hw1

Homework #1

Exercise 3.1

Using the famous Galton data set from the mosaicData package:

```r
library(mosaic)
```

```
## Loading required package: 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

## Loading required package: lattice

## Loading required package: ggformula

## Loading required package: ggplot2

## Loading required package: ggstance

##
## Attaching package: 'ggstance'

## The following objects are masked from 'package:ggplot2':
##
##     geom_errorbarh, GeomErrorbarh

##
## New to ggformula?  Try the tutorials:
##  learnr::run_tutorial("introduction", package = "ggformula")
##  learnr::run_tutorial("refining", package = "ggformula")

## Loading required package: mosaicData

## Loading required package: Matrix
## Registered S3 method overwritten by 'mosaic':
##   method                           from
##   fortify.SpatialPolygonsDataFrame ggplot2
```

```
##
## The 'mosaic' package masks several functions from core packages in order to add
## additional features.  The original behavior of these functions should not be affected by this.
##
## Note: If you use the Matrix package, be sure to load it BEFORE loading mosaic.
```

```
##
## Attaching package: 'mosaic'
```

```
## The following object is masked from 'package:Matrix':
##
##     mean
```

```
## The following object is masked from 'package:ggplot2':
##
##     stat
```

```
## The following objects are masked from 'package:dplyr':
##
##     count, do, tally
```

```
## The following objects are masked from 'package:stats':
##
##     binom.test, cor, cor.test, cov, fivenum, IQR, median,
##     prop.test, quantile, sd, t.test, var
```

```
## The following objects are masked from 'package:base':
##
##     max, mean, min, prod, range, sample, sum
```

```r
head(Galton, n=5)
```

```
##   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
```

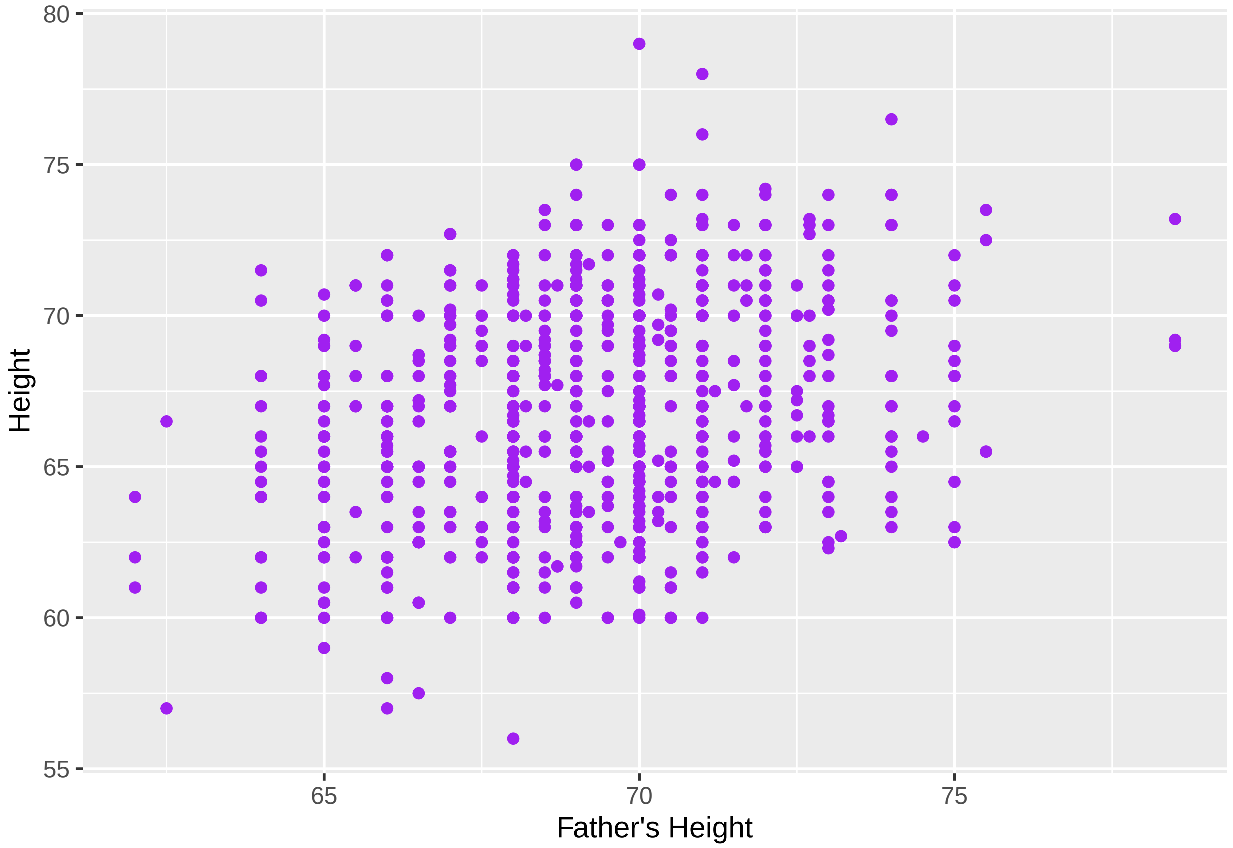
```r
summary(Galton)
```

```
##      family        father          mother      sex         height
##  185    : 15   Min.   :62.00   Min.   :58.00   F:433   Min.   :56.00
##  166    : 11   1st Qu.:68.00   1st Qu.:63.00   M:465   1st Qu.:64.00
##  66     : 11   Median :69.00   Median :64.00           Median :66.50
##  130    : 10   Mean   :69.23   Mean   :64.08           Mean   :66.76
##  136    : 10   3rd Qu.:71.00   3rd Qu.:65.50           3rd Qu.:69.70
##  140    : 10   Max.   :78.50   Max.   :70.50           Max.   :79.00
##  (Other):831
##      nkids
##  Min.   : 1.000
##  1st Qu.: 4.000
##  Median : 6.000
##  Mean   : 6.136
##  3rd Qu.: 8.000
##  Max.   :15.000
##
```

```
?Galton
```

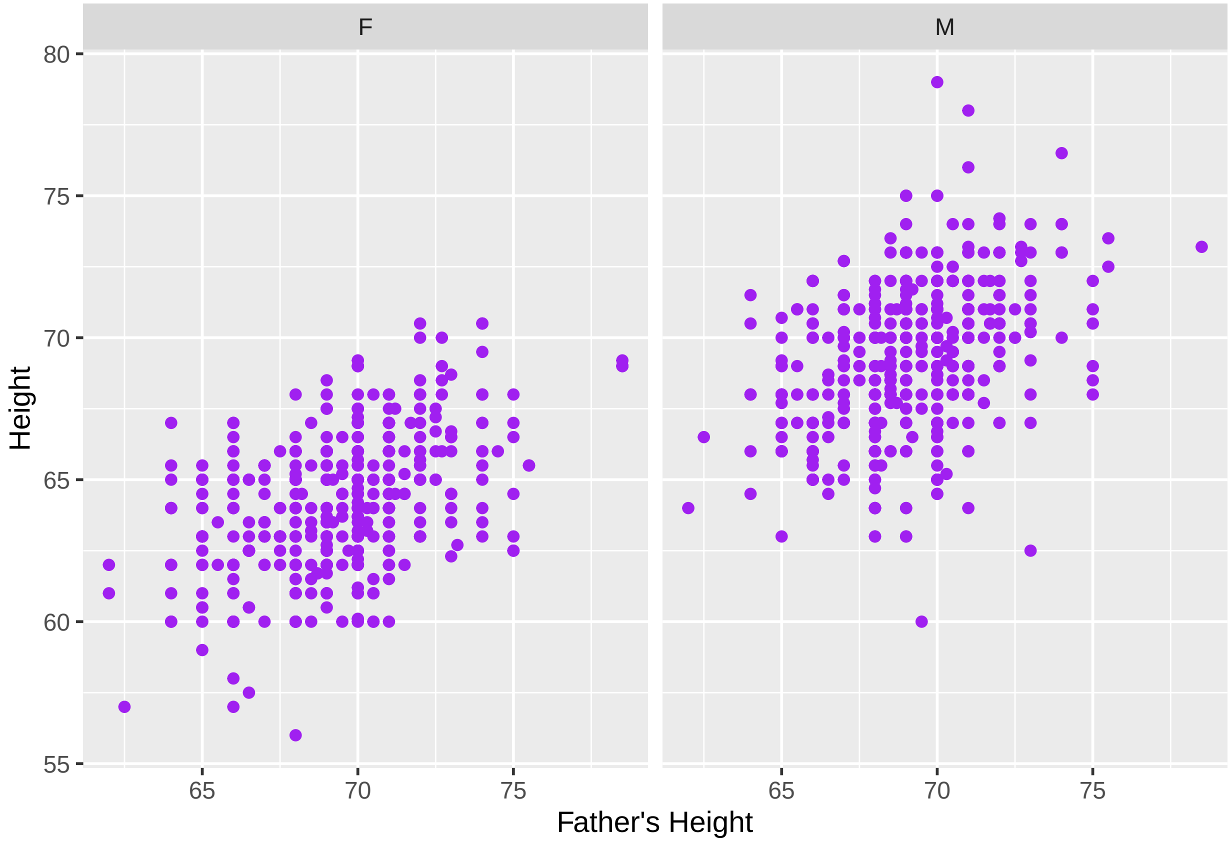
1. Create a scatterplot of each person's height against their father's height

```
ggplot(data = Galton, aes(x = father, y = height)) + geom_point(colour="purple") + xlab("Father's Heigh
```



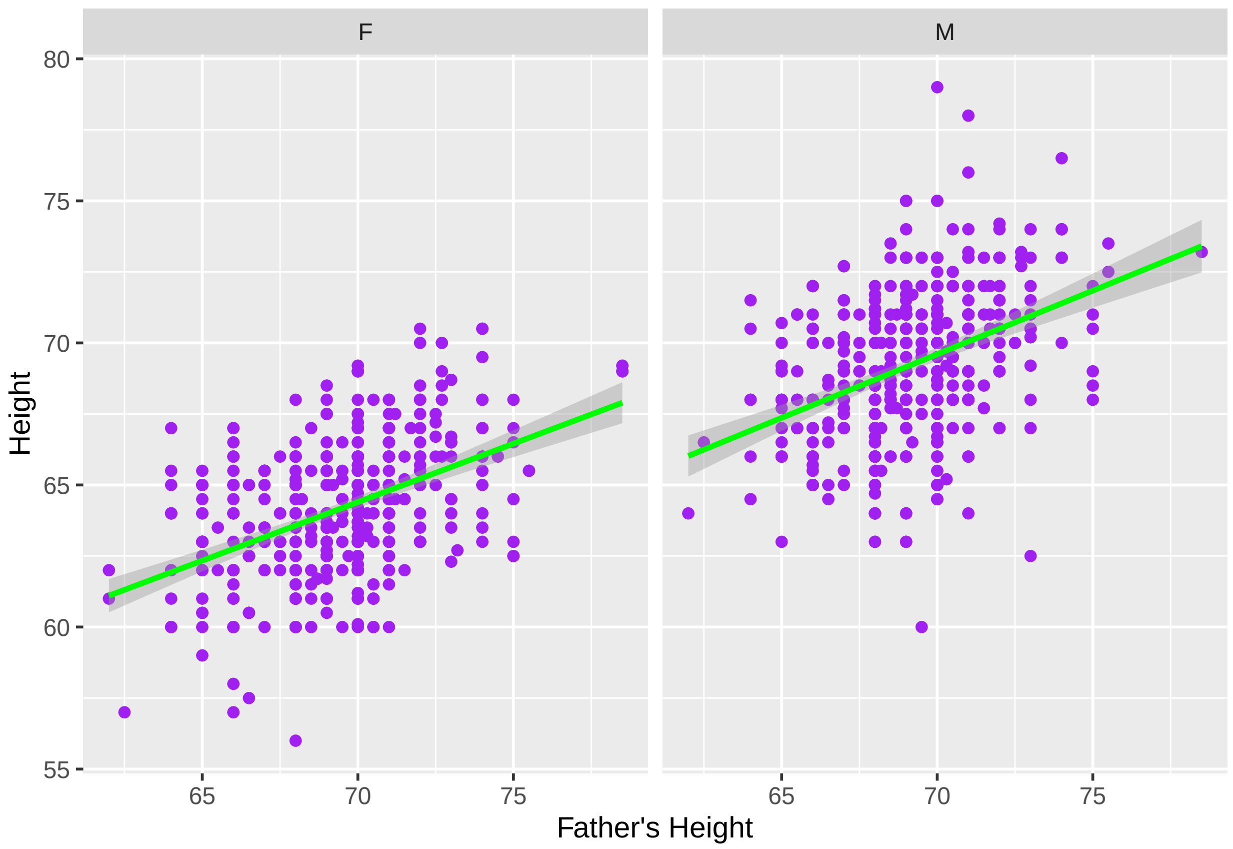
2. Separate your plot into facets by sex

```
ggplot(data = Galton, aes(x = father, y = height)) + geom_point(colour="purple") + xlab("Father's Heigh
```



3. Add regression lines to all of your facets

```
ggplot(data = Galton, aes(x = father, y = height)) + geom_point(colour="purple") + xlab("Father's Heigh
```



Exercise 3.2

Using the RailTrail data set from the mosaicData package:

```
head(RailTrail, n=5)
```

```
##   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
##   weekday dayType
## 1    TRUE weekday
## 2    TRUE weekday
## 3    TRUE weekday
## 4   FALSE weekend
## 5    TRUE weekday
```

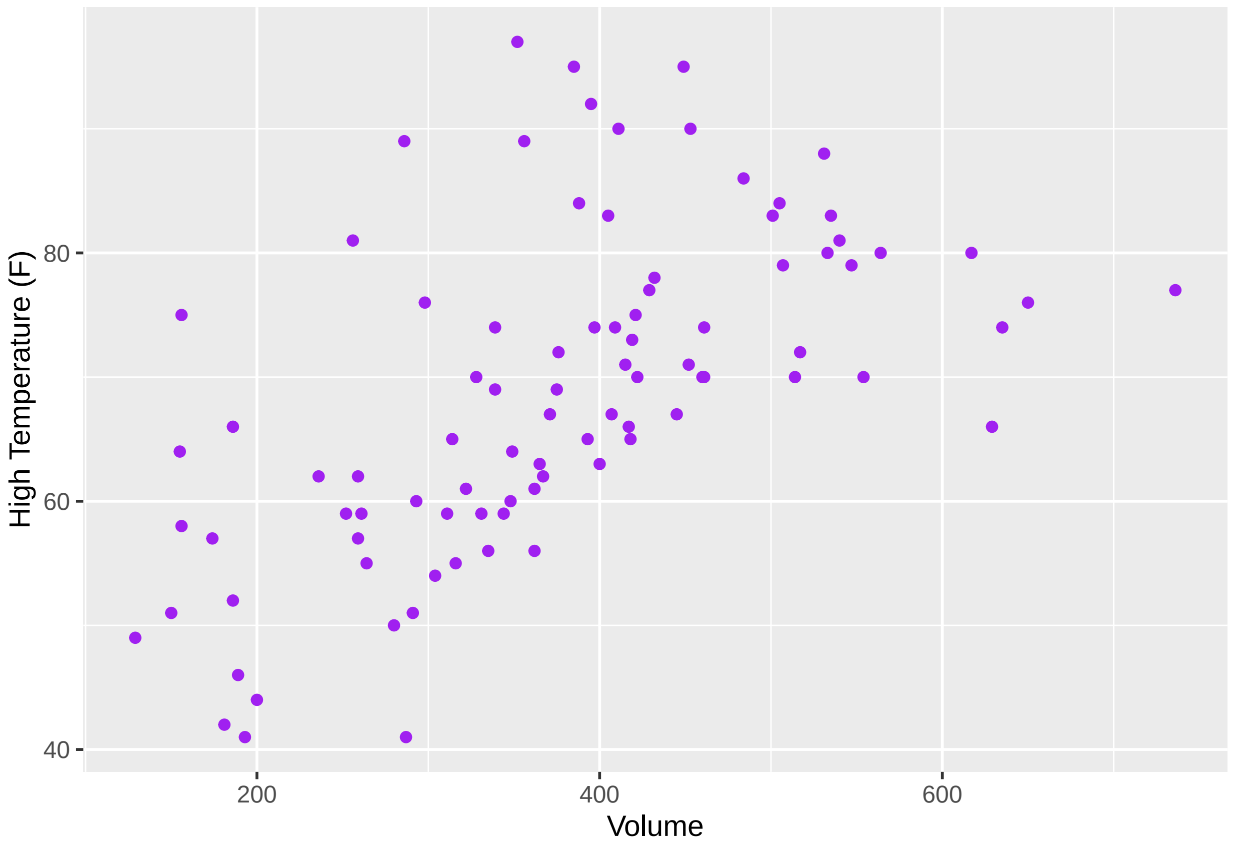
```
summary(RailTrail)
```

```
##     hightemp        lowtemp         avgtemp          spring
##  Min.   :41.00   Min.   :19.00   Min.   :33.00   Min.   :0.0000
##  1st Qu.:59.25   1st Qu.:38.00   1st Qu.:48.62   1st Qu.:0.0000
##  Median :69.50   Median :44.50   Median :55.25   Median :1.0000  
##  Mean   :68.83   Mean   :46.03   Mean   :57.43   Mean   :0.5889  
##  3rd Qu.:77.75   3rd Qu.:53.75   3rd Qu.:64.50   3rd Qu.:1.0000  
##  Max.   :97.00   Max.   :72.00   Max.   :84.00   Max.   :1.0000  
##      summer             fall          cloudcover         precip       
##  Min.   :0.0000   Min.   :0.0000   Min.   : 0.000   Min.   :0.00000  
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 3.650   1st Qu.:0.00000  
##  Median :0.0000   Median :0.0000   Median : 6.400   Median :0.00000  
##  Mean   :0.2778   Mean   :0.1333   Mean   : 5.807   Mean   :0.09256  
##  3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.: 8.475   3rd Qu.:0.02000  
##  Max.   :1.0000   Max.   :1.0000   Max.   :10.000   Max.   :1.49000  
##      volume        weekday          dayType         
##  Min.   :129.0   Mode :logical   Length:90         
##  1st Qu.:291.5   FALSE:28        Class :character  
##  Median :373.0   TRUE :62        Mode  :character  
##  Mean   :375.4                                     
##  3rd Qu.:451.2                                     
##  Max.   :736.0                                     
```

```
?RailTrail
```

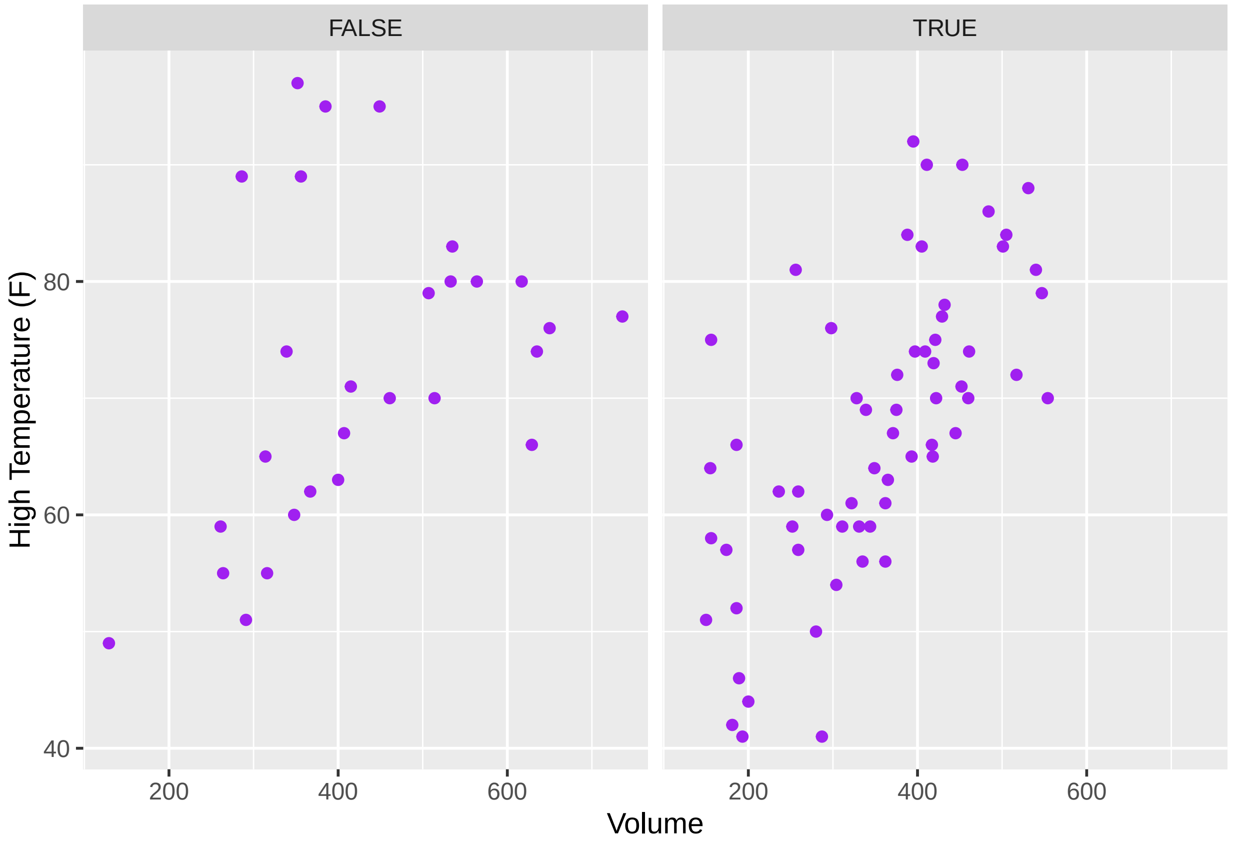
1. Create a scatterplot of the number of crossings per day volume against the high temperature that day

```
ggplot(data = RailTrail, aes(x = volume, y = hightemp)) + geom_point(colour="purple") + xlab("Volume") +
```



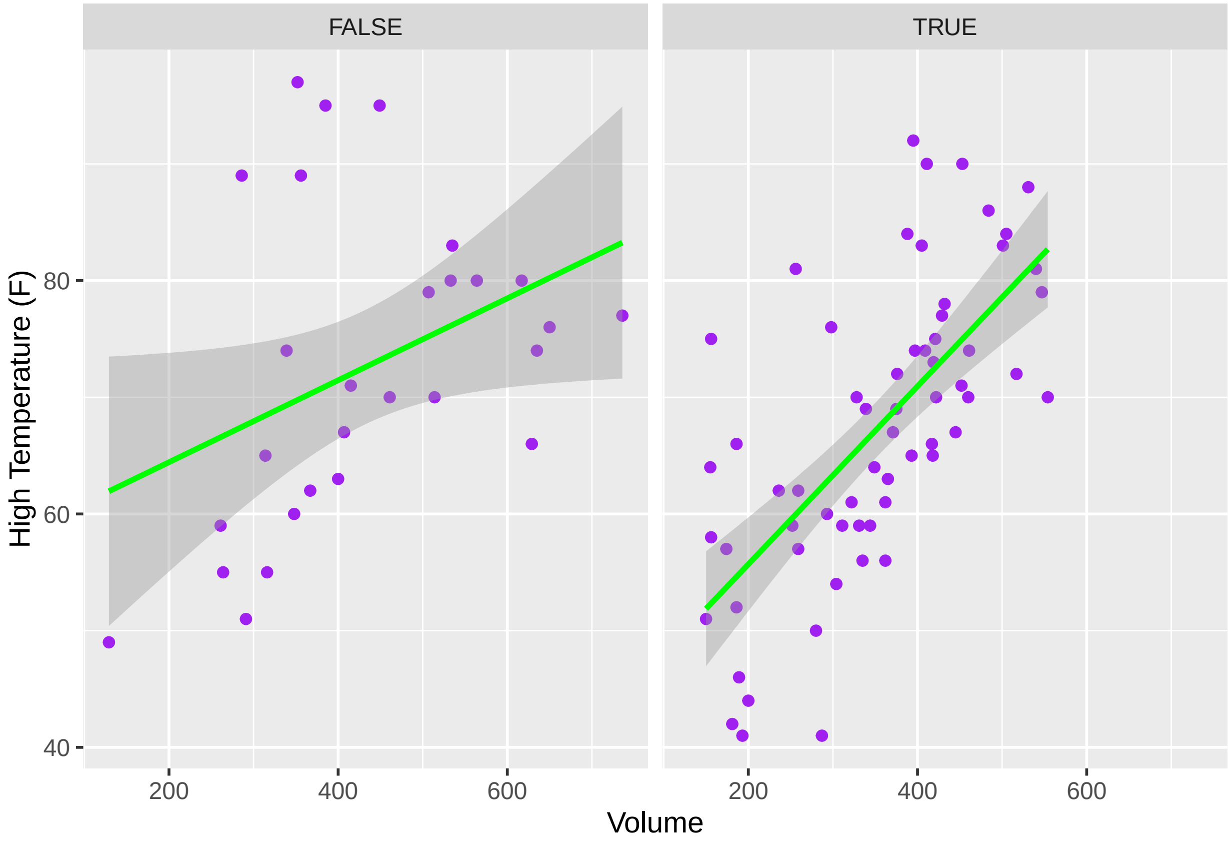
2. Separate your plot into facets by weekday

```
ggplot(data = RailTrail, aes(x = volume, y = hightemp)) + geom_point(colour="purple") + xlab("Volume") +
```



3. Add regression lines to the two facets

```
ggplot(data = RailTrail, aes(x = volume, y = hightemp)) + geom_point(colour="purple") + xlab("Volume") +
```



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.

```
library(babynames)
head(babynames, n=5)
```

```
## # A tibble: 5 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
```

```
summary(babynames)
```

```
##       year          sex                name                 n        
##  Min.   :1880   Length:1924665     Length:1924665     Min.   :    5.0  
##  1st Qu.:1951   Class :character   Class :character   1st Qu.:    7.0  
##  Median :1985   Mode  :character   Mode  :character   Median :   12.0  
##  Mean   :1975                                         Mean   :  180.9
##  3rd Qu.:2003                                         3rd Qu.:   32.0
##  Max.   :2017                                         Max.   :99686.0
##       prop
##  Min.   :2.260e-06
##  1st Qu.:3.870e-06
##  Median :7.300e-06
##  Mean   :1.363e-04
##  3rd Qu.:2.288e-05
##  Max.   :8.155e-02
```

```
?babynames
```

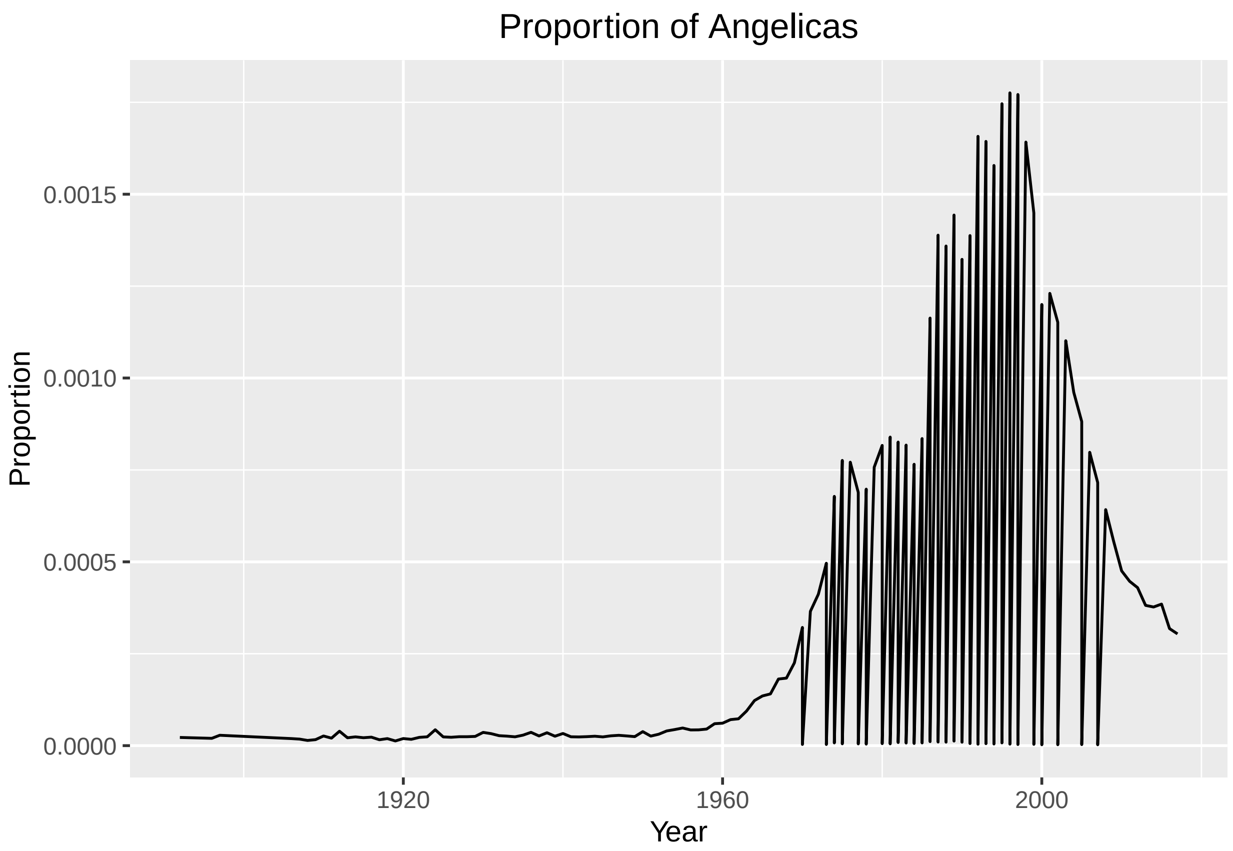
```
library(tidyr)
```

```
##
## Attaching package: 'tidyr'

## The following objects are masked from 'package:Matrix':
##
##     expand, pack, unpack
```

```
library(magrittr)
```

```
##
## Attaching package: 'magrittr'

## The following object is masked from 'package:tidyr':
##
##     extract
```

```
ggplot(data = babynames %>%
filter(name=="Angelica"), aes(x = year, y = prop)) +
geom_line() + xlab("Year") + ylab("Proportion") +
ggtitle("Proportion of Angelicas") + theme(plot.title = element_text(hjust=0.5))
```



Exercise 3.4

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

```
head(Marriage, n = 5)
```

```
##   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
## 5   B230p994  12/9/96      12/14/96     5          MINISTER  Groom 12/14/66
##        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
## 5 30.02192    White         0     <NA> 12       0      348.5 Saggitarius
```

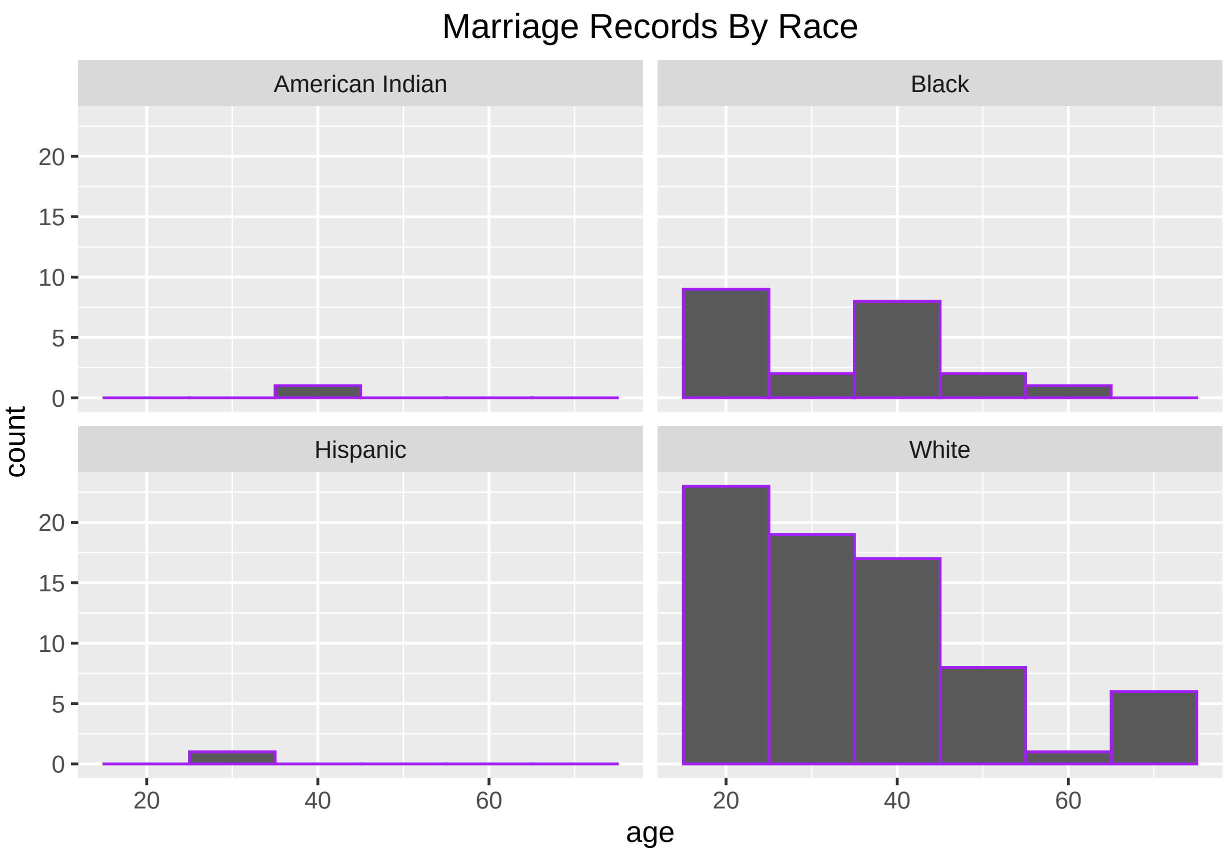
```
summary(Marriage)
```

```
##      bookpageID     appdate     ceremonydate     delay
##  B230p1209: 2   1/22/99 : 2   1/24/97 : 2    Min.   : 0.000
##  B230p1354: 2   1/30/98 : 2   1/30/98 : 2    1st Qu.: 0.000
##  B230p1665: 2   1/8/97  : 2   1/31/99 : 2    Median : 3.000
##  B230p1948: 2   10/14/98: 2   10/2/98 : 2    Mean   : 5.673
##  B230p539 : 2   10/2/98 : 2   10/20/97: 2    3rd Qu.: 9.000
##  B230p677 : 2   10/20/97: 2   10/23/98: 2    Max.   :28.000
##  (Other)  :86   (Other) :86   (Other) :86
##            officialTitle   person         dob          age
##  MARRIAGE OFFICIAL:44    Bride:49   1/21/76 : 1   Min.   :16.27
##  PASTOR           :22    Groom:49   1/30/66 : 1   1st Qu.:21.66
##  MINISTER         :20               1/31/62 : 1   Median :31.90
##  BISHOP           : 2               1/6/60  : 1   Mean   :34.51
##  CATHOLIC PRIEST  : 2               10/1/52 : 1   3rd Qu.:42.82
##  CHIEF CLERK      : 2               10/10/79: 1   Max.   :74.25
##  (Other)          : 6               (Other) :92
##               race      prevcount         prevconc        hs
##  American Indian: 1   Min.   :0.0000   Death  : 7   Min.   : 8.00
##  Black          :22   1st Qu.:0.0000   Divorce:43   1st Qu.:12.00
##  Hispanic       : 1   Median :1.0000   NA's   :48   Median :12.00
##  White          :74   Mean   :0.7755                Mean   :11.68
##                       3rd Qu.:1.0000                3rd Qu.:12.00
##                       Max.   :5.0000                Max.   :12.00
##
##     college        dayOfBirth             sign
##  Min.   :0.000   Min.   :  6.00   Pisces     :16
##  1st Qu.:0.000   1st Qu.: 81.88   Aries      :10
##  Median :1.000   Median :167.00   Virgo      :10
##  Mean   :1.625   Mean   :178.07   Gemini     : 9
##  3rd Qu.:2.000   3rd Qu.:263.94   Saggitarius: 9
##  Max.   :7.000   Max.   :358.00   Cancer     : 8
##  NA's   :10                       (Other)    :36
```

```
?Marriage
```

1. Create an informative and meaningful data graphic.

```
ggplot(data = Marriage, aes(x = age)) + facet_wrap(~race) + geom_histogram(binwidth = 10, colour="purpl
```

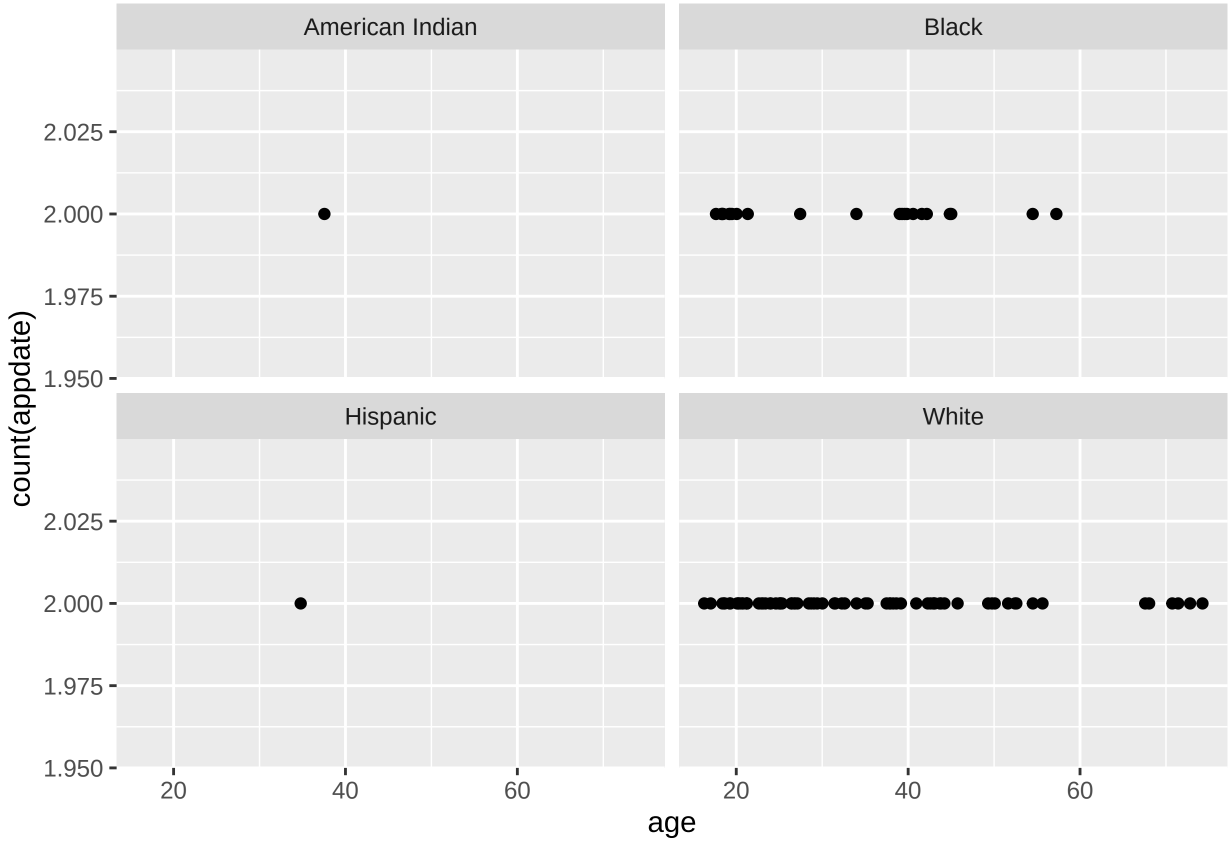


2. Identify each of the visual cues that you are using, and describe how they are related to each variable.

- Position: Title in the center (horizontally). Each facet is grouped by race
- Length: Age (x axis) ranges from 16.27 to 74.25. Min count (y axis) is 1 (hispanic) and max count is 74 (white).
- Direction: For american Indian and hispanic there isn't a clear trend (not enough data) but for white and black the count decreases as the person ages.
- Color: outlining purple

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.

```
g1 <-Marriage %>% mutate(col_bool = ifelse(college == 0 | is.na(college), FALSE, TRUE))

ggplot(data = Marriage %>% mutate(group = paste(race,sign,g1, sep="-")), aes(x = age, y=count(appdate),
```



Here I'm using 1. race 2. college 3. age 4. sign 5. appdate