Week Two: Descriptive Statistics

CS 217

Course Objectives

By the end of the course, students should be proficient at:

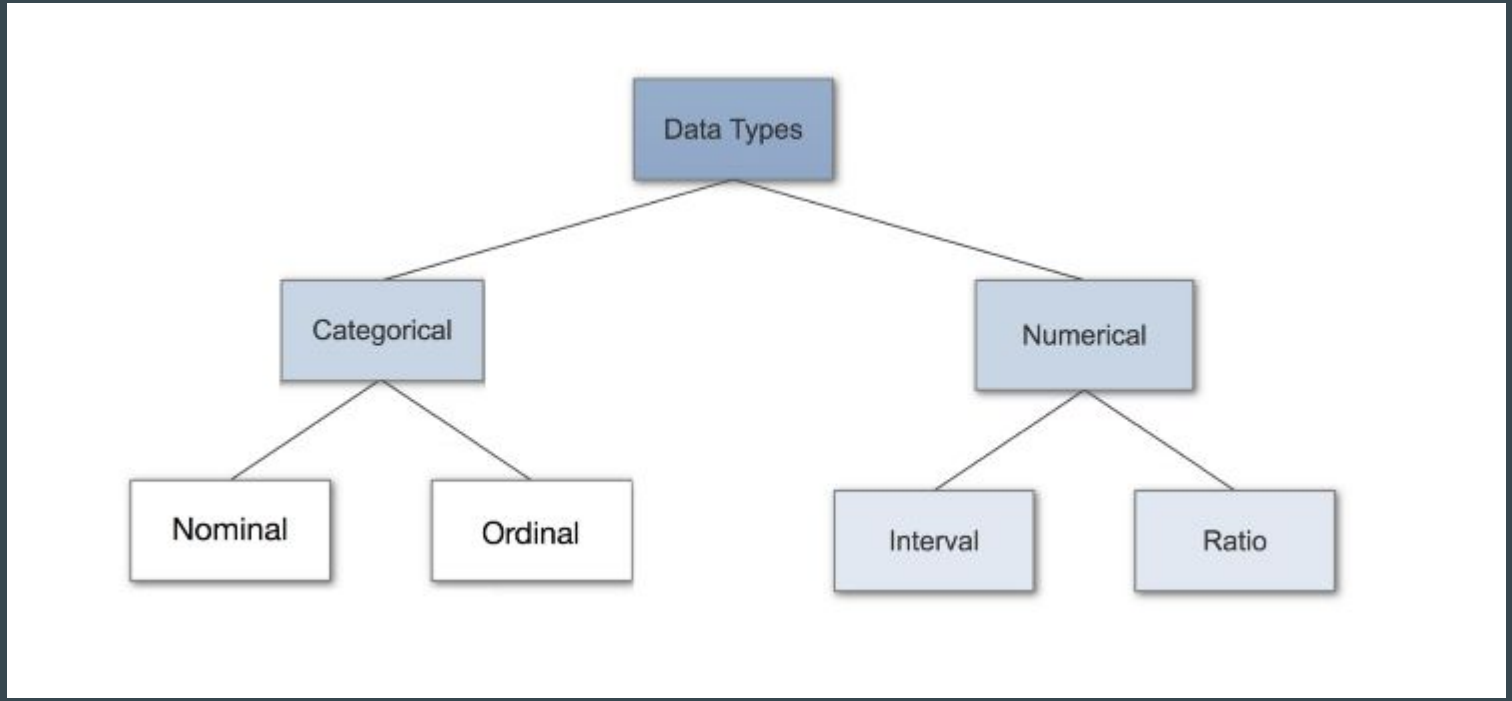
1. **Single Variable Explorations**: Examine a single variable, understand its underlying distribution, and choose the appropriate summary statistics for it.
2. **Pair-Wise Exploration**: Identify possible relationships between variables and compute correlations and linear fits.
3. **Estimation and Hypothesis Testing**: Understand the following three questions when reporting statistical results: 1) How big is the effect? 2) How much variability should we expect if we run the same measurement again? 3) Is it possible that the apparent effect is due to chance?
4. **Visualization**: Use data visualization as a tool for examining data and communicating results

Course Objectives

By the end of the course, students should be proficient at:

1. **Single Variable Explorations**: Examine a single variable, understand its underlying distribution, and choose the appropriate summary statistics for it.
2. **Pair-Wise Exploration**: Identify possible relationships between variables and compute correlations and linear fits.
3. **Estimation and Hypothesis Testing**: Understand the following three questions when reporting statistical results: 1) How big is the effect? 2) How much variability should we expect if we run the same measurement again? 3) Is it possible that the apparent effect is due to chance?
4. **Visualization**: Use data visualization as a tool for examining data and communicating results

Types of Data



Categorical Data

- Categorical data is **qualitative**
- Nominal Data:
  - Discrete units with no order
  - **What's an example of nominal data?**
- Ordinal Data:
  - Discrete units with order
  - Distance between categories not clear
  - **What's an example of ordinal data?**

Categorical Data

- Categorical data is **qualitative**
- Nominal Data:
  - Discrete units with no order
  - **What's an example of nominal data?**
- Ordinal Data:
  - Discrete units with order
  - Distance between categories not clear
  - **Example**: What is your current class standing?

Freshman	Sophomore	Junior	Senior
1	2	3	4

Numerical Data

- Interval Data
  - Ordered units with the same distance
  - No absolute zero
  - **Example**: Temperature, which can be negative
- Ratio Data
  - Ordered units with the same distance
  - Absolute zero
  - **Example**: Height, which can’t be negative

Descriptive Statistics

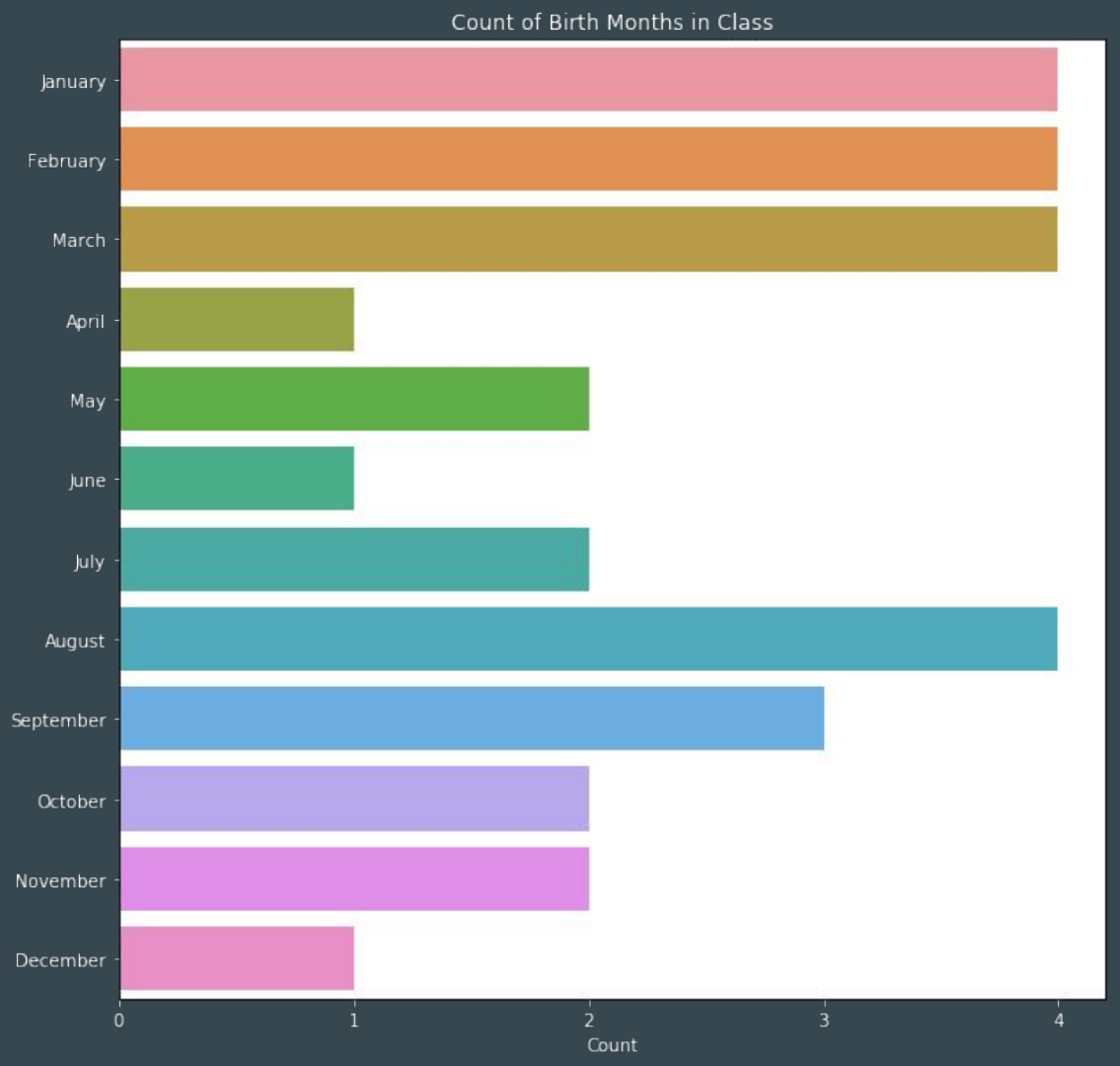
- **Descriptive Statistics** involve **describing the data** without drawing conclusions
- Descriptive Statistics **do not** allow us to make conclusions about the data beyond any hypotheses that we may have
- What information about our data do we even want to know?

Descriptive Statistics

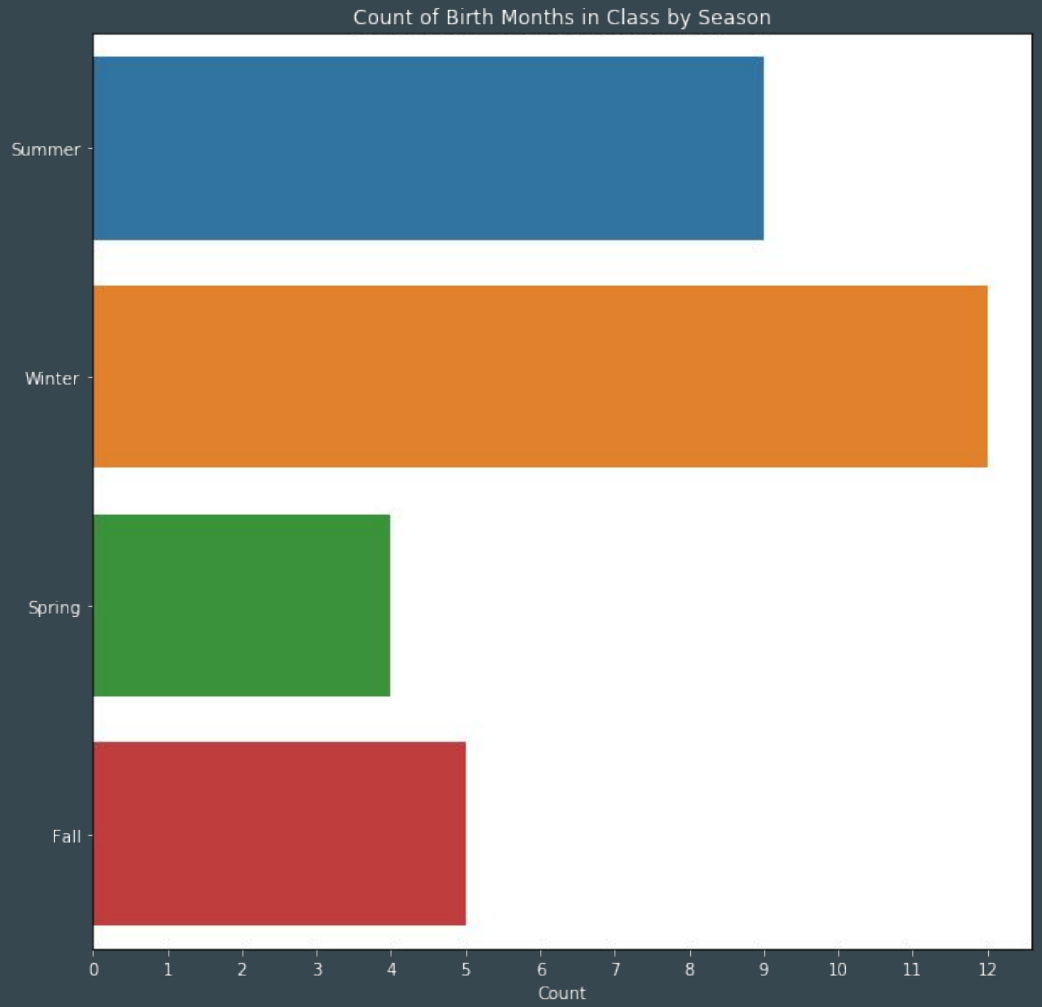
- Say I collect the birthdate for every student in the class. Below is a sample of what my survey dataset looks like.
- How could I visualize this data?

Name	Month	Day	Year
Bob	June	1	2000
Linda	March	17	1997
Tina	September	30	1999

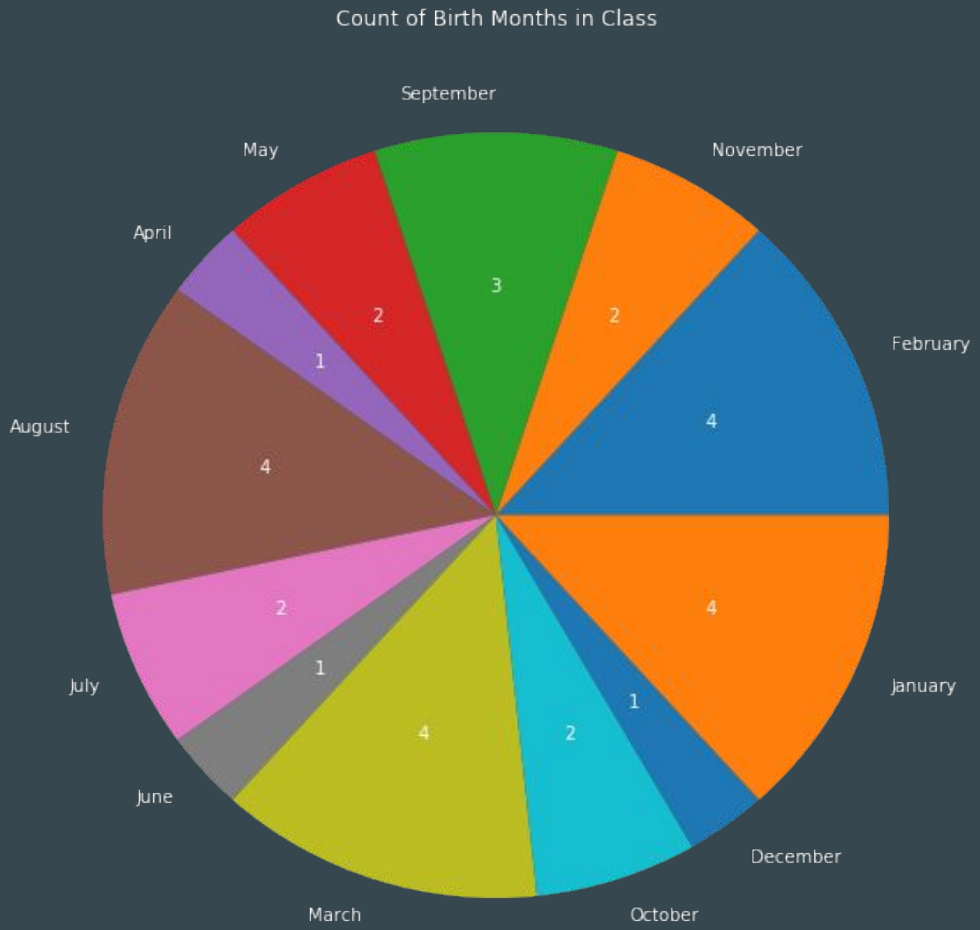
- A **bar graph** is a good way to visually explore categorical data.
- Here we can see the count of recorded birth months for the students in our (imaginary) dataset.
- The **mode** is the **most frequent value** in a distribution and is a useful statistic to describe categorical data.
- What is the mode here? (hint: there can be multiple nodes)



- We could also **group** our categorical data and examine it that way.
- For example, we could group student birth months by the season they were born in and look at the **bar graph** of the results.
- What is the mode here? (hint: there can be multiple nodes)



- Pie charts are another common way to display discrete data, though bar graphs are preferred by data scientists
- Why may this be true?
- Note that either a pie chart or a histogram can show either the **count** or **frequency** of a given discrete variable or the **percent of frequency** for a given discrete variable relative to the entire distribution.

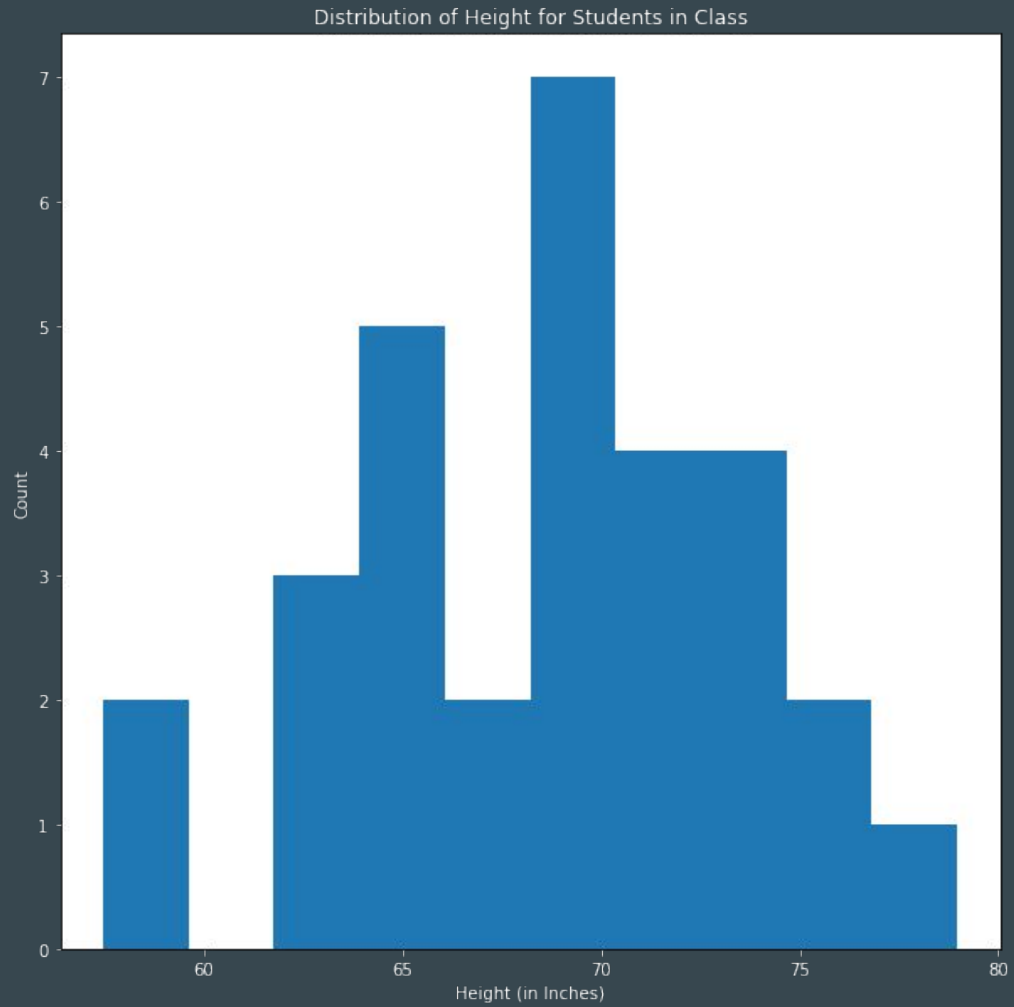


Descriptive Statistics

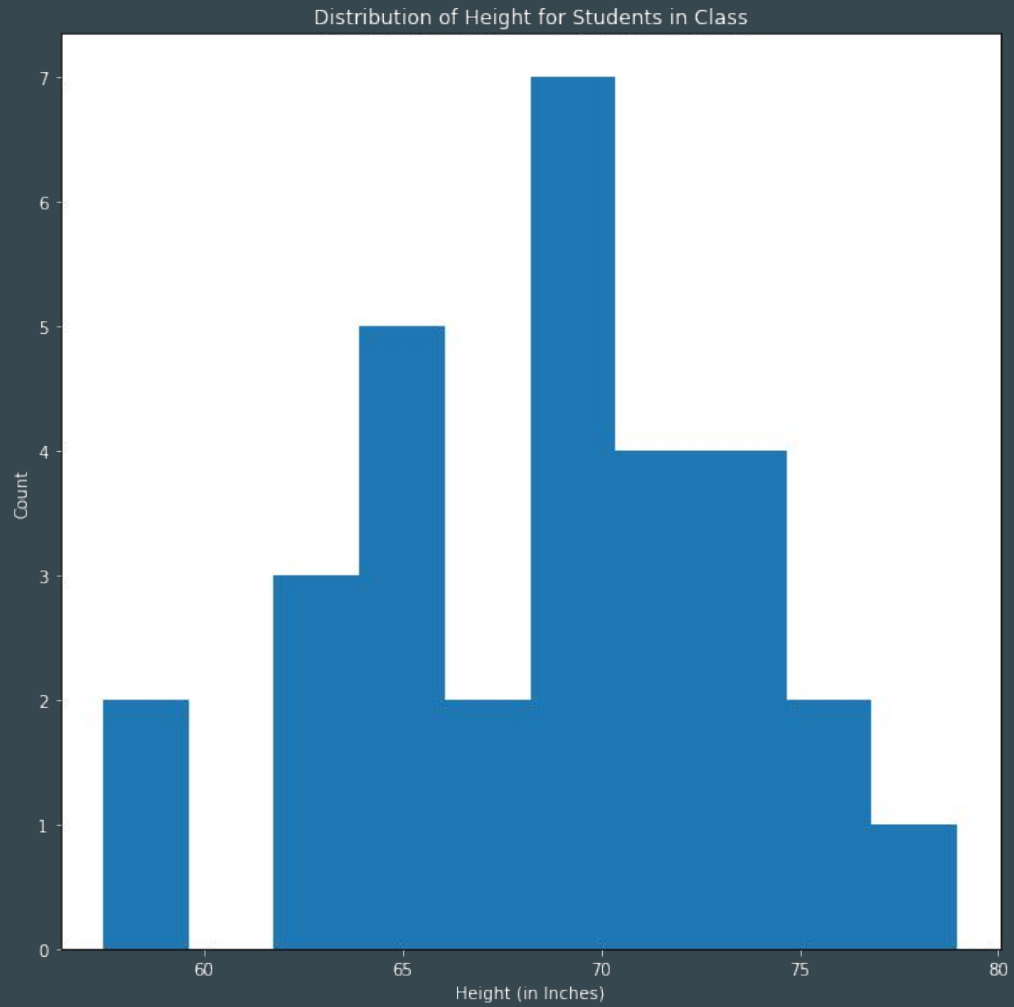
- Say I collect the height for every student in the class. Below is a sample of what my survey dataset looks like.
- How could I visualize this data? What's different about this data than our previous dataset?

Name	Height
Gene	5'2"
Louise	4'1"
Teddy	5'11"

- A **histogram** is a good way to visually explore continuous data.
- While it is visually similar to a bar graph, note that it is different in that we are grouping, or 'binning' categorical data that's close to one another.
- While we can show the count for every discrete value, it is much less useful for a distribution of continuous variables.



- Because discrete variables have no inherent order, we simply want to see the count for each of them.
- With a histogram of a continuous distribution, we want to use it as a tool to understand the overall distribution. Where is the mean? Is there a big range? Does it look like there are outliers?



Descriptive Statistics

- Say I collect the height for every student in the class. What statistics (data points) could I draw from having this collective data?

Descriptive Statistics

- Say I collect the height for every student in the class. What statistics (data points) could I draw from having this collective data?
  - **Mean** - average height
  - **Median** - middle height when put in order
  - **Mode** - most common height
  - **Maximum/Minimum -** tallest/shortest height
  - **Range** - difference between shortest and tallest height
  - **Standard Deviation/Variance** - measure of how spread out height is
  - **Outliers -** is anyone particularly tall or short?

Questions for Continuous Data

- Central Tendency
  - Do the values tend to cluster around a particular point?
- Modes
  - Is there more than one cluster?
- Spread
  - How much variability is there in the values?
- Outliers:
  - Are there extreme values far from the modes?

Central Tendency

- Mean
  - The mean tells us the average of a distribution
  - It is most useful when our data does not have outliers or a big range

$$\bar{x} = \frac{1}{n} \sum_i x_i$$

Central Tendency

- Median
  - The median also tells us the central tendency of a distribution
  - It is middle score for a set of data arranged in order of value, or the value in the position of (n + 1) / 2
    - If the count of the distribution is odd, the ‘middle’ value is used, i.e. for a distribution with 11 values, the 6th value ((11 + 1) / 2) is the median
    - If the count of the distribution is even, the two ‘middle’ values are used, i.e. for a distribution with 10 values, the mean of the 5th and 6th values ((10 + 1) / 2) is the median
  - It is more useful for skewed data than the mean

14	35	45	55	55	**56**	56	65	87	89	92

14	35	45	55	**55**	**56**	56	65	87	89

Central Tendency

- Say four people are in a bar. Their net worths are $40,000, $50,000, $60,000 and $70,000.
- What is the mean of their net worths?
- What is the median?

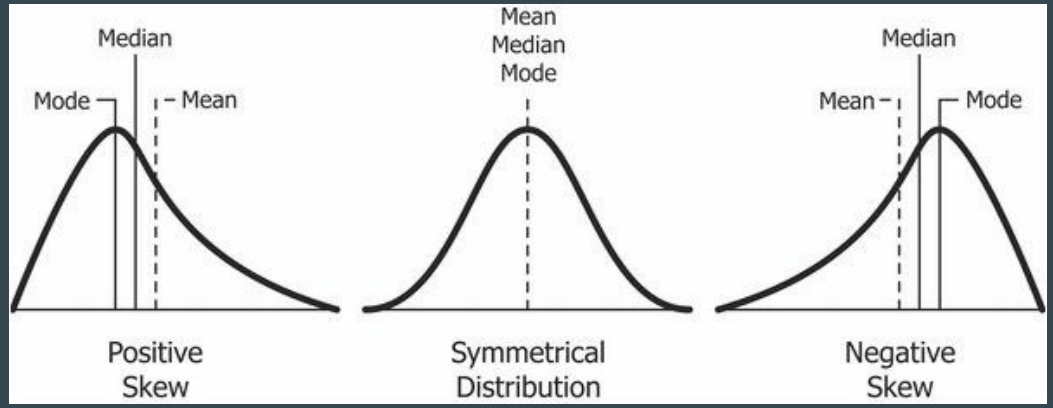
Central Tendency

- Say four people are in a bar. Their net worths are $40,000, $50,000, $60,000 and $70,000.
- Jeff Bezos walks into a bar. His net worth is $140,000,000,000.
- What is the mean of their net worths?
- What is the median?

Central Tendency

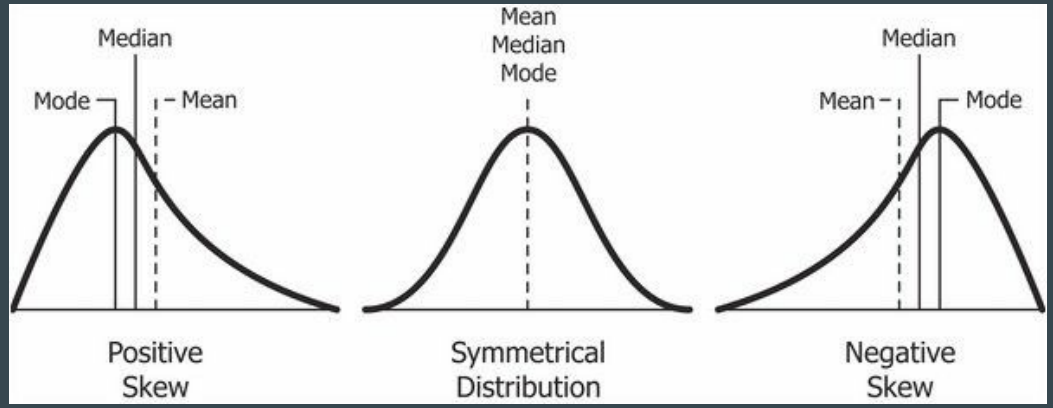
Skew

- A **normal distribution** is unskewed, or is symmetrically distributed on either side of the mean.
- Here, the mean value should be close to the median, and either can be used as a measure of the center of the distribution.
- Hence why the mean is a good measure of central value for an unskewed distribution



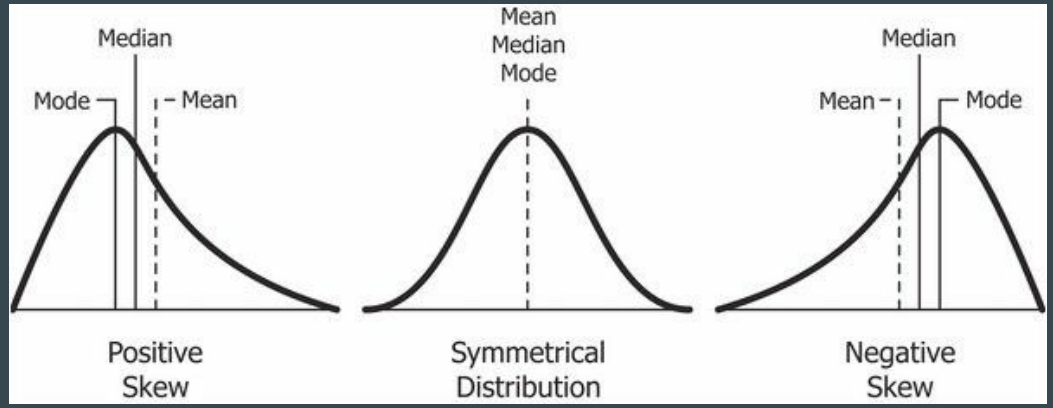
Skew

- A distribution has a **negative skew** if there are **values** on the lower end of the distribution that are significantly less than anything else in the distribution.
- Here, the mean will be lower than the median of the distribution, though the majority of the values in the distribution will be greater than the mean
- Hence why the **median is a better measure of central tendency** here.



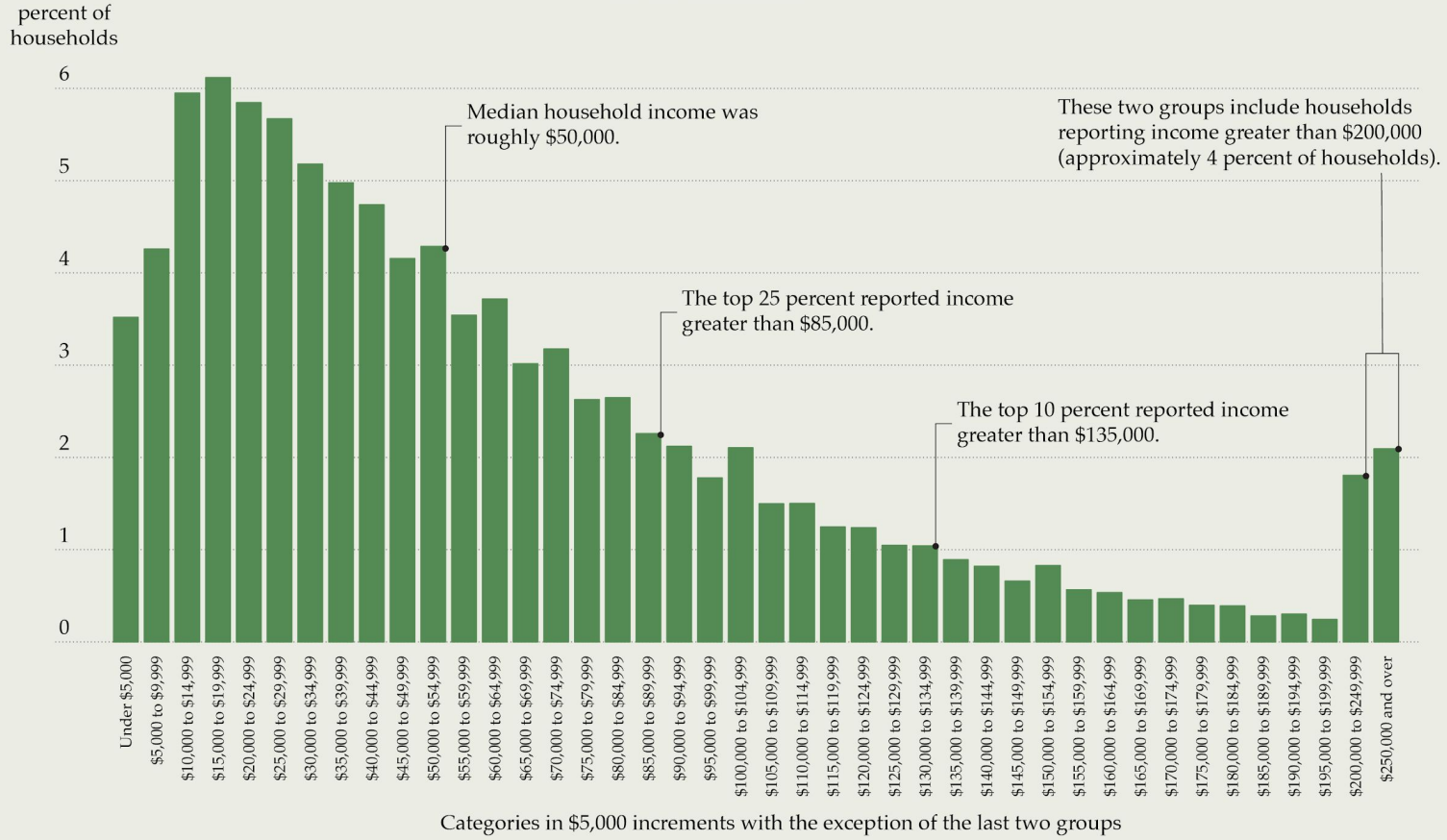
Skew

- A distribution has a **positive skew** if there are **values** on the upper end of the distribution that are significantly greater than anything else in the distribution.
- Here, the mean will be greater than the median of the distribution, though the majority of the values in the distribution will be less than the mean
- Hence why the **median is a better measure of central tendency** here.



Distribution of annual household income in the United States

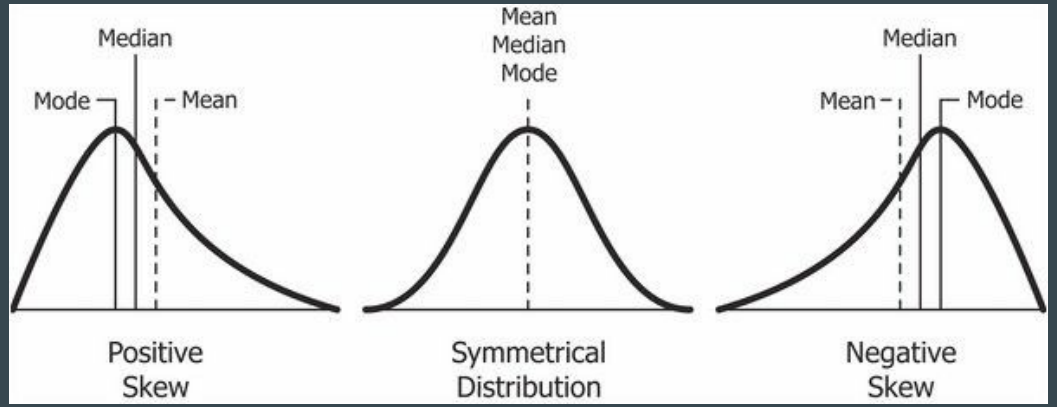
2010 estimate



Source: U.S. Census Bureau, Current Population Survey, 2011 Annual Social and Economic Supplement

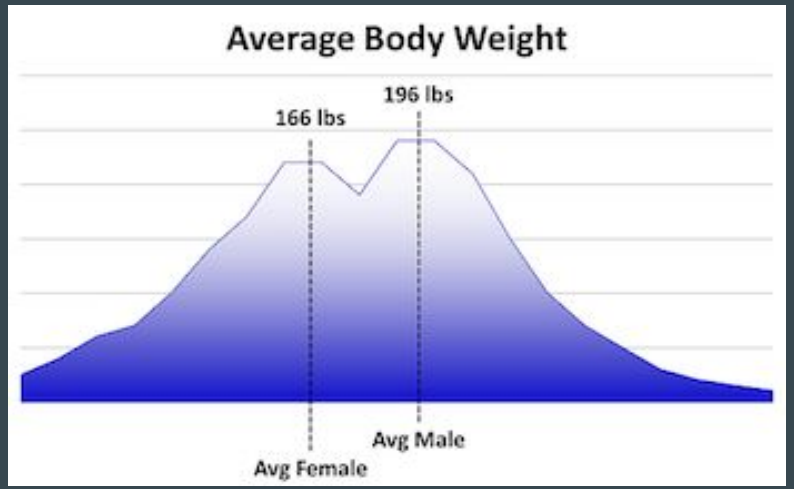
Central Tendency

- Much like in a discrete distribution, the mode in a continuous distribution is the most common value(s)
- It is less imåportant to calculate this for a distribution than note it visually: the mode is the value that corresponds to the top of the ‘peak’ of the histogram.



Central Tendency

- Much like in a discrete distribution, the mode in a continuous distribution is the most common value(s)
- It is less important to calculate this for a distribution than note it visually: the mode is the value that corresponds to the top of the ‘peak’ of the histogram.
- A **bimodal distribution** is one that has two peaks.



Spread

- Range
  - The difference between the **minimum** (smallest) and **maximum** (largest) values in a distribution
  - Maximum: 92
  - Minimum: 4
  - Range: 92-4 = **88**

4	35	45	55	55	56	56	65	87	89	**92**

Spread

- Quartiles
  - The value of each quarter of your distribution when arranged in order of value
  - Split the dataset by the **median** and find the median of each divided set
  - What is the median below?

4	35	45	55	55	56	56	65	87	89	92

Spread

- Quartiles
  - The value of each quarter of your distribution when arranged in order of value
  - Split the dataset by the **median (including the median)** and find the median of each divided set
  - What is the median below?
  - The median is the **6th** value of the dataset, here 56. It's also the **second quartile.**

4	35	**45**	55	55	**56**	56	65	**87**	89	92

Spread

- Quartiles
  - The value of each quarter of your distribution when arranged in order of value
  - Split the dataset by the **median (including the median)** and find the median of each divided set
  - What is the median below?
  - The median is the **6th** value of the dataset, here 56. It's also the **second quartile.**
  - What is the median of the first dataset?
  - What is the median of the second dataset?

4	35	**45**	**55**	55	56

56	56	**65**	**87**	89	92

Spread

- Quartiles
  - The value of each quarter of your distribution when arranged in order of value
  - Split the dataset by the **median (including the median)** and find the median of each divided set
  - What is the median below?
  - The median is the **6th** value of the dataset, here 56. It's also the **second quartile.**
  - The first quartile is (45 + 55) / 2, or 50.
  - The third quartile is (65 + 87) / 2, or 76
  - The **interquartile range** is the difference between the first and third quartiles, or 76-50=26

4	35	**45**	**55**	55	56

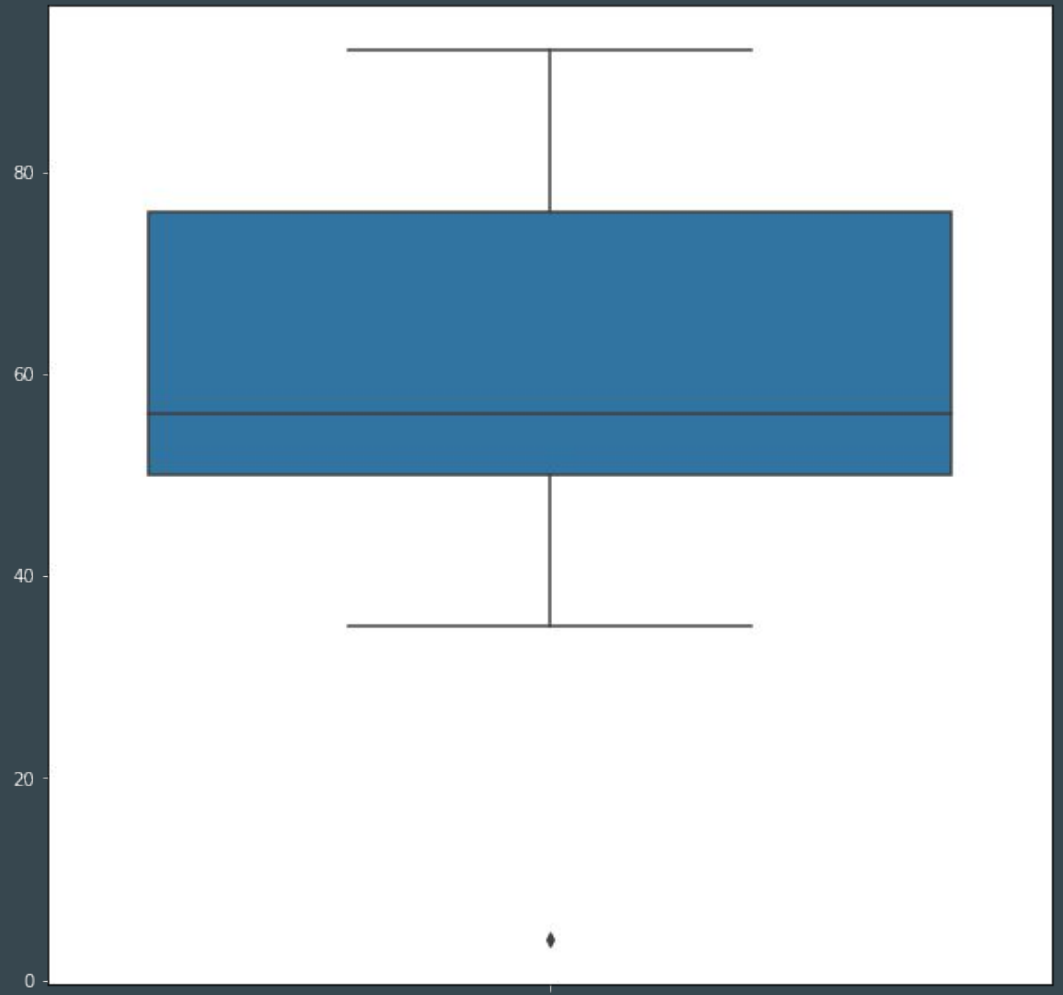
56	56	**65**	**87**	89	92

Spread

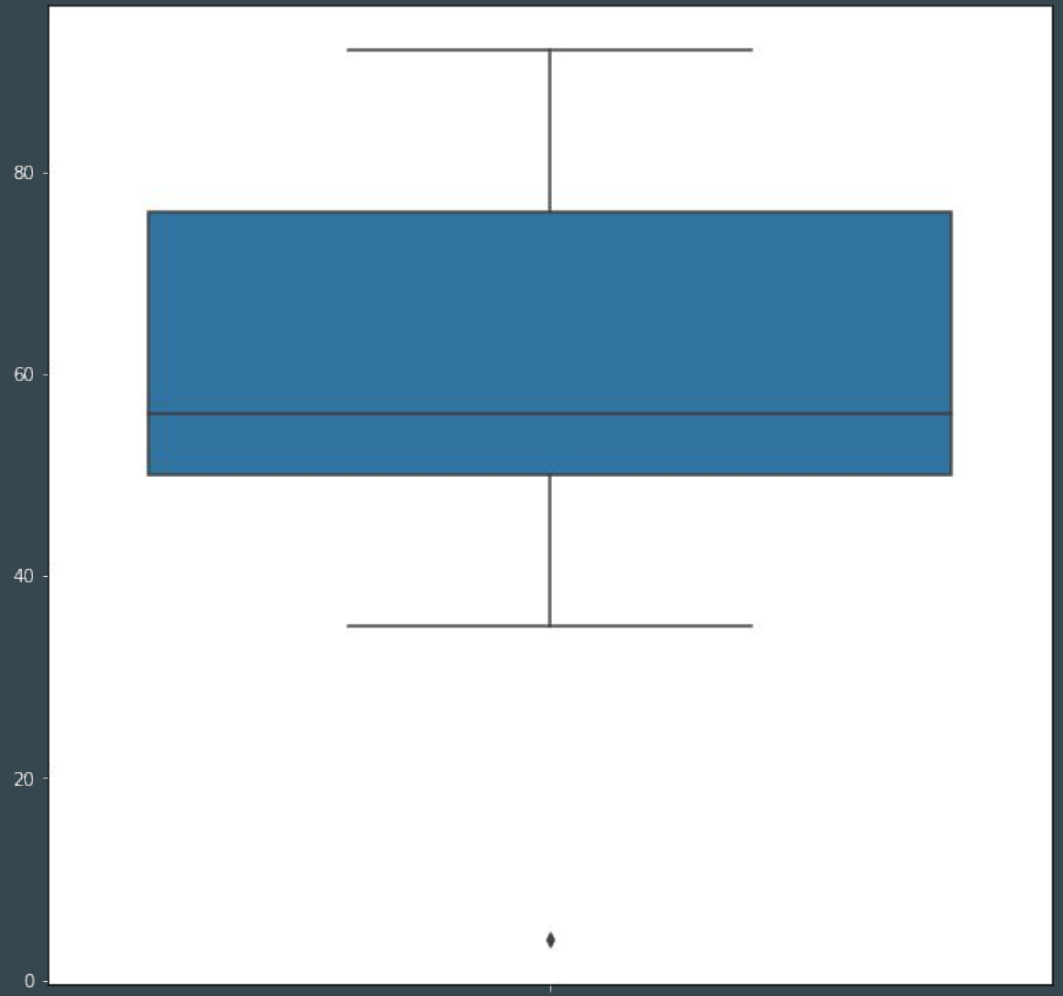
- The first quartile is equivalent to the **25th percentile**, as it's greater than 25% of your data and less than 75% of your data
- The second quartile is equivalent to the **50th percentile**, as it's greater than 50% of your data and less than 50% of your data.
- The third quartile is equivalent to the **75th percentile,** as it's greater than 75% of your data and less than 25% of your data
- You can calculate the percentile rank for any value in a dataset by dividing the rank order of that value by N + 1 values in a dataset.
- For example 2 / 12 is 0.1667, meaning that the second value of the dataset (35) is in the 17th percentile of the dataset.

4	35	45	55	55	56	56	65	87	89	92

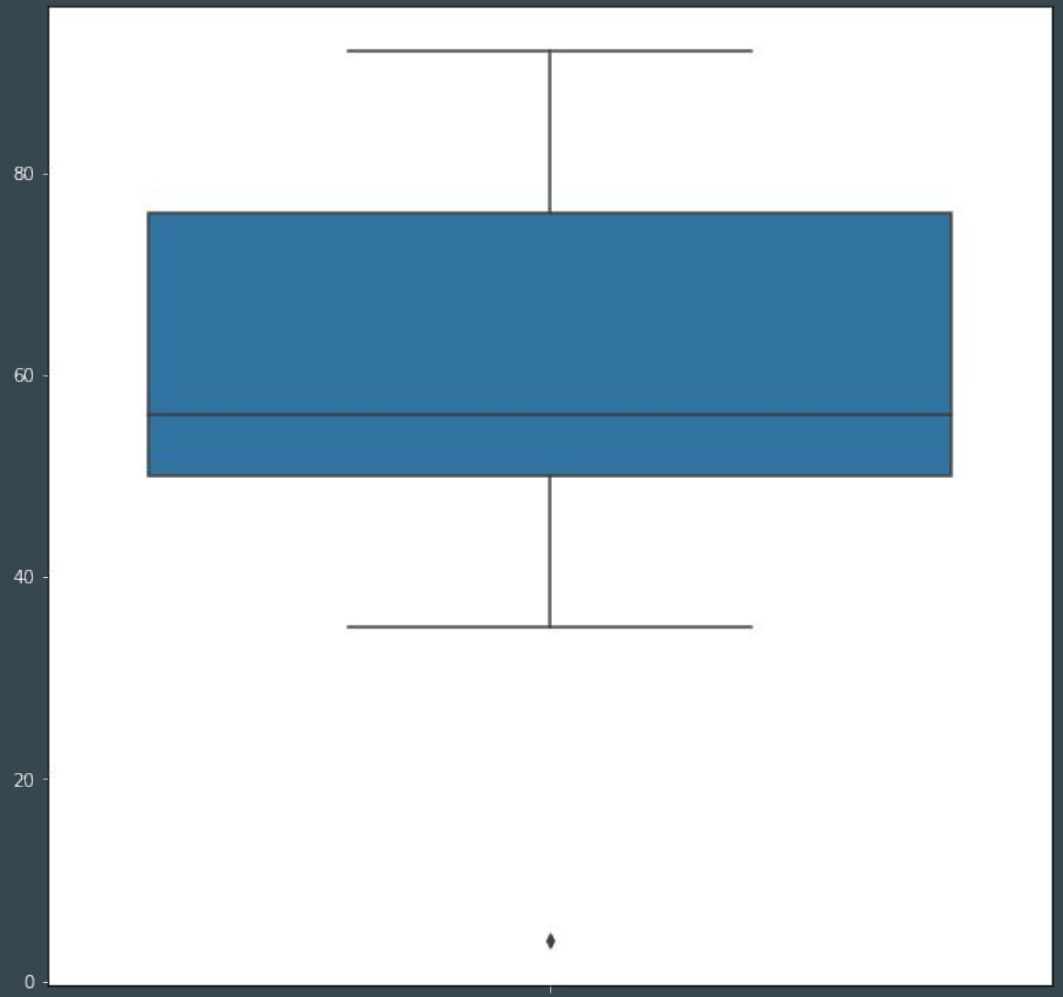
- A **boxplot** is a good way to see the quartiles in a distribution.
- Here, the line in the middle of the box is the median (56, as we saw earlier)
- The top of the box is the third quartile (76)
- The bottom of the box is the first quartile (50)



- The top whisker represents the highest datapoint less than the third quartile + 1.5 * the Interquartile Range
- Our interquartile range, as we found, is 26
- The third quartile is 76
- Thus the third quartile plus 26 * 1.5 is 76 + (26 * 1.5), or 115.
- The highest datapoint in this set less than 115 is the maximum value in the dataset, or 92
- If the dataset had a value greater than 115, it'd be an **outlier.**



- The bottom whisker represents the smallest datapoint greater than the first quartile - 1.5 * the Interquartile Range
- Our interquartile range, as we found, is 26
- The first quartile is 50
- Thus the first quartile minus 26 * 1.5 is 50 - (26 * 1.5), or 11.
- The smallest datapoint in this set greater than 11 is 35.
- Since the dataset has a value less than 35 (4, the smallest value in the dataset), it is an outlier and represented here as a dot.



Spread

- The absolute difference between a data point and the mean of a distribution is that point's **deviation**.
- The mean of our distribution is 58.1.
- The deviation of 45 is thus the absolute value of the difference between 58.1 - 45, or 13.1.

$$D_i = | x_i - m(X) |$$

Spread

- The absolute difference between a data point and the mean of a distribution is that point's **deviation**.
- The mean of our distribution is 58.1.
- The deviation of 45 is thus the absolute value of the difference between 58.1 - 45, or 13.1.

$$D_i = | x_i - m(X) |$$

Value	Mean	Deviation
4	58.1	54.1
35	58.1	23.1
45	58.1	13.1
55	58.1	3.1
55	58.1	3.1
56	58.1	2.1
56	58.1	2.1
65	58.1	6.9
87	58.1	28.9
89	58.1	30.9
92	58.1	33.9

Spread

- The **variance** is the mean of every **squared deviation** in a distribution
- Why the squared deviation rather than the absolute deviation?

$$\sigma^2 = \frac{\sum (X-\mu)^2}{N}$$

Spread

- The **variance** is the mean of every squared deviation in a distribution
- The variance to our dataset here is 604.26.

$$\sigma^2 = \frac{\sum (X-\mu)^2}{N}$$

Value	Mean	Deviation	Sq. Deviation
4	58.1	54.1	2926.21
35	58.1	23.1	533.61
45	58.1	13.1	171.61
55	58.1	3.1	9.61
55	58.1	3.1	9.61
56	58.1	2.1	4.41
56	58.1	2.1	4.41
65	58.1	6.9	47.61
87	58.1	28.9	835.21
89	58.1	30.9	954.81
92	58.1	33.9	1149.21

Spread

- The **standard deviation** is the square root of the variance and is a much more common measure of spread than the variance.
- A standard deviation that is equal to the mean or higher is considered “high” but this is an extremely loose measure
- It’s much more important to take the value in context, relative to contextual expectations or other, similar distributions.

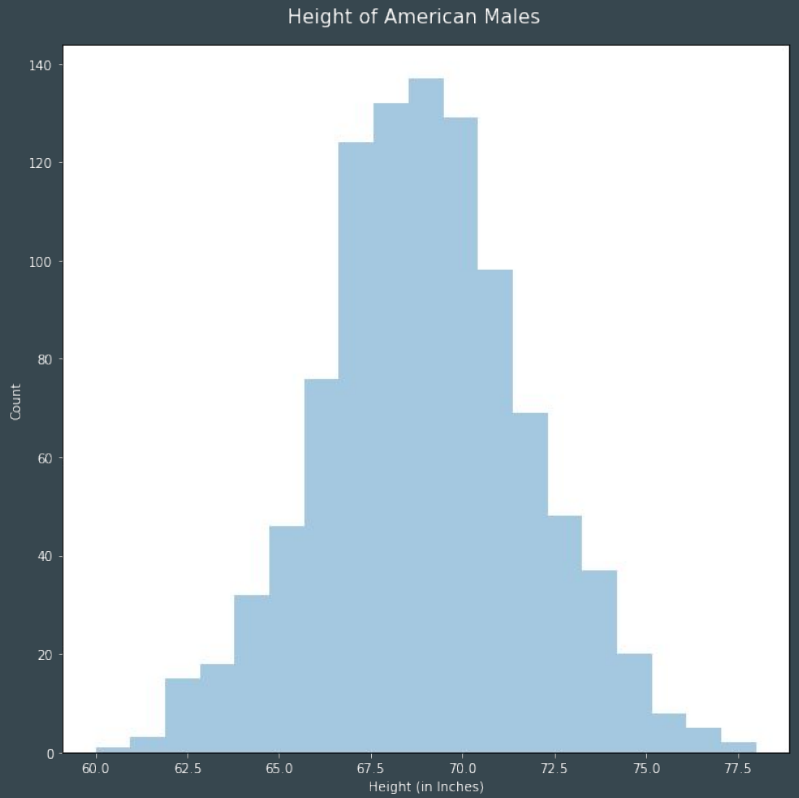
Spread

- The **standard deviation** is the square root of the variance.
- The standard deviation of our distribution here is the square root of 604.26, or 24.58.

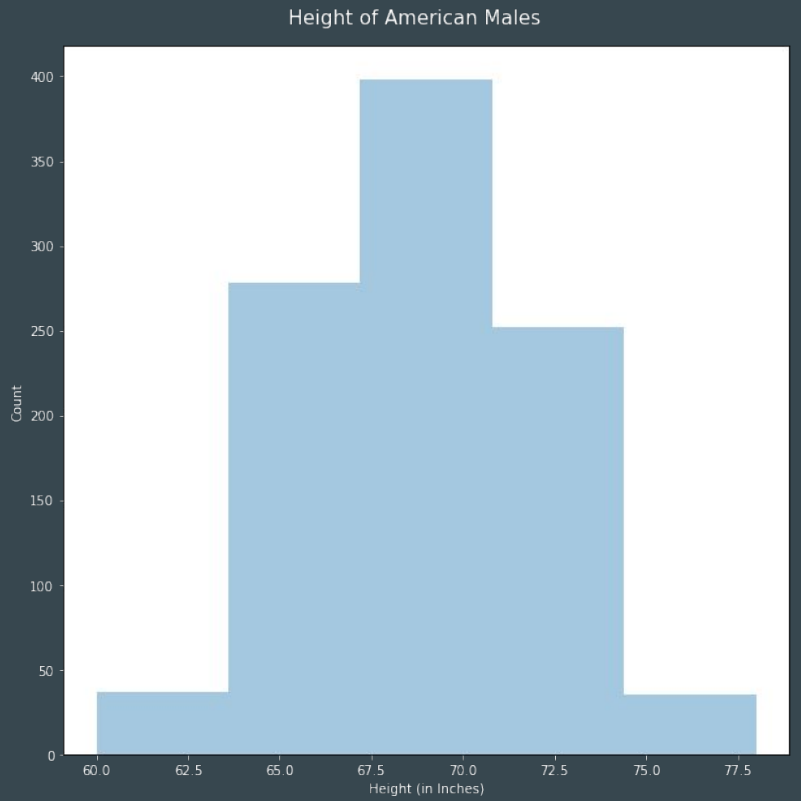
$$\sigma = \sqrt{\frac{\sum (X-\mu)^2}{N}}$$

Value	Mean	Deviation	Sq. Deviation
4	58.1	54.1	2926.21
35	58.1	23.1	533.61
45	58.1	13.1	171.61
55	58.1	3.1	9.61
55	58.1	3.1	9.61
56	58.1	2.1	4.41
56	58.1	2.1	4.41
65	58.1	6.9	47.61
87	58.1	28.9	835.21
89	58.1	30.9	954.81
92	58.1	33.9	1149.21

- A histogram plots the frequency of each value in a distribution
- It is by far the most common visual plotting tool for a distribution
- It is an easy way to see the metrics we talked about - where does the center seem to be? Are there outliers? Is the dataset skewed? Are there multiple peaks?
- One big limitation of the histogram is that the data is grouped into bins, which can affect the look of the graph
- For instance, the data currently has 20 bins.

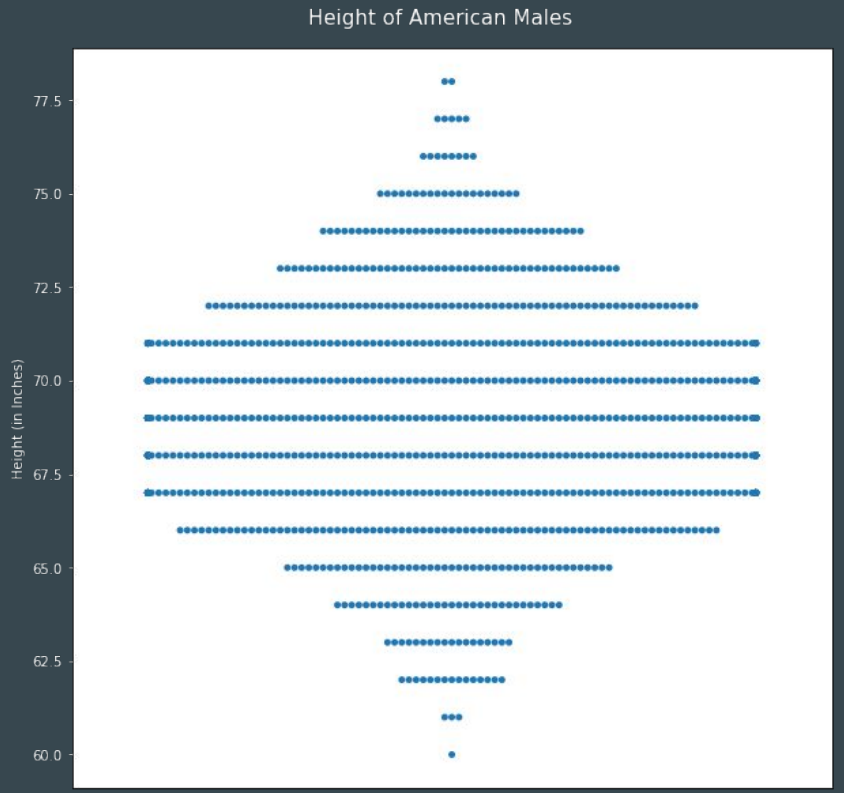


- A histogram plots the frequency of each value in a distribution
- It is by far the most common visual plotting tool for a distribution
- It is an easy way to see the metrics we talked about - where does the center seem to be? Are there outliers? Is the dataset skewed? Are there multiple peaks?
- One big limitation of the histogram is that the data is grouped into bins, which can affect the look of the graph
- ...And now it has 5.



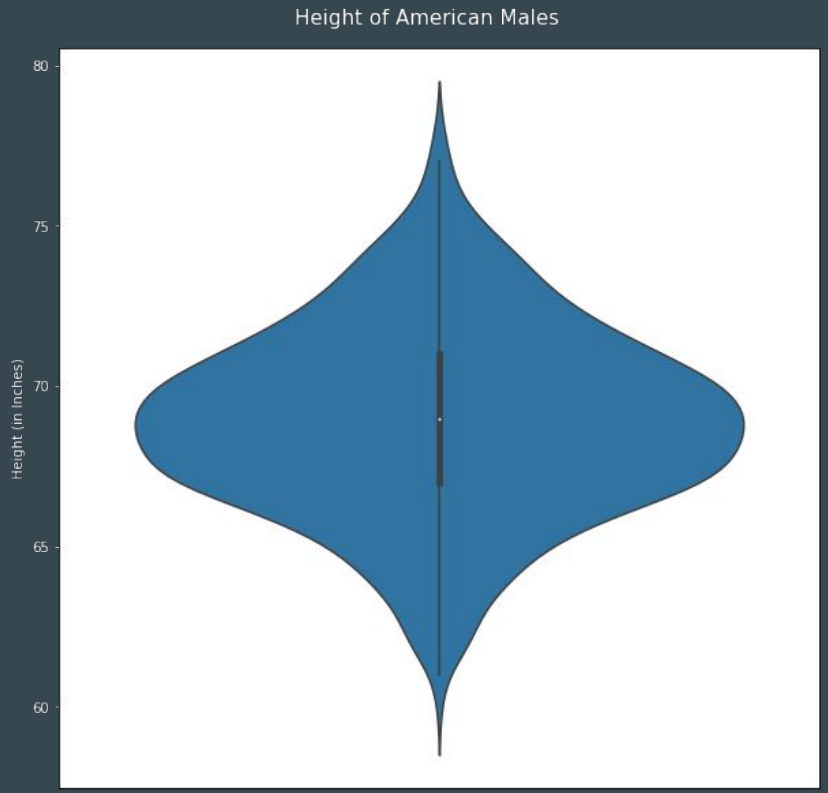
Visualizations

- A **swarm plot** is an alternative method of visualizing the data that doesn't have this limitation
- Swarm plots can be slow on larger datasets



Visualizations

- A violin plot achieves the same effect as a swarm plot and is computationally friendlier to larger datasets.



Visualizations

- A boxplot is a good way to see outliers in a distribution
- The box contains the Interquartile Range (25% - 75%) with a horizontal line representing the median
- The whiskers represent the first quartile minus 1.5 * IQR on the lower end, and the third quartile plus 1.5 * IQR on the other end
- Any value outside of the whiskers, such as the dots to the right, can be considered an outlier