Homework 1

Ola Tranum Arnegaard - 10/08/2019

Exercise 3.1

Intro

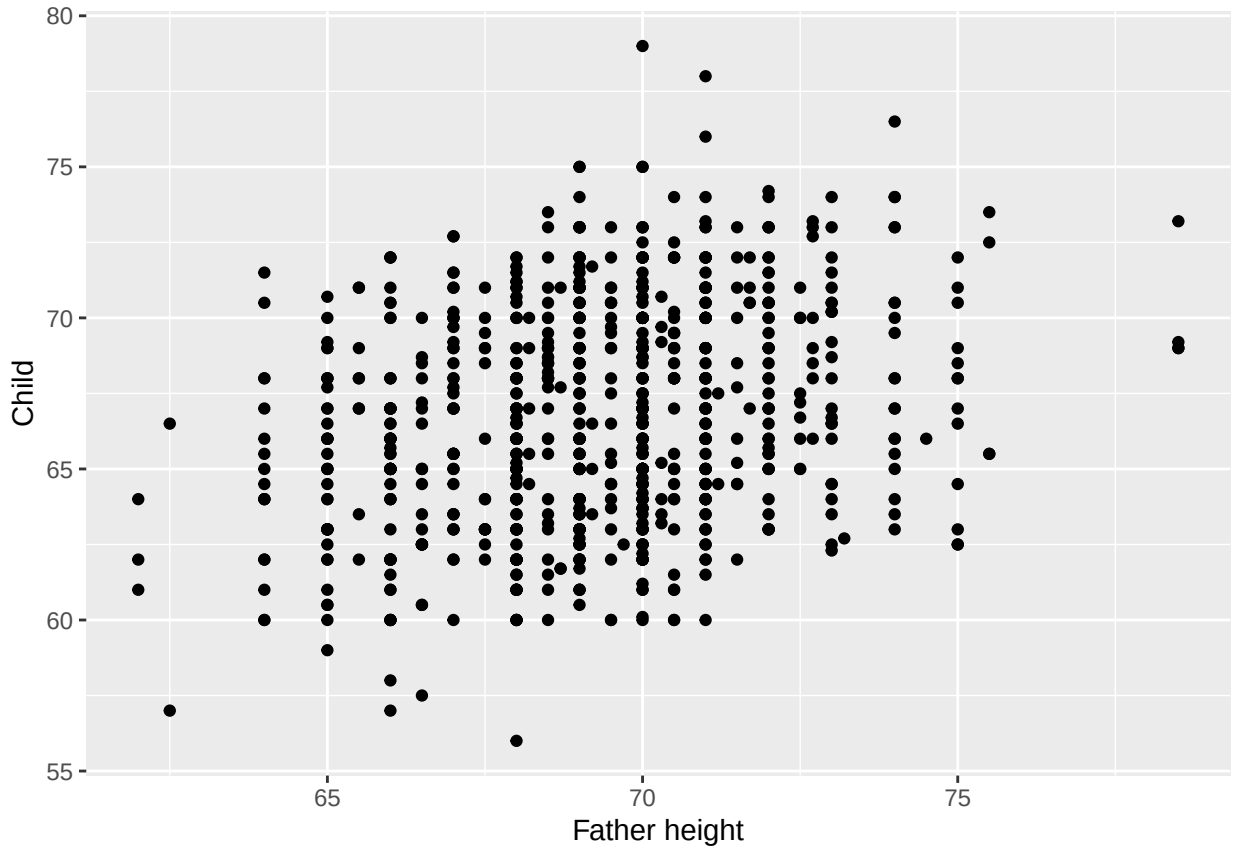
Using the famous Galton data set from the mosaicData package:

```
head(Galton)
```

```
##   family father mother sex height nkids
## 1      1   78.5   67.0   M   73.2     4
## 2      1   78.5   67.0   F   69.2     4
## 3      1   78.5   67.0   F   69.0     4
## 4      1   78.5   67.0   F   69.0     4
## 5      2   75.5   66.5   M   73.5     4
## 6      2   75.5   66.5   M   72.5     4
```

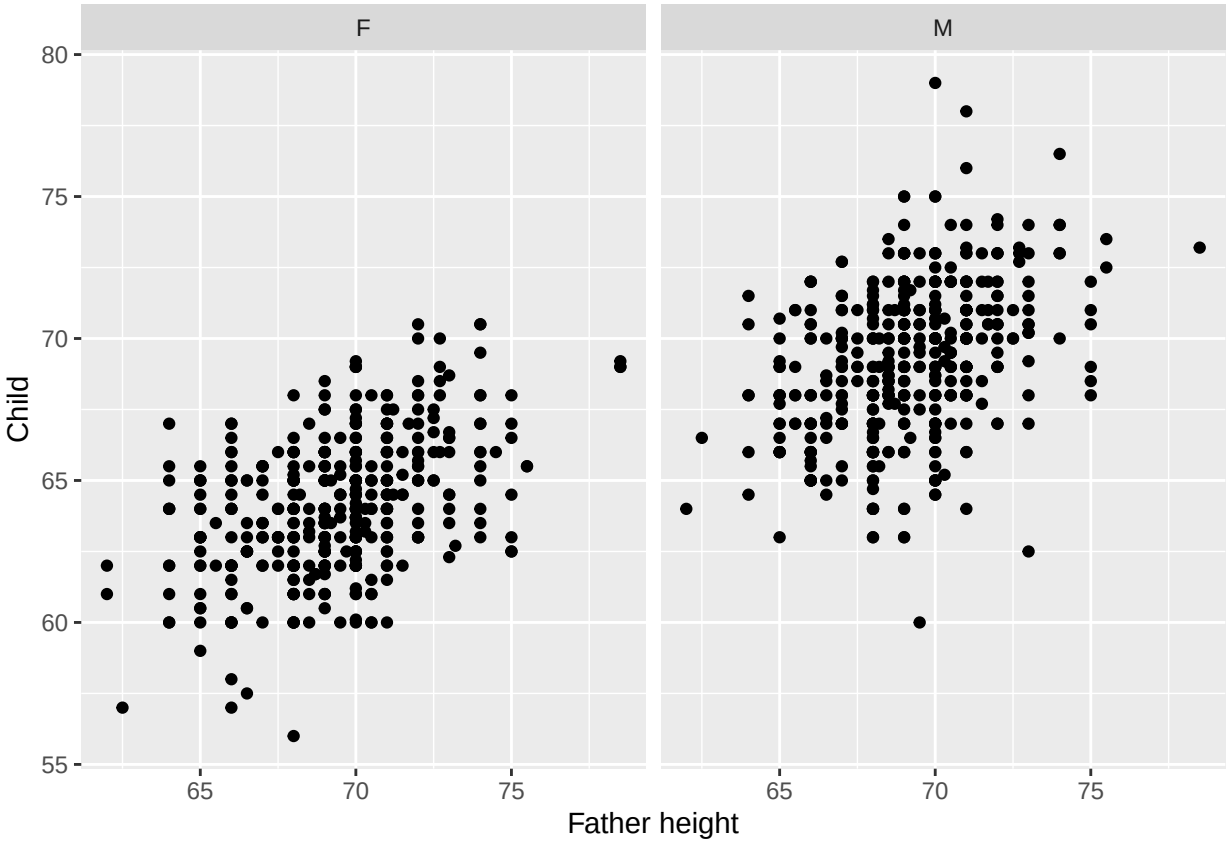
1. Create a scatterplot of each person's height against their father's height

```
plot1 <- ggplot(Galton, aes(x = Galton$father, y = Galton$height)) + geom_point()
plot1 <- plot1 + xlab("Father height") + ylab("Child")
plot1
```



2. Separate your plot into facets by sex

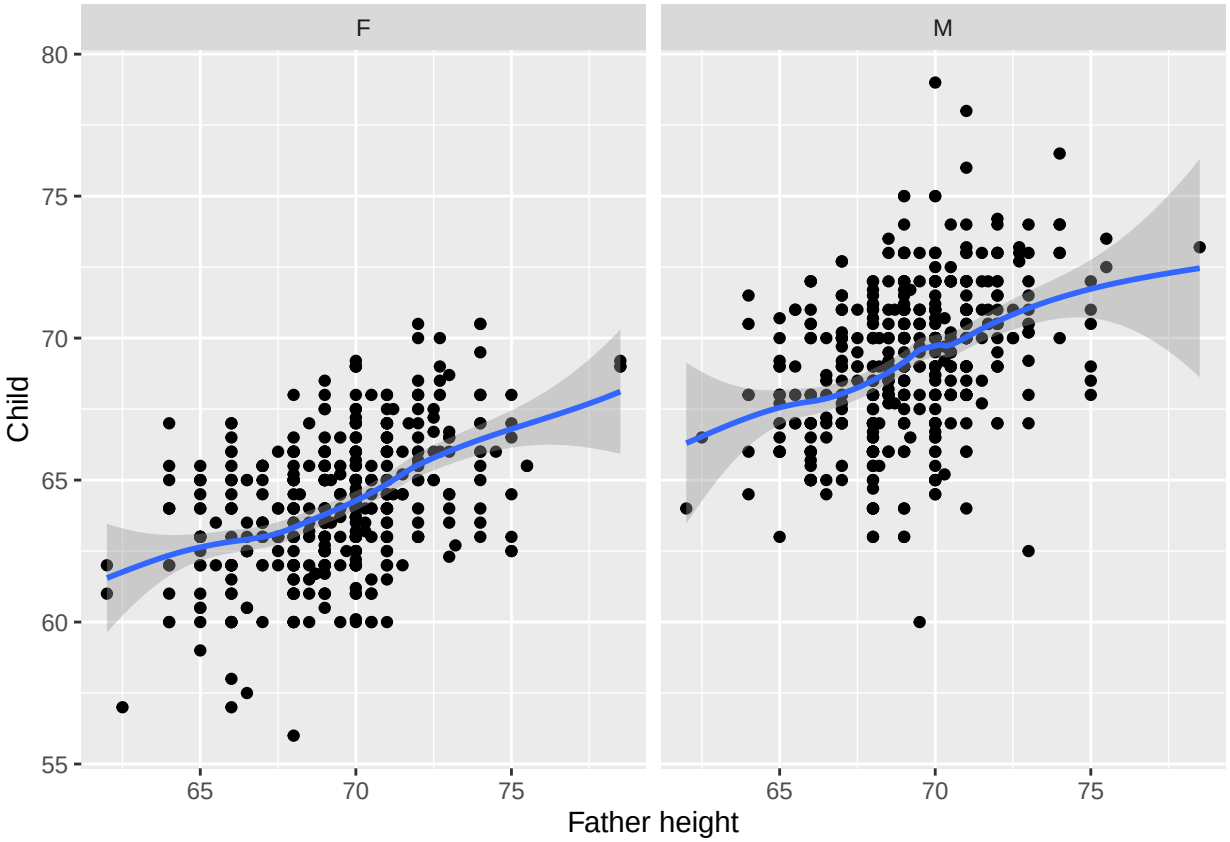
```
plot1 <- plot1 + facet_wrap(Galton$sex,nrow=1,ncol=2)
plot1
```



3. Add regression lines to all of your facets

```
plot1 <- plot1 + geom_smooth(method = 'auto')
plot1
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Exercise 3.2

Intro

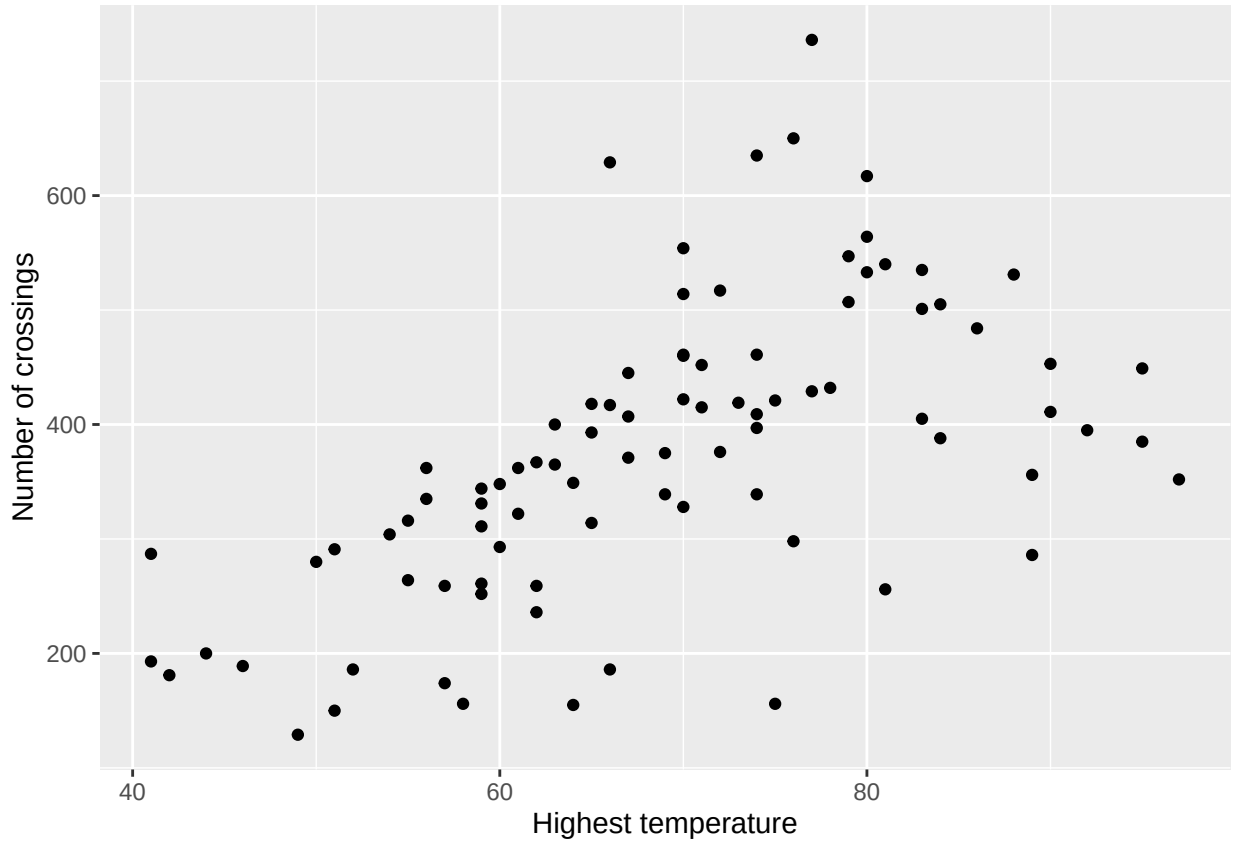
Using the RailTrail data set from the mosaicData package:

```
head(RailTrail)
```

```
##   hightemp lowtemp avgtemp spring summer fall cloudcover precip volume
## 1       83      50    66.5      0      1    0        7.6   0.00    501
## 2       73      49    61.0      0      1    0        6.3   0.29    419
## 3       74      52    63.0      1      0    0        7.5   0.32    397
## 4       95      61    78.0      0      1    0        2.6   0.00    385
## 5       44      52    48.0      1      0    0       10.0   0.14    200
## 6       69      54    61.5      1      0    0        6.6   0.02    375
##   weekday dayType
## 1    TRUE weekday
## 2    TRUE weekday
## 3    TRUE weekday
## 4   FALSE weekend
## 5    TRUE weekday
## 6    TRUE weekday
```

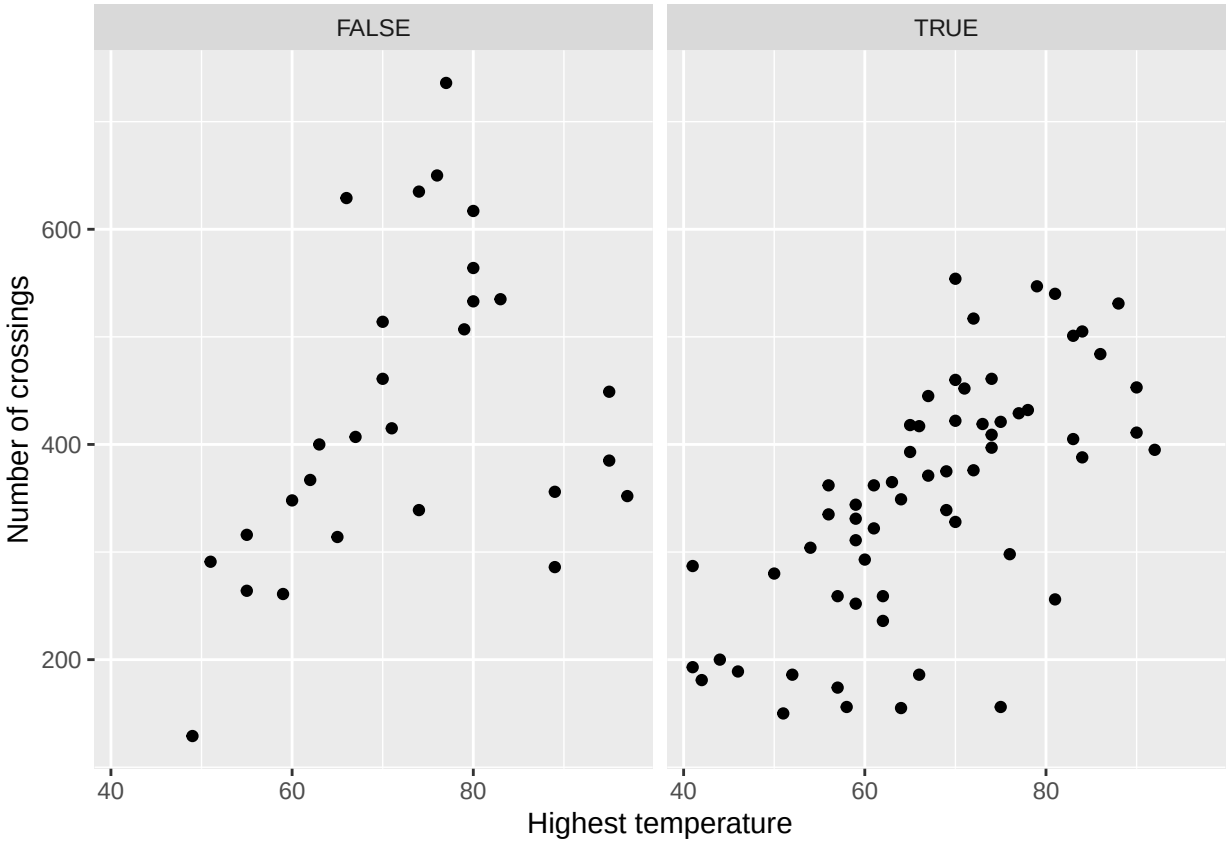
1. Create a scatterplot of the number of crossings per day volume against the high temperature that day

```
plot2 <- ggplot(RailTrail, aes(x = RailTrail$hightemp, y = RailTrail$volume)) + geom_point()
plot2 <- plot2 + xlab("Highest temperature") + ylab("Number of crossings")
plot2
```



2. Separate your plot into facets by `weekday`

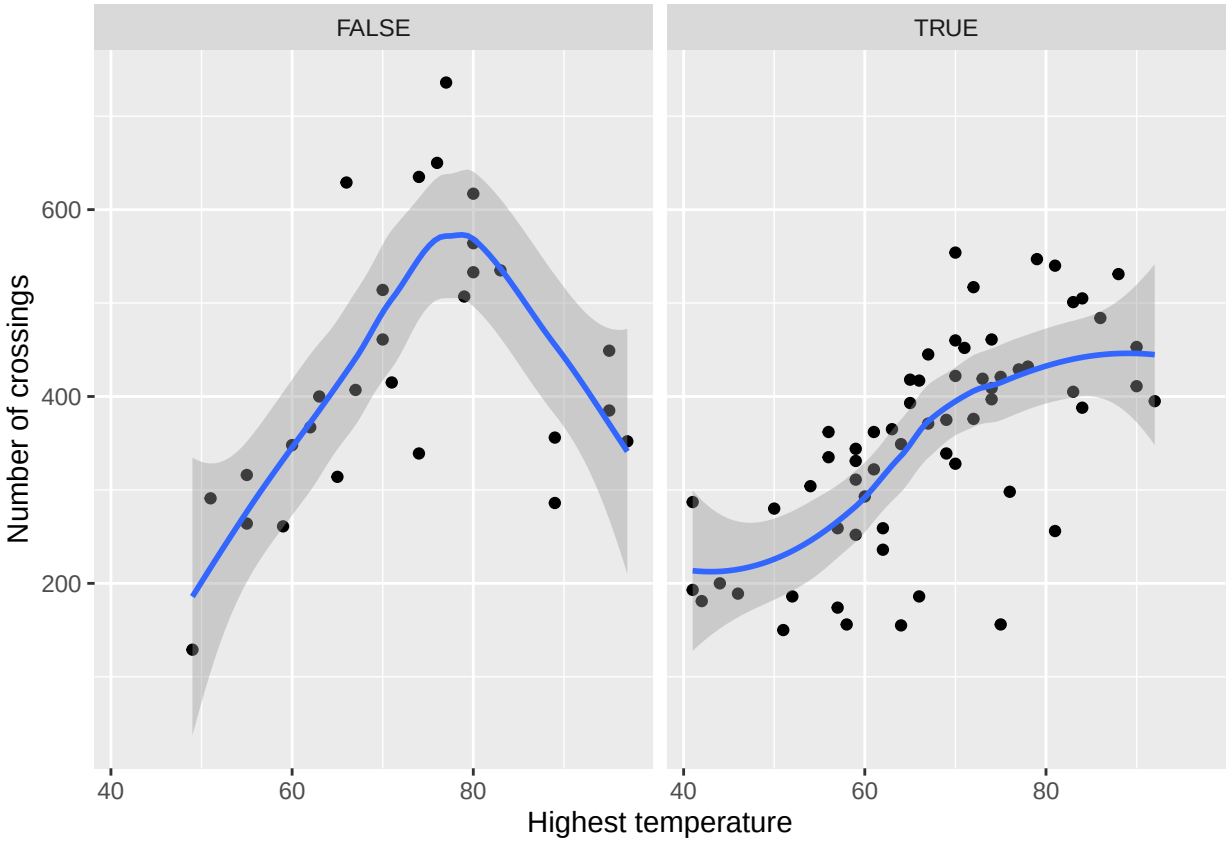
```
plot2 <- plot2 + facet_wrap(RailTrail$weekday,nrow=1,ncol=2)
plot2
```



3. Add regression lines to all of your facets

```
plot2 <- plot2 + geom_smooth(method = 'auto')
plot2
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Exercise 3.3

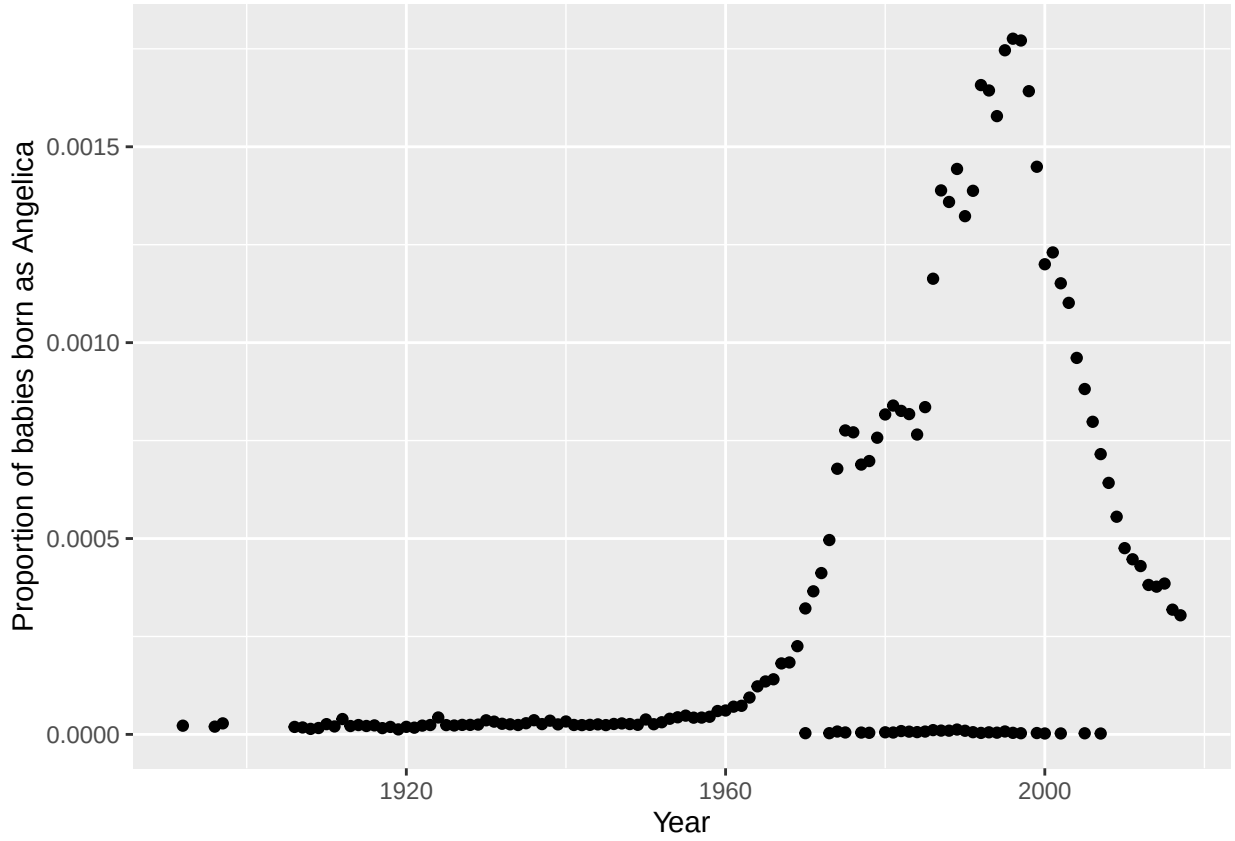
Angelica Schuyler Church (1756{1814) was the daughter of New York Governer Philip Schuyler and sister of Elizabeth Schuyler Hamilton. Angelica, New York was named after her. Generate a plot of the reported **proportion of babies born with the name Angelica over time and interpret the figure.**

```
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
```

```
df <- as.data.frame(c(babynames[babynames$name=='Angelica','year'],babynames[babynames$name=='Angelica'

plot3 <- ggplot(df, aes(x = df$year, y = df$prop)) + geom_point()
plot3 <- plot3 + xlab("Year") + ylab("Proportion of babies born as Angelica")
plot3
```



Result You see that it is a clear rise in the proportion of babies named Angelica throughout the second half of the ninenteenth century. However, throughout the 2000's the named have experienced a downfall.

Exercise 3.4

Intro

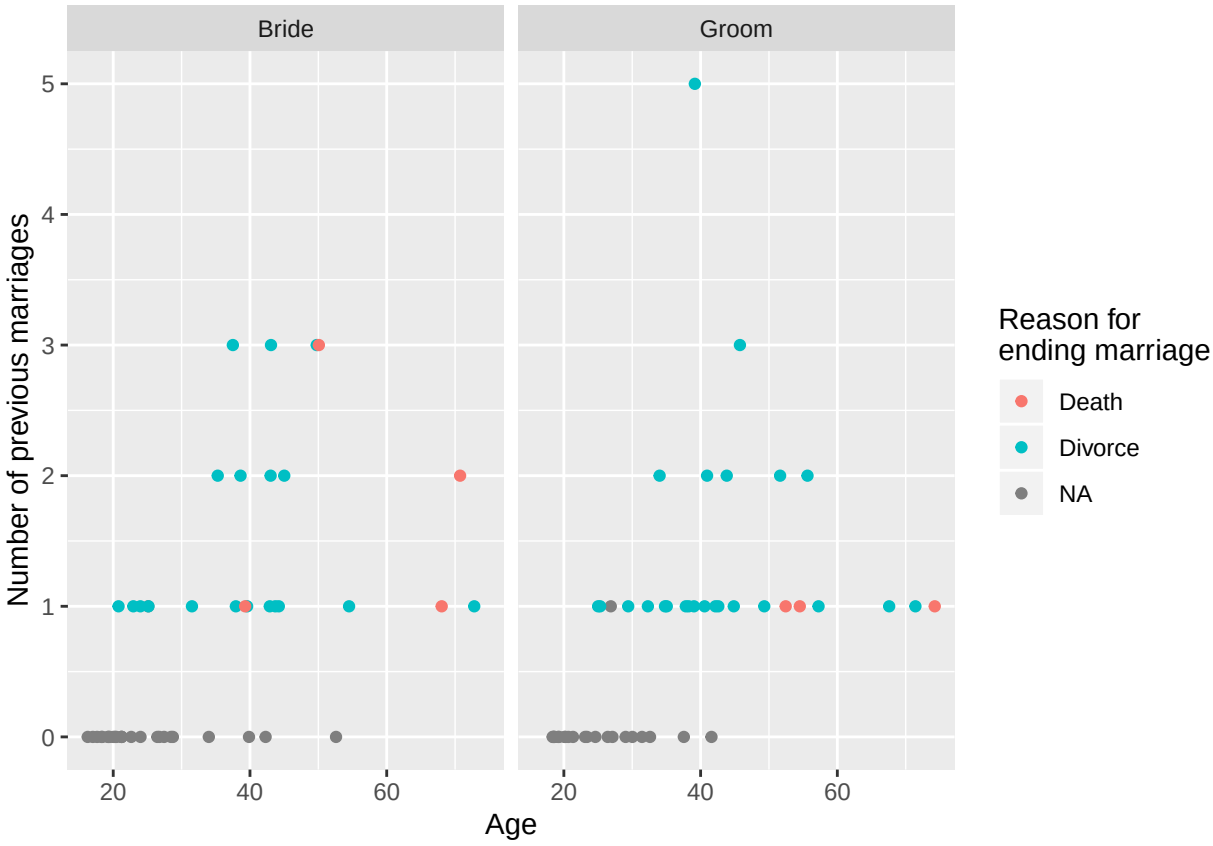
The following questions use the Marriage data set from the mosaicData package.

```
head(Marriage, 4)
```

```
##   bookpageID  appdate ceremonydate delay     officialTitle person     dob
## 1   B230p539 10/29/96      11/9/96    11    CIRCUIT JUDGE   Groom 4/11/64
## 2   B230p677 11/12/96     11/12/96     0 MARRIAGE OFFICIAL  Groom  8/6/64
## 3   B230p766 11/19/96     11/27/96     8 MARRIAGE OFFICIAL  Groom 2/20/62
## 4   B230p892  12/2/96      12/7/96     5          MINISTER  Groom 5/20/56
##        age     race prevcount prevconc hs college dayOfBirth   sign
## 1 32.60274    White         0     <NA> 12       7      102.0  Aries
## 2 32.29041    White         1  Divorce 12       0      219.0    Leo
## 3 34.79178 Hispanic         1  Divorce 12       3       51.5 Pisces
## 4 40.57808    Black         1  Divorce 12       4      141.0 Gemini
```

1. Create an informative and meaningful data graphic.

```
plot4 <- ggplot(Marriage, aes(x = Marriage$age, y = Marriage$prevcount)) + geom_point(aes(color = prevc
plot4 <- plot4 + xlab("Age") + ylab("Number of previous marriages")
plot4
```



2. Identify each of the visual cues that you are using, and describe how they are related to each variable.

- Colors are used as a visual for the feature `reason for ending marriage`
- Seperation of plot is used to seperate `brides` and `grooms`
- Scatter plot is used to compare `previous marriages` and `age`

3. Create a data graphic with at least five variables (either quantitative or categorical). For the purposes of this exercise, do not worry about making your visualization meaningful|just try to encode five variables into one plot.

```
plot4 + geom_point(size = 1.5, aes(color = prevconc,shape = race))
```