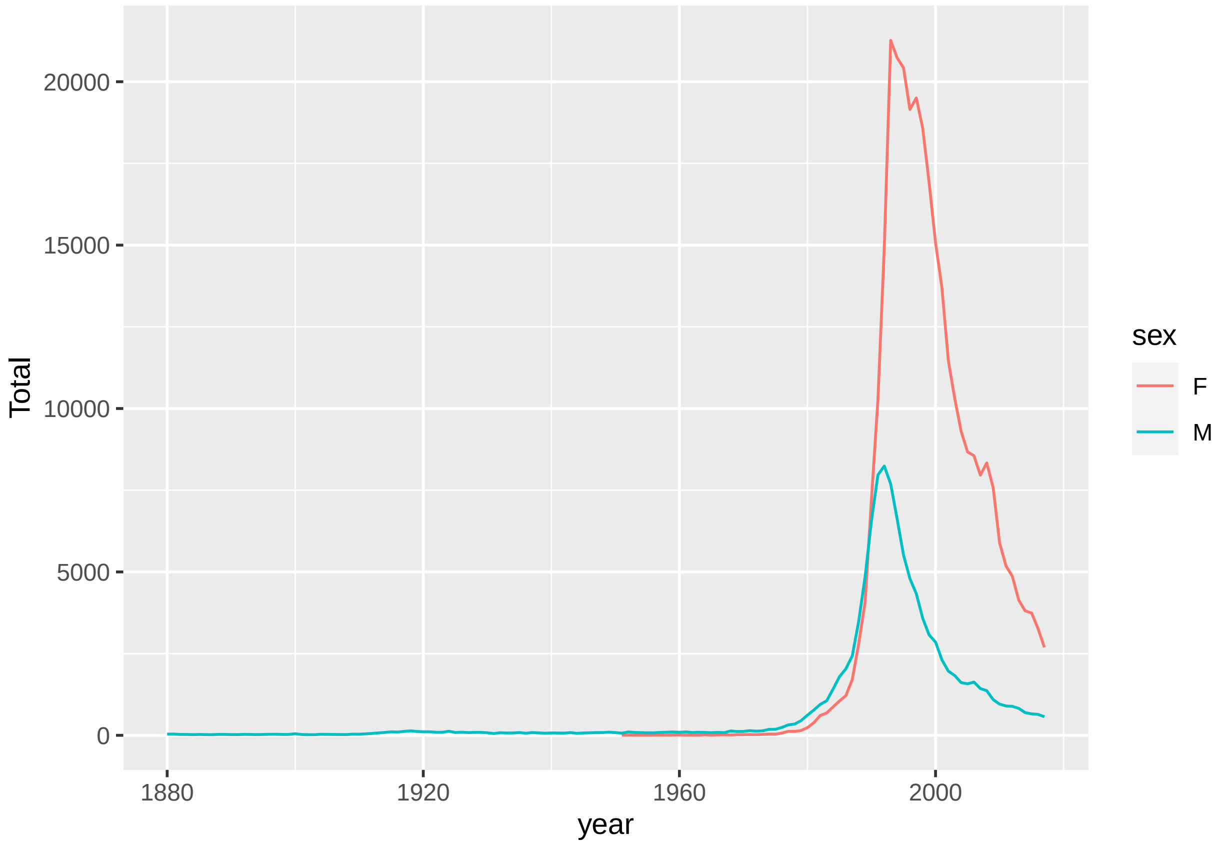
```
babyname_taylor
```

```
## # A tibble: 201 x 3
## # Groups:   year [138]
##     year sex   Total
##    <dbl> <chr> <int>
##  1  1880 M        37
##  2  1881 M        39
##  3  1882 M        27
##  4  1883 M        27
##  5  1884 M        21
##  6  1885 M        26
##  7  1886 M        22
##  8  1887 M        20
##  9  1888 M        29
## 10  1889 M        28
## # ... with 191 more rows
```

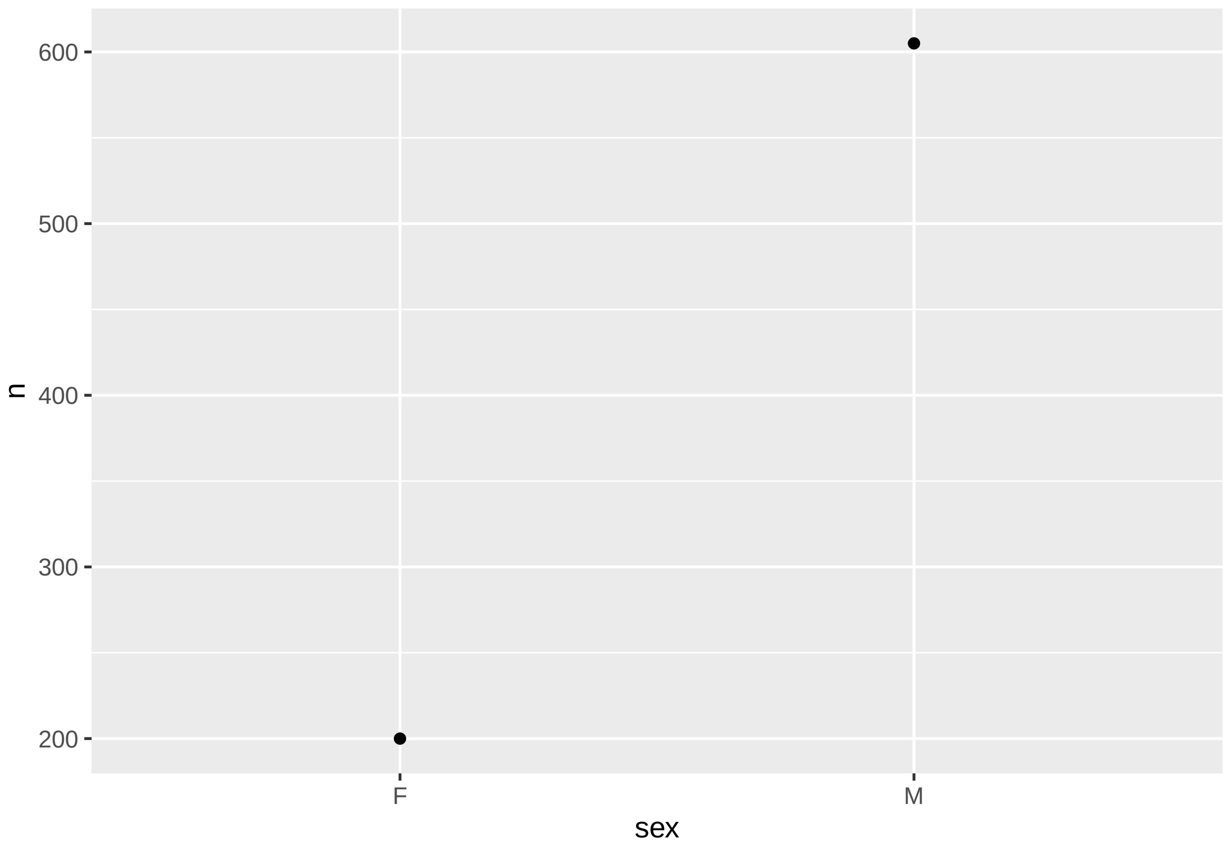
```
head(babyname_taylor)
```

```
## # A tibble: 6 x 3
## # Groups:   year [6]
##    year sex   Total
##   <dbl> <chr> <int>
## 1  1880 M        37
## 2  1881 M        39
## 3  1882 M        27
## 4  1883 M        27
## 5  1884 M        21
## 6  1885 M        26
```

```
#1. Plot the number of male and female babies named Taylor by year
ggplot(babyname_taylor, aes(x =year , y = Total, color = sex)) +
geom_line()
```

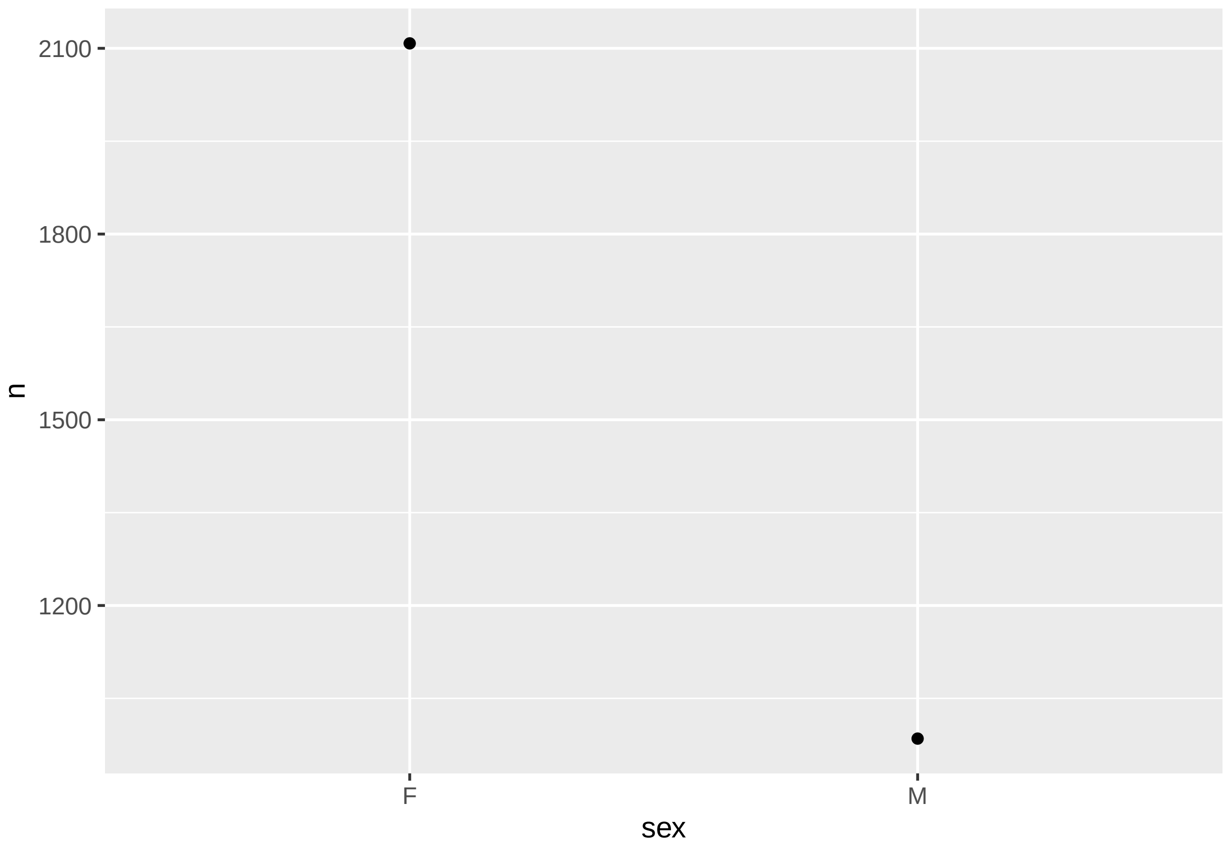


```
#2 a) Is a 23-year old named Quinn more likely to be a boy or a girl?
babyname_Quinn_23=filter(babynames, name=="Quinn" & year==(2018-23))%>% select(sex,n)
ggplot(babyname_Quinn_23, aes(sex,n)) +
geom_point()
```



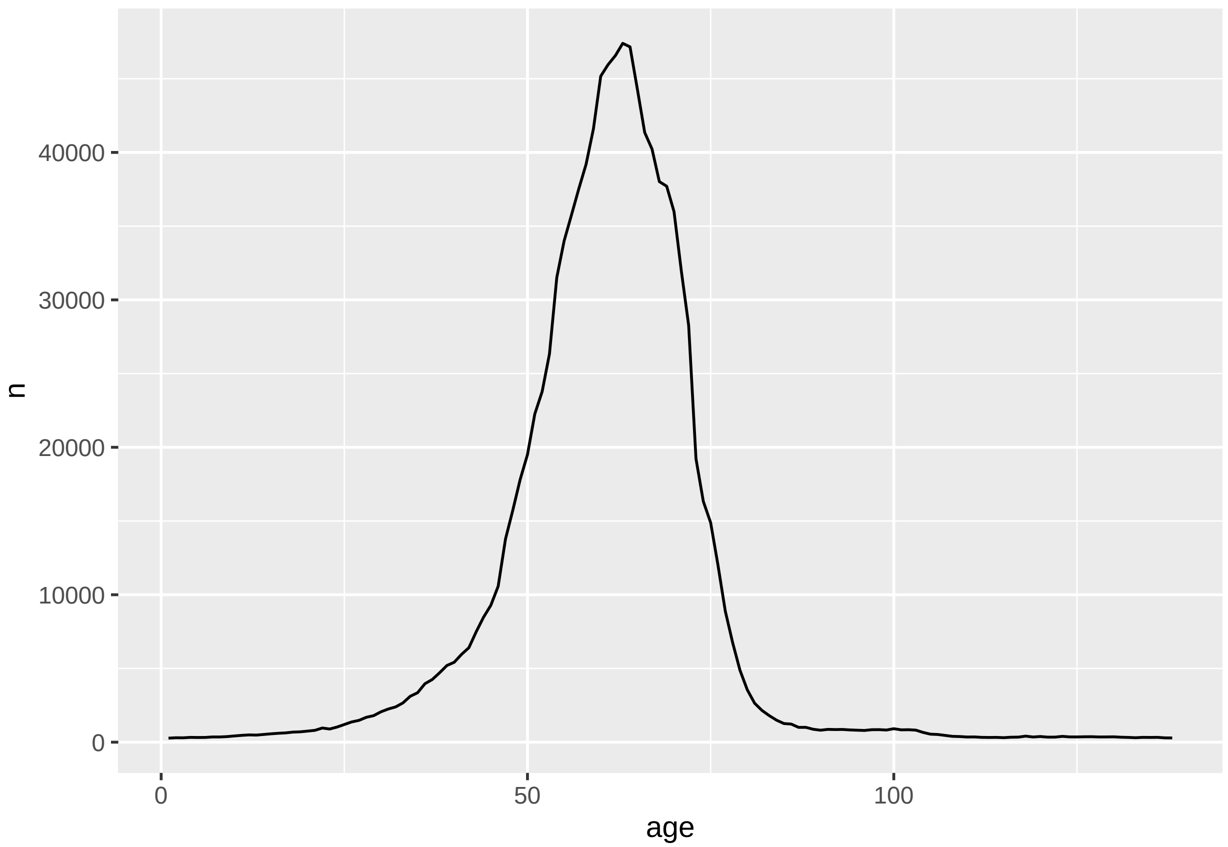
#23-year old named Quinn is more likely to be MALE(boy)

```
#2 b) Is a 6 year old named Quinn more likely to be a boy or a girl?
babyname_Quinn_6=filter(babynames, name=="Quinn" & year==(2018-6))%>% select(sex,n)
ggplot(babyname_Quinn_6, aes(sex,n)) +
geom_point()
```



# 6 year old named Quinn is more likely to be FEMALE(girl)

```
#2 c) What is your best guess as to how old a woman named Susan is?
babyname_Susan=filter(babynames, name=="Susan" & sex=="F")%>%
summarise(age=(2018-year),n)
ggplot(babyname_Susan, aes(x =age , y = n)) +
geom_line()
```



#best guess for woman named susan's age is 62-70 yo

```
#2 d) Find the five most popular female names in the year 2017.
baby_female=filter(babynames, year==2018 & sex=="F")%>% select(name,n) %>% arrange(desc(n))
baby_top5 = baby_female[1:5,]
baby_top5
```

```
## # A tibble: 5 x 2
##   name      n
##   <chr> <int>
## 1 <NA>     NA
## 2 <NA>     NA
## 3 <NA>     NA
## 4 <NA>     NA
## 5 <NA>     NA
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.