CSE 435: Pattern Recognition
Lecture 3-4
Exploratory Data Analysis

Tanvir Azhar
Lecturer, East Delta University

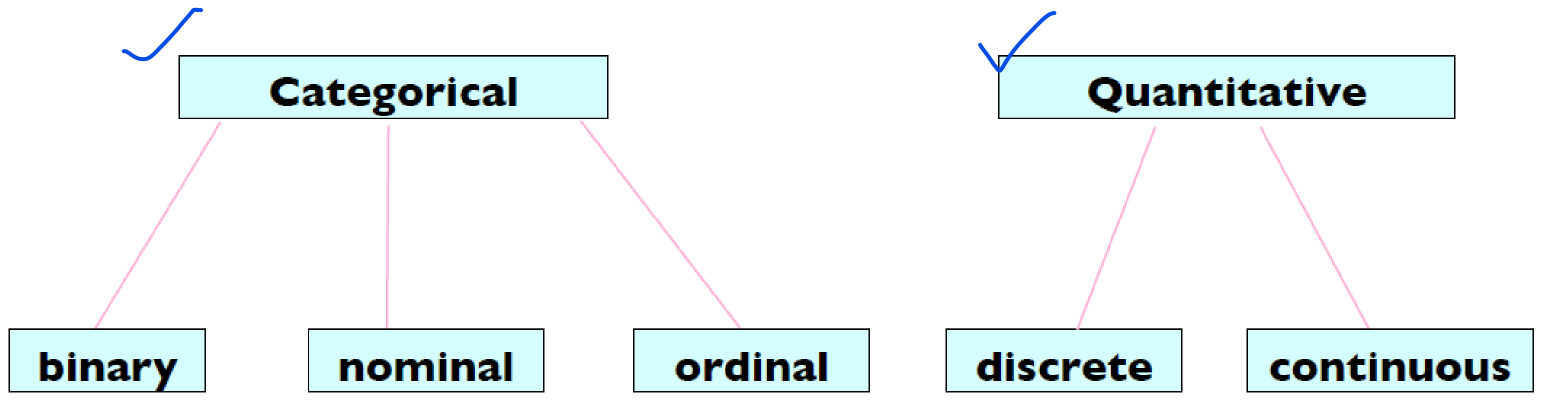
Outline

- Exploratory Data Analysis
  - Chart types
  - Some important Stat Review
  - Hypothesis Testing

Descriptive vs. Inferential Statistics

- **Descriptive:** e.g., Median; describes data you have but can't be generalized beyond that
  - We'll talk about Exploratory Data Analysis
- **Inferential:** e.g., t-test, that enable inferences about the population beyond our data
  - These are the techniques we'll leverage for Machine Learning and Prediction

Types of Data



Categorical

Quantitative

binary

nominal

ordinal

discrete

continuous

2 categories

more categories

order matters

numerical

uninterrupted

Dimensionality of Data Sets

- **Univariate:** Measurement made on one variable per subject

- **Bivariate:** Measurement made on two variables per subject

- **Multivariate:** Measurement made on many variables per subject

Examples of Business Questions

- **Simple (descriptive) Stats**
  - “Who are the most profitable customers?”
- **Hypothesis Testing**
  - “Is there a difference in value to the company of these customers?”
- **Segmentation/Classification**
  - What are the common characteristics of these customers?
- **Prediction**
  - Will this new customer become a profitable customer? If so, how profitable?

adapted from Provost and Fawcett, “Data Science for Business”

Applying techniques

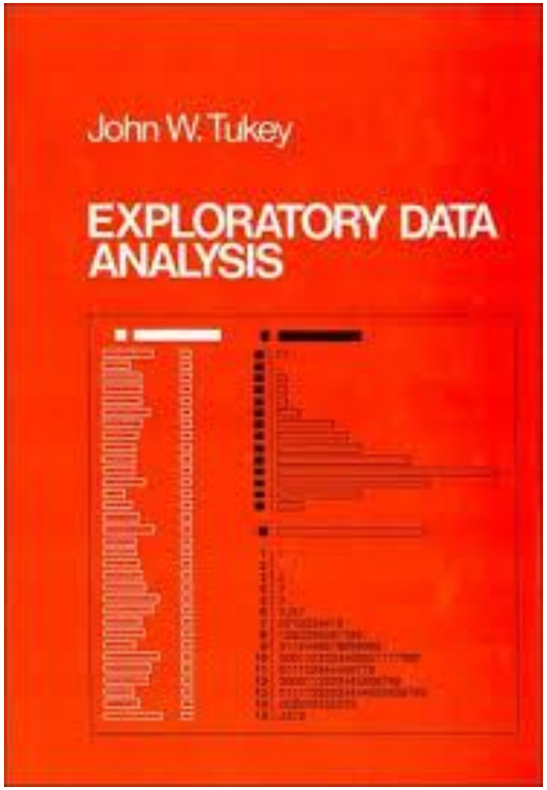
- Most business questions are causal: what would happen if? (e.g. I show this ad)
- But its easier to ask correlational questions, (what happened in this past when I showed this ad).
- **Supervised Learning:**
  - Classification and Regression
- **Unsupervised Learning:**
  - Clustering and Dimension reduction
- Note: Unsupervised Learning is often used inside a larger Supervised learning problem.
  - E.g. auto-encoders for image recognition neural nets.

Applying techniques

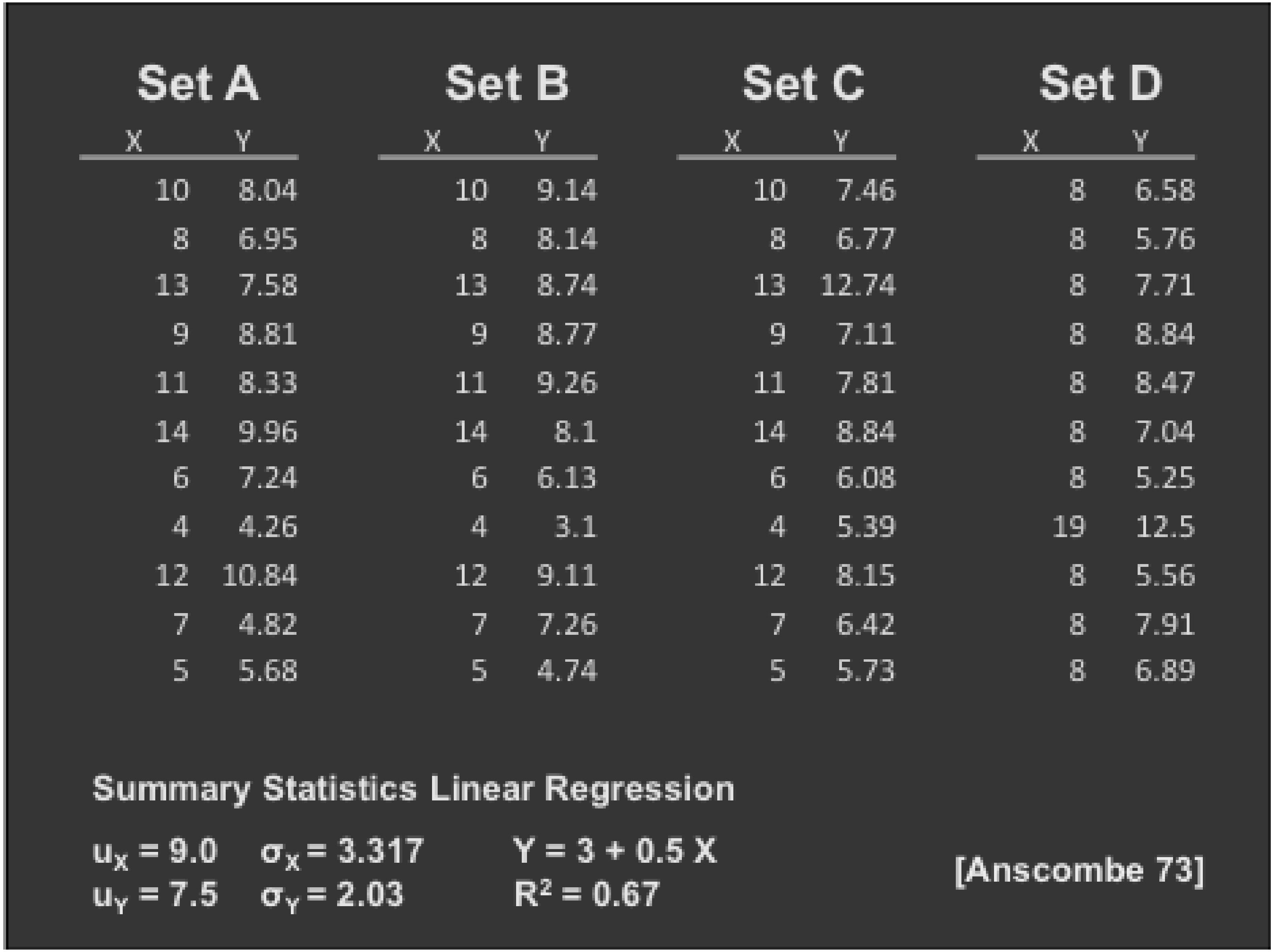
- **Supervised Learning:**
  - kNN (k Nearest Neighbors)
  - Naïve Bayes
  - Logistic Regression
  - Support Vector Machines
  - Random Forests
- **Unsupervised Learning:**
  - Clustering
  - Factor analysis
  - Latent Dirichlet Allocation

Exploratory Data Analysis 1977

- Based on insights developed at Bell Labs in the 60’s
- Techniques for visualizing and summarizing data
- What can the data tell us? (in contrast to “confirmatory” data analysis)
- Introduced many basic techniques:
  - 5-number summary, box plots, stem and leaf diagrams,…
- 5 Number summary:
  - extremes (min and max)
  - median & quartiles
  - More robust to skewed & longtailed distributions



The Trouble with Summary Stats



Set A		Set B		Set C		Set D	
X	Y	X	Y	X	Y	X	Y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.11	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

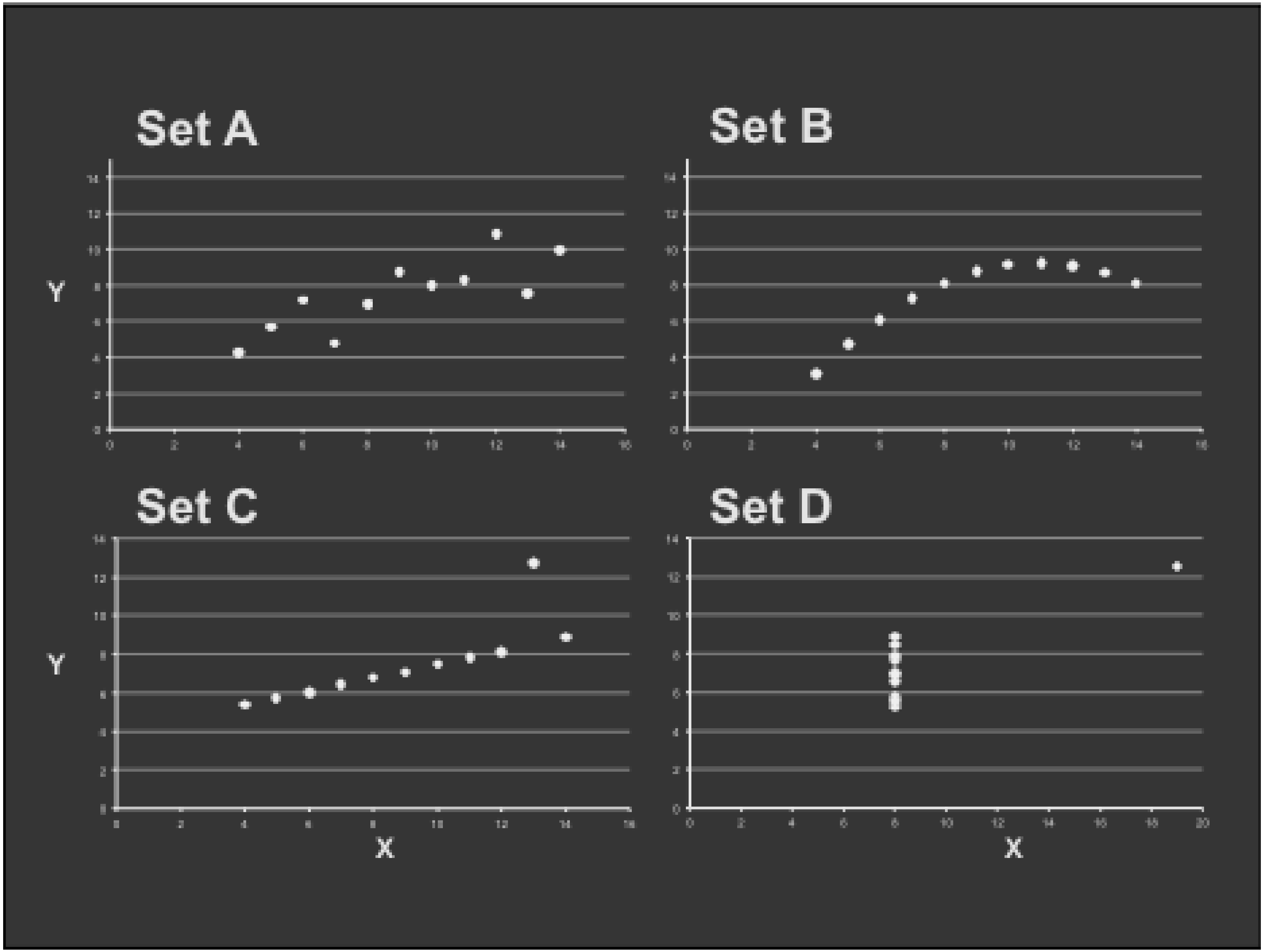
Summary Statistics Linear Regression

$u_X = 9.0$ $\sigma_X = 3.317$ $Y = 3 + 0.5\,X$

$u_Y = 7.5$ $\sigma_Y = 2.03$ $R^2 = 0.67$

[Anscombe 73]

Looking at Data



Data Presentation

- Data Art

facebook
December 2010

Chart types

- Single variable
  - Dot plot
  - Histogram
  - Error bar plot
  - Box-and-whisker plot
  - KDE Plot
  - Distribution Plot

Chart types

- **Count Plot:** Countplot is basically a count of frequency plot in form of a bar graph. It plots the count of each category in a separate bar. When we use the pandas' value counts function on any column, It is the same visual form of the value counts function. Mostly used for Categorial Data (Univariate Analysis)

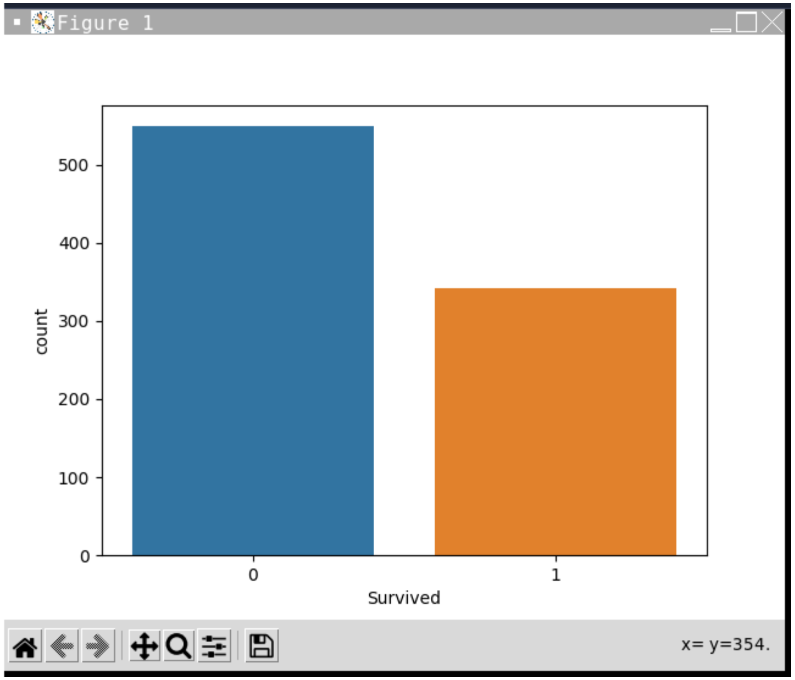
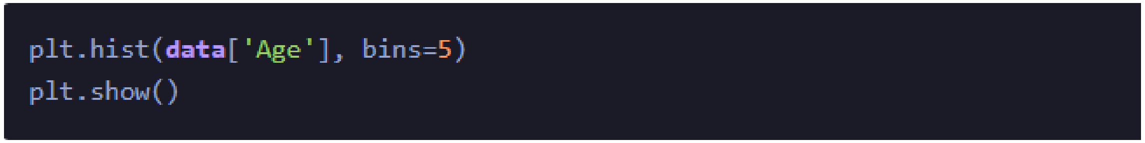


Chart types

- **Histogram:**

A histogram is a value distribution plot of numerical columns. It basically creates bins in various ranges in values and plots it where we can visualize how values are distributed. We can have a look where more values lie like in positive, negative, or at the center(mean). We use histogram to analyse numerical data.



```
plt.hist(data['Age'], bins=5)
plt.show()
```

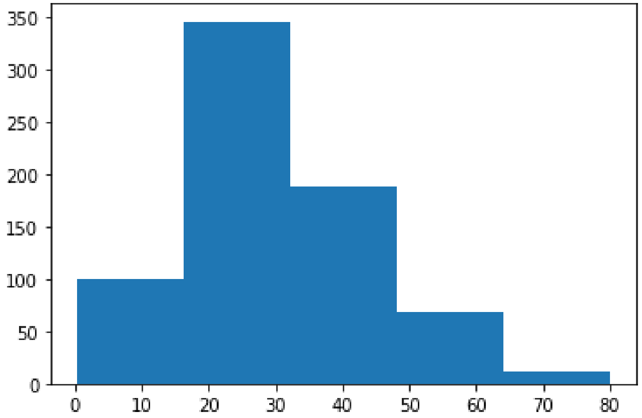
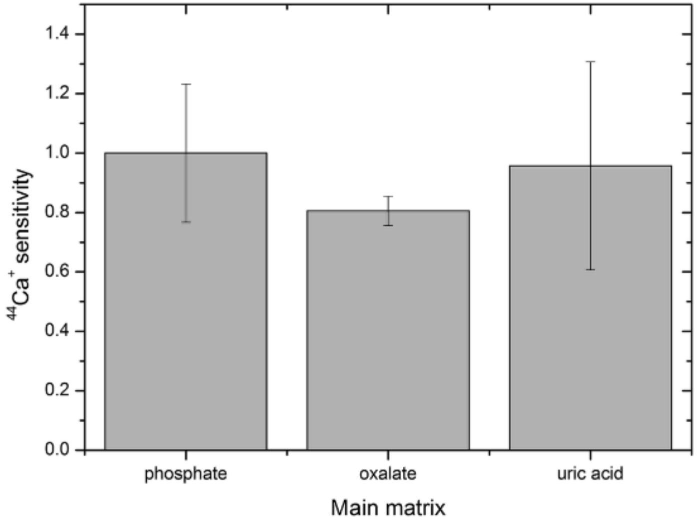


Chart types

- **Error bars:** An error bar is a line through a point on a graph, parallel to one of the axes, which represents the uncertainty or variation of the corresponding coordinate of the point.



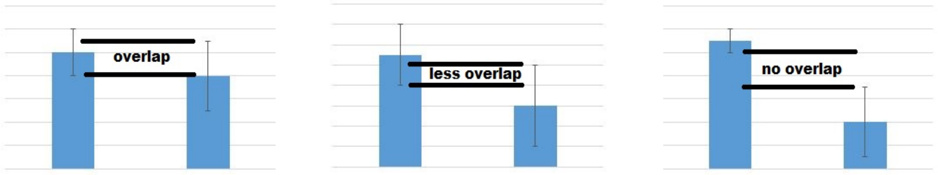


Chart types

- **Box-and-whisker plot** : a graphical form of 5-number summary (Tukey)

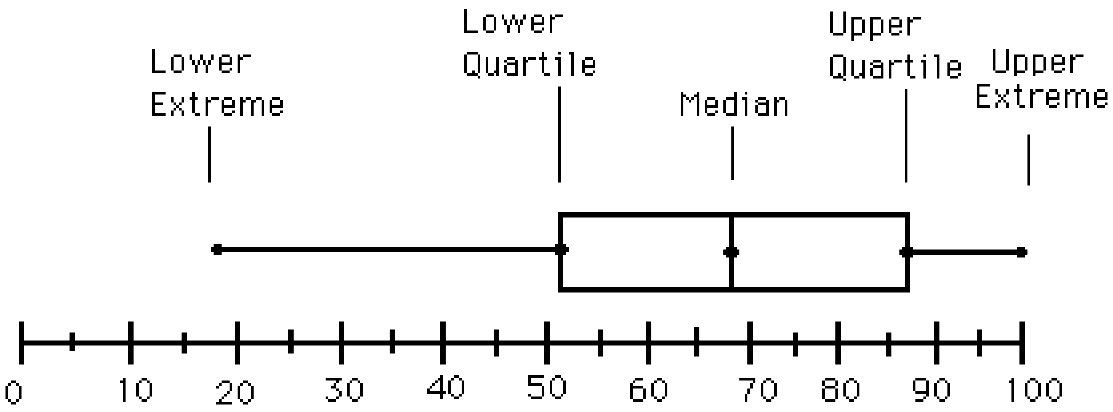
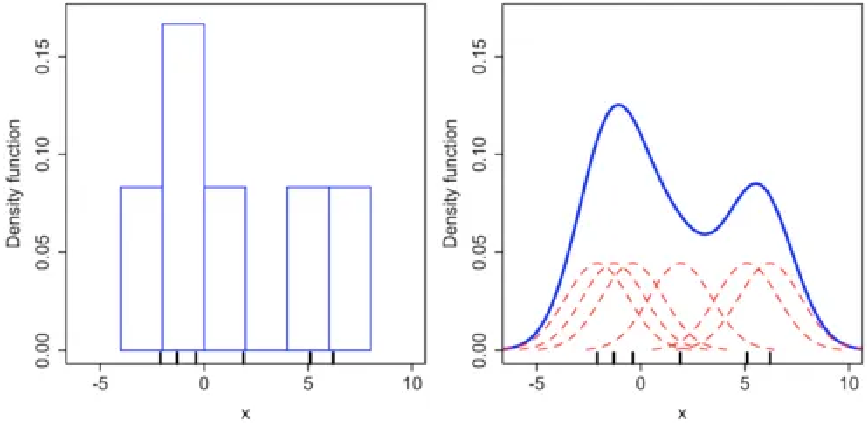


Chart types

- **KDE Plot:** **Kernel Density Estimation often referred to as KDE is a technique that lets you create a smooth curve given a set of data.**



Comparison of histogram and kernel function.

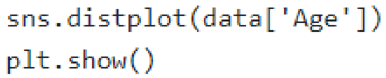
Chart types

- Histogram and Kernel Density Estimates
    - Histogram
        - Proper selection of bin width is important
        - Outliers should be discarded
    - KDE (like a smooth histogram)
        - Kernel function
            - Box, Epanechnikov, Gaussian
        - Kernel bandwidth

Chart types

- **Distrubution Plot**

- Distplot is also known as the second Histogram because it is a slight improvement version of the Histogram. Distplot gives us a KDE(Kernel Density Estimation) over histogram which explains PDF(Probability Density Function) which means what is the probability of each value occurring in this column. If you have study statistics before then definitely you should know about PDF function.



```
sns.distplot(data['Age'])
plt.show()
```

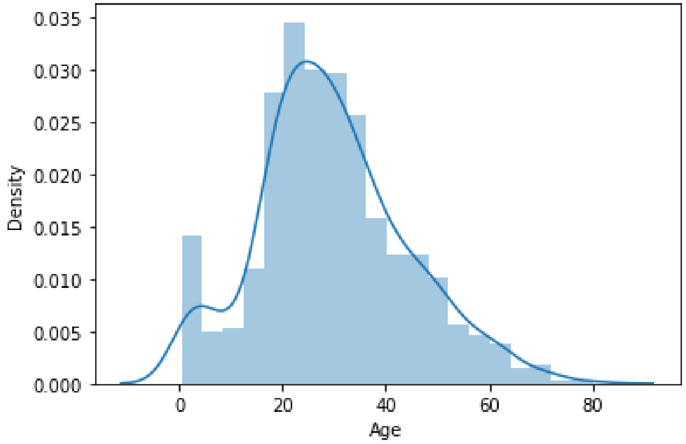


Chart types

- Two variables
  - Bar chart
  - Scatter plot
  - Line plot
  - Log-log plot

Chart types

- **Bar plot:** one variable is discrete

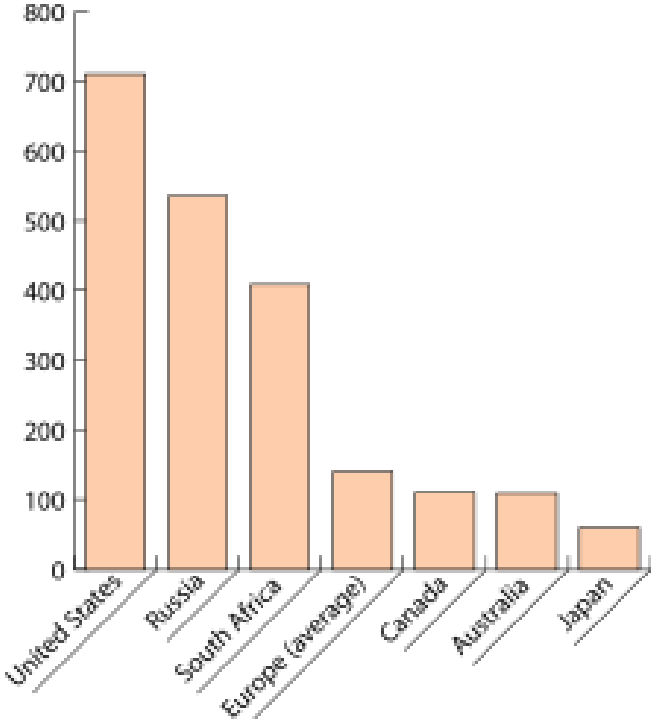
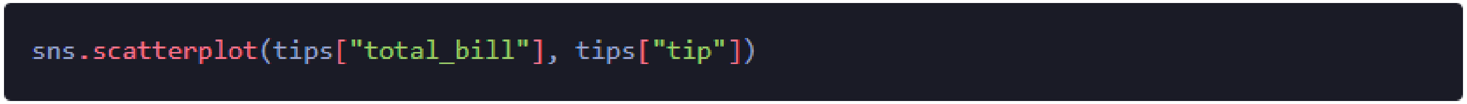


Chart types

- **Scatter plot**

To plot the relationship between two numerical variables scatter plot is a simple plot to do. Let us see the relationship between the total bill and tip provided using a scatter plot.



```
sns.scatterplot(tips["total_bill"], tips["tip"])
```

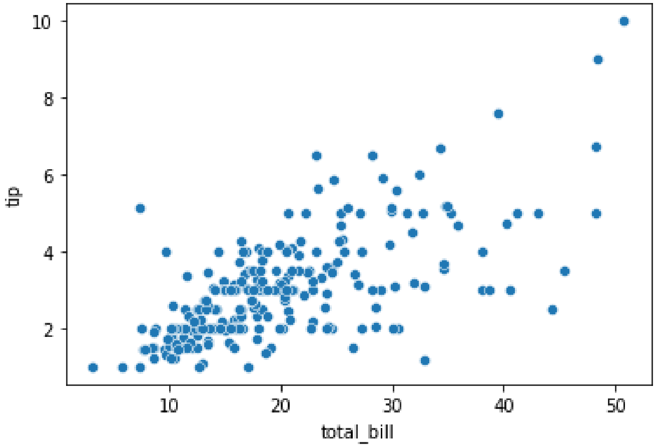
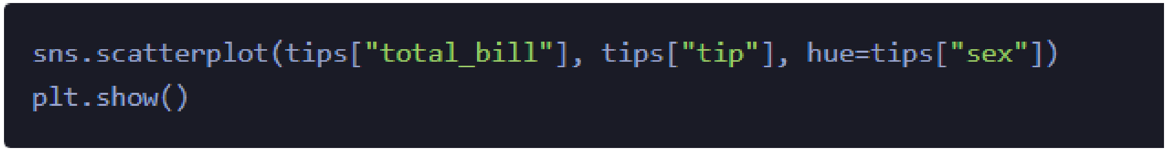


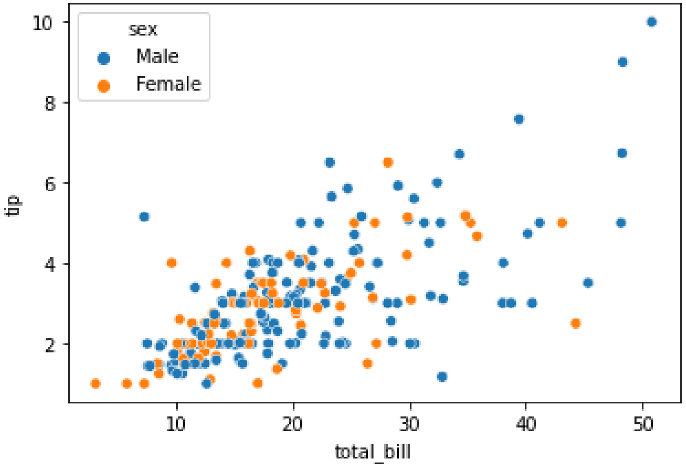
Chart types

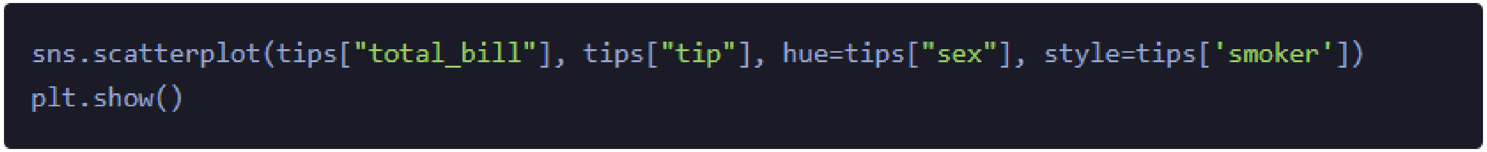
- **Scatter plot**

It can be used for both bi-variate and multivariate analysis



```
sns.scatterplot(tips["total_bill"], tips["tip"], hue=tips["sex"])
plt.show()
```





```
sns.scatterplot(tips["total_bill"], tips["tip"], hue=tips["sex"], style=tips['smoker'])
plt.show()
```

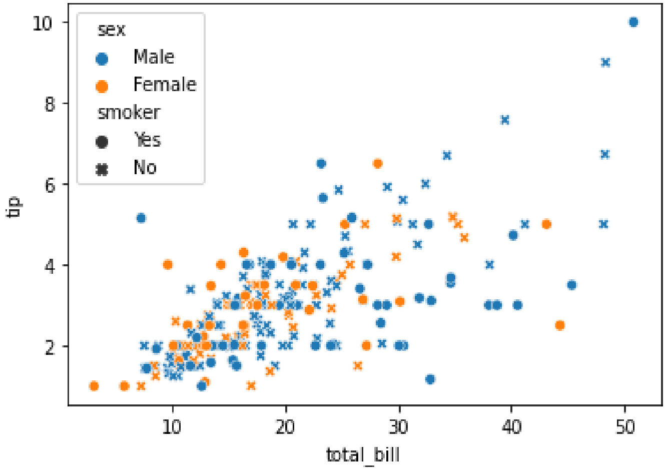


Chart types

- **Line plot**

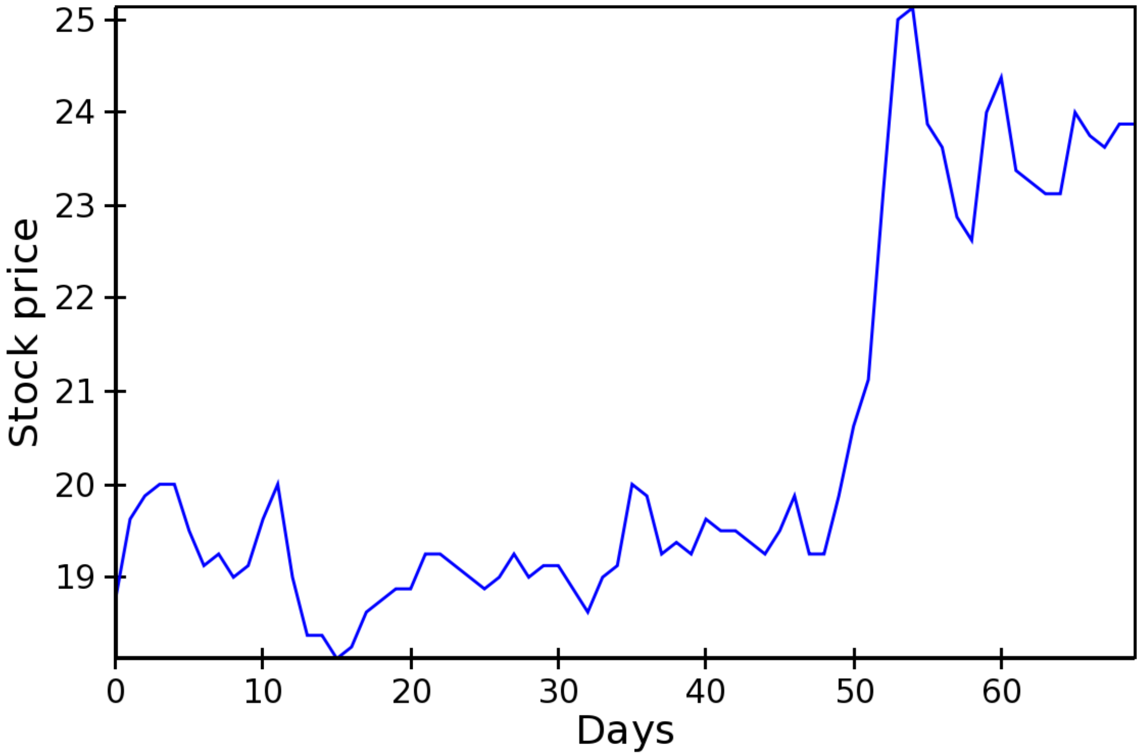
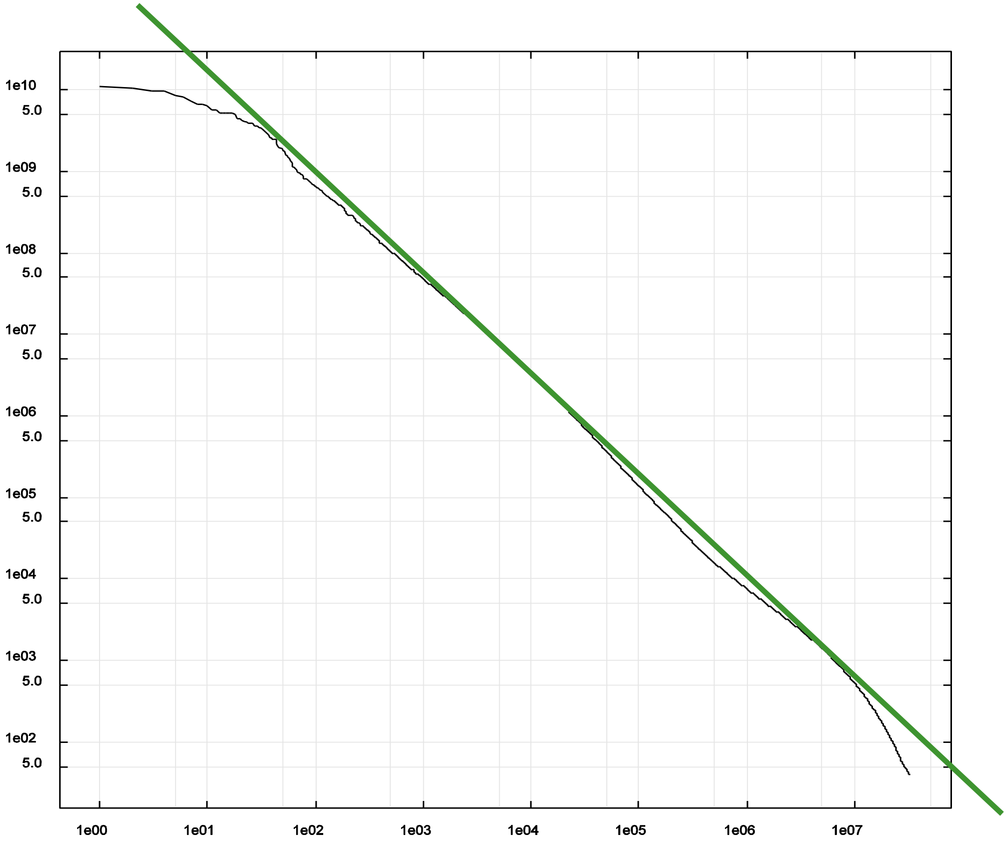


Chart types

- **Log-log plot:** Very useful for power law data

Frequency of words in tweets



slope ~ -1

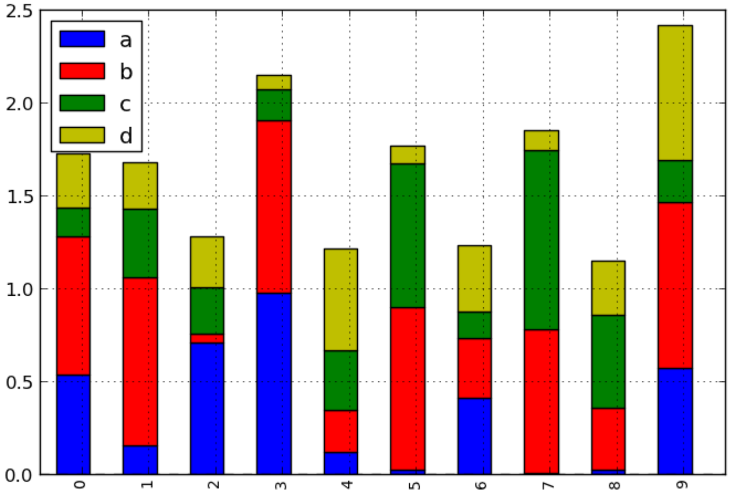
Rank of words in tweets, most frequent to least:
I, the, you,…

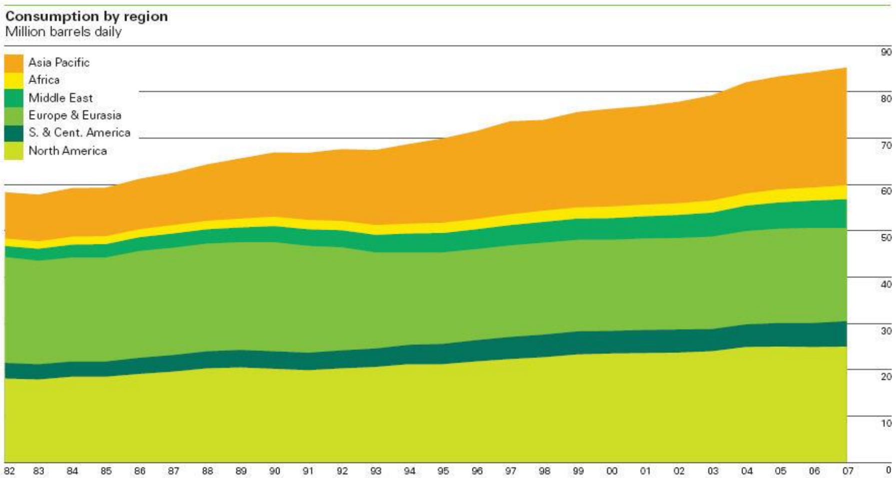
Chart types

- More than two variables
  - Stacked plots
  - Parallel coordinate plot
  - Box Plot

Chart types

- **Stacked plot:** stack variable is discrete:

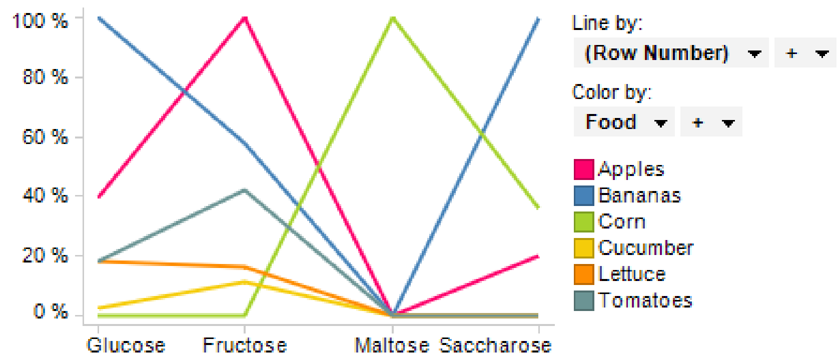




World oil consumption rose by about 1mmb/d in 2007, just below the 10-year average. OECD consumption declined nearly 400,000b/d. China accounted for the largest increment to consumption even though the growth rate was below average. Consumption in oil exporting regions was robust.

Chart types

- **Parallel coordinate plot:** one discrete variable, an arbitrary number of other variables:



Line by: (Row Number)

Color by: Food

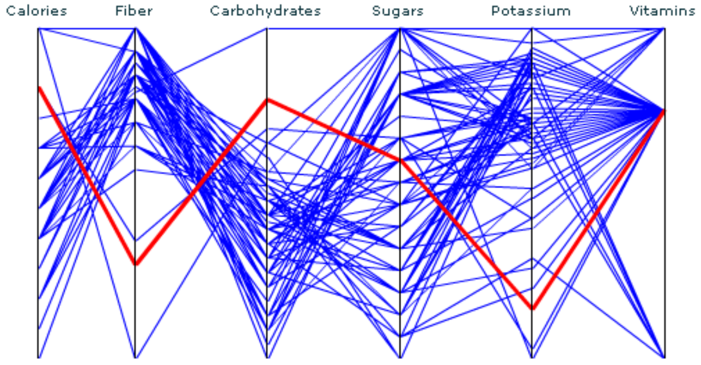
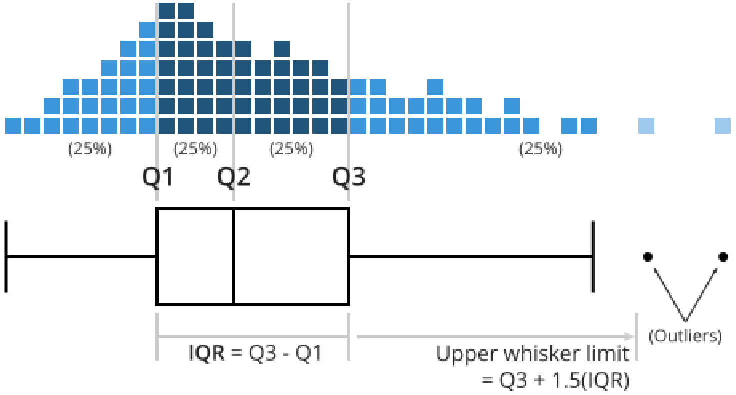


Chart types

- **Box plot:** Box plots are used to show distributions of numeric data values, especially when you want to compare them between multiple groups.



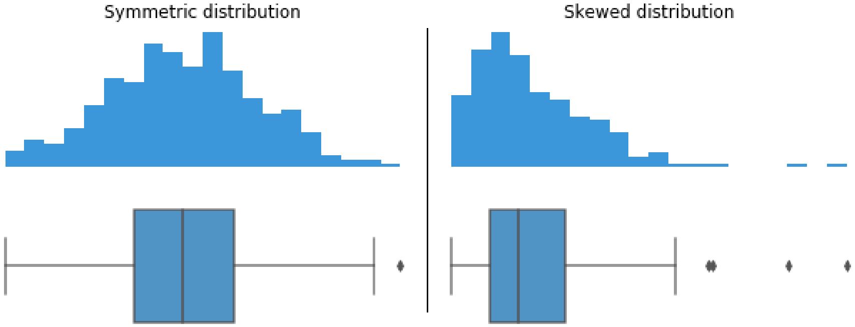
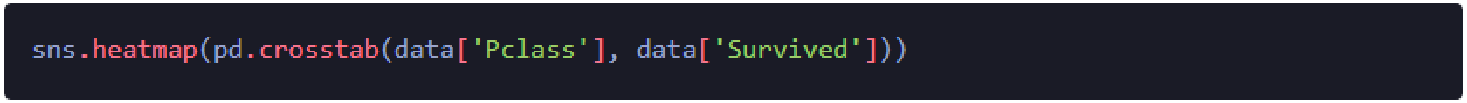
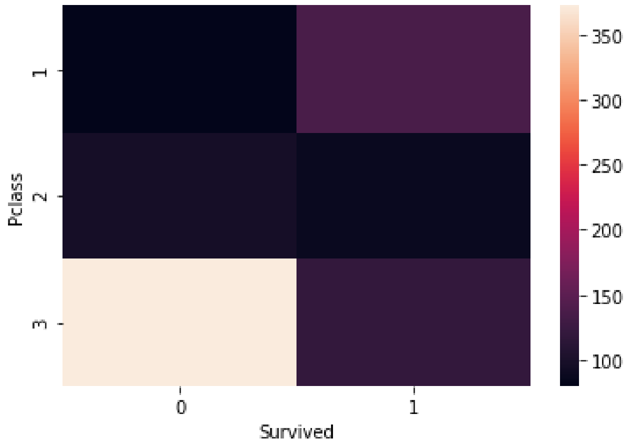


Chart types

- **Heatmap:** It basically shows that how much presence of one category concerning another category is present in the dataset.



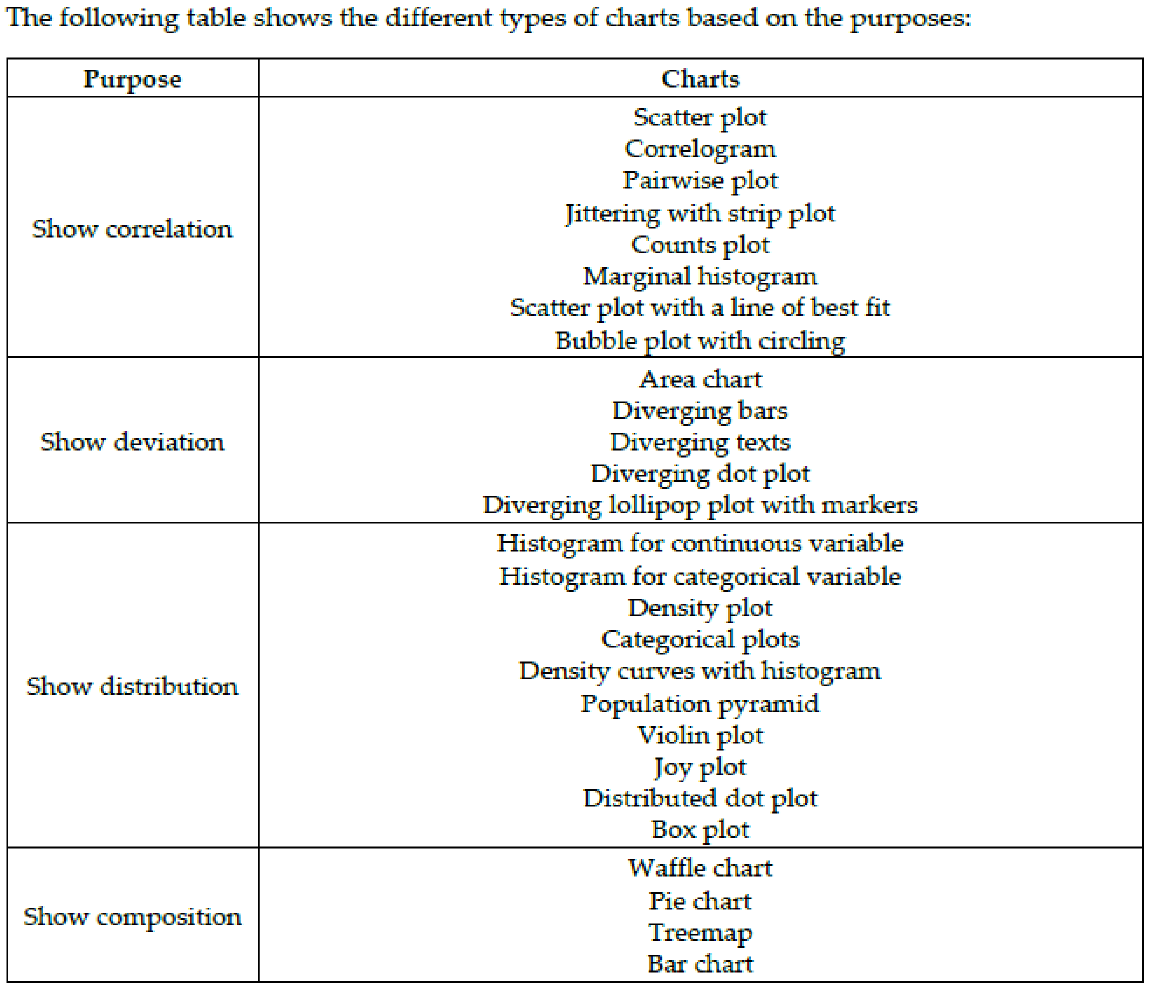
```
sns.heatmap(pd.crosstab(data['Pclass'], data['Survived']))
```



Choosing The Best Chart

- As mentioned with each of the preceding charts that we have seen, it is important to understand what type of data you have. If you have continuous variables, then a histogram would be a good choice. Similarly, if you want to show ranking, an ordered bar chart would be a good choice.
- Choose the chart that effectively conveys the right and relevant meaning of the data without actually distorting the facts.
- Simplicity is best. It is considered better to draw a simple chart that is comprehensible than to draw sophisticated ones that require several reports and text order to understand.
- Choose a diagram that does not overload the audience with information. Our purpose should be to illustrate abstract information in a clear way.

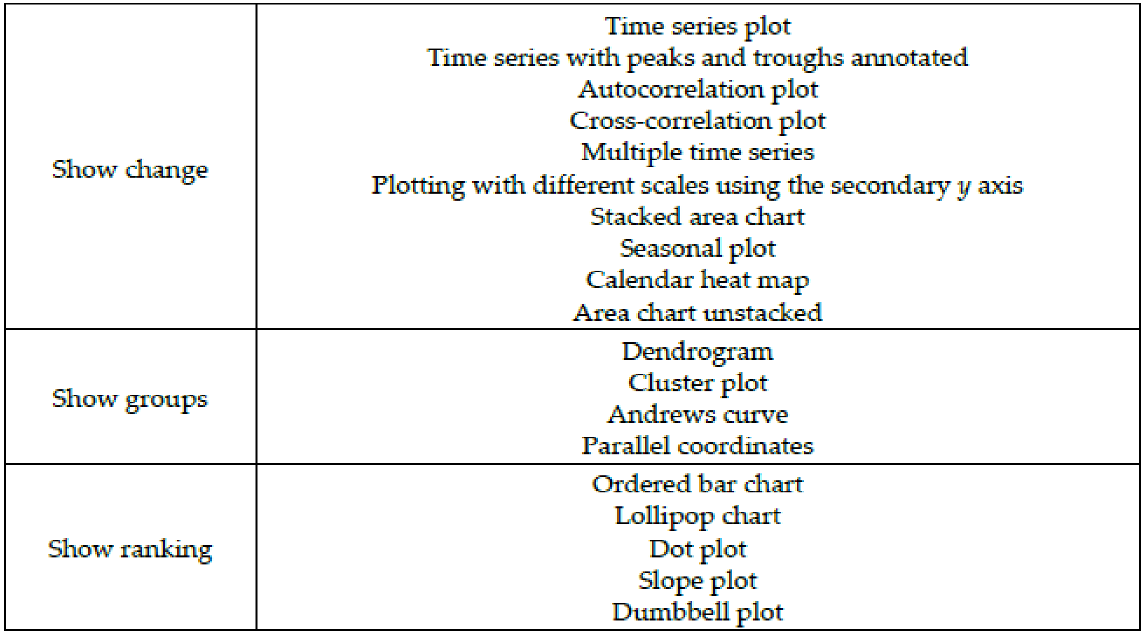
Choosing The Best Chart



The following table shows the different types of charts based on the purposes:

Purpose	Charts
Show correlation	Scatter plot Correlogram Pairwise plot Jittering with strip plot Counts plot Marginal histogram Scatter plot with a line of best fit Bubble plot with circling
Show deviation	Area chart Diverging bars Diverging texts Diverging dot plot Diverging lollipop plot with markers
Show distribution	Histogram for continuous variable Histogram for categorical variable Density plot Categorical plots Density curves with histogram Population pyramid Violin plot Joy plot Distributed dot plot Box plot
Show composition	Waffle chart Pie chart Treemap Bar chart

Choosing The Best Chart



Show change	Time series plot Time series with peaks and troughs annotated Autocorrelation plot Cross-correlation plot Multiple time series Plotting with different scales using the secondary *y* axis Stacked area chart Seasonal plot Calendar heat map Area chart unstacked
Show groups	Dendrogram Cluster plot Andrews curve Parallel coordinates
Show ranking	Ordered bar chart Lollipop chart Dot plot Slope plot Dumbbell plot

Essential Statistical Foundation

Topics Covered

- We will discuss briefly the following topics:
- Population, Sample, Sampling
- Descriptive Statistics Basics
- Probability Distribution
- Hypothesis Testing

Normal Distributions, Mean, Variance

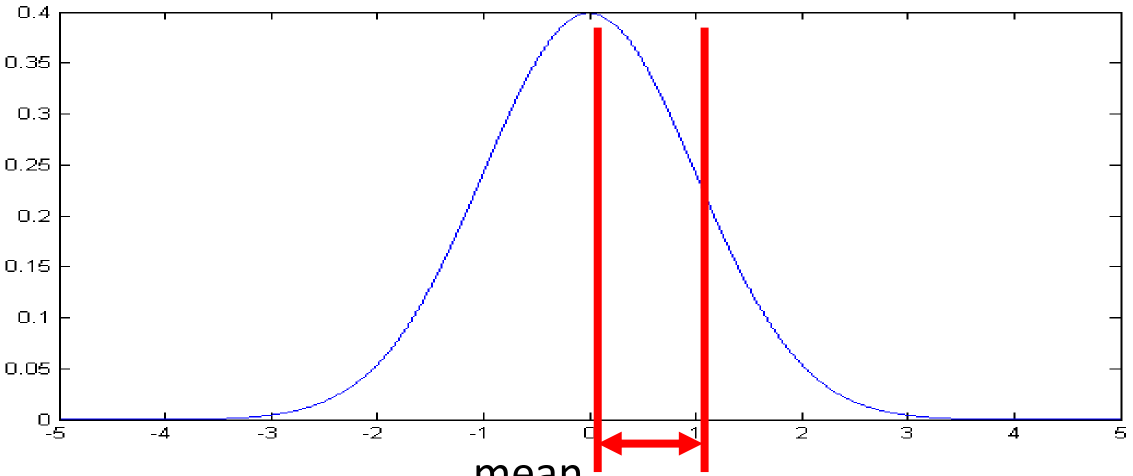
The mean of a set of values is just the average of the values.

Variance a measure of the width of a distribution. Specifically, the variance is the mean squared deviation of samples from the sample mean:

$$Var(X) = \frac{1}{n} \, {}_{i=1}^{n} \, (X_i - \bar{X})^2$$

The standard deviation is the square root of variance.

The normal distribution is completed characterized by mean and variance.



mean

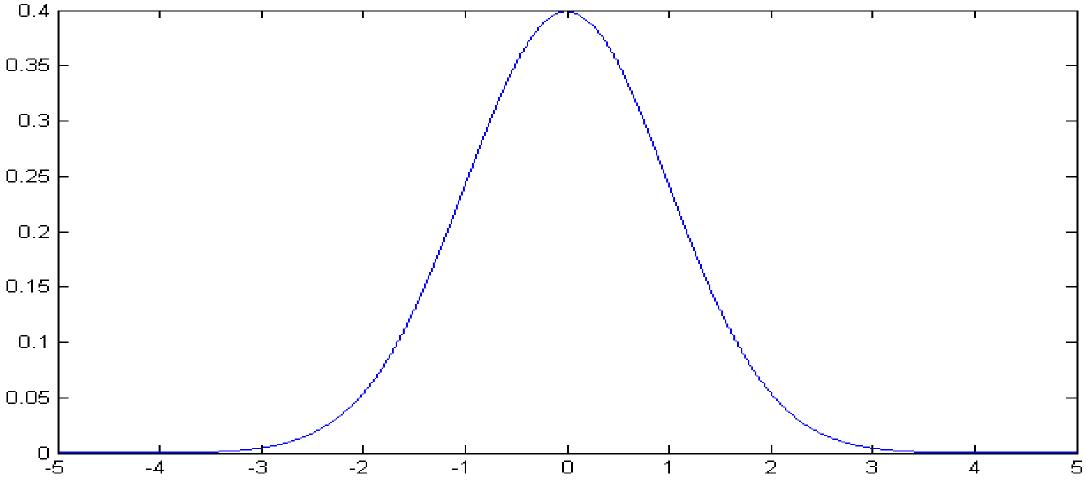
Standard deviation

Central Limit Theorem

The distribution of the sum (or mean) of a set of n identically-distributed random variables Xi approaches a normal distribution as n → ∞.

The common parametric statistical tests, like t-test and ANOVA assume normally-distributed data, but depend on sample mean and variance measures of the data.

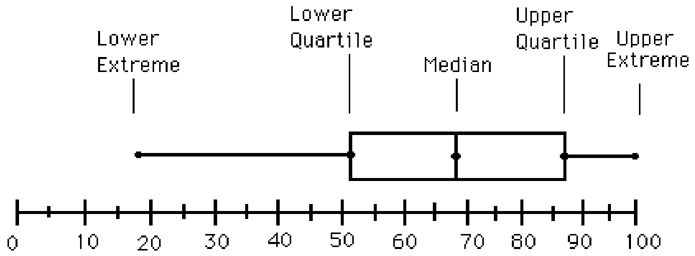
They typically work reasonably well for data that are not normally distributed as long as the samples are not too small.



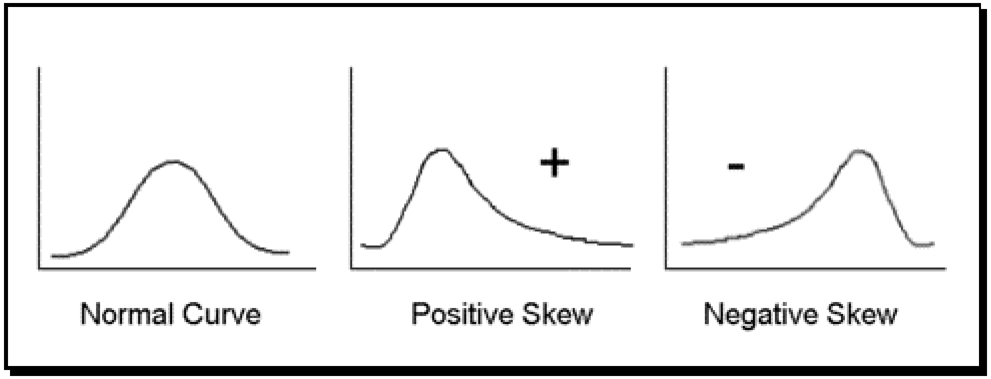
Correcting distributions

Many statistical tools, including mean and variance, t-test, ANOVA etc. **assume data are normally distributed**.

Very often this is not true. The box-and-whisker plot is a good clue



Whenever its asymmetric, the data cannot be normal. The histogram gives even more information

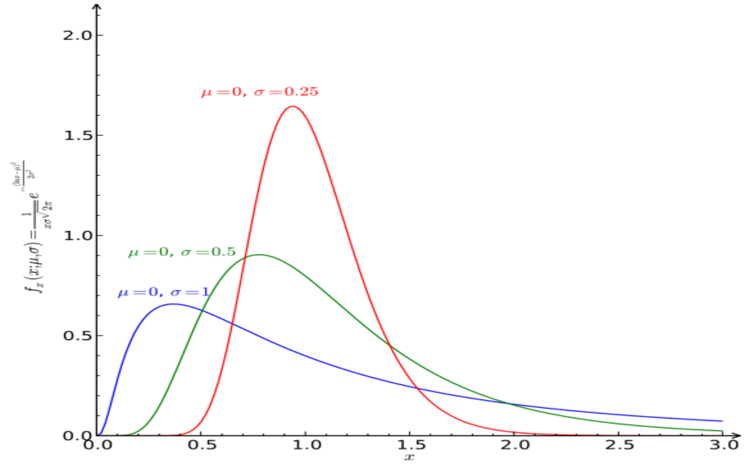


Correcting distributions

In many cases these distribution can be corrected before any other processing.

Examples:

- X satisfies a log-normal distribution, Y=log(X) has a normal dist.



- X poisson with mean k and sdev. sqrt(k). Then sqrt(X) is approximately normally distributed with sdev 1.

Distributions

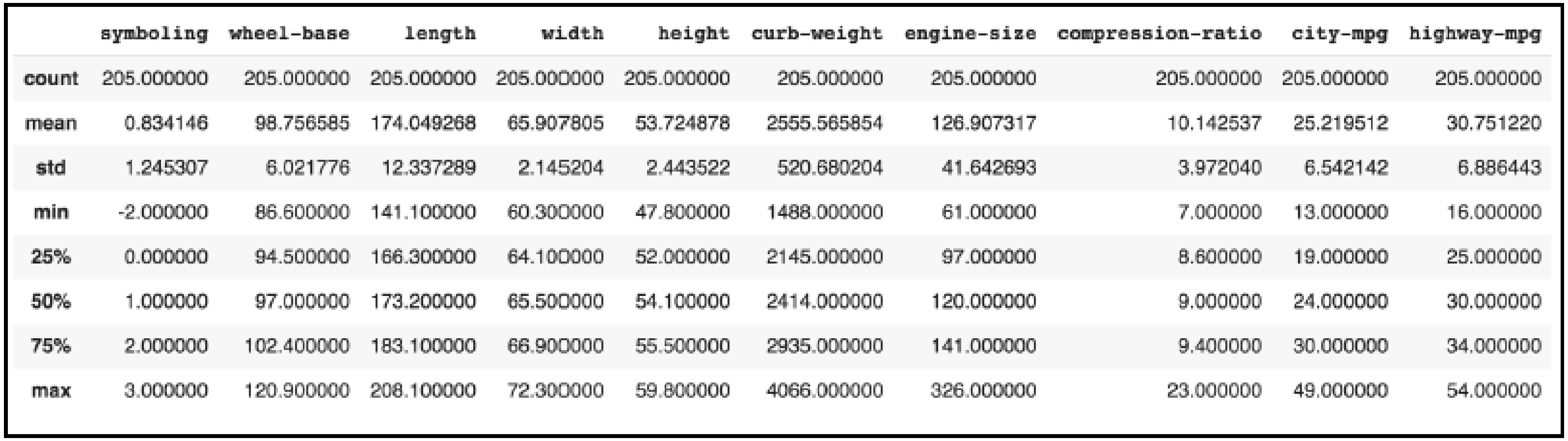
Some other important distributions:

- **Poisson:** the distribution of counts that occur at a certain “rate”.
  - Observed frequency of a given term in a corpus.
  - Number of visits to a web site in a fixed time interval.
  - Number of web site clicks in an hour.
- **Exponential:** the interval between two such events.
- **Zipf/Pareto/Yule distributions:** govern the frequencies of different terms in a document, or web site visits.
- **Binomial/Multinomial:** The number of counts of events (e.g. die tosses = 6) out of n trials.

- You should understand the distribution of your data before applying any model.

```
df.describe()
```

The output of the preceding code is shown in the following screenshot:



	symboling	wheel-base	length	width	height	curb-weight	engine-size	compression-ratio	city-mpg	highway-mpg
count	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000
mean	0.834146	98.756585	174.049268	65.907805	53.724878	2555.565854	126.907317	10.142537	25.219512	30.751220
std	1.245307	6.021776	12.337289	2.145204	2.443522	520.680204	41.642693	3.972040	6.542142	6.886443
min	-2.000000	86.600000	141.100000	60.300000	47.800000	1488.000000	61.000000	7.000000	13.000000	16.000000
25%	0.000000	94.500000	166.300000	64.100000	52.000000	2145.000000	97.000000	8.600000	19.000000	25.000000
50%	1.000000	97.000000	173.200000	65.500000	54.100000	2414.000000	120.000000	9.000000	24.000000	30.000000
75%	2.000000	102.400000	183.100000	66.900000	55.500000	2935.000000	141.000000	9.400000	30.000000	34.000000
max	3.000000	120.900000	208.100000	72.300000	59.800000	4066.000000	326.000000	23.000000	49.000000	54.000000

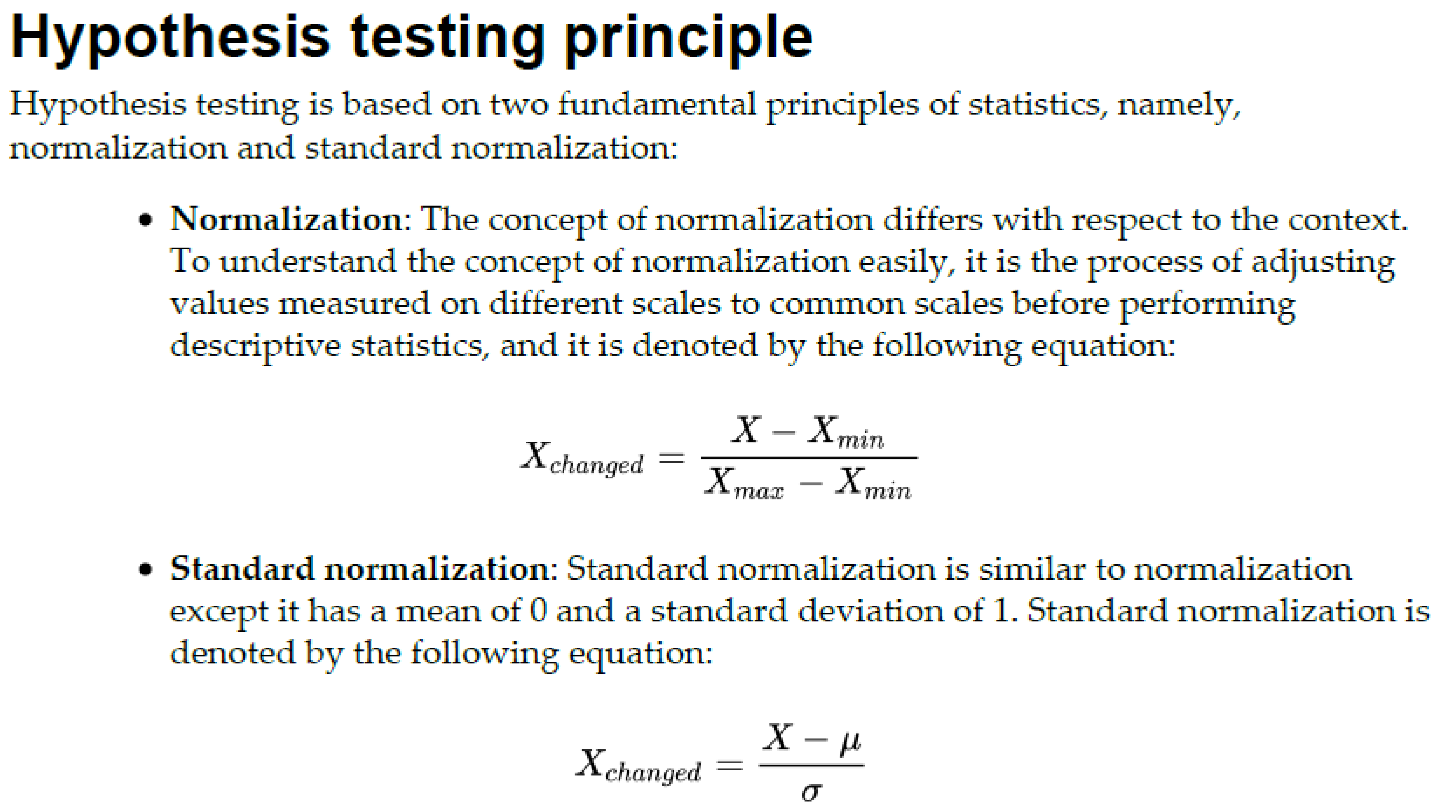
Hypothesis testing

Hypothesis testing is often used to facilitate statistical decisions using experimental datasets. The testing is used to validate assumptions about a population parameter. For example, consider the following statements:

- The average score of students taking the Machine Learning course at the University of Nepal is 78.
- The average height of boys is higher than that of girls among the students taking the Machine Learning course.

In all these examples, we assume some statistical facts to prove those statements. A situation like this is where hypothesis testing helps. A hypothesis test assesses two mutually exclusive statements about any particular **population** and determines which statement is best established by the **sample** data. Here, we used two essential keywords: population and sample. A population includes all the elements from a set of data, whereas a sample consists of one or more observations taken from any particular population.

Hypothesis Testing



Hypothesis testing principle

Hypothesis testing is based on two fundamental principles of statistics, namely, normalization and standard normalization:

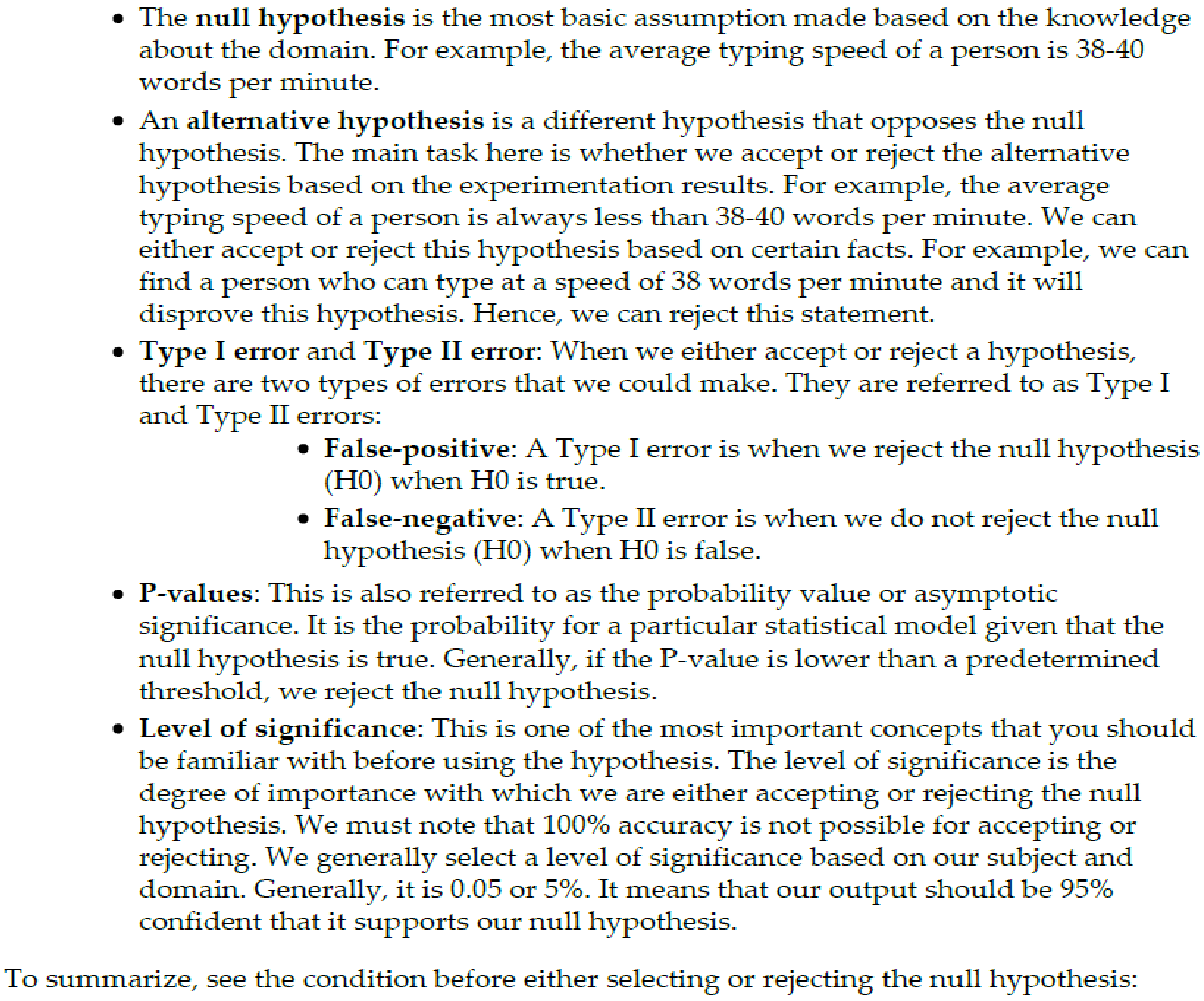
- **Normalization**: The concept of normalization differs with respect to the context. To understand the concept of normalization easily, it is the process of adjusting values measured on different scales to common scales before performing descriptive statistics, and it is denoted by the following equation:

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

- **Standard normalization**: Standard normalization is similar to normalization except it has a mean of 0 and a standard deviation of 1. Standard normalization is denoted by the following equation:

$$X_{changed} = \frac{X - \mu}{\sigma}$$

Hypothesis Testing – contd.



- The **null hypothesis** is the most basic assumption made based on the knowledge about the domain. For example, the average typing speed of a person is 38-40 words per minute.
- An **alternative hypothesis** is a different hypothesis that opposes the null hypothesis. The main task here is whether we accept or reject the alternative hypothesis based on the experimentation results. For example, the average typing speed of a person is always less than 38-40 words per minute. We can either accept or reject this hypothesis based on certain facts. For example, we can find a person who can type at a speed of 38 words per minute and it will disprove this hypothesis. Hence, we can reject this statement.
- **Type I error** and **Type II error**: When we either accept or reject a hypothesis, there are two types of errors that we could make. They are referred to as Type I and Type II errors:
    - **False-positive**: A Type I error is when we reject the null hypothesis (H0) when H0 is true.
    - **False-negative**: A Type II error is when we do not reject the null hypothesis (H0) when H0 is false.
- **P-values**: This is also referred to as the probability value or asymptotic significance. It is the probability for a particular statistical model given that the null hypothesis is true. Generally, if the P-value is lower than a predetermined threshold, we reject the null hypothesis.
- **Level of significance**: This is one of the most important concepts that you should be familiar with before using the hypothesis. The level of significance is the degree of importance with which we are either accepting or rejecting the null hypothesis. We must note that 100% accuracy is not possible for accepting or rejecting. We generally select a level of significance based on our subject and domain. Generally, it is 0.05 or 5%. It means that our output should be 95% confident that it supports our null hypothesis.

To summarize, see the condition before either selecting or rejecting the null hypothesis:

Average reading time

Let's say a reading competition was conducted with some adults. The data looks like the following:

```
[236, 239, 209, 246, 246, 245, 215, 212, 242, 241, 219, 242, 236, 211, 216,
214, 203, 223, 200, 238, 215, 227, 222, 204, 200, 208, 204, 230, 216, 204,
201, 202, 240, 209, 246, 224, 243, 247, 215,249, 239, 211, 227, 211, 247,
235, 200, 240, 213, 213, 209, 219,209, 222, 244, 226, 205, 230, 238, 218,
242, 238, 243, 248, 228,243, 211, 217, 200, 237, 234, 207, 217, 211, 224,
217, 205, 233, 222, 218, 202, 205, 216, 233, 220, 218, 249, 237, 223]
```

Now, our hypothesis question is this: **Is the average reading speed of random students (adults) more than 212 words per minute?**

We can break down the preceding concept into the following parameters:

- **Population**: All adults
- **Parameter of interest**: μ, the population of a classroom
- **Null hypothesis**: $\mu = 212$
- **Alternative hypothesis**: $\mu > 212$
- **Confidence level**: $\alpha = 0.05$

We know all the required parameters. Now, we can use a Z-test from the `statsmodels` package with `alternate="larger"`:

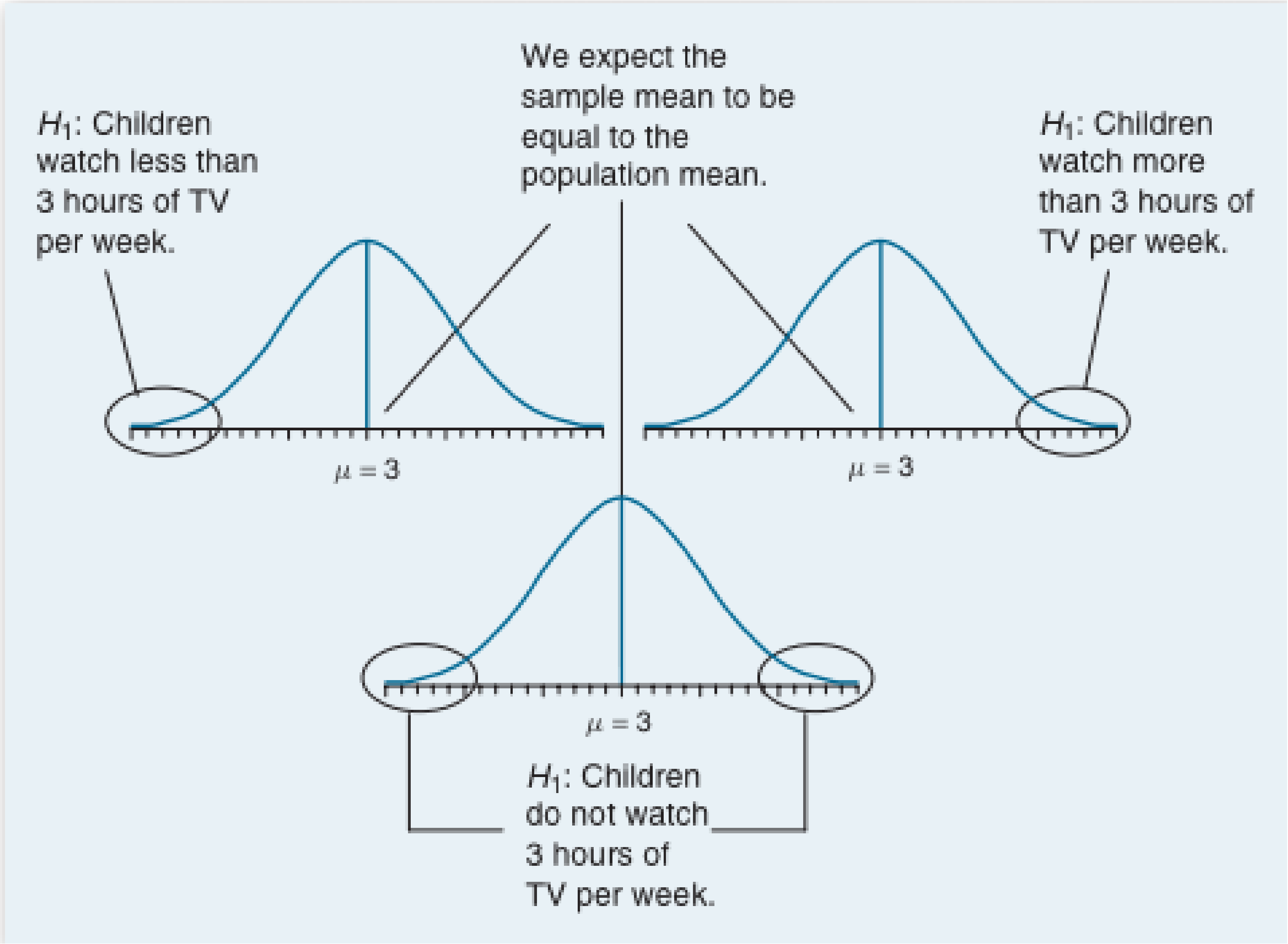
```
import numpy as np

sdata = np.random.randint(200, 250, 89)
sm.stats.ztest(sdata, value = 80, alternative = "larger")
```

The output of the preceding code is as follows:

```
(91.63511530225408, 0.0)
```

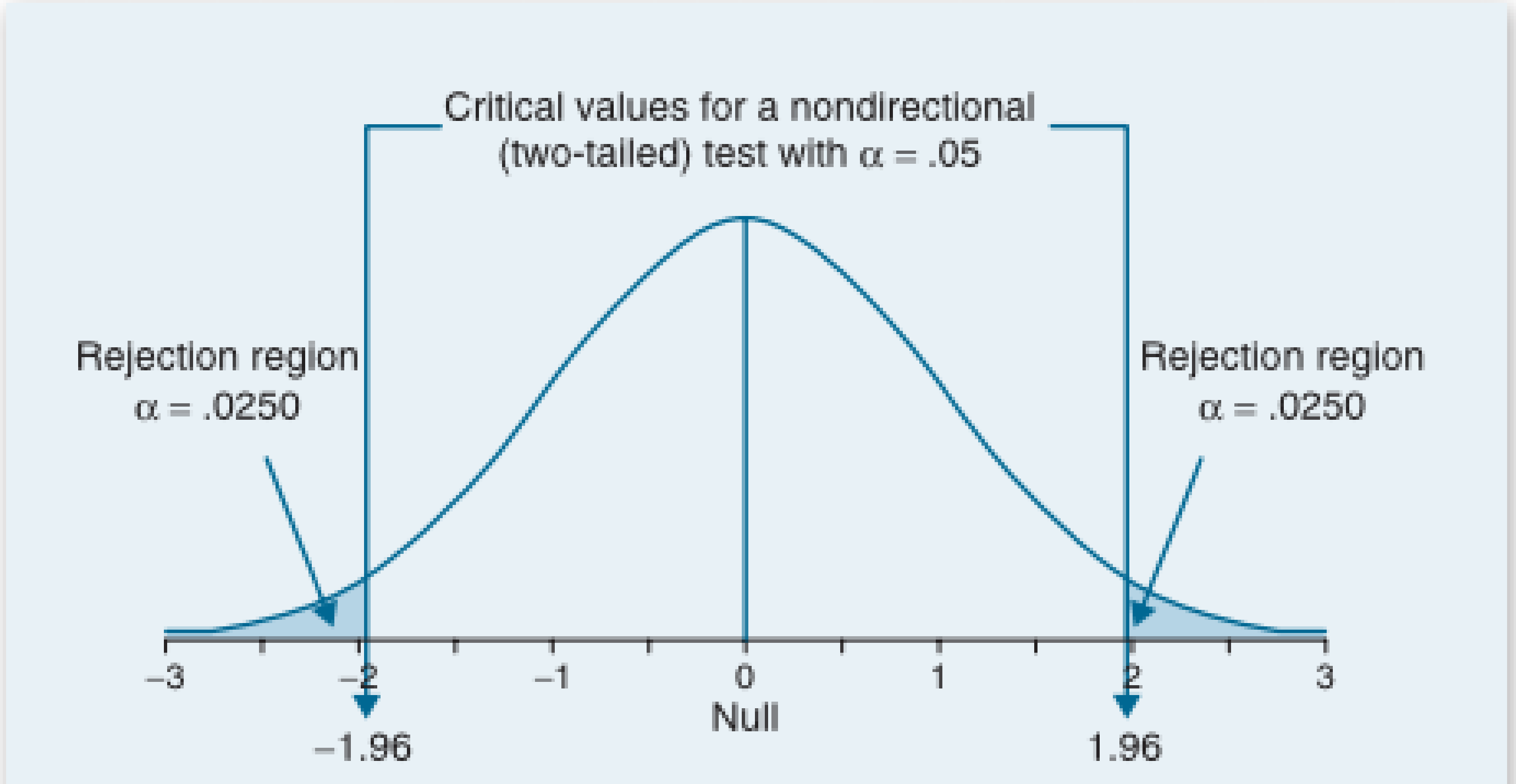
Since the computed P-value (0.0) is lower than the standard confidence level ($\alpha = 0.05$), we can **reject the null hypothesis**. That means the statement *the average reading speed of adults is 212 words per minute* is rejected.



H_1: Children watch less than 3 hours of TV per week.
We expect the sample mean to be equal to the population mean.
H_1: Children watch more than 3 hours of TV per week.
$\mu = 3$
$\mu = 3$
$\mu = 3$
H_1: Children do not watch 3 hours of TV per week.

From G.J. Primavera, “Statistics for the Behavioral Sciences”

Two-tailed Significance



From G.J. Primavera, “Statistics for the Behavioral Sciences”

When the p value is less than 5% (p < .05), we reject the null hypothesis

Closing Words

All the tests so far are parametric tests that assume the data are **normally distributed**, and that the samples are **independent of each other and all have the same distribution** (IID).

They may be arbitrarily inaccurate is those assumptions are not met. Always make sure your data satisfies the assumptions of the test you're using. e.g. watch out for:

- Outliers – will corrupt many tests that use variance estimates.
- Correlated values as samples, e.g. if you repeated measurements on the same subject.
- Skewed distributions – give invalid results.

Non-parametric tests

These tests make no assumption about the distribution of the input data, and can be used on very general datasets:

- K-S test

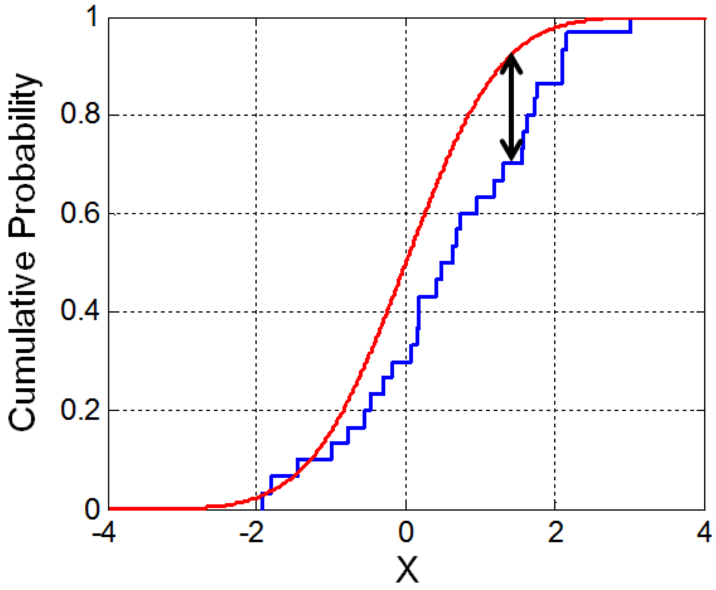
K-S test

The K-S (Kolmogorov-Smirnov) test is a very useful test for checking whether two (continuous or discrete) distributions are the same.

In the **one-sided test**, an observed distribution (e.g. some observed values or a histogram) is compared against a reference distribution.

In the **two-sided test**, two observed distributions are compared.

The K-S statistic is just the **max distance between the CDFs** of the two distributions.



While the statistic is simple, its distribution is not!

But it is available in most stat packages.

K-S test

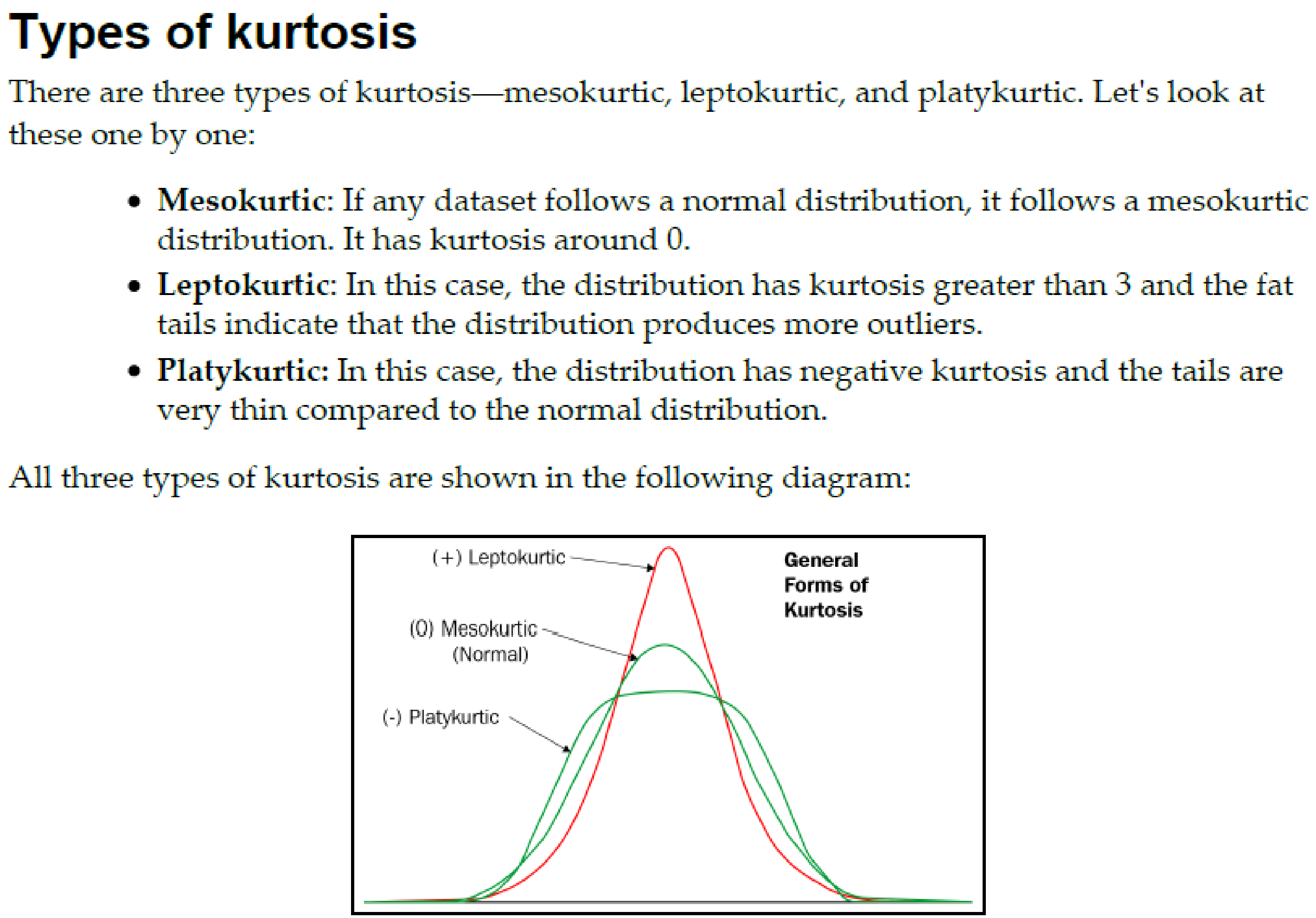
The K-S test can be used to test **whether a data sample has a normal distribution** or not.

Thus it can be used as a sanity check for any common parametric test (which assumes normally-distributed data).

It can also be used to compare distributions of data values in a large data pipeline: **Most errors will distort the distribution of a data parameter and a K-S test can detect this**.

Kurtosis

- Kurtosis, unlike skewness, is not about the peakedness or flatness. It is the measure of outlier presence in a given distribution. Both high and low kurtosis are an indicator that data needs further investigation. The higher the kurtosis, the higher the outliers.



Types of kurtosis

There are three types of kurtosis—mesokurtic, leptokurtic, and platykurtic. Let's look at these one by one:

- **Mesokurtic**: If any dataset follows a normal distribution, it follows a mesokurtic distribution. It has kurtosis around 0.
- **Leptokurtic**: In this case, the distribution has kurtosis greater than 3 and the fat tails indicate that the distribution produces more outliers.
- **Platykurtic:** In this case, the distribution has negative kurtosis and the tails are very thin compared to the normal distribution.

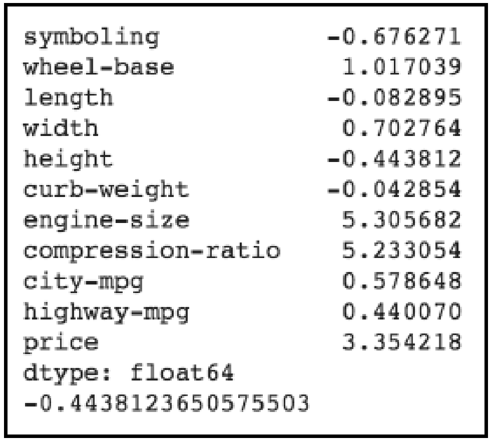
All three types of kurtosis are shown in the following diagram:

Different Python libraries have functions to get the kurtosis of the dataset. The SciPy library has the `scipy.stats.kurtosis(dataset)` function. Using the pandas library, we calculate the kurtosis of our `df` data frame using the `df.kurt()` function:

```
# Kurtosis of data in data using skew() function
kurtosis =df.kurt()
print(kurtosis)

# Kurtosis of the specific column
sk_height=df.loc[:,"height"].kurt()
print(sk_height)
```

The output of the preceding code is given here:



```
symboling            -0.676271
wheel-base            1.017039
length               -0.082895
width                 0.702764
height               -0.443812
curb-weight          -0.042854
engine-size           5.305682
compression-ratio     5.233054
city-mpg              0.578648
highway-mpg           0.440070
price                 3.354218
dtype: float64
-0.4438123650575503
```

Similarly, we can compute the kurtosis of any particular data column. For example, we can compute the kurtosis of the column height as `df.loc[:,"height"].kurt()`.