Broadly, there are three different types of derived metrics:

1.    Type-driven metrics

2.    Business-driven metrics

3.    Data-driven metrics

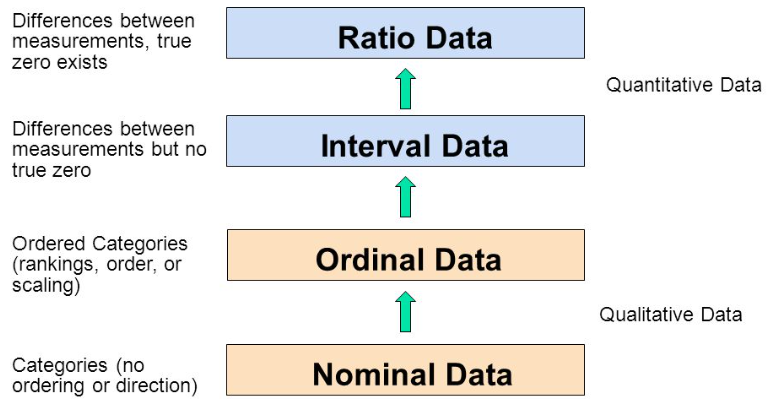
**Type-Driven Metrics**

These metrics can be derived by understanding the variable’s typology. You have already learnt one simple way of classifying variables/attributes — **categorical (ordered, unordered)** and **quantitative or numeric**. Similarly, there are various other ways of classification, one of which is Steven's typology.

Steven’s typology classifies variables into four types — nominal, ordinal, interval and ratio:

* **Nominal variables**: Categorical variables, where the categories **differ only by their names**; there is **no order**among categories, e.g. colour (red, blue, green), gender (male, female), department (HR, analytics, sales)
  + These are the most basic form of categorical variables
* **Ordinal variables**: Categories follow a certain **order**, but the **mathematical difference** **between categories is not meaningful**, e.g. education level (primary school, high school, college), height (high, medium, low), performance (bad, good, excellent), etc.
  + Ordinal variables are **nominal as well**
* **Interval variables**: Categories follow a certain order, and the **mathematical difference between categories is meaningful**but division or multiplication is not, e.g. temperature in degrees celsius ( the difference between 40 and 30 degrees C is meaningful, but 30 degrees x 40 degrees is not), dates (the difference between two dates is the number of days between them, but 25th May / 5th June is meaningless), etc.
  + Interval variables are**both nominal and ordinal**
* **Ratio variables**: Apart from the mathematical difference, the ratio (division/multiplication) is possible, e.g. sales in dollars ($100 is twice $50), marks of students (50 is half of 100), etc.
  + Ratio variables are **nominal, ordinal and interval type**

In the 1940s, Stanley Smith Stevens introduced four scales of measurement: nominal, ordinal, interval, and ratio. These are still widely used today as a way to describe the characteristics of a variable. Knowing the scale of measurement for a variable is an important aspect in choosing the right statistical analysis.



**Nominal**

A nominal scale describes a variable with categories that do not have a natural order or ranking. You can code nominal variables with numbers if you want, but the order is arbitrary and any calculations, such as computing a mean, median, or standard deviation, would be meaningless.

Examples of nominal variables include:

* genotype, blood type, zip code, gender, race, eye color, political party

**Ordinal**

An ordinal scale is one where the order matters but not the difference between values.

Examples of ordinal variables include:

* socio economic status (“low income”,”middle income”,”high income”), education level (“high school”,”BS”,”MS”,”PhD”), income level (“less than 50K”, “50K-100K”, “over 100K”), satisfaction rating (“extremely dislike”, “dislike”, “neutral”, “like”, “extremely like”).

Note the differences between adjacent categories do not necessarily have the same meaning. For example, the difference between the two income levels “less than 50K” and “50K-100K” does not have the same meaning as the difference between the two income levels “50K-100K” and “over 100K”.

[Make more informed and accurate analysis choices with Prism. Start your free Prism trial](http://www.graphpad.com/demos/).

**Interval**

An interval scale is one where there is order and the difference between two values is meaningful.

Examples of interval variables include:

* temperature (Farenheit), temperature (Celcius), pH, SAT score (200-800), credit score (300-850).

**Ratio**

A ratio variable, has all the properties of an interval variable, and also has a clear definition of 0.0. When the variable equals 0.0, there is none of that variable.

Examples of ratio variables include:

* enzyme activity, dose amount, reaction rate, flow rate, concentration, pulse, weight, length, temperature in Kelvin (0.0 Kelvin really does mean “no heat”), survival time.

When working with ratio variables, but not interval variables, the ratio of two measurements has a meaningful interpretation. For example, because weight is a ratio variable, a weight of 4 grams is twice as heavy as a weight of 2 grams. However, a temperature of 10 degrees C should not be considered twice as hot as 5 degrees C. If it were, a conflict would be created because 10 degrees C is 50 degrees F and 5 degrees C is 41 degrees F. Clearly, 50 degrees is not twice 41 degrees.  Another example, a pH of 3 is not twice as acidic as a pH of 6, because pH is not a ratio variable.

**🔹 Similarities Across Derived Metrics Methods**

All derived metric methods:

* **Use existing data** to generate new insights.
* **Summarize trends and relationships** more effectively.
* **Help in decision-making** by providing actionable insights.

However, how they’re calculated and their use cases differ. Let’s break it down:

**1️⃣ Ratio-Based Metrics (Proportions & Rates)**

* **Definition:** These metrics express relationships between two quantities, usually as a fraction or percentage.
* **Example:** **Batting Average in Cricket** Batting Average=Total Runs ScoredTotal Times Out\text{Batting Average} = \frac{\text{Total Runs Scored}}{\text{Total Times Out}}Batting Average=Total Times OutTotal Runs Scored​
* **Use Case:** Helps compare performance across different conditions (e.g., strike rate in T20 vs. Test cricket).

✅ **Why it's useful?**

* Helps normalize data across different scales.
* Useful for **comparison** (e.g., comparing customer retention rates across different regions).

**2️⃣ Aggregated Metrics (Summarized Totals)**

* **Definition:** These metrics are simple summaries like sums, means, medians, and percentiles.
* **Example:** **Total Revenue Generated per Month** Total Revenue=∑(Sales from all transactions in a month)\text{Total Revenue} = \sum (\text{Sales from all transactions in a month})Total Revenue=∑(Sales from all transactions in a month)
* **Use Case:** Useful for spotting overall trends (e.g., **monthly sales trends**, **total runs scored in IPL**).

✅ **Why it's useful?**

* Provides a **big-picture** view.
* Helps **track growth and trends** over time.

**3️⃣ Indexed Metrics (Relative to a Baseline)**

* **Definition:** These metrics compare data relative to a baseline (such as an industry average or a previous period).
* **Example:** **Consumer Price Index (CPI)** compares today’s prices to a base year. Index Value=(Current ValueBase Value)×100\text{Index Value} = \left( \frac{\text{Current Value}}{\text{Base Value}} \right) \times 100Index Value=(Base ValueCurrent Value​)×100
* **Use Case:** Used for benchmarking (e.g., **team performance in IPL compared to league average**).

✅ **Why it's useful?**

* Helps identify **relative improvements or declines**.
* Useful for **trend analysis and seasonality** (e.g., stock market indices, sports rankings).

**🔍 Key Differences Summary**

| **Derived Metric Type** | **What It Measures** | **Example** |
| --- | --- | --- |
| **Ratio-Based** | Relationship between two values | **Conversion Rate** = Leads / Visitors |
| **Aggregated** | Overall summary of data | **Total Sales**, **Total Runs in IPL** |
| **Indexed** | Relative comparison to a baseline | **Inflation Index**, **Team Performance Index** |

**Difference Between Univariate and Bivariate Analysis**

| **Feature** | **Univariate Analysis** | **Bivariate Analysis** |
| --- | --- | --- |
| **Definition** | Analysis of a **single variable** | Analysis of **two variables** to find relationships |
| **Purpose** | Summarizes and describes data distribution | Identifies relationships or dependencies between variables |
| **Examples of Techniques** | **Histograms, Box plots, Frequency tables, Mean, Median, Mode, Variance** | **Scatter plots, Correlation, Regression, Heatmaps, Two-way tables** |
| **Example Scenario** | Analyzing the distribution of students' test scores | Examining the relationship between students' test scores and study hours |
| **Visualization** | **Histogram, Boxplot, Bar chart, Pie chart** | **Scatter plot, Line plot, Heatmap** |

For this exercise, we will use the news popularity data set again which is a set of articles published by the digital media website Mashable over a period of two years.

There are more than 60 attributes in the data set which describe the articles, such as the **total number of words**, the number of **positive** and **negative** **words**, the **length of the title**, the article **genre** (e.g. business, travel or entertainment) and, more importantly, the output variable, i.e. the number of **shares**.

In the following questions, you will attempt to understand the **distribution of shares** (the number of times an article was shared on the internet) using measures of spread. Suppose you work for a content marketing company as an analyst and want to understand the ‘distribution of shares'. Based on your understanding, you will share with your peers some rough metrics such as the **representative number of shares**, the **spread of shares**, etc.

categorical (ordered / unordered) and quantitative (or numeric). In this segment, you will learn how to conduct univariate analysis on unordered categorical variables.

It is important to note that **rank-frequency plots** enable you to extract meaning even from seemingly trivial **unordered categorical variables** such as country, name of an artist, name of a github user etc.

Roughly speaking, outliers are abnormally large or small values. There is no one fixed rule in deciding outliers. However, you would have noticed that going from the 99th to the 100th percentile, there is an approximately 25x increase in the number of shares.

You may have also noted that in the lower quartiles, there is only about 5% increase in the number of shares per percentile. This increases to about 10% per percentile in the higher quartiles and 20% beyond the 95th percentile. In this case, some articles in the top quartiles are clearly outliers.

To classify some articles as outliers, you may need to consult your client or the business to understand the reasons behind abnormal values. In some cases, they are justifiable, whereas in others they cannot be explained and are thus labelled as outliers.

For example, let’s say that you decide to label all the articles beyond the **95th percentile** as outliers since you observe a roughly 20% increase in the number of shares beyond that. Thus, the last article you include will be the one at the 95th percentile. You remove every article beyond that point from the data set.

Here is a short explanation to [remove outliers in Python](https://www.kdnuggets.com/2017/02/removing-outliers-standard-deviation-python.html) (this method uses the concept of standard deviation). Also, you can take a look at this[StackOverflow answer](https://stackoverflow.com/questions/35827863/remove-outliers-in-pandas-dataframe-using-percentiles)(this uses the concept of percentiles which is required for solving the question below). And if you want to get a general idea about using percentiles in Python, take a look [here](https://docs.scipy.org/doc/numpy-1.15.1/reference/generated/numpy.percentile.html#numpy.percentile) and [here](https://docs.scipy.org/doc/numpy-1.15.1/reference/generated/numpy.quantile.html#numpy.quantile).

Let’s say you work at a **central government body** and want to compare the literacy rates of people across all the states in India. Your task is to understand the meanings of the rows and the attributes and find the number of literates, illiterates, literacy rate, etc. You have already downloaded the **Census Data** for the data sourcing exercise. In case you have not downloaded yet, do so from the link [here](https://cdn.upgrad.com/UpGrad/temp/1dcf9310-2223-4242-a8f9-807360f659d2/EDA_census.xlsx) before attempting the following questions.

Open the file in Excel and try to understand the column names. For this exercise, you will only need the following columns:

1. **C1-C4**: The first four columns are self-explanatory
2. **C5**: Total/Rural/Urban population
3. **C6**: Age group (e.g. age- group = 7 would mean the number of people of age 7 in each column)
4. **C7-C9**: Total population in that row
5. **C10-C12**: Illiterate population in that row
6. **C13-C15**: Literate population in that row

Delete the **top few rows** to convert it into a standard format with appropriate column names (you may have to rename some columns manually). Then, convert the file into **a CSV** and import into Python. Answer the following questions.

The data set provided [here](https://cdn.upgrad.com/UpGrad/temp/05e69f76-61d8-4774-8dec-78dad8015018/grades.csv) contains information on the dates and times of students’ submissions of an assignment. It has two variables — “**submission**” (the unique URL of a student’s submission on the platform) and “**submit\_time**” (the date-time of submission).

This was the Association Rule Mining assignment whose submission deadline was Jan 3, 2017 - 11:59:59 PM. The second deadline was Jan 9, 2017 - 11:59 PM. Submissions between the first and the second deadline attract a 30% penalty in marks.

The variable ‘submission’ has several components separated by slashes. The last component contains the submitted filename in the format ‘roll\_number.xxx’ where xxx is the **file extension (.zip, .R etc.)**

You want to understand:

1. The typical dates when students submit assignments
2. The typical times (hour of the day) of submissions

Extract all the relevant metrics (such as the date, day, month, year, hour, and minutes) of each person’s submission and attempt the following questions.