5. Data essentials and Basic Plotting in `R`

Principles of Data Science with R

Dr. Uma Ravat

PSTAT 10

->

Summary:

- Factors (Textbook Chapter 4)
- Logical values (Textbook Chapter 4)

Even More data structures

- Lists and Data frames (Textbook Chapter 5)
- Special values (Textbook Chapter 6)

Maintain a glossary of functions used.

Next we will see...

- Basics for summarizing data visually
- Exploratory Data Analysis with `starwars` dataset
  - Plotting in `R`
- Importing and exporting data

->

->

->

->

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.
  - Use R essentials for accessing, subsetting

- **Quantitative data summary:** Calculate descriptive statistics of each column such as mean, standard deviation to know the center and spread for each variable(column)
  - Details when we move onto Module 2 on Probability and Statistics

- **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

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.

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

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

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

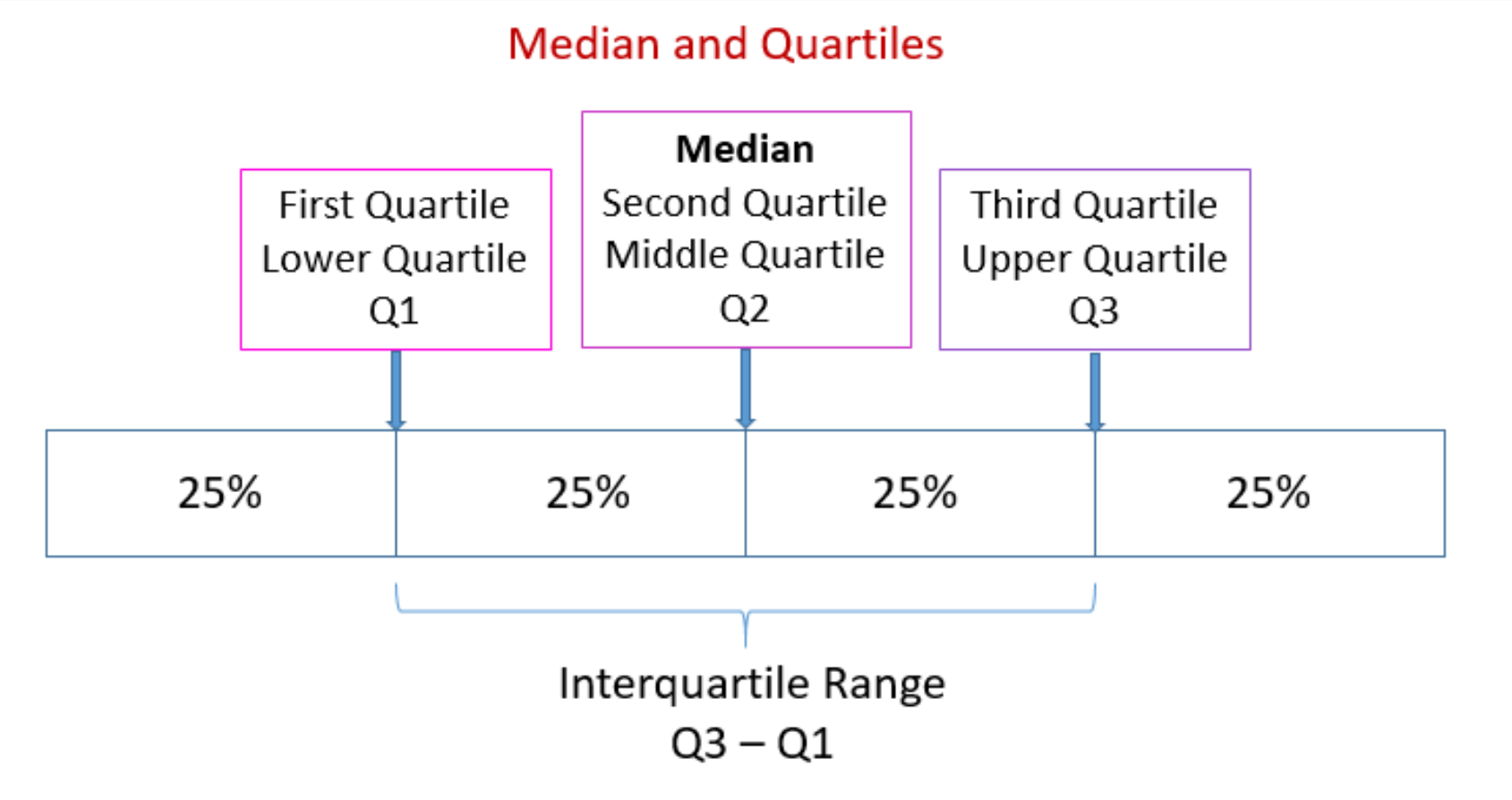
```
##
##    Adelie Chinstrap    Gentoo
## 0.4418605 0.1976744 0.3604651
```

There are 152 or 44.19% penguins of Adelie species etc

Summarizing Numerical Data with numbers

Summarizing Numerical Data with numbers

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



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

Summarizing numeric data : 5 number summary

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

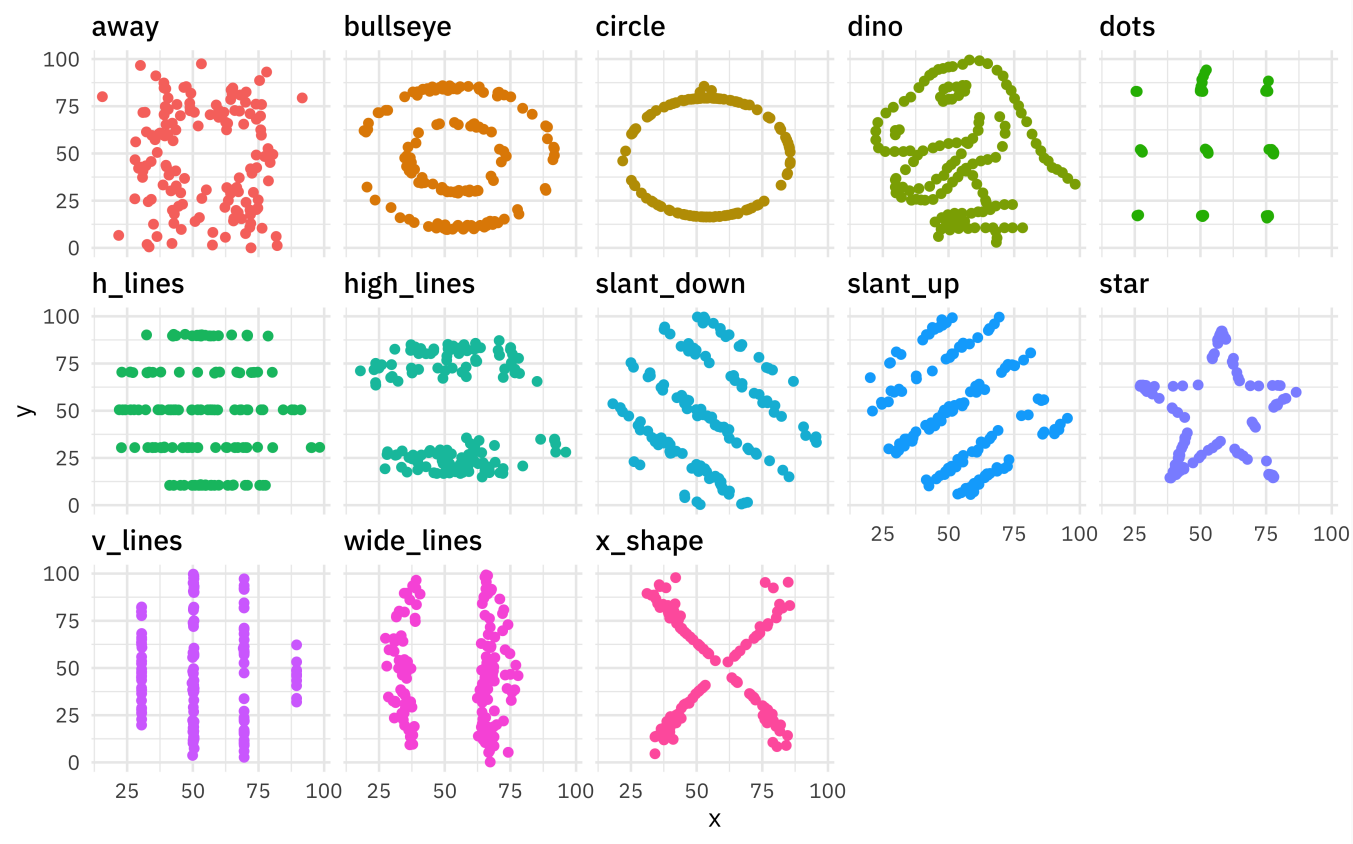
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   32.10   39.23   44.45   43.92   48.50   59.60       2
```

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

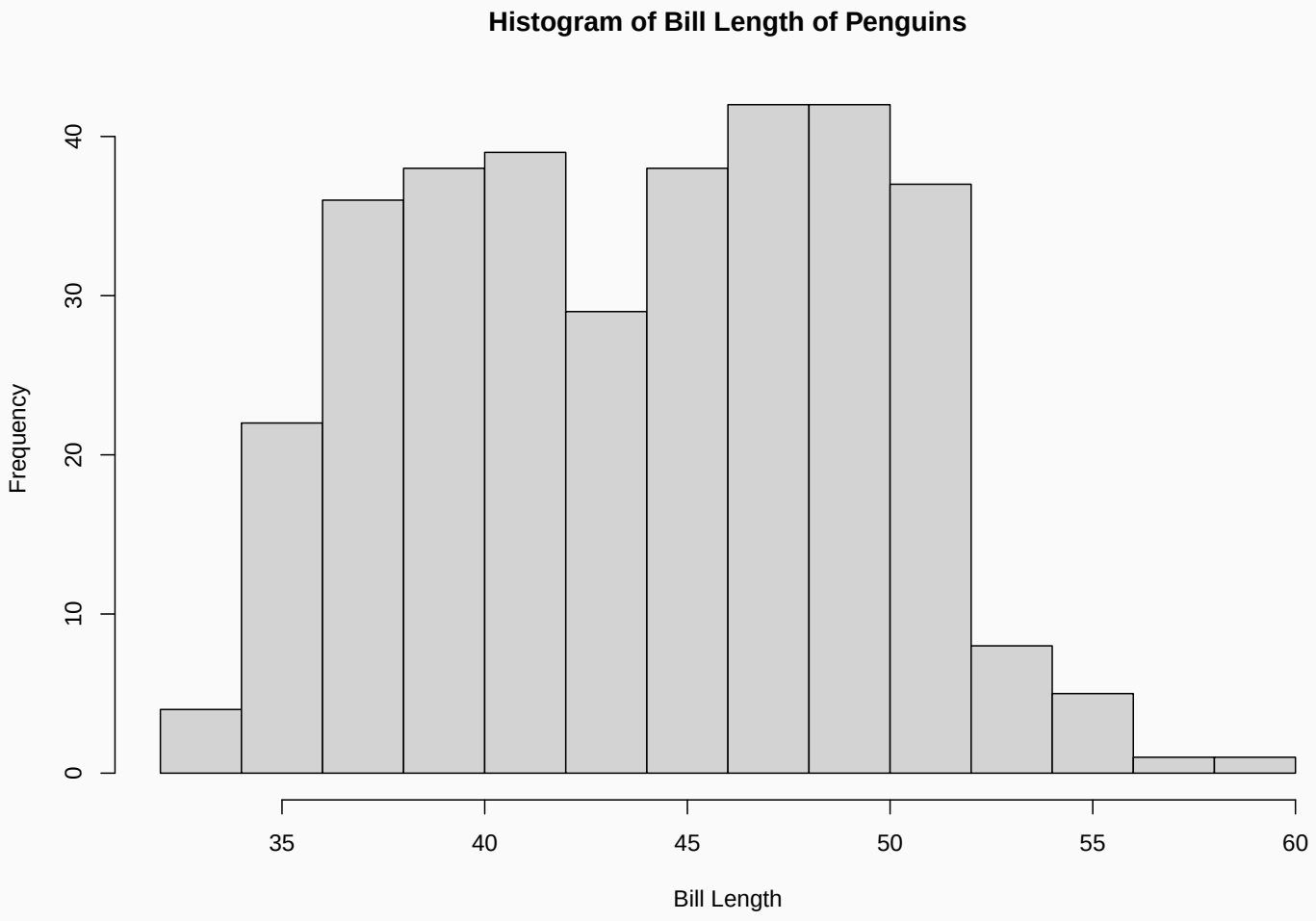
Describing/Summarizing data graphically by creating simple visualizations

Visualizing numeric data

Visualizing numerical data : Histograms

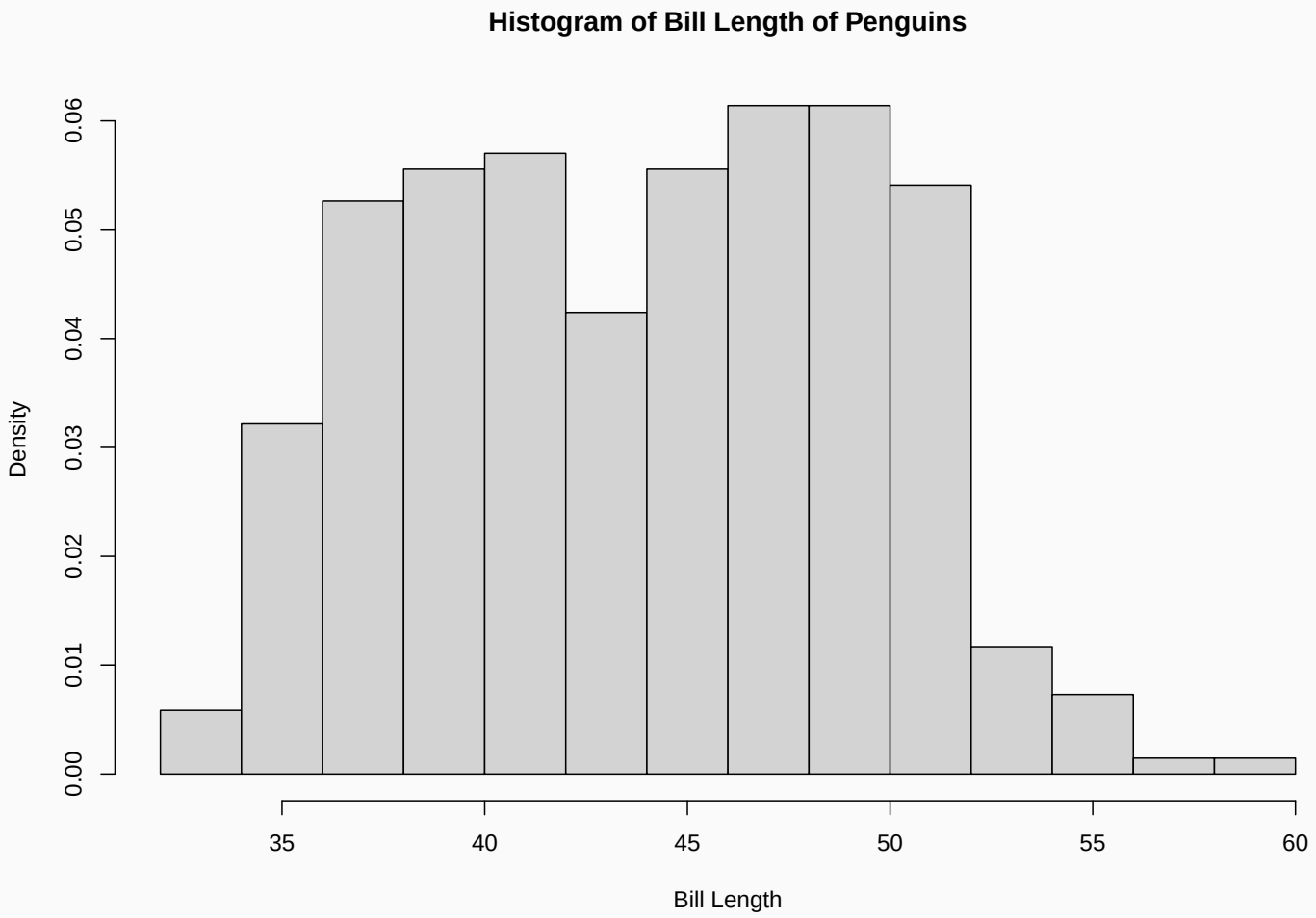
Count/Frequency histogram

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



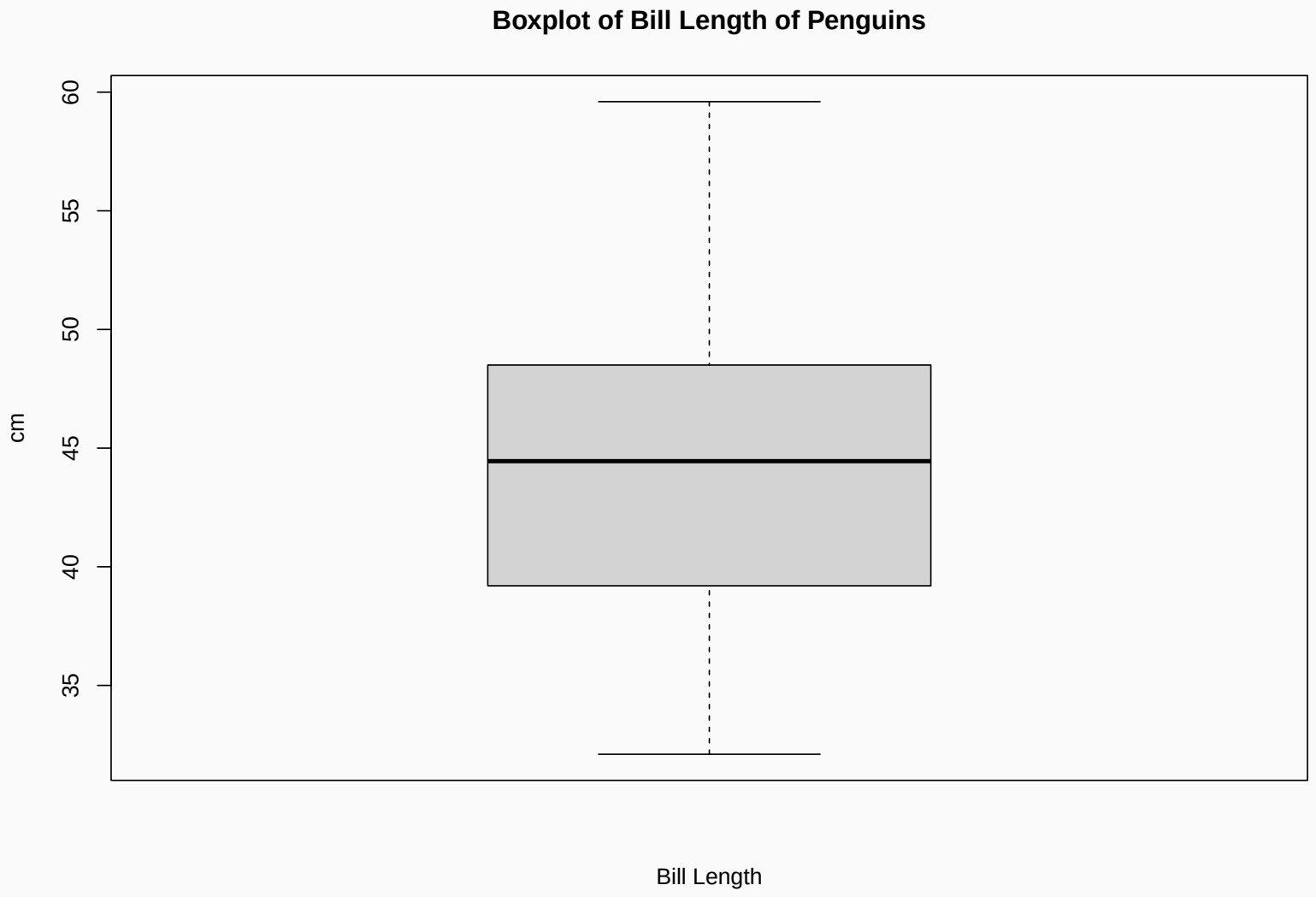
Probability/Density Histogram

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



Visualizing 5-number summary: box plot

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



Whiskers

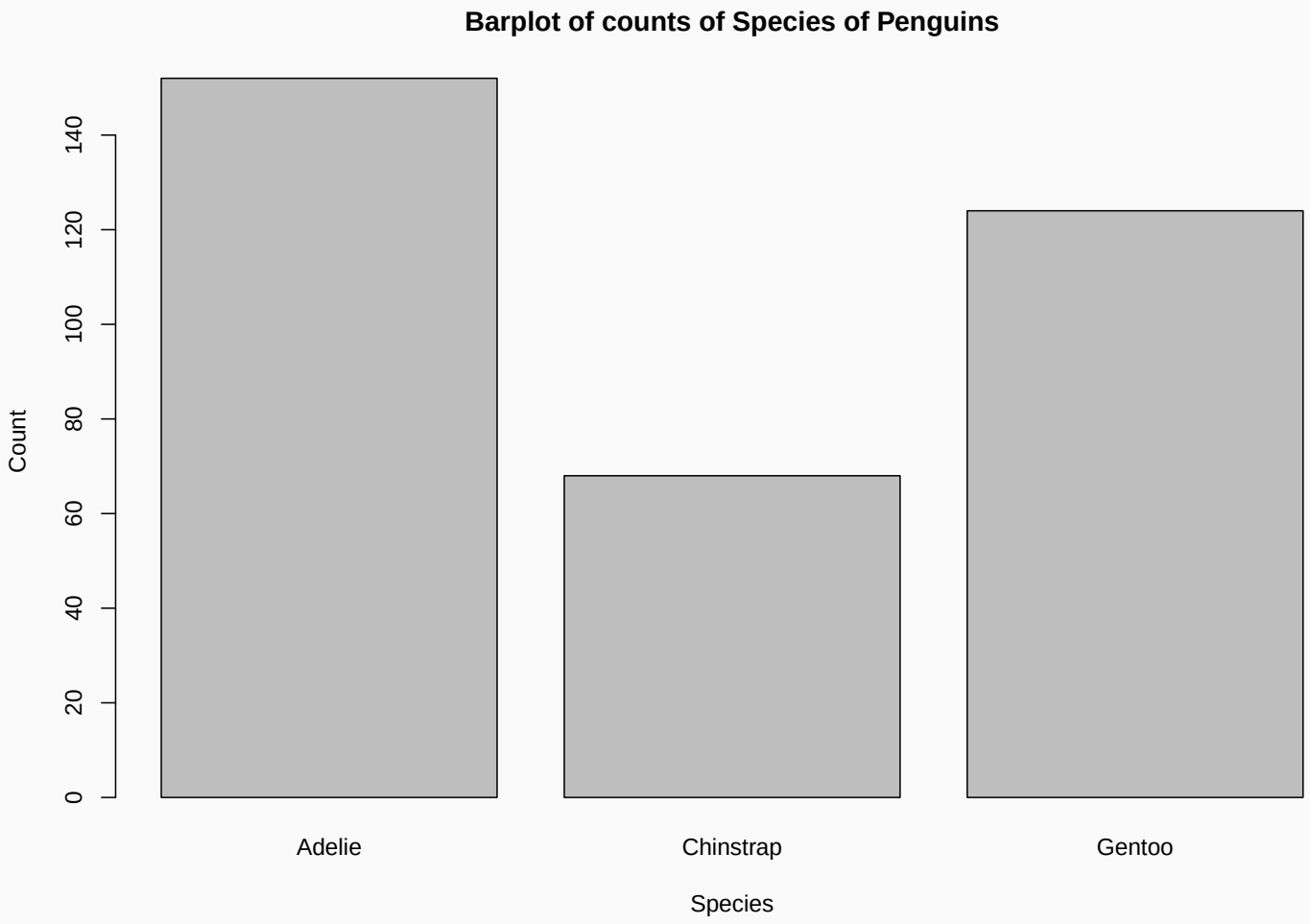
are drawn at Q1 - 1.5 IQR and Q3 + 1.5 IQR. Data beyond whiskers are considered outliers

Visualising categorical data

Visualising categorical data : Bar plots

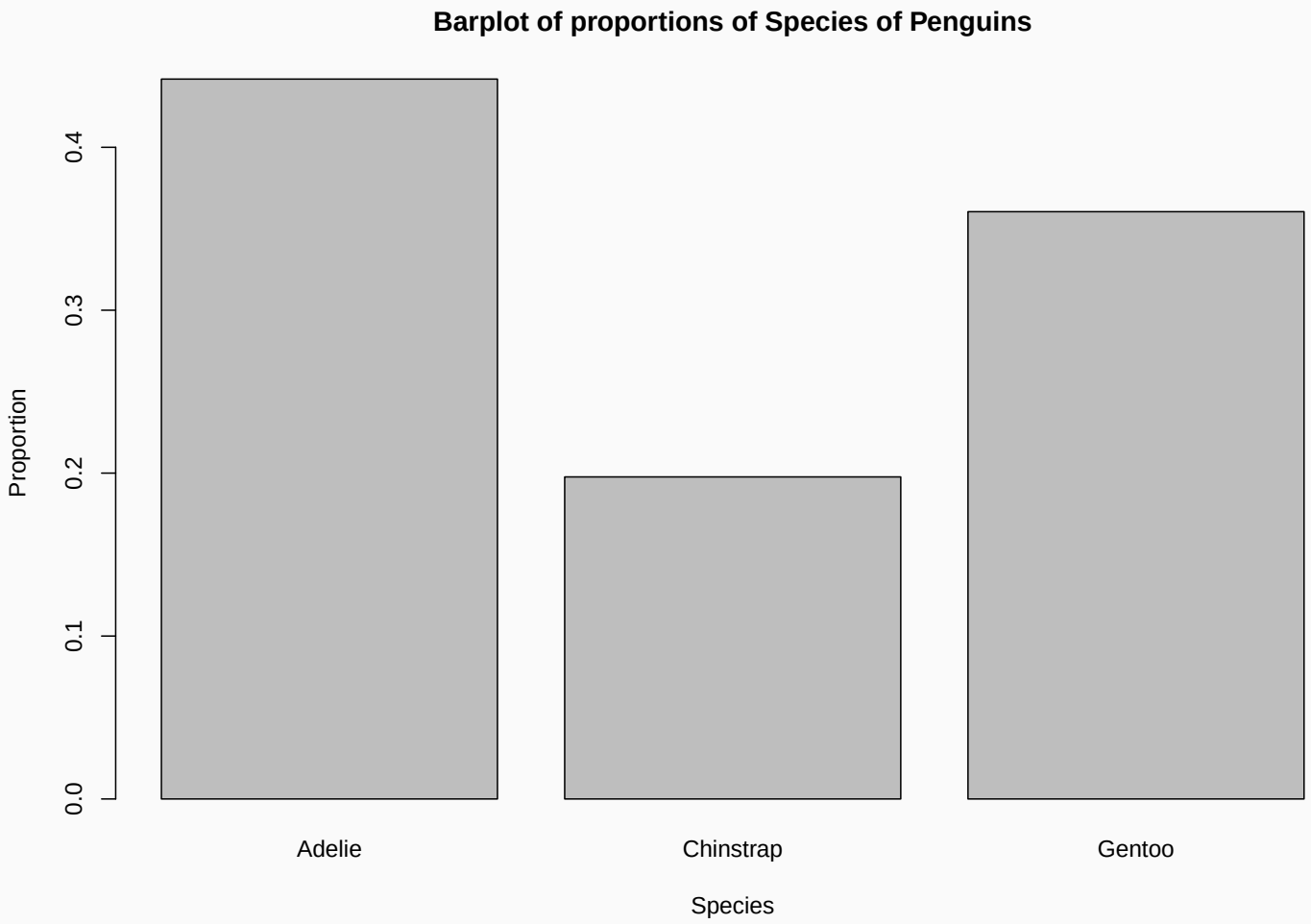
Frequency/Count Bar Plots

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

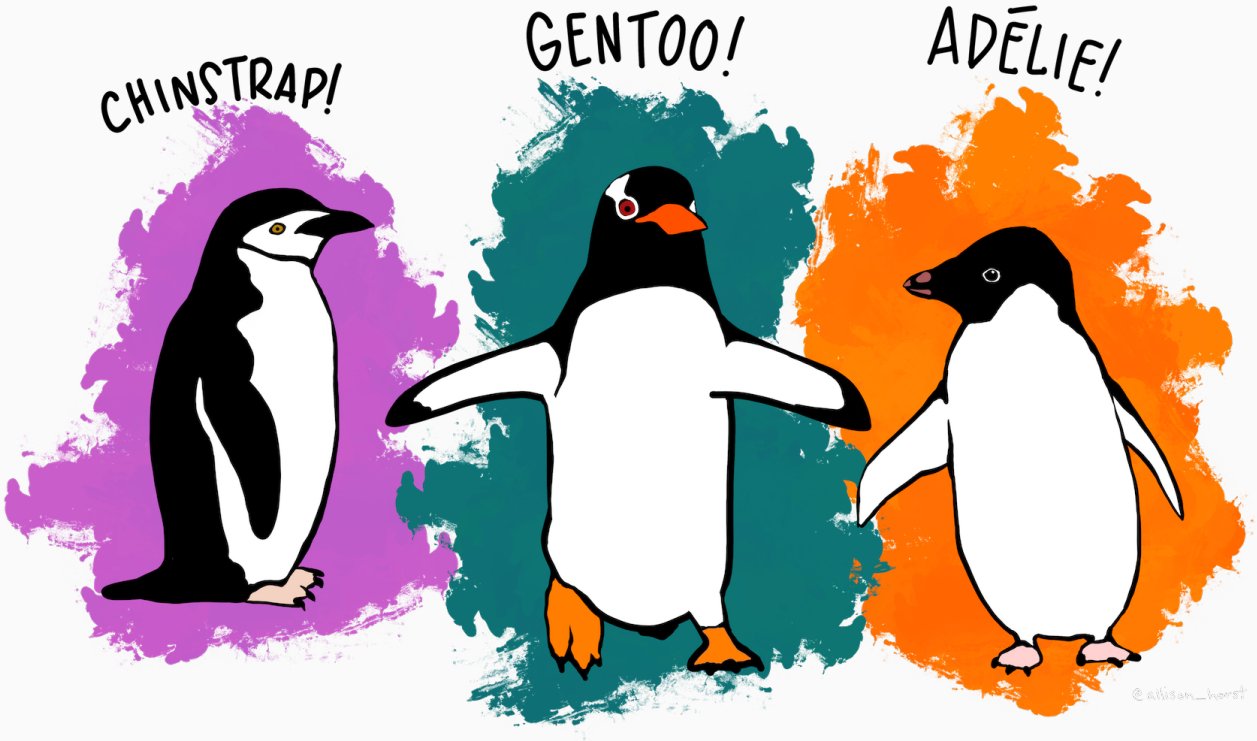


Proportion Bar Plots

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



Palmer Penguins[1]

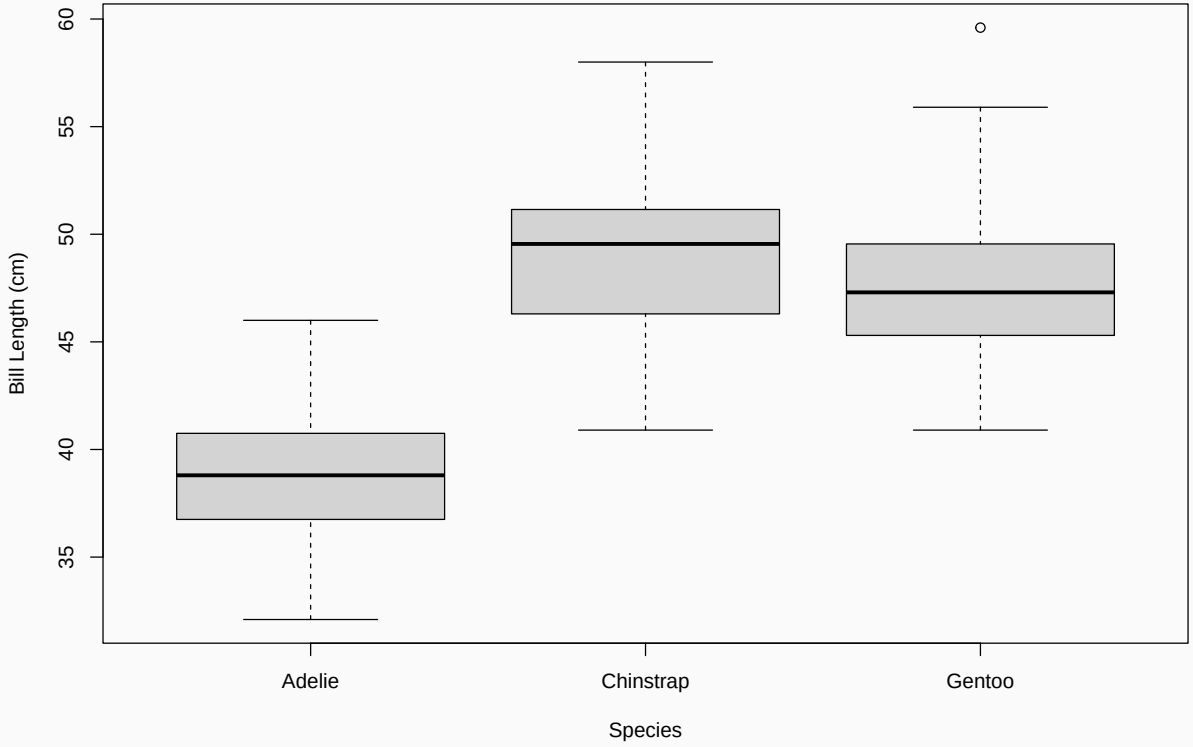


[1] Penguin artworks by @allison_horst.

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)" )
```

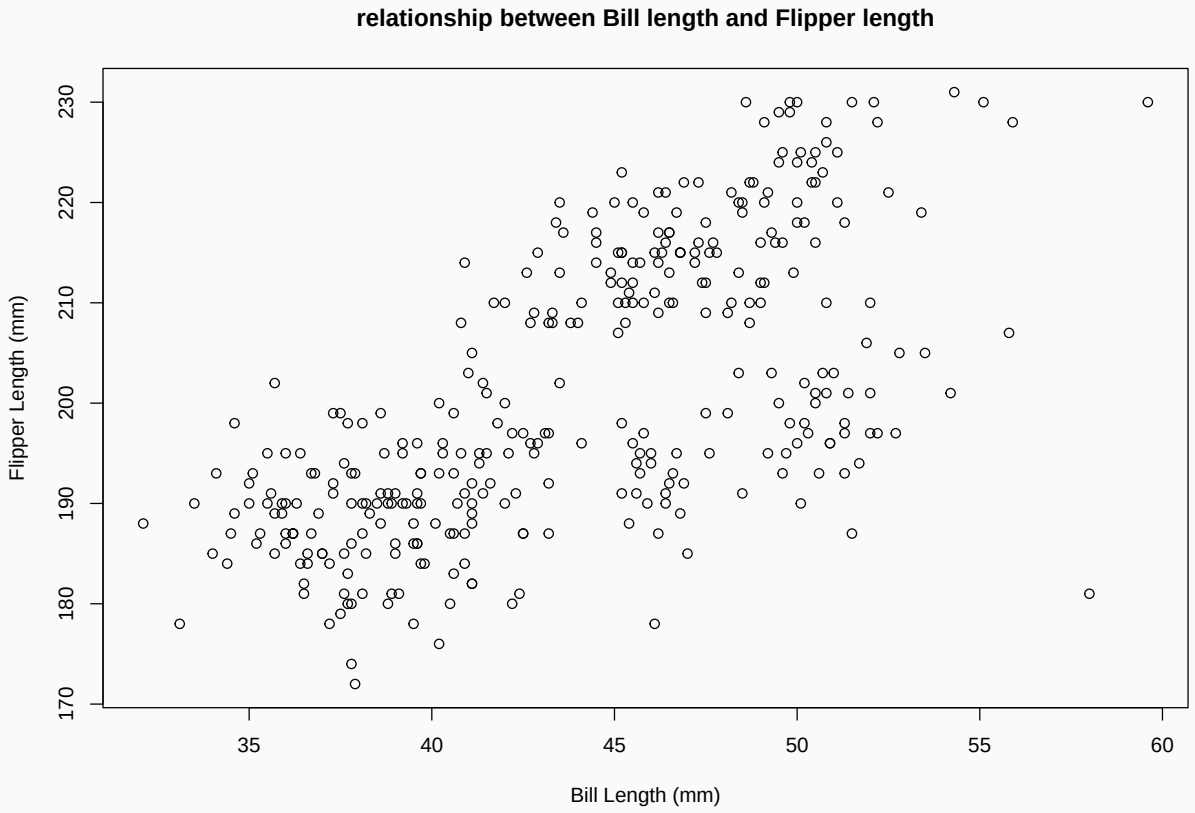


What can you conclude from this plot?

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?

PSTAT 10: Lecture 5 : Basic Plotting Tutorial

- Information
- Exploring a dataset
- Exploratory data analysis
- Basic plotting
- Numerical data
- Categorical Data
- Download answers

Start Over

Information

Name:

Enter answer here...

Submit Answer

Email:

Enter answer here...

Submit Answer

Continue

questions you should be able to answer

- What is EDA?
  - How do you do it?
- How do you plot, summarize different data types?
- How do you import, export data for your analysis?

Summary:

- EDA
- Plotting
- Importing and Exporting data

Maintain a glossary of functions used.

Learning Programming is HARD!

E. Kale Edmiston PhD
@EKaleEdmiston

Follow

A friend/colleague who is an excellent programmer offhandedly told me the other day that coding is 90% googling error messages & 10% writing code. Until this point, I thought that all the time I spent googling error messages meant I was bad at coding. What a perspective change!

8:12 AM - 4 Jan 2019

151 Retweets **1,069** Likes

27 151 1.1K