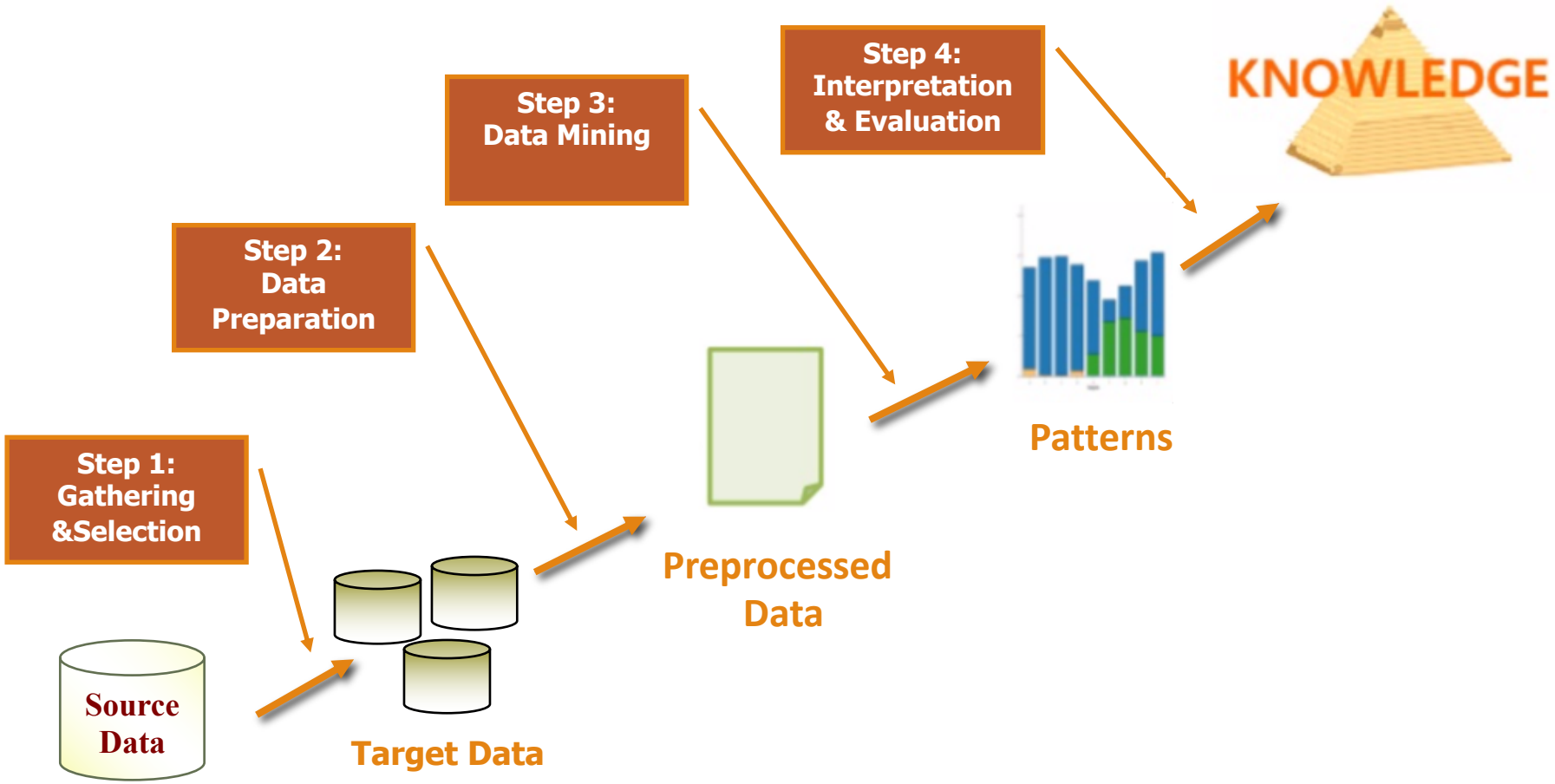
Lecture 2: Managing and Understanding Data

Data Mining Process: Recap



Data Mining Process: Recap

- **Data understanding:** With preliminary analysis, data exploration provides a high level overview of each attribute in the data set and interaction between the attributes.

#	Name	Type 1	Type 2	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
1	Bulbasaur	Grass	Poison	45	49	49	65	65	45	1	FALSE
2	Ivysaur	Grass	Poison	60	62	63	80	80	60	1	FALSE
3	Venusaur	Grass	Poison	80	82	83	100	100	80	1	FALSE
4	Mega Venus	Grass	Poison	80	100	123	122	120	80	1	FALSE
5	Charmander	Fire		39	52	43	60	50	65	1	FALSE
6	Charmeleon	Fire		58	64	58	80	65	80	1	FALSE
7	Charizard	Fire	Flying	78	84	78	109	85	100	1	FALSE
8	Mega Chariz	Fire	Dragon	78	130	111	130	85	100	1	FALSE
9	Mega Chariz	Fire	Flying	78	104	78	159	115	100	1	FALSE
10	Squirtle	Water		44	48	65	50	64	43	1	FALSE
11	Wartortle	Water		59	63	80	65	80	58	1	FALSE
12	Blastoise	Water		79	83	100	85	105	78	1	FALSE
13	Mega Blasto	Water		79	103	120	135	115	78	1	FALSE
14	Caterpie	Bug		45	30	35	20	20	45	1	FALSE
15	Metapod	Bug		50	20	55	25	25	30	1	FALSE
16	Butterfree	Bug	Flying	60	45	50	90	80	70	1	FALSE

If a feature represents a characteristic measured in numbers, it is unsurprisingly called **numeric.**

if a feature is an attribute that consists of a set of categories, the feature is called **categorical** or **nominal**.

Overview

- The better you understand your data, the better you will be able to match a machine learning model to your learning problem.

Overview

- Exploring the Structure of Data
- Exploring Numeric Variables
- Exploring Categorical Variables
- Exploring Relationships between Variables

Overview

- Exploring the Structure of Data
- Exploring Numeric Variables
- Exploring Categorical Variables
- Exploring Relationships between Variables

Exploring the Structure of Data

The str() function provides a method for displaying the structure of a data frame, or any R data structure including vectors and lists.

```
'data.frame':   800 obs. of  12 variables:
 $ X.         : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Name       : chr  "Bulbasaur" "Ivysaur" "Venusaur" "Mega Venusaur" ...
 $ Type.1     : chr  "Grass" "Grass" "Grass" "Grass" ...
 $ Type.2     : chr  "Poison" "Poison" "Poison" "Poison" ...
 $ HP         : int  45 60 80 80 39 58 78 78 78 44 ...
 $ Attack     : int  49 62 82 100 52 64 84 130 104 48 ...
 $ Defense    : int  49 63 83 123 43 58 78 111 78 65 ...
 $ Sp..Atk    : int  65 80 100 122 60 80 109 130 159 50 ...
 $ Sp..Def    : int  65 80 100 120 50 65 85 85 115 64 ...
 $ Speed      : int  45 60 80 80 65 80 100 100 100 43 ...
 $ Generation: int  1 1 1 1 1 1 1 1 1 1 ...
 $ Legendary : chr  "False" "False" "False" "False" ...
```

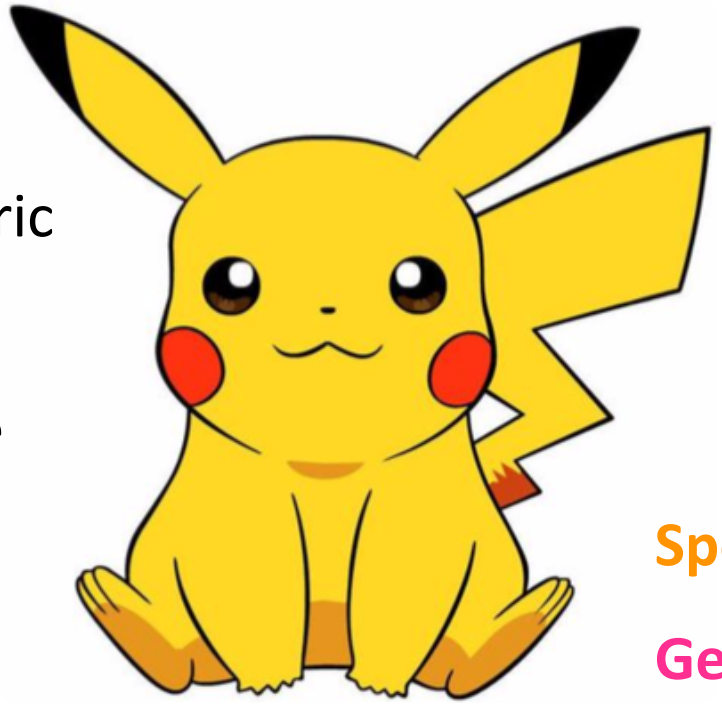
Number of instances

Number of variables

The type of each variable

Features of One Pokemon (e.g., Pikachu)

Name: Pikachu



Attack: 55

Defense: 40

Sp. Atk: 50

Sp. Def: 50

Speed: 90

Generation: 1

Type 1: Electric

Type 2: None

HP: 35

Legendary: False

Exploring the Structure of Data

```
> summary(pokemon)
       X.             Name              Type.1             Type.2                HP              Attack          Defense      
 Min.   :  1.0   Length:800         Length:800         Length:800         Min.   :  1.00   Min.   :  5   Min.   :  5.00  
 1st Qu.:200.8   Class :character   Class :character   Class :character   1st Qu.: 50.00   1st Qu.: 55   1st Qu.: 50.00  
 Median :400.5   Mode  :character   Mode  :character   Mode  :character   Median : 65.00   Median : 75   Median : 70.00  
 Mean   :400.5                                                           Mean   : 69.26   Mean   : 79   Mean   : 73.84  
 3rd Qu.:600.2                                                           3rd Qu.: 80.00   3rd Qu.:100   3rd Qu.: 90.00  
 Max.   :800.0                                                           Max.   :255.00   Max.   :190   Max.   :230.00  
    Sp..Atk          Sp..Def          Speed          Generation     Legendary        
 Min.   : 10.00   Min.   : 20.0   Min.   :  5.00   Min.   :1.000   Length:800        
 1st Qu.: 49.75   1st Qu.: 50.0   1st Qu.: 45.00   1st Qu.:2.000   Class :character  
 Median : 65.00   Median : 70.0   Median : 65.00   Median :3.000   Mode  :character  
 Mean   : 72.82   Mean   : 71.9   Mean   : 68.28   Mean   :3.324                     
 3rd Qu.: 95.00   3rd Qu.: 90.0   3rd Qu.: 90.00   3rd Qu.:5.000                     
 Max.   :194.00   Max.   :230.0   Max.   :180.00   Max.   :6.000                     
```

Exploring the Structure of Data

- Remove variables
- Change categorical variables to factors

```
> summary(pokemon)
     Type.1         Type.2          HP              Attack          Defense          Sp..Atk          Sp..Def          Speed       
 Water  :112            :386   Min.   :  1.00   Min.   :  5   Min.   :  5.00   Min.   : 10.00   Min.   : 20.0   Min.   :  5.00  
 Normal : 98   Flying   : 97   1st Qu.: 50.00   1st Qu.: 55   1st Qu.: 50.00   1st Qu.: 49.75   1st Qu.: 50.0   1st Qu.: 45.00  
 Grass  : 70   Ground   : 35   Median : 65.00   Median : 75   Median : 70.00   Median : 65.00   Median : 70.0   Median : 65.00  
 Bug    : 69   Poison   : 34   Mean   : 69.26   Mean   : 79   Mean   : 73.84   Mean   : 72.82   Mean   : 71.9   Mean   : 68.28  
 Psychic: 57   Psychic  : 33   3rd Qu.: 80.00   3rd Qu.:100   3rd Qu.: 90.00   3rd Qu.: 95.00   3rd Qu.: 90.0   3rd Qu.: 90.00  
 Fire   : 52   Fighting: 26   Max.   :255.00   Max.   :190   Max.   :230.00   Max.   :194.00   Max.   :230.0   Max.   :180.00  
 (Other):342   (Other) :189                                                                                                      
 Generation Legendary
 1:166      False:735
 2:106      True : 65
 3:160
 4:121
 5:165
 6: 82
```

Exploring the Structure of Data

- Missing values
  - Missing mechanism: understanding the reasons why data are missing
  - Deal with missingness

```
> summary(pokemon)
     Type.1         Type.2          HP              Attack          Defense          Sp..Atk          Sp..Def          Speed       
 Water  :112   none    :386   Min.   :  1.00   Min.   :  5   Min.   :  5.00   Min.   : 10.00   Min.   : 20.0   Min.   :  5.00  
 Normal : 98   Flying  : 97   1st Qu.: 50.00   1st Qu.: 55   1st Qu.: 50.00   1st Qu.: 49.75   1st Qu.: 50.0   1st Qu.: 45.00  
 Grass  : 70   Ground  : 35   Median : 65.00   Median : 75   Median : 70.00   Median : 65.00   Median : 70.0   Median : 65.00  
 Bug    : 69   Poison  : 34   Mean   : 69.26   Mean   : 79   Mean   : 73.84   Mean   : 72.82   Mean   : 71.9   Mean   : 68.28  
 Psychic: 57   Psychic : 33   3rd Qu.: 80.00   3rd Qu.:100   3rd Qu.: 90.00   3rd Qu.: 95.00   3rd Qu.: 90.0   3rd Qu.: 90.00  
 Fire   : 52   Fighting: 26   Max.   :255.00   Max.   :190   Max.   :230.00   Max.   :194.00   Max.   :230.0   Max.   :180.00  
 (Other):342   (Other) :189                                                                                                      
 Generation Legendary
 1:166      False:735
 2:106      True : 65
 3:160
 4:121
 5:165
 6: 82
```

Overview

- Exploring the Structure of Data
- Exploring Numeric Variables
- Exploring Categorical Variables
- Exploring Relationships between Variables

Exploring Numeric Variables

- We employ a common set of measurements to describe values known as **summary statistics**.

```
> summary(pokemon$Attack)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      5      55      75      79     100     190
```

```
> summary(pokemon$Sp..Atk + pokemon$Sp..Def)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   35.0   105.0   140.0   144.7   180.0   340.0
```

Exploring Numeric Variables

- Measures of **central tendency** are a class of statistics used to identify a value that falls in the middle of a set of data.
  - Mean: the sum of all values divided by the number of values
  - Median: the value that occurs halfway through an ordered list of values
    - $median(x) = \begin{cases} x_{r+1}, if\ m\ is\ odd. m = 2r + 1 \\ \frac{1}{2}(x_r + x_{r+1}), if\ m\ is\ even. m = 2r \end{cases}$

Exploring Numeric Variables

- Mean and median are affected differently by the values falling at the far ends of the range.
  - Which one is more sensitive to **outliers**/extreme values?

```
> mean(pokemon$Attack)
[1] 79.00125
> median(pokemon$Attack)
[1] 75
> range(pokemon$Attack)
[1]   5 190
```

Exploring Numeric Variables

- Measuring **spread –** quartiles and the five-number summary
  - Mean and median tell us little about whether or not there is diversity in the measurements.
  - The **five-number summary** is a set of five statistics that roughly depict the spread of a feature's values.

1. Minimum (`Min.`)
2. First quartile, or Q1 (`1st Qu.`)
3. Median, or Q2 (`Median`)
4. Third quartile, or Q3 (`3rd Qu.`)
5. Maximum (`Max.`)

The **quartiles** divide a dataset into four portions, each with the same number of values.

Exploring Numeric Variables

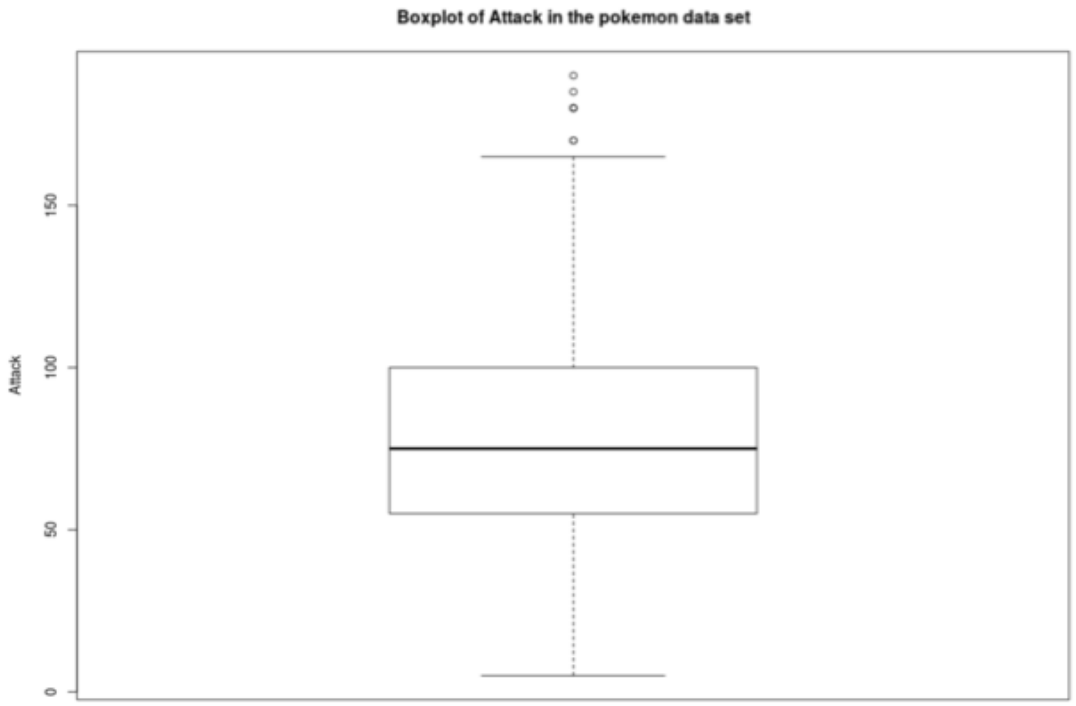
```
> quantile(pokemon$Attack)
  0%  25%  50%  75% 100%
   5   55   75  100  190

> summary(pokemon$Attack)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      5      55      75      79     100     190
```

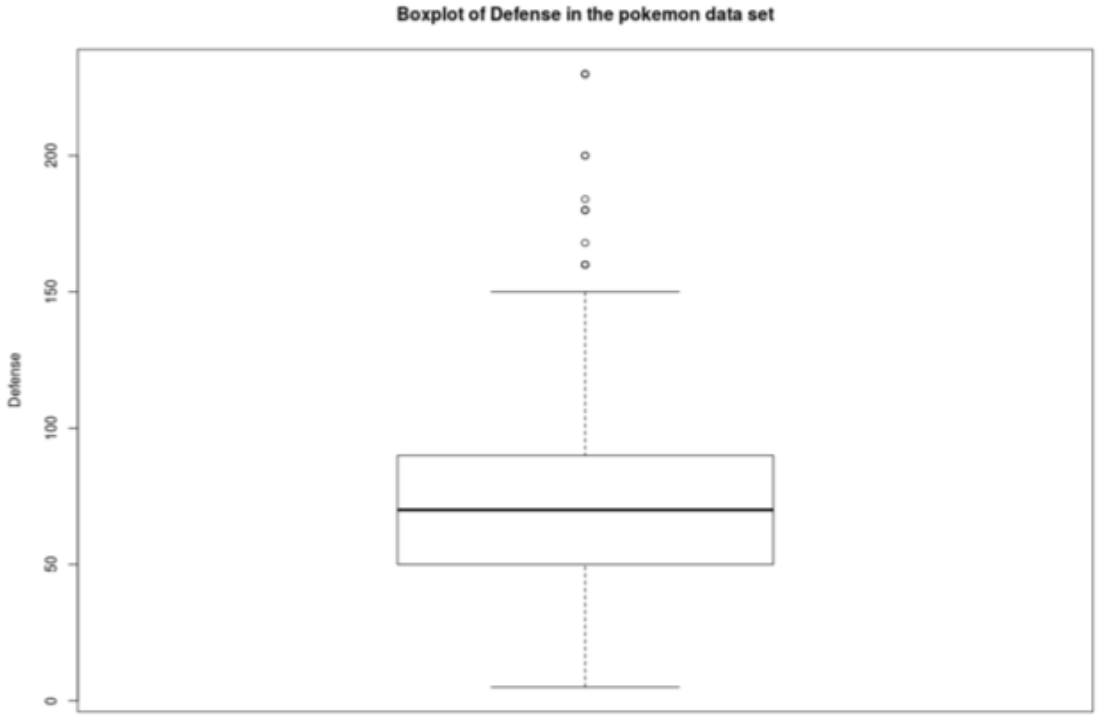
- The difference between Q1 and Q3 is known as the **Interquartile Range** (**IQR**)

```
> IQR(pokemon$Attack)
[1] 45
```

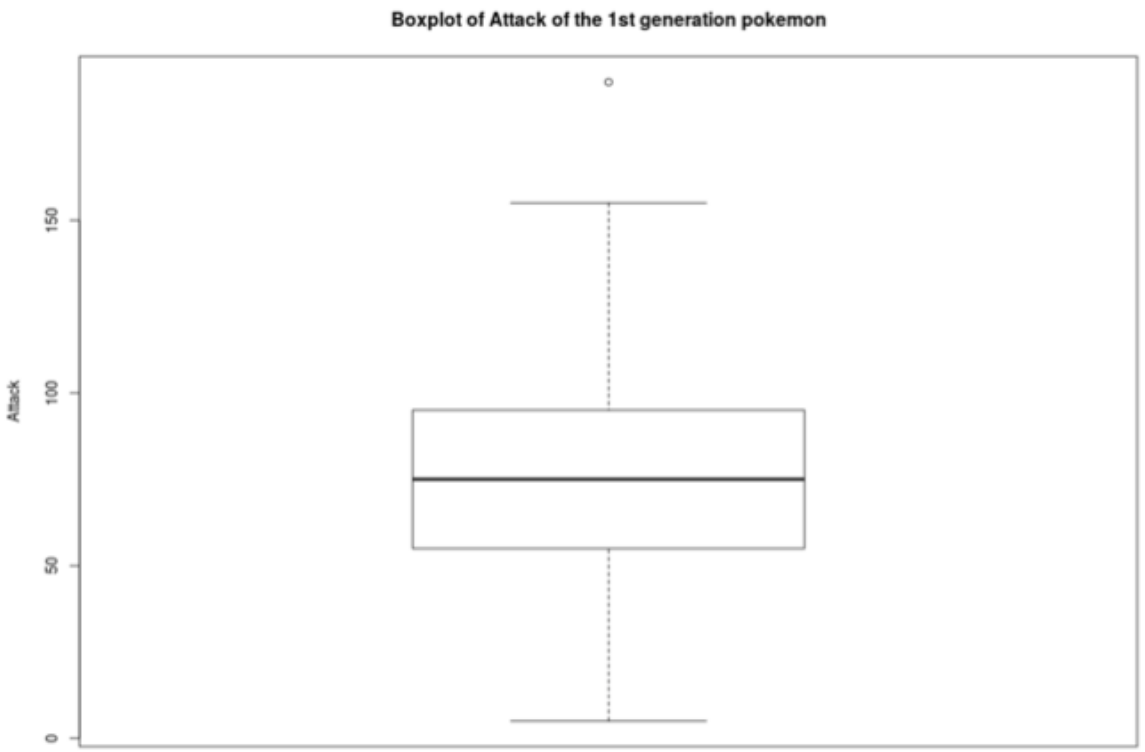
Exploring Numeric Variables

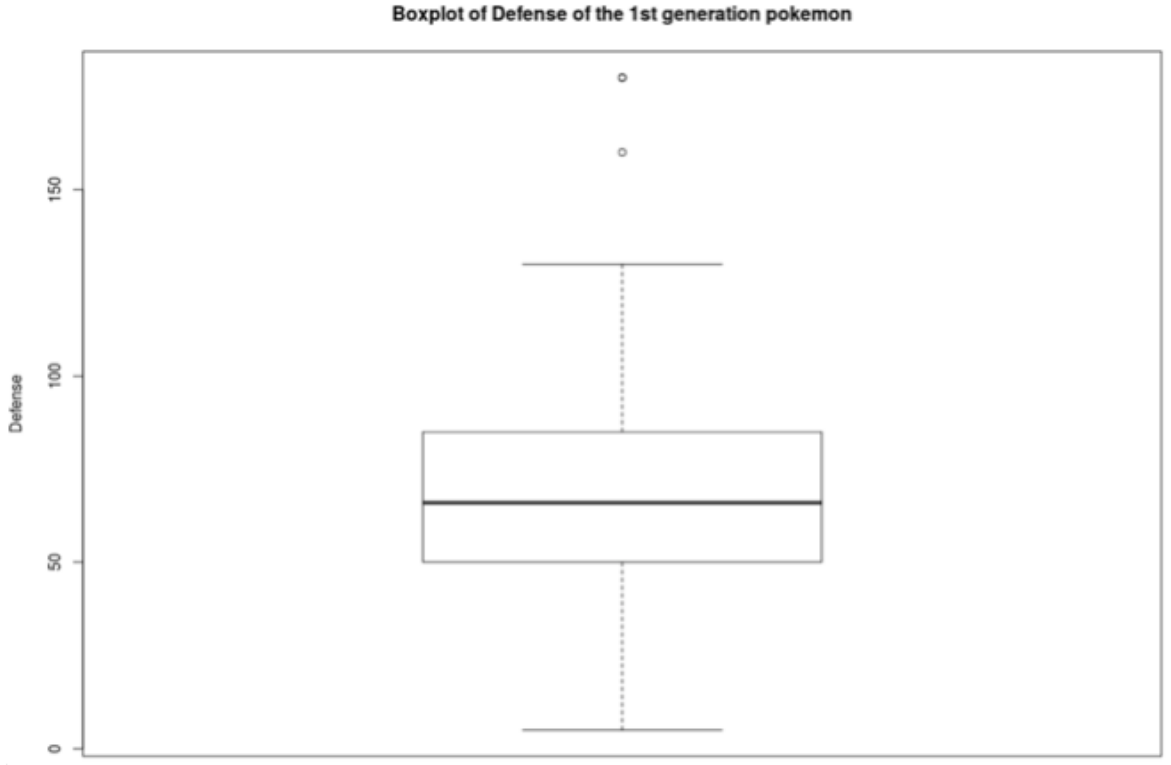
- Visualizing numeric variables – boxplot
  - A common visualization of the five-number summary is **boxplot**.





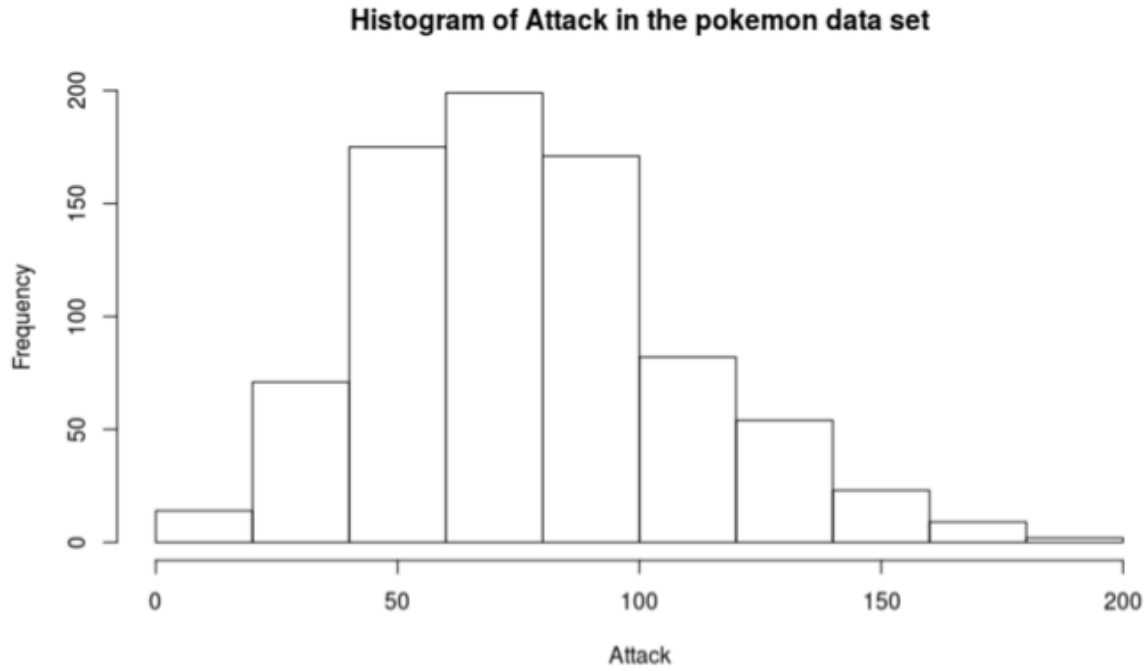
Exploring Numeric Variables

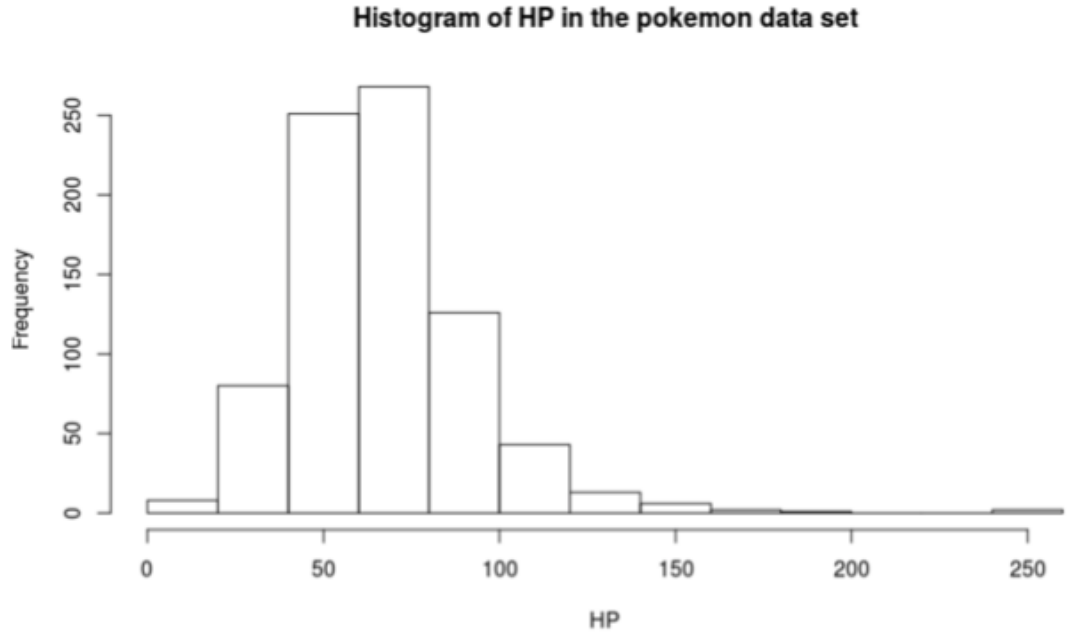




Exploring Numeric Variables

- Visualizing numeric variables – histograms
  - It divides the variable's values into a predefined number of portions or **bins** that act as containers for values.





Exploring Numeric Variables

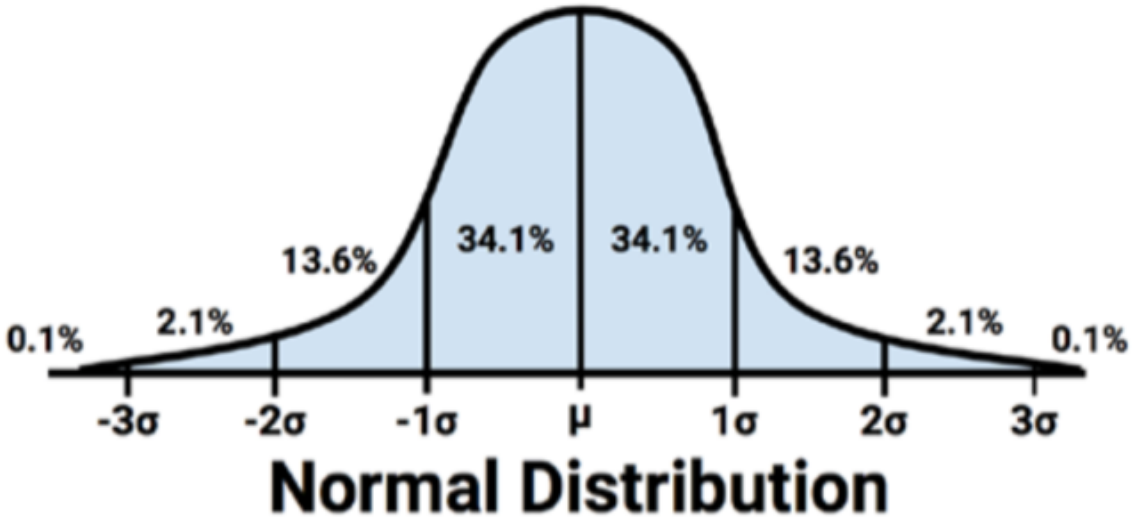
- Visualizing numeric variables – histograms
  - The histogram is composed of a series of bars with heights indicating the count, or **frequency** of values falling within each of the equal width bins partitioning the values.
  - The vertical lines that separate the bars, as labeled on the horizontal axis, indicate the start and end points of the range of values for the bin.

Exploring Numeric Variables

- This characteristic is known as **skew**, or more specifically right skew, because the values on the high end (right side) are far more spread out than the values on the low end (left side).

Exploring Numeric Variables

- The **68-95-99.7 rule** states that 68 percent of the values in a normal distribution fall within one standard deviation of the mean, while 95 percent and 99.7 percent of the values fall within two and three standard deviations, respectively.



Overview

- Exploring the Structure of Data
- Exploring Numeric Variables
- Exploring Categorical Variables
- Exploring Relationships between Variables

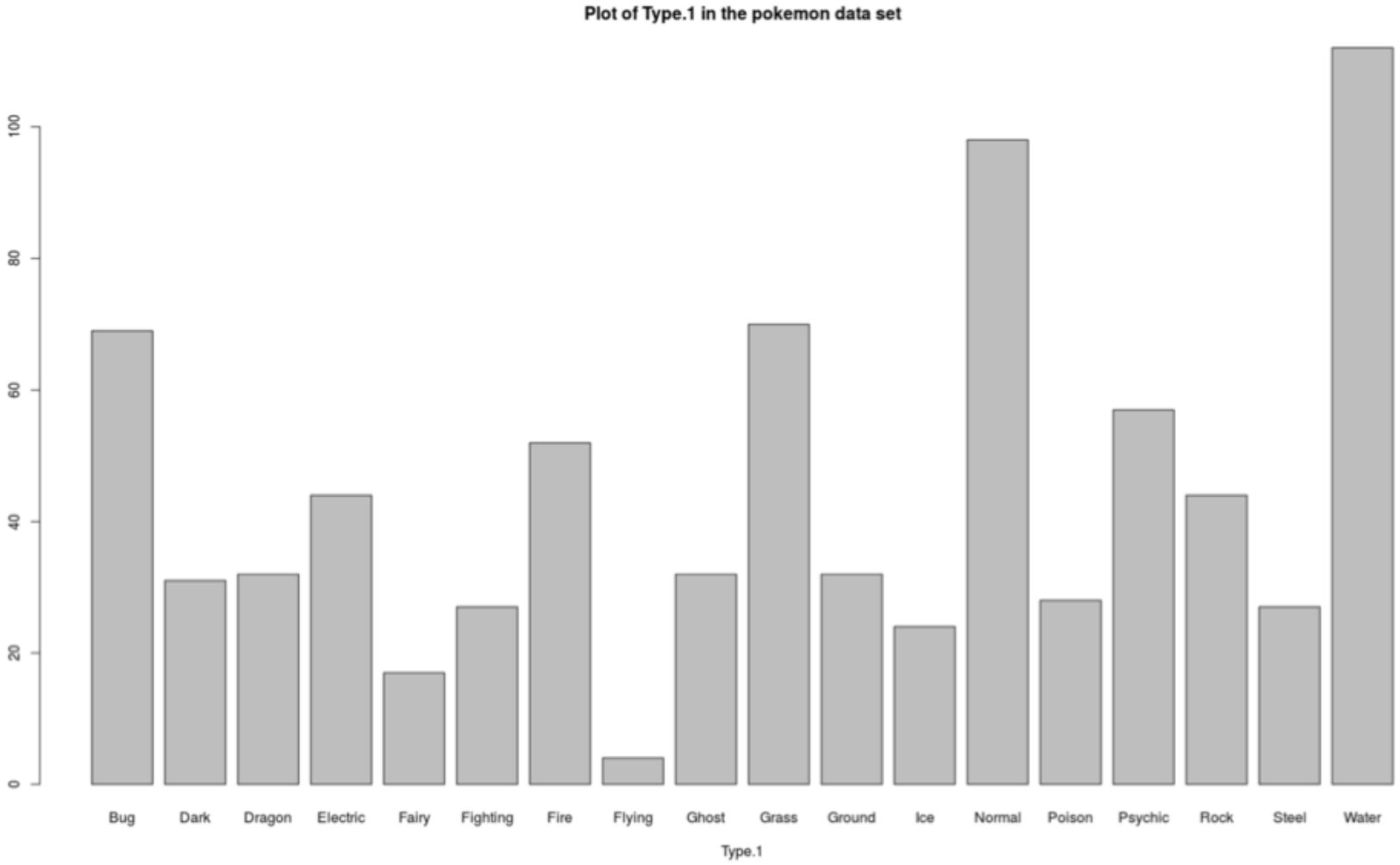
Exploring Categorical Variables

- **Frequency** of particular values

```
> summary(pokemon$Type.1)
    Bug    Dark  Dragon Electric   Fairy Fighting    Fire  Flying   Ghost   Grass  Ground     Ice  Normal  Poison Psychic
     69      31      32      44      17      27      52       4      32      70      32      24      98      28      57
   Rock   Steel   Water
     44      27     112
```

- A table that presents a single categorical variable is known as a **one-way table**.
  - table() function
  - prop.table() function

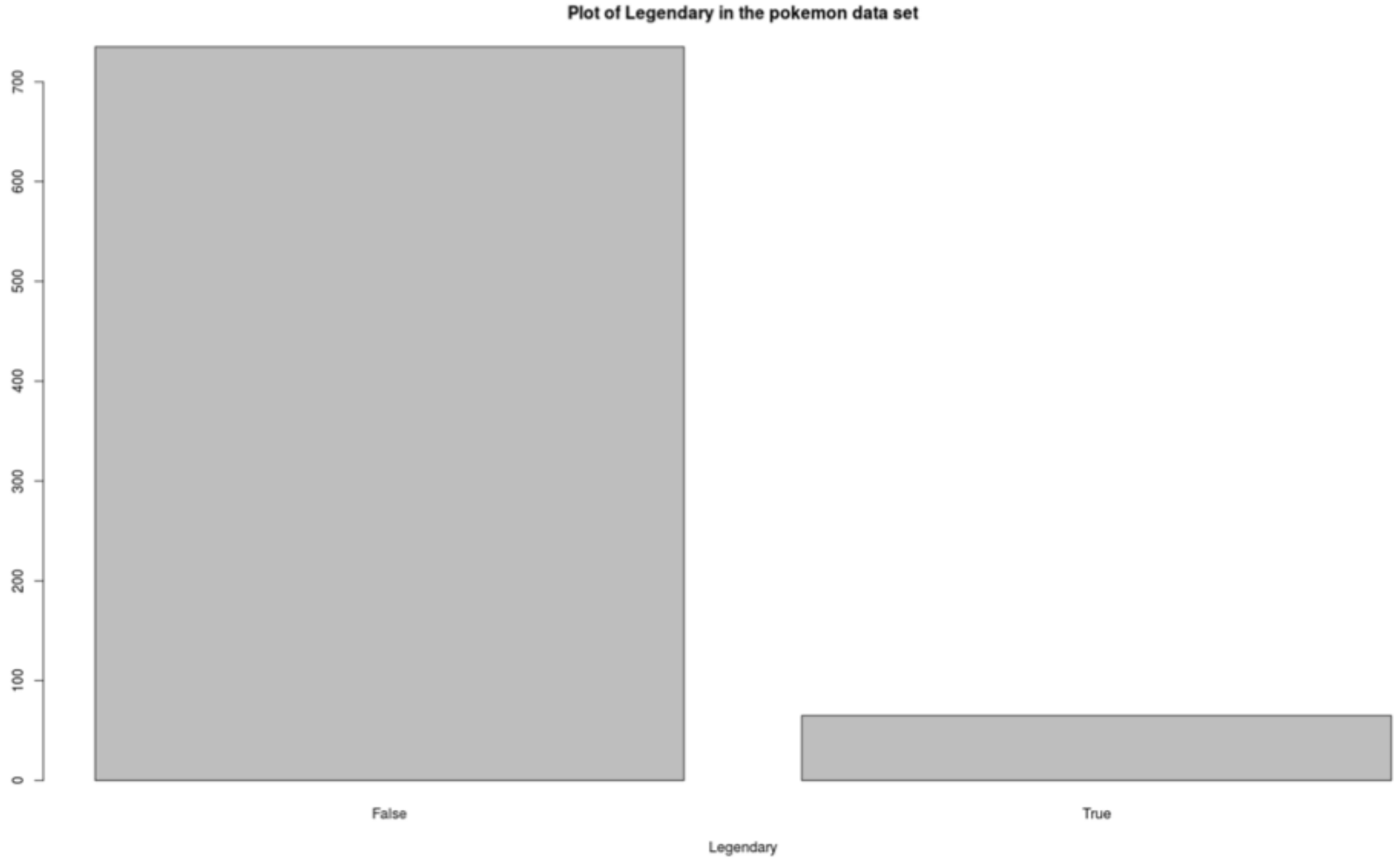
Exploring Categorical Variables



Exploring Categorical Variables

- Measuring the central tendency – the mode
  - The **mode** of a feature is the value occurring most often.
  - It would be dangerous to place too much emphasis on the mode, since the most common value is not necessarily a majority.

Exploring Categorical Variables



Overview

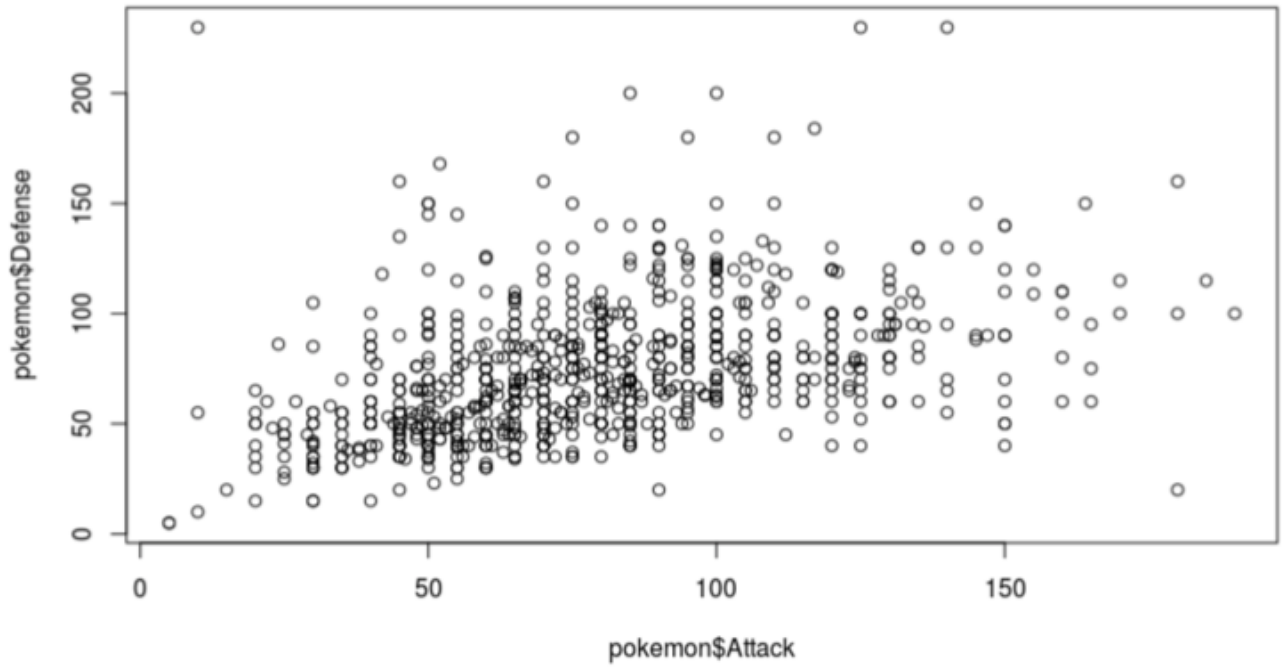
- Exploring the Structure of Data
- Exploring Numeric Variables
- Exploring Categorical Variables
- Exploring Relationships between Variables

Exploring relationships between variables

- So far, we have examined variables one at a time, calculating only **univariate** statistics.
  - **Bivariate** relationships consider the relationship between two variables. Relationships of more than two variables are called **multivariate** relationships.

Exploring relationships between variables

- Scatterplots: two numeric variables
  - A **scatterplot** is a diagram that visualizes a bivariate relationship.
    - 2-dimensional feature space
    - plot() function



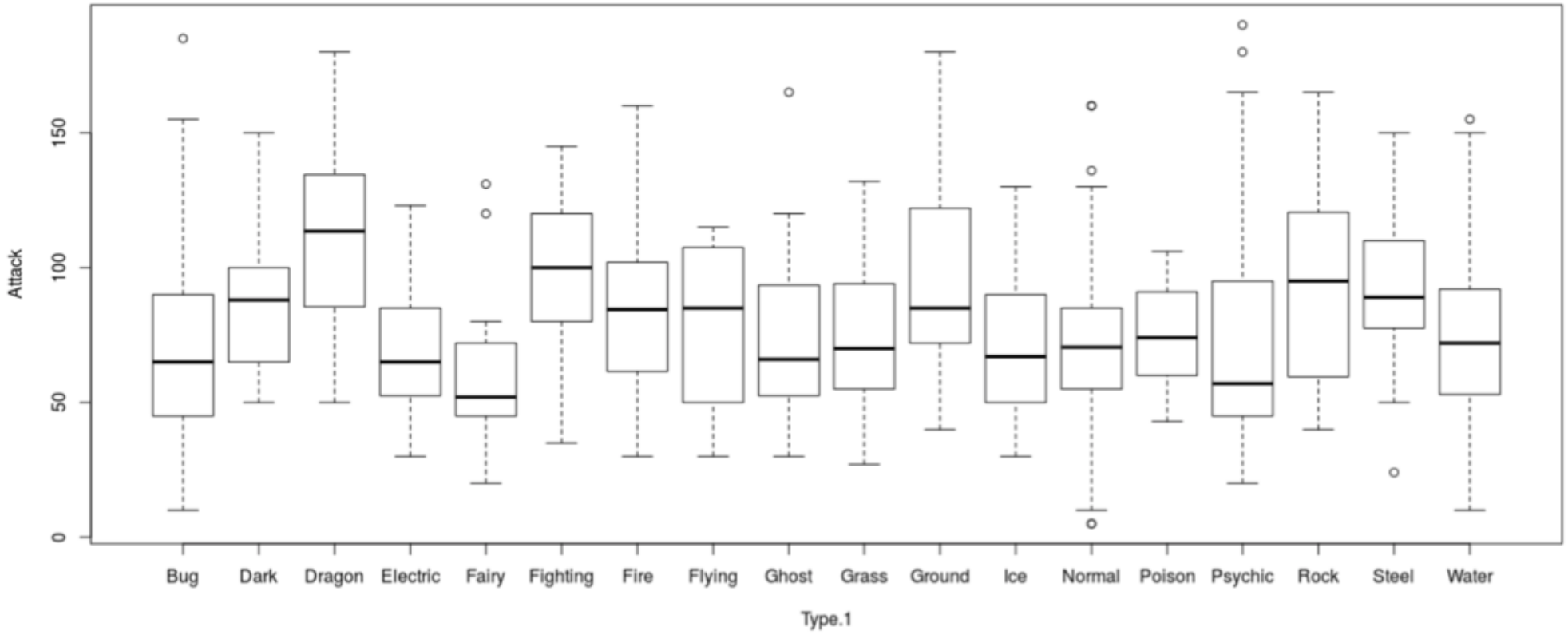
Exploring relationships between variables

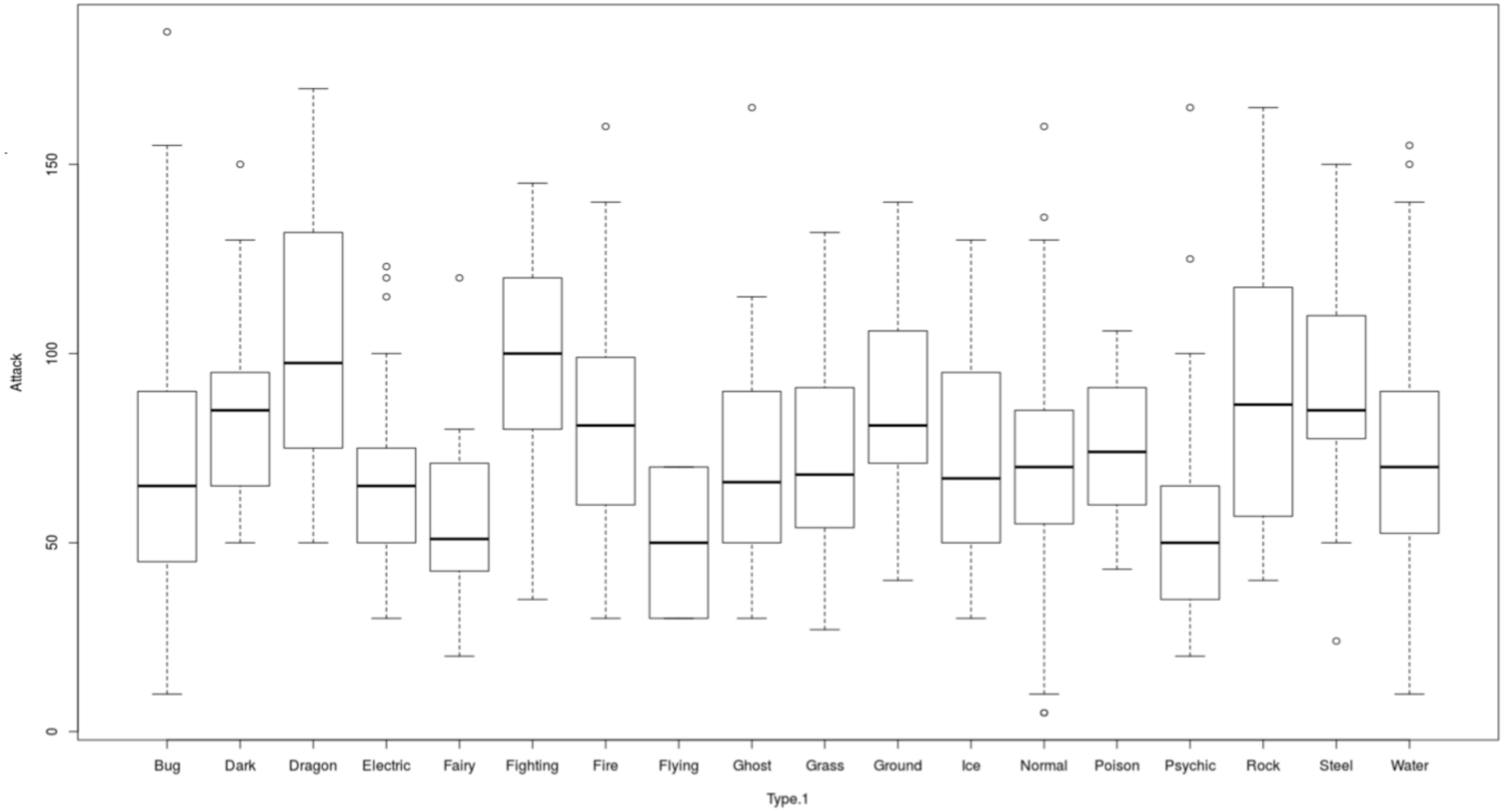
- Scatterplots: two numeric variables
  - Positive/Negative association.
  - The strength of a linear association between two variables is measured by a statistic known as **correlation** (range [-1, 1]).

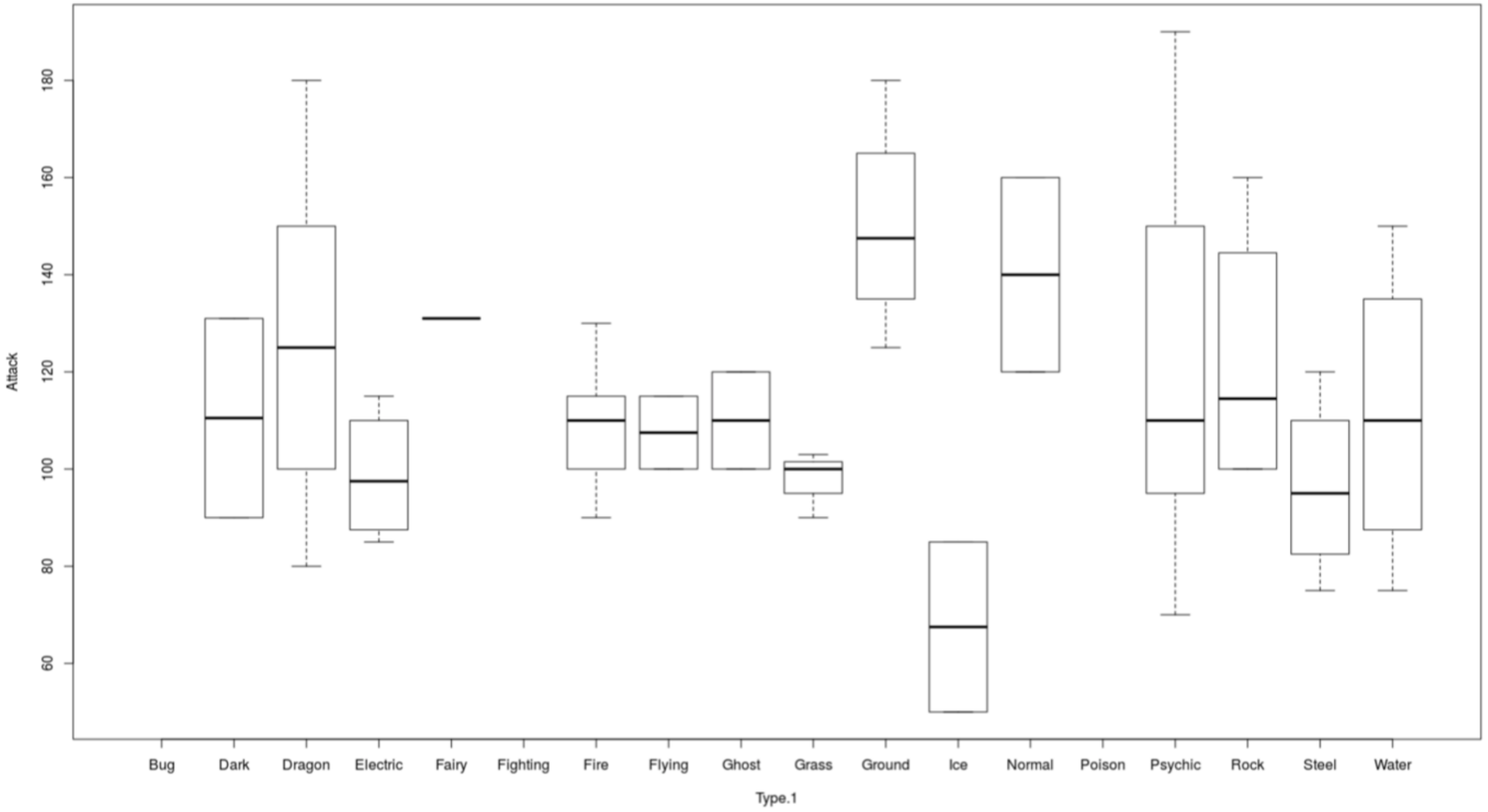
```
> cor(pokemon[,c("Attack", "Defense")])
           Attack   Defense
Attack  1.0000000 0.4386871
Defense 0.4386871 1.0000000
> cor(pokemon[,3:8])
                HP    Attack   Defense   Sp..Atk   Sp..Def     Speed
HP      1.0000000 0.4223860 0.2396223 0.3623799 0.3787181 0.1759521
Attack  0.4223860 1.0000000 0.4386871 0.3963618 0.2639896 0.3812397
Defense 0.2396223 0.4386871 1.0000000 0.2235486 0.5107466 0.0152266
Sp..Atk 0.3623799 0.3963618 0.2235486 1.0000000 0.5061214 0.4730179
Sp..Def 0.3787181 0.2639896 0.5107466 0.5061214 1.0000000 0.2591331
Speed   0.1759521 0.3812397 0.0152266 0.4730179 0.2591331 1.0000000
```

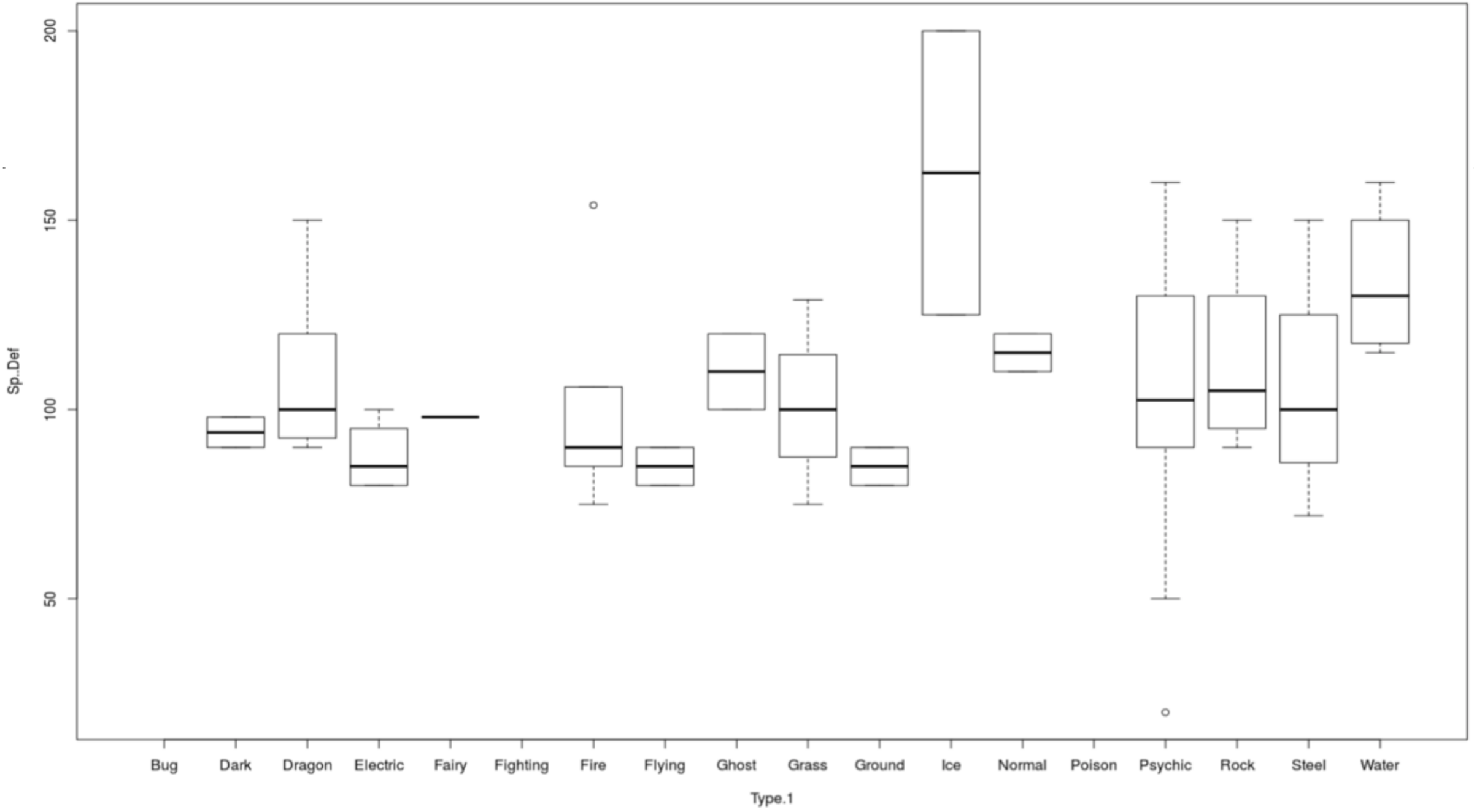
Exploring relationships between variables

- Relationships between numeric variable and categorical variable




Attack
0
50
100
150
Bug
Dark
Dragon
Electric
Fairy
Fighting
Fire
Flying
Ghost
Grass
Ground
Ice
Normal
Poison
Psychic
Rock
Steel
Water
Type.1

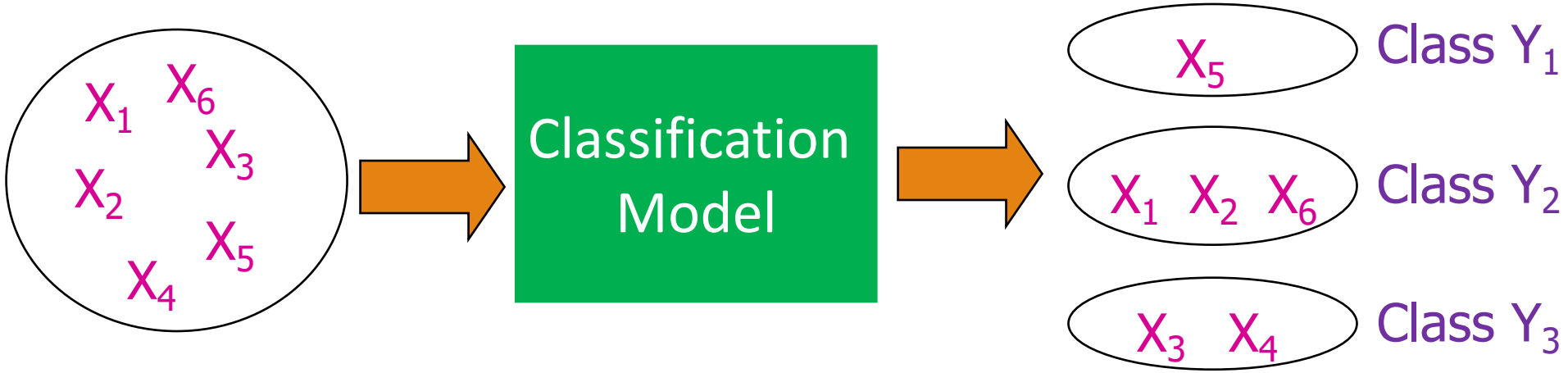

Attack
60
80
100
120
140
160
180
Bug
Dark
Dragon
Electric
Fairy
Fighting
Fire
Flying
Ghost
Grass
Ground
Ice
Normal
Poison
Psychic
Rock
Steel
Water
Type.1


Sp..Def
200
150
100
50
Bug
Dark
Dragon
Electric
Fairy
Fighting
Fire
Flying
Ghost
Grass
Ground
Ice
Normal
Poison
Psychic
Rock
Steel
Water
Type.1

Introduction to Classification

What is Classification?

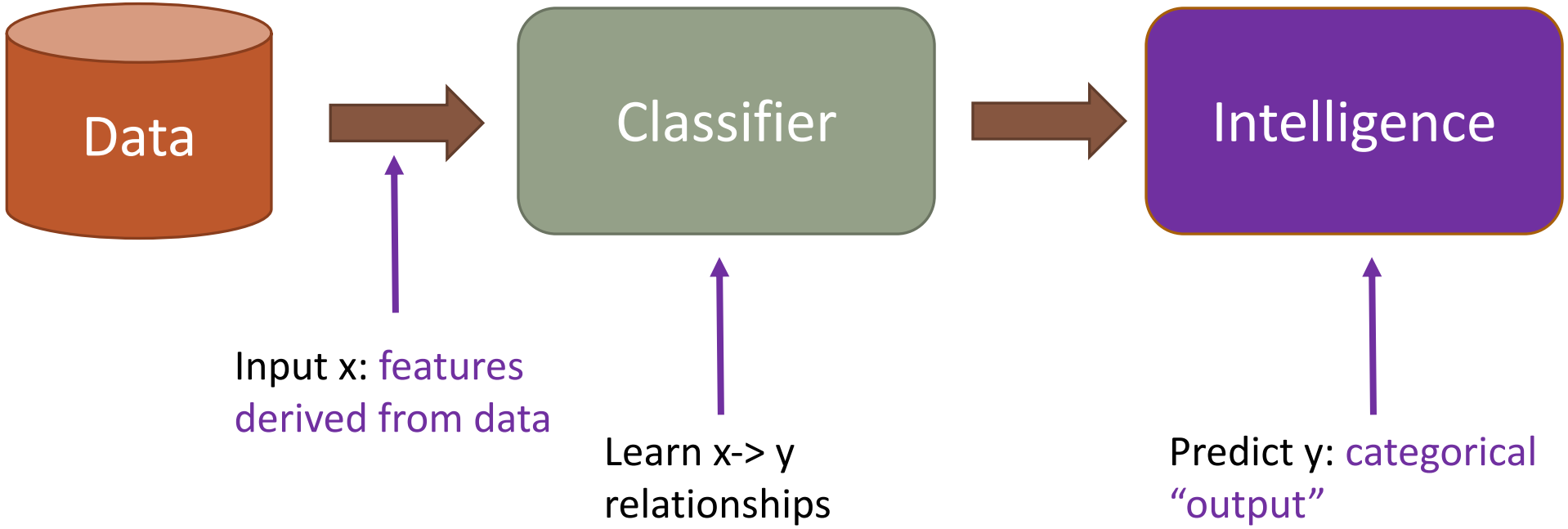
- Classify objects into a set of **pre-specified** classes (or categories) based on the values of relevant object attributes (features).



Classes Y_1, Y_2, and Y_3 are pre-determined

What is Classification?

From features to predictions



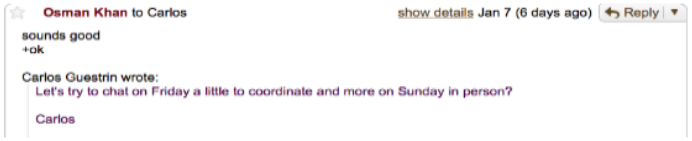
Classification Motivation

- Why classification?
  - Estimating/predicting the class or category of action to support **time and cost-effective decision making**.
    - Identifying the class with a single or a small number of data attributes (e.g., gender, age) is manageable by human decision makers.
    - But not when the number of attributes or the number of instances is large, or if the "mapping" is complex.

Classification Motivation

- Decisions on actions depend on knowing the category or class to which the receiver of actions belongs. E.g.,
  - Whether any and what treatment a patient needs for a disease depends on the prediction of possible patient outcomes of different treatment plans.
  - To dis-approve a credit card transaction depends the prediction that the transaction is likely to be fraudulent.
  - To block an IP should from accessing a server depends on the prediction that this IP is likely to be responsible for an intrusive attack.

Binary Classification

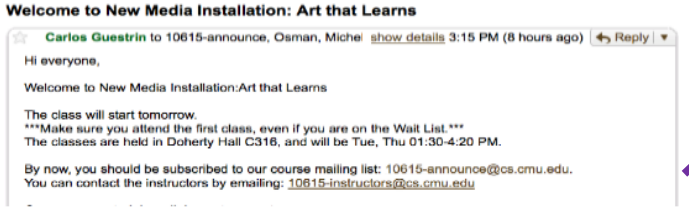


Osman Khan to Carlos show details Jan 7 (6 days ago) Reply

sounds good
+ok

Carlos Guestrin wrote:
Let's try to chat on Friday a little to coordinate and more on Sunday in person?

Carlos



Welcome to New Media Installation: Art that Learns

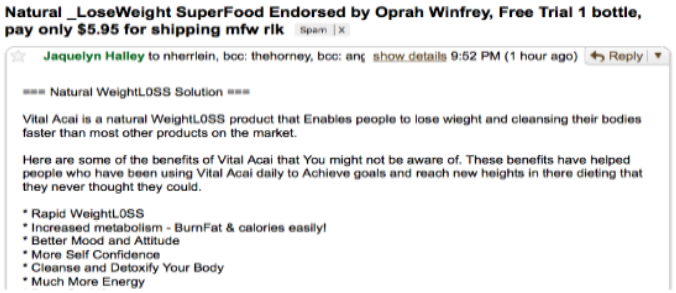
Carlos Guestrin to 10615-announce, Osman, Michel show details 3:15 PM (8 hours ago) Reply

Hi everyone,

Welcome to New Media Installation:Art that Learns

The class will start tomorrow.
Make sure you attend the first class, even if you are on the Wait List.
The classes are held in Doherty Hall C316, and will be Tue, Thu 01:30-4:20 PM.

By now, you should be subscribed to our course mailing list: 10615-announce@cs.cmu.edu.
You can contact the instructors by emailing: 10615-instructors@cs.cmu.edu



Natural _LoseWeight SuperFood Endorsed by Oprah Winfrey, Free Trial 1 bottle, pay only $5.95 for shipping mfw rlk Spam x

Jaquelyn Halley to nherrlein, bcc: thehorney, bcc: anq show details 9:52 PM (1 hour ago) Reply

=== Natural WeightL0SS Solution ===

Vital Acai is a natural WeightL0SS product that Enables people to lose wieght and cleansing their bodies faster than most other products on the market.

Here are some of the benefits of Vital Acai that You might not be aware of. These benefits have helped people who have been using Vital Acai daily to Achieve goals and reach new heights in there dieting that they never thought they could.

* Rapid WeightL0SS
* Increased metabolism - BurnFat & calories easily!
* Better Mood and Attitude
* More Self Confidence
* Cleanse and Detoxify Your Body
* Much More Energy

Not Spam

Spam

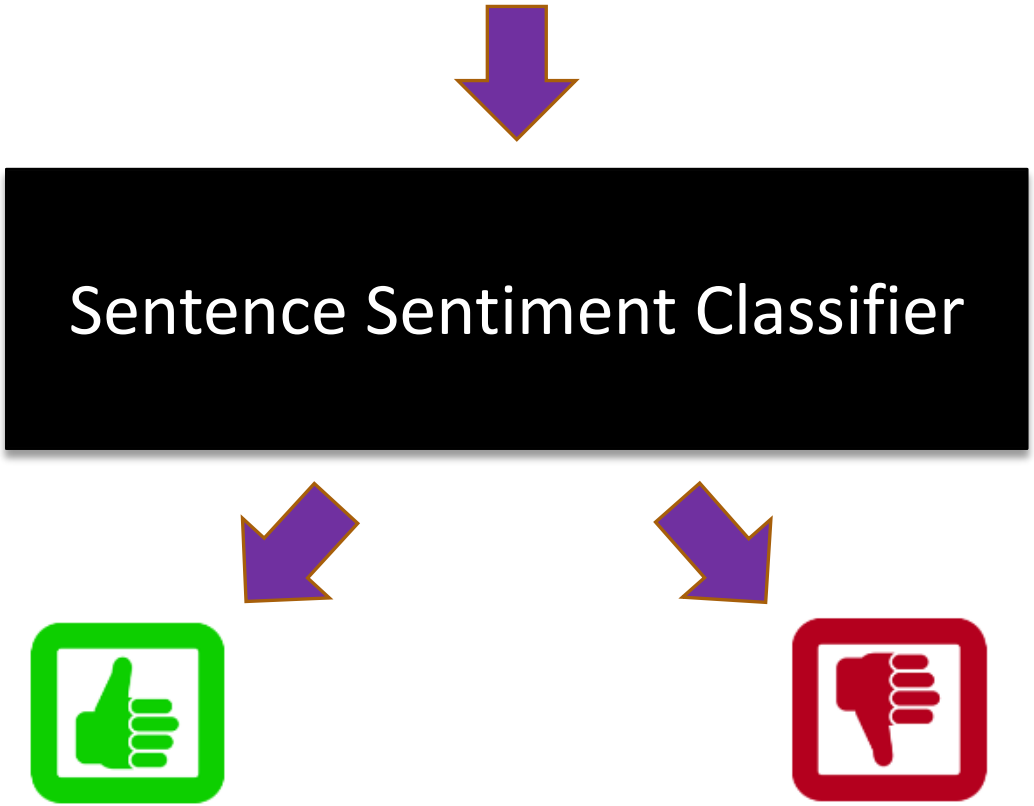
Input: x

Text of email, sender, IP, …

Output: y

Binary Classification

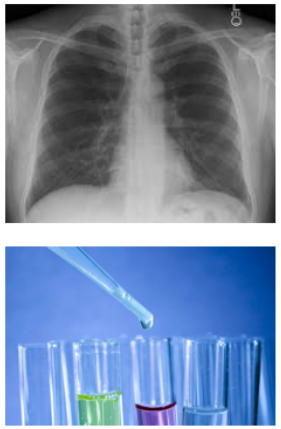
Input **x**: Easily best sushi in Salt Lake City.



Output: y
Sentiment

Multiclass Classification

Input: x



Disease Classification Model

Output: y

- Healthy
- Cold
- Flu
- Pneumonia
- …

Decision Tree