Lesson 01 Intro to Data Mining and Data

Lusine Zilfimian

February 10 (Monday), 2020

Welcome to the world of Data Mining!

- Syllabus Highlights and Info about the Course

Welcome to the world of Data Mining!

- Syllabus Highlights and Info about the Course
- Intro to Data and Data Types

Welcome to the world of Data Mining!

- Syllabus Highlights and Info about the Course
- Intro to Data and Data Types
- Exploring Data: Summary Statistics

Welcome to the world of Data Mining!

- Syllabus Highlights and Info about the Course
- Intro to Data and Data Types
- Exploring Data: Summary Statistics
- Exploring Data: Visualization

Welcome to the world of Data Mining!

- Syllabus Highlights and Info about the Course
- Intro to Data and Data Types
- Exploring Data: Summary Statistics
- Exploring Data: Visualization
- Ungraded Quiz

Syllabus highlights

- Course name: **Data Mining**

Syllabus highlights

- Course name: **Data Mining**
- Number of Credits: **3**

Syllabus highlights

- Course name: **Data Mining**
- Number of Credits: **3**
- Github Page: **click on me**

Syllabus highlights

- Course name: **Data Mining**
- Number of Credits: **3**
- Github Page: **click on me**
- Syllabus: **uploaded to our Github page**

Syllabus highlights

- Course name: **Data Mining**
- Number of Credits: **3**
- Github Page: **click on me**
- Syllabus: **uploaded to our Github page**
- Textbooks: **let me know if you can not find/download them**

Syllabus highlights

- Course name: **Data Mining**
- Number of Credits: **3**
- Github Page: **click on me**
- Syllabus: **uploaded to our Github page**
- Textbooks: **let me know if you can not find/download them**
- Language and Software: **R** and **R Studio**

Syllabus highlights

- Course name: **Data Mining**
- Number of Credits: **3**
- Github Page: **click on me**
- Syllabus: **uploaded to our Github page**
- Textbooks: **let me know if you can not find/download them**
- Language and Software: **R** and **R Studio**
- R Textbooks: **see in syllabus**

Syllabus highlights

- Course name: **Data Mining**
- Number of Credits: **3**
- Github Page: **click on me**
- Syllabus: **uploaded to our Github page**
- Textbooks: **let me know if you can not find/download them**
- Language and Software: **R** and **R Studio**
- R Textbooks: **see in syllabus**
- Question: **Do we need lab sessions for R, R Markdown, R Shiny, and Github? (think about it).**

Syllabus more important highlights

- HW: **(almost) weekly**

$$Final\ Grade = 0.2(HW + ME + FP) + 0.3FE + 0.1Q$$

Syllabus more important highlights

- HW: **(almost) weekly**
- Exams: **Midterm + Final Exam**

$$Final\ Grade = 0.2(HW + ME + FP) + 0.3FE + 0.1Q$$

Syllabus more important highlights

- HW: **(almost) weekly**
- Exams: **Midterm + Final Exam**
- Final Project: **the guideline will be uploaded on Github**

$$Final\ Grade = 0.2(HW + ME + FP) + 0.3FE + 0.1Q$$

Syllabus more important highlights

- HW: **(almost) weekly**
- Exams: **Midterm + Final Exam**
- Final Project: **the guideline will be uploaded on Github**
- Quizzes: **:)**

$$Final\ Grade = 0.2(HW + ME + FP) + 0.3FE + 0.1Q$$

Syllabus more important highlights

- HW: **(almost) weekly**
- Exams: **Midterm + Final Exam**
- Final Project: **the guideline will be uploaded on Github**
- Quizzes: **:)**
- Grading policy:

$$Final\ Grade = 0.2(HW + ME + FP) + 0.3FE + 0.1Q$$

Based on my previous experience...

- No Makeups for Quizzes.

Based on my previous experience...

- No Makeups for Quizzes.
- No late HWs: I respect your time and expect the same from you.

Based on my previous experience...

- No Makeups for Quizzes.
- No late HWs: I respect your time and expect the same from you.
- Cheating: Be honest! Any similarities, which can be considered as cheated, will not be graded.

Based on my previous experience...

- No Makeups for Quizzes.
- No late HWs: I respect your time and expect the same from you.
- Cheating: Be honest! Any similarities, which can be considered as cheated, will not be graded.
- Feel free to ask questions and have comments.

Questions?

Intro to Data Mining

What is DM? What is the difference between DM and Statistics?

- Data mining is the process of automatically discovering useful information in large data repositories.

Intro to Data Mining

What is DM? What is the difference between DM and Statistics?

- Data mining is the process of automatically discovering useful information in large data repositories.
- **Not all information** discovery tasks are considered to be data mining.

The process of converting raw data into useful information:

- Data Cleaning

The process of converting raw data into useful information:

- Data Cleaning
- Data Integration and Selection

The process of converting raw data into useful information:

- Data Cleaning
- Data Integration and Selection
- Data Transformation

The process of converting raw data into useful information:

- Data Cleaning
- Data Integration and Selection
- Data Transformation
- Data Mining Algorithms

The process of converting raw data into useful information:

- Data Cleaning
- Data Integration and Selection
- Data Transformation
- Data Mining Algorithms
- Evaluation (+ Interpretation)

The process of converting raw data into useful information:

- Data Cleaning
- Data Integration and Selection
- Data Transformation
- Data Mining Algorithms
- Evaluation (+ Interpretation)
- Visualization

Example of using DM in

- Bioinformatics

Example of using DM in

- Bioinformatics
- Marketing

Example of using DM in

- Bioinformatics
- Marketing
- Macroeconomics

Example of using DM in

- Bioinformatics
- Marketing
- Macroeconomics
- Education

Data Mining Tasks

Data mining tasks are generally divided into two major categories:

- Predictive task

There are two types of predictive modeling tasks:

Data Mining Tasks

Data mining tasks are generally divided into two major categories:

- Predictive task
- Descriptive task

There are two types of predictive modeling tasks:

Data Mining Tasks

Data mining tasks are generally divided into two major categories:

- Predictive task
- Descriptive task

There are two types of predictive modeling tasks:

- Classification

Data Mining Tasks

Data mining tasks are generally divided into two major categories:

- Predictive task
- Descriptive task

There are two types of predictive modeling tasks:

- Classification
- Regression

Course Structure: Topics at a glance

- Types of Data, Data Preprocessing

Course Structure: Topics at a glance

- Types of Data, Data Preprocessing
- Exploring Data

Course Structure: Topics at a glance

- Types of Data, Data Preprocessing
- Exploring Data
- Linear Regression

Course Structure: Topics at a glance

- Types of Data, Data Preprocessing
- Exploring Data
- Linear Regression
- Logistic Regression

Course Structure: Topics at a glance

- Types of Data, Data Preprocessing
- Exploring Data
- Linear Regression
- Logistic Regression
- Poisson Regression

Course Structure: Topics at a glance

- Types of Data, Data Preprocessing
- Exploring Data
- Linear Regression
- Logistic Regression
- Poisson Regression
- Regularization

Course Structure: Topics at a glance

- Types of Data, Data Preprocessing
- Exploring Data
- Linear Regression
- Logistic Regression
- Poisson Regression
- Regularization
- Classification

Course Structure: Topics at a glance

- Types of Data, Data Preprocessing
- Exploring Data
- Linear Regression
- Logistic Regression
- Poisson Regression
- Regularization
- Classification
- Cluster Analysis

Course Structure: Topics at a glance

- Types of Data, Data Preprocessing
- Exploring Data
- Linear Regression
- Logistic Regression
- Poisson Regression
- Regularization
- Classification
- Cluster Analysis
- Dimensionality Reduction

Course Structure: Topics at a glance

- Types of Data, Data Preprocessing
- Exploring Data
- Linear Regression
- Logistic Regression
- Poisson Regression
- Regularization
- Classification
- Cluster Analysis
- Dimensionality Reduction
- Hmm... other interesting topics

Types of Data, Data Preprocessing

What are an attribute and a measurement scale?

- **An attribute** is a property or characteristic of an object that may vary; either from one object to another or from one time to another.

Types of Data, Data Preprocessing

What are an attribute and a measurement scale?

- **An attribute** is a property or characteristic of an object that may vary; either from one object to another or from one time to another.
- **A measurement scale** is a rule (function) that associates a numerical or symbolic value with an attribute of an object.

Types of Attributes

We **CAN** define 4 types of attributes:

- Nominal

The definition of the attribute types is cumulative.

Types of Attributes

We **CAN** define 4 types of attributes:

- Nominal
- Ordinal

The definition of the attribute types is cumulative.

Types of Attributes

We **CAN** define 4 types of attributes:

- Nominal
- Ordinal
- Interval

The definition of the attribute types is cumulative.

Types of Attributes

We **CAN** define 4 types of attributes:

- Nominal
- Ordinal
- Interval
- Ratio

The definition of the attribute types is cumulative.

Types of Data Sets

- **Record data**

Long format:

```
##         Name  HW Grade
## 1     Lusine HW1    15
## 2      David HW1    16
## 3 Shoghakat HW1    17
## 4     Lusine HW2    18
## 5      David HW2    19
## 6 Shoghakat HW2    20
## 7     Lusine HW3    18
## 8      David HW3    17
## 9 Shoghakat HW3    20
```

Types of Data Sets

Wide format:

```
##          Name HW.1 HW.2 HW.3
## 1      Lusine   15   18   18
## 2       David   16   19   17
## 3 Shoghakat   17   20   20
```

Transaction data

```
##   ID                  Items
## 1  1     Lays, Coca-Cola
## 2  2 Lays, Beer, Sprite
## 3  3     Chocolate, Milk
```

Document-term Matrix

```
##   Document math is life
## 1       D1    1  2    3
## 2       D2    4  5    6
## 3       D3    0  7    8
```

- etc.

Data Quality

- Noise and Outlier

Data Quality

- Noise and Outlier
- Missing Value

Data Quality

- Noise and Outlier
- Missing Value
- Inconsistent Value

Data Quality

- Noise and Outlier
- Missing Value
- Inconsistent Value
- Duplicated data and Deduplication

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 (Redundant and Irrelevant features)

Data Preprocessing

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

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

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

Flequencies 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

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.

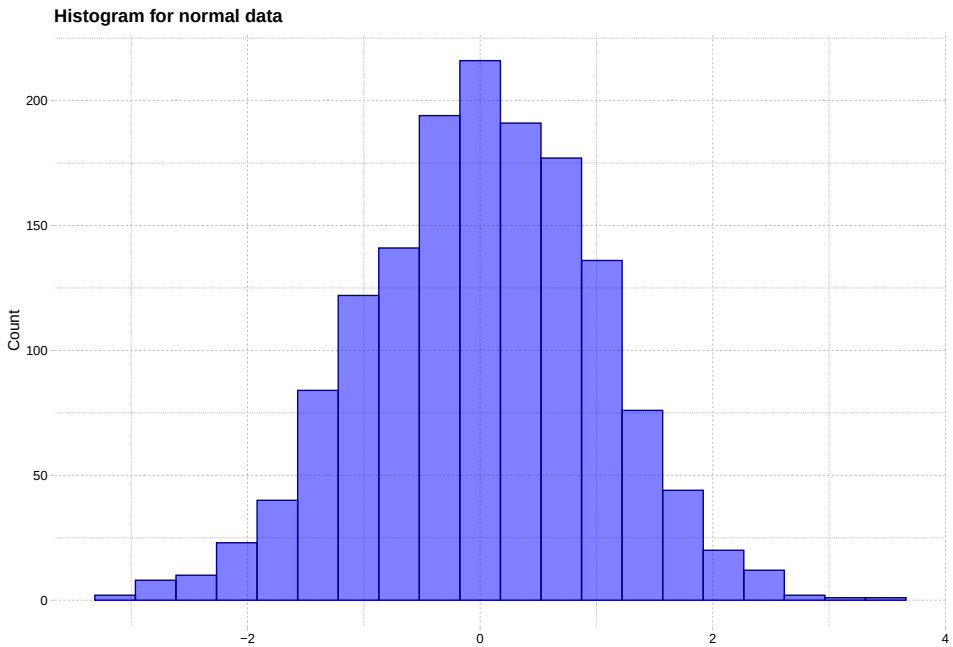
- Histogram, Stem and Leaf plot
- Bar Plot
- Box Plot
- Scatter Plot

Exploring Data: Visualization

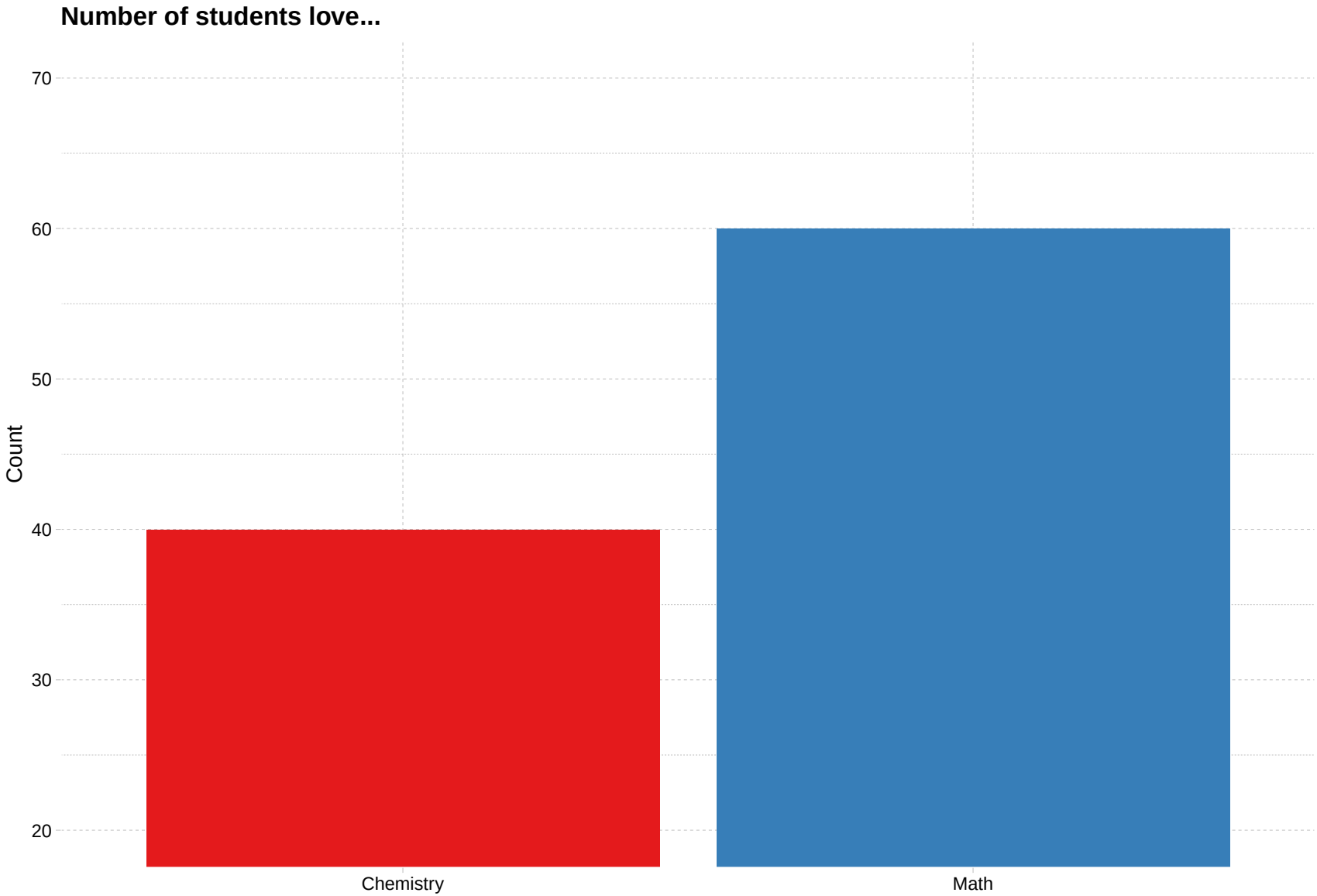
Visualizations of the data may be **the best way** of finding patterns of interest since a person cannot get an insight from the list of numbers.

- 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 |
##
##   -0 | 42554320
##    0 | 113338
##    2 |
##    4 | 2
```



- Bar plot

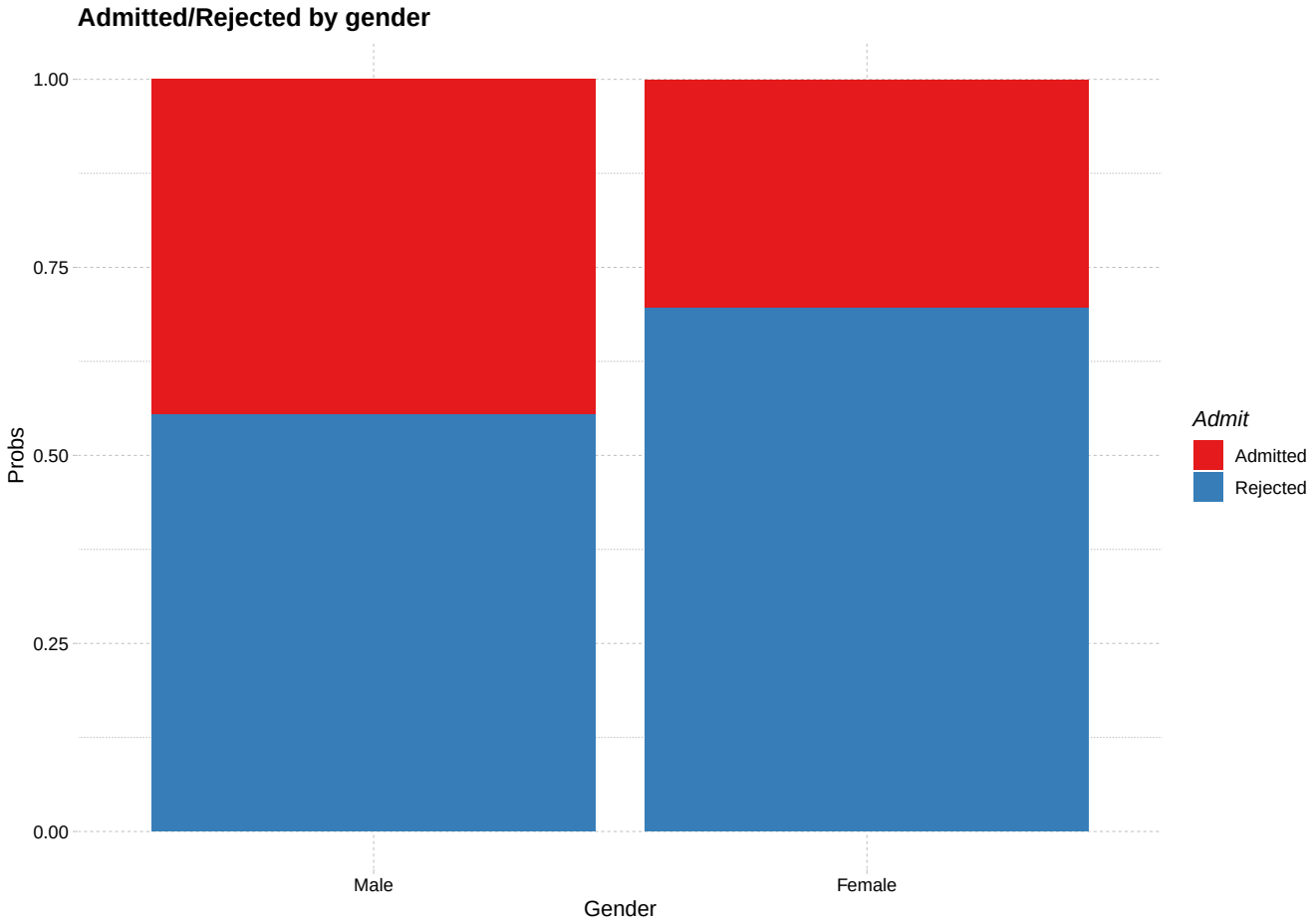

Number of students love...
70
60
50
40
30
20
Count
Chemistry
Math

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
```

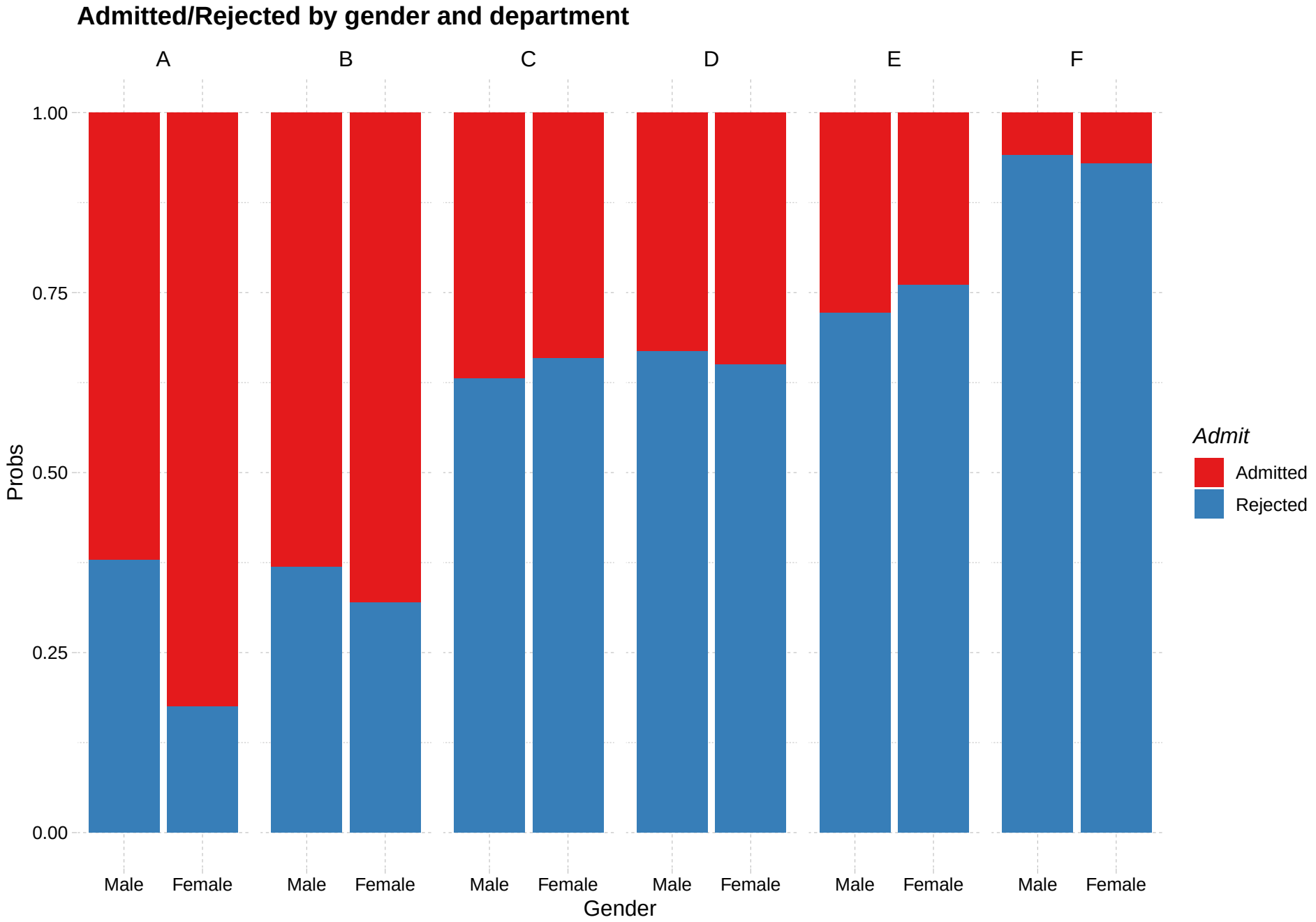
Cross tabs

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

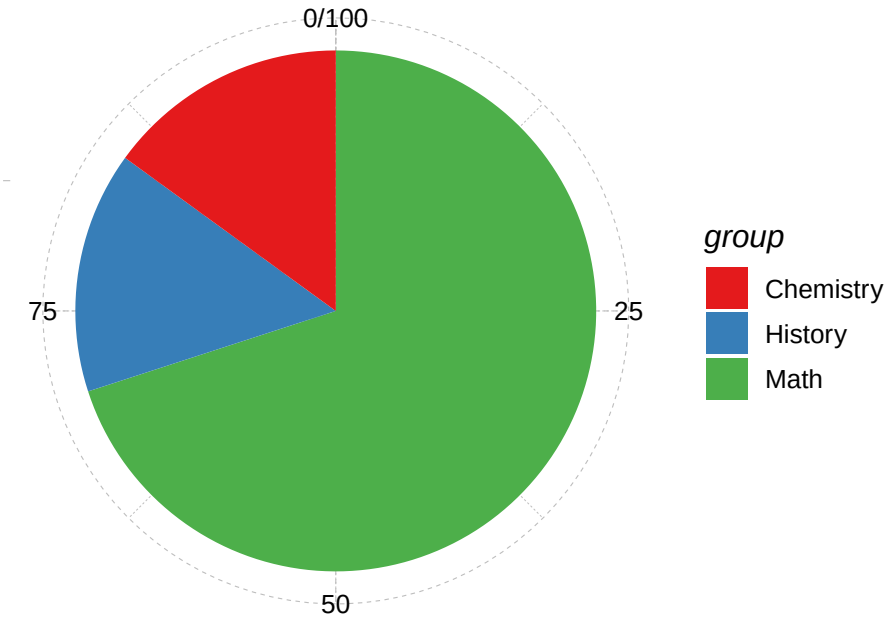


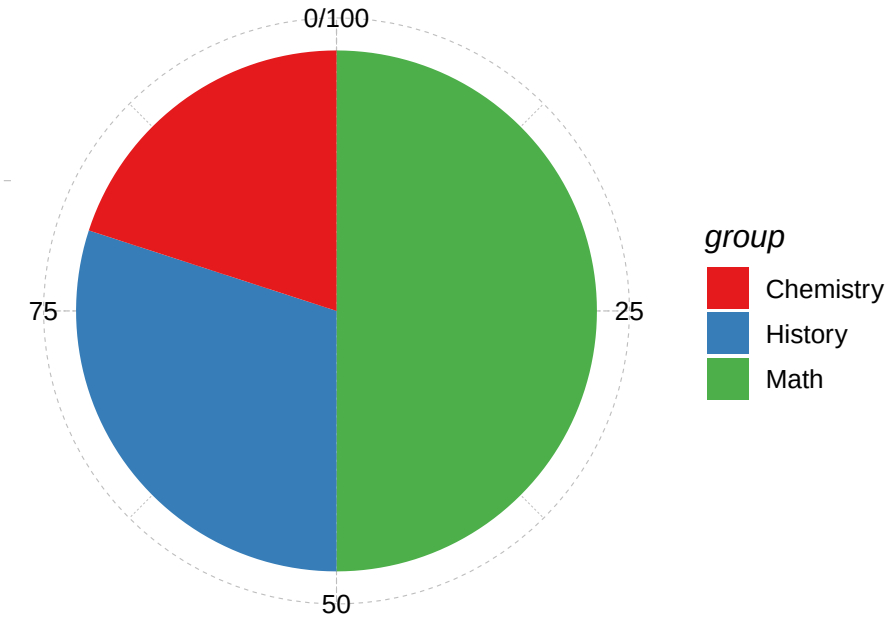
Proportional cross tabs

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

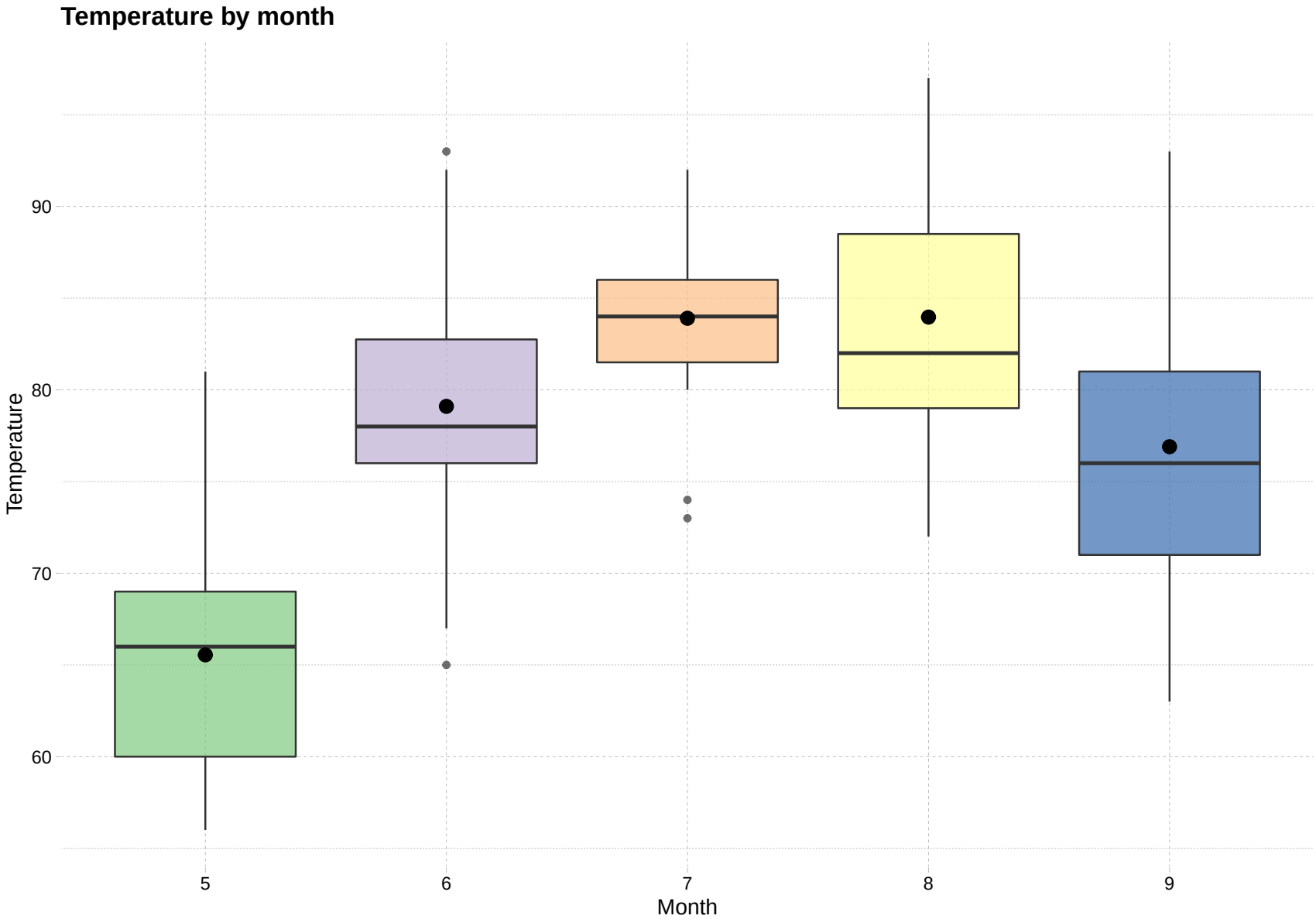

Admitted/Rejected by gender and department
A
B
C
D
E
F
1.00
0.75
0.50
0.25
0.00
Probs
Male
Female
Gender
Admit
Admitted
Rejected

- 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

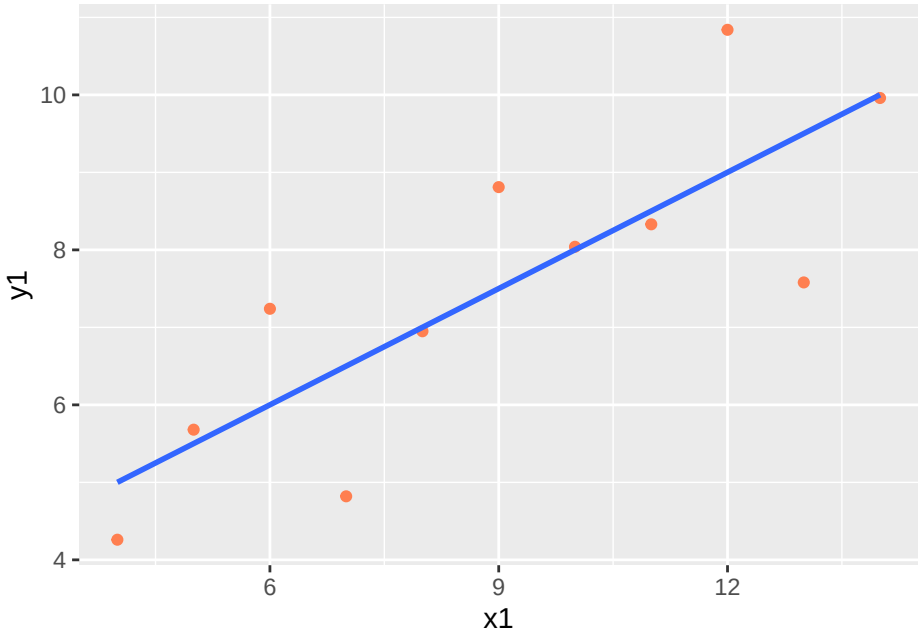
```
##  x1  x2  x3  x4  y1  y2  y3  y4
## 9.0 9.0 9.0 9.0 7.5 7.5 7.5 7.5
```

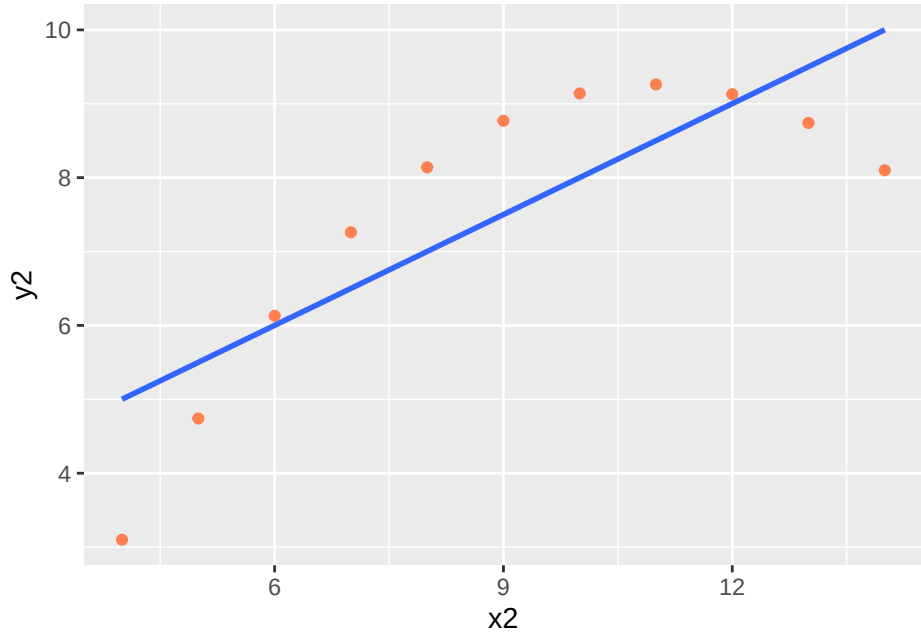
SD

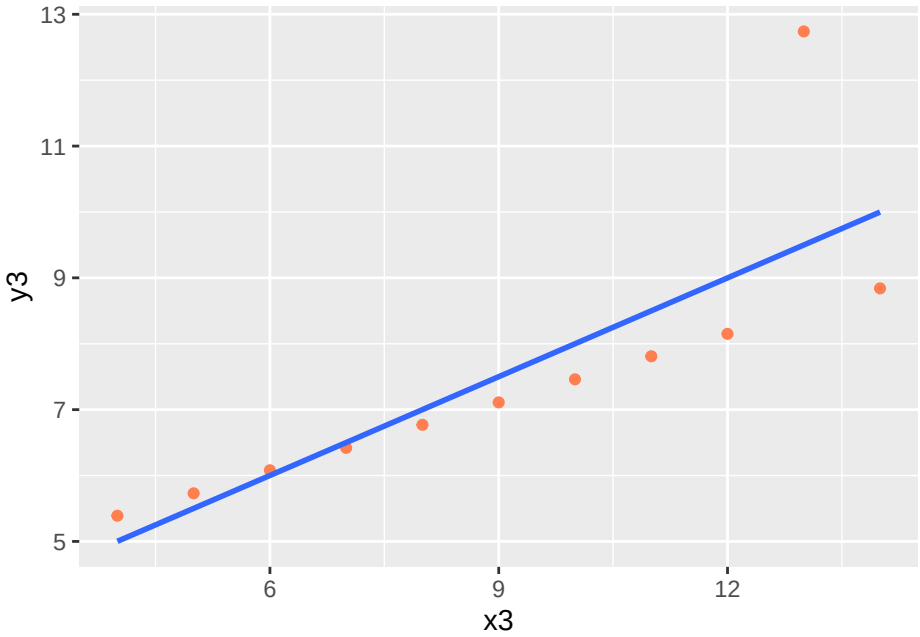
```
##   x1   x2   x3   x4   y1   y2   y3   y4
## 3.32 3.32 3.32 3.32 2.03 2.03 2.03 2.03
```

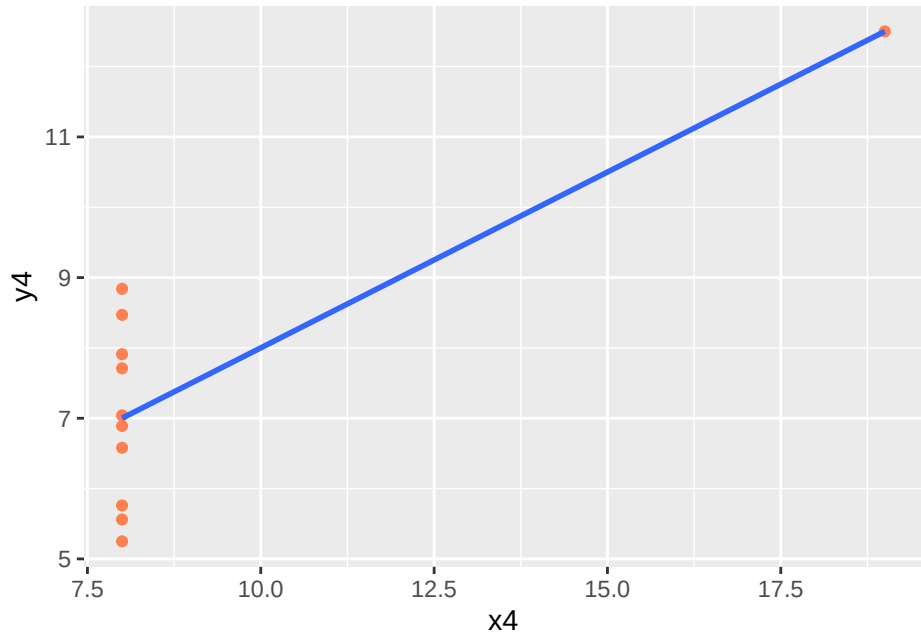
Correlation

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

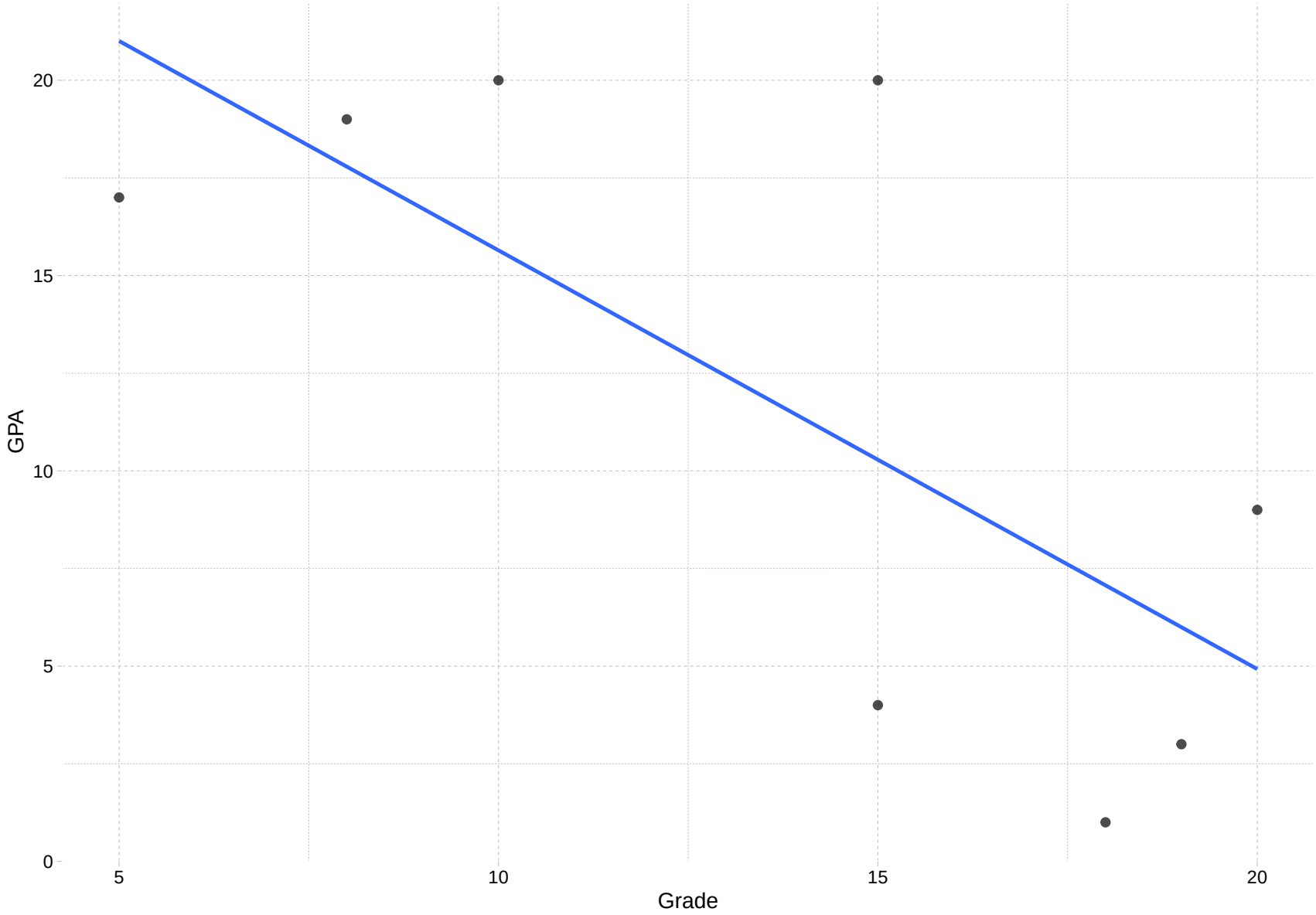

y1
x1
4
6
8
10
6
9
12

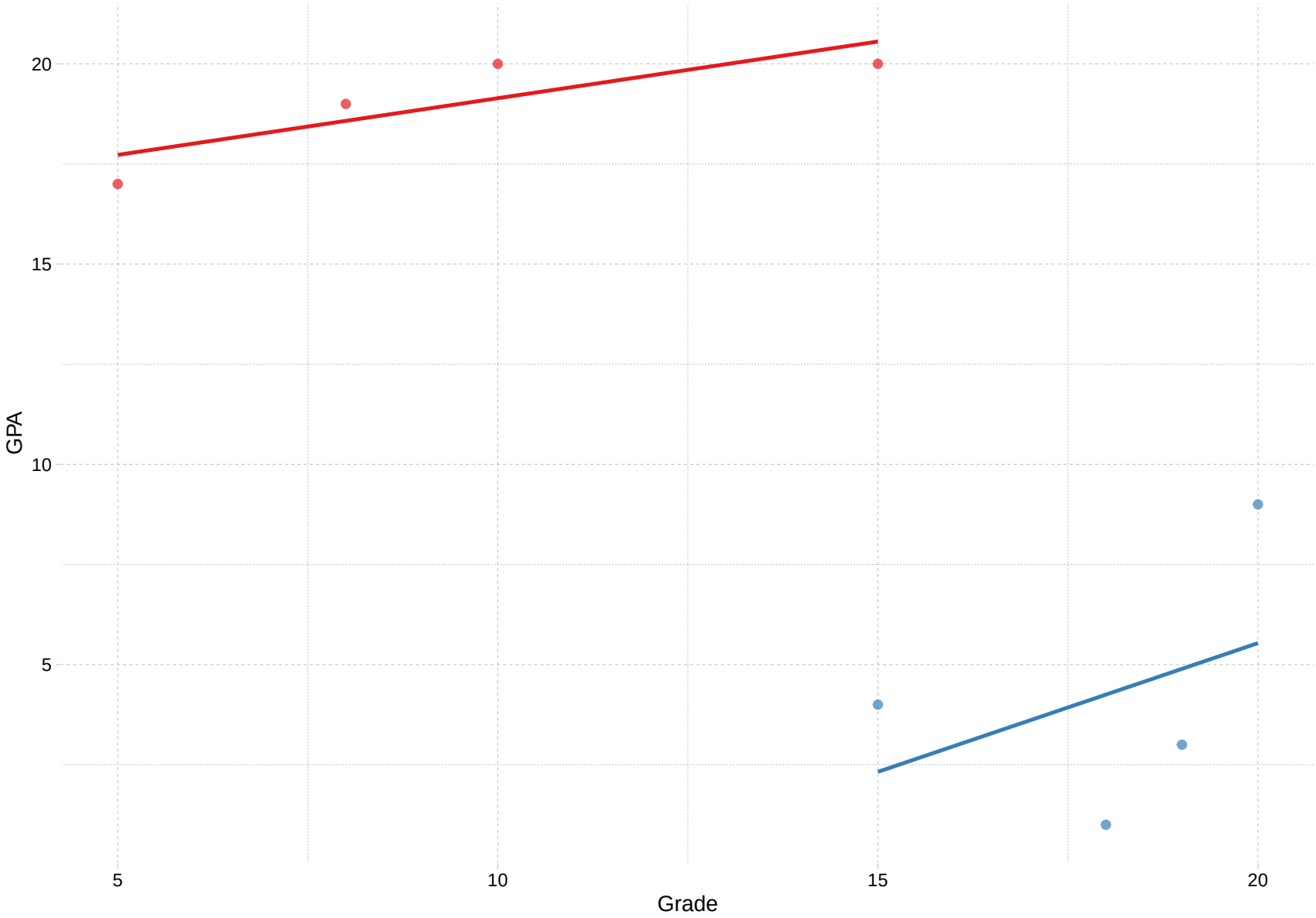

y2
x2
4
6
8
10
6
9
12


y3
x3
5
7
9
11
13
6
9
12

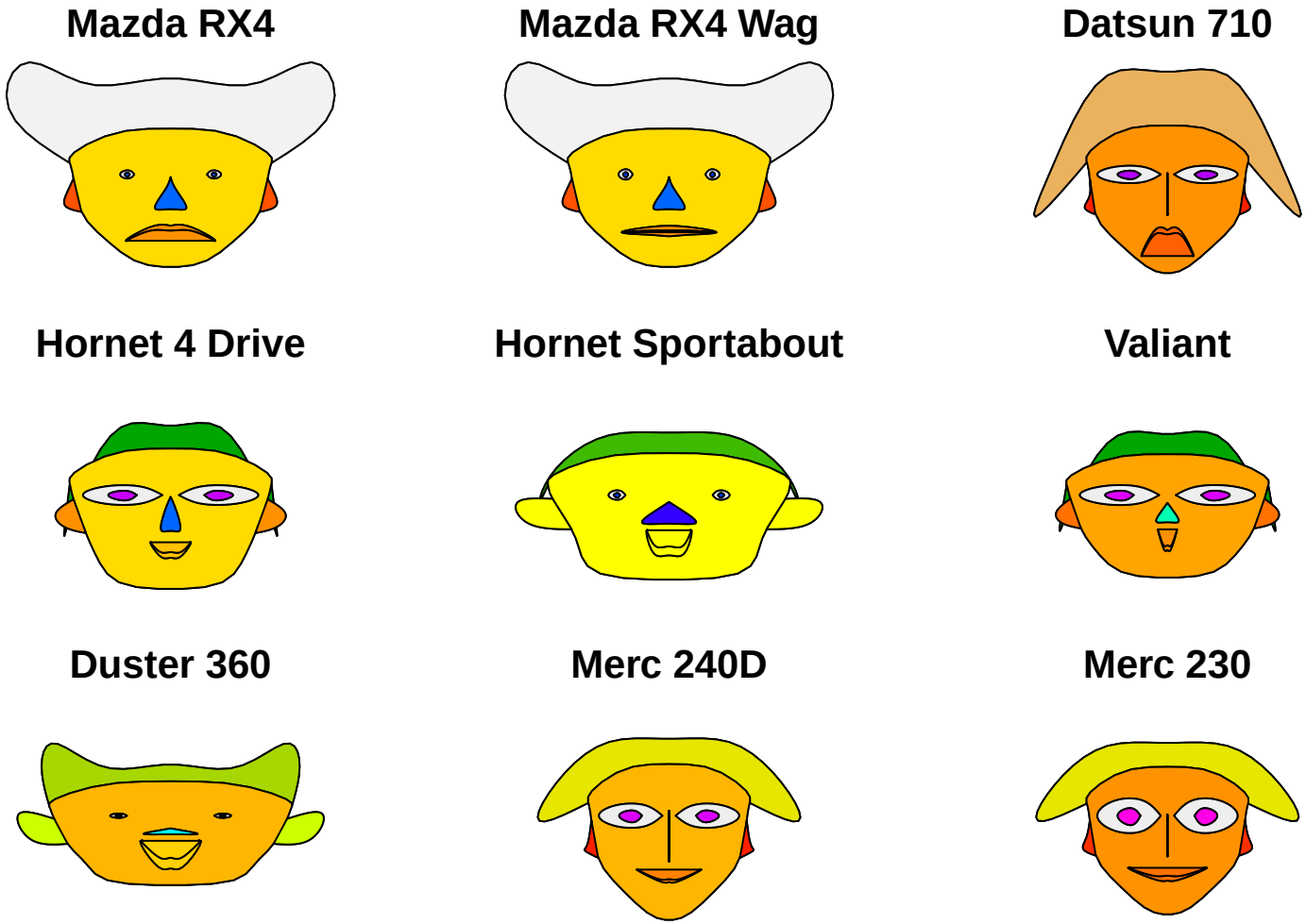

y4
x4
5
7
9
11
7.5
10.0
12.5
15.0
17.5

- Simpson's paradox




20
15
10
5
GPA
5
10
15
20
Grade

- Chernoff faces



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