2. Data essentials

Transfer exploration seminar: Statistics and Data Science

Dr. Uma Ravat

PSTAT 194TR

R essentials: summary

- Console and Environment Panes, Command Prompt
- Objects, Assignment Operator : <-
  - Variables: nouns
  - Functions: verbs
  - Naming conventions
- Packages: ready made functions and datasets from others
  - Install once
  - Load every time you need it
- Help : ?
- Comments: #
  - **use them!** for yourself, the grader
- Coding style : **have one** and be consistent
  - See chapters 1-3 of the tidyverse style guide
- Environment

Next we will see...

Types of Statistical Data

- Numerical
- Categorical

EDA - Simple Techniques

- Data wrangling
- Quantative data summary
- Visual data summary

Disclaimer: Lot's of new terminology. Focus on how `R` handles things

Review after lecture

Maintain a glossary of functions used.

What is data?

What is data?

Data can be any unprocessed fact, value, text, sound file, image, video . . .

Examples

- your homework files
- photos
- each click on a website
- each transation at your bank, credit card, grocery store

Types of statistical data and formats

Types of Statistical Data

Type of data determines the analysis or models you can use

Quantitative or numeric: Numeric information

- Example: height, weight, age, GPA
- can use math functions eg. average, max, min, sum
- not everything represented by a number is a quantitative/numeric variable
  - Zipcode, StudentID

Qualitative or categorical: descriptions (usually words)

- Example: Eye color, state of residence,
- can't use math functions
- categorical variables have **levels**

```
levels(penguins$species)
```

```
## [1] "Adelie"    "Chinstrap" "Gentoo"
```

Types of Numerical data

- **Discrete** : data that can be counted and therefore can take on only certain values
  - eg. shoe size, number of questions answered correctly

- **Continuous** : data that is measured on an infinite scale
  - eg. Height, weight, temperature

Types of Categorical data

- **Ordinal** : data that can be ordered or data on a scale
  - eg. income (low, medium, high)

- **Nominal** : data with no apparent order to it
  - eg. gender

Let's Practice!: Identify the type of data

- Age : 12, 13, 17 years old
- Spice level: mild, medium, hot
- Temperature: 77.5 , 80.2, 73
- Eye color: green, blue, brown, black

Structure or format of data

What is a dataset?

A dataset is any collection of data

Typically, a dataset contains data in tabular form:

- **Variables** across the columns
- **Observations** or data points down the rows

What does data *look like* ? Where is the dataset?

- generally a .csv file

```
species,island,bill_length_mm,bill_depth_mm,flipper_length_mm,body_mass_g,sex
Adelie,Torgersen,39.1,18.7,181,3750,MALE
Adelie,Torgersen,39.5,17.4,186,3800,FEMALE
Adelie,Torgersen,40.3,18,195,3250,FEMALE
Adelie,Torgersen,,,,,
Adelie,Torgersen,36.7,19.3,193,3450,FEMALE
Adelie,Torgersen,39.3,20.6,190,3650,MALE
Adelie,Torgersen,38.9,17.8,181,3625,FEMALE
Adelie,Torgersen,39.2,19.6,195,4675,MALE
Adelie,Torgersen,34.1,18.1,193,3475,
Adelie,Torgersen,42,20.2,190,4250,
Adelie,Torgersen,37.8,17.1,186,3300,
```

- A .csv file can be loaded into an `R` as a data frame.

.csv file as a data frame in R

Go to file/function Addins

penguins

Filter

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
1	Adelie	Torgersen	39.1	18.7	181	3750	male	2007
2	Adelie	Torgersen	39.5	17.4	186	3800	female	2007
3	Adelie	Torgersen	40.3	18.0	195	3250	female	2007
4	Adelie	Torgersen	NA	NA	NA	NA	NA	2007
5	Adelie	Torgersen	36.7	19.3	193	3450	female	2007
6	Adelie	Torgersen	39.3	20.6	190	3650	male	2007
7	Adelie	Torgersen	38.9	17.8	181	3625	female	2007
8	Adelie	Torgersen	39.2	19.6	195	4675	male	2007
9	Adelie	Torgersen	34.1	18.1	193	3475	NA	2007
10	Adelie	Torgersen	42.0	20.2	190	4250	NA	2007
11	Adelie	Torgersen	37.8	17.1	186	3300	NA	2007
12	Adelie	Torgersen	37.8	17.3	180	3700	NA	2007
13	Adelie	Torgersen	41.1	17.6	182	3200	female	2007
14	Adelie	Torgersen	38.6	21.2	191	3800	male	2007
15	Adelie	Torgersen	34.6	21.1	198	4400	male	2007
16	Adelie	Torgersen	36.6	17.8	185	3700	female	2007
17	Adelie	Torgersen	38.7	19.0	195	3450	female	2007

Showing 1 to 17 of 344 entries, 8 total columns

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA)

Goals:

- Understand data type, shape & structure
- Investigate important variables and groups
- Identify potential outliers
- Explore patterns in data

Simple EDA Techniques

Simple EDA Techniques

- **Data wrangling:** Inspect the data and data types, handle missing data.

- **Quantitative data summary:** Calculate descriptive statistics of each column such as mean, standard deviation to know the center and spread for each variable(column)

- **Visual data summary:** Create simple visualizations of the data such as histograms, box plots, bar plots, scatter plots to see distribution of data.

Exploring `penguins` dataset in `R`

Go to file/function | Addins

penguins

Filter

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
1	Adelie	Torgersen	39.1	18.7	181	3750	male	2007
2	Adelie	Torgersen	39.5	17.4	186	3800	female	2007
3	Adelie	Torgersen	40.3	18.0	195	3250	female	2007
4	Adelie	Torgersen	NA	NA	NA	NA	NA	2007
5	Adelie	Torgersen	36.7	19.3	193	3450	female	2007
6	Adelie	Torgersen	39.3	20.6	190	3650	male	2007
7	Adelie	Torgersen	38.9	17.8	181	3625	female	2007
8	Adelie	Torgersen	39.2	19.6	195	4675	male	2007
9	Adelie	Torgersen	34.1	18.1	193	3475	NA	2007
10	Adelie	Torgersen	42.0	20.2	190	4250	NA	2007
11	Adelie	Torgersen	37.8	17.1	186	3300	NA	2007
12	Adelie	Torgersen	37.8	17.3	180	3700	NA	2007
13	Adelie	Torgersen	41.1	17.6	182	3200	female	2007
14	Adelie	Torgersen	38.6	21.2	191	3800	male	2007
15	Adelie	Torgersen	34.6	21.1	198	4400	male	2007
16	Adelie	Torgersen	36.6	17.8	185	3700	female	2007
17	Adelie	Torgersen	38.7	19.0	195	3450	female	2007

Showing 1 to 17 of 344 entries, 8 total columns

The `penguins` dataset is stored in a **data frame** with

- **344 observations/samples/cases/subjects** (rows)
  - each case represents a penguin
- **8 variables** (columns)
  - species, island, bill_length_mm, bill_depth_mm etc
  - each corresponds to some measurement of the penguin

Work Along

Explore the penguins dataset using `R` functions

Quantitative Data Summary

Quantitative Data Summary

Describing/Summarizing data with numbers

Summarizing Categorical Data with numbers

Summarizing categorical data : table

Categorical data are summarized with counts or proportions.

```
table(penguins$species)
```

```
##
##    Adelie Chinstrap    Gentoo
##       152        68       124
```

```
prop.table(table(penguins$species))
```

```
##
##    Adelie Chinstrap    Gentoo
## 0.4418605 0.1976744 0.3604651
```

There are 152 or 44.19% penguins of Adelie species etc

Summarizing Numerical Data with numbers

Summarizing Numerical Data: Descriptive Statistics

- Measures of center
  - mean, median, mode
- Measures of spread
  - range, variance, standard deviation

Measures of center

- Mean : Average value of the data
- Median: Middle value of the data
- Mode : Most frequently occuring value

Measures of center: Mean

Mean of a list of numbers $= \bar{x} = \frac{\text{sum of numbers}}{\text{how many numbers in the list}}$

```r
x <- c(1, 5, 1, 2, 5, 4, 6)
mean(x)
```

```
## [1] 3.428571
```

```r
sum(x)/length(x)
```

```
## [1] 3.428571
```

If you change the numbers in the list then the mean(average) changes

Measures of center: Median

```r
x <- c(1, 5, 3, 2, 5, 4, 6)
x <- sort(x)
x
```

```
## [1] 1 2 3 4 5 5 6
```

```r
median(x)
```

```
## [1] 4
```

Median is not sensitive to changes in extreme values

```r
x <- c(1, 5, 3, 2, 5, 4, 600)
median(x)
```

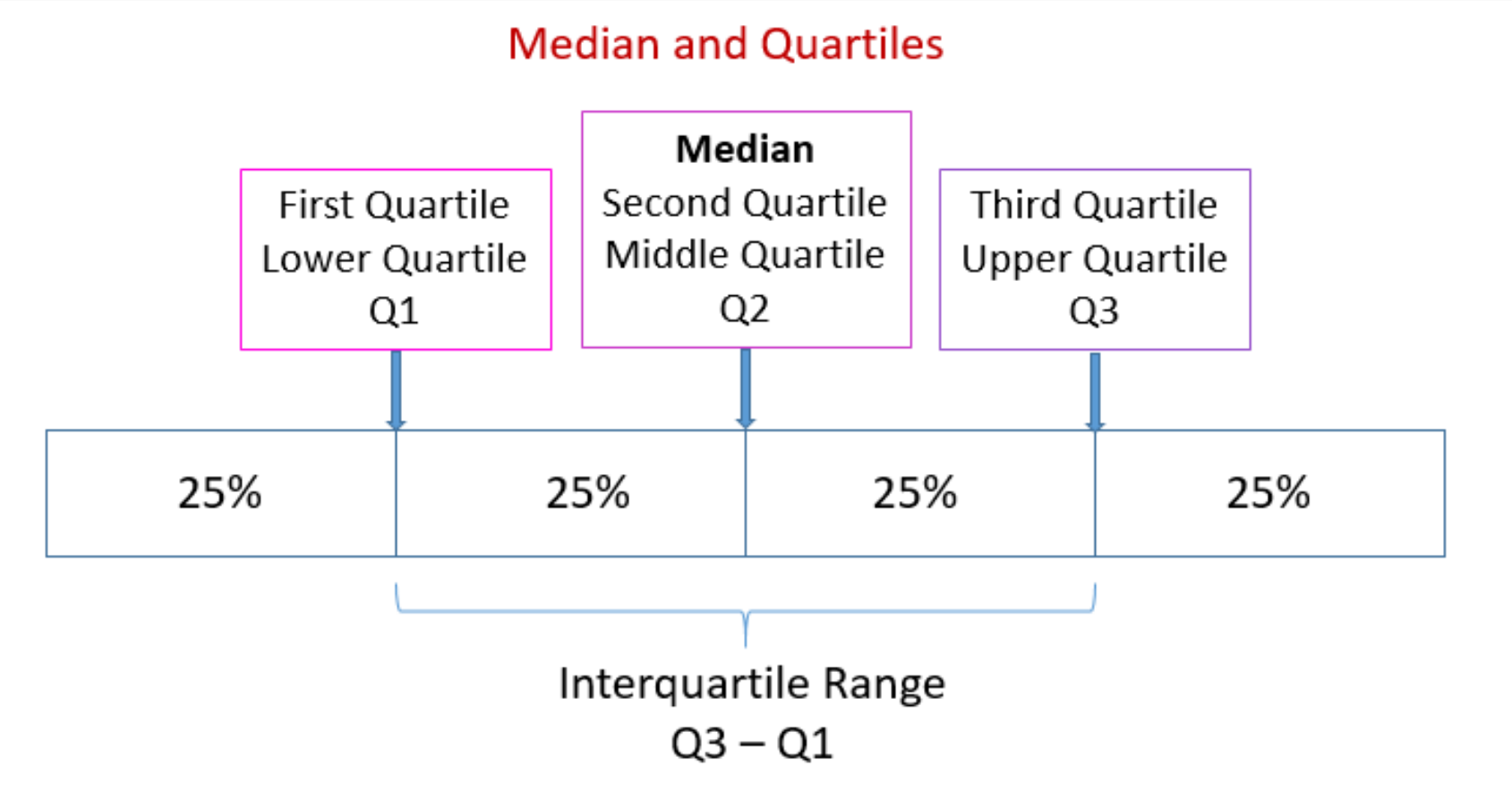
```
## [1] 4
```

Measures of center: Mode

```r
x <- c(1, 5, 3, 2, 5, 4, 6)
sort(table(x))
```

```
## x
## 1 2 3 4 6 5
## 1 1 1 1 1 2
```

Descriptive Summary: Quartiles: Q1, Q3, and Interquartile Range



Interquartile Range (IQR) = Q3 - Q1 which represents the middle 50% of the data.

Quartiles in R

```r
x <- c(1, 5, 3, 2, 5, 4, 6)
quantile(x, .25) # quantile function not quartile!
```

```
## 25%
## 2.5
```

```r
quantile(x, .75)
```

```
## 75%
##   5
```

Measures of spread

Below are midterm results from three classes.

Class 1: 80 80 80 80 80

Class 2: 76 78 80 82 84

Class 3: 60 70 80 90 100

What do you notice about midterm results?

Measures of spread

- Range
- Variance
- Standard Deviation

Give insight into how "spread out" the data is from the mean or average

- small spread: the data is packed near the center
- large spread, the data is spread spread out or not concentrated near the center.

Measures of spread: Range

- difference between the lowest and highest values.

```
x <- c(1, 5, 3, 2, 5, 4, 6)
sort(x)
```

```
## [1] 1 2 3 4 5 5 6
```

```
max(x) - min(x)
```

```
## [1] 5
```

Measures of spread: Variance and standard deviation

sample variance = (average) of squared distance from the mean

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Where s^2 is the sample variance, x_i is a sample observation value, $\bar{x}$ is the sample mean, and n is the number of observations.

```
var(x)
```

```
## [1] 3.238095
```

Measures of spread: standard deviation

sample standard deviation = square root of variance

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

```
sqrt(var(x))
```

```
## [1] 1.799471
```

```
sd(x)
```

```
## [1] 1.799471
```

Summarizing numeric data : 5 number summary

```r
summary(penguins$bill_length_mm)
```

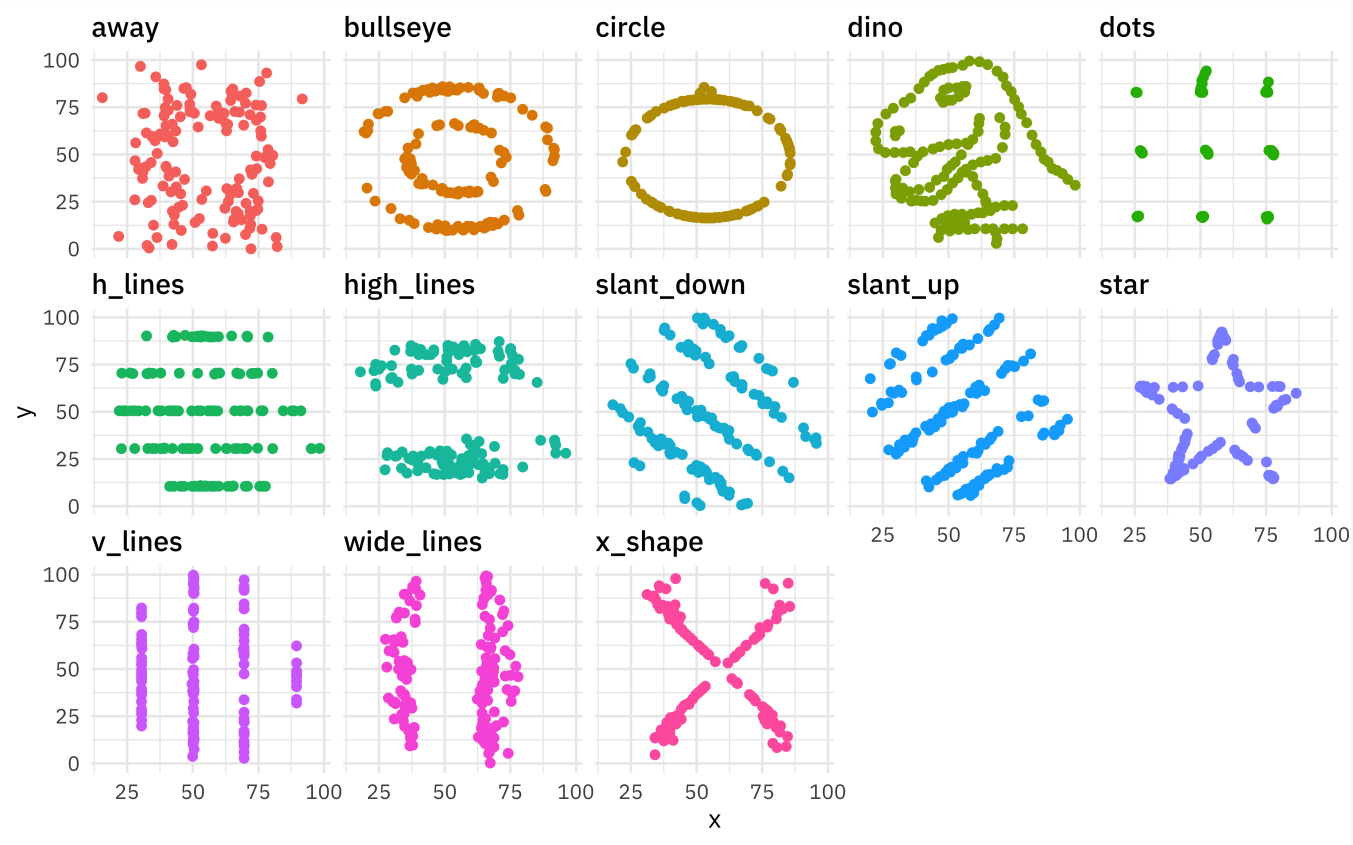
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   32.10   39.23   44.45   43.92   48.50   59.60       2
```

Includes measures of

- center - mean, median
- spread - range, quartiles,

Wrapping up summary statistics:

Beware summary statistics alone. . . meet the DINO DOZEN



Simple EDA Techniques

- **Data wrangling:** Inspect the data and data types, handle missing data.

- **Quantitative data summary:** Calculate descriptive statistics of each column such as mean, standard deviation to know the center and spread for each variable(column)

- **Visual data summary:** Create simple visualizations of the data such as histograms, box plots, bar plots, scatter plots to see distribution of data.

Visual data summary

Visual data summary

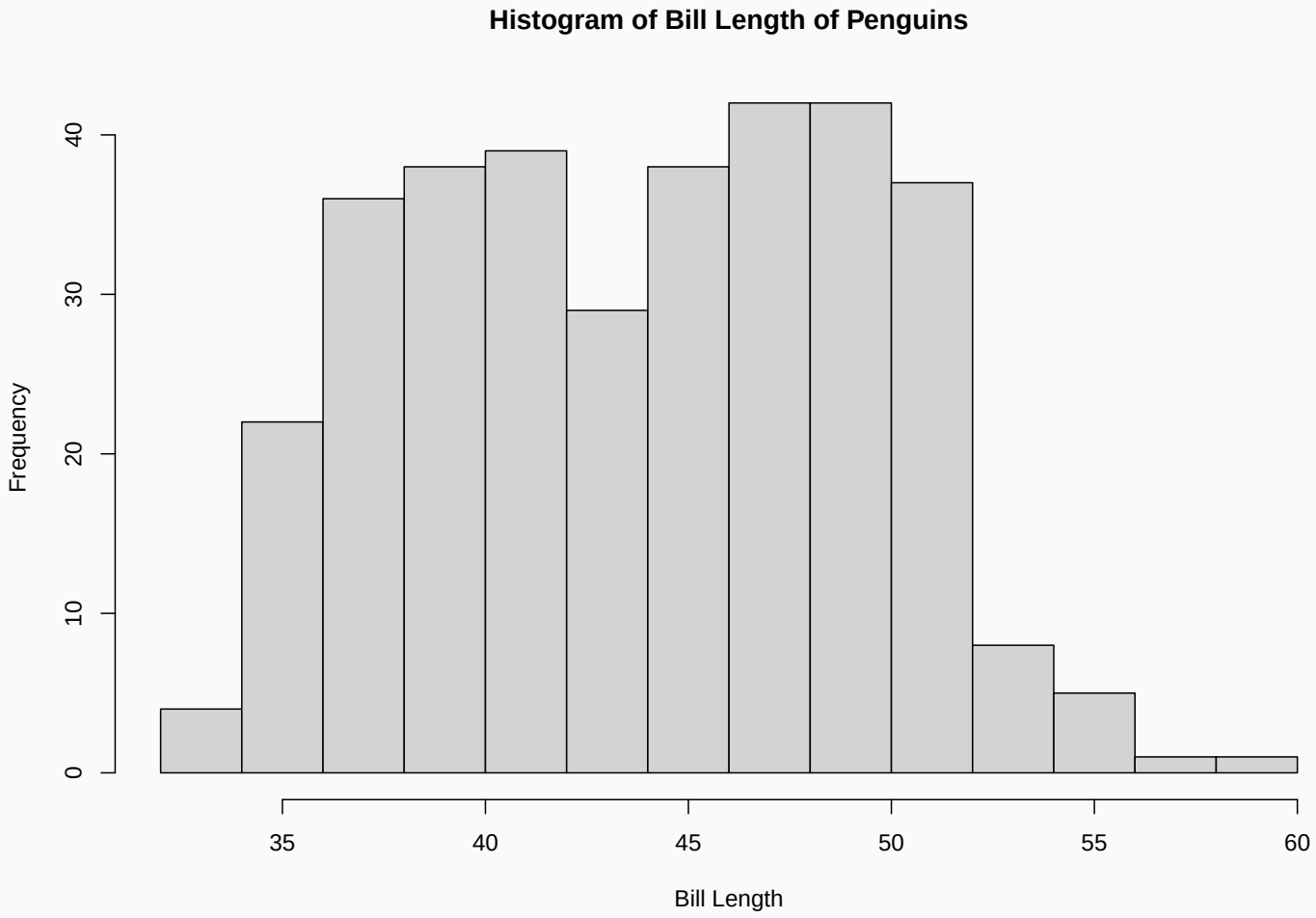
Describing/Summarizing data graphically by creating simple visualizations

Visualizing numeric data

Visualizing numerical data : Histograms

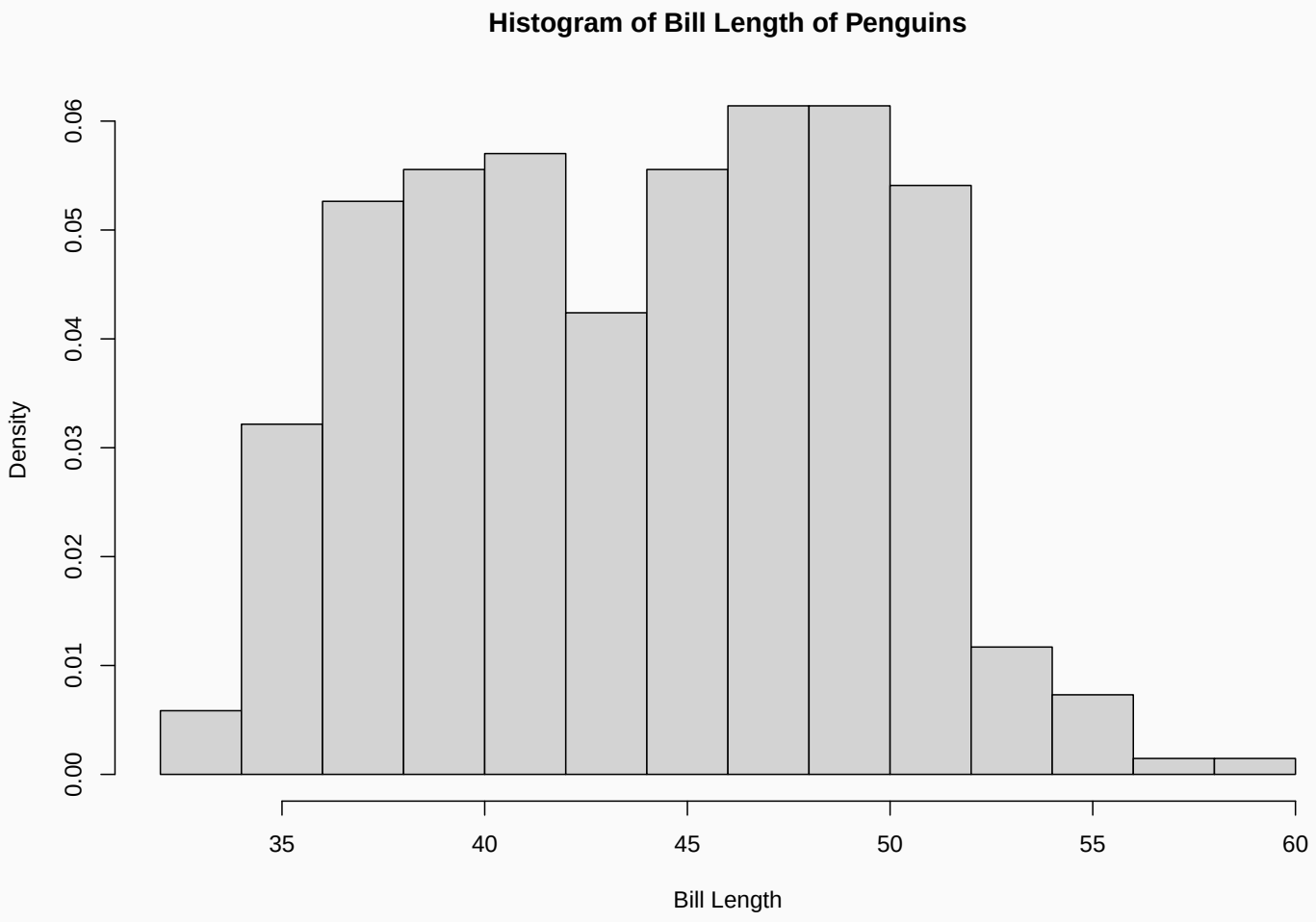
Count/Frequency histogram

```
hist(penguins$bill_length_mm,
     main = "Histogram of Bill Length of Penguins",
     xlab = "Bill Length")
```

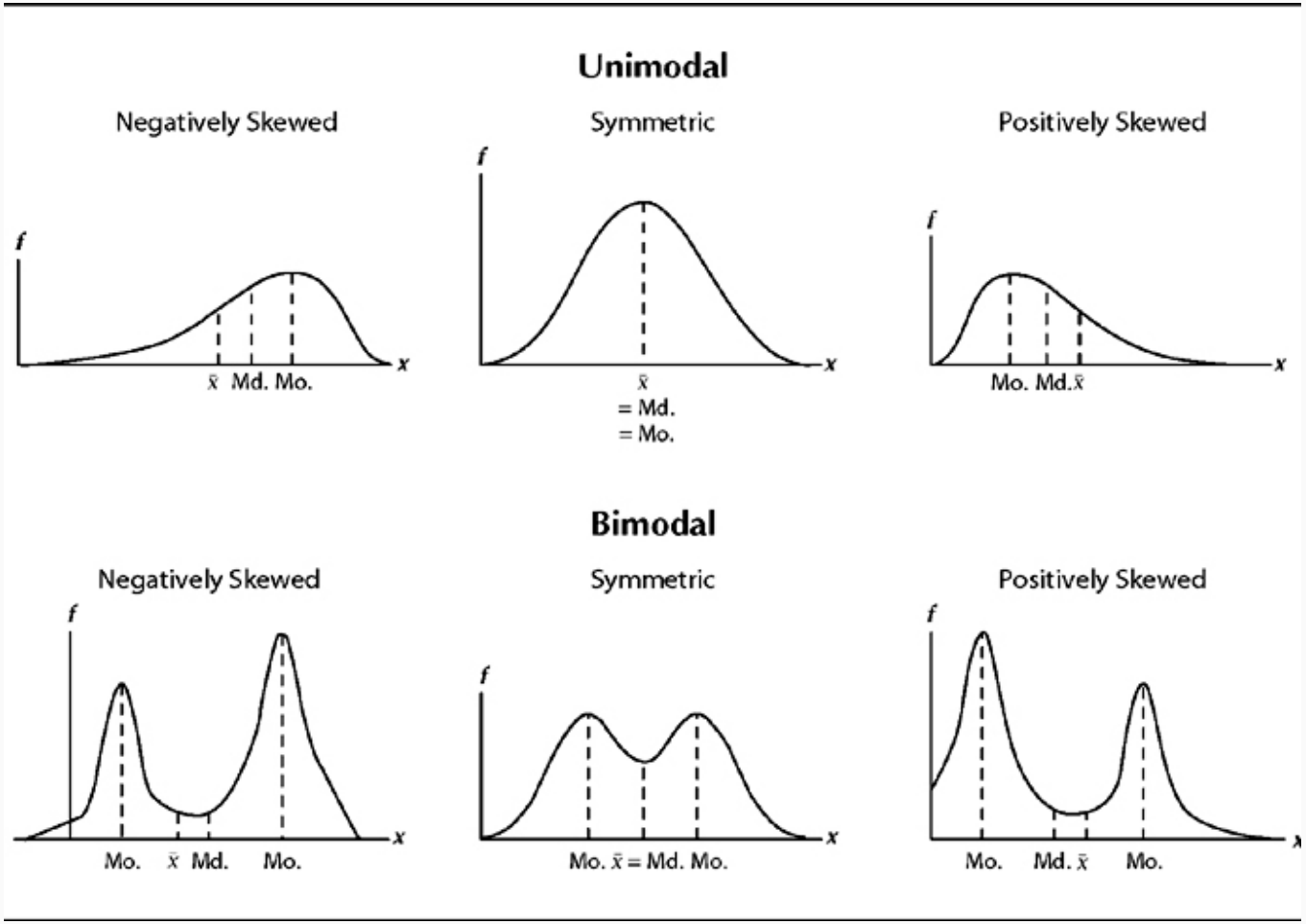


Probability/Density Histogram

```
hist(penguins$bill_length_mm, probability = TRUE,
     main = "Histogram of Bill Length of Penguins",
     xlab = "Bill Length")
```



Skewed data

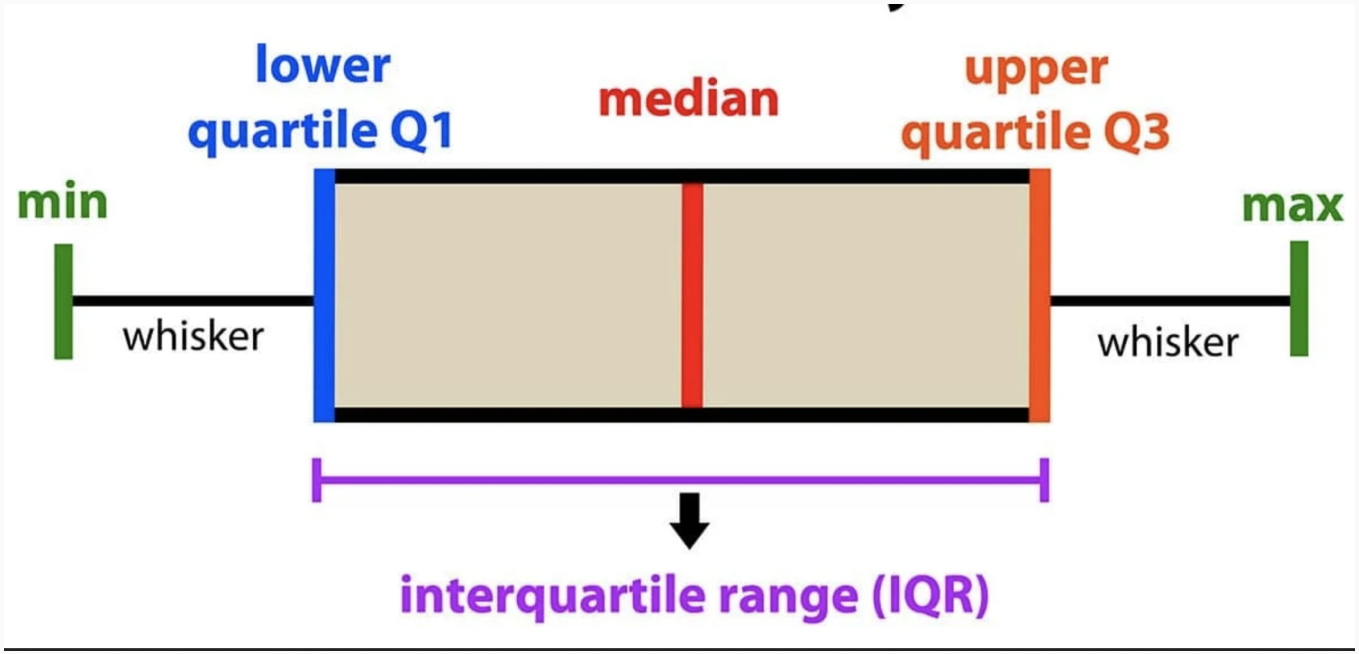


When data display a skewed distribution we rely on median rather than the mean to understand the center of the distribution.

Explore this cool site: Exploring Histograms Visually

Image: Sirkin, R. M. (2006). Measuring central tendency. In Statistics for the social sciences (pp. 83-126). SAGE Publications, Inc., https://www.doi.org/10.4135/9781412985987

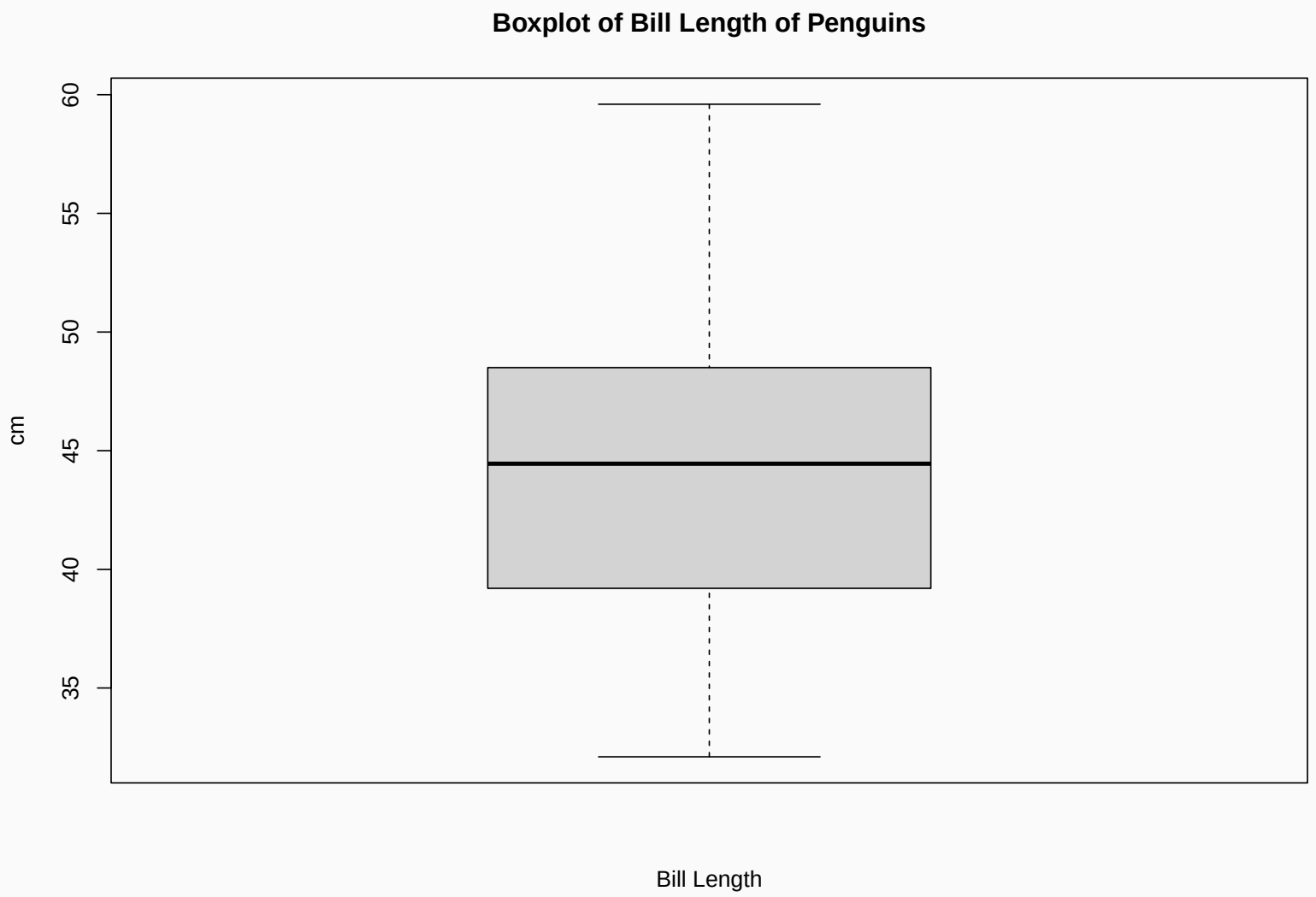
Visualizing 5-number summary: box plot



Whiskers can go as high as Q3 + 1.5 IQR and as low as Q1 - 1.5 IQR

Anything beyond whiskers indicated with a dot at the observation value are potential outliers

```r
boxplot(penguins$bill_length_mm,
        main ="Boxplot of Bill Length of Penguins",
     xlab = "Bill Length", ylab = "cm")
```

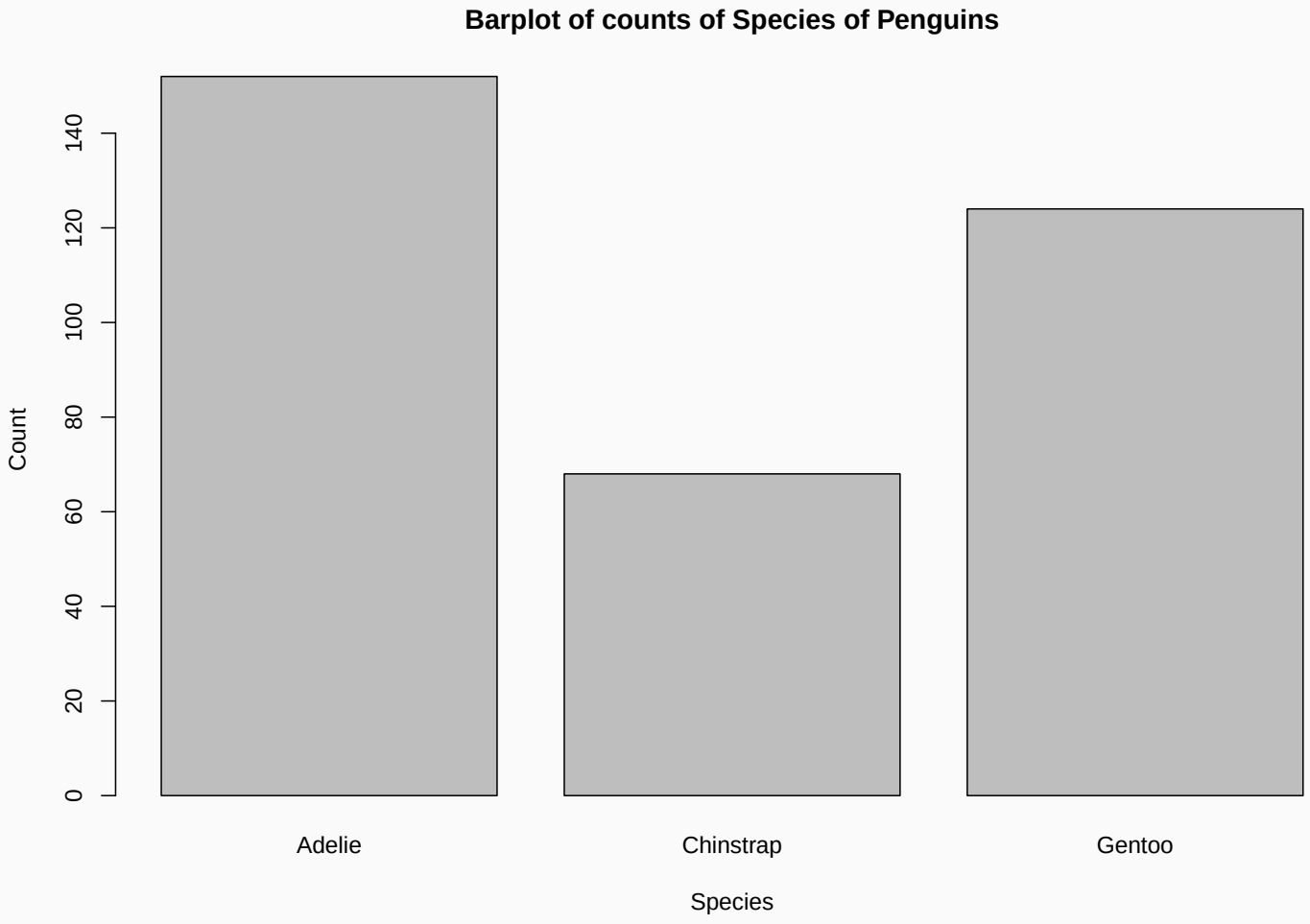


Visualising categorical data

Visualising categorical data : Bar plots

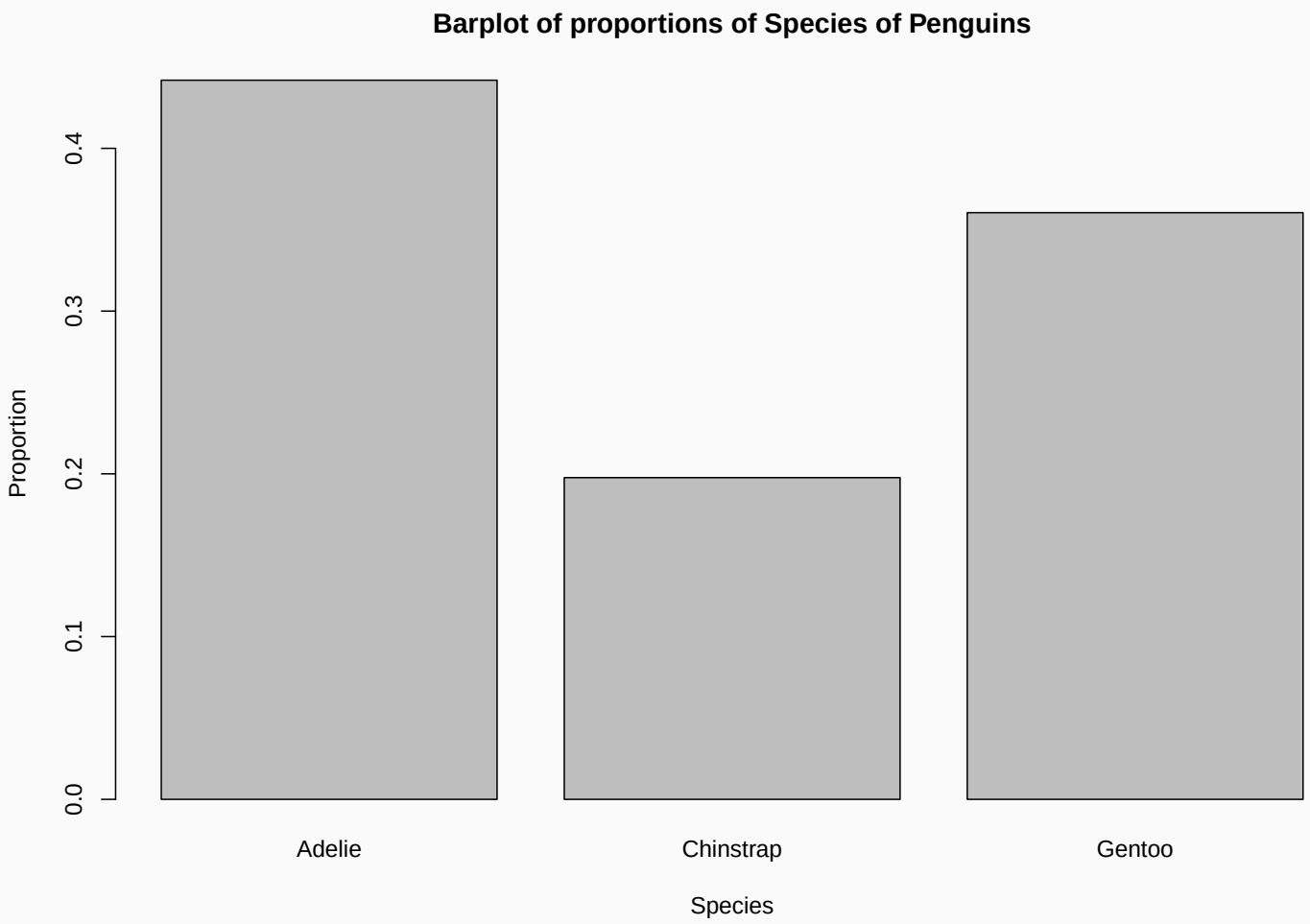
Frequency/Count Bar Plots

```
barplot(table(penguins$species ),
        main = "Barplot of counts of Species of Penguins",
     xlab = "Species", ylab = "Count")
```

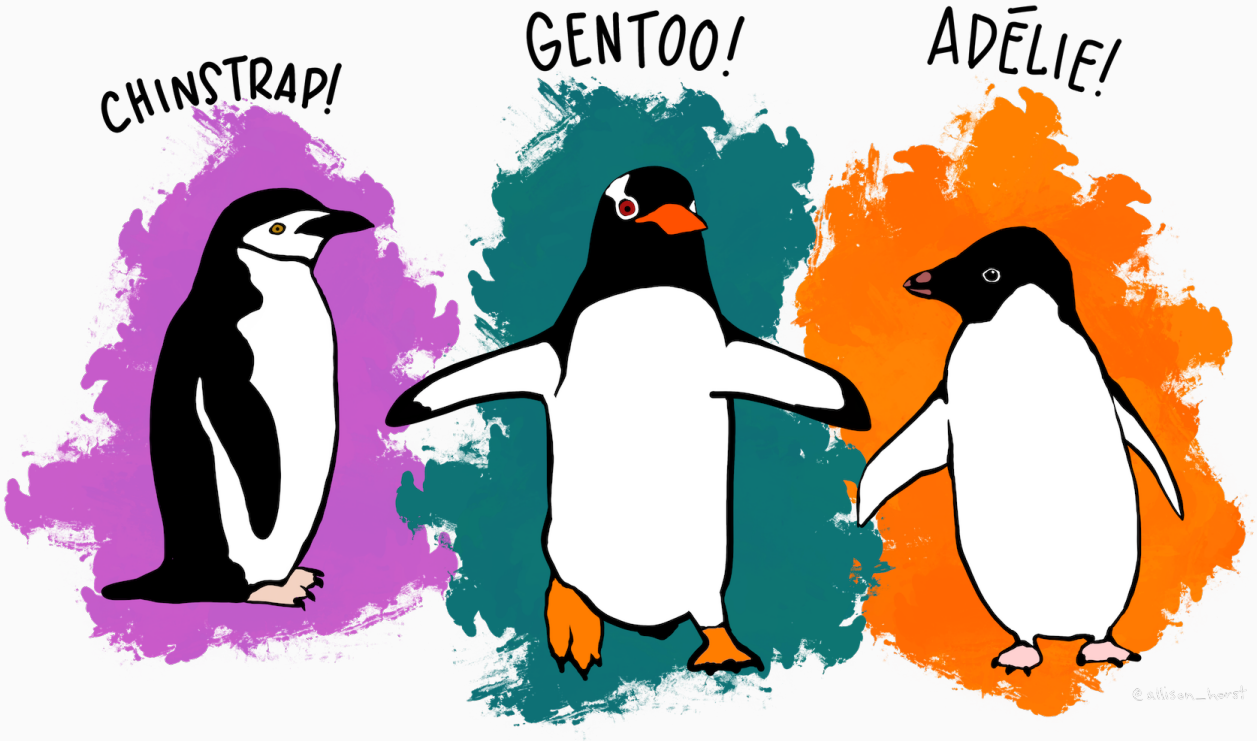


Proportion Bar Plots

```
barplot(prop.table(table(penguins$species) ),
        main = "Barplot of proportions of Species of Penguins",
     xlab = "Species", ylab = "Proportion")
```



Palmer Penguins[1]

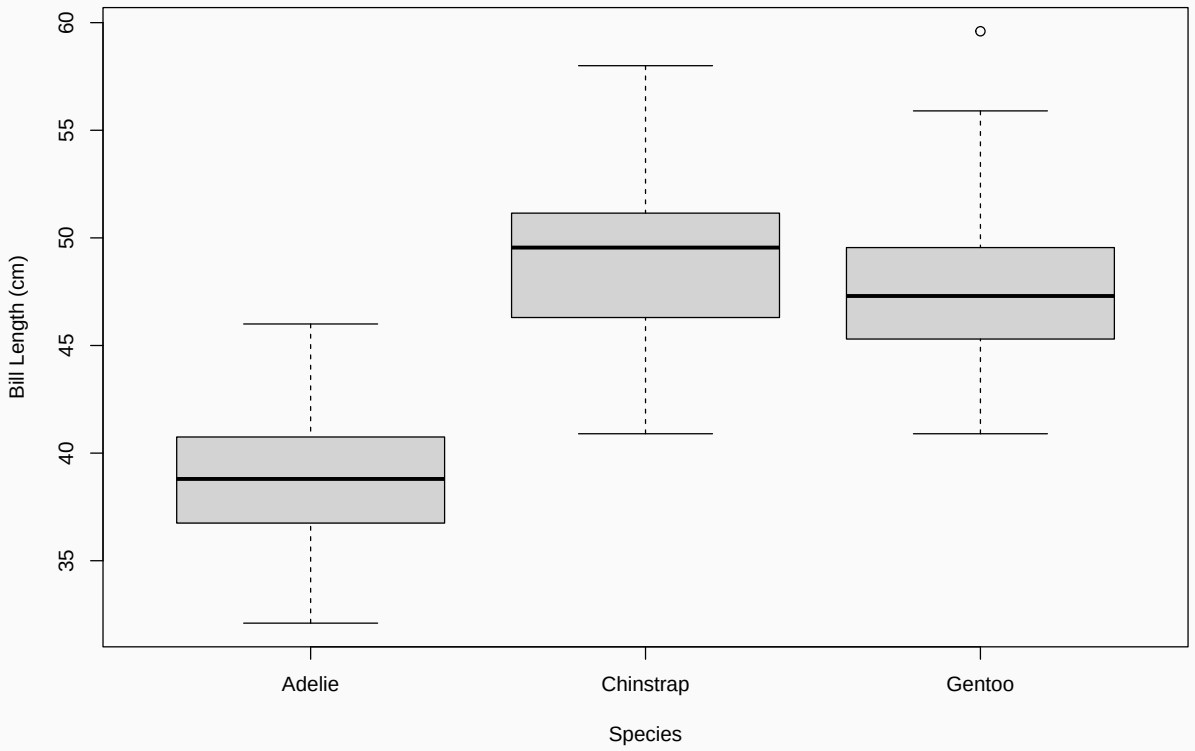


[1]Penguin artworks by @allison_horst.

Comparing more than one variable

Side-by-side Boxplots: Visualizing numerical ~ categorical variable together

```
boxplot(bill_length_mm ~ species, penguins,
        xlab = "Species", ylab = "Bill Length (cm)" )
```

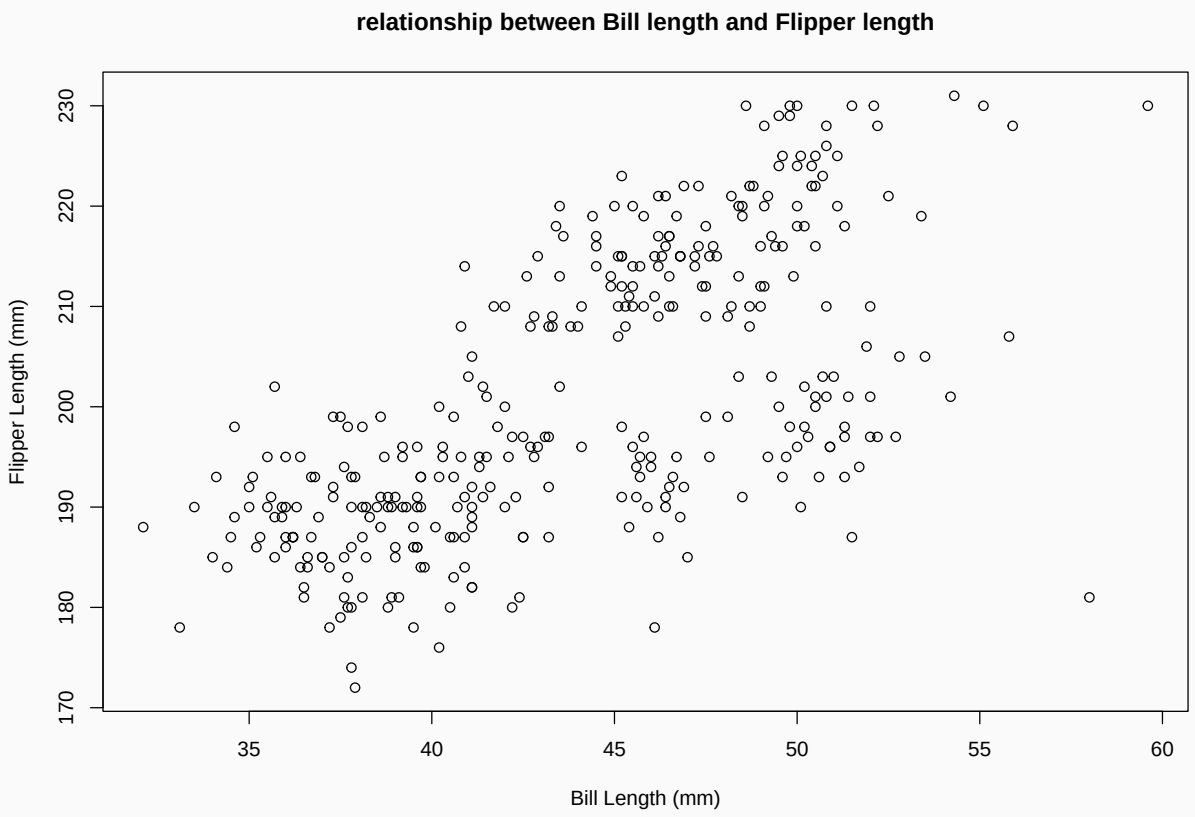


What can you conclude from this plot?

Side-by-side boxplots use the `y ~ x` notation. In R, this construct is called a `formula`

Scatterplot: Visualizing two numerical variables

```
plot(penguins$bill_length_mm,
     penguins$flipper_length_mm,
     xlab ="Bill Length (mm)", ylab ="Flipper Length (mm)",
     main = "relationship between Bill length and Flipper length")
```



What can you conclude from this plot?

Data Essentials Summary

Types of Statistical Data

- Numerical - discrete or continous
- Categorical - ordinal or nominal

EDA - Simple Techniques

- Data wrangling - variables, observations, data types
- Quantative data summary - center, spread, 5 number summary
- Visual data summary - bar plots, histogram, box plots

Disclaimer: Lot's of new terminology. Focus on how `R` handles things

Review after lecture

Maintain a glossary of functions used.

Next we will see...

- Conditionals
- Functions