# Lesson 03 EDA

Lusine Zilfimian

February 20 (Tuesday), 2024

# Data Preprocessing

- Aggregation: to reduce the memory and provide high-level view

# Data Preprocessing

- Aggregation: to reduce the memory and provide high-level view
- Sampling (types)

# Data Preprocessing

- Aggregation: to reduce the memory and provide high-level view
- Sampling (types)
- Feature subset selection (approaches)

# Data Preprocessing

- Aggregation: to reduce the memory and provide high-level view
- Sampling (types)
- Feature subset selection (approaches)
  - Redundant

# Data Preprocessing

- Aggregation: to reduce the memory and provide high-level view
- Sampling (types)
- Feature subset selection (approaches)
  - Redundant
  - Irrelevant features

# Data Preprocessing

- Aggregation: to reduce the memory and provide high-level view
- Sampling (types)
- Feature subset selection (approaches)
  - Redundant
  - Irrelevant features
- Feature creation

# Data Preprocessing

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

# Data Preprocessing

- Aggregation: to reduce the memory and provide high-level view
- Sampling (types)
- Feature subset selection (approaches)
  - Redundant
  - Irrelevant features
- Feature creation
- Discretization and binarization
  - Unsupervised Discretization

# Data Preprocessing

- Aggregation: to reduce the memory and provide high-level view
- Sampling (types)
- Feature subset selection (approaches)
  - Redundant
  - Irrelevant features
- Feature creation
- Discretization and binarization
  - Unsupervised Discretization
  - Supervised Discretization

# Data Preprocessing

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

# Exploring Data

- Exploring Data: Univariate Summary Statistics

# Exploring Data

- Exploring Data: Univariate Summary Statistics
- Exploring Data: Multivariate Summary Statistics

# 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%
##   1   2   4
```

## Mean and Median

```
## Mean:  7.888889
```

```
## Median:  2
```

# Questions

- Is the sample mean $x$ a weighted average of the set of its distinct values.

# Questions

- Is the sample mean $x$ a weighted average of the set of its distinct values.
- Is the statement correct? If each data value is increased/multiplied by a constant number c, then this causes the sample mean also to be increased/multiplied by c.

# Questions

- Is the sample mean $x$ a weighted average of the set of its distinct values.
- Is the statement correct? If each data value is increased/multiplied by a constant number c, then this causes the sample mean also to be increased/multiplied by c.
- Suppose that we have two distinct samples of sizes $n_1$ and $n_2$. If the sample mean of the first sample is $x_1$ and that of the second is $x_2$, what is the sample mean of the combined sample of size $n_1 + n_2$?

## Exploring Data: Summary Statistics

### Range and Variance

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

```
## Variance:  251.3611
```

```
## SD:  15.85437
```

### IQR and MAD

```
## IQR:  3
```

```
## MAD:  1.4826
```

# Exploring Data: Summary Statistics

- **Covariance and Correlation**

# Exploring Data: Summary Statistics

- **Covariance and Correlation**
- Their properties

# Questions:

- We have 2 Datasets: $x : 0, 1, 2$, and $y : -1, 2, \alpha$. Find all values of $\alpha$ such that the correlation coefficient $cor(x, y)$ is maximal (equals to 1).

# Questions:

- We have 2 Datasets: $x : 0, 1, 2$, and $y : -1, 2, \alpha$. Find all values of $\alpha$ such that the correlation coefficient $cor(x, y)$ is maximal (equals to 1).
- (T/F/ND) The sample variance remains unchanged when a constant is added to each data value.

# Questions:

- We have 2 Datasets: $x : 0, 1, 2$, and $y : -1, 2, \alpha$. Find all values of $\alpha$ such that the correlation coefficient $cor(x, y)$ is maximal (equals to 1).
- (T/F/ND) The sample variance remains unchanged when a constant is added to each data value.
- (T/F/ND) The variance of the sample mean increases while the sample size increases.

# 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
## x1    1.000  1.000  1.000 -0.500  0.816  0.816  0.816 -0.314
## x2    1.000  1.000  1.000 -0.500  0.816  0.816  0.816 -0.314
## x3    1.000  1.000  1.000 -0.500  0.816  0.816  0.816 -0.314
## x4   -0.500 -0.500 -0.500  1.000 -0.529 -0.718 -0.345  0.817
## y1    0.816  0.816  0.816 -0.529  1.000  0.750  0.469 -0.489
## y2    0.816  0.816  0.816 -0.718  0.750  1.000  0.588 -0.478
## y3    0.816  0.816  0.816 -0.345  0.469  0.588  1.000 -0.155
## y4   -0.314 -0.314 -0.314  0.817 -0.489 -0.478 -0.155  1.000
```

# Linear Regression

```
results_reg <- function(y,x){

  mod <- lm(y~x)
  sum<- summary(mod)
  cat("Coeff:", paste(round(mod$coefficients,5),
                      collapse = " "),
      "R^2:", round(sum$r.squared, 4))
}
```

# Linear Regression

```
results_reg(anscombe$y1, anscombe$x1)
```

```
## Coeff: 3.00009 0.50009 R^2: 0.6665
```

```
results_reg(anscombe$y2, anscombe$x2)
```

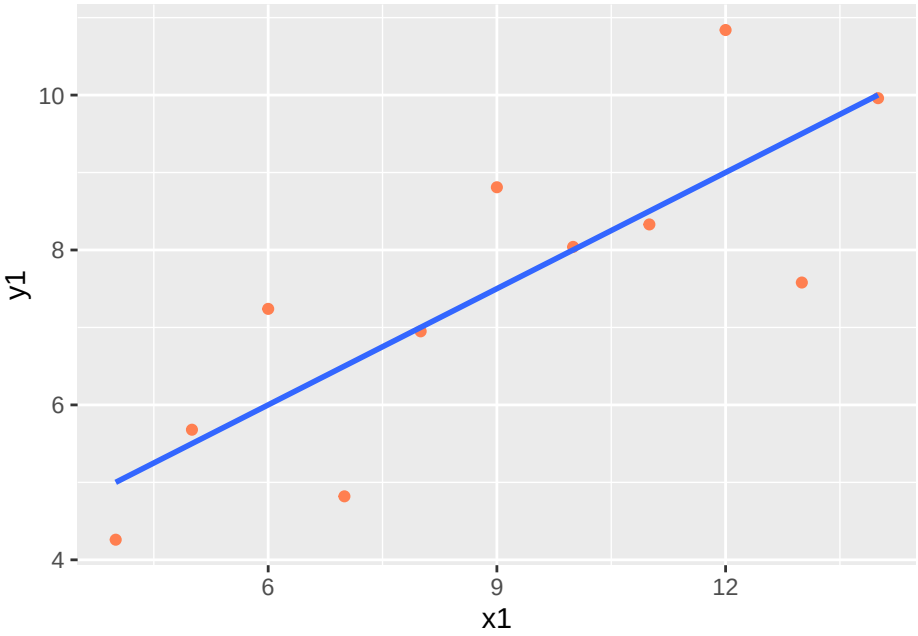
```
## Coeff: 3.00091 0.5 R^2: 0.6662
```

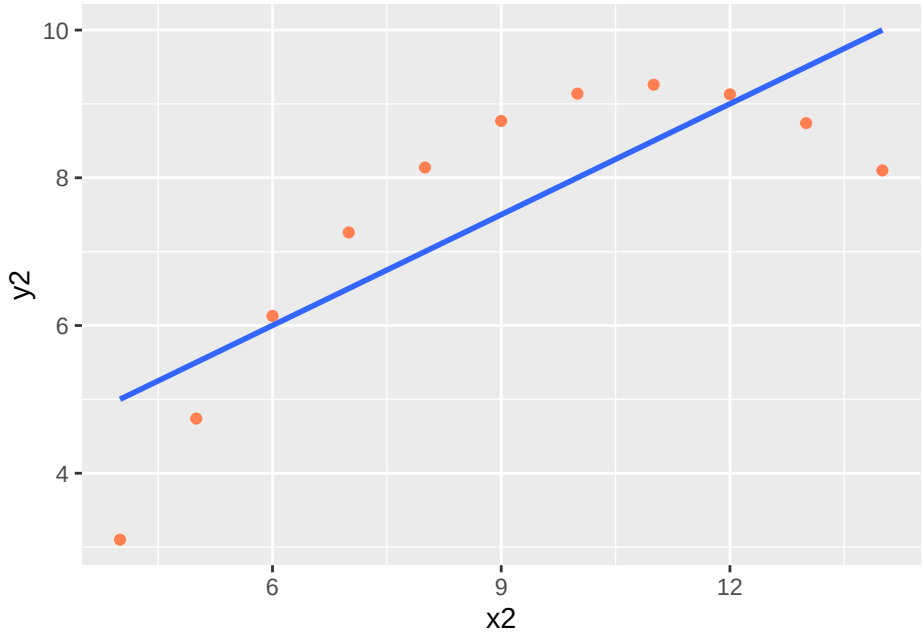
```
results_reg(anscombe$y3, anscombe$x3)
```

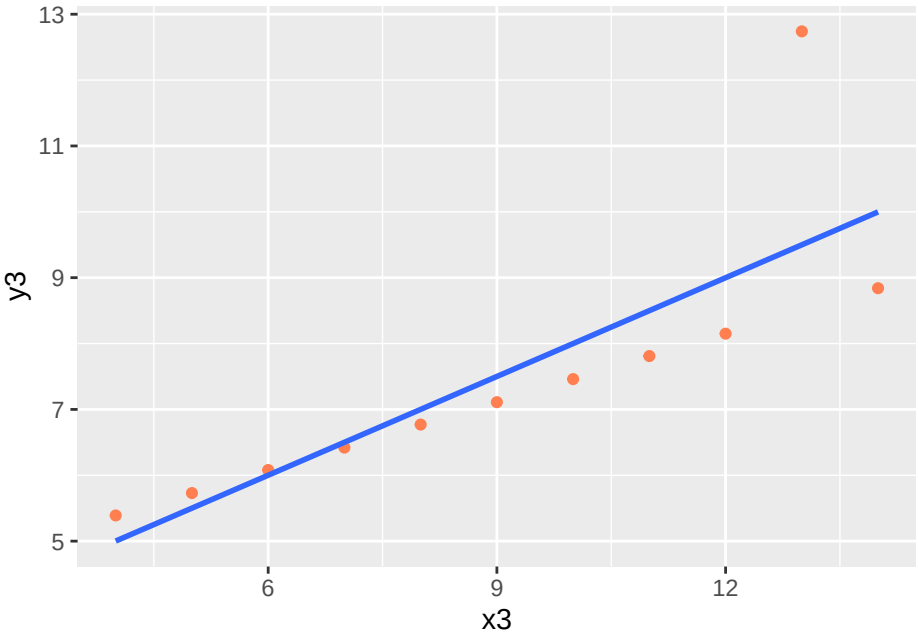
```
## Coeff: 3.00245 0.49973 R^2: 0.6663
```

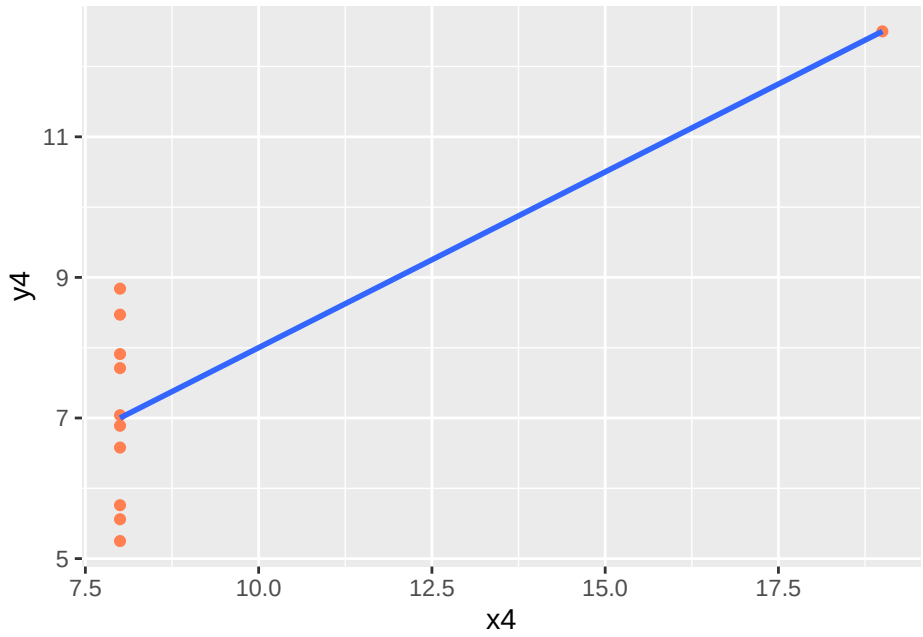
```
results_reg(anscombe$y4, anscombe$x4)
```

```
## Coeff: 3.00173 0.49991 R^2: 0.6667
```


y1
x1
10
8
6
4
6
9
12


y2
x2
10
8
6
4
6
9
12


y3
x3
13
11
9
7
5
6
9
12


y4
x4
11
9
7
5
7.5
10.0
12.5
15.0
17.5

# Correlation

- Correlation is not causation!

# Correlation

- Correlation is not causation!
- Be aware of spurious correlations.

# 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

# 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

# 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

# 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.

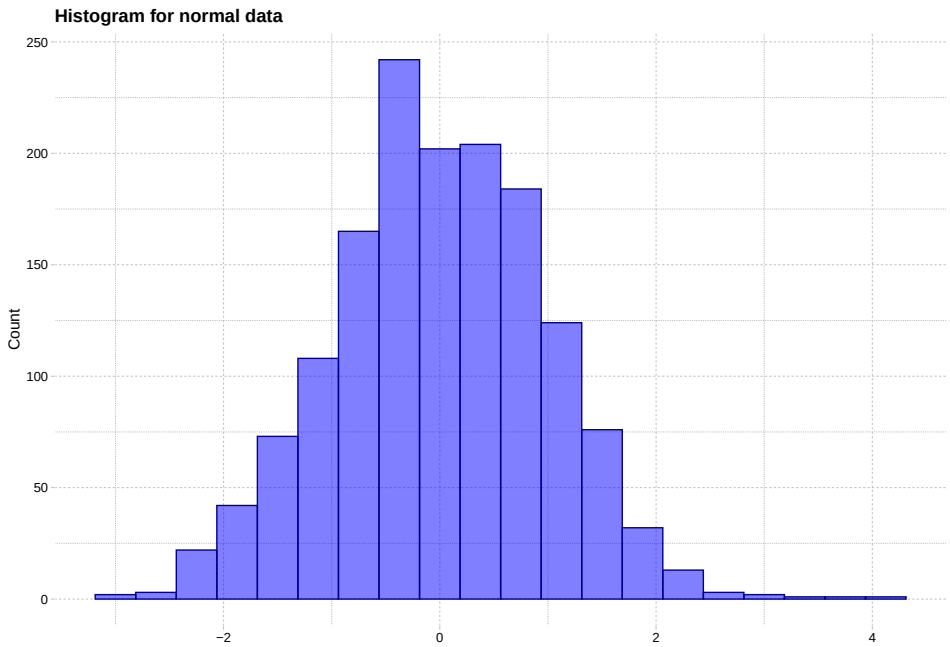
- Histogram, Stem and Leaf plot
- Bar Plot
- Box Plot
- Scatter Plot

# Exploring Data: Visualization

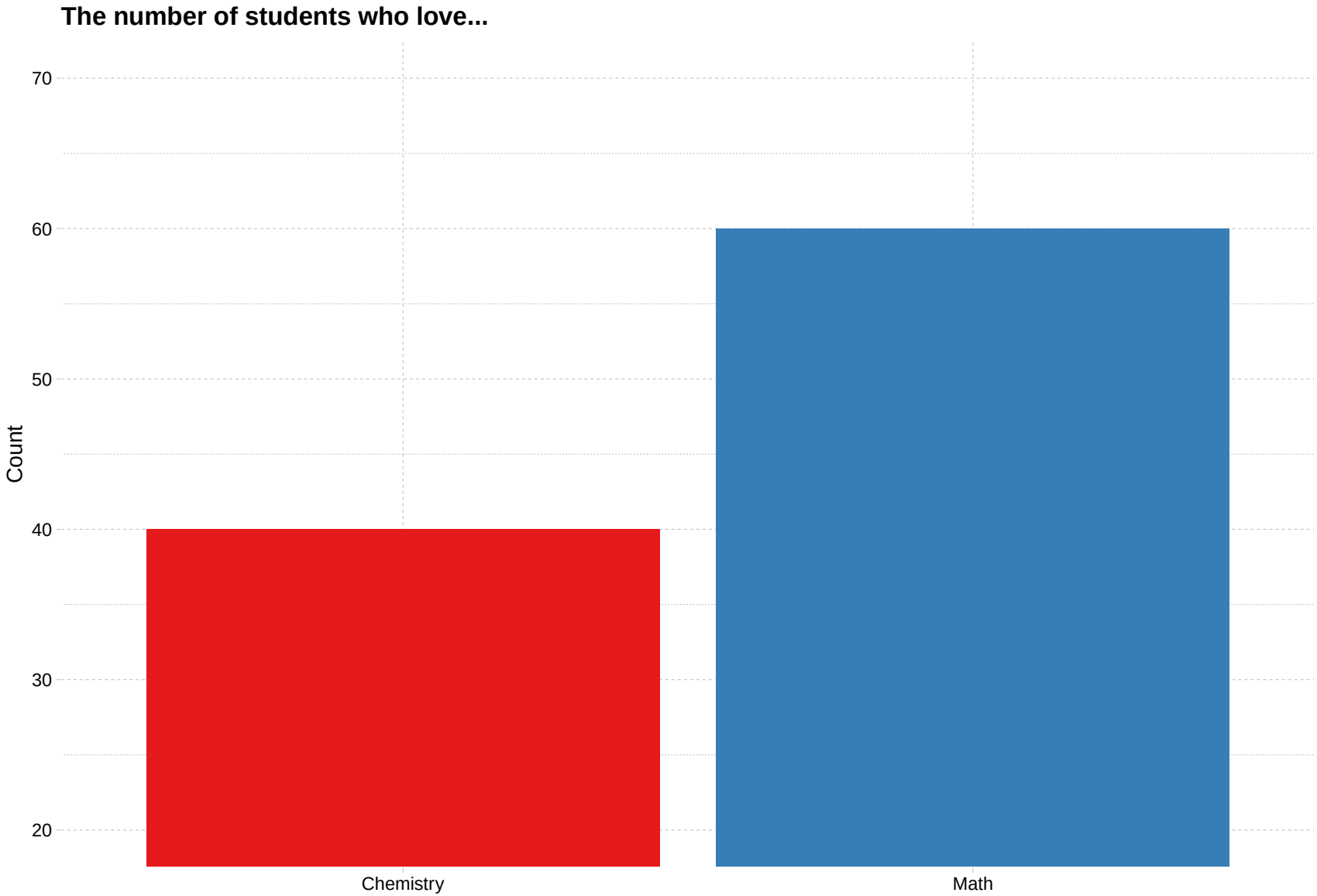
Visualizations of the data may be **the best way** of finding patterns of interest since a person cannot get an insight from the list of numbers.

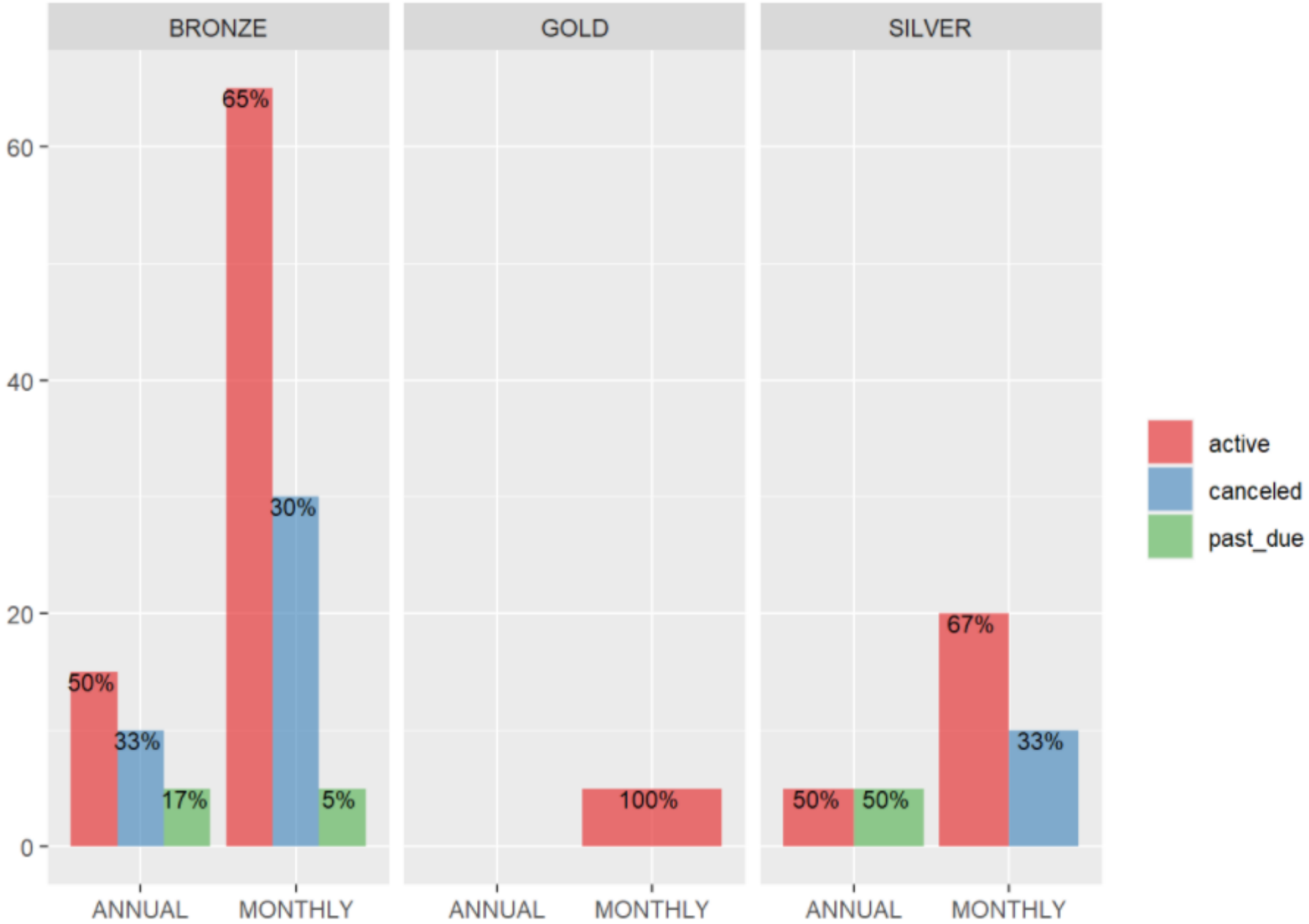
- Histogram, Stem and Leaf plot
- Bar Plot
- Box Plot
- Scatter Plot
- Time Series (Line Graph (Do we need to separate it?))

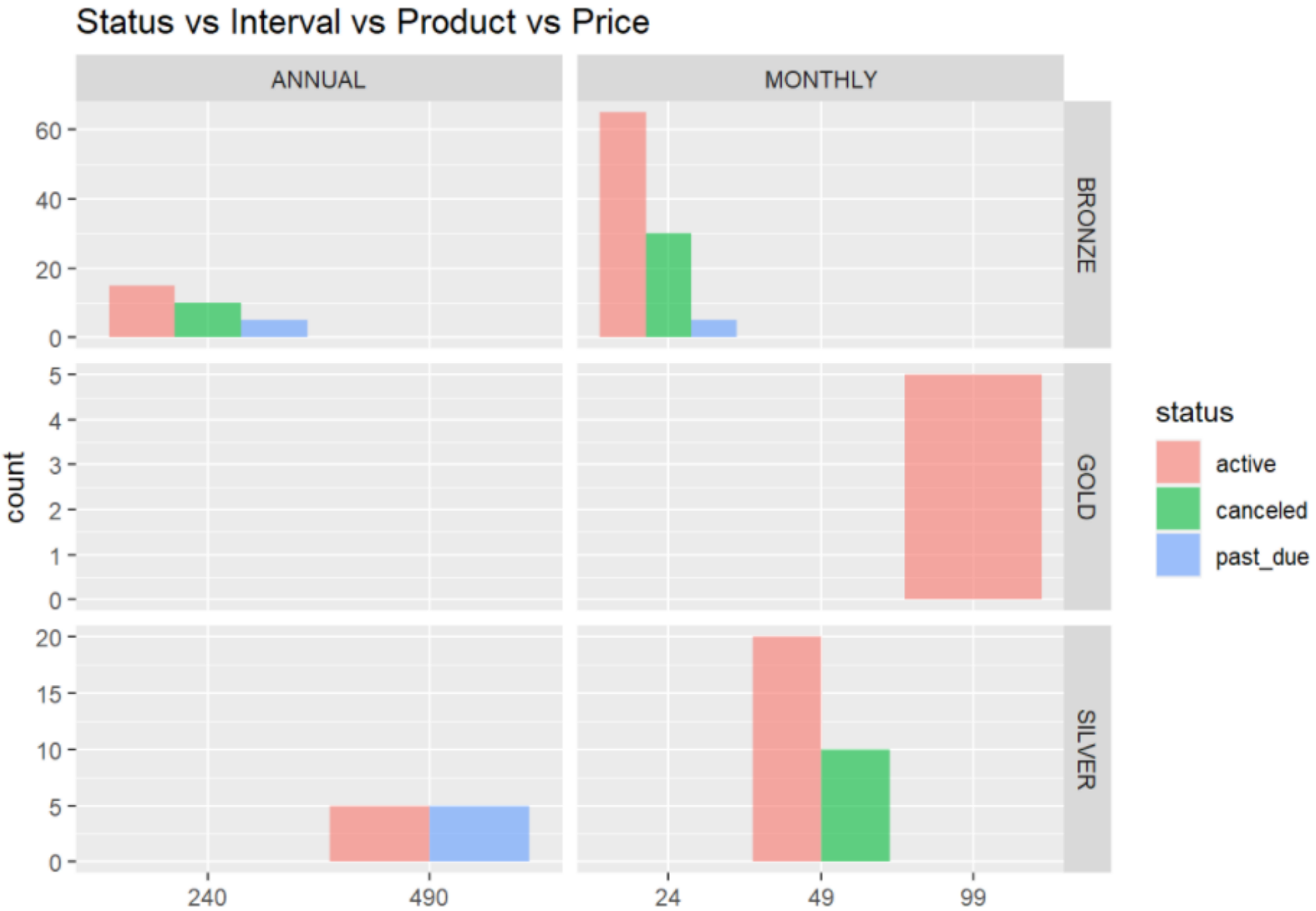
```
##
##   The decimal point is at the |
##
##   -1 | 0
##   -0 | 888866
##   -0 | 32
##    0 | 12344
##    0 | 5
```



- Bar plot


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


BRONZE
GOLD
SILVER
60
40
20
0
65%
30%
50%
33%
17%
5%
100%
50%
50%
67%
33%
ANNUAL
MONTHLY
ANNUAL
MONTHLY
ANNUAL
MONTHLY
active
canceled
past_due

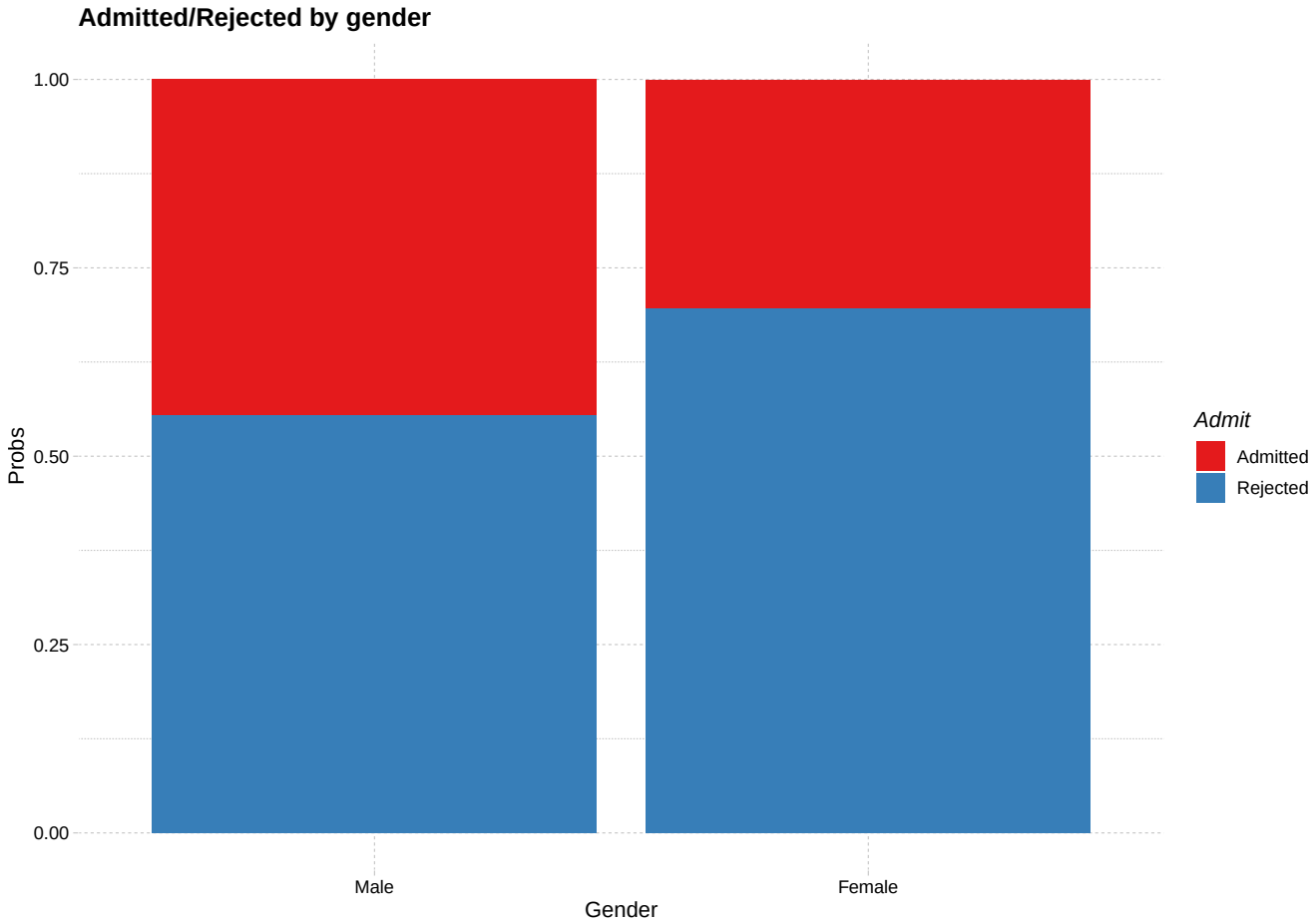

Status vs Interval vs Product vs Price
ANNUAL
MONTHLY
BRONZE
GOLD
SILVER
count
60
40
20
0
5
4
3
2
1
0
20
15
10
5
0
240
490
24
49
99
status
active
canceled
past_due

**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

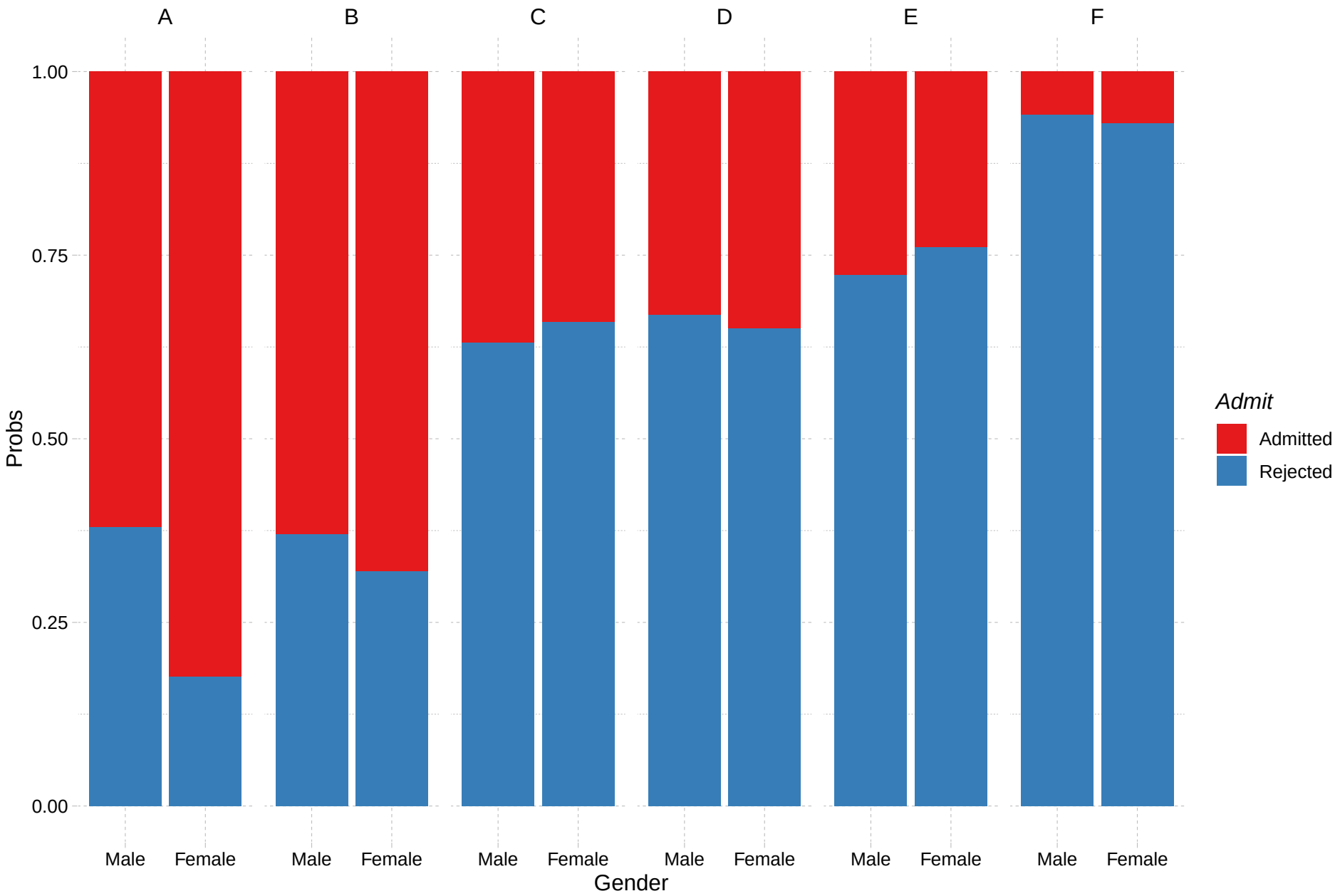
```
##          Admit
## Gender    Admitted Rejected
##   Male        1198     1493
##   Female       557     1278
```




# Proportional cross tabs

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

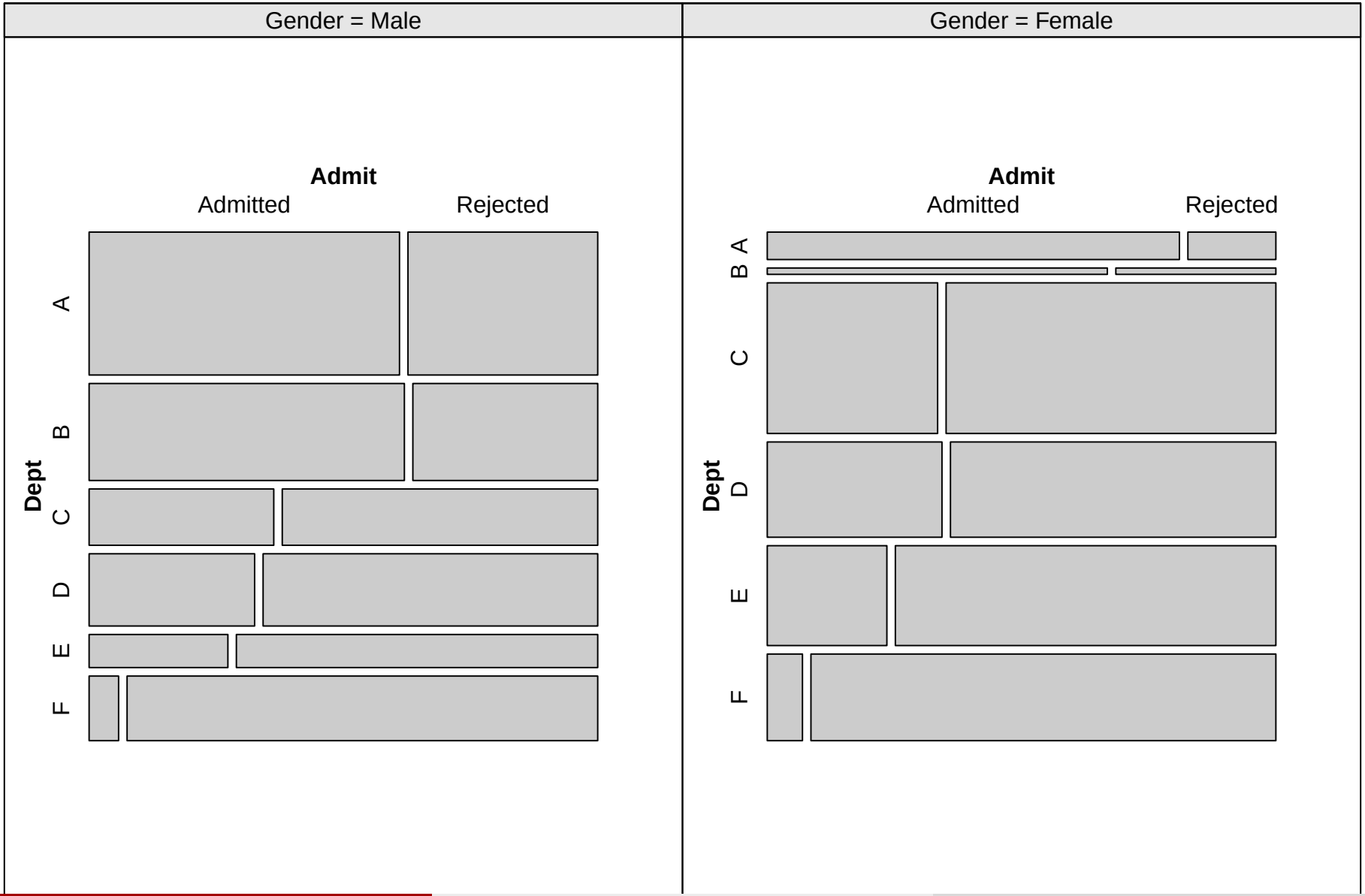
## Admitted/Rejected by gender and department



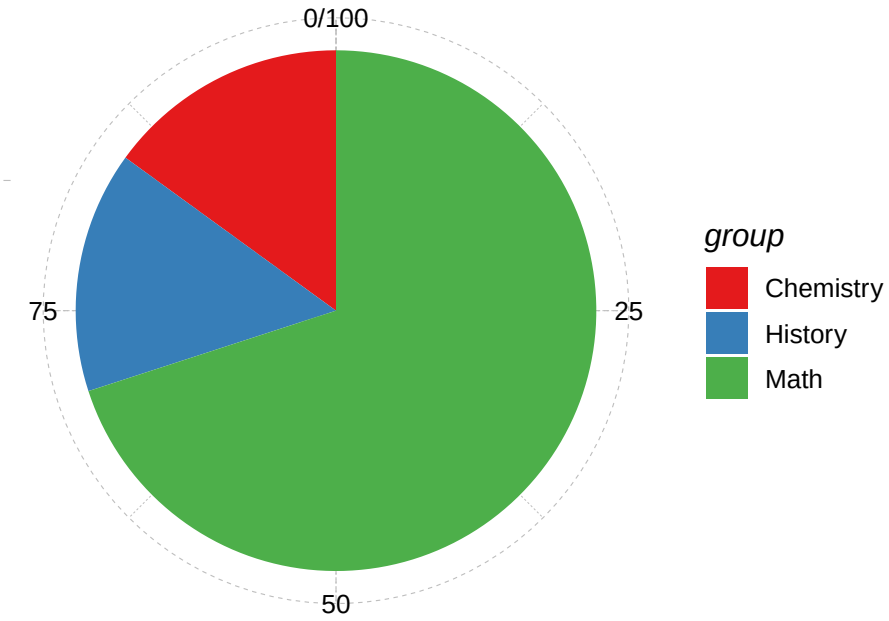
- Females used to apply more to departments with higher rejection rates

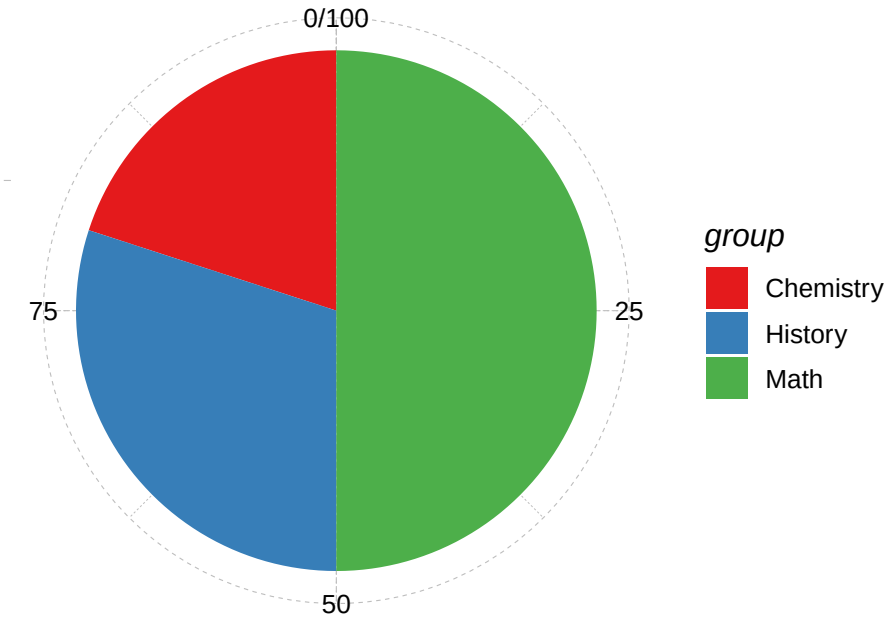
- Females used to apply more to departments with higher rejection rates
- Here we see that many more men apply to departments A and B which have high acceptance rates, while women apply to departments that are harder to get into.

```
## Loading required package: grid
```



- Pie Chart

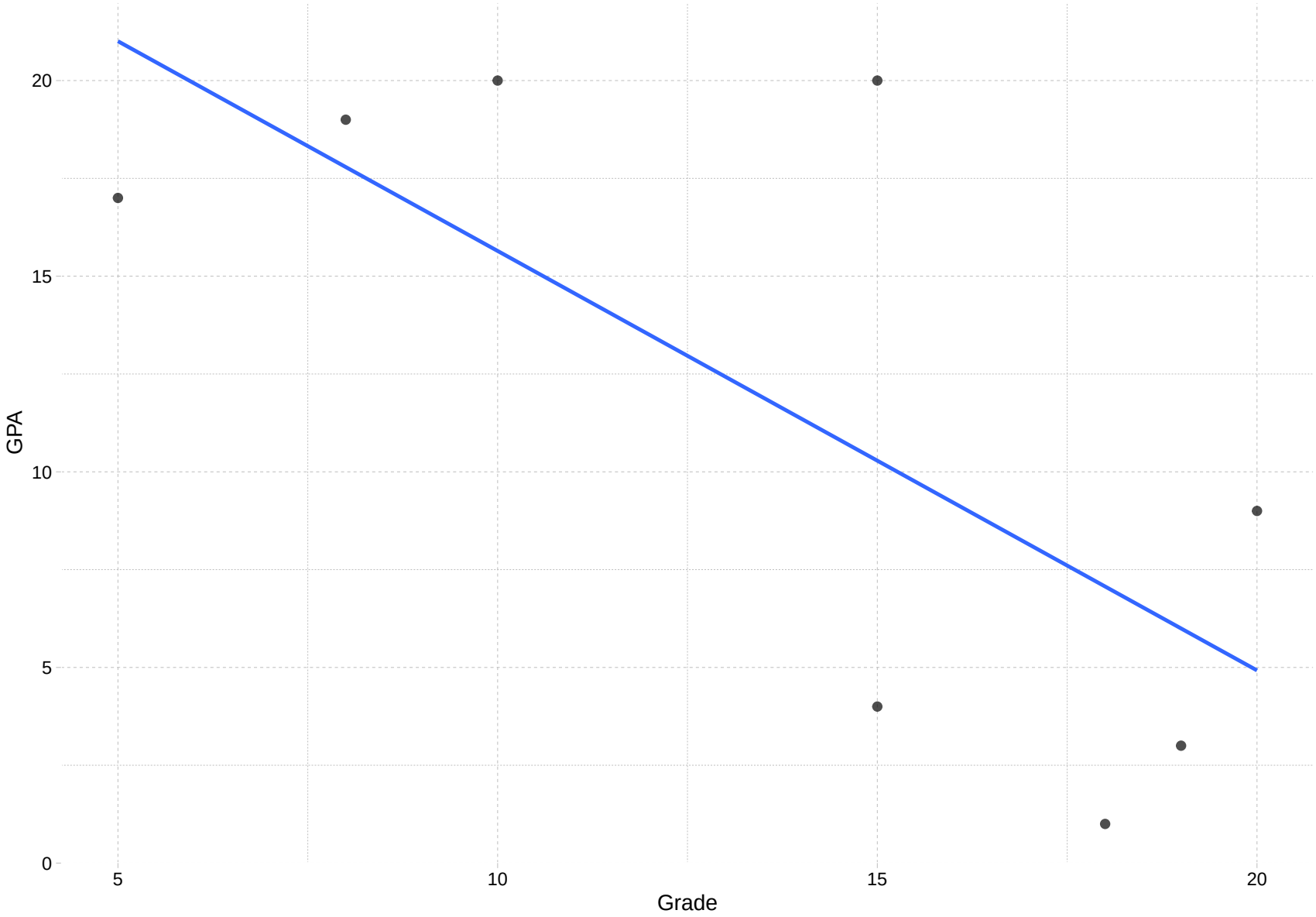

0/100
25
50
75
group
Chemistry
History
Math

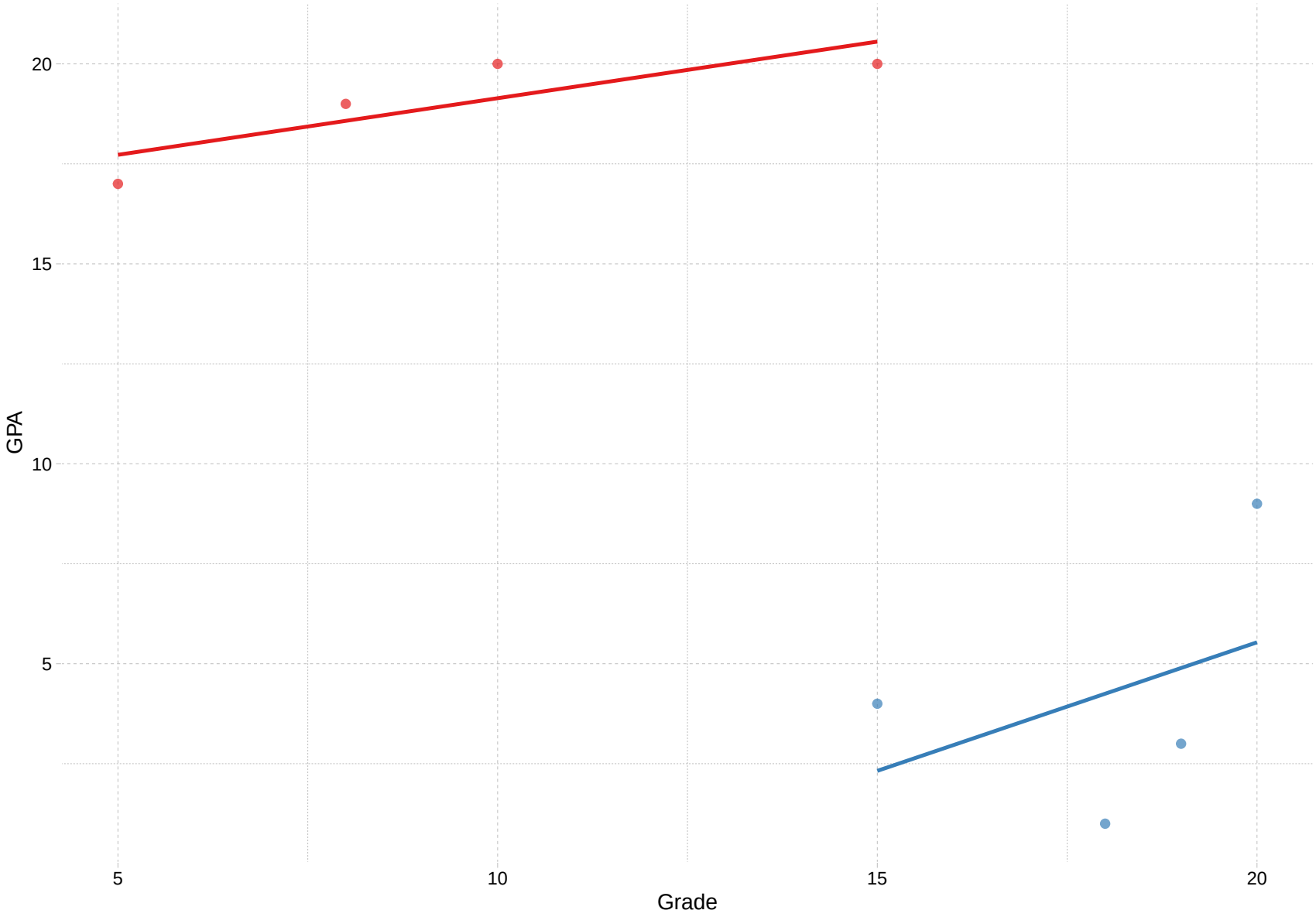

0/100
25
50
75
group
Chemistry
History
Math

- Boxplot

- Boxplot
- Scatter Plot

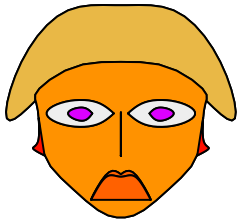
- Simpson's paradox



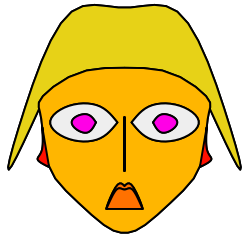

20
15
10
5
GPA
5
10
15
20
Grade

- Chernoff faces - each attribute is associated with a specific feature of a face.

**Honda Civic**

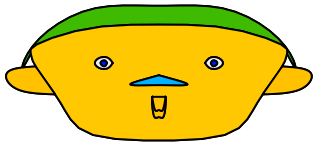


**Toyota Corolla**

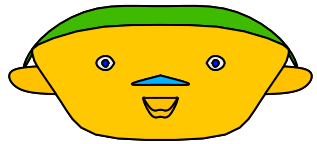


**Toyota Corona**

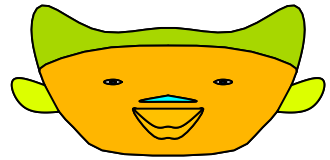
**Dodge Challenger**



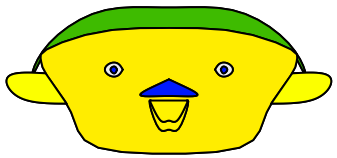
**AMC Javelin**



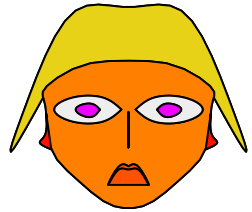
**Camaro Z28**



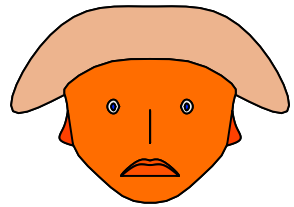
**Pontiac Firebird**



**Fiat X1−9**



**Porsche 914−2**



```
## 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 just looking at data or summary statistics?