Part of practicing good data literacy means asking…

- Who participated in the data?
- Who is left out?
- Who made the data?

Ethical issues regarding data collection may be divided into the following categories:

- Consent: Individuals must be informed and give their consent for information to be collected.
- Ownership: Anyone collecting data must be aware that individuals have ownership over their information.
- Intention: Individuals must be informed about what information will be taken, how it will be stored, and how it will be used.
- Privacy: Information about individuals must be kept secure. This is especially important for any and all personally identifiable information.

continuous variables : Height (ft)

nominal variable : Honeylocust, or Pin Oak

dichotomous variable : “on/off”, “yes/no”

ordinal variable : scale of 1 to 5, how pretty we think each tree is

Categorical Variables : variables that contain qualitative information like (proportion: the frequency divided by the total number of musicians)

- Defined your variables to create tidy datasets
- Classified the variables based on their types
- Reconciled messy data
- Decided how to deal with missing data
- Addressed issues of accuracy
- Aligned your questions to the available data to ensure validity
- Created representative samples

These are important techniques for anyone working with data to always be conscious of

The IQR (**interquartile range)** is the difference between Q3 and Q1 (quartile 1, quartile 3 which is 25% and 75%), marking the range for just the middle 50% of the data (median).

The mode is defined as the value with the highest frequency, but we can also think of the mode as the value where the peak of the distribution occurs.

The **standard deviation** describes the spread of values in a numeric distribution relative to the mean.

Aesthetic properties are the attributes we use to communicate data visually:

- Position

- Size
- Shape
- Color / pattern

information redundancy : Information redundancy helps key data points to stand out. visualizes the same information using multiple different aesthetic properties. It's important for readability, organization and prioritization of information, and accessibility.

Chart types

bar graph : compare an amount between different categories

histogram : Histograms measure the distribution or spread, of a variable

Box plot : make percentile and quartile values obvious

Violin plot : show the peaks in data

Bivariate chart : a statistical method examining how two different things are related

multivariate charts : how correlated, or dependent, variables jointly affect a process or outcome

scatter plot : uses dots to represent values for two different numeric variables (only makes sense for numeric variables, not categorical)

line chart : measuring a variable changing over time

stock chart : measures the value of a company over time

bivariate map : displays two or more variables on a single map

Universal Design

making our work available and easier to access for more people

- Readability: keep the reading level to a high school level whenever possible
- Prior knowledge: define unfamiliar terms and avoid unnecessary jargon
- Information overload: introduce new information with intentional pacing and organization

logarithmic scale : exponential growth that won't fit on the page with a linear scale (like from 1 – 10 -180 on axis)

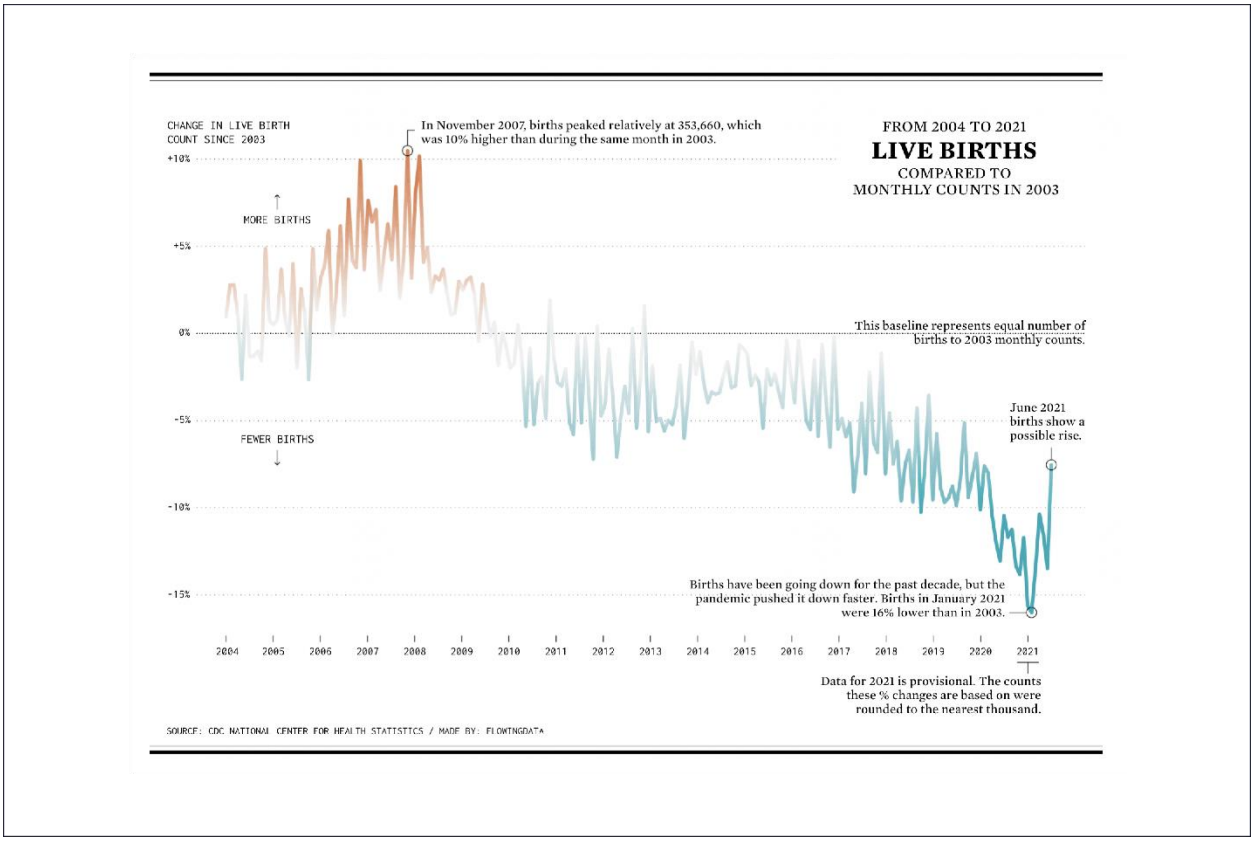
Color Scales

Sequential scales : the same hue with more and more white added to or taken away from the color. Sequential scales are used to show a variable increasing or decreasing in intensity or amount, like income

Divergent scales : opposite sides of the color wheel. A divergent scale is used to visualize data where the middle is a baseline, and either side represents a contrasting change

Categorical scales : use a variety of colors to differentiate categories without assigning a rank or order to them.

Annotations value



this *Live Births* graph from FlowingData to see how much value the annotations add. They…

- add detail to the highest and lowest points on the graph
- explain what the 0% baseline means
- provide a caveat for the 2021 data

- reinforce in words that the percents on the y-axis show “more births” and “fewer births”

Descriptive analyses include measures of central tendency (e.g., mean, median, mode) and spread (e.g., range, quartiles, variance, standard deviation, distribution), which are referred to as descriptives or summary statistics.

Exploratory analysis explores relationships between variables within a dataset and can group subsets of data. Exploratory analyses might reveal correlations between variables. Principal Component Analysis or PCA, which compresses the variables into principal components that can be plotted against each other

Inferential analysis lets us test a hypothesis on a sample of a population and then extend our conclusions to the whole population.

- Sample size must be big enough compared to the total population size (10% is a good rule-of-thumb).
- Our sample must be randomly selected and representative of the total population.
- We can only test one hypothesis at a time.

Causal analysis lets us go beyond correlation and actually assign causation when we carefully design and conduct experiments. In addition, causal inference sometimes allows us to determine causal effects even when experimentation is not possible.

Predictive analysis goes beyond understanding the past and present and allows us to make data-driven predictions about the future uses data and supervised machine learning techniques to identify the likelihood of future outcomes.

Biases are systematic errors in thinking influenced by cultural and personal experiences.

Selection bias occurs when study subjects (i.e., the sample) are not representative of the population.

Algorithmic bias arises when an algorithm produces systematic and repeatable errors that lead to unfair outcomes, such as privileging one group over another.

Testing an algorithm with a non-representative dataset leads to evaluation bias.

Confirmation bias influences data analysis when we consciously or unconsciously interpret results in a way that supports our original hypothesis.

Overgeneralization bias is inappropriately extending observations made with one dataset to other datasets, leading to overinterpreting results and unjustified extrapolation.

Reporting bias is the human tendency to only report or share results that affirm our beliefs or hypotheses, also known as "positive" results

multi-step process where some users leave at each step is called a funnel.

A churn rate is the percent of subscribers to a monthly service who have canceled.

Data acquisition (also called data mining) is the process of gathering data.

Making API request from browser

`https://api.census.gov/data/2020/acs/acs5?get=NAME,B08303_001E&for=state:*`.

The part of the URL after `?` contains the query parameters, each parameter is separated by `&`.

- The `get` parameter specifies a comma-separated list of variables we want to fetch
  - `NAME` is the name of the geographic level
  - `B08303_001E` is the number of commuters
- The `for` parameter specifies the geographic level
  - we are requesting state-level data
  - and we want all states, as indicated by the `*`.

Examples: https://api.census.gov/data/2020/acs/acs5/examples.html

JSON is a great universal format for data interchange

The JSON data we got from the Census API is a list of lists in Python, where each inner list corresponds to a single row of data.

Binomial events always have 2 possible outcomes, which we refer to as *success* and *failure*. The probability of a successful outcome is represented by the parameter `p`. For example, for the event of a coin toss using a fair coin, `p` would be `0.5`

Linear Regression is when you have a group of points on a graph, and you find a line that approximately resembles that group of points. A good Linear Regression algorithm minimizes the *error*, or the distance from each point to the line. A line

with the least error is the line that fits the data the best. We call this a line of *best fit*.

The ways we were able to explore this data set in preparation for a regression model:

- We previewed the first few rows of the data set using the .head() method.
- We checked the data type of each variable in the data set using .dtypes and corrected variables with incorrect data types.
- We investigated our categorical data to inform categorical encoding.
- We investigated the scale of our quantitative variables and considered whether standardizing/scaling might be appropriate.
- We investigated missing data.
- We checked for outliers.
- We inspected the distributions of our quantitative variables.
- We looked at the relationships between pairs of features using both scatter plots and box plots.

how to scope data science/analytics projects and what questions they need to answer before launching the project.

Step 1: Goals – Define the goal(s) of the project

Step 2: Actions – What actions/interventions do you have that this project will inform?

Step 3: Data – What data do you have access to internally? What data do you need? What can you augment from external and/or public sources?

Step 4: Analysis – What analysis needs to be done? Does it involve description, detection, prediction, optimization, or behavior change? How will the analysis be validated?

Ethical Considerations: How have you thought through privacy, transparency, discrimination/equity, and accountability issues around this project?

A **DataFrame** is structured like a table or spreadsheet. The rows and the columns both have indexes

NumPy arrays are unique in that they are more flexible than normal Python lists. They are called **ndarrays** since they can have any number (n) of dimensions (d).

the **Series** is the core object of the pandas library.

NumPy arrays have one dtype for the entire array, while pandas DataFrames have one dtype per column.

An ***aggregate* statistic** is a way of creating a single number that describes a group of numbers.

Exploratory Data Analysis (EDA) is all about getting curious about your data – finding out what is there, what patterns you can find, and what relationships exist.

EDA techniques

The EDA process generally involves strategies that fall into the following three categories:

1. **Data inspection**: Data inspection is an important first step of any analysis. This can help illuminate potential issues or avenues for further investigation. Like .head()
2. **Numerical summarization:** For numerical data, allows us to get a sense of scale, spread, and central tendency. For categorical data, this gives us information about the number of categories and frequencies of each. Like .describe(include = 'all')
3. **Data visualization:** visual summaries can provide even more context and detail in a small amount of space.

Assessing Variable

there are only two types of variables: numerical and categorical. In "flat" file formats (like tables, csvs, or DataFrames), the observations are the rows, the variables are the columns, and the values are at the intersection.

Categorical variables come in 3 types:

- Nominal variables, which describe something like red, yellow, blue or hot, cold
- Ordinal variables, which have an inherent ranking like 1st, 2nd, 3rd.
- Binary variables, which have only two possible variations.

Quantitative Variables (numerical)

- Continuous variables come from measurements. Like Length, time, and temperature (float)
- Discrete variables come from counting. Like People, cars, and dogs (int)

One-Hot Encoding (OHE) *:* This technique is useful when managing nominal variables because it encodes the variable without creating an order among the categories.

Central Tendency for Quantitative Data

The central location (also called central tendency) is often used to communicate the “typical” value of a variable. Recall that there are a few different ways of calculating the central location:

- **Mean**: Also called the “average”; calculated as the sum of all values divided by the number of values.
- **Median**: The middle value of the variable when sorted.
- **Mode**: The most frequent value in the variable.
- **Trimmed Mean**: The mean excluding x percent of the lowest and highest data points.

-Boxplots and histograms are often used for visualization.

Spread (Quantitative Data)

Spread, or dispersion, describes the variability within a feature. there are a few values that can describe the spread:

- **Range**: The difference between the maximum and minimum values in a variable.
- **Inter-Quartile Range (IQR):** The difference between the 75th and 25th percentile values.
- **Variance:** The average of the squared distance from each data point to the mean.
- **Standard Deviation (SD):** The square root of the variance.
- **Mean Absolute Deviation (MAD):** The mean absolute value of the distance between each data point and the mean.

-Bar charts and pie charts are often used for visualization.

The **variance** is a measure of variability. It is calculated by taking the average of squared deviations from the mean. Variance tells you the degree of spread in your data set. The more spread the data, the larger the variance is in relation to the mean.

Covariance measures the direction of a relationship between two variables, while **correlation** measures the strength of that relationship. Both correlation and covariance are positive when the variables move in the same direction and negative when they move in opposite directions.

covariance between two variables shows no visible linear relationship

A covariance matrix for two variables looks something like this:

	variable 1	**variable 2**
variable 1	variance(variable 1)	covariance
variable 2	covariance	variance(variable 2)

Correlation:

Definition: Correlation measures the strength and direction of a linear relationship between two variables. It's a standardized measure that ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation), with 0 indicating no linear relationship.

Use: Correlation is used to assess the strength and direction of the linear relationship between two variables. It's valuable for feature selection, regression analysis, and understanding associations in data.

The proportion of respondents in each category of a single question is called a **marginal proportion**.

Chi-Square

The chi-squared test is a statistical test used to determine whether there is a significant association or independence between two categorical variables. The null hypothesis assumes independence between the variables, while the alternative hypothesis suggests a significant association.

compare the p-value to a significance level (alpha: for ex 0.05) to make a decision about whether to reject the null hypothesis.

If the p-value is less than alpha, you can conclude that there is a significant association between the variables. If it's greater than alpha, you fail to reject the null hypothesis, indicating no significant association.

use the *Chi-Square statistic* to summarize how different these two tables are.

$$ChiSquare = \sum \frac{(observed - expected)^2}{expected}$$

different methods to assess whether there is an association between categorical variables :

- Contingency tables of frequencies
- Contingency tables of proportions
- Marginal proportions
- Expected contingency tables
- The Chi-Square statistic

Feature	Correlation	Contingency Table	Covariance
Variable types	Numerical	Categorical	Numerical
Relationship	Linear	Non-linear patterns	Average co-variation
Direction	Positive/negative	Not applicable	Positive/negative
Standardization	Yes	No	No
Interpretation	Strength and direction of linear relationship	Frequency distribution and association	Average tendency to vary together

Two events are considered **mutually exclusive** if they cannot occur at the same time. For example, consider a single coin flip: the events “tails” and “heads”

If we want to calculate the probability that a pair of dependent events both occur, we need to define **conditional probability**.

conditional probability measures the probability of one event occurring, given that another one has already occurred.

$$P(A \mid B) = P(A) \quad \text{and} \quad P(B \mid A) = P(B)$$

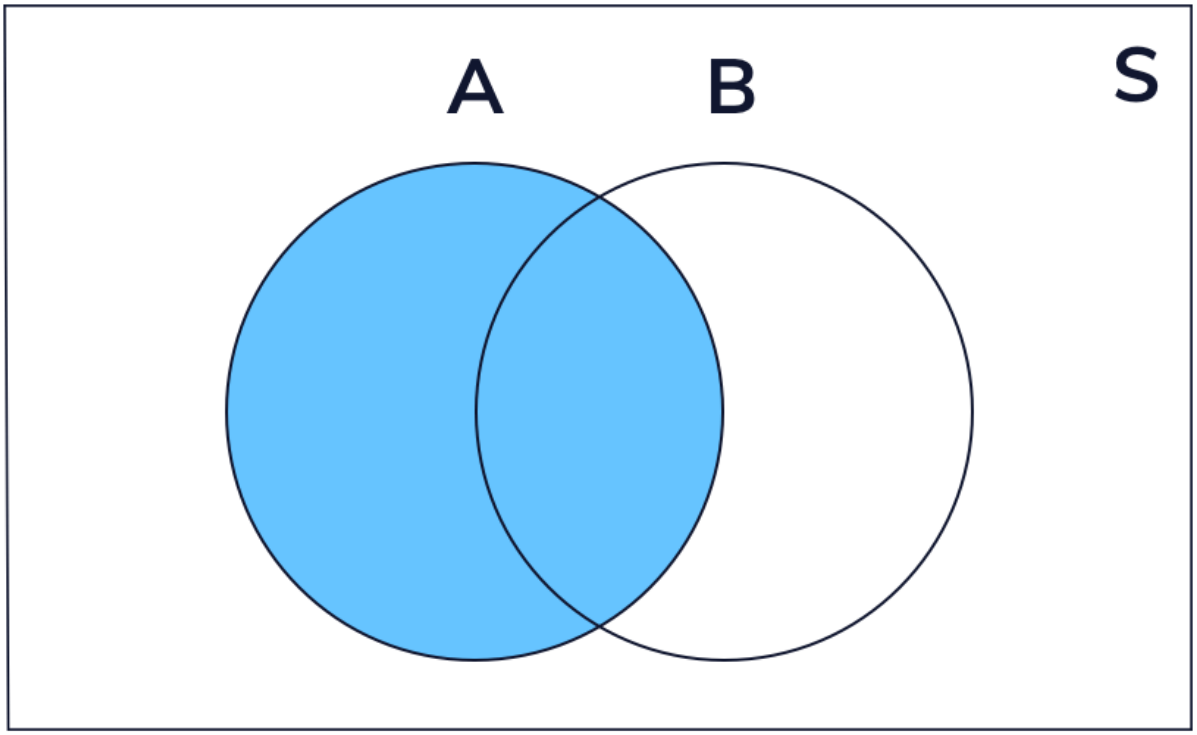
Multiplication Rule

probability that two events happen simultaneously

$P(A \text{ and } B) = P(A) \cdot P(B \mid A)$

For Independent Events is :
$P(A \text{ and } B) = P(A) \cdot P(B)$

Addition Rule For not mutually exclusive



P(A or B) = P(A)

For not mutually exclusive ： $P(A \text{ or } B)=P(A)+P(B)$

A **probability mass function (PMF)** is a type of *probability distribution* that defines the probability of observing a particular value of a discrete random variable.

The probability mass function that describes the likelihood of each possible outcome (eg., 0 heads, 1 head, 2 heads, etc.) is called the **binomial distribution**. The parameters for the binomial distribution are:

- `n` for the number of trials (eg., n=10 if we flip a coin 10 times)
- `p` for the probability of success in each trial (probability of observing a particular outcome in each trial. In this example, p= 0.5 because the probability of observing heads on a fair coin flip is 0.5)

The **cumulative distribution function** for a discrete random variable can be derived from the probability mass function. However, instead of the probability of observing a specific value, the cumulative distribution function gives the probability of observing a specific value OR LESS.

Cumulative distribution functions are constantly increasing, so for two different numbers that the random variable could take on, the value of the function will always be greater for the larger number. Mathematically, this is represented as:

If $x_1 < x_2 \rightarrow CDF(x_1) < CDF(x_2)$
We showed how the probability mass function can be used to calculate the probability of observing less than 3 heads out of 10 coin flips by adding up the probabilities of observing 0, 1, and 2 heads. The cumulative distribution function produces the same answer by evaluating the function at `CDF(X=2)`. In this case, using the CDF is simpler than the PMF because it requires one calculation rather than three.

Probability Density Functions

They define the probability distributions of continuous random variables and span across all possible values that the given random variable can take on.

When trying to evaluate the area under the curve at a specific point, the width of that area becomes 0, and therefore the probability equals 0.

heights fall under a type of probability distribution called a *normal distribution*. The parameters for the normal distribution are the mean and the standard deviation, and we use the form *Normal(mean, standard deviation)* as shorthand.

The **Poisson distribution** is another common distribution, and it is used to describe the number of times a certain event occurs within a fixed time or space interval. For example, the Poisson distribution can be used to describe the number of cars that pass through a specific intersection between 4pm and 5pm on a given day.

Expected Value of the Binomial Distribution

$Expected(\#ofHeads) = E(X) = n \times p$ ➔ n, representing the number of events and p, representing the probability of "success"

Variance of the Binomial Distribution

$Variance(\#ofHeads) = Var(X) = n \times p \times (1-p)$

$Variance(\#ofHeads) = 10 \times 0.5 \times (1-0.5) = 2.5$

The **Central Limit Theorem (CLT)** allows us to specifically describe the sampling distribution of the mean.

The CLT states that the sampling distribution of the mean is normally distributed as long as the population is not too skewed or the sample size is large enough.

Note that the CLT only applies to the sampling distribution of the mean and not other statistics like maximum, minimum, and variance!

- mean $\bar{x}$ approximately equal to the population mean μ
- standard deviation equal to the population standard deviation divided by the square root of the sample size. We can write this out as:

$$Sampling\ Distribution\ St.Dev = \frac{\sigma}{\sqrt{n}}$$

The standard deviation of a sampling distribution is also known as the **standard error** *of the estimate of the mean*. In many instances, we cannot know the population standard deviation, so we estimate the standard error using the sample standard deviation:

$$\frac{standard\ deviation\ of\ our\ sample}{\sqrt{\text{sample size}}}$$

Keep in mind that:

- As sample size increases, the standard error will decrease.
- As the population standard deviation increases, so will the standard error.

A statistic is called an **unbiased estimator** of a population parameter if the mean of the sampling distribution of the statistic is equal to the value of the statistic for the population.

population variance is calculated as:

$$population\ variance = \frac{\sum(observation - \mu)^2}{n}$$

When we measure the sample variance using the same formula, it turns out that we tend to underestimate the population variance. Because of this, we divide by *n-1* instead of *n*:

$$sample\ variance = \frac{\sum(observation - sample\ mean)^2}{n - 1}$$

Introduction to Hypothesis Testing

Step 1: Ask a Question

Step 2: Define the Null and Alternative Hypotheses

null hypothesis is a type of statistical hypothesis that proposes that no statistical significance exists in a set of given observations

An alternative hypothesis is an opposing theory to the null hypothesis

Step 3: Determine the Null Distribution

Now that we have our null hypothesis, we can generate a null distribution: the distribution (across repeated samples) of the statistic we are interested in if the null hypothesis is true.

Step 4: Calculate a P-Value (or Confidence Interval)

- One Sided or One-Tailed test
- Two Sided or Two-Tailed test, referencing the two tails of the null distribution that are counted in the p-value
- observing a sample average less than or equal to

Step 5: Interpret the Results

Significance thresholds

The significance threshold can be any number between 0 and 1, but a common choice is 0.05. P-values that are less than this threshold are considered "significant", while larger p-values are considered "not significant".

P-values below the chosen threshold are declared significant and lead the data scientist to "reject the null hypothesis in favor of the alternative". A common choice for this threshold, which is also sometimes referred to as Alpha, is 0.05 —

but this is an arbitrary choice! Using a lower threshold means you are less likely to find significant results, but also less likely to mistakenly report a significant result when there is none

* the true probability of a learner answering the question correctly **is** 70% (if we showed the question to ALL learners, exactly 70% would answer it correctly). This puts us in the first column of the table above (the null hypothesis "is true"). If we run a test and calculate a significant p-value, we will make type I error (also called a false positive because the p-value is falsely significant), leading us to remove the question when we don't need to. If we run a test and calculate a non-significant p-value, we make a type II error, leading us to leave the question on our site when we should have taken it down.

Null hypothesis:	is true	is false
P-value significant	Type I Error	Correct!
P-value not significant	Correct!	Type II error

*Impact of the Alternative Hypothesis

the one-sided test described above (p = 0.031) would lead a data scientist to reject the null at a 0.05 significance level. Meanwhile, a two-sided test on the same data leads to a p-value of 0.062, which is greater than the 0.05 threshold. Thus, for the two-sided test, a data scientist could not reject the null hypothesis.

***Summary**

- A p-value is a probability, usually reported as a decimal between zero and one.
- A small p-value means that an observed sample statistic (or something more extreme) would be unlikely to occur if the null hypothesis is true.
- A significance threshold can be used to translate a p-value into a "significant" or "non-significant" result.

- In practice, the alternative hypothesis and significance threshold should be chosen prior to data collection.

The binomial distribution describes the number of expected "successes" in an experiment with some number of "trials".

hypothesis tests in general:

- All hypothesis tests start with a null and alternative hypothesis
- Outcomes of a hypothesis test that might be reported include:
    - confidence intervals
    - p-values
- A hypothesis test can be simulated by:
    - taking repeated random samples where the null hypothesis is assumed to be true
    - using those simulated samples to generate a null distribution
    - comparing an observed sample statistic to that null distribution

Linear Regression

It can be used to understand the relationship between a quantitative variable and one or more other variables, sometimes with the goal of making predictions.

$y=mx+b$. In this equation:

- x and y represent variables, such as height and weight or hours of studying and quiz scores.
- b represents the *y-intercept* of the line. This is where the line intersects with the y-axis (a vertical line located at $x = 0$).

- *m* represents the slope. This controls how steep the line is. If we choose any two points on a line, the slope is the ratio between the vertical and horizontal distance between those points; this is often written as rise/run.

Finding the "Best" Line

In simple OLS (*ordinary least squares*) regression, we assume that the relationship between two variables *x* and *y* can be modeled as:

$$y=mx+b+error$$

Linear Interpolation

Linear interpolation is a method in which we can predict the value of missing data by connecting a line through two adjacent data points. The basic idea with linear interpolation is that, if we can understand what the data is at point A and point C, we can take the mean of those points and provide a value at point B.

$$y=mx+b$$

- y = the value we are trying to predict
- m = the slope of the line we are seeing
- x = the known value for which we are predicting y
- b = the intercept, or starting point on the y-axis

In the case of (most) datasets, we won't have values for these variables.

$$y = y_1 + (x - x_1) * \frac{(y_2 - y_1)}{(x_2 - x_1)}$$

Linear interpolation is a simple technique that can be very useful when our data is linear.

Normality assumption

The normality assumption states that the residuals should be normally distributed. To check this assumption, we can inspect a histogram of the residuals and make sure that the distribution looks approximately normal

Homoscedasticity assumption

Homoscedasticity is a fancy way of saying that the residuals have equal variation across all values of the predictor variable. A common way to check this is by plotting the residuals against the fitted values.

The X Matrix

when we use statsmodels.api.OLS.from_formula() to create a model. When we pass a formula to this function (like 'weight ~ height' or 'rent ~ borough'), it actually creates a new data set, which we don’t see. This new data set is often referred to as the X matrix, and it is used to fit the model.

When we have multiple axes in the same picture, we call each set of axes a ***subplot.*** The picture or object that contains all of the subplots is called a ***figure***.

when to use different types of plots:

- Compare categories of data with bar graphs
- Show uncertainty in data using error bars and shading

- Compare proportional datasets using pie charts
- Analyze frequency data using histograms

Each **bin** is represented by a different rectangle whose height is the number of elements from the dataset that fall within that bin.

The *width* of each bin is the distance between the minimum and maximum values of each bin.

Chart categories

Composition charts

Focusing Question: What are the parts of some whole? What is the data made of?

Distribution Charts

Datasets that work well: Data in large quantities and/or with an array of attributes works well for these types of charts. Visualizations in this category will allow you to see patterns, re-occurrences, and a clustering of data points.

Note: In statistics, a commonly seen distribution is a bell curve, also known as a normal distribution. A bell curve is a bell-shaped distribution where most of the values in the dataset crowd around the average

Relationship Charts

Focusing Question: How do variables relate to each other?

Datasets that work well: Data with two or more variables can be displayed in these charts. These charts typically illustrate a correlation between two or more variables.

Comparison Charts

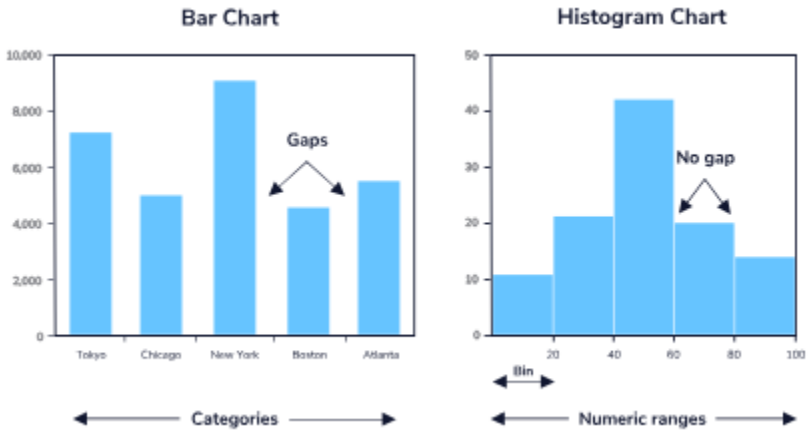
Focusing Question: How do variables compare to each other?

Datasets that work well: Data must have multiple variables, and the visualizations in this category allow readers to compare those items against the others. For example, a line graph that has multiple lines, each belonging to a different variable. Multi-colored bar charts are also a great way to compare items in data.

Bar Charts vs. Histograms

- Bar charts are used for categorical variables, while histograms are used for quantitative data.

- Histograms must always be arranged in a specific order because they represent a range of numerical values. For a bar chart, each bar represents frequencies of category variables within a category.



Bar Charts vs. pie charts

While bar charts are usually used for visualizing a table of frequencies, pie charts are an alternative when you want to visualize a table of proportions.

Univariate analysis

Univariate analysis focuses on a single variable at a time. Univariate data visualizations can help us answer questions like:

What is the typical price of a rental in New York City?

Quantitative variables

Box plots (or violin plots) and histograms are common choices for visually summarizing a quantitative variable. Show information about minimum and maximum values, central location, and spread. Histograms can additionally illuminate patterns that can impact an analysis

Categorical variables

For categorical variables, we can use a bar plot (instead of a histogram) to quickly visualize the frequency (or proportion) of values in each category. Alternatively, we could use a pie chart to communicate the same

Bivariate analysis

In many cases, a data analyst is interested in the relationship between two variables in a dataset. For example:

- Do apartments in different boroughs tend to cost different amounts?

One quantitative variable and one categorical variable

Two good options for investigating the relationship between a quantitative variable and a categorical variable are side-by-side box plots and overlapping histograms.

Two quantitative variables

A scatter plot is a great option for investigating the relationship between two quantitative variables.

Two categorical variables

Side by side (or stacked) bar plots are useful for visualizing the relationship between two categorical variables.

Multivariate analysis

exploring the relationship between three or more variables in a single visualization. scatter plot using visual cues such as colors, shapes, and patterns. Another common data visualization for multivariate analysis is a heat map of a correlation matrix for all quantitative variables

Time Series Data

Data represented in a single point in time is known as cross-sectional data. As a Data Scientist or Analyst, sometimes you might encounter data that is collected over periods of time, known as time series data.

time series data plots:

Line plot

Box plot

heatmap

Lag scatter plot:

We can use a lag scatter plot to explore the relationship between an observation and a lag of that observation.

Autocorrelation plot:

An autocorrelation plot is used to show whether the elements of a time series are positively correlated, negatively correlated, or independent of each other.

Different types of missing data

But there's more to missing data than missingness. Missing data comes in four varieties:

- **Structurally Missing Data** we expect this data to be missing for some logical reason
- **Missing Completely at Random (MCAR)** *the probability of any datapoint being MCAR is the same for all data points – this type of missing data is mostly hypothetical*
- **Missing at Random (MAR)** the probability of any data point being MAR is the same within groups of the *observed* data – this is much more realistic than MCAR

- **Missing Not at Random (MNAR)** there is some reason why the data is missing

When trying to **diagnose the type of missingness**, data about the data (aka meta data) can be invaluable. The date/time data was collected, how it was collected, who collected it, where it was collected, etc. can all give invaluable clues to solving the problem of missing data.

Data is safe to delete when:

1. It is either MAR or MCAR missing data. We can remove data that falls into either of these categories without affecting the rest of the data, since we assume that the data is missing at random. However, if the percentage of missing data is too high, then we can't delete the data — we would be reducing our sample size too much.
2. The missing data has a low correlation with other features in the data. If the missing data is not important for what we're doing, then we can safely remove that data.

Types of deletion

- ### Listwise Deletion

  Listwise deletion, also known as complete-case analysis, is a technique in which we remove the entire observation when there is missing data. This particular technique is usually employed when the missing variable(s) will directly impact the analysis we are trying to perform, A safe place to start is assuming that if less than 5% of data is missing.

```
data.dropna(inplace=True)
```

- **Pairwise Deletion**
  looks for context to what we are trying to analyze. In pairwise deletion, we only remove rows when there are missing values in the variables we are directly analyzing. Unlike listwise deletion, we do not care if other variables are missing, and can retain those rows.

  data.dropna(subset=['Height','Education'], inplace=True, how='any')

LOCF stands for Last Observation Carried Forward. With this technique, we are going to fill in the missing data with the previous value.

```
df['comfort'].ffill(axis=0, inplace=True)
```

NOCB stands for Next Observation Carried Backward, and it solves imputation in the opposite direction of LOCF.

```
df['comfort'].bfill(axis=0, inplace=True)
```

BOCF, or Baseline Observation Carried Forward. In this approach, the initial values for a given variable are applied to missing values.

```
baseline = df['concentration'][0]
```

```
baseline = df['concentration'][0]
```

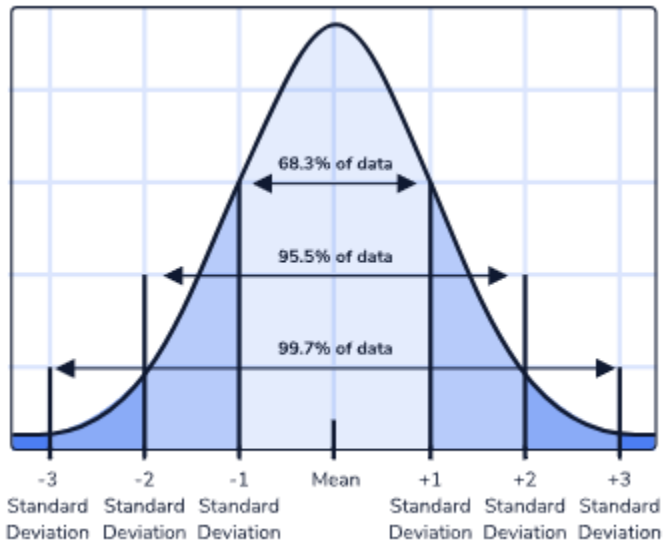
WOCF, or Worst Observation Carried Forward. With this kind of imputation, we want to assume that the data is the worst possible value. This would be useful if the purpose of our analysis was to record improvement in some value (for example, if we wanted to study if a treatment was helping a particular patient's condition)de

```
worst = df['pain'].max()
```

```
df['pain'].fillna(value=worst, inplace=True)
```

Using Standard Deviation

we can begin to investigate how unusual that datapoint truly is. In fact, you can usually expect around 68% of your data to fall within one standard deviation of the mean, 95% of your data to fall within two standard deviations of the mean, and 99.7% of your data to fall within three standard deviations of the mean.



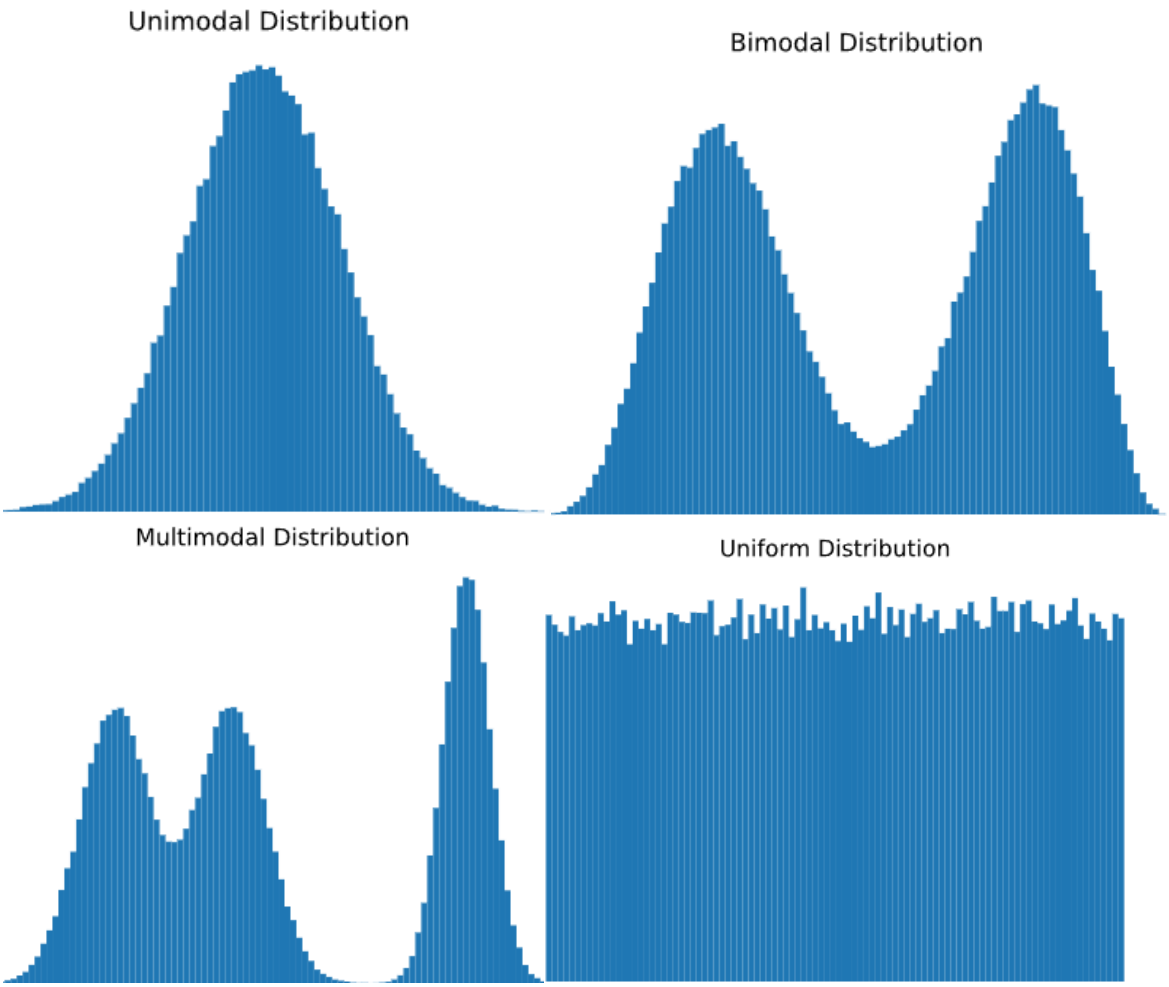
If you have a data point that is over three standard deviations away from the mean, that's an incredibly unusual piece of data!

How to interpret a distribution using the following five features of a dataset:

- Center
- Spread
- Skew
- Modality
- Outliers

Modality

The modality describes the number of peaks in a dataset.



interpret summaries of nominal categorical and ordinal categorical variables.

- For nominal categorical variables, there is no ordering to the categories. Because of this, we're limited to using the mode to describe central tendency and there is no way to summarize the spread.

- For ordinal categorical variables, there is an implied ordering to the categories. In Python, we can use pd.Categorical() to transform a variable to a categorical type. The Categorical type allows us to access a numeric value for each category by using .cat.codes. From there, we may perform operations on this variable as if it were a regular, numeric variable.

- For ordinal categorical variables, median and mode can be used to summarize the central tendency, and the IQR (or any difference between percentiles) can be used to summarize the spread.

Ordinal Categories: Spread

the mean is not interpretable for ordinal categorical variables because the mean relies on the assumption of equal spacing between categories.

Remember that the standard deviation and variance both depend on the mean, without a mean, we can't have a reliable standard deviation or variance either!

Instead, we can rely on other summary statistics, like the proportion of the data within a range, or percentiles/quantiles.

transform your data before visualizing

Data Centering

Data centering involves subtracting the mean of a data set from each data point so that the new mean is 0

Centered data is useful because it tells us how far above or below the mean each data point is.

Data Scaling

Two of the most commonly used data scaling techniques are:

- Min-max normalization
- Standardization

1. Min-Max Normalization

The goal of normalization is to put features with different ranges onto the same scale.

the minimum value of that feature is transformed into 0, the maximum value is transformed into 1, and every other value is transformed into a decimal between 0 and 1.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

2. Standardization

Standardization (also known as Z-score normalization) involves subtracting the mean of each observation and then dividing by the standard deviation:

$$z = \frac{value - mean}{stdev}$$

This will help avoid outlier, all the features will have a mean of zero, a standard deviation of one, and therefore, the same scale.

When to Normalize vs. Standardize?

if you need your data to be on a 0-1 scale, then it makes sense to use min-max normalization. If you have outliers in your data, then it is best to use standardization (Z-score normalization) since it does not have a bounding range like min-max normalization does.

Binning Data

The process of transforming numerical variables into categorical counterparts is called “binning.”

Binning is a way to group a number of continuous values into a smaller number of “bins”.

SQL

Date and Time Functions

```
SELECT DATETIME(timestring, modifier1, modifier2, ...);
```

SELECT DATETIME('now', 'localtime');

shift the date backwards to a specified part of the date.

- `start of year`: shifts the date to the beginning of the current year.
- `start of month`: shifts the date to the beginning of the current month.
- `start of day`: shifts the date to the beginning of the current day.

```
SELECT DATE('2005-09-15', 'start of month');
```

add a specified amount to the date and time of the time string.

- `'+-N years'`: offsets the year
- `'+-N months'`: offsets the month
- `'+-N days'`: offsets the day
- `'+-N hours'`: offsets the hour
- `'+-N minutes'`: offsets the minute
- `'+-N seconds'`: offsets the second

```
SELECT DATETIME('2020-02-10', 'start of month', '-1 day', '+7 hours');
```

The substitutions to extract each part of the date and time are the following:

- `%Y` returns the year (YYYY)
- `%m` returns the month (01-12)
- `%d` returns the day of month (01-31)
- `%H` returns the hour (00-23)
- `%M` returns the minute (00-59)
- `%S` returns the second (00-59)

Storytelling

Part I: Designing a Story

Many successful data visualizations and dashboards start by working on paper first. This is where the storytelling process begins.

Think about the following, and write or sketch out what comes to mind:

- What is the objective, or core question(s), of the viz? In other words, What question do you want to answer?
- Who is the intended audience?
- What data will be used and what insights can be drawn from it?

For a narrative, we may want to give viewers the option to explore the data to come up with their own conclusion, or we might choose to lead the viewer to a specific conclusion.

Tableau offers some commonly used approaches to storytelling with data that can help us find focus in a visualization or dashboard.

- Change Over Time: Uses temporal data to illustrate trends or shifts
- Drill Down: Starts with the bigger picture, then gets into finer details
- Zoom Out: Starts with a smaller detail then gets into broad and high-level picture
- Contrast: Depicts differences between 2+ subjects
- Intersections: Explores how 2 initially separate dimensions may converge
- Factors: Takes a broader subject and divides it into factors (types/categories)
- Outliers: Highlights and profiles outlier in the data

Best Practices for Telling Great Stories

Try to take on a beginner's mindset, and imagine that you're seeing your dashboard for the first time. Does it answer the questions you hoped to address? If there's a central argument, is it clear to the viewer? If the dashboard is made for exploration, is it easy to navigate?

Part II: Best Practices

1. Choose the Right Chart
   - Temporal changes: Line chart, area charts
   - Parts of a whole: Pie chart, Treemap
   - Relationship between variables: Scatterplot

- Distribution: Histogram, box plot
- Ranking or Magnitude: Bar chart, packed bubbles

Choose the Right Chart Type for Your Data

2. Don’t Overcomplicate It

   It may be helpful to stick to 3-4 vizzes in a dashboard

3. Visual Balance

   In art and design, visual balance is the sense of “weighted clarity in a composition.”

   we should aim to maintain a sense of balance in our vizzes by utilizing elements such as positioning, size, color, shapes, and textures.

4. Using Aesthetic Properties to Convey a Story Theme
5. Tying It All Together

   Once you’ve finished creating your viz, revisit it and ask yourself these questions:

   - Does the visualization serve its purpose, i.e. answer the core question posed?
   - Are my narrative approach and chart choices appropriate and easy to follow?
   - Am I following visualization best practices?
   - Have I used aesthetic properties to enhance my story?

Histogram

Histograms are used to understand the “shape” of a single column of numeric data. A histogram breaks the data-points into bins and then counts the number of data points in each bin.

Scatterplots

Scatterplots are used for visualizing **correlation** between two columns. As one column increases, does the other also increase? Decrease? Fluctuate?

It’s important to remember that the existence of a correlation like this does not mean the one column is directly influencing the other.

Line charts

Line charts plot points in the data connected with a line to visualize the trend. Multiple lines can be plotted on the same graph to compare trends between groups.

Sparklines

Sparklines are a feature in Excel that places a trend line (like a mini line chart) next to each row of a table.

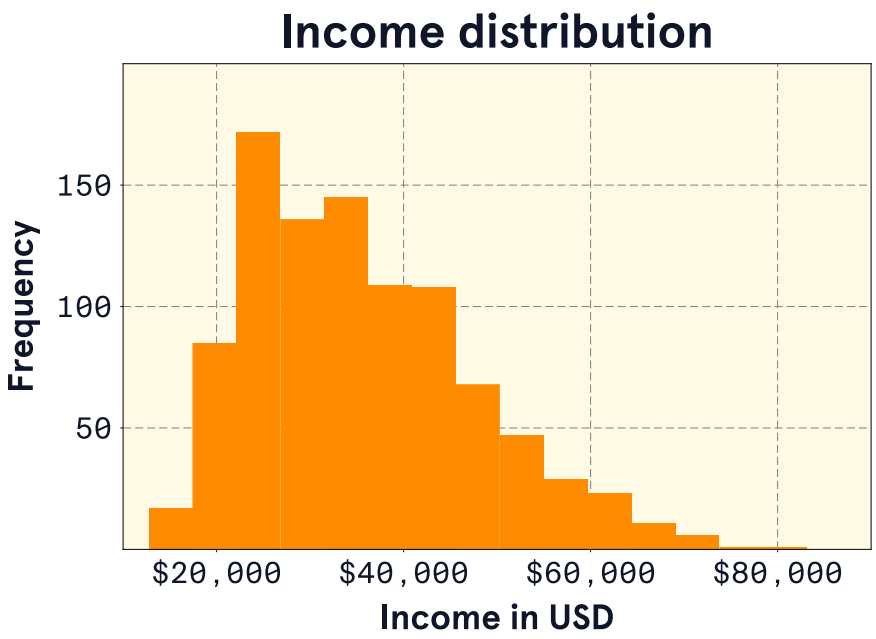
Sparklines can be helpful when we want to see general trends for individual categories.

You are now able to:

- Create programs with Python
- Query and manipulate data with SQL and Python pandas
- Create data visualizations with Python matplotlib
- Summarize and analyze datasets
- Conduct hypothesis testing
- Clean and tidy datasets
- Communicate your findings clearly
- Create dashboards to display your results
- Leverage BI Tools such as Excel when appropriate
- Use Jupyter Notebooks for experimentation and communication
- Tell a story about data and share it with the world.

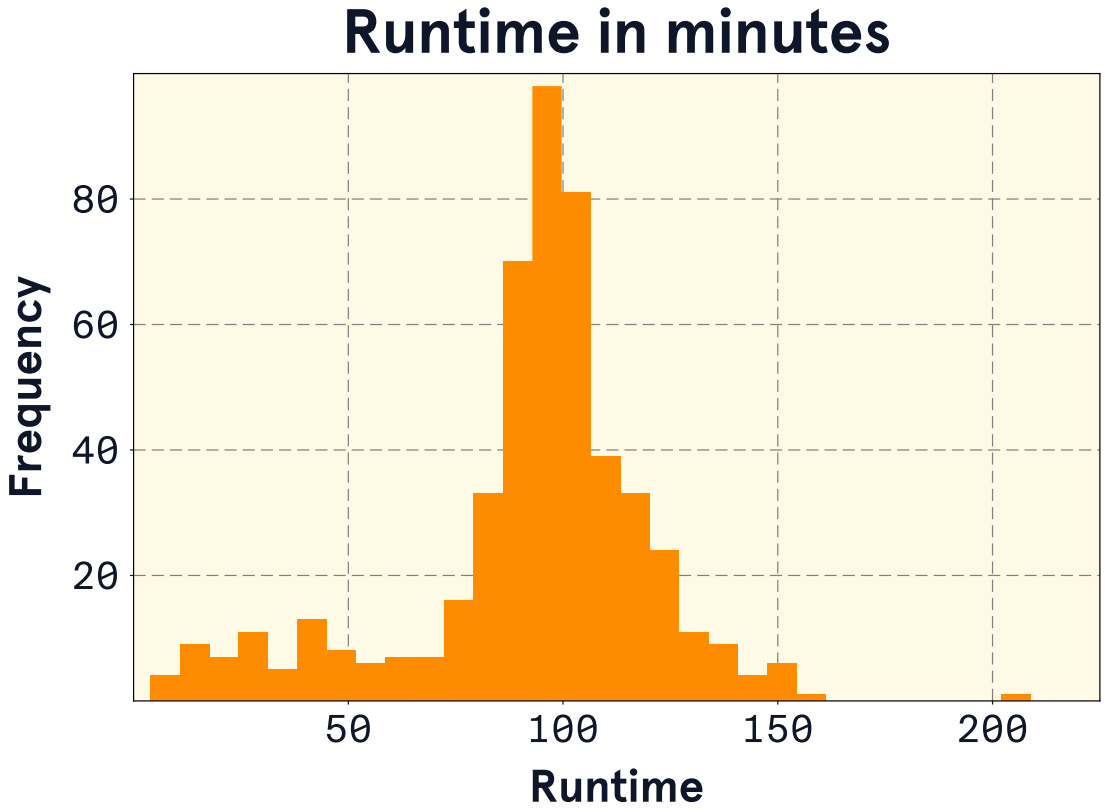
Idiom

mean μ (mu) and standard deviation σ (sigma).



A skewed distribution is asymmetrical with a steep change in frequency on one side and a flatter, trailing change in frequency on the other.

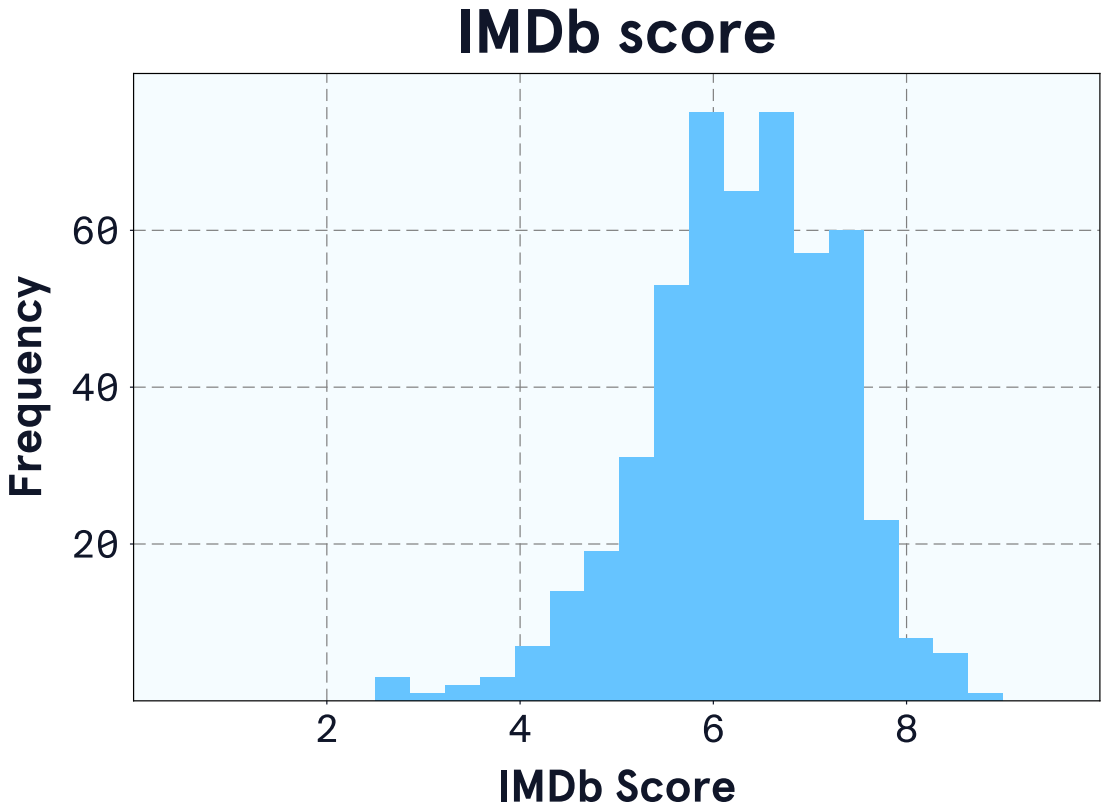
Specifically, the income distribution is right-skewed (also called positively-skewed) because the tail is on the right side



Based on the distribution plot, what concerns might you have with the default summary statistics used by the analytics software?

There are two aspects of this distribution plot that might lead to concern about using the mean and standard deviation:

1. The distribution is left-skewed — it has a long tail of low values on the left side. These values might influence the mean to be lower.
2. There is a single high value of just above 200 minutes. This value might be an outlier that influences the mean to be higher.



Using the plot and summary statistics, describe the distribution of IMDb scores.

The distribution of IMDb scores is mostly symmetrical in a bell shape, indicating a normal distribution. There are a couple of very low scores, but they are not far from the rest of the distribution, so they may not be extreme enough to be considered outliers. Since the distribution is fairly symmetrical, we can rely on the mean of 6.3 to give us a good idea of what a typical IMDb rating is. With a standard deviation of 1, we know there is some variation in scores, but most scores fall between 4 and 8 on the 1-10 scale.