STAT 2593

Lecture 002 - Pictorial and Tabular Methods in Statistics

Dylan Spicker

Pictorial and Tabular Methods in Statistics

Learning Objectives

1. Define and characterize the distribution of a dataset.
2. Understand sample notation.
3. Explain the purposes of visualization.
4. Interpret and explain stem-and-leaf plots, dotplots, and histograms.

Raw data are effectively useless for interpretation.

```
"id","ht","age","baseht","baseage","logfev1"
1,1.200000004768372,9.341500282287 6,1.200000004768372,9.341500282287 6,0.215110003948212
1,1.279999997138977,10.39290044669189,1.200000004768372,9.341500282287 6,0.371560007333755
1,1.330000004291534,11.45240002075195,1.200000004768372,9.341500282287 6,0.488579988479614
1,1.419999995708466,12.46000000381 47,1.200000004768372,9.341500282287 6,0.751420021057129
1,1.480000001907349,13.41819995391846,1.200000004768372,9.341500282287 6,0.832910001277924
1,1.5,15.4743003845215,1.200000004768372,9.341500282287 6,0.892000019550323
1,1.519999989092651,16.3722991943359,1.200000004768372,9.341500282287 6,0.871290028095245
2,1.129999995231628,6.587299823760 99,1.129999995231628,6.587299823760 99,0.307480007410049
2,1.190000005722046,7.649600028991 7,1.129999995231628,6.587299823760 99,0.350659996271133
2,1.490000009536743,12.73919963836670 7,1.129999995231628,6.587299823760 99,0.756120026111603
2,1.529999971389771 3,13.77410003036499,1.129999995231628,6.587299823760 99,0.867100000381 47
2,1.549999952316284,14.69400002441406,1.129999995231628,6.587299823760 99,1.047320008827789
2,1.559999942779541,15.82199954986572,1.129999995231628,6.587299823760 99,1.153730003482819
2,1.570000005245209,16.66799926757812,1.129999995231628,6.587299823760 99,0.924260020256042
2,1.570000005245209,17.63179969787598,1.129999995231628,6.587299823760 99,1.134619951248169
3,1.179999947547913,6.913099765777599,1.179999947547913,6.913099765777599,0.431780010461807
3,1.230000019073486,7.975399971008301,1.179999947547913,6.913099765777599,0.385259985923767
3,1.299999952316284,8.966500282287 6,1.179999947547913,6.913099765777599,0.598839998245239
3,1.350000023841858,9.987700462341309,1.179999947547913,6.913099765777599,0.751420021057129
3,1.470000028610229,11.07730007171631,1.179999947547913,6.913099765777599,0.966979980846875
3,1.570000005245209,13.06779956817627,1.179999947547913,6.913099765777599,0.896089971065521
3,1.590000033378601,14.10270023345947,1.179999947547913,6.913099765777599,1.018849968891022
3,1.600000023841858,15.08010005950928,1.179999947547913,6.913099765777599,1.105260014534
3,1.600000023841858,16.01639938354492,1.179999947547913,6.913099765777599,1.085190057754 52
4,1.149999976158142,6.759799957275391,1.149999976158142,6.759799957275391,0.858269999213219
4,1.210000038146973,7.822000026702881,1.149999976158142,6.759799957275391,0.182319998741 15
4,1.259999990463257,8.813099861145020,1.149999976158142,6.759799957275391,0.277630001306534
4,1.309999942779541,9.834400177001953,1.149999976158142,6.759799957275391,0.444689989089966
4,1.389999985694885,10.92399978637695,1.149999976158142,6.759799957275391,0.576610028743744
4,1.460000038146973,11.93159961700439,1.149999976158142,6.759799957275391,0.672940015792847
4,1.539999961853027,12.91170024871826,1.149999976158142,6.759799957275391,0.722710013389587
4,1.590000033378601,13.94659996032715,1.149999976158142,6.759799957275391,1.022449970245361
4,1.600000023841858,14.86649990081787,1.149999976158142,6.759799957275391,1.036739969444275
4,1.629999995231628,15.78779983520508,1.149999976158142,6.759799957275391,1.187839984893799
5,1.110000014305115,6.502399921417236,1.110000014305115,6.502399921417236,0.829559993743896
5,1.149999976158142,7.564700126647949,1.110000014305115,6.502399921417236,0.113299989891243
5,1.519999980926514,13.73109993743896,1.110000014305115,6.502399921417236,0.896089971065521
5,1.539999961853027,14.70499992370605,1.110000014305115,6.502399921417236,0.862890005111694
5,1.549999952316284,15.58660030364990,1.110000014305115,6.502399921417236,0.955510020256042
5,1.549999952316284,16.64609909057617,1.110000014305115,6.502399921417236,0.924260020256042
5,1.559999942779541,17.52219963073730,1.110000014305115,6.502399921417236,0.832910001277924
6,1.240000009536743,6.899400234222412,1.240000009536743,6.899400234222412,0.262360006570816
6,1.299999952316284,7.961699962615967,1.240000009536743,6.899400234222412,0.476229995488912
6,1.360000014305115,8.985600471496582,1.240000009536743,6.899400234222412,0.565310001373291
6,1.409999966621399,9.973999977111816,1.240000009536743,6.899400234222412,0.712949991226196
6,1.470000028610229,11.06369972229,1.240000009536743,6.899400234222412,0.774730026721954
6,1.559999942779541,12.07120037078857,1.240000009536743,6.899400234222412,0.900160014629364
```

We use **graphical displays** to summarize the useful information instead.

The **distribution** of a dataset describes the possible values that are in a sample, and the relative frequency of those values.

Notation

- Our data consist of n observations.
- Each observation is denoted with a lowercase x.
- We use subscripts to enumerate sample observations, x_i, for $i = 1, \ldots, n$.

Goals for Visualizations

1. Identify what is **typical** in the data.

Goals for Visualizations

1. Identify what is **typical** in the data.
2. Identify the **spread** of the data.

Goals for Visualizations

1. Identify what is **typical** in the data.
2. Identify the **spread** of the data.
3. Determine if there are any **gaps** in the data.

Goals for Visualizations

1. Identify what is **typical** in the data.
2. Identify the **spread** of the data.
3. Determine if there are any **gaps** in the data.
4. Identify the shape of the **distribution** of the data.

Goals for Visualizations

1. Identify what is **typical** in the data.
2. Identify the **spread** of the data.
3. Determine if there are any **gaps** in the data.
4. Identify the shape of the **distribution** of the data.
5. Identify **peaks** in the data.

Goals for Visualizations

1. Identify what is **typical** in the data.
2. Identify the **spread** of the data.
3. Determine if there are any **gaps** in the data.
4. Identify the shape of the **distribution** of the data.
5. Identify **peaks** in the data.
6. Determine whether there are any **outliers**.

Stem-and-Leaf Plots

2, 17, 14, 35, 37, 44, 2, 47, 41, 35, 50, 40, 16, 20, 22, 22, 23, 9, 23, 23, 48, 34, 8, 26, 11, 4, 49, 11, 39, 34, 29, 4, 4, 35, 6, 3, 5, 40, 28, 40, 45, 27, 5, 37, 1, 27, 16, 20, 19, 50, 10, 50, 19, 20, 28, 45, 40, 4, 32, 25

2, 17, 14, 35, 37, 44, 2, 47, 41, 35, 50, 40, 16, 20, 22, 22, 23, 9, 23, 23, 48, 34, 8, 26, 11, 4, 49, 11, 39, 34, 29, 4, 4, 35, 6, 3, 5, 40, 28, 40, 45, 27, 5, 37, 1, 27, 16, 20, 19, 50, 10, 50, 19, 20, 28, 45, 40, 4, 32, 25

```
The decimal point is 1 digit(s) to the right of the |

0 | 12234444
0 | 55689
1 | 0114
1 | 66799
2 | 00022333
2 | 5677889
3 | 244
3 | 555779
4 | 000014
4 | 55789
5 | 000
```

Constructing Stem-and-Leaf Plots

1. Identify the stem (leading digit) and leaves (remaining digits)

Constructing Stem-and-Leaf Plots

1. Identify the stem (leading digit) and leaves (remaining digits)

2. Place the stems from smallest to largest in a vertical column

Constructing Stem-and-Leaf Plots

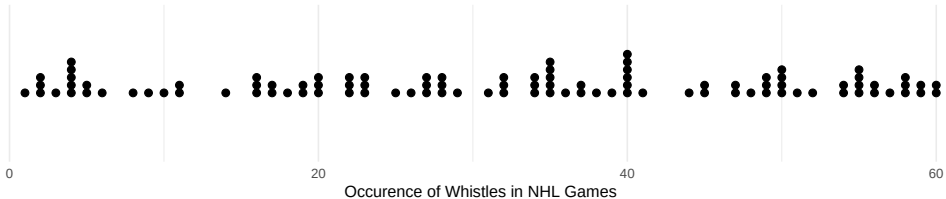
1. Identify the stem (leading digit) and leaves (remaining digits)
2. Place the stems from smallest to largest in a vertical column
3. Place each leaf with the corresponding stem in ascending order

Constructing Stem-and-Leaf Plots

1. Identify the stem (leading digit) and leaves (remaining digits)
2. Place the stems from smallest to largest in a vertical column
3. Place each leaf with the corresponding stem in ascending order
4. Indicate the units (where is the decimal point?)

Dot Plots

2, 51, 17, 56, 14, 35, 37, 44, 2, 47, 41, 35, 50, 40, 16, 20, 22, 22, 23, 9, 23, 23, 48, 34, 8, 26, 56, 11, 4, 58, 58, 49, 11, 60, 39, 34, 29, 4, 4, 55, 35, 52, 6, 3, 5, 40, 28, 40, 45, 59, 58, 27, 5, 37, 1, 27, 55, 16, 20, 19, 50, 10, 50, 19, 20, 28, 55, 54, 45, 40, 57, 54, 4, 32, 59, 25, 35, 60, 27, 28, 4, 32, 50, 55, 47, 31, 49, 2, 22, 32, 38, 18, 17, 40, 35, 34, 16, 49, 40, 36



Constructing Dot Plots

1. Plot all observations on the vertical axis at their value.

Constructing Dot Plots

1. Plot all observations on the vertical axis at their value.

2. Stack dots for repeated observations.

Histograms

- Histograms represent the **frequency distribution** using bar charts.

Histograms

- Histograms represent the **frequency distribution** using bar charts.
- For discrete variables, we use each category, and count the observations in each category.

Histograms

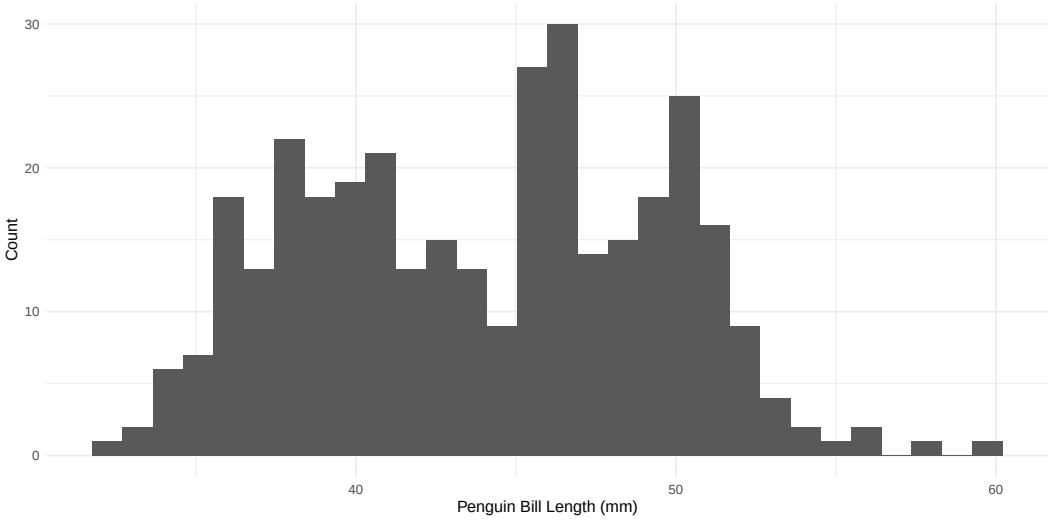
- Histograms represent the **frequency distribution** using bar charts.
- For discrete variables, we use each category, and count the observations in each category.
- For continuous variables, we **bin** the observations, and count the observations in each bin.

Histograms

- Histograms represent the **frequency distribution** using bar charts.
- For discrete variables, we use each category, and count the observations in each category.
- For continuous variables, we **bin** the observations, and count the observations in each bin.
- Can also use the **relative frequency**, which is given by

$$\text{relative frequency} = \frac{\text{frequency}}{\text{total number of observations}}.$$

Histograms



Considerations for Histograms

- Classes should be the same width.

Considerations for Histograms

- Classes should be the same width.
- Classes should not overlap.

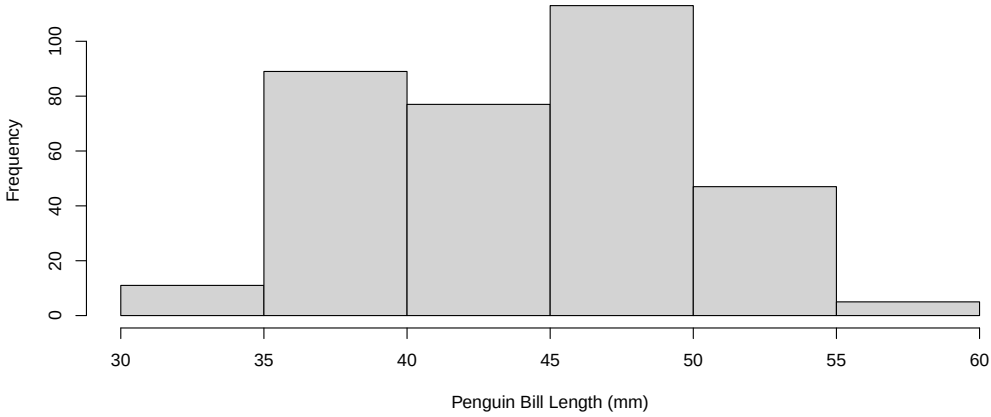
Considerations for Histograms

- Classes should be the same width.
- Classes should not overlap.
- Classes should include all possible values.

Histograms

	Bins	Counts	Relative Frequency
1	30 - 35	11	0.03216374
2	35 - 40	89	0.26023392
3	40 - 45	77	0.22514620
4	45 - 50	113	0.33040936
5	50 - 55	47	0.13742690
6	55 - 60	5	0.01461988

Histograms


Frequency
0
20
40
60
80
100
30
35
40
45
50
55
60
Penguin Bill Length (mm)

What to Look for in a Histogram

- Shape, including **modality**, **symmetry**, and **skewness**

What to Look for in a Histogram

- Shape, including **modality**, **symmetry**, and **skewness**
  - How many peaks (modes) are there: one is **unimodal**, two is **bimodal**, three or more is **multimodal**

What to Look for in a Histogram

- Shape, including **modality**, **symmetry**, and **skewness**
  - How many peaks (modes) are there: one is **unimodal**, two is **bimodal**, three or more is **multimodal**
  - Does the distribution look like a mirror image? If not, it is skewed.

What to Look for in a Histogram

- Shape, including **modality**, **symmetry**, and **skewness**
  - How many peaks (modes) are there: one is **unimodal**, two is **bimodal**, three or more is **multimodal**
  - Does the distribution look like a mirror image? If not, it is skewed.
  - Left (or negative) skew has a tail pointing left; right (or positive) skew has a tail pointing right.

What to Look for in a Histogram

- Shape, including **modality**, **symmetry**, and **skewness**
  - How many peaks (modes) are there: one is **unimodal**, two is **bimodal**, three or more is **multimodal**
  - Does the distribution look like a mirror image? If not, it is skewed.
  - Left (or negative) skew has a tail pointing left; right (or positive) skew has a tail pointing right.
- The **centre** of the distribution

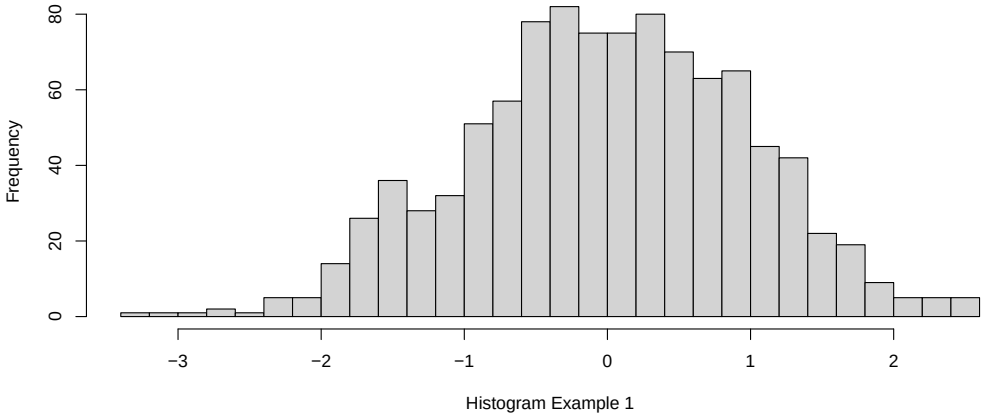
What to Look for in a Histogram

- Shape, including **modality**, **symmetry**, and **skewness**
  - How many peaks (modes) are there: one is **unimodal**, two is **bimodal**, three or more is **multimodal**
  - Does the distribution look like a mirror image? If not, it is skewed.
  - Left (or negative) skew has a tail pointing left; right (or positive) skew has a tail pointing right.
- The **centre** of the distribution
- The **spread** of the distribution

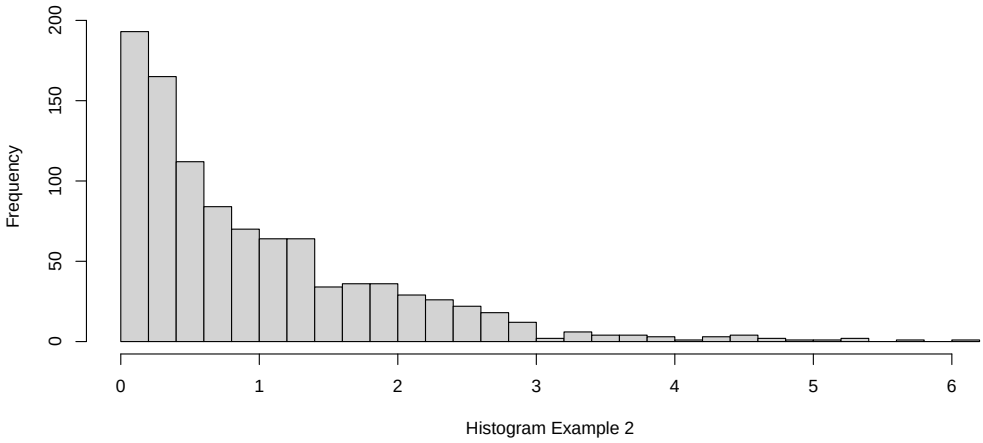
What to Look for in a Histogram

- Shape, including **modality**, **symmetry**, and **skewness**
  - How many peaks (modes) are there: one is **unimodal**, two is **bimodal**, three or more is **multimodal**
  - Does the distribution look like a mirror image? If not, it is skewed.
  - Left (or negative) skew has a tail pointing left; right (or positive) skew has a tail pointing right.
- The **centre** of the distribution
- The **spread** of the distribution
- **Outliers** or deviations from the general pattern

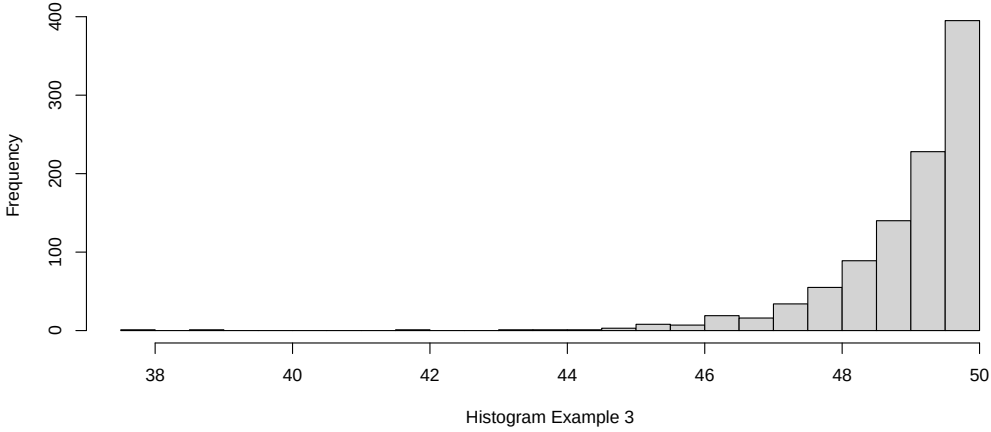
Examples



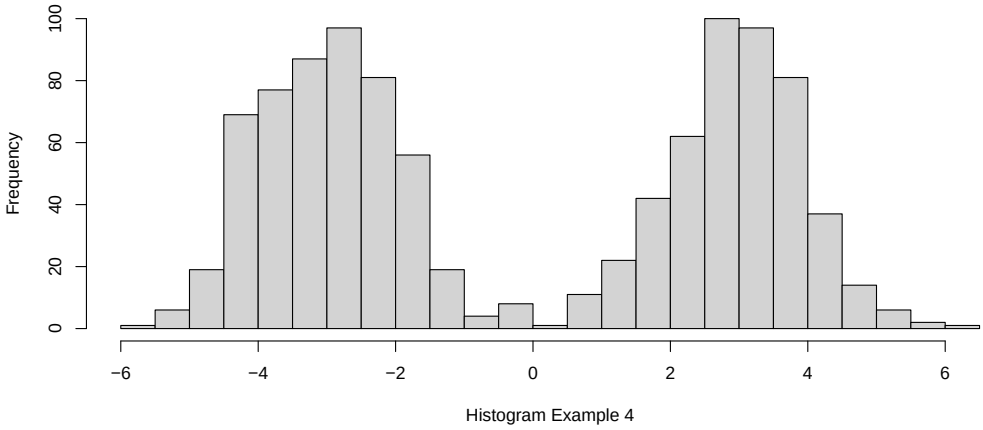
Examples



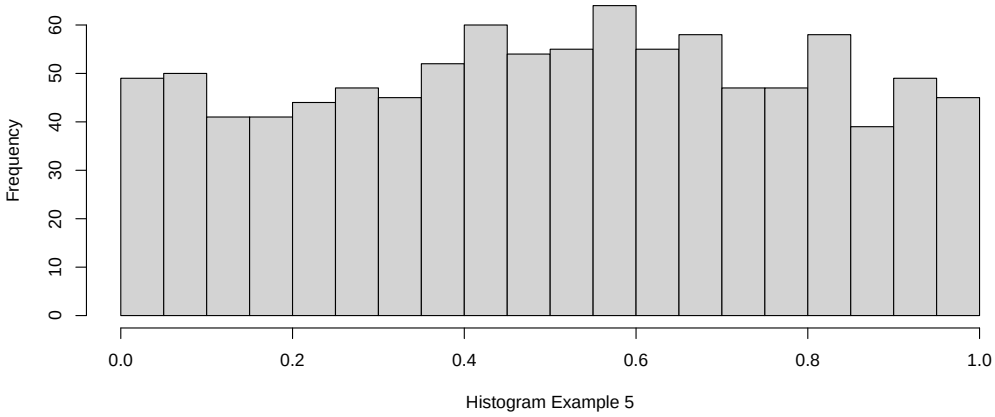
Examples


Frequency
0
100
200
300
400
38
40
42
44
46
48
50
Histogram Example 3

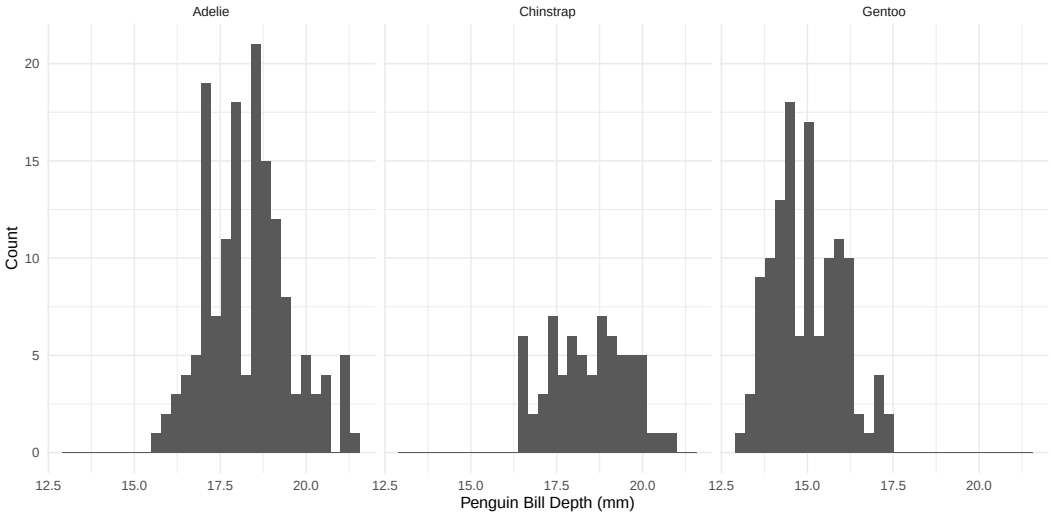
Examples



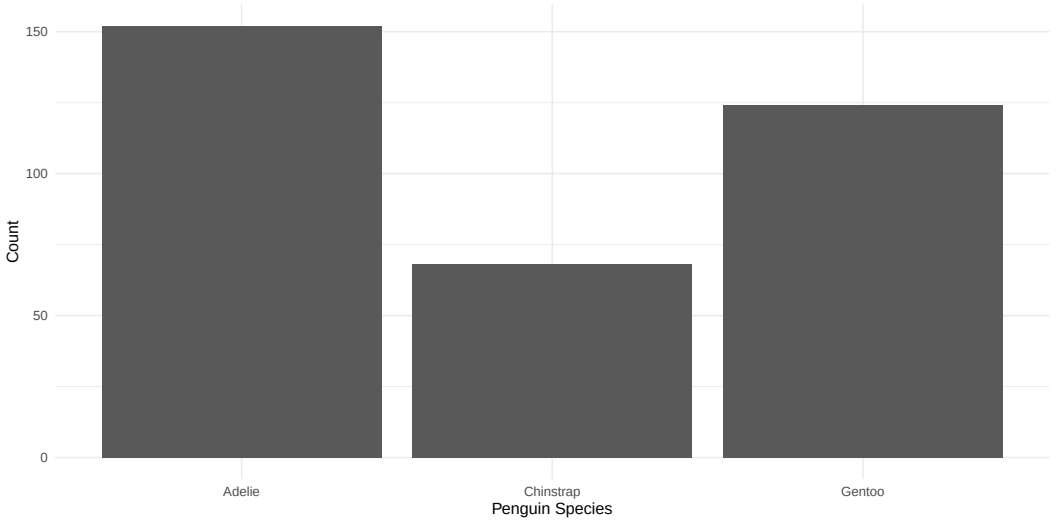
Examples


Frequency
0
10
20
30
40
50
60
0.0
0.2
0.4
0.6
0.8
1.0
Histogram Example 5

Histograms: Comparing Distributions



Bar Charts: Histograms for Categorical Data



Summary

- Raw data are difficult to interpret on their own.
- Visualizations can help display the character of a distribution.
- Stem-and-leaf plots, dot plots, and histograms are all useful for quantitative variables.
- Histograms (as bar plots) can be used for categorical variables.