Week Two: Descriptive Statistics

CS 217

Course Objectives

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

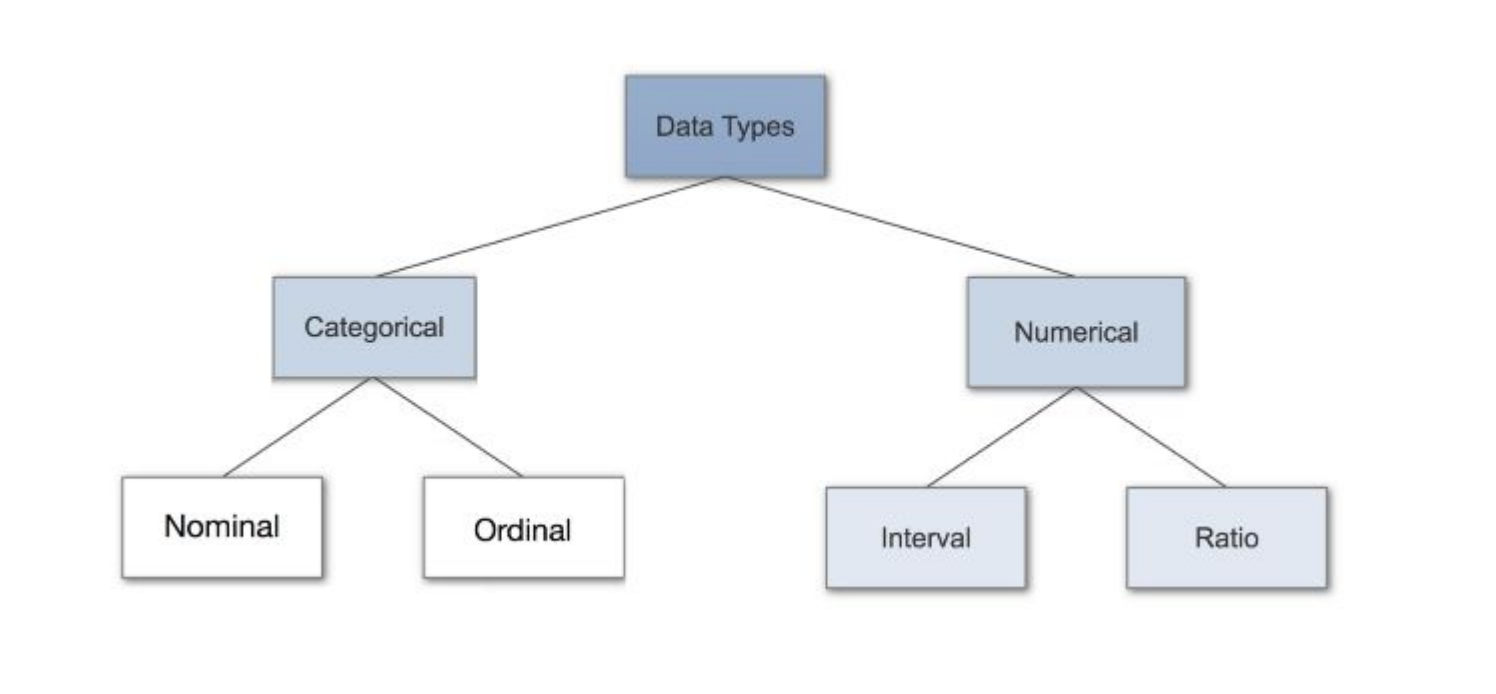
1. **Single Variable Explorations**: Examine a single variable, understand its underlying distribution, and choose the appropriate summary statistics for it.
2. **Pair-Wise Exploration**: Identify possible relationships between variables and compute correlations and linear fits.
3. **Estimation and Hypothesis 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
  - **Example**: What is your major?
- 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 birth month for every student in the class. What information could I draw from having this collective data?

Descriptive Statistics

- Say I collect the birth month for every student in the class. What information could I draw from having this collective data?
  - **Mode** - Which month is most common
  - **Grouped Mode** - Which season is the most common

Descriptive Statistics

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

Descriptive Statistics

- Say I collect the height for every student in the class. What information 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?

Central Tendency

- Mean
  - The mean tells us the average of a distribution
  - It's the sum of all values in a distribution divided by the number of values in a dataset
  - It is most useful when our data does not have outliers or a big range
  - If I tell you that the mean height of American men is 5'10", that is a good application of the mean
  - If I tell you that the mean income of an American family is $75,000, that is a less useful application of the mean

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

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
  - 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

- Mode
  - The most frequent score in a distribution
  - Most useful for nominal variables
  - Less useful for continuous variables

Pet	Frequency
Dog	**15**
Cat	**15**
Fish	5
Hamster	8
Gerbil	4
Rabbit	2

Central Tendency

- Measures of Central Tendency are obviously useful, but are only one of the things we are deducing from a distribution
- If a distribution is extremely skewed, a median is a more useful measure of its central tendency than the mean but a measure of the skew is a much more useful takeaway than just the median!
- Always think about datasets **holistically** and **with curiosity** rather than checking off boxes in your analysis

Spread

- Range
  - The difference between the **minimum** (smallest) and **maximum** (largest) scores in a distribution
  - The range below is: 92-14 = **78**

14	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
  - The second quartile is equivalent to the median
  - Below the first quartile is 45, the second quartile/median is 56, and the third quartile is 87
  - The **interquartile range** is the difference between the first quartile and the third quartile, which here is 42

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

Spread

- Percentiles
  - The first quartile is equivalent to the **25th percentile**, as it's higher than 25% of your data and lower than 75% of your data
  - The third quartile is equivalent to the **75**th percentile, as it's higher than 75% of your data and lower than 25% of your data
  - You can calculate the percentile rank for any value in a dataset.

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

Spread

- Deviation
  - The absolute difference between a data point and the mean of a distribution
  - For instance, if the mean of the below distribution is **59**, the deviation of 15 is the absolute value of 15 - 59, or 34

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

14	35	45	55	55	56	56	65	87	89	92

Spread

- Variance
  - The mean of every squared deviation in my distribution
  - The population variance has N - the number of data points in a dataset - as the denominator
  - The sample variance has N -1 as the denominator (we will cover the difference between these two later in the course)

$$\sigma^2 = \frac{\sum (X-\mu)^2}{N} \qquad S^2 = \frac{\sum (X-\mu)^2}{N-1}$$

Spread

- Variance
  - What is the (population) variance of the below dataset?

5	6	7	8	9

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

Spread

- Variance
  - What is the (population) variance of the below dataset?
  - The mean of the below dataset is **7**.
  - The number of observations in the dataset is **5**.
  - Squared Deviations:
    - (5 - 7) ^ 2 = 4
    - (6 - 7) ^ 2 = 1
    - (7 - 7) ^ 2 = 0
    - (8 - 7) ^ 2 = 1
    - (9 - 7) ^ 2 = 4
  - The sum of all squared deviations is 10. This divided by 5 is 2.
  - Our variance is 2!

5	6	7	8	9

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

Spread

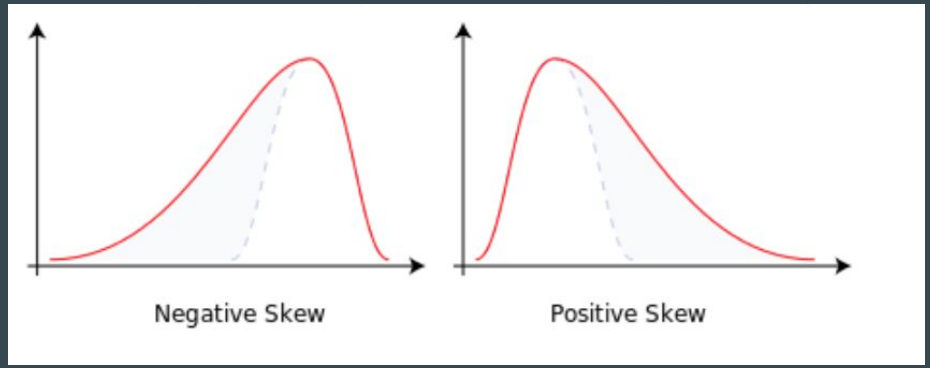
- Standard Deviation
  - The square root of the variance
  - A more common measure of spread than the variance (we will also go into why later in the course)
  - A standard deviation that is equal to the mean or higher is considered “high” but this is an extremely loose measure
  - It’s much more important to take the value in context, relative to contextual expectations or other, similar data sets

$$\sigma = \sqrt{\frac{\sum (X-\mu)^2}{N}} \qquad S = \sqrt{\frac{\sum (X-\mu)^2}{N-1}}$$

Spread

- Skewness
  - A measure of the asymmetry of a dataset around its mean
  - 0 indicates no skew - the dataset is centered around the mean
  - Positive value indicates the data is skewed right - the data is centered below the mean
  - Negative value indicates the data is skewed left - the data is centered above the mean
  - If the skew value is less than -1, or greater than 1, the data is “highly skewed”, though this rule isn’t written in stone

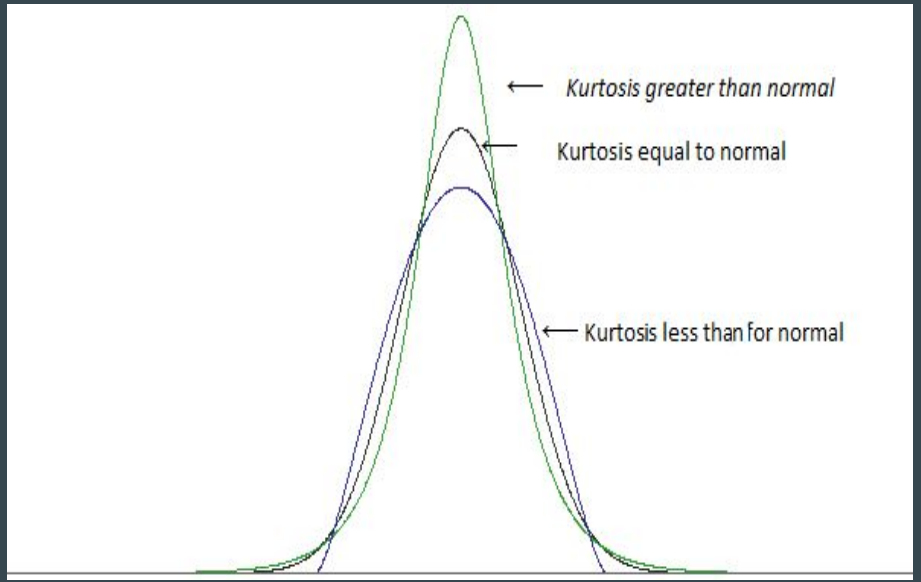
$$g_1 = \frac{\sum (X-\mu)^3 / N}{s^3}$$



Outliers

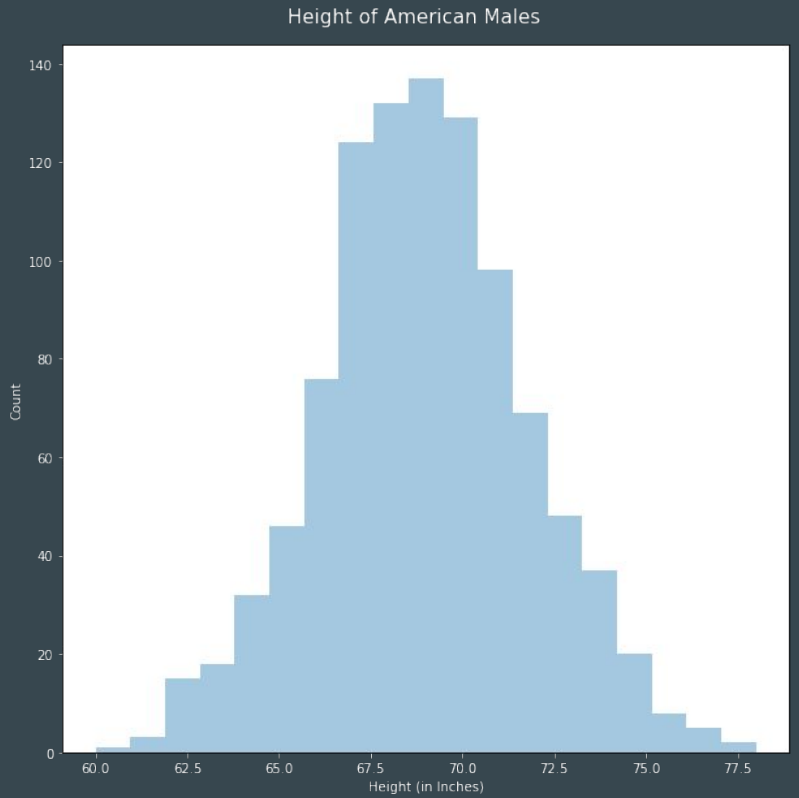
- Kurtosis
  - Measure of how long the tails are in a dataset
  - A kurtosis of 0 indicates a normal distribution (the baseline may also be 3 but in Scipy.Stats the baseline is 0)
  - Higher kurtosis indicates more likely evidence of outliers
  - Commonly misconceived to measure the size of a distribution's 'peak' but more relevant for outliers

$$kurtosis = \frac{\sum (X-\mu)^4 / N}{s^4}$$



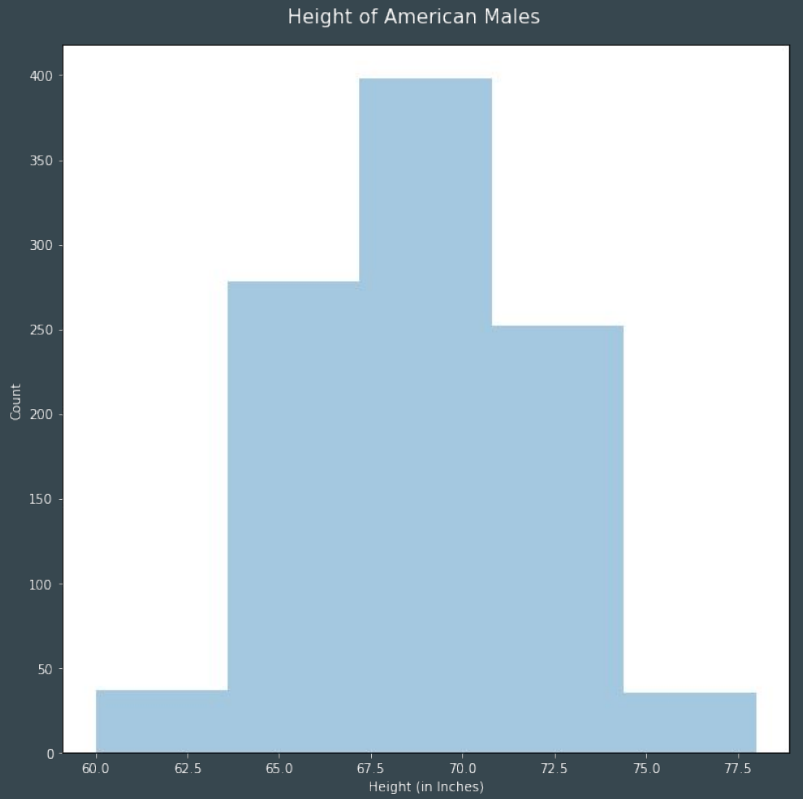
Visualizations

- Histogram
  - 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.



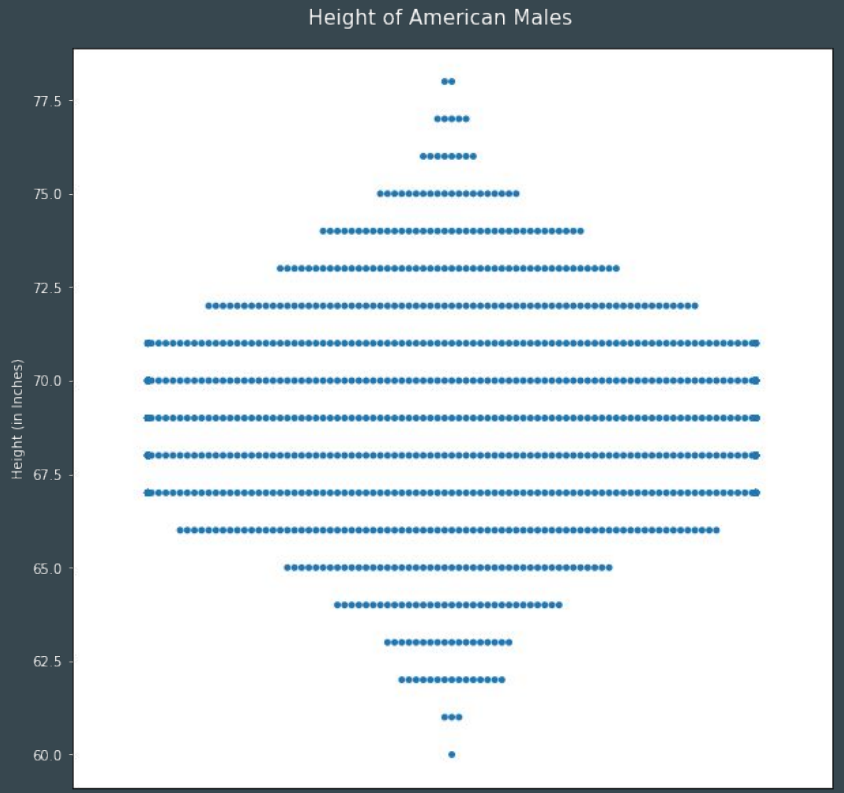
Visualizations

- Histogram
  - 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.



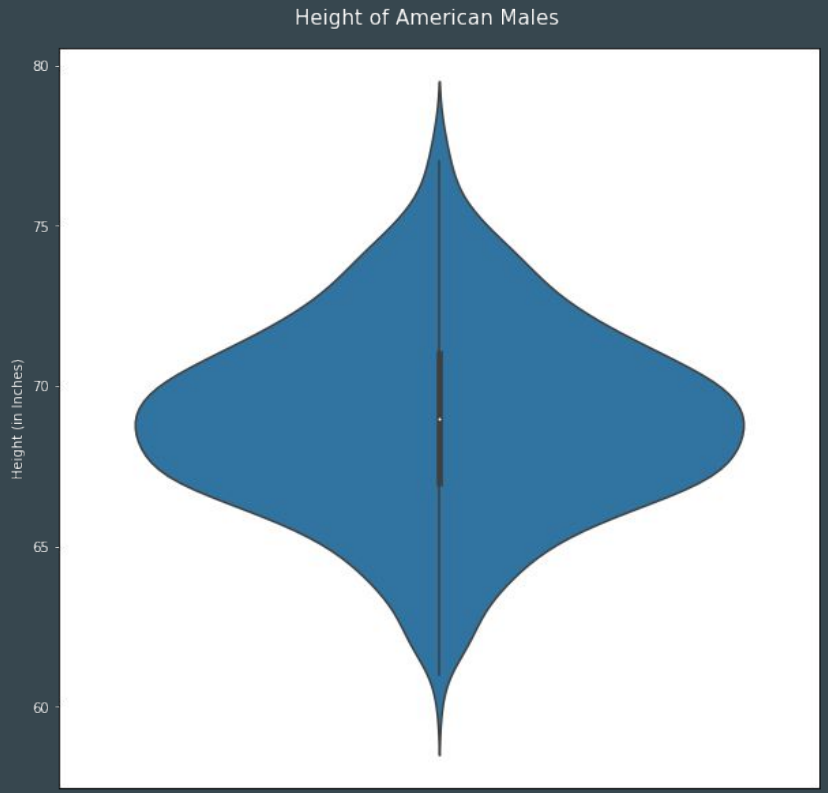
Visualizations

- Swarm Plot
  - 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



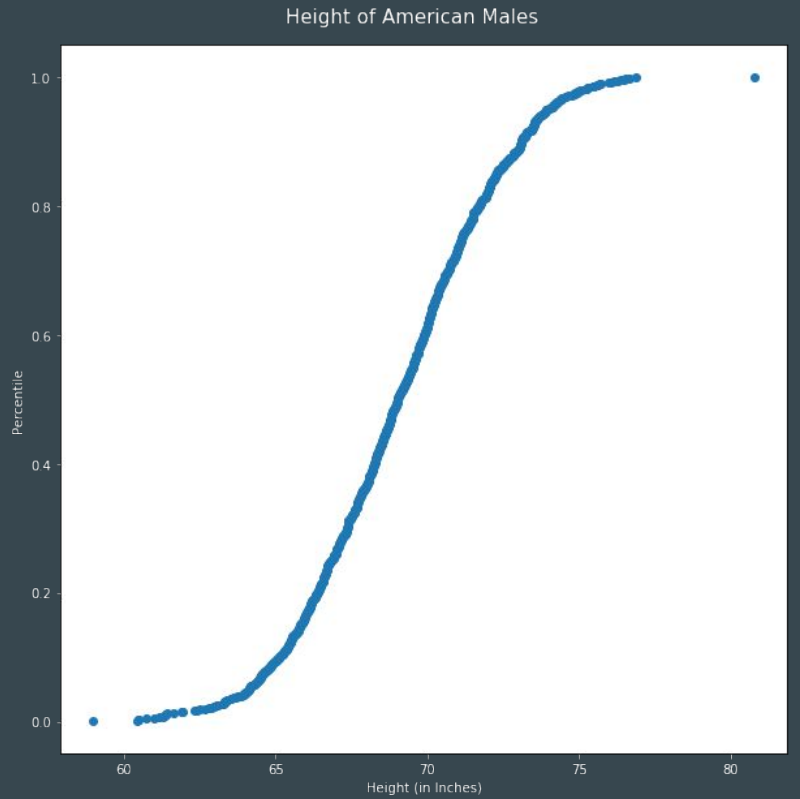
Visualizations

- Violin Plot
  - A violin plot achieves the same effect



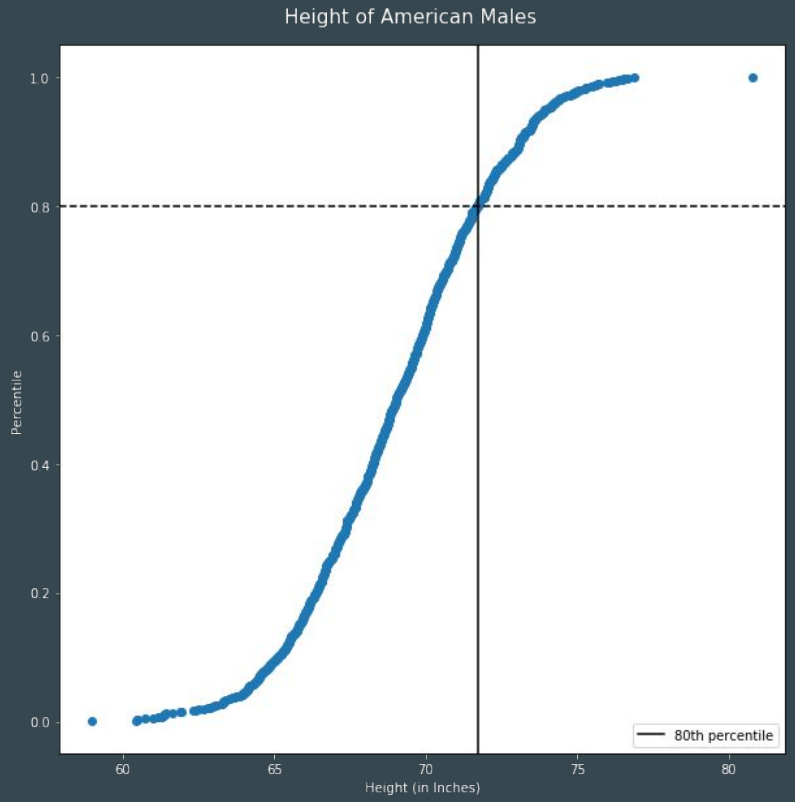
Visualizations

- Empirical Cumulative Distribution Function (ECDF)
  - An ECDF is a good way to see the percentile ranks of a distribution
  - We'll see later in the course that it's also a good way to compare different distributions compared to histograms and swarm plots



Visualizations

- Empirical Cumulative Distribution Function (ECDF)
  - An ECDF is a good way to see the percentile ranks of a distribution
  - We'll see later in the course that it's also a good way to compare different distributions compared to histograms and swarm plots



Visualizations

- Boxplot
  - 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