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)

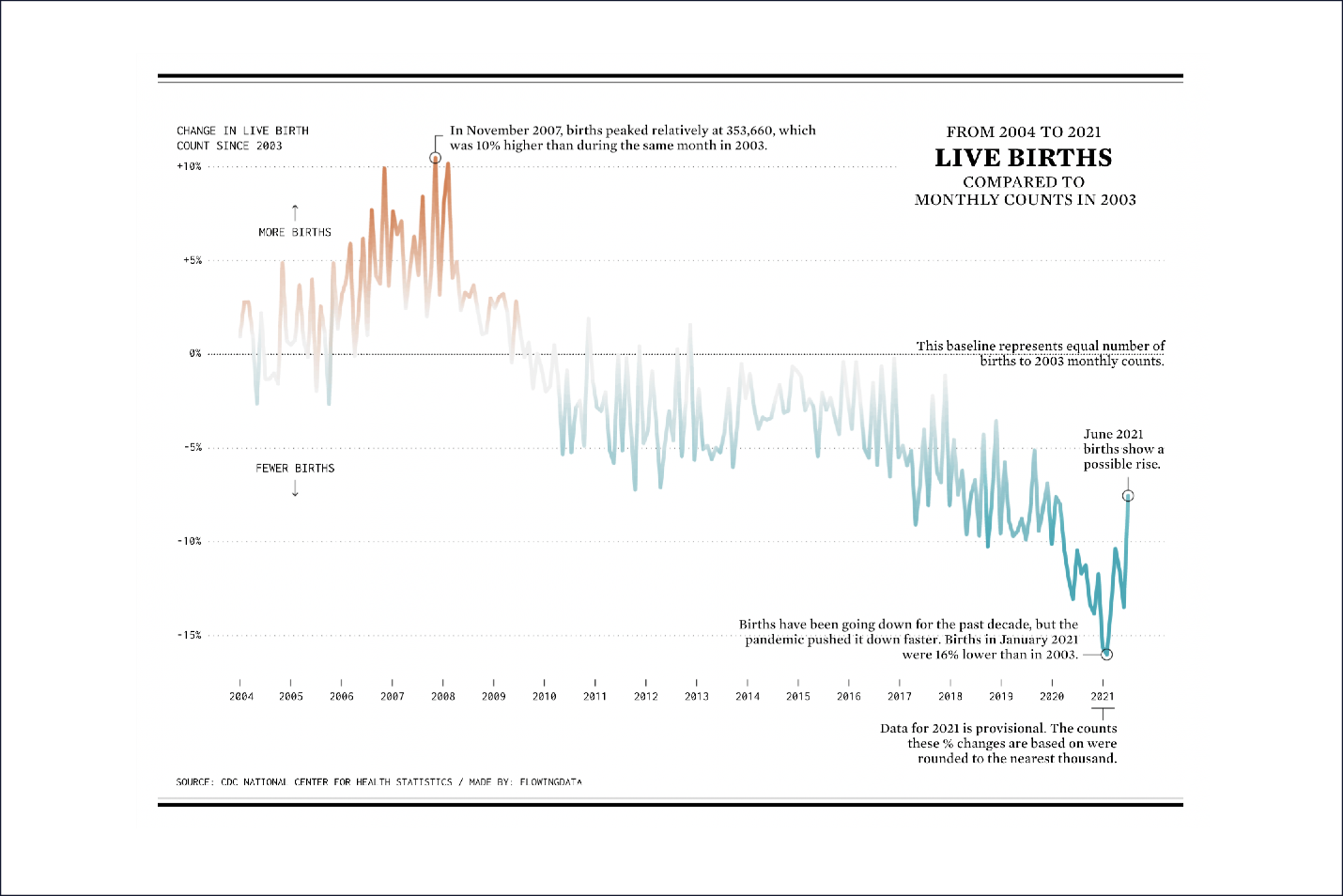
## Color Scales

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.

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.

تعریف: واریانس اندازه گیری پخش یا گسترش مقادیر یک متغیر در دیتاست است.

استفاده: واریانس برای درک پخش یا تنوع یک متغیر به کار می‌رود و معمولاً در آمار توصیفی و آزمون فرضیه استفاده می‌شود.

# **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.

تعریف: کوواریانس اندازه گیری تغییر مشترک دو متغیر است. این مقدار نشان می‌دهد که آیا یک متغیر افزایش می‌یابد هنگامی که متغیر دیگر افزایش می‌یابد یا کاهش می‌یابد هنگامی که متغیر دیگر کاهش می‌یابد.

استفاده: کوواریانس برای تحلیل رابطه بین دو متغیر به کار می‌رود. با این حال، ارزش معنایی کوواریانس به دلیل وابستگی آن به واحد متغیرها معمولاً تفسیر نمی‌شود.

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**.

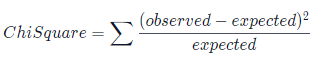
### **Chi-Square**

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.



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

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* ∣ *B*) = *P*(*A*) and *P*(*B* ∣ *A*) = *P*(*B*)​

### **Multiplication Rule**

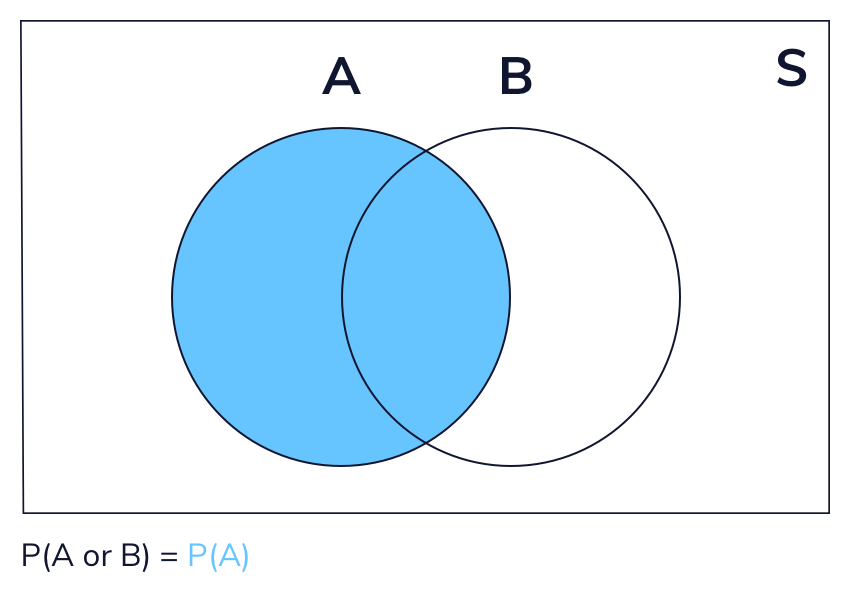
probability that two events happen simultaneously

*P*(*A* and *B*) = *P*(*A*) ⋅ *P*(*B* ∣ *A*)

##### For Independent Events is :

*P*(*A* and *B*) = *P*(*A*) ⋅ *P*(*B*)

### **Addition Rule** For not mutually exclusive

For not mutually exclusive : P(A 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.

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​ → *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*×*p 🡺* n, representing the number of events and p, representing the probability of “success”

**Variance of the Binomial Distribution**

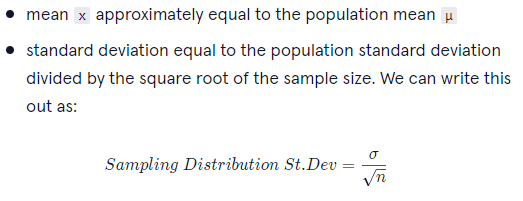
*Variance*(#*ofHeads*) = *Var*(*X*) = *n*×*p*×(1−*p*)

*Variance*(#*ofHeads*)=10×0.5×(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!



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:



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:



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*:



**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**

Itcan 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*

##### **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**

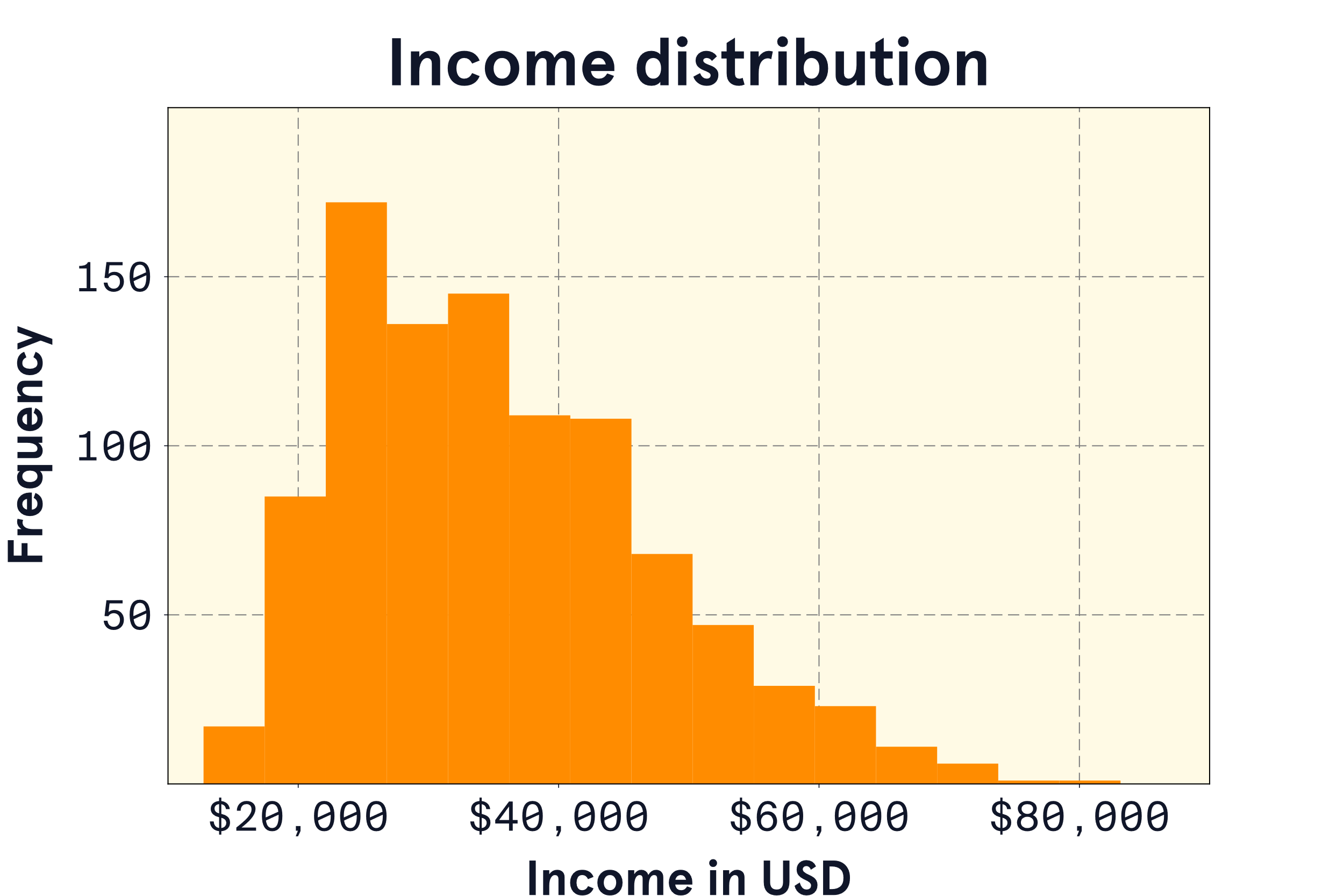
Focusing Question: How do variables compare to each other?

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.



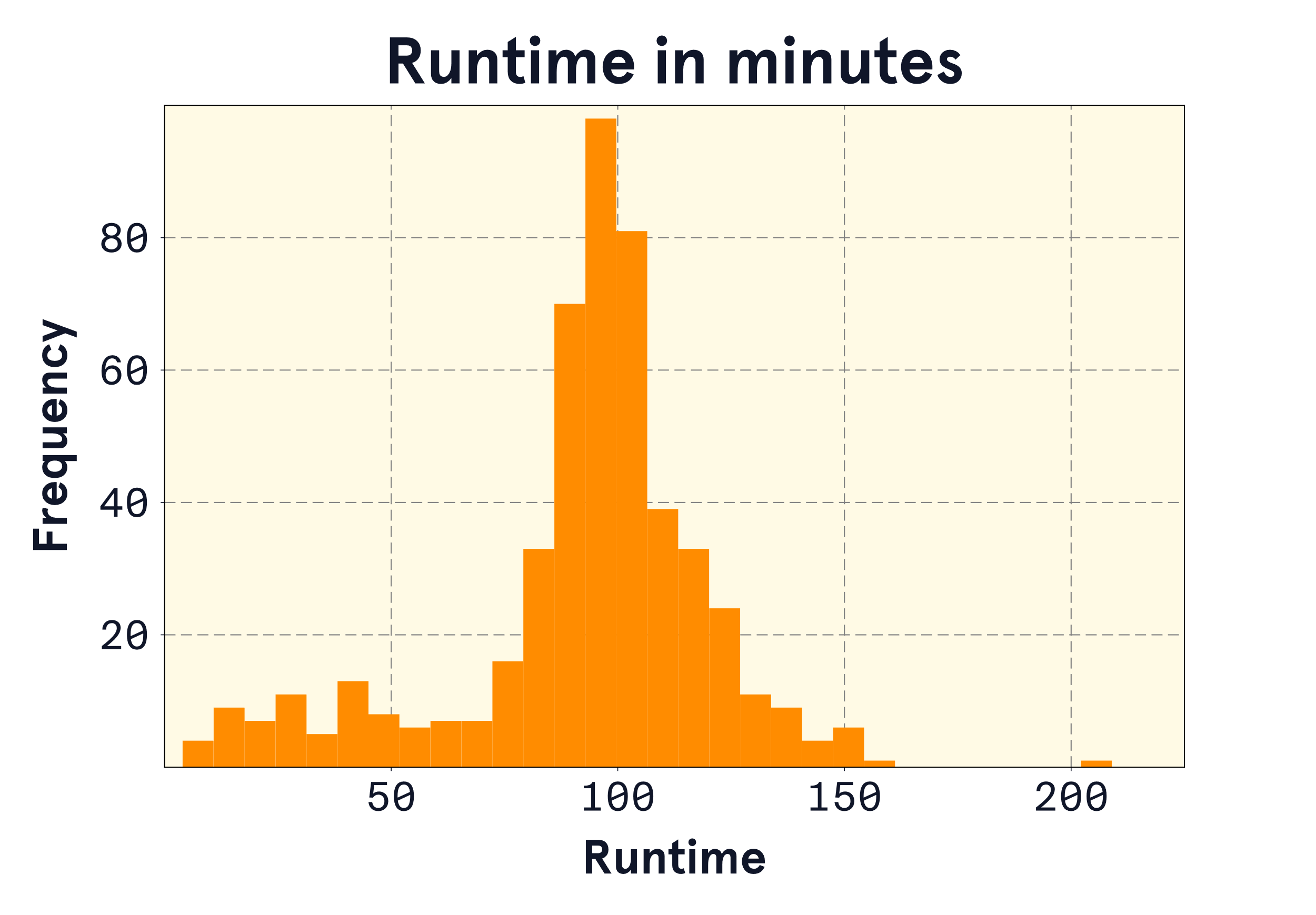
# **Idiom**

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.

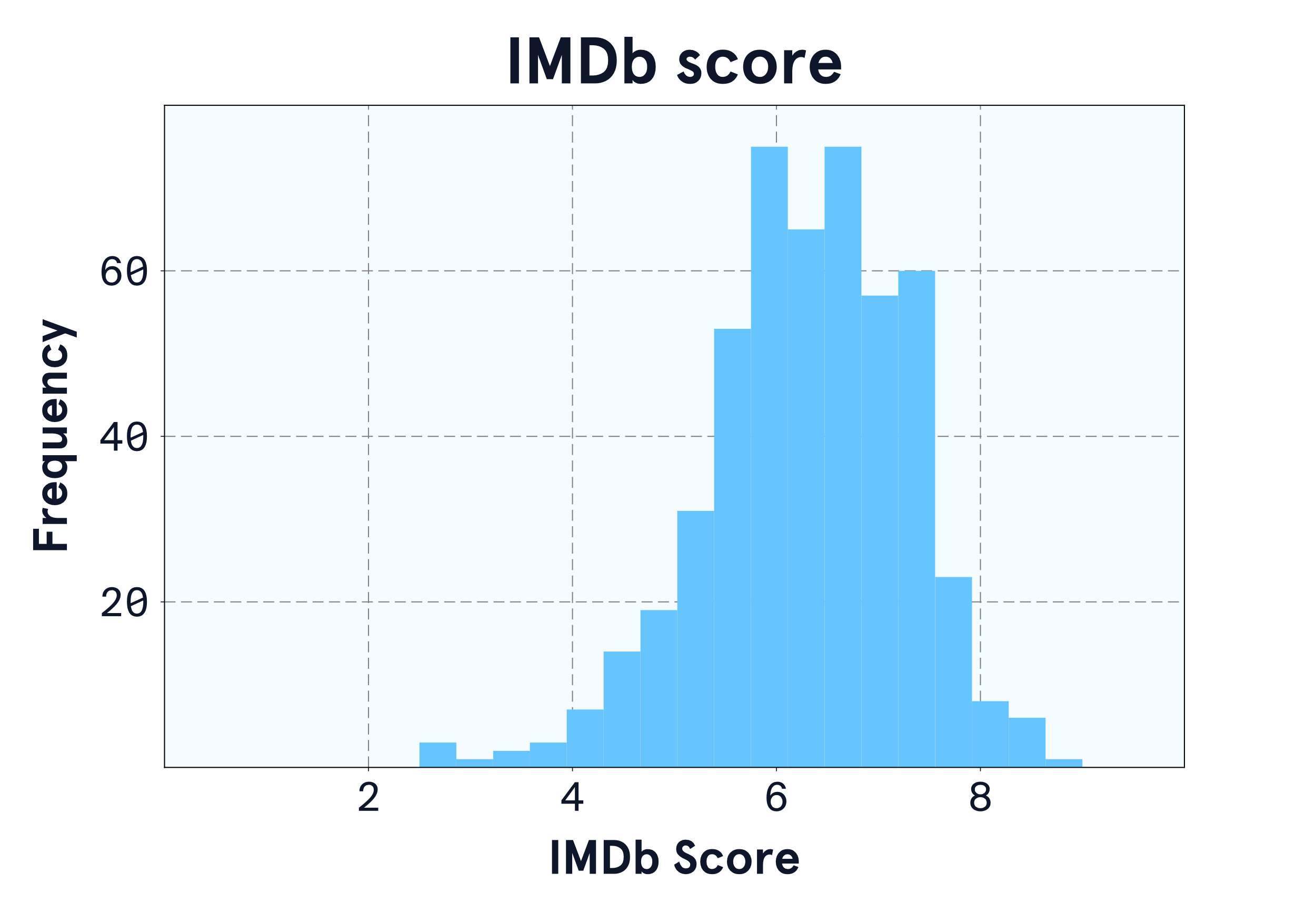
Specifically, the income distribution is right-skewed (also called positively-skewed) because the tail is on the right side



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.