Konopacka_list1_corrected.R

martyna

2021-10-17

```
#Load libraries-----
library("tidyverse")
```

```
## -- Attaching packages ---------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.0
## v tidyr   1.1.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library("data.table")
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

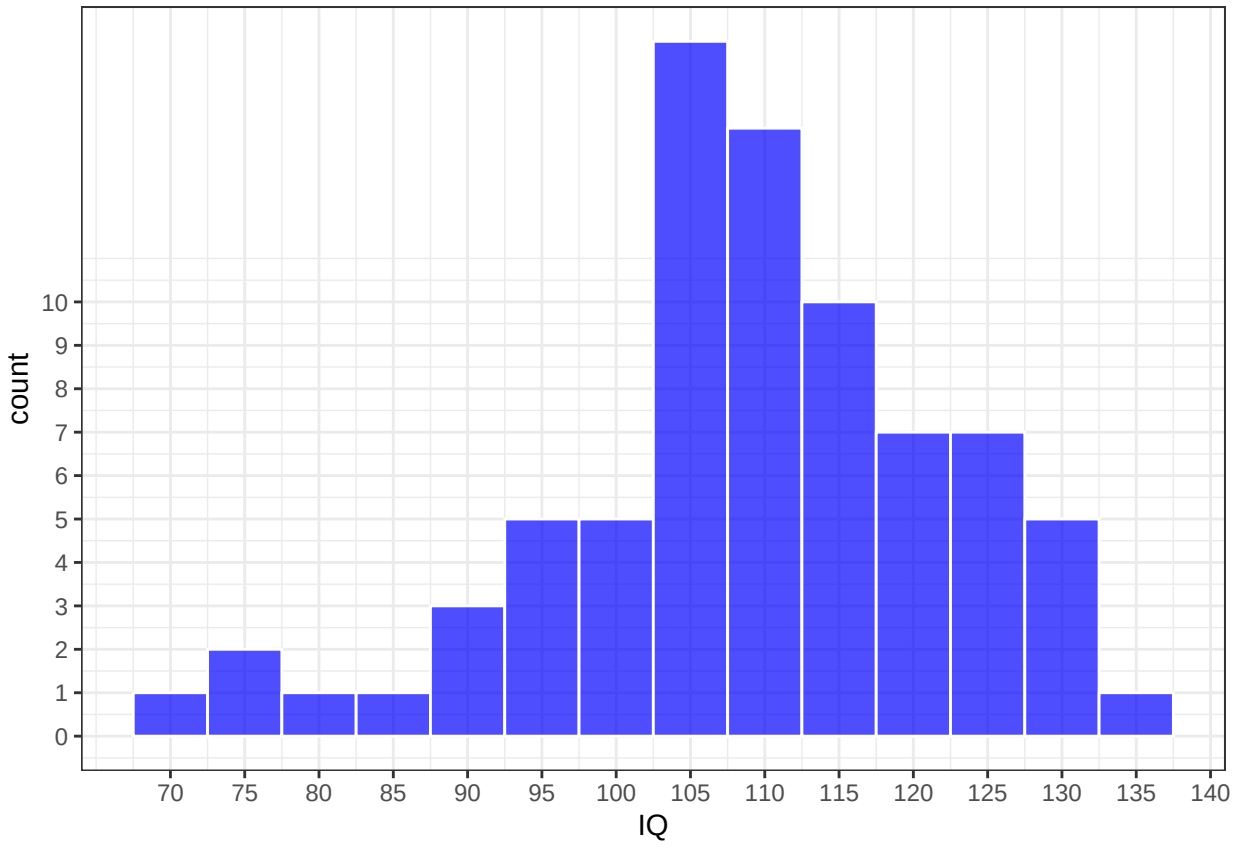
```
## The following object is masked from 'package:purrr':
##
##     transpose
```

```
#READ AND ORGANIZE DATA-----------------------------------------
grades = read.table(url("http://www.math.uni.wroc.pl/~elsner/dydaktyka/dane/grades.txt"),
                    col.names = c("id", "GPA", "IQ", "Gender", "Psych"))

income = read.table(url("http://www.math.uni.wroc.pl/~elsner/dydaktyka/dane/income.dat"),
                    col.names = c("id", "Age", "Education", "Gender", "Income", "Job"),
                    colClasses = c("integer", "integer", "factor", "factor", "integer", "factor"))

levels(income$Gender) <- c("Male", "Female")
levels(income$Job) <- c("Private sector", "Public sector", "Self-employed")
levels(income$Education) <- c("did not reach high school",
                              "some high school, without diploma",
                              "high school diploma",
                              "some college, no bachelor's degree",
                              "bachelor's degree", "postgraduate degree")

#TASK 1a-------------------------------------------------------
#funtion for printing stats and base for all plots of grades dtf
show_statistics <- function(vec){
```

```
  words <- c("min: ", "max: ", "spread: ", "median: ", "1st quartile: ", "3rd quartile: ", "mean: ", "s
  stats <- c(min(vec, na.rm = TRUE),
             max(vec, na.rm = TRUE),
             max(vec, na.rm = TRUE) - min(vec, na.rm = TRUE), #spread
             median(vec, na.rm = TRUE),
             quantile(vec, na.rm = TRUE)[2],
             quantile(vec, na.rm = TRUE)[4],
             mean(vec, na.rm = TRUE),
             sd(vec, na.rm = TRUE),
             var(vec, na.rm = TRUE),
             sd(vec, na.rm = TRUE) / abs(mean(vec, na.rm = TRUE)))
  print(deparse(substitute(vec)))
  print(paste(words, stats, sep = ""))
}

base_plot <- function(vec, binw){
  return (
    ggplot(grades, aes(x = vec)) +
      geom_histogram(binwidth =binw,
                     color = "white",
                     fill = "blue",
                     alpha = 0.7) +
      theme_bw()
  )
}

#Histogram 1: IQ---------------
#skewed left and unimodal, with its peak around 115 points.
#It's spread can be read from show_statistics. Center = median = 110.
show_statistics(grades$IQ)
```

```
## [1] "grades$IQ"
##  [1] "min: 72"
##  [2] "max: 136"
##  [3] "spread: 64"
##  [4] "median: 110"
##  [5] "1st quartile: 103"
##  [6] "3rd quartile: 117.5"
##  [7] "mean: 108.923076923077"
##  [8] "standard deviation: 13.1709728370582"
##  [9] "variance: 173.474525474525"
## [10] "coefficient of variation: 0.120919948362823"
```

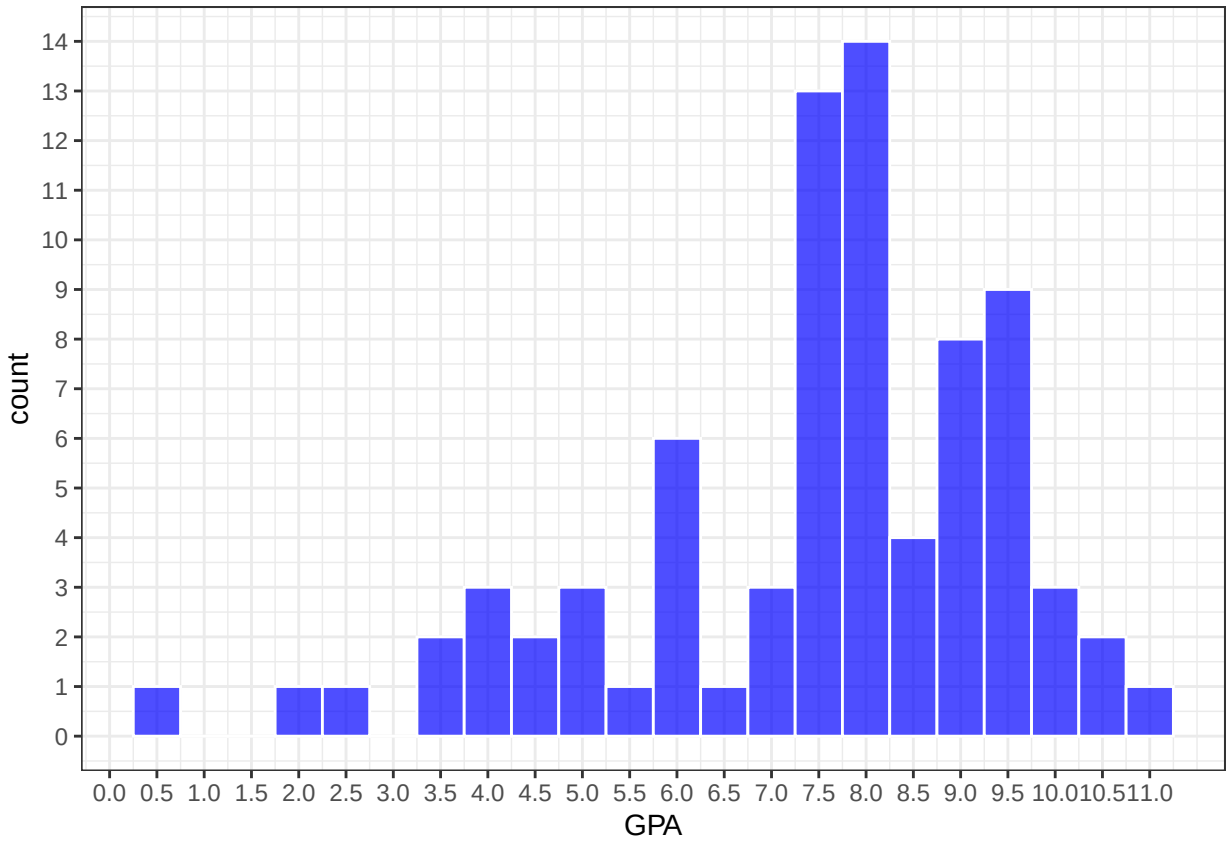
```
base_plot(grades$IQ, 5) +
  scale_x_continuous(breaks = seq(70, 140, by = 5)) +
  scale_y_continuous(breaks = seq(0, 10, by = 1)) +
  xlab("IQ")
```




```
#Histogram 2: GPA----------------
#irregular shape, multimodal with main peak around 7. Skewed left.
show_statistics(grades$GPA)
```

```
## [1] "grades$GPA"
##  [1] "min: 0.53"
##  [2] "max: 10.76"
##  [3] "spread: 10.23"
##  [4] "median: 7.829"
##  [5] "1st quartile: 6.278"
##  [6] "3rd quartile: 8.983"
##  [7] "mean: 7.44653846153846"
##  [8] "standard deviation: 2.09955744064159"
##  [9] "variance: 4.40814144655345"
## [10] "coefficient of variation: 0.281950795189718"
```

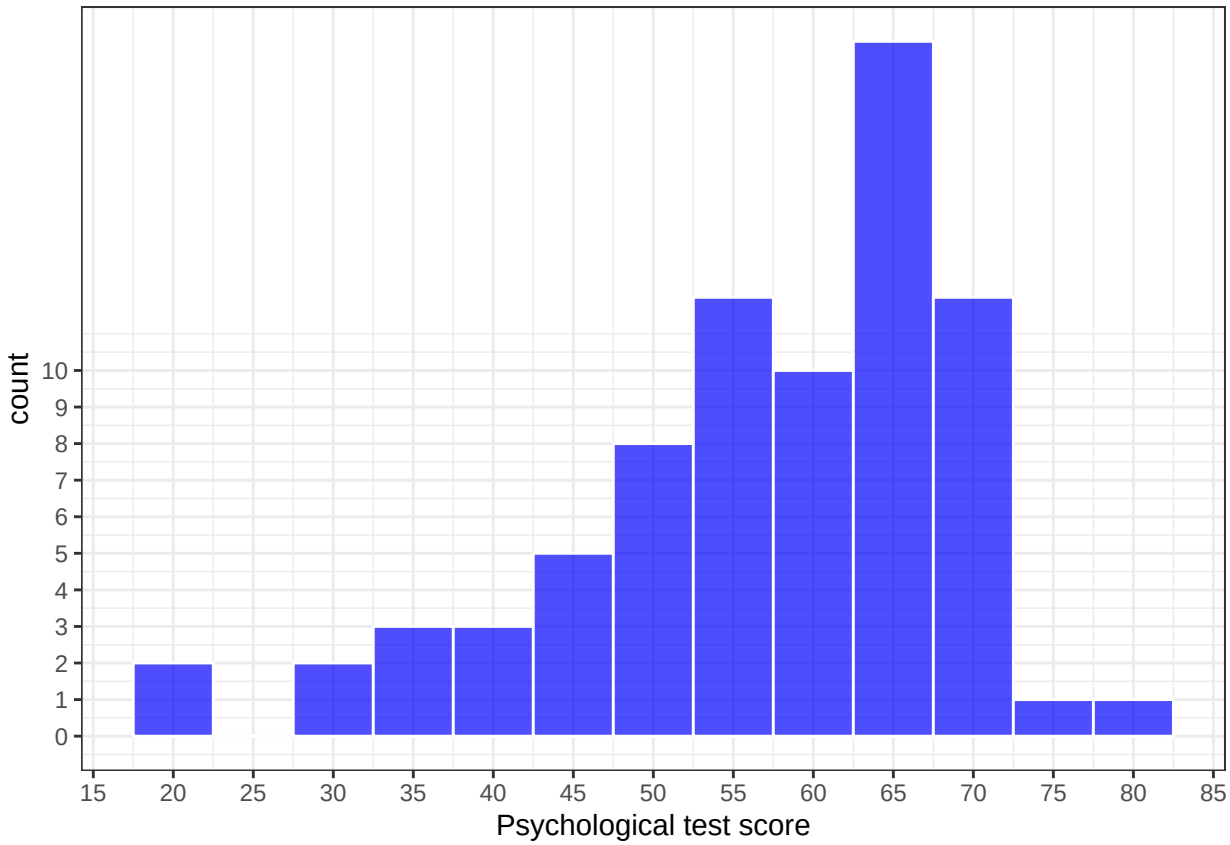
```
base_plot(grades$GPA, 0.5) +
  scale_x_continuous(breaks = seq(0, 11, by = 0.5)) +
  scale_y_continuous(breaks = seq(0, 14, by = 1)) +
  xlab("GPA")
```




```
#PHistogram 3: Psych----------------
#skewed left, multimodal histogram with its main peak around 66 points.
show_statistics(grades$Psych)
```

```
## [1] "grades$Psych"
##  [1] "min: 20"
##  [2] "max: 80"
##  [3] "spread: 60"
##  [4] "median: 59.5"
##  [5] "1st quartile: 51"
##  [6] "3rd quartile: 66"
##  [7] "mean: 56.9615384615385"
##  [8] "standard deviation: 12.4122293148103"
##  [9] "variance: 154.063436563437"
## [10] "coefficient of variation: 0.217905443744138"
```

```
base_plot(grades$Psych, 5) +
  scale_x_continuous(breaks = seq(15, 85, by = 5)) +
  scale_y_continuous(breaks = seq(0, 10, by = 1)) +
  xlab("Psychological test score")
```




```
#TASK 1b-----------------------------------------------------------------
grades_f <- filter(grades, Gender == "F")
grades_m <- filter(grades, Gender == "M")

#Histogram 1: IQ---------
#histogram for female students is unimodal and more symmetric, while histogram for male
#students is bimodal and skewed left. peaks of both histograms are located around value
#of 105-110 points.
show_statistics(grades_f$IQ)
```

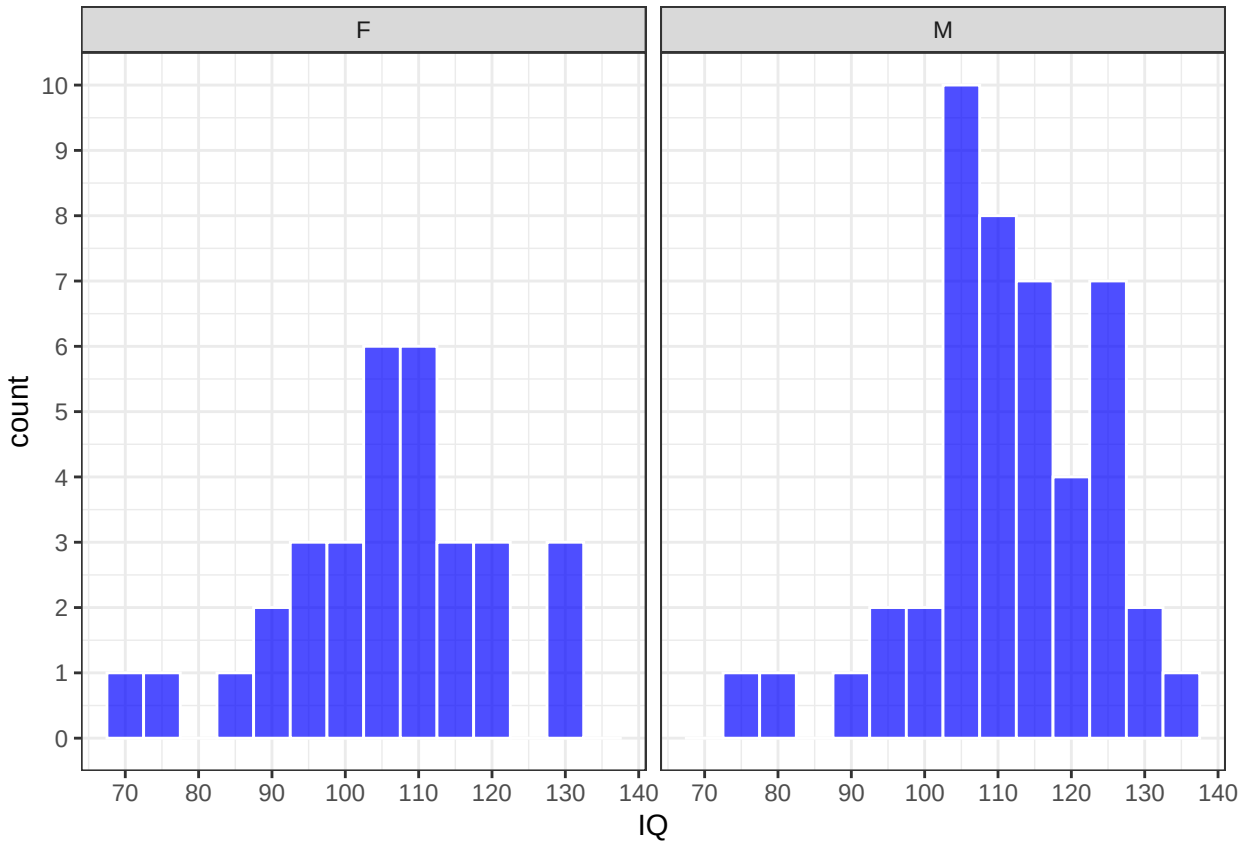
```
## [1] "grades_f$IQ"
##  [1] "min: 72"
##  [2] "max: 132"
##  [3] "spread: 60"
##  [4] "median: 106"
##  [5] "1st quartile: 97.5"
##  [6] "3rd quartile: 114"
##  [7] "mean: 105.4375"
##  [8] "standard deviation: 14.2216044192652"
##  [9] "variance: 202.254032258065"
## [10] "coefficient of variation: 0.134881843929012"
```

```
show_statistics(grades_m$IQ)
```

```
## [1] "grades_m$IQ"
##  [1] "min: 77"
##  [2] "max: 136"
```

```
##  [3] "spread: 59"
##  [4] "median: 111"
##  [5] "1st quartile: 106"
##  [6] "3rd quartile: 119"
##  [7] "mean: 111.347826086957"
##  [8] "standard deviation: 11.9521590635413"
##  [9] "variance: 142.854106280193"
## [10] "coefficient of variation: 0.107340749106384"
```

```
base_plot(grades$IQ, 5) +
  scale_x_continuous(breaks = seq(60, 145, by = 10)) +
  scale_y_continuous(breaks = seq(0, 10, by = 1)) +
  xlab("IQ") +
  facet_wrap(~ Gender)
```




```
#Histogram2: GPA------------
#Both histograms are multimodal with main peaks around 7.5 points, skewed left.
show_statistics(grades_f$GPA)
```

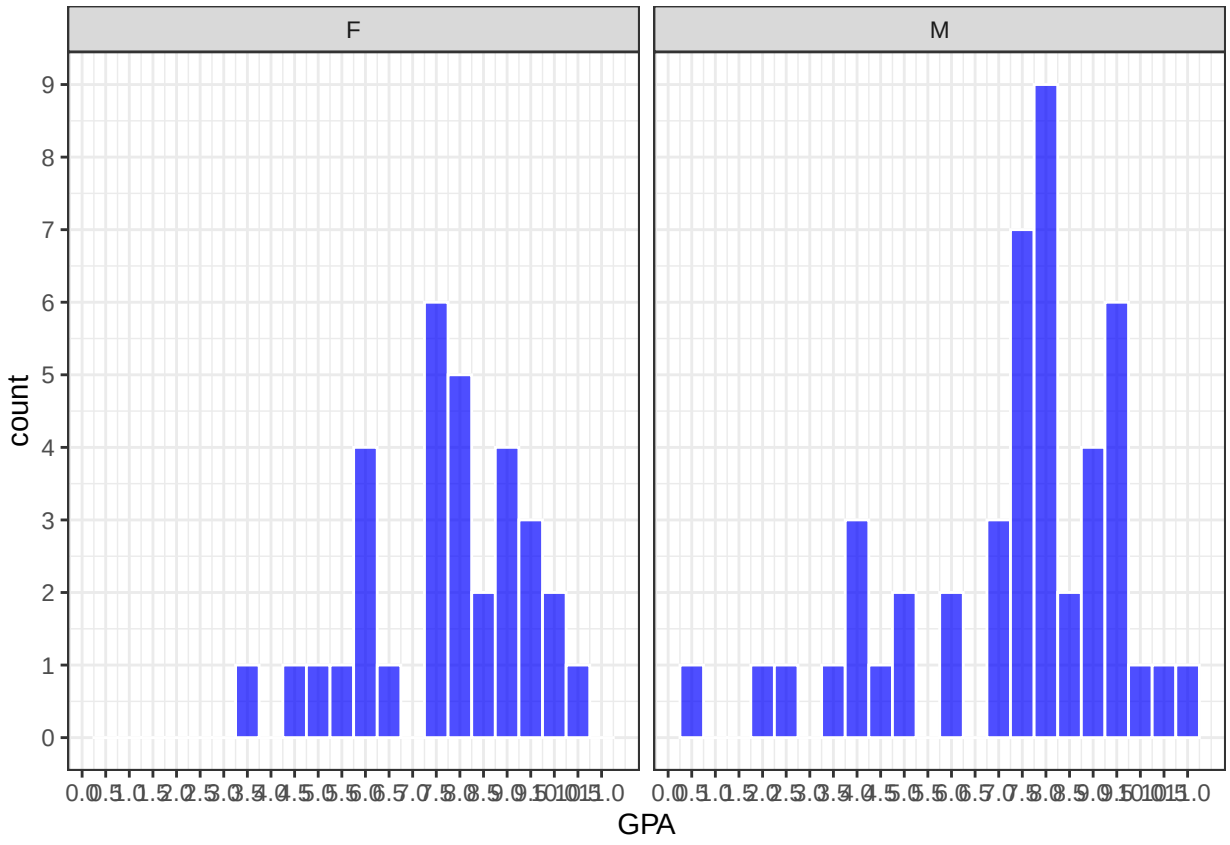
```
## [1] "grades_f$GPA"
##  [1] "min: 3.408"
##  [2] "max: 10.7"
##  [3] "spread: 7.292"
##  [4] "median: 7.8275"
##  [5] "1st quartile: 6.372"
##  [6] "3rd quartile: 8.953"
##  [7] "mean: 7.684"
```

```
##  [8] "standard deviation: 1.6943183423316"
##  [9] "variance: 2.87071464516129"
## [10] "coefficient of variation: 0.220499523989016"
```

```
show_statistics(grades_m$GPA)
```

```
## [1] "grades_m$GPA"
##  [1] "min: 0.53"
##  [2] "max: 10.76"
##  [3] "spread: 10.23"
##  [4] "median: 7.8825"
##  [5] "1st quartile: 6.36425"
##  [6] "3rd quartile: 9.0835"
##  [7] "mean: 7.28134782608696"
##  [8] "standard deviation: 2.34461594312863"
##  [9] "variance: 5.49722392077295"
## [10] "coefficient of variation: 0.322003013608078"
```

```
base_plot(grades$GPA, 0.5) +
  scale_x_continuous(breaks = seq(0, 11, by = 0.5)) +
  scale_y_continuous(breaks = seq(0, 14, by = 1)) +
  xlab("GPA") +
  facet_wrap(~Gender)
```



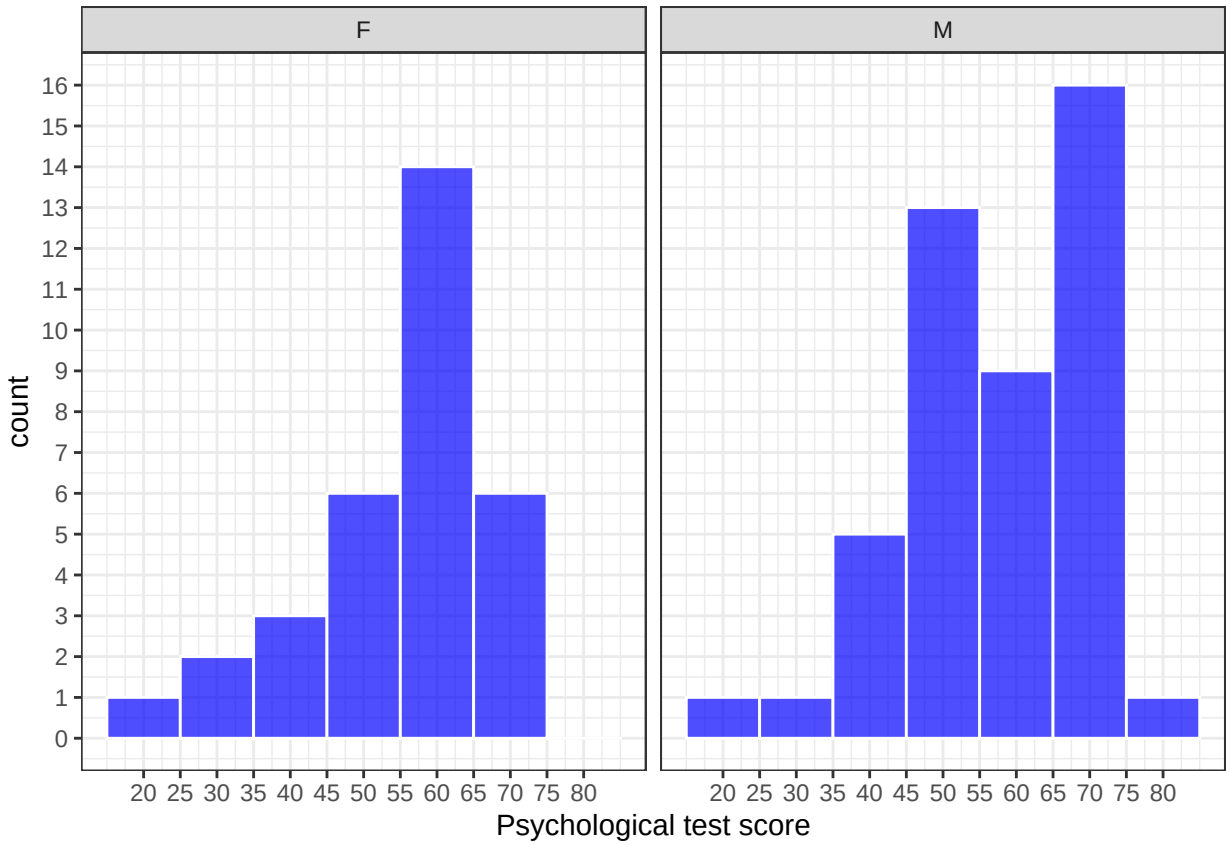

```
#Histogram3: Psychological test-------------
#Histogram for female students is skewed left and unimodal, while histogram for
#male students is more irregular, bimodal.
show_statistics(grades_f$Psych)
```

```
## [1] "grades_f$Psych"
##  [1] "min: 21"
##  [2] "max: 72"
##  [3] "spread: 51"
##  [4] "median: 60"
##  [5] "1st quartile: 52.75"
##  [6] "3rd quartile: 64"
##  [7] "mean: 55.6875"
##  [8] "standard deviation: 12.5272284090657"
##  [9] "variance: 156.931451612903"
## [10] "coefficient of variation: 0.224955841240237"
```

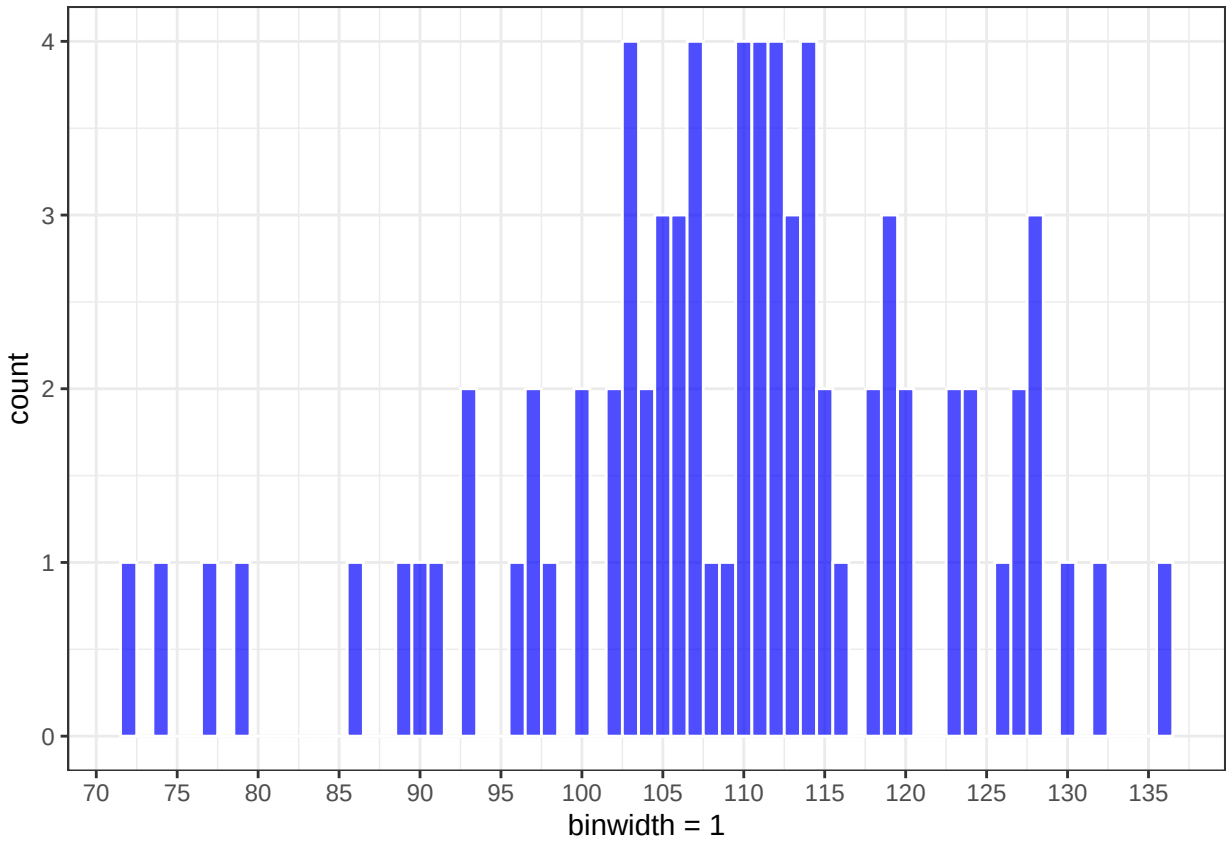
```
show_statistics(grades_m$Psych)
```

```
## [1] "grades_m$Psych"
##  [1] "min: 20"
##  [2] "max: 80"
##  [3] "spread: 60"
##  [4] "median: 59"
##  [5] "1st quartile: 51"
##  [6] "3rd quartile: 67"
##  [7] "mean: 57.8478260869565"
##  [8] "standard deviation: 12.3916950527437"
##  [9] "variance: 153.554106280193"
## [10] "coefficient of variation: 0.214211940032398"
```

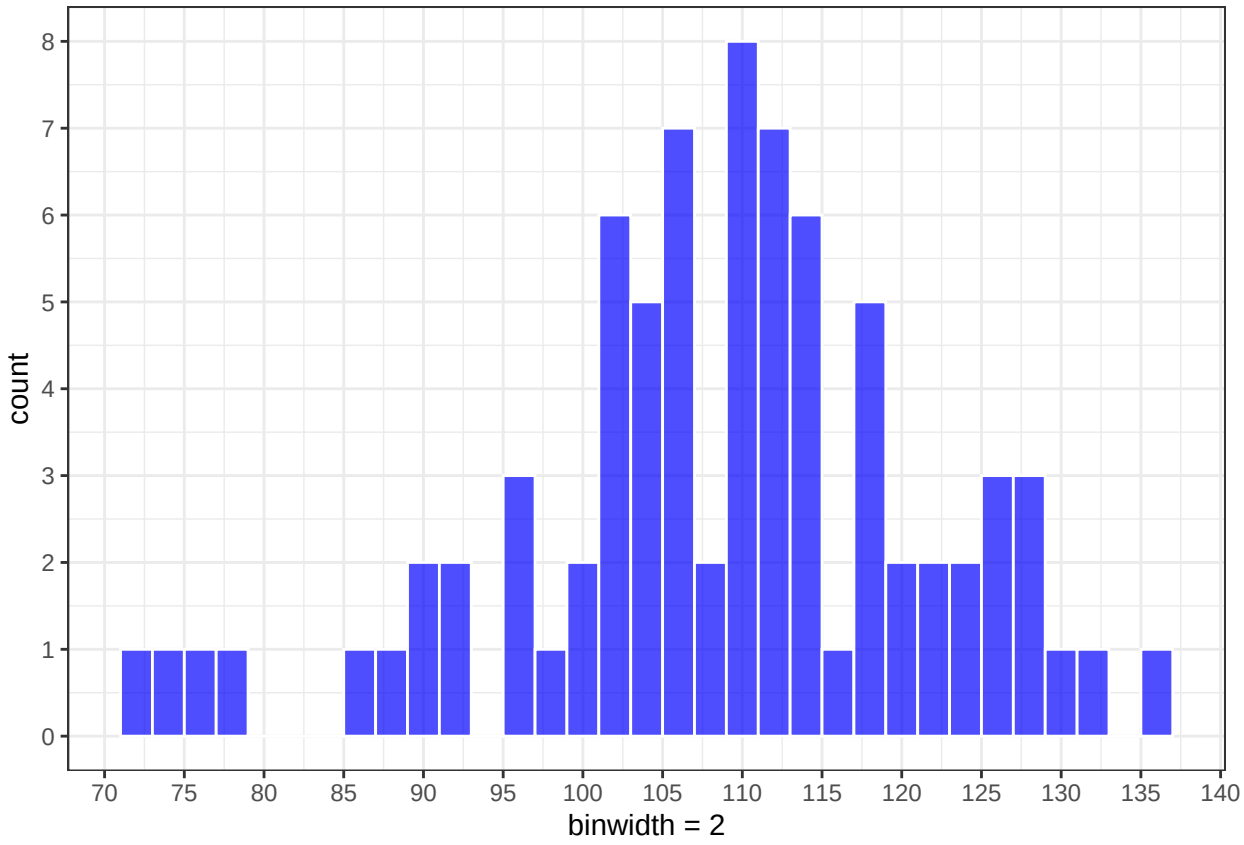
```
base_plot(grades$Psych, 10) +
  scale_x_continuous(breaks = seq(20, 80, by = 5)) +
  scale_y_continuous(breaks = seq(0, 18, by = 1)) +
  xlab("Psychological test score") +
  facet_wrap(~Gender)
```



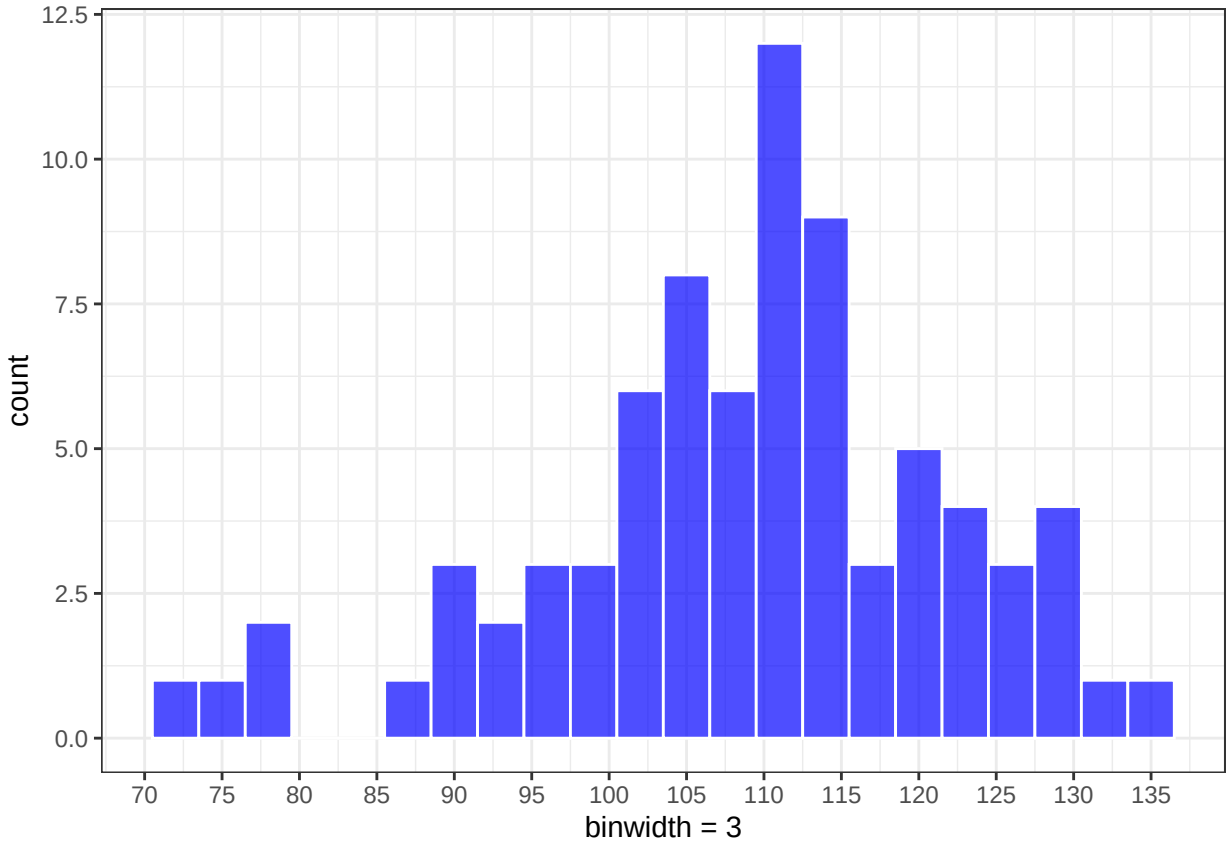

```
#TASK 2----------------------------------------------------------------------
base_plot(grades$IQ, 1) +
  scale_x_continuous(breaks = seq(70, 140, by = 5)) +
  scale_y_continuous(breaks = seq(0, 10, by = 1)) +
  xlab("binwidth = 1") #it's really unreadable and makes it harder to recognize statistical tendencies
```



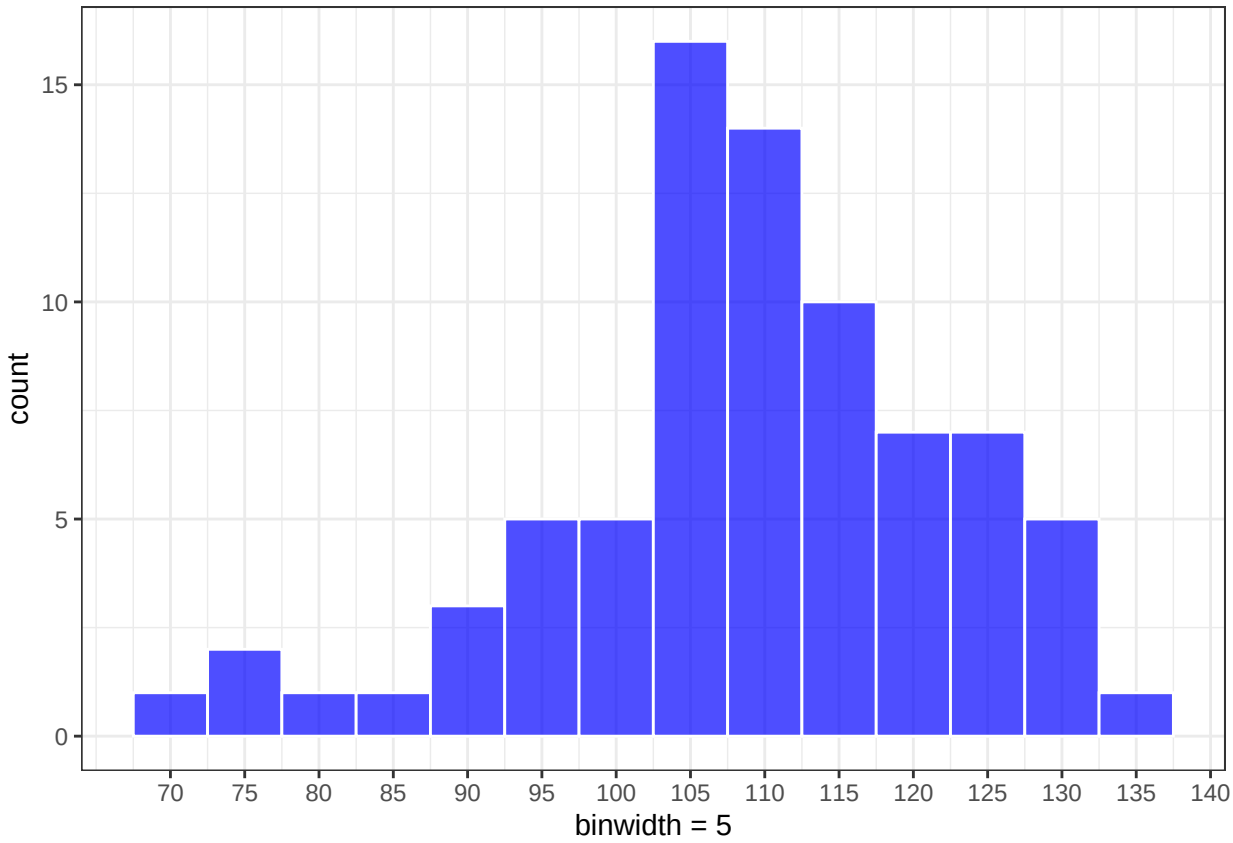

```
base_plot(grades$IQ, 2) +
  scale_x_continuous(breaks = seq(70, 140, by = 5)) +
  scale_y_continuous(breaks = seq(0, 10, by = 1)) +
  xlab("binwidth = 2") #it's still a bit unreadable, but definitely better
```



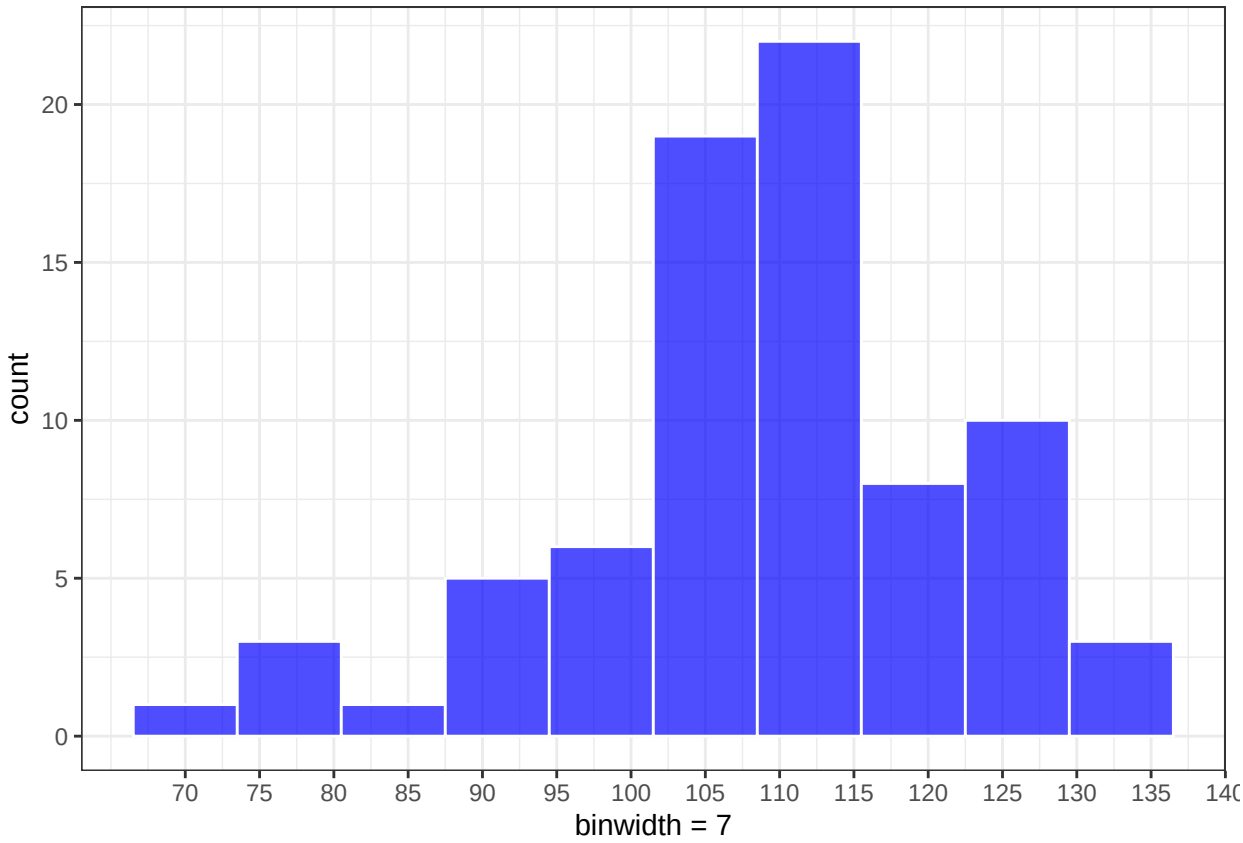

```
base_plot(grades$IQ, 3) +
  scale_x_continuous(breaks = seq(70, 140, by = 5)) +
  xlab("binwidth = 3") #I find it acceptable number of classes
```



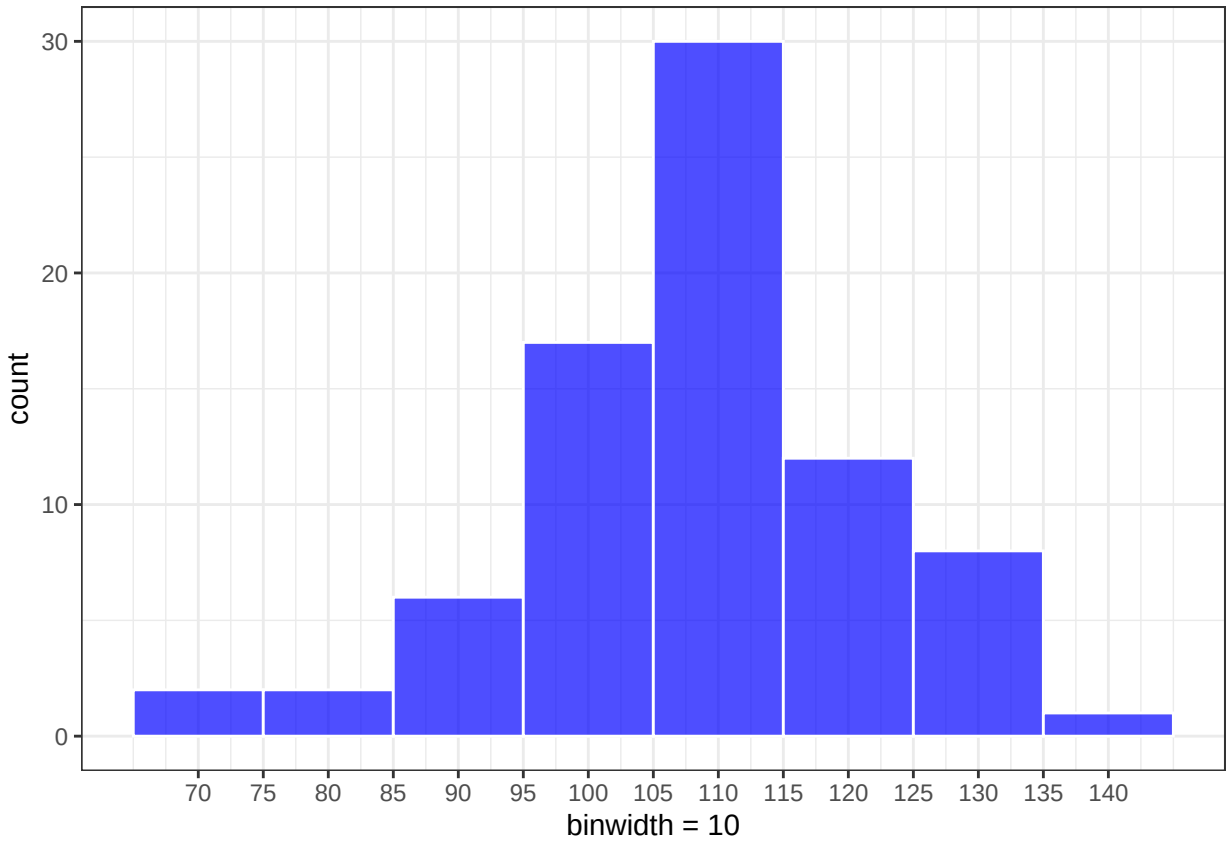

```
base_plot(grades$IQ, 5) +
  scale_x_continuous(breaks = seq(70, 140, by = 5)) +
  xlab("binwidth = 5") #I find this the best number of classes, but binwidth = 3 or 4 is also good
```




```
base_plot(grades$IQ, 7) +
  scale_x_continuous(breaks = seq(70, 140, by = 5)) +
  xlab("binwidth = 7") #here we're losing too much data
```



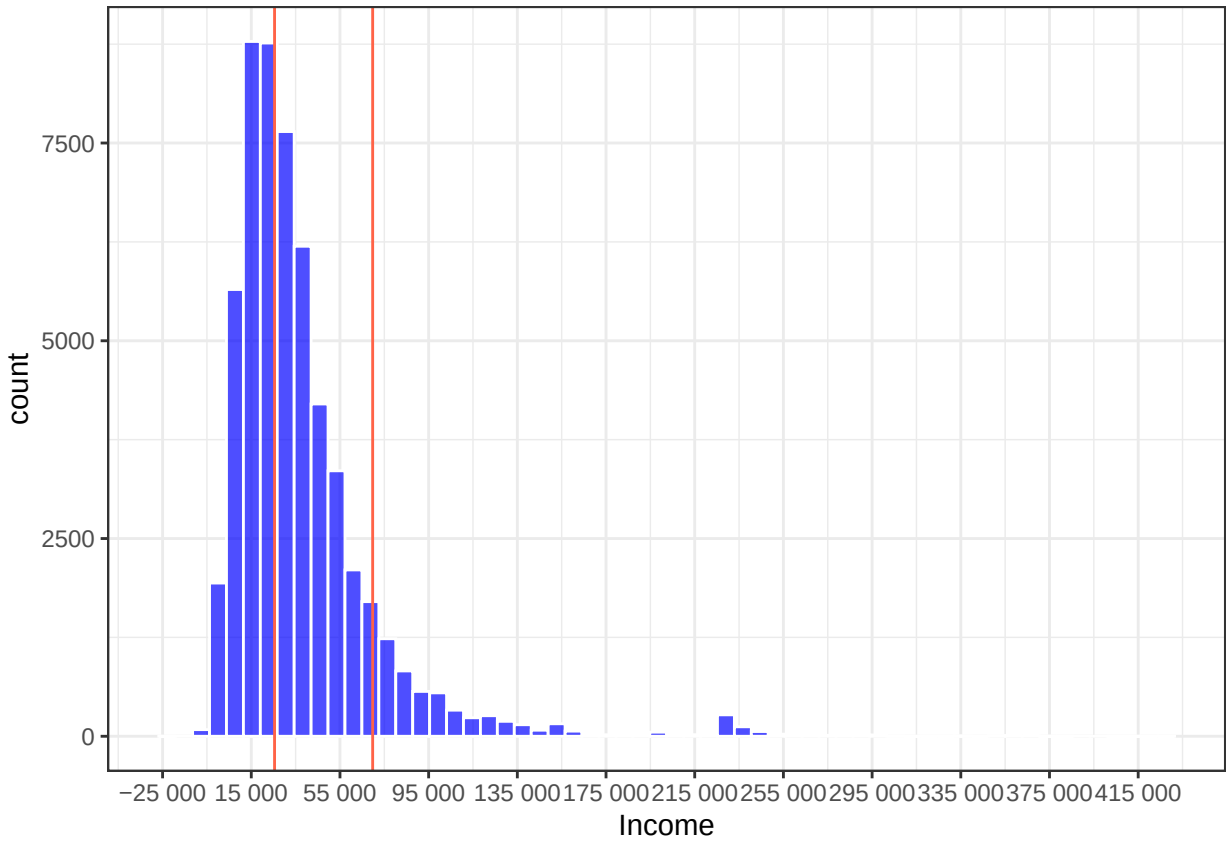

```
base_plot(grades$IQ, 10) +
  scale_x_continuous(breaks = seq(70, 140, by = 5)) +
  xlab("binwidth = 10")
```




```r
#now students with IQ difference of 10 points are in the same bin. This is too much,
#considering the fact that it is 10/64 > 15% of range.

#TASK 3--------------------------------------
mark_outliers <- function(vec, a = 1.5){ #returns boolean vector showing wheter observation is an outli
  left = a*quantile(vec)[2]
  right = a*quantile(vec)[4]
  cond <- vec <= right & vec >= left
  return(cond)
} #It can be used for recognizing or removing outliers

#I tried to create a function drawing lines that separate outliers, doesn't work.
#It was supposed to be added as a layer to ggplot (with +)
# outlier_lines <- function(vec, a = 1.5){
#   left = a*quantile(vec)[2]
#   right = a*quantile(vec)[4]
#   return(
#     geom_vline(xintercept = left, color = "tomato") +
#     geom_vline(xintercept = right, color = "tomato")
#   )
# }

#Histogram 1: all people. It's skewed right and unimodal and looks a little bit
#like exponential drop from its peak at value around 25000. Some observations are
#invisible, because thir values are too small compared to the highest ones.
show_statistics(income$Income)
```

```
## [1] "income$Income"
##  [1] "min: -24998"
##  [2] "max: 425510"
##  [3] "spread: 450508"
##  [4] "median: 29717"
##  [5] "1st quartile: 17000"
##  [6] "3rd quartile: 46503.5"
##  [7] "mean: 37864.6094563409"
##  [8] "standard deviation: 36158.0278116033"
##  [9] "variance: 1307402975.22468"
## [10] "coefficient of variation: 0.954929374177084"
```

```r
#outliers lines
left = 1.5*quantile(income$Income)[2]
right = 1.5*quantile(income$Income)[4]

ggplot(income, aes(x = Income)) +
  geom_histogram(bins = 60,
                 color = "white",
                 fill = "blue",
                 alpha = 0.7) +
  theme_bw() +
  geom_vline(xintercept = left, color = "tomato") +
  geom_vline(xintercept = right, color = "tomato") +
  scale_x_continuous(labels = scales::number,
                     breaks = seq(-25000, 430000, by = 40000))
```



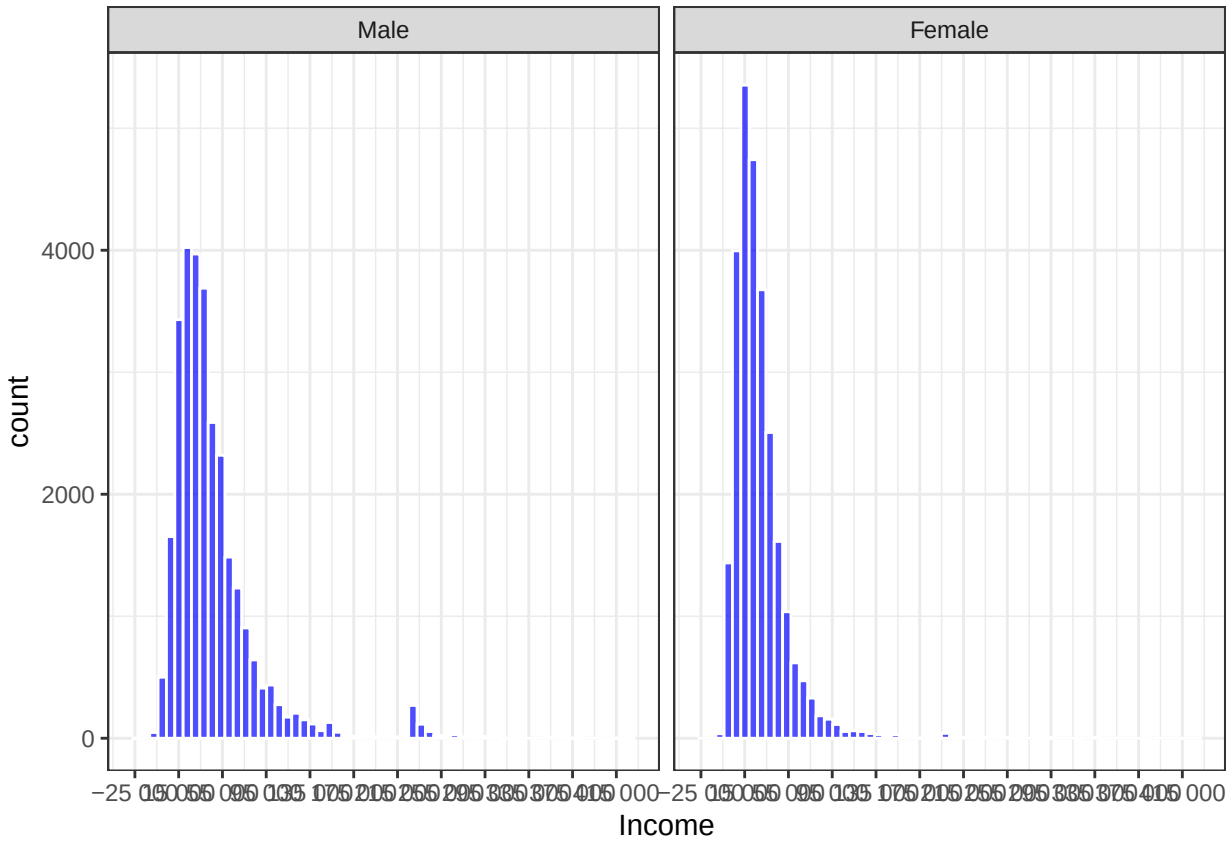
```r
#Males and females separately:
income_f <- filter(income, Gender == "Female")
income_m <- filter(income, Gender == "Male")

show_statistics(income_f$Income)
```

```
## [1] "income_f$Income"
##  [1] "min: -9999"
##  [2] "max: 385068"
##  [3] "spread: 395067"
##  [4] "median: 23012"
##  [5] "1st quartile: 13004"
##  [6] "3rd quartile: 36200"
##  [7] "mean: 28422.3532696271"
##  [8] "standard deviation: 25279.1599072655"
##  [9] "variance: 639035925.617101"
## [10] "coefficient of variation: 0.889411220367861"
```

```r
show_statistics(income_m$Income)
```

```
## [1] "income_m$Income"
##  [1] "min: -24998"
##  [2] "max: 425510"
##  [3] "spread: 450508"
##  [4] "median: 36000"
##  [5] "1st quartile: 22146"
##  [6] "3rd quartile: 55852.25"
##  [7] "mean: 46489.4676182652"
##  [8] "standard deviation: 41977.6358413775"
##  [9] "variance: 1762121910.8313"
## [10] "coefficient of variation: 0.902949377395859"
```

```r
#outliers lines
left_f = 1.5*quantile(income_f$Income)[2]
right_f = 1.5*quantile(income_f$Income)[4]
left_m = 1.5*quantile(income_m$Income)[2]
right_m = 1.5*quantile(income_m$Income)[4]

#Shape of the histograms is similar to shape of plot for both genders.
ggplot(income, aes(x = Income)) +
  geom_histogram(bins = 60,
                 color = "white",
                 fill = "blue",
                 alpha = 0.7) +
  theme_bw() +
  scale_x_continuous(labels = scales::number,
                     breaks = seq(-25000, 430000, by = 40000)) +
  facet_wrap(~Gender)
```




```
#maybe logarithmic scale would be better? + I don't know how to add outlier lines to facets
```