Lesson 02 Exploring Data

Lusine Zilfimian

February 17 (Monday), 2020

Contents

Questions (1)

- Did you find books/sources?
- Did you find HW?
- Will we have a session on Wednesday? When? Where?
- Did you join/fork me on Github?

Hmmm. . .

data-mining-ysu-spring-2020

Repositories 4 | Packages | People 1 | Teams | Projects | Settings

Find a repository... | Type: All | Customize pins | New

Lecture-slides-Data

0 0 0 0 Updated 2 minutes ago

Syllabus-and-Grades

Here is the syllabus for DM course Spring 2020 (YSU DSFB).

0 0 0 0 Updated 7 days ago

Announcements-Blog

0 0 0 0 Updated on Jan 17

Homework-Quiz

People 1 >

Invite someone

Last Lecture ReCap

- Give the definition of categorical and numeric data.
- What is the difference between long and wide data?

Questions (2)

Your questions?

Questions (3)

Question from Smbat:

What is the difference between **noise** and inconsistent value?

- Noise is the random component of a measurement error.
- Usually happens in signal or image processing.
- Noisy data is data with a large amount of additional meaningless information in it.
- Examples.

Data Preprocessing

- Aggregation: to reduce the memory and provide high-level view
- Sampling (types)
- Feature subset selection (Redundant and Irrelevant features)
- Feature creation
- Discretization and binarization
- Variable transformation (Normalization or Standardization)

Exploring Data: Summary Statistics

Frequencies and the Mode

```
## DM
## Drop Fail Pass  Sum
##    2    4   10   16
```

Percentiles

```
## The dataset is  1 1 2 2 2 4 4 5 50
```

```
## 25% 50% 75%
##   2   2   4
```

Mean and Median

```
## Mean:  7.888889
```

```
## Median:  2
```

Range and Variance

```
## Range:  1 50
```

```
## Variance:  251.3611
```

```
## SD:  15.85437
```

IQR and MAD

```
## IQR:  2
```

```
## MAD:  1.4826
```

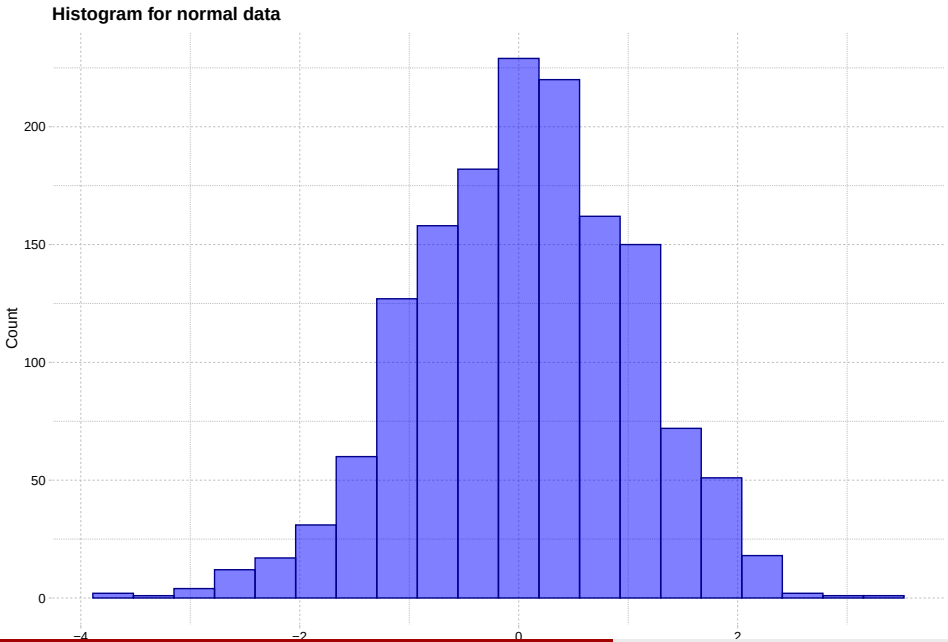
Covariance and Correlation

Exploring Data: Visualization

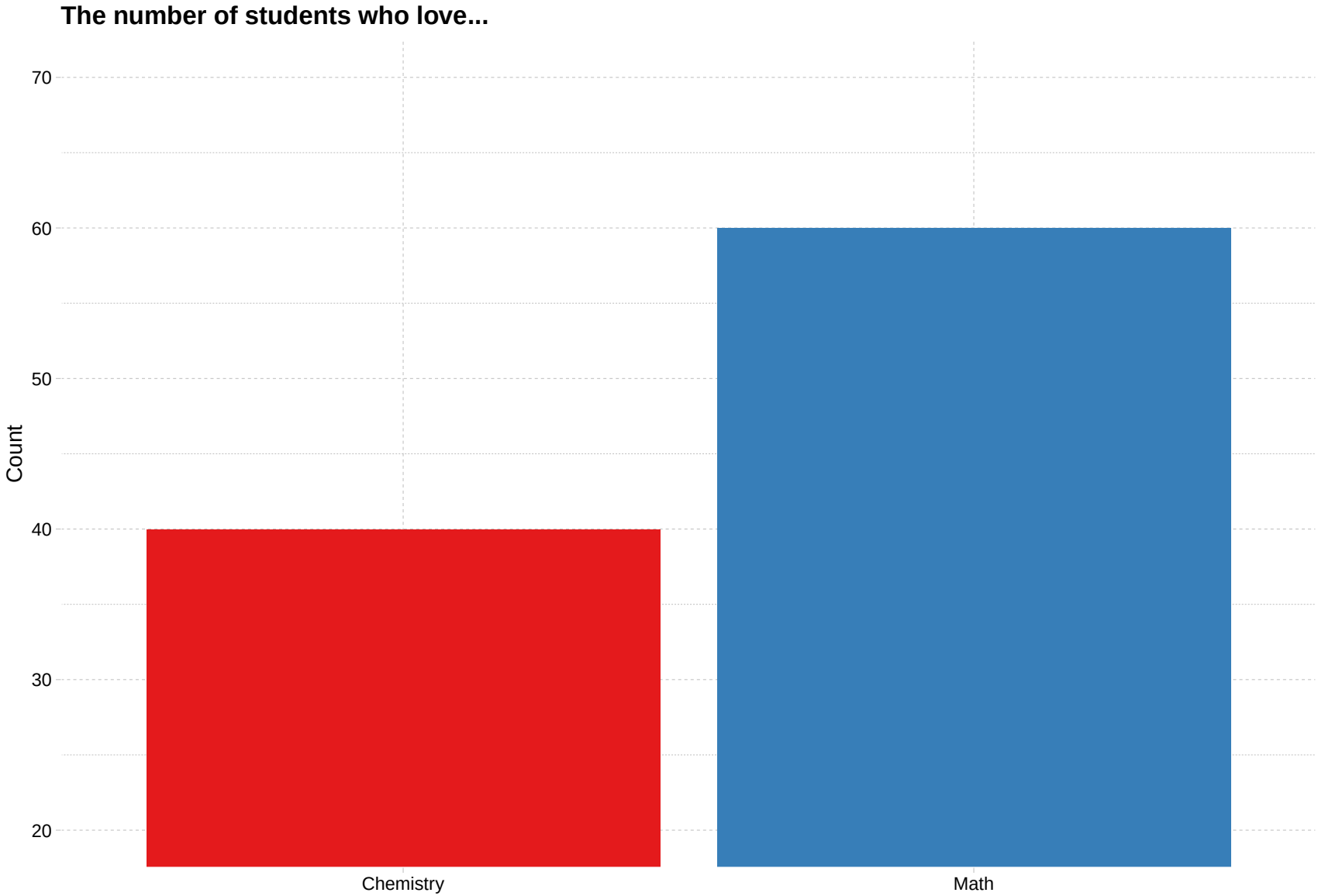
Visualizations of the data may be **the best way** of finding patterns of interest since a person cannot get an insight from the list of numbers.

- Histogram, Stem and Leaf plot
- Bar Plot
- Box Plot
- Scatter Plot
- Time Series (Line Graph (Do we need to separate it?))

```
##
##   The decimal point is at the |
##
##   -2 | 2
##   -1 | 421
##   -0 | 9764
##    0 | 247
##    1 | 156
##    2 | 0
```



- Bar plot

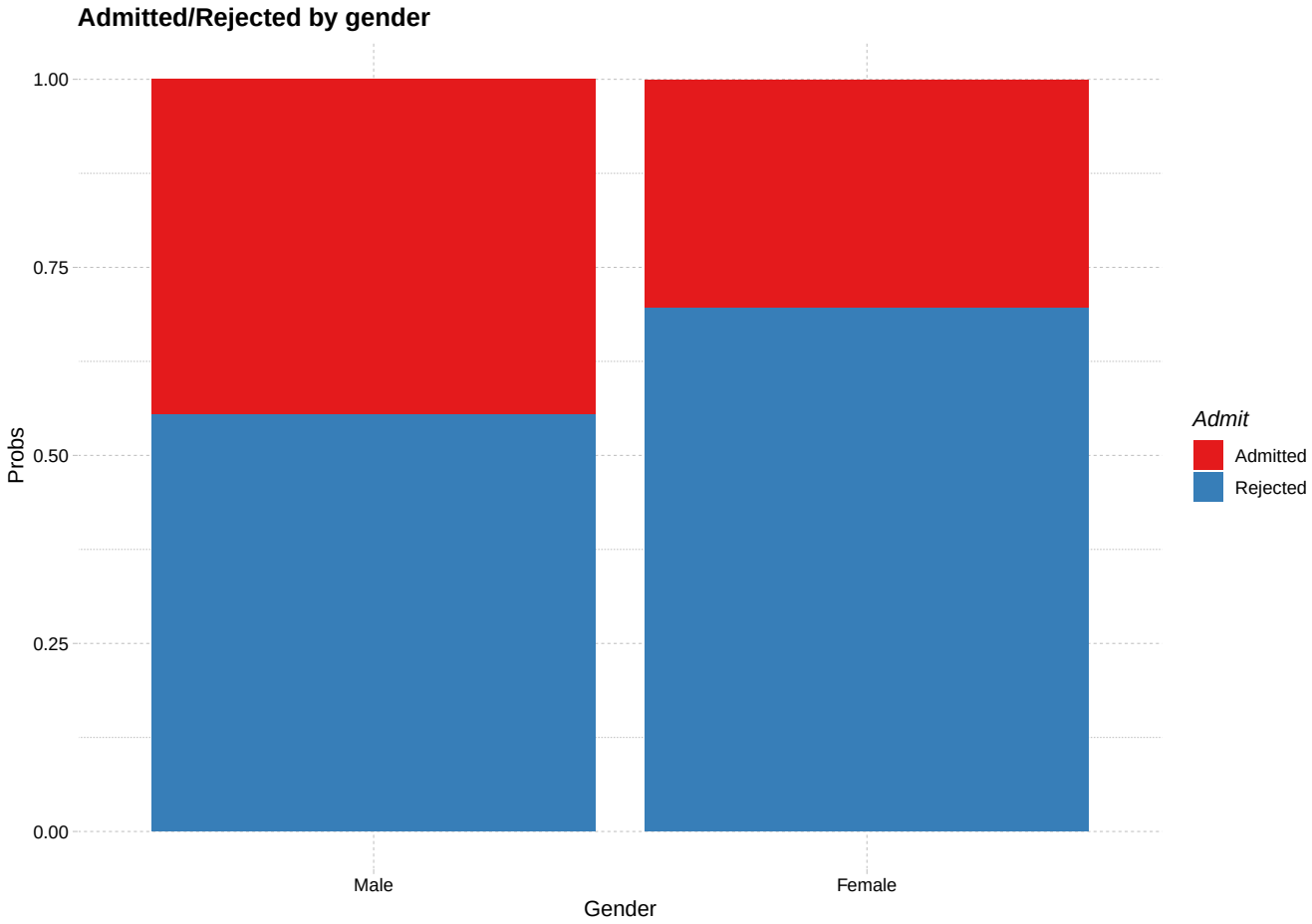

The number of students who love...
70
60
50
40
30
20
Count
Chemistry
Math

UCBAdmissions - aggregate data on applicants to graduate school at Berkeley for the **six** largest departments in 1973.

```
##       Admit Gender Dept Freq
## 1  Admitted   Male    A  512
## 2  Rejected   Male    A  313
## 3  Admitted Female    A   89
## 4  Rejected Female    A   19
## 5  Admitted   Male    B  353
## 6  Rejected   Male    B  207
```

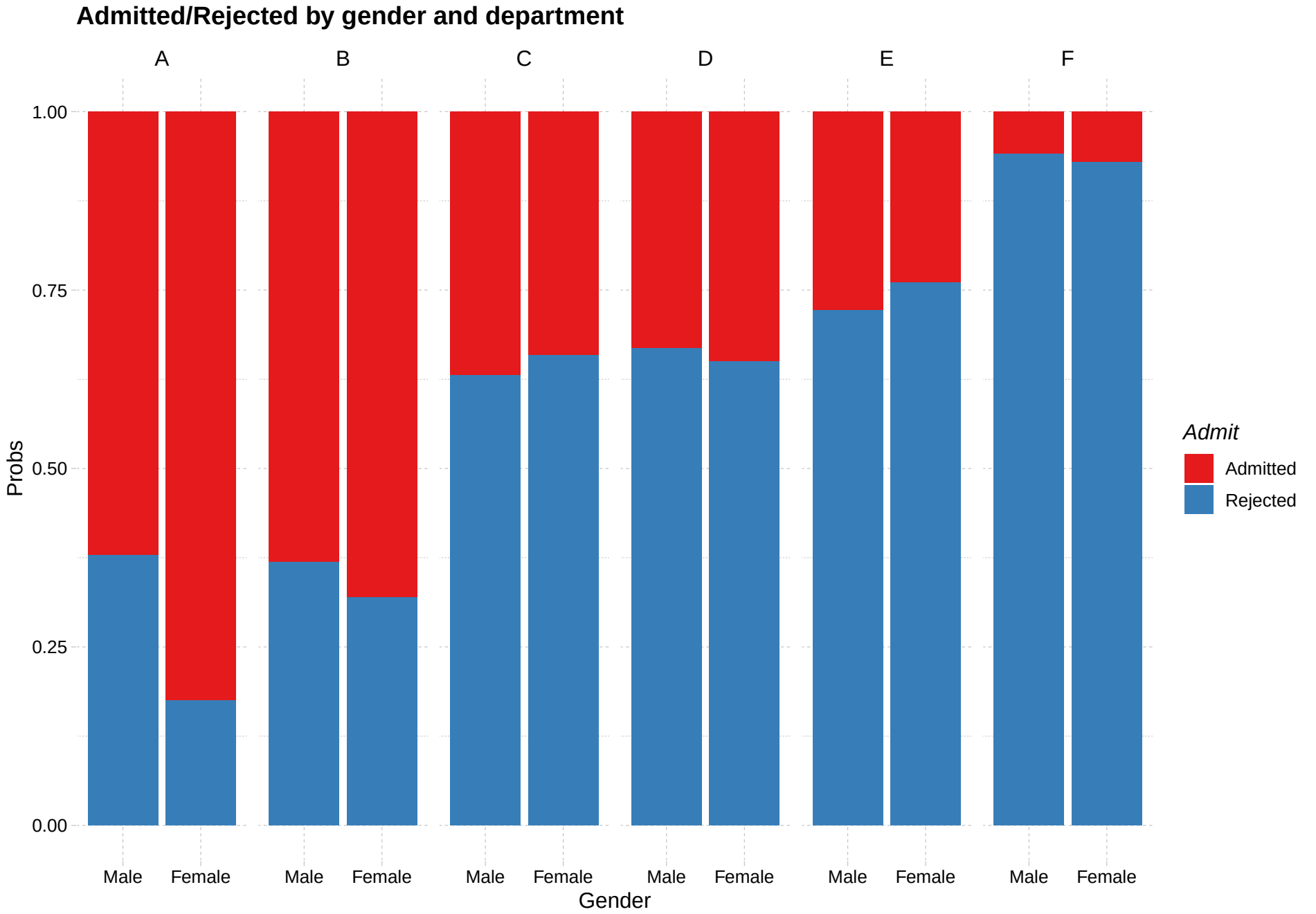
Cross tabs

```
##          Admit
## Gender    Admitted Rejected
##   Male        1198     1493
##   Female       557     1278
```

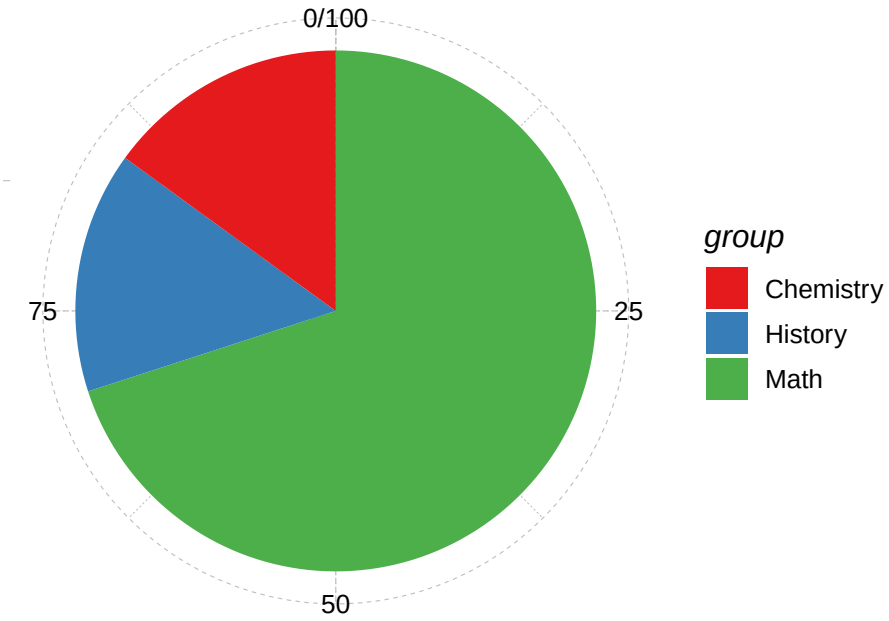


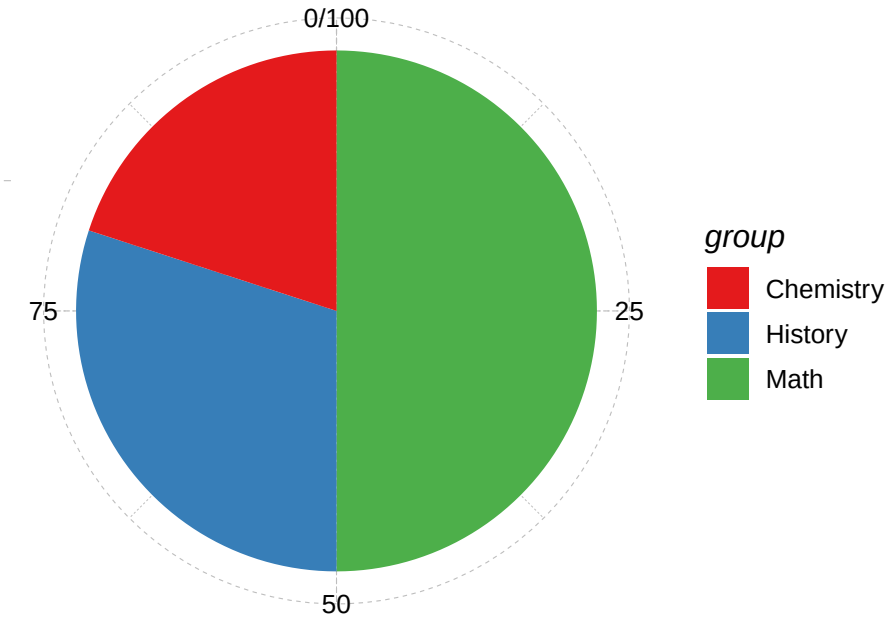
Proportional cross tabs

```
##          Admit
## Gender     Admitted  Rejected
##   Male    0.4451877 0.5548123
##   Female 0.3035422 0.6964578
```

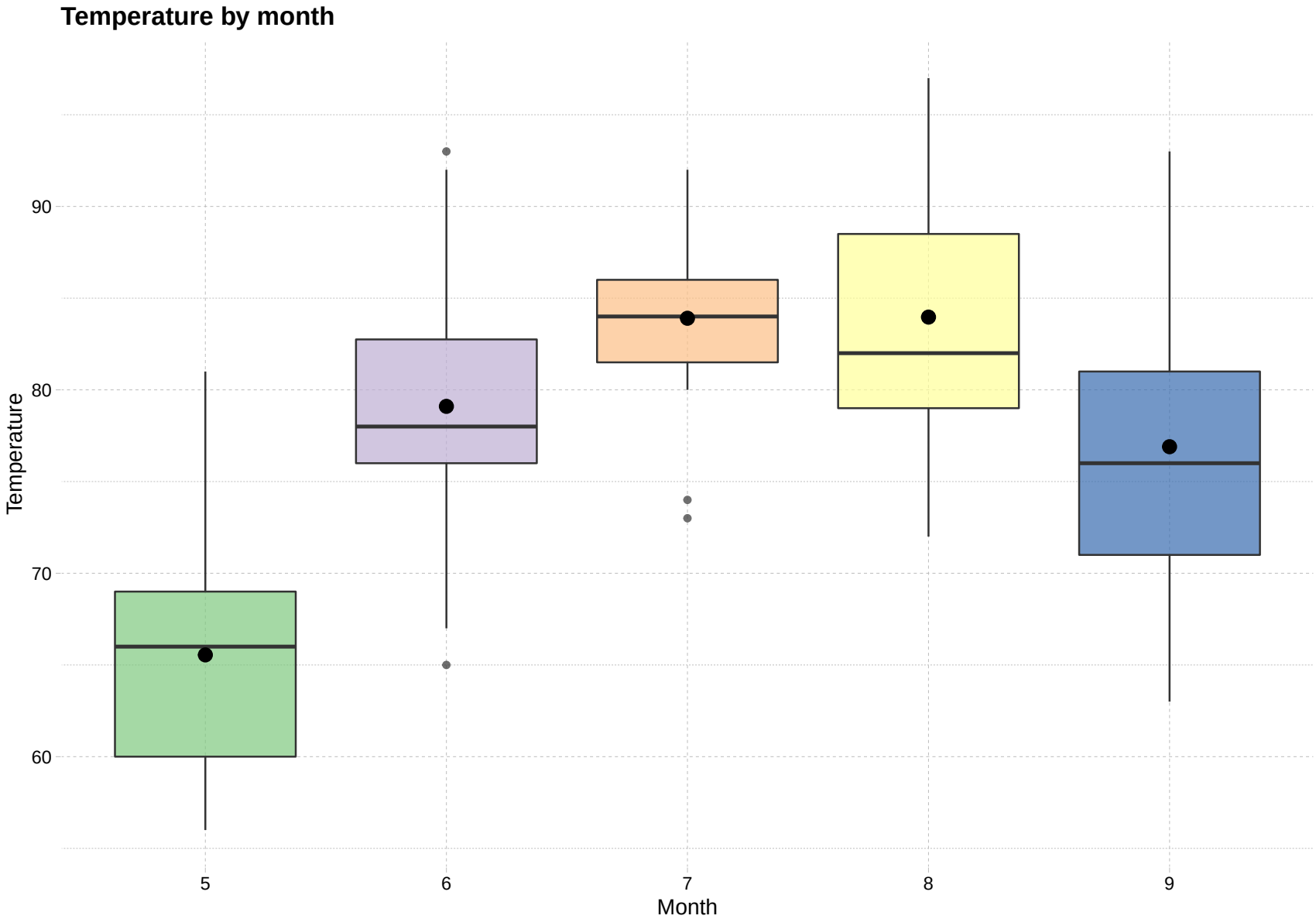

Admitted/Rejected by gender and department
A
B
C
D
E
F
1.00
0.75
0.50
0.25
0.00
Probs
Male
Female
Gender
Admit
Admitted
Rejected

- Pie Chart


0/100
25
50
75
group
Chemistry
History
Math


0/100
25
50
75
group
Chemistry
History
Math

- Box plot



- Scatter Plot and Anscombe's quarters

```
##    x1 x2 x3 x4    y1   y2    y3    y4
## 1  10 10 10  8  8.04 9.14  7.46  6.58
## 2   8  8  8  8  6.95 8.14  6.77  5.76
## 3  13 13 13  8  7.58 8.74 12.74  7.71
## 4   9  9  9  8  8.81 8.77  7.11  8.84
## 5  11 11 11  8  8.33 9.26  7.81  8.47
## 6  14 14 14  8  9.96 8.10  8.84  7.04
## 7   6  6  6  8  7.24 6.13  6.08  5.25
## 8   4  4  4 19  4.26 3.10  5.39 12.50
## 9  12 12 12  8 10.84 9.13  8.15  5.56
## 10  7  7  7  8  4.82 7.26  6.42  7.91
## 11  5  5  5  8  5.68 4.74  5.73  6.89
```

Mean

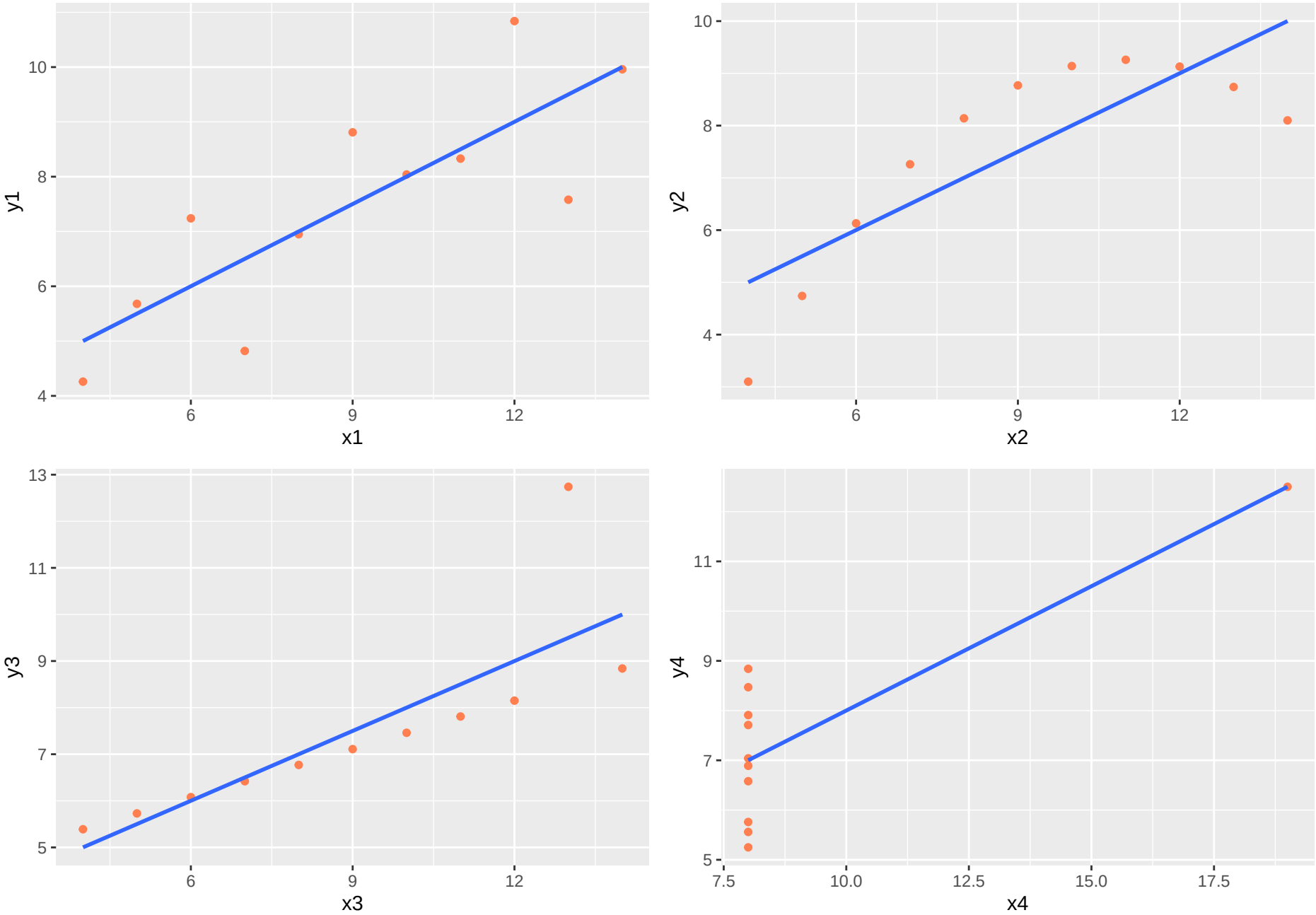
```
##  x1  x2  x3  x4  y1  y2  y3  y4
## 9.0 9.0 9.0 9.0 7.5 7.5 7.5 7.5
```

SD

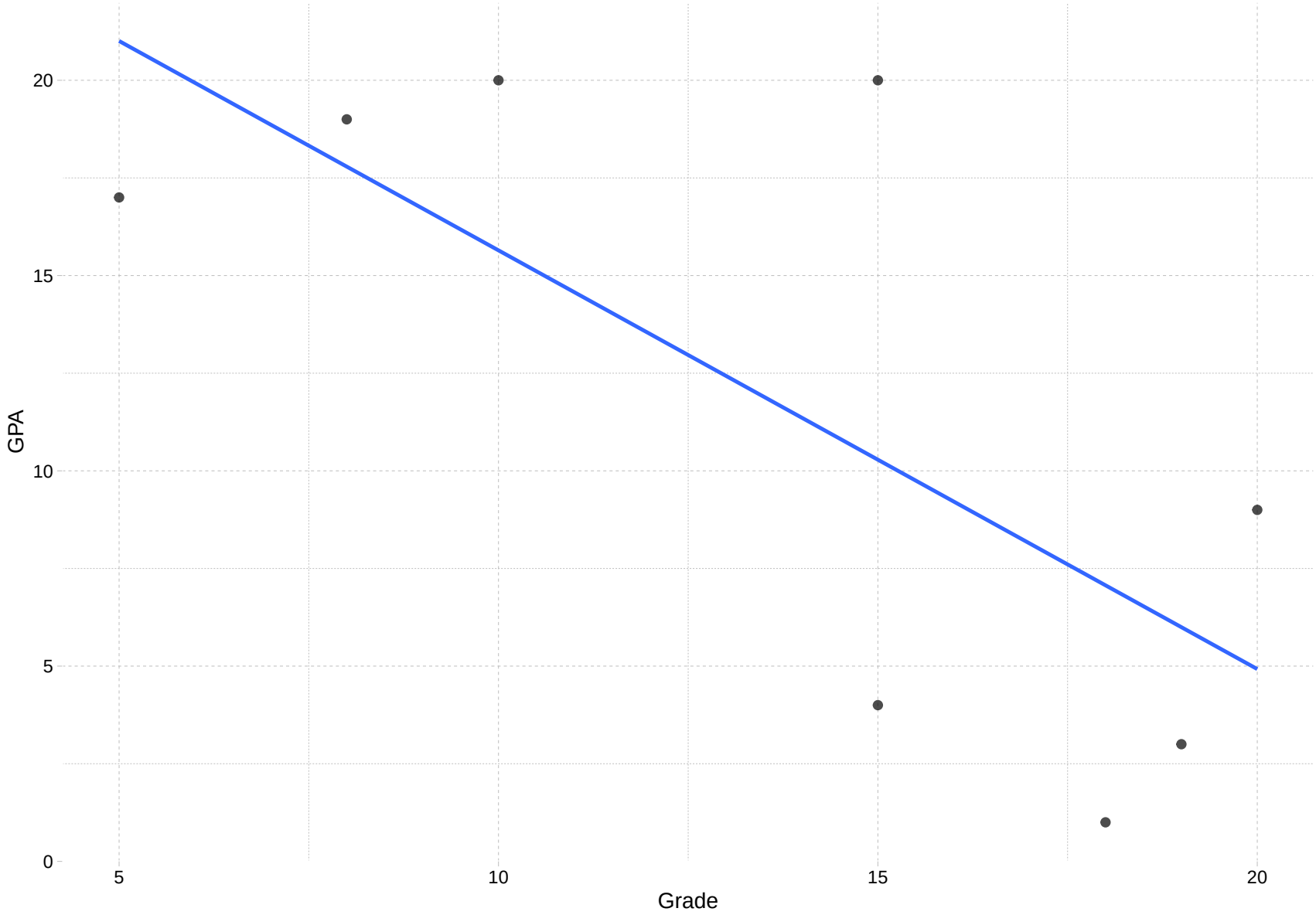
```
##   x1   x2   x3   x4   y1   y2   y3   y4
## 3.32 3.32 3.32 3.32 2.03 2.03 2.03 2.03
```

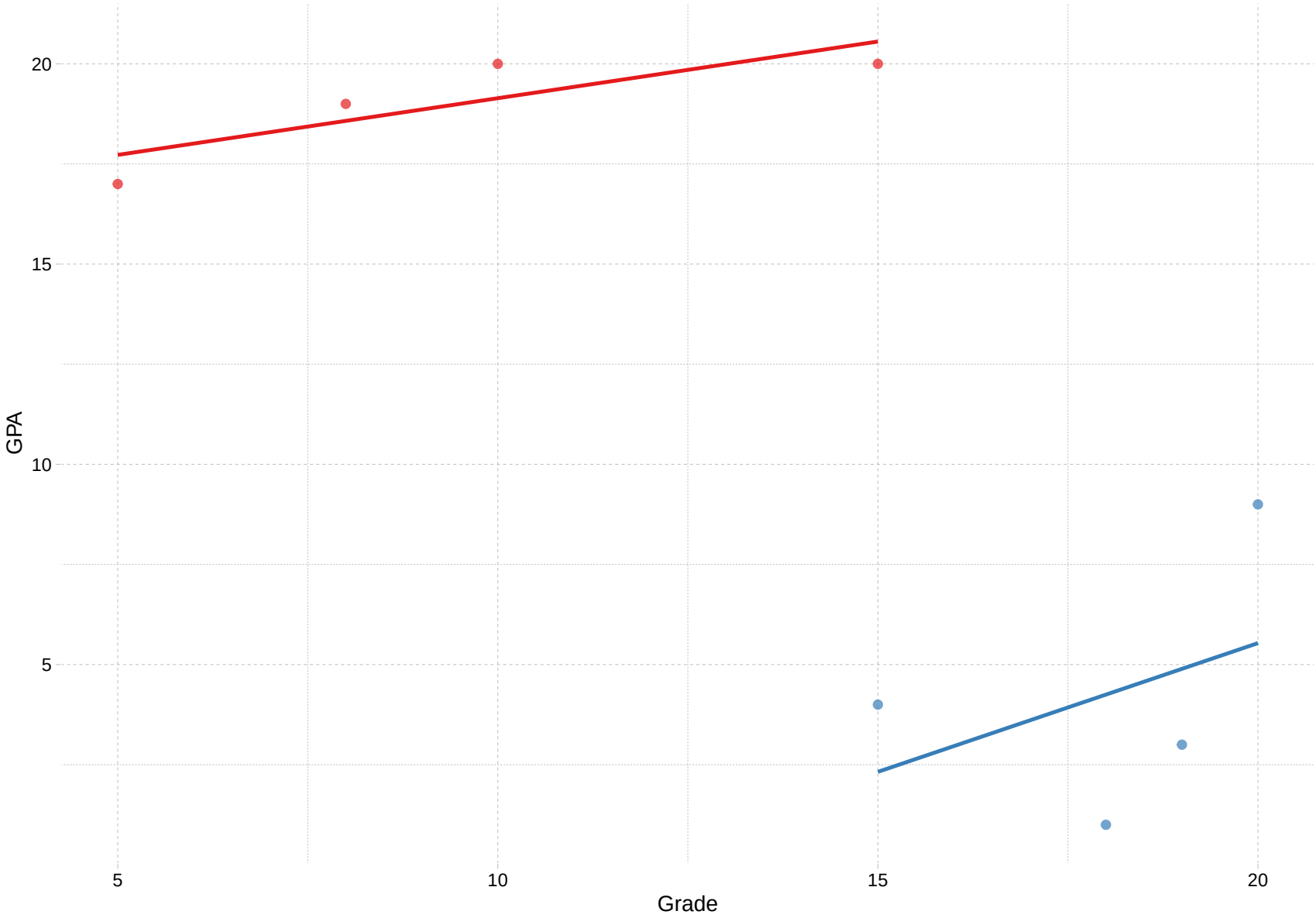
Correlation

```
##     x1     x2     x3     x4     y1     y2     y3     y4
##  1.000  1.000  1.000 -0.500  0.816  0.816  0.816 -0.314
```

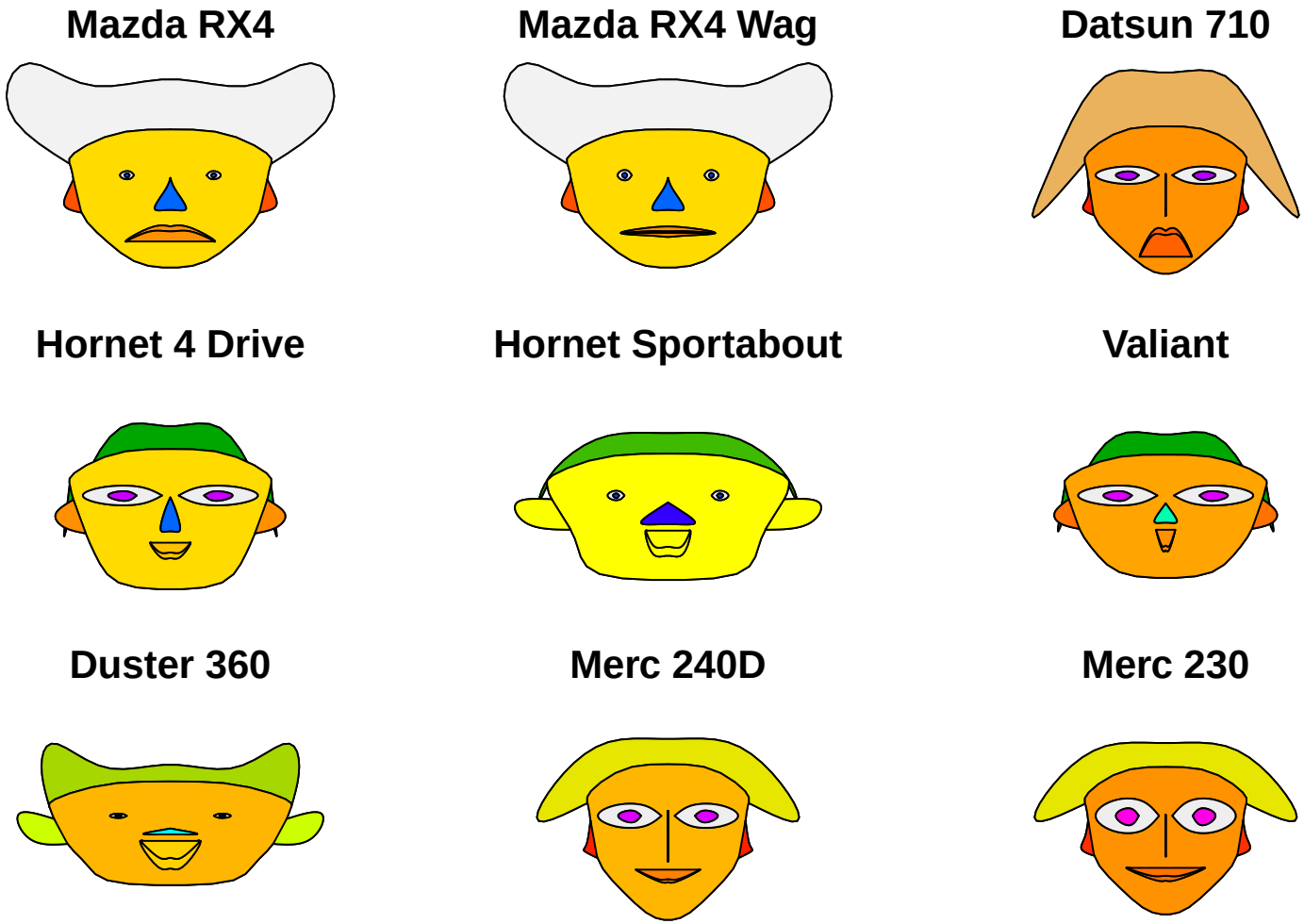

y1
x1
y2
x2
y3
x3
y4
x4

- Simpson's paradox




20
15
10
5
GPA
5
10
15
20
Grade

- Chernoff faces



```
## effect of variables:
##  modified item       Var
##  "height of face   " "mpg"
##  "width of face    " "cyl"
##  "structure of face" "disp"
##  "height of mouth  " "hp"
##  "width of mouth   " "drat"
##  "smiling          " "wt"
##  "height of eyes   " "qsec"
##  "width of eyes    " "vs"
##  "height of hair   " "am"
##  "width of hair   "  "gear"
##  "style of hair   "  "carb"
##  "height of nose  "  "mpg"
##  "width of nose   "  "cyl"
##  "width of ear    "  "disp"
##  "height of ear   "  "hp"
```

And finally, do you agree that visualization and summary stats are stronger than our brains?

Time for Quiz

Go to **socrative.com** to check your knowledge :)