Assignment 1

Feiyu Zheng (fz114)

1/25/2022

Note: In this assignment, I am using 2022 as "the current year" instead of 2017.

Install the package "babynames"

```
#install.packages("babynames") # install the data package
library(babynames) # load the data
```

Show the data

```
babynames
```

```
## # A tibble: 1,924,665 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
##  7  1880 F     Ida        1472 0.0151
##  8  1880 F     Alice      1414 0.0145
##  9  1880 F     Bertha     1320 0.0135
## 10  1880 F     Sarah      1288 0.0132
## # ... with 1,924,655 more rows
```

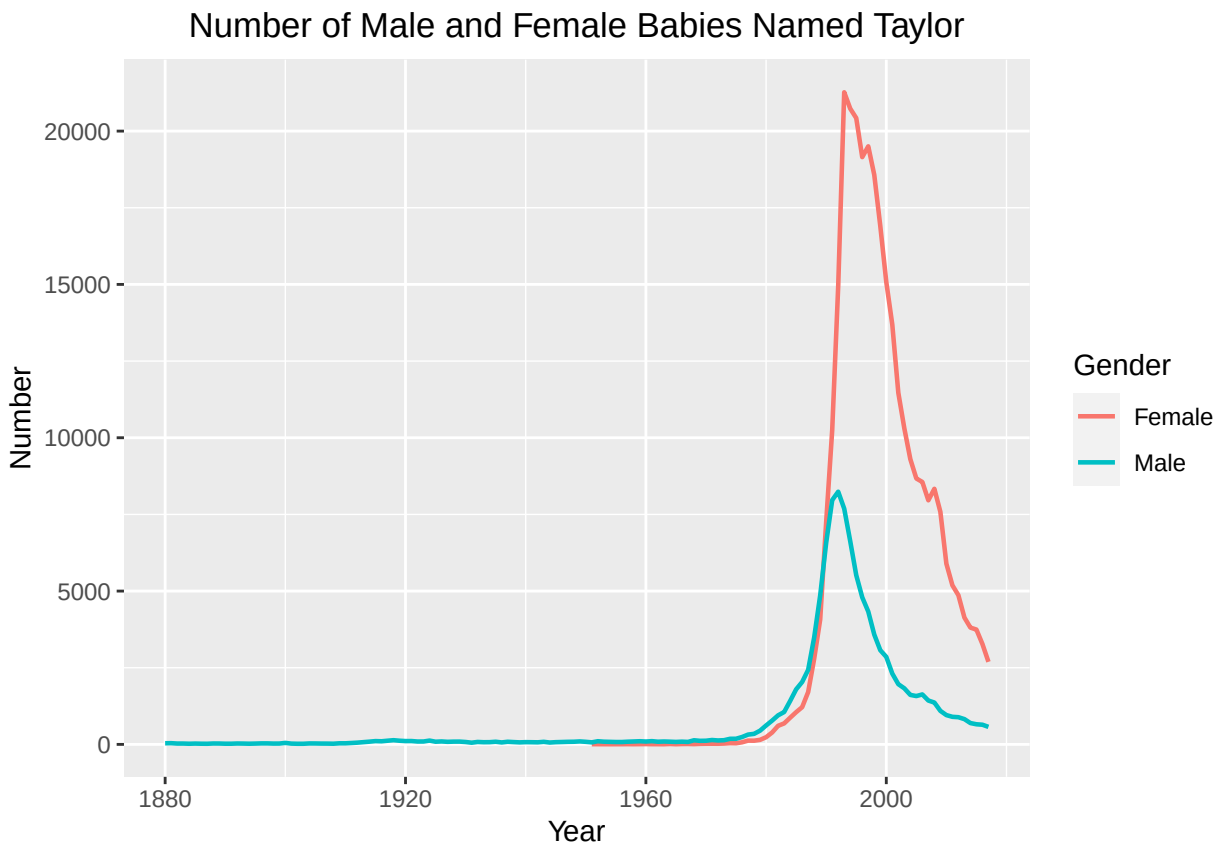
Plot the number of male and female babies named Taylor *by year*

```
# filter the data
babiesNamedTaylor <- babynames %>%
  filter(name=="Taylor") %>%
  mutate(sex = recode(sex, "F"="Female", "M" = "Male"))
# show the data
head(babiesNamedTaylor)
```

```
## # A tibble: 6 x 5
##    year sex   name       n     prop
##   <dbl> <chr> <chr>  <int>    <dbl>
## 1  1880 Male  Taylor    37 0.000312
## 2  1881 Male  Taylor    39 0.000360
## 3  1882 Male  Taylor    27 0.000221
## 4  1883 Male  Taylor    27 0.000240
## 5  1884 Male  Taylor    21 0.000171
```

```
## 6  1885 Male  Taylor    26 0.000224
```

```
# plot
ggplot(babiesNamedTaylor) +
  geom_line(mapping = aes(year, n, color=sex), size = 0.8) +
  labs(x = "Year", y = "Number", color = "Gender") +
  ggtitle("Number of Male and Female Babies Named Taylor") +
  theme(plot.title = element_text(hjust = 0.5))
```



Answer the following questions, showing plots to substantiate your answers:

Is a 16 year old named Quinn more likely to be a boy or a girl?

```
# filter the data
babiesNamedQuinn16 <- babynames %>%
  filter(name == "Quinn", year == (year(now()))-16) %>%
  mutate(sex = recode(sex, "F" = "Female", "M" = "Male"))
# show the data
head(babiesNamedQuinn16)
```

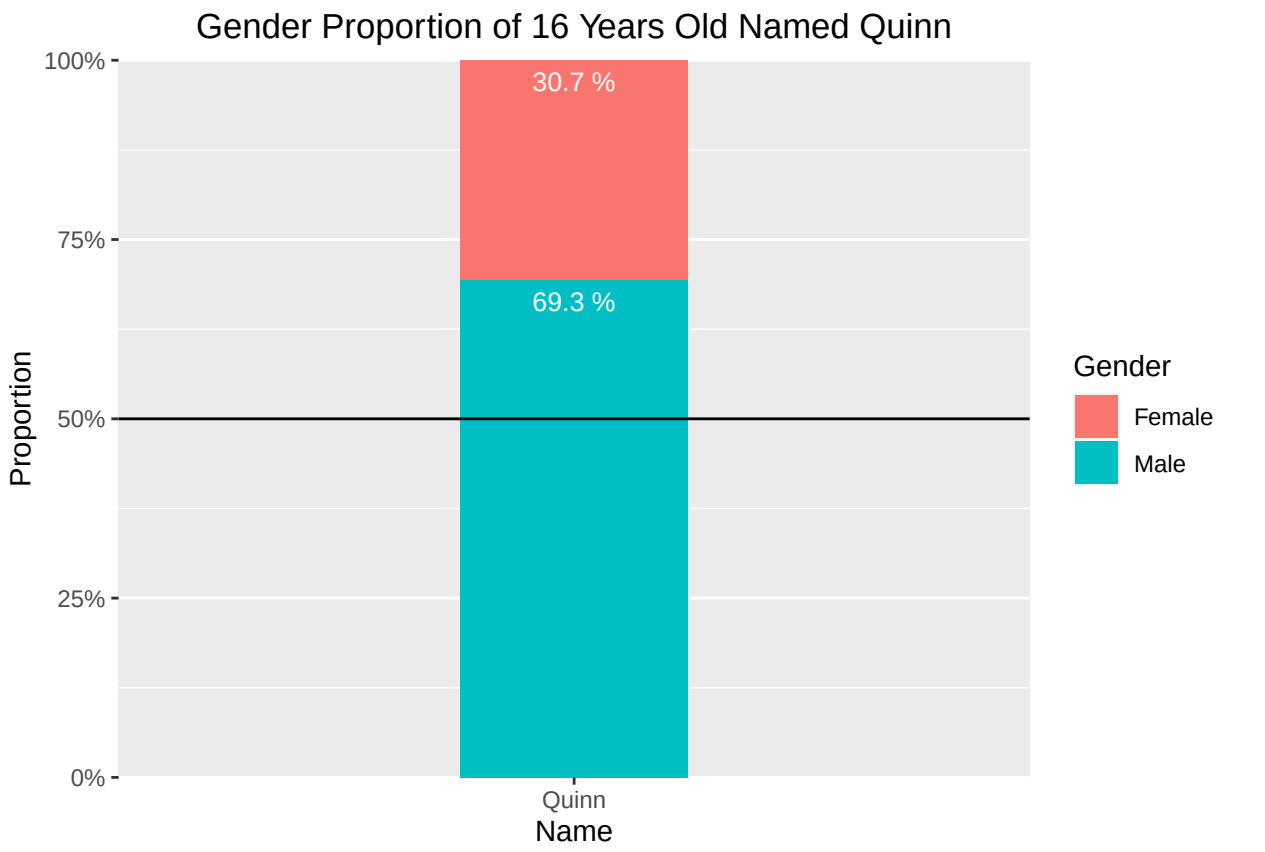
```
## # A tibble: 2 x 5
##    year sex    name      n     prop
##   <dbl> <chr>  <chr> <int>    <dbl>
## 1  2006 Female Quinn   545 0.000261
## 2  2006 Male   Quinn  1233 0.000563
```

```
# visualization
# compute the position of geom_text
label_data = babiesNamedQuinn16 %>%
```

```
  arrange(name, desc(sex)) %>%
  mutate(ylabel_pos = cumsum(n) / sum(n), ylabel = n / sum(n)) %>%
  group_by(sex, add = TRUE) %>%
  mutate(ylabel = sum(ylabel)) %>%
  slice(n())
```

```
## Warning: The `add` argument of `group_by()` is deprecated as of dplyr 1.0.0.
## Please use the `.add` argument instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was generated.
```

```
# draw the graph
ggplot(babiesNamedQuinn16, aes(x = name, y = n, fill = sex)) +
  geom_bar(stat = "identity", position = "fill", width = 0.3) +
  geom_text(data = label_data,
            aes(
              y = ylabel_pos,
              label = paste(round(ylabel * 100, 1), "%")),
            vjust = 1.6,
            color = "white",
            size = 3.5) +
  geom_hline(yintercept = 0.5) +
  scale_y_continuous(labels = scales::percent, expand = c(0, 0)) +
  ggtitle("Gender Proportion of 16 Years Old Named Quinn") +
  labs(x = "Name", y = "Proportion", fill = "Gender") +
  theme(plot.title = element_text(hjust = 0.5))
```



Based on the graph above, we can see that about 69.3% of 16-year-old children named Quinn are male. This

shows that a 16-year-old child named Quinn is more likely to be a boy.

Is a 2 year old named Quinn more likely to be a boy or a girl?

```
babiesNamedQuinn2 <- filter(babynames, name == "Quinn",
                            year == (year(now())) - 2) # filter the data
head(babiesNamedQuinn2) # show the data
```

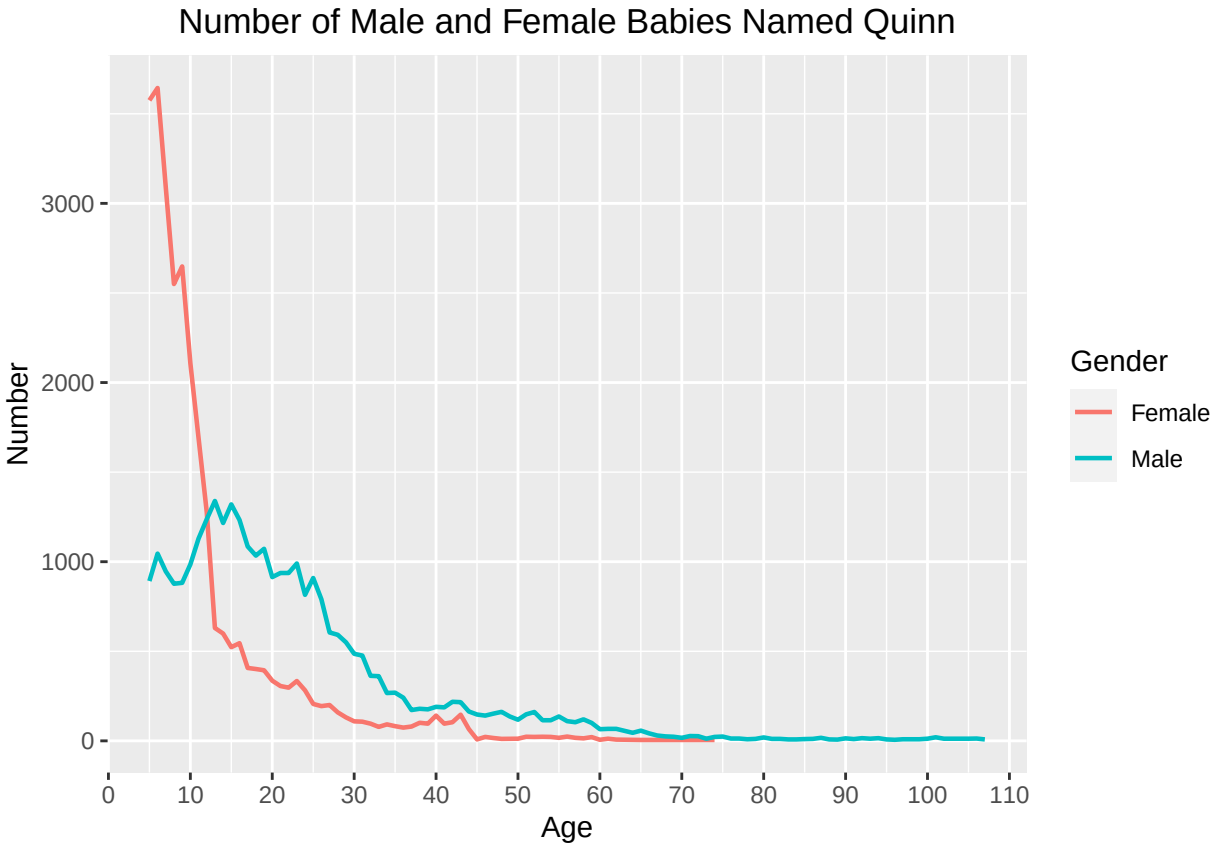
```
## # A tibble: 0 x 5
## # ... with 5 variables: year <dbl>, sex <chr>, name <chr>, n <int>, prop <dbl>
```

Based on the above result, we see that there is no data for 2-year-old babies named Quinn. So instead of making guess based on this data, we can try to find some clues on the whole data by executing the code below.

```
# filter the data
babiesNamedQuinn <- babynames %>%
  filter(name=="Quinn") %>%
  mutate(sex = recode(sex, "F"="Female", "M" = "Male"))
# show the data
head(babiesNamedQuinn)
```

```
## # A tibble: 6 x 5
##    year sex   name      n       prop
##   <dbl> <chr> <chr> <int>      <dbl>
## 1  1915 Male  Quinn     8 0.00000908
## 2  1916 Male  Quinn    13 0.0000141
## 3  1917 Male  Quinn    12 0.0000125
## 4  1918 Male  Quinn    12 0.0000114
## 5  1919 Male  Quinn    12 0.0000118
## 6  1920 Male  Quinn    12 0.0000109
```

```
# plot
ggplot(babiesNamedQuinn) +
  geom_line(mapping = aes((year(now()) - year), n, color=sex), size = 0.8) +
  scale_x_continuous(breaks=seq(0, 110, 10)) +
  labs(x = "Age", y = "Number", color = "Gender") +
  ggtitle("Number of Male and Female Babies Named Quinn") +
  theme(plot.title = element_text(hjust = 0.5))
```



Based on the above graph, we can see that there is an increasing trend that for babies named Quinn under 15-year-old, female babies named Quinn are taking more and more proportion than male babies named Quinn as their ages fall. So based on this trend, I think a 2-year-old named Quinn is more likely to be a girl.

What is your best guess as to how old a woman named Susan is?

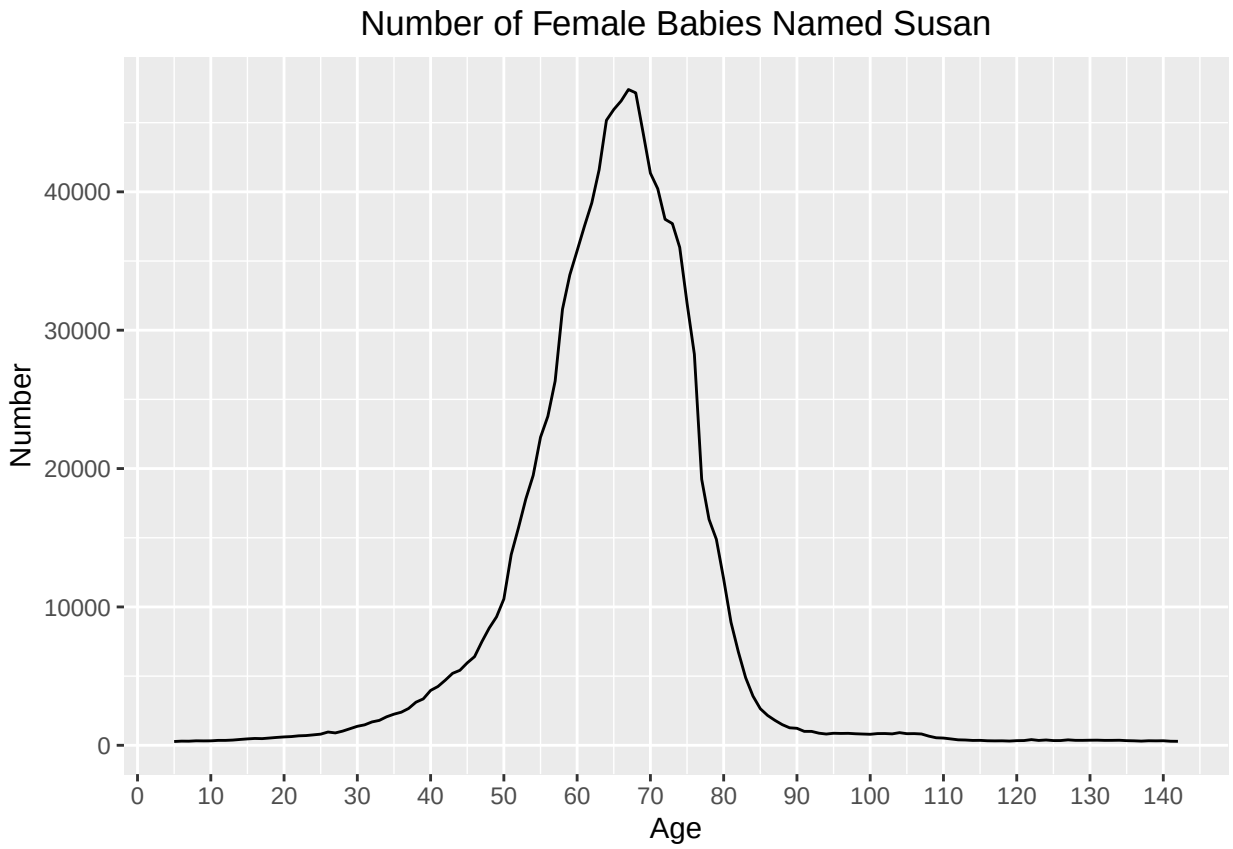
```
# filter the data
femaleBabiesNamedSusan <- babynames %>%
  filter(name=="Susan", sex=="F")

 # show the data
head(femaleBabiesNamedSusan)
```

```
## # A tibble: 6 x 5
##    year sex   name      n    prop
##   <dbl> <chr> <chr> <int>   <dbl>
## 1  1880 F     Susan   286 0.00293
## 2  1881 F     Susan   292 0.00295
## 3  1882 F     Susan   326 0.00282
## 4  1883 F     Susan   322 0.00268
## 5  1884 F     Susan   326 0.00237
## 6  1885 F     Susan   302 0.00213
```

```
# plot
ggplot(femaleBabiesNamedSusan) +
  geom_line(mapping=aes(x = (year(now()) - year), y = n)) +
  scale_x_continuous(breaks=seq(0, 150, 10)) +
  labs(
```

```
    x = "Age",
    y = "Number",
    title = "Number of Female Babies Named Susan",
    color = "Gender") +
  theme(plot.title = element_text(hjust = 0.5))
```




```
# show the sorted data according to the amount of babies in descending order
femaleBabiesNamedSusan <- femaleBabiesNamedSusan %>%
  mutate(age = year(now()) - year)
arrange(femaleBabiesNamedSusan, desc(n))
```

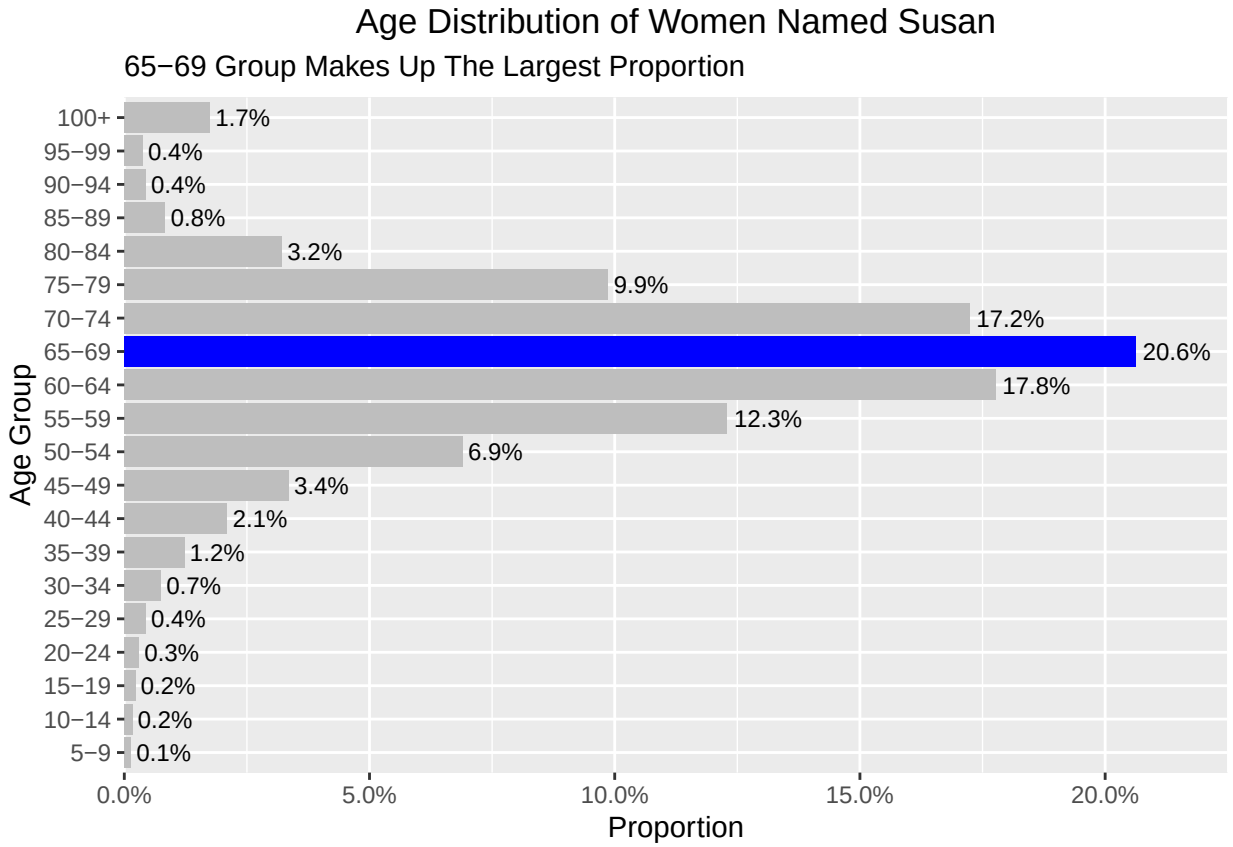
```
## # A tibble: 138 x 6
##     year sex   name      n   prop   age
##    <dbl> <chr> <chr> <int>  <dbl> <dbl>
##  1  1955 F     Susan 47397 0.0236    67
##  2  1954 F     Susan 47158 0.0237    68
##  3  1956 F     Susan 46567 0.0226    66
##  4  1957 F     Susan 45951 0.0219    65
##  5  1958 F     Susan 45172 0.0219    64
##  6  1953 F     Susan 44285 0.0230    69
##  7  1959 F     Susan 41598 0.0200    63
##  8  1952 F     Susan 41350 0.0217    70
##  9  1951 F     Susan 40227 0.0218    71
## 10  1960 F     Susan 39200 0.0188    62
## # ... with 128 more rows
```

```
# creating age groups
ageBreaks <- seq(0, 100, 5)
```

```r
ageLabels <- c()
for(age in ageBreaks){
  if(age == 100){
    ageBreaks <- append(ageBreaks, 200)
    ageLabels <- append(ageLabels, paste(100, "+", sep = ""))
    break
  }
  ageLabels <- append(ageLabels, paste(age, age + 4, sep = "-"))
}

femaleBabiesNamedSusan <- femaleBabiesNamedSusan %>%
  mutate(ageGroup = cut((year(now()) - year), breaks = ageBreaks, right = F, labels = ageLabels)) %>%
  group_by(ageGroup) %>%
  summarise(n = sum(n)) %>%
  mutate(max_prop = ifelse(n == max(n), "1", "0"))


ggplot(femaleBabiesNamedSusan, aes(x = ageGroup, y = n / sum(n), fill = max_prop)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = paste(round(n / sum(n) * 100, 1), "%", sep = "")), hjust = -0.1, size = 3) +
  scale_y_continuous(labels = scales::percent, expand = c(0, 0), lim = c(0, 0.225)) +
  scale_fill_manual(values = c("1" = "blue", "0" = "grey"), guide = "none") +
  labs(
    x = "Age Group",
    y = "Proportion",
    title = "Age Distribution of Women Named Susan",
    subtitle = "65-69 Group Makes Up The Largest Proportion") +
  coord_flip() +
  theme(plot.title = element_text(hjust = 0.5))
```



Based on the results above, we can see that women named Susan with an age between 65 to 69 make up the largest proportion, 20.6%, of all women named Susan. Based on these data, we can make a best guess that the age of a woman named Susan is more likely around 65 to 69, or more precisely, 67 according to the sorted data above.