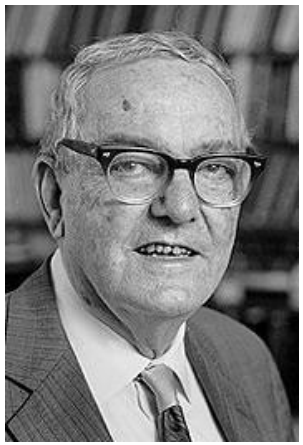
Outline

- Part 0: What is ML?
- Part 1: Data, data types, formats, basic analysis (today)
- Part 2: Data Visualization (Next Tuesday)
- Part 3: Dimensionality reduction (Week after)

Machine Learning

- **Herbert Alexander Simon**: "Learning is any process by which a system improves performance from experience."
- "Machine Learning is concerned with computer programs that automatically improve their performance through experience. "



Herbert Simon
Turing Award 1975
Nobel Prize in Economics 1978

Machine Learning

- Machine Learning
  - Study of algorithms that
  - improve their performance
  - at some task
  - with experience
- Optimize a performance criterion using example data or past experience.
- Role of Statistics: Inference from a sample
- Role of Computer science: Efficient algorithms to
  - Solve the optimization problem
  - Representing and evaluating the model for inference

What We Talk About When We Talk About“Learning”

- Learning general models from a data of particular examples
- Data is cheap and abundant (data warehouses, data marts); knowledge is expensive and scarce.
- Build a model that is *a good and useful approximation* to the data.

Why Machine Learning?

- Develop systems that can automatically adapt and customize themselves to individual users.
  - Personalized news or mail filter
- Discover new knowledge from large databases (***data mining***).
  - Market basket analysis
- Ability to mimic human and replace certain monotonous tasks - which require some intelligence.
  - like recognizing handwritten characters
- Develop systems that are too difficult/expensive to construct manually because they require specific detailed skills or knowledge tuned to a specific task (knowledge engineering bottleneck).

Why now?

- Flood of available data (especially with the advent of the Internet)
- Increasing computational power
- Growing progress in available algorithms and theory developed by researchers
- Increasing support from industries

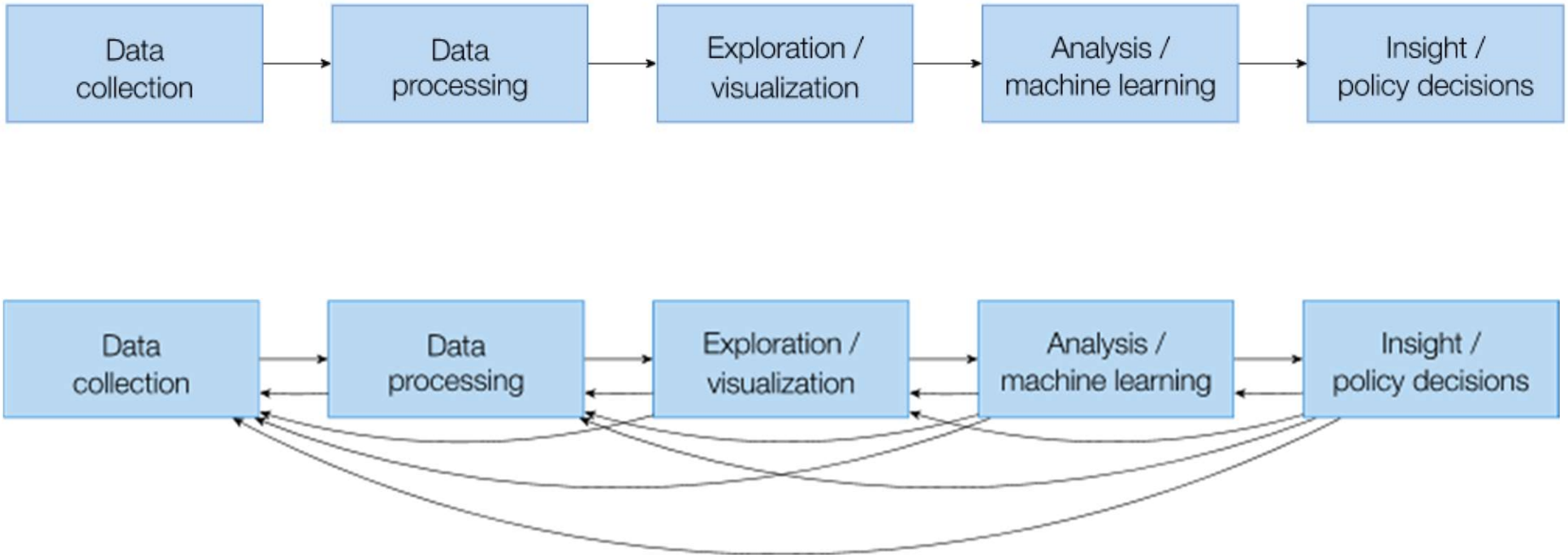
The data revolution

- As “big data” efforts amass more data… the need for new data science methodologies increases

- Data today have more volume, velocity, variety, etc.

Drowning in Data
but starving for Knowledge

Course point of view: Data Pipeline



Lecture 2: Exploratory Data Analysis (EDA)

Olexandr Isayev

Department of Chemistry, CMU

olexandr@cmu.edu

Data

In 2017, *The Economist* published a story titled:

"The world's most valuable resource is no longer oil, but data."

https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data

data is new oil

All Images News Shopping Videos More Settings Tools

About 2,040,000,000 results (0.80 seconds)

www.forbes.com › sites › forbestechcouncil › 2019/11/15 › data-is-th...

Council Post: Data Is The New Oil -- And That's A Good Thing

Nov 15, 2019 - Back in 2017, The Economist published a story titled, "The world's most valuable resource is no longer **oil**, but **data**." Since its publication, the ...

www.wired.com › story › no-data-is-not-the-new-oil

No, Data Is Not the New Oil | WIRED

Feb 26, 2019 - "**Data** is the **new oil**" is one of those deceptively simple mantras for the modern world. Whether in The **New** York Times, The Economist, ...

hackernoon.com › data-is-the-new-oil-1227197762b2

Data is the New Oil - By Giuliano Giacaglia - Hacker Noon

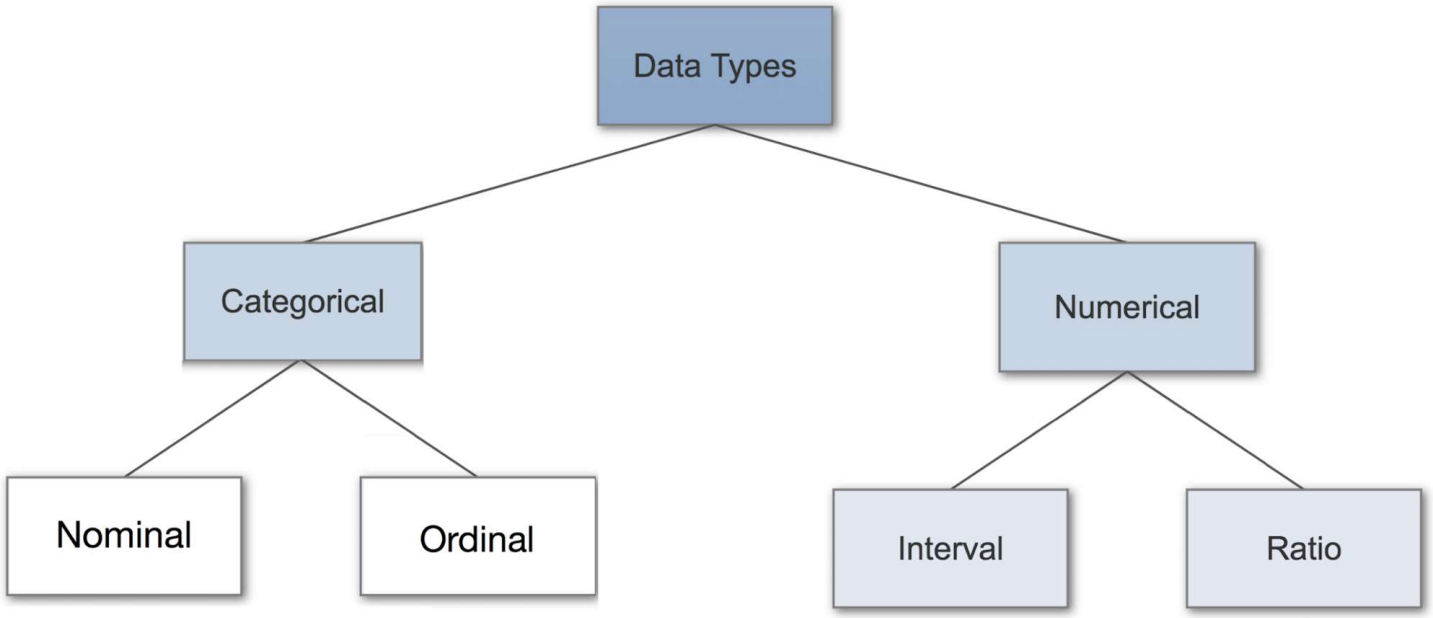
"**Data** is the **new oil**. It's valuable, but if unrefined it cannot really be used. It has to be changed into gas, plastic, chemicals, etc to create a valuable entity that ...

www.economist.com › leaders › 2017/05/06 › the-worlds-most-valua...

The world's most valuable resource is no longer oil, but data

May 6, 2017 - The **data** economy demands a **new** approach to antitrust rules.

Basic Data Types in Statistics



Categorical Data

- A categorical or discrete variable is one that has two or more categories (values).
- Categorical data represents some kind of characteristics.
- Categorical data can also take on numerical values (Example: 1 for female and 0 for male). Note that those numbers don't have mathematical meaning.
- It's often referred as **Nominal Data**, implying no order

Analysis of categorical data generally involves the use of data tables.

A ***two-way table*** presents categorical data by counting the number of observations that fall into each group for two variables, one divided into rows and the other divided into columns.

	Eye Color				
Hair Color	Blue	Green	Brown	Black	Total
Blonde	2	1	2	1	6
Red	1	1	2	0	4
Brown	1	0	4	2	7
Black	1	0	2	0	3
Total	5	2	10	3	20

The totals for each category, also known as marginal distributions, provide the number of individuals in each row or column without accounting for the effect of the other variable

Ordinal Data

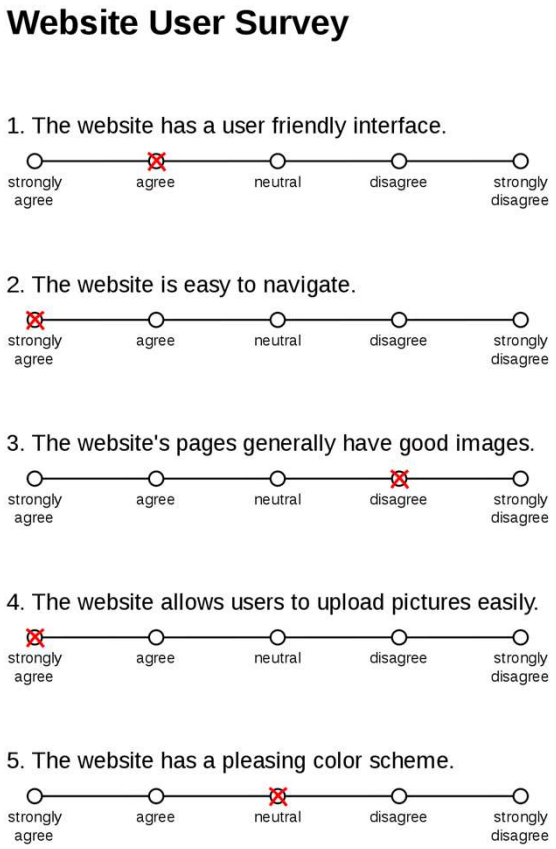
An ordinal variable is a special type of categorical variable.

The difference between the two is that there is a clear ordering of the variables.

- For example, suppose you have a variable, signal strength, with three categories (low, medium and high).
- In addition to being able to classify spectra into these three categories, you can order the categories as low, medium and high.
- Ordinal data can also take on numerical values (Example: 0 – no signal, 1 – low, 2 - medium, 3 – high). Note that those numbers have mathematical meaning.

Likert & Interval scales

- Likert scales could be arbitrary.
- The value assigned to a Likert item has no objective numerical basis
- A good Likert scale, will present a *symmetry* of categories about a midpoint with clearly defined linguistic qualifiers.
- **Interval Scale:** each category is interval (age groups, class grades)



Analysis

Wilcoxon signed-rank test: This is a qualitative statistical test used to compare the 2 groups of matched samples to assess their differences.

Friedman 2-way ANOVA: This is a non-parametric way of finding differences in matched sets of 3 or more groups.

Ordered probit (***prob****ability* + *un****it***)

The purpose of the model is to estimate the probability that an observation with particular characteristics will fall into a specific one of the categories

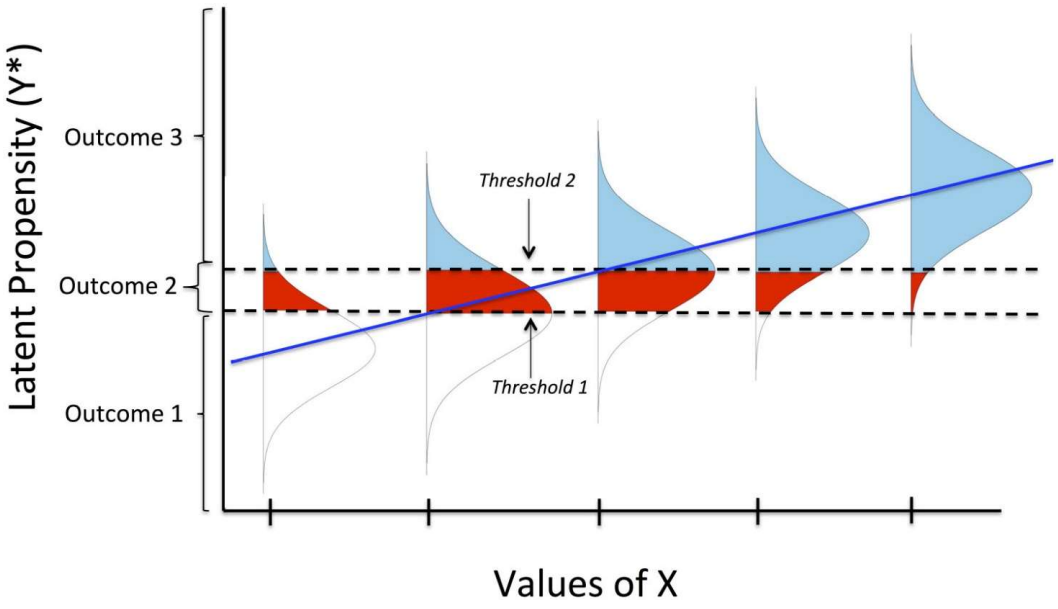


Figure from: https://dx.doi.org/10.4135/9781526473455

Numerical Data

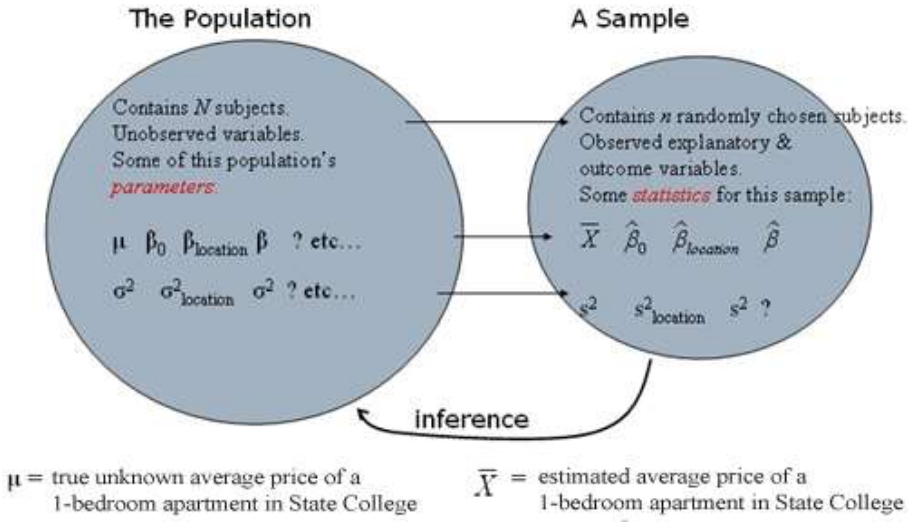
Discrete Data

- This type of data **can't be measured but it can be counted**.
  - Five peaks in spectra
  - Experiment was repeated 100 times

Continuous Data

- Continuous Data represents measurements and therefore their values **can't be counted but they can be measured**.

Discrete Data Analysis



Whole kitchen sink of methods!

Dependent on hypothesis testing: parametric or not

- **Chi-squared test -** whether an association (or relationship) between categorical variables in a sample is likely to reflect a real association between these 2 variables

Continuous Data

Interval

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

- temperature (Farenheit), temperature (Celcius), 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.

- 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.

Summary Statistics

Provides:	Nominal	Ordinal	Interval	Ratio
The "order" of values is known		✔	✔	✔
"Counts," aka "Frequency of Distribution"	✔	✔	✔	✔
Mode	✔	✔	✔	✔
Median		✔	✔	✔
Mean			✔	✔
Can quantify the difference between each value			✔	✔
Can add or subtract values			✔	✔
Can multiple and divide values				✔
Has "true zero"				✔

What kind of variable is Zip Code?

Group frequencies [edit]

Tables of vibrational transitions of stable[4] and transient molecules[5] are also available.

Bond	Type of bond	Specific type of bond	Absorption peak (cm^{-1})	Appearance
C–H	alkyl	methyl	1260	strong
			1380	weak
			2870	medium to strong
			2960	medium to strong
		methylene	1470	strong
			2850	medium to strong
			2925	medium to strong
		methine	2890	weak
	vinyl	$C{=}CH_2$	900	strong
			2975	medium
			3080	medium
		C=CH	3020	medium
		monosubstituted alkenes	900	strong
			990	strong
		cis-disubstituted alkenes	670–700	strong
		trans-disubstituted alkenes	965	strong
		trisubstituted alkenes	800–840	strong to medium
	aromatic	benzene/sub. benzene	3070	weak
		monosubstituted benzene	700–750	strong
			690–710	strong
		ortho-disub. benzene	750	strong
		meta-disub. benzene	750–800	strong
			860–900	strong
		para-disub. benzene	800–860	strong
	alkynes	any	3300	medium
	aldehydes	any	2720	medium
			2820	

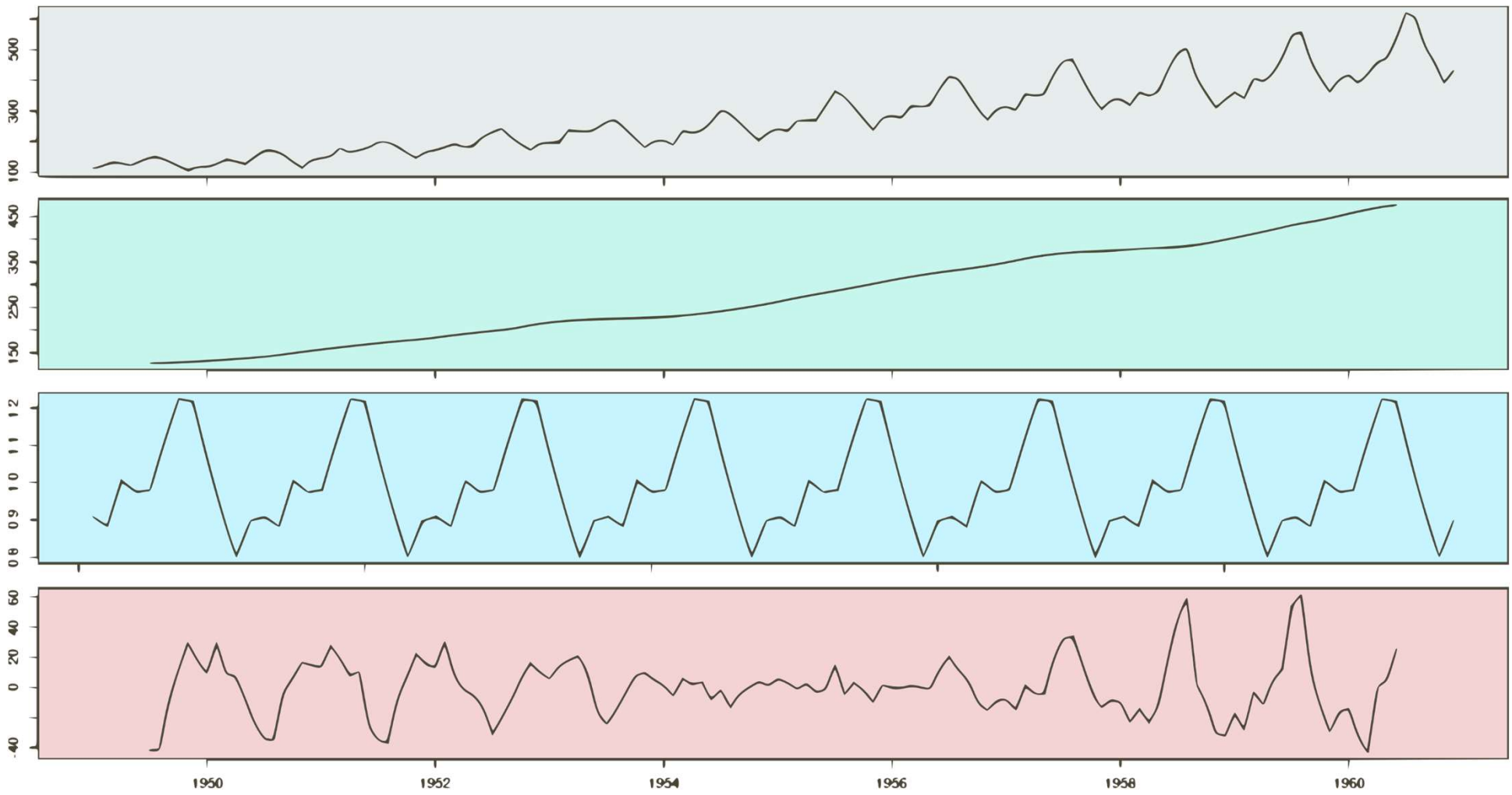
https://en.wikipedia.org/wiki/Infrared_spectroscopy_correlation_table

Time series

A time series is a series of data points indexed in time order.

Most commonly, a time series is a sequence taken at successive equally spaced points in time.

In a time series, time is often the independent variable and the goal is usually to make a forecast for the future.


100
300
500
150
250
350
450
0.8
0.9
1.0
1.1
1.2
-40
0
20
40
60
1950
1952
1954
1956
1958
1960

Other typical questions

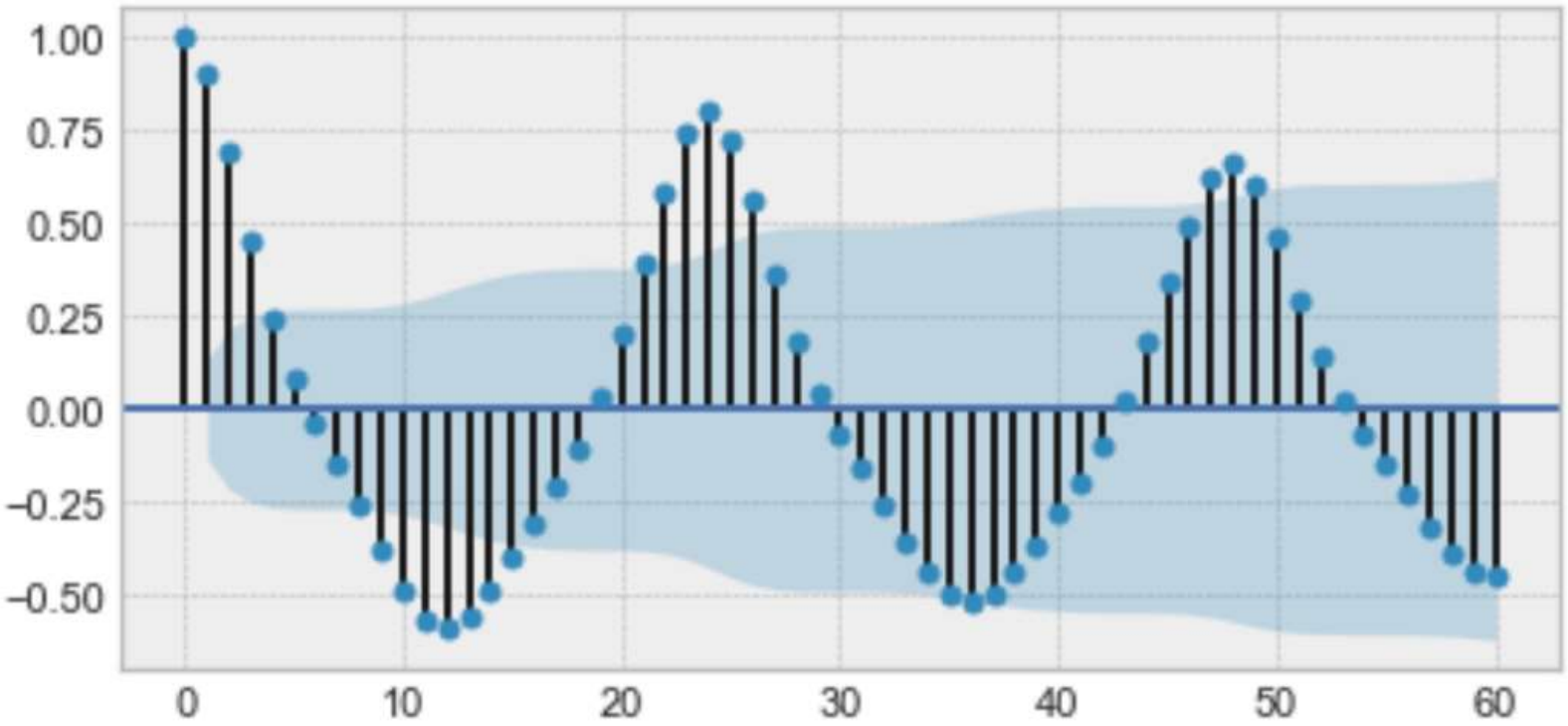
Is it **stationary**?

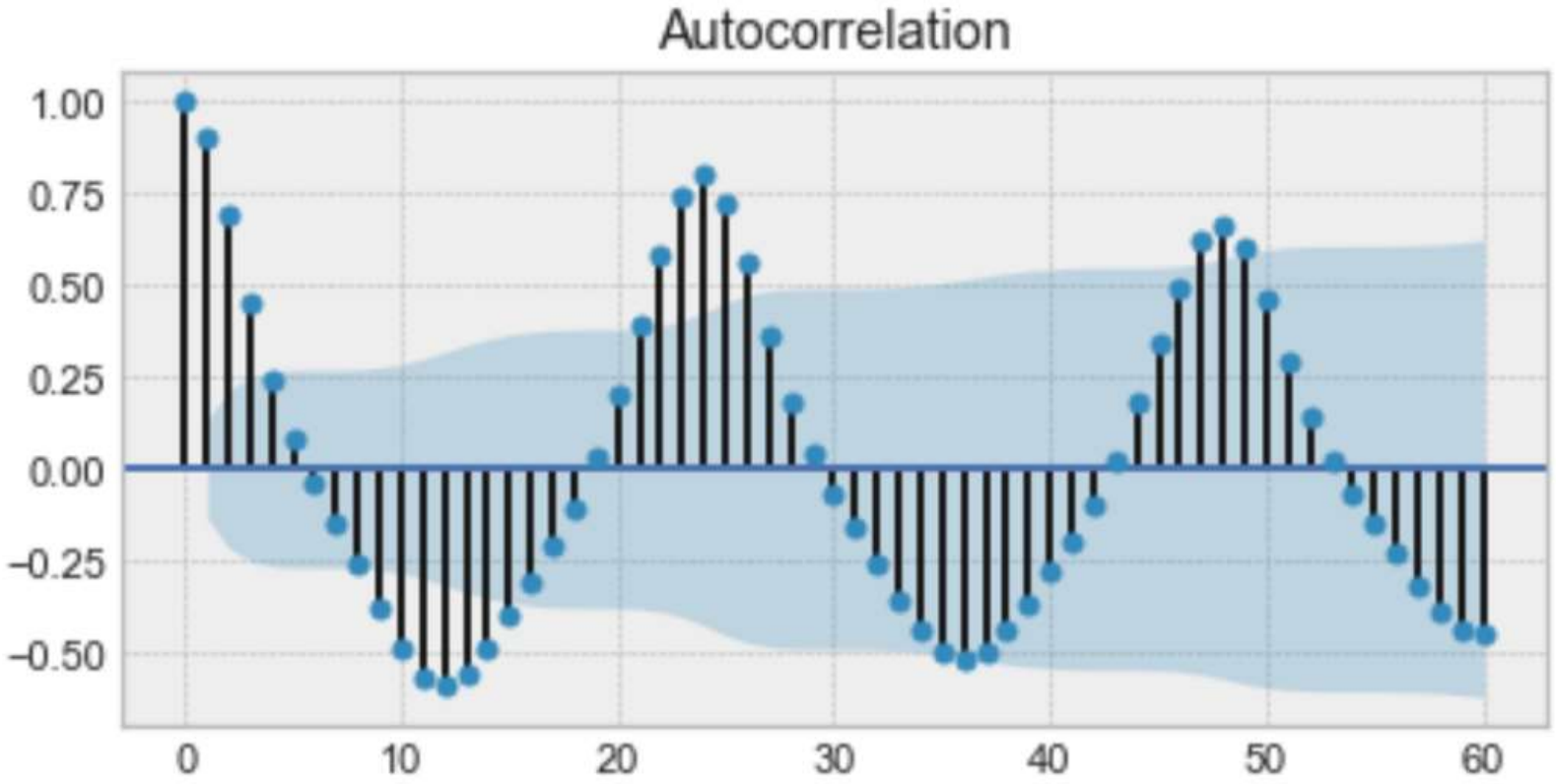
Is there a **seasonality**?

Is the target variable **autocorrelated**?

Autocorrelation

autocorrelation is the similarity between observations as a function of the time lag between them.

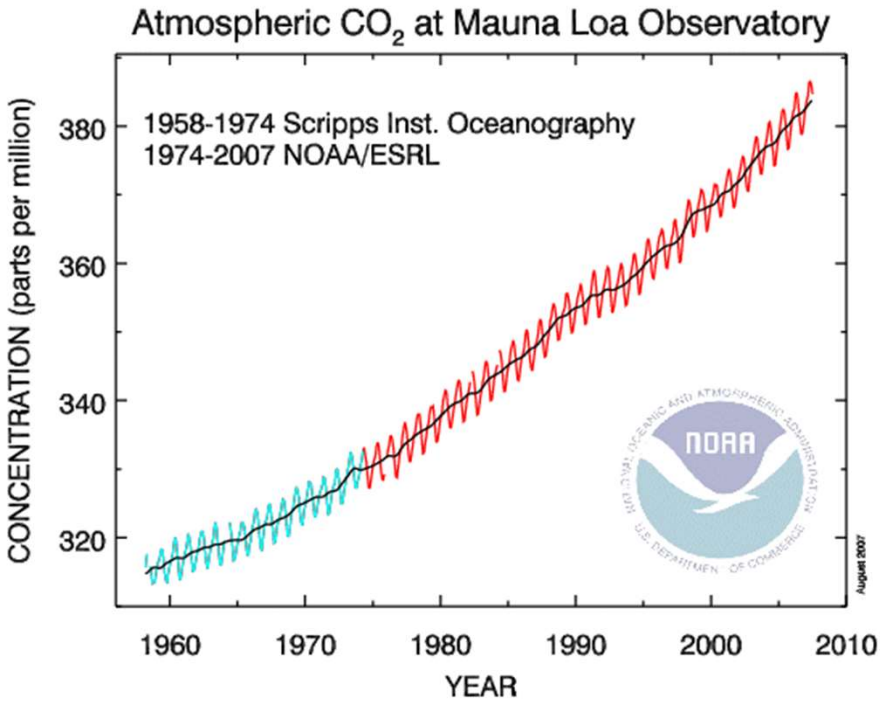

1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
0
10
20
30
40
50
60



Notice how the plot looks like sinusoidal function. This is a hint for **seasonality,** and you can find its value by finding the period in the plot above, which would give 24h.

Seasonality

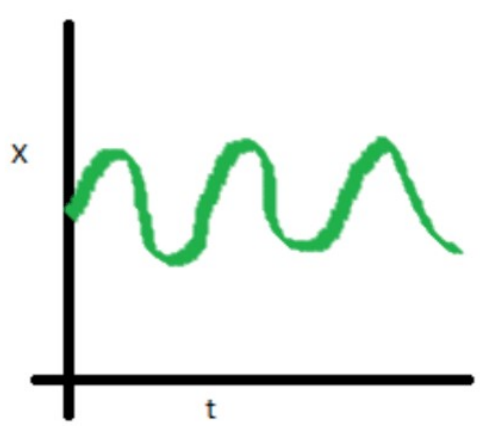
Seasonality refers to periodic fluctuations.



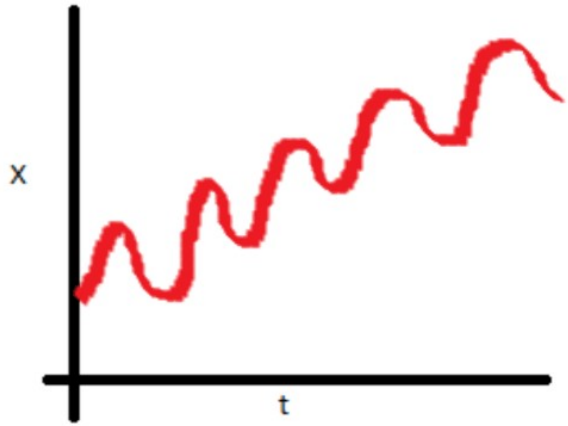
Seasonality can be derived from an autocorrelation plot if it has a sinusoidal shape. Simply look at the period, and it gives the length of the season.

Stationarity

Stationarity is an important characteristic of time series. A time series is said to be stationary if its statistical properties do not change over time. In other words, it has **constant mean and variance**, and covariance is independent of time.

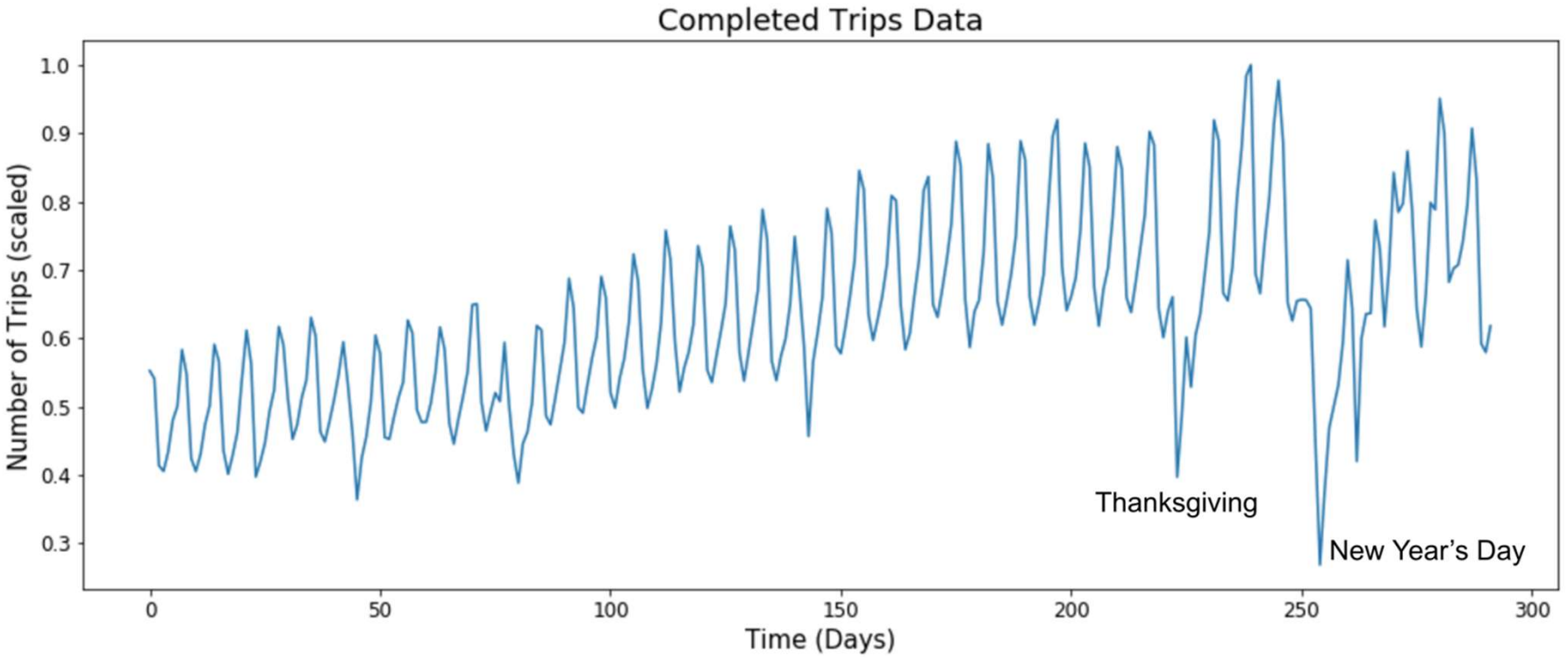


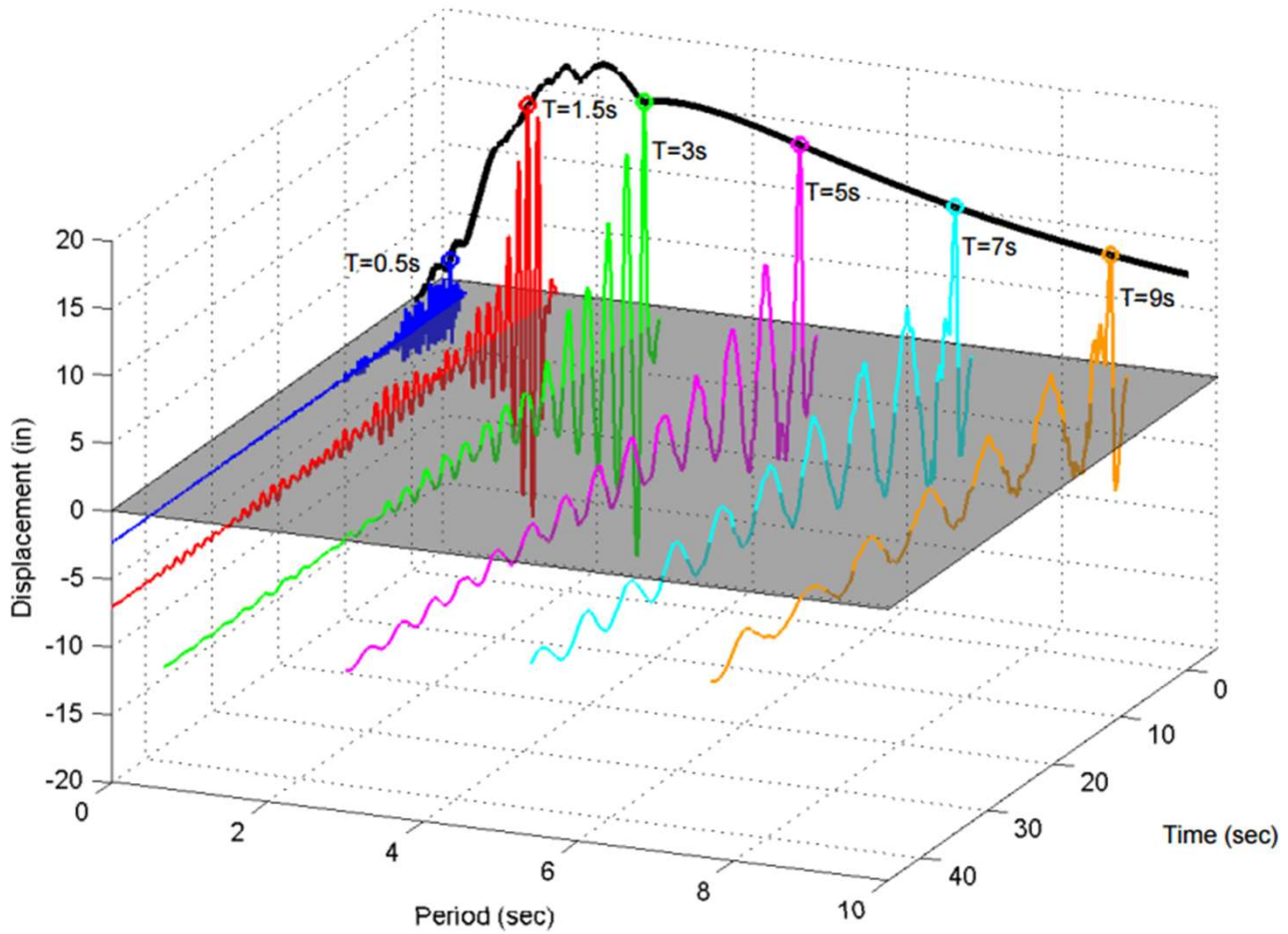
Stationary series



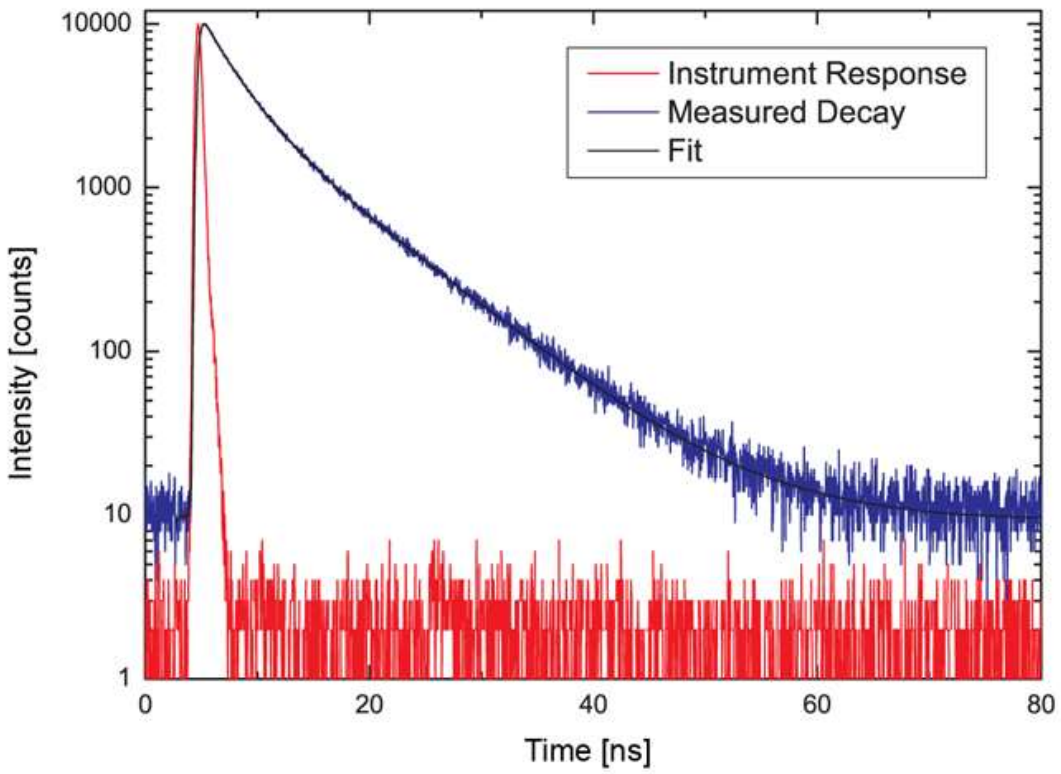
Non-Stationary series

Uber trips

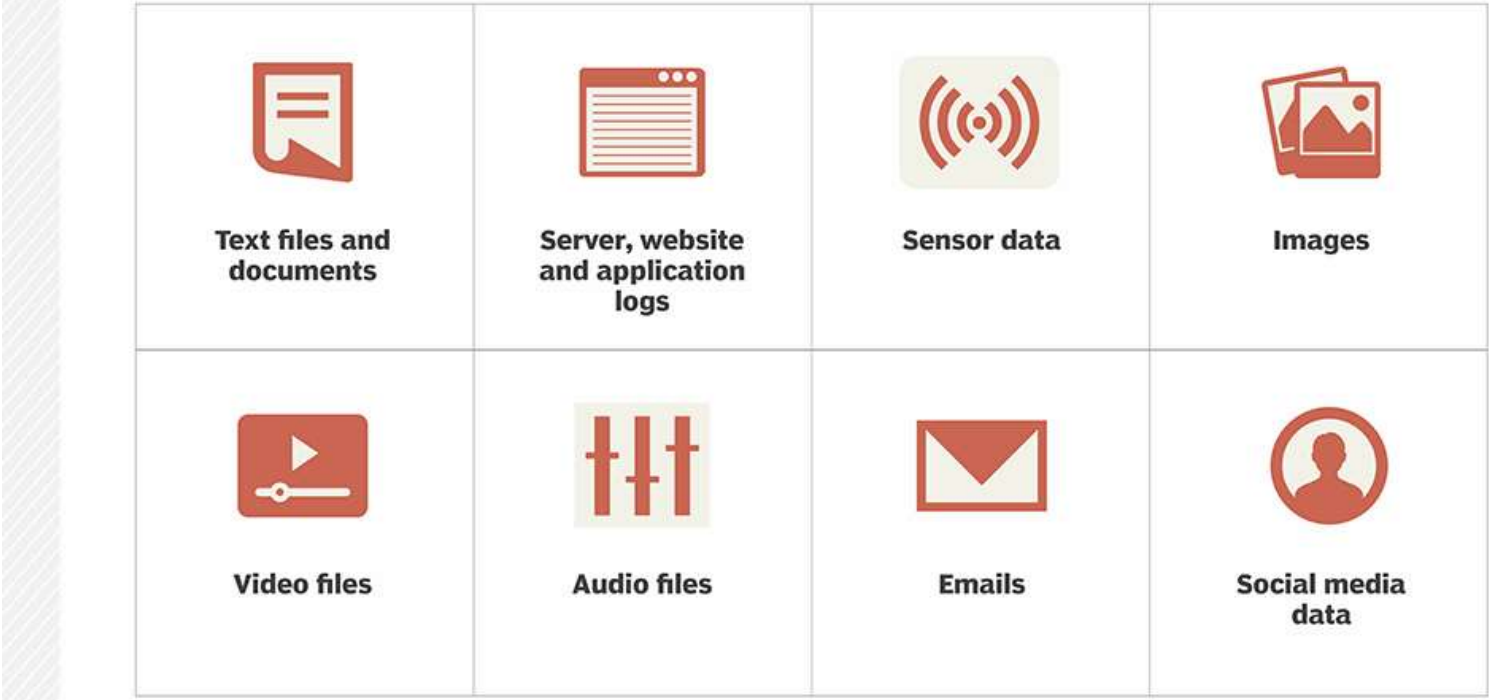

Completed Trips Data
Number of Trips (scaled)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0
50
100
150
200
250
300
Time (Days)
Thanksgiving
New Year's Day


T=0.5s
T=1.5s
T=3s
T=5s
T=7s
T=9s
Displacement (in)
20
15
10
5
0
-5
-10
-15
-20
0
2
4
6
8
10
Period (sec)
0
10
20
30
40
Time (sec)

Time-resolved Fluorescence



Unstructured data types



Missing Values and Data Imputation

		0	1	2	3	4	5	6	7	8
0	0	6	148.0	72.0	35.0	NaN	33.6	0.627	50	1
1	1	1	85.0	66.0	29.0	NaN	26.6	0.351	31	0
2	2	8	183.0	64.0	NaN	NaN	23.3	0.672	32	1
3	3	1	89.0	66.0	23.0	94.0	28.1	0.167	21	0
4	4	0	137.0	40.0	35.0	168.0	43.1	2.288	33	1
5	5	5	116.0	74.0	NaN	NaN	25.6	0.201	30	0
6	6	3	78.0	50.0	32.0	88.0	31.0	0.248	26	1
7	7	10	115.0	NaN	NaN	NaN	35.3	0.134	29	0
8	8	2	197.0	70.0	45.0	543.0	30.5	0.158	53	1
9	9	8	125.0	96.0	NaN	NaN	NaN	0.232	54	1
10	10	4	110.0	92.0	NaN	NaN	37.6	0.191	30	0
11	11	10	168.0	74.0	NaN	NaN	38.0	0.537	34	1
12	12	10	139.0	80.0	NaN	NaN	27.1	1.441	57	0
13	13	1	189.0	60.0	23.0	846.0	30.1	0.398	59	1
14	14	5	166.0	72.0	19.0	175.0	25.8	0.587	51	1
15	15	7	100.0	NaN	NaN	NaN	30.0	0.484	32	1
16	16	0	118.0	84.0	47.0	230.0	45.8	0.551	31	1
17	17	7	107.0	74.0	NaN	NaN	29.6	0.254	31	1
18	18	1	103.0	30.0	38.0	83.0	43.3	0.183	33	0
19	19	1	115.0	70.0	30.0	96.0	34.6	0.529	32	1

Analyze NaN values

```
# count the number of NaN values in each column
print(dataset.isnull().sum())

# count the number of NaN values in each row
print(dataset.isnull().sum(axis=1))
```

Remove NaNs

Remove Rows With Missing Values

```
# drop rows with missing values
dataset.dropna(inplace=True)
```

Remove Columns With Missing Values

```
dataset.dropna(inplace=True, axis=1)
```

		0	1	2	3	4	5	6	7	8
0	0	6	148.0	72.0	35.0	NaN	33.6	0.627	50	1
1	1	1	85.0	66.0	29.0	NaN	26.6	0.351	31	0
2	2	8	183.0	64.0	NaN	NaN	23.3	0.672	32	1
3	3	1	89.0	66.0	23.0	94.0	28.1	0.167	21	0
4	4	0	137.0	40.0	35.0	168.0	43.1	2.288	33	1
5	5	5	116.0	74.0	NaN	NaN	25.6	0.201	30	0
6	6	3	78.0	50.0	32.0	88.0	31.0	0.248	26	1
7	7	10	115.0	NaN	NaN	NaN	35.3	0.134	29	0
8	8	2	197.0	70.0	45.0	543.0	30.5	0.158	53	1
9	9	8	125.0	96.0	NaN	NaN	NaN	0.232	54	1
10	10	4	110.0	92.0	NaN	NaN	37.6	0.191	30	0
11	11	10	168.0	74.0	NaN	NaN	38.0	0.537	34	1
12	12	10	139.0	80.0	NaN	NaN	27.1	1.441	57	0
13	13	1	189.0	60.0	23.0	846.0	30.1	0.398	59	1
14	14	5	166.0	72.0	19.0	175.0	25.8	0.587	51	1
15	15	7	100.0	NaN	NaN	NaN	30.0	0.484	32	1
16	16	0	118.0	84.0	47.0	230.0	45.8	0.551	31	1
17	17	7	107.0	74.0	NaN	NaN	29.6	0.254	31	1
18	18	1	103.0	30.0	38.0	83.0	43.3	0.183	33	0
19	19	1	115.0	70.0	30.0	96.0	34.6	0.529	32	1

Most (but not all) algorithms fail when there is missing data

Impute Missing Values

There are many options we could consider when replacing a missing value.
No right answer

Any imputing performed on the training dataset will have to be performed on new data in the future when predictions are needed from the finalized model.

This needs to be taken into consideration when choosing how to impute the missing values.

Can affect results in unpredictable way

Imputation options

A constant value that has meaning within the domain, such as 0, distinct from all other values.

A value from another randomly selected record.

A mean, median or mode value for the column.

A value estimated by another predictive model.

		0	1	2	3	4	5	6	7	8
0	0	6	148.0	72.0	35.0	NaN	33.6	0.627	50	1
1	1	1	85.0	66.0	29.0	NaN	26.6	0.351	31	0
2	2	8	183.0	64.0	NaN	NaN	23.3	0.672	32	1
3	3	1	89.0	66.0	23.0	94.0	28.1	0.167	21	0
4	4	0	137.0	40.0	35.0	168.0	43.1	2.288	33	1
5	5	5	116.0	74.0	NaN	NaN	25.6	0.201	30	0
6	6	3	78.0	50.0	32.0	88.0	31.0	0.248	26	1
7	7	10	115.0	NaN	NaN	NaN	35.3	0.134	29	0
8	8	2	197.0	70.0	45.0	543.0	30.5	0.158	53	1
9	9	8	125.0	96.0	NaN	NaN	NaN	0.232	54	1
10	10	4	110.0	92.0	NaN	NaN	37.6	0.191	30	0
11	11	10	168.0	74.0	NaN	NaN	38.0	0.537	34	1
12	12	10	139.0	80.0	NaN	NaN	27.1	1.441	57	0
13	13	1	189.0	60.0	23.0	846.0	30.1	0.398	59	1
14	14	5	166.0	72.0	19.0	175.0	25.8	0.587	51	1
15	15	7	100.0	NaN	NaN	NaN	30.0	0.484	32	1
16	16	0	118.0	84.0	47.0	230.0	45.8	0.551	31	1
17	17	7	107.0	74.0	NaN	NaN	29.6	0.254	31	1
18	18	1	103.0	30.0	38.0	83.0	43.3	0.183	33	0
19	19	1	115.0	70.0	30.0	96.0	34.6	0.529	32	1

Example

Pandas provides the fillna() function for replacing missing values with a specific value.

For example, we can use fillna() to replace missing values with the mean value for each column, as follows:

```
# fill missing values with mean column values
dataset.fillna(dataset.mean(), inplace=True)
```

Does measurement scale matter for data analysis?

Knowing the measurement scale for your variables can help prevent mistakes

- Can I take the average of a group of zip (postal) codes?
- ratio of two temperature values?

Knowing the measurement scale for your variables could really help you plan your analyses or interpret the results.

Note that sometimes, the measurement scale for a variable is not clear cut. What kind of variable is color?

- In psychology: different colors would be regarded as nominal.
- In physics: color is quantified by wavelength, so color would be considered a ratio variable.

Data conversion

There are occasions when you will have some control over the measurement scale.

- For example, with temperature, you can choose degrees C or F and have an interval scale or choose degrees Kelvin and have a ratio scale.

With income level, instead of offering categories and having an ordinal scale, you can try to get the actual income and have a ratio scale.

Generally speaking, you want to strive to have a scale towards the ratio end as opposed to the nominal end.

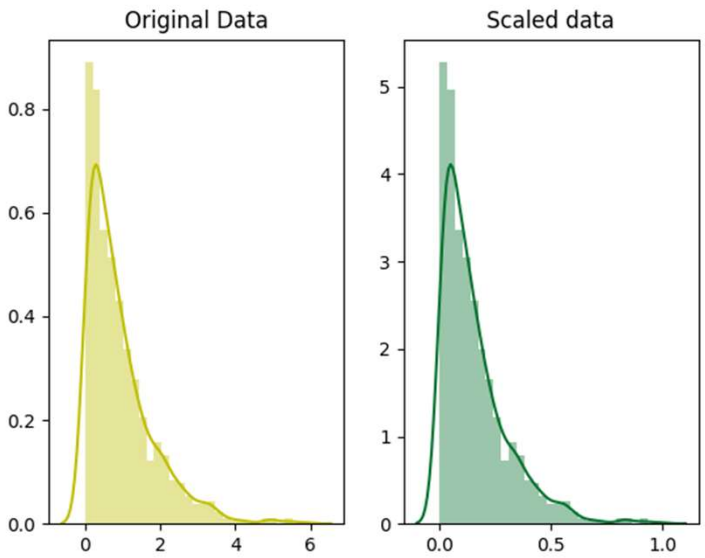
Data Normalization and Standardization

- Data Scaling
- Data Normalization
- Data Standardization
- Data Curation (Lecture in about scientific data)

Scaling

In scaling *(also called **min-max scaling**)*, you transform the data such that the features are within a specific range e.g. [0, 1].

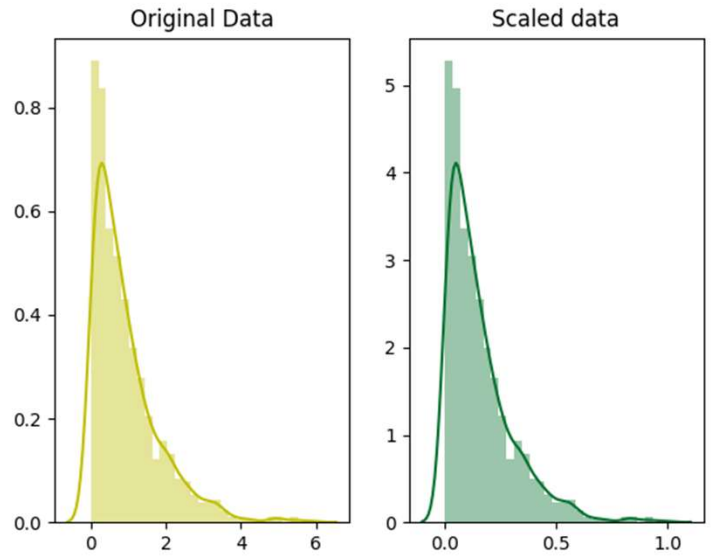
$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$



Since, the range of values of data may vary widely, it becomes a necessary step in data preprocessing while using machine learning algorithms.

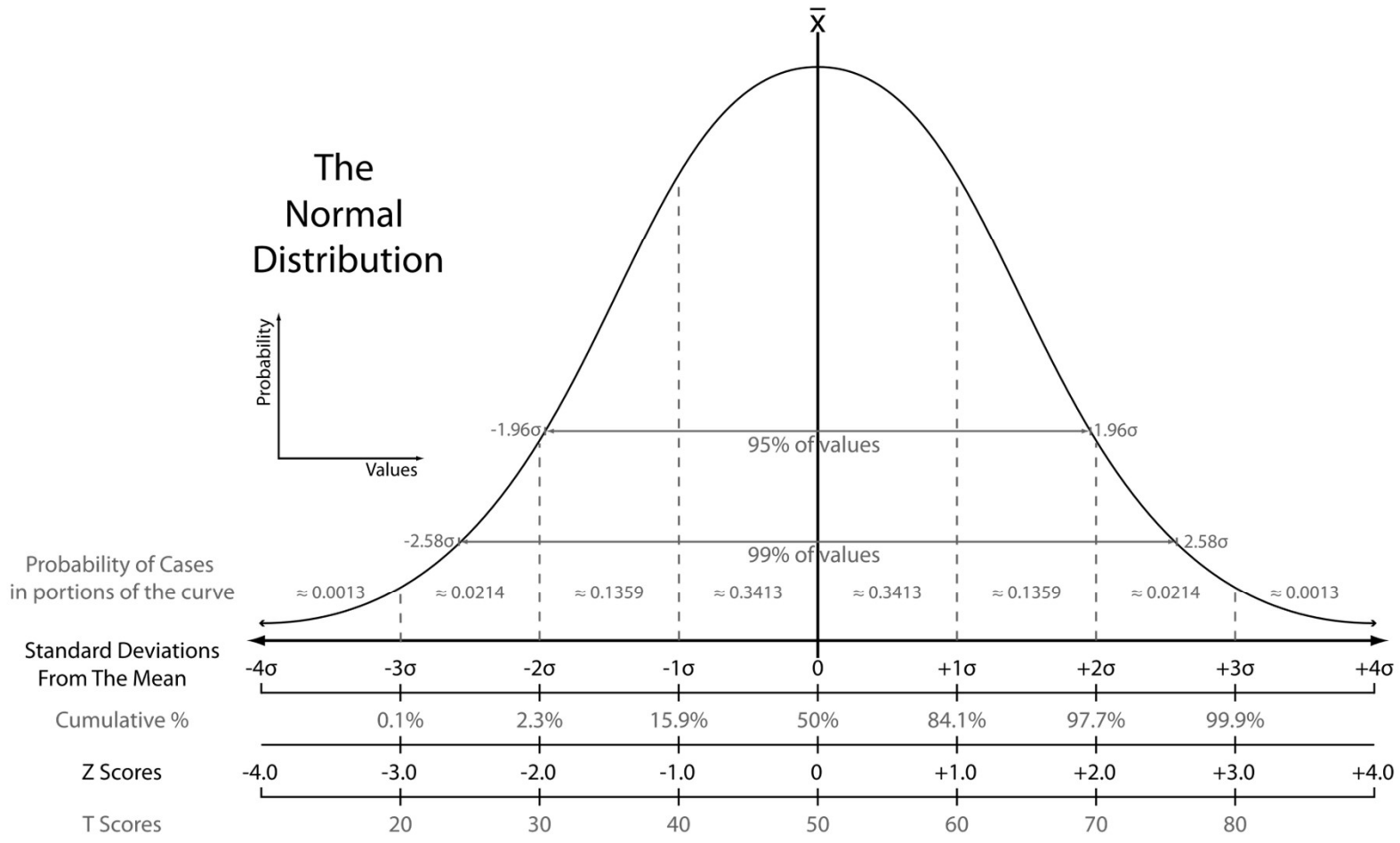
Scaling just changes the range of your data.

Good practice: always scale continuous data



Normalization and Standardization

The point of normalization is to change your observations so that they can be described as a normal distribution.

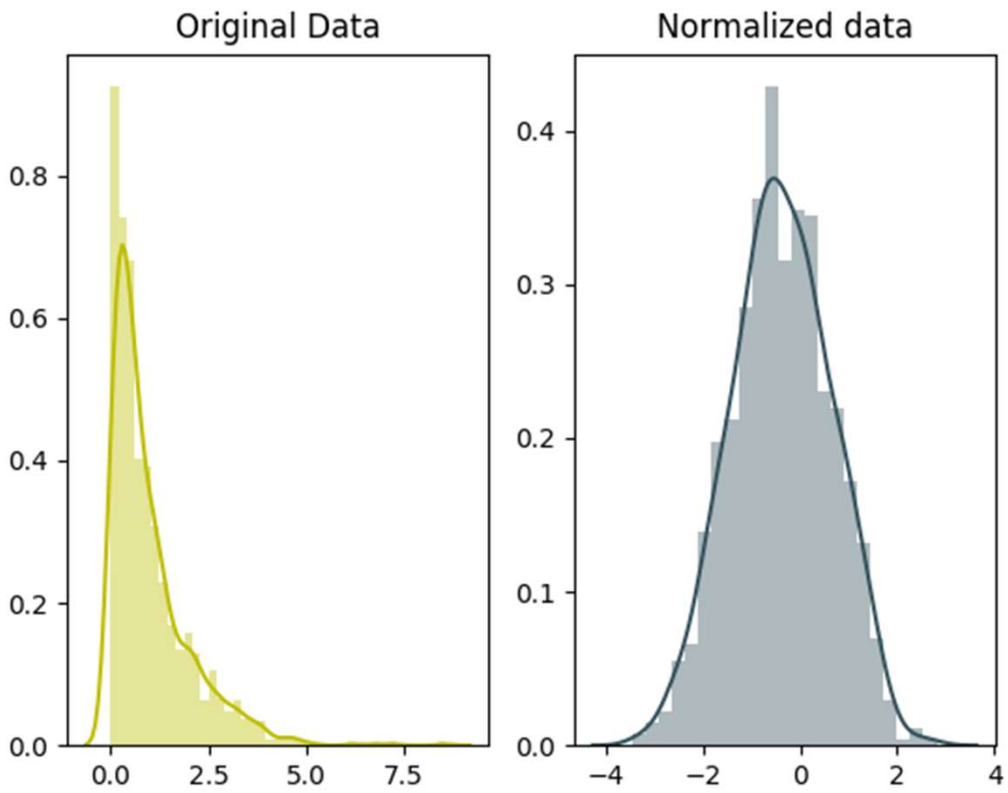

The Normal Distribution
Probability
Values
$\bar{x}$
-1.96σ
1.96σ
95% of values
-2.58σ
2.58σ
99% of values
Probability of Cases in portions of the curve
≈ 0.0013
≈ 0.0214
≈ 0.1359
≈ 0.3413
≈ 0.3413
≈ 0.1359
≈ 0.0214
≈ 0.0013
Standard Deviations From The Mean
-4σ
-3σ
-2σ
-1σ
0
+1σ
+2σ
+3σ
+4σ
Cumulative %
0.1%
2.3%
15.9%
50%
84.1%
97.7%
99.9%
Z Scores
-4.0
-3.0
-2.0
-1.0
0
+1.0
+2.0
+3.0
+4.0
T Scores
20
30
40
50
60
70
80

Standardization

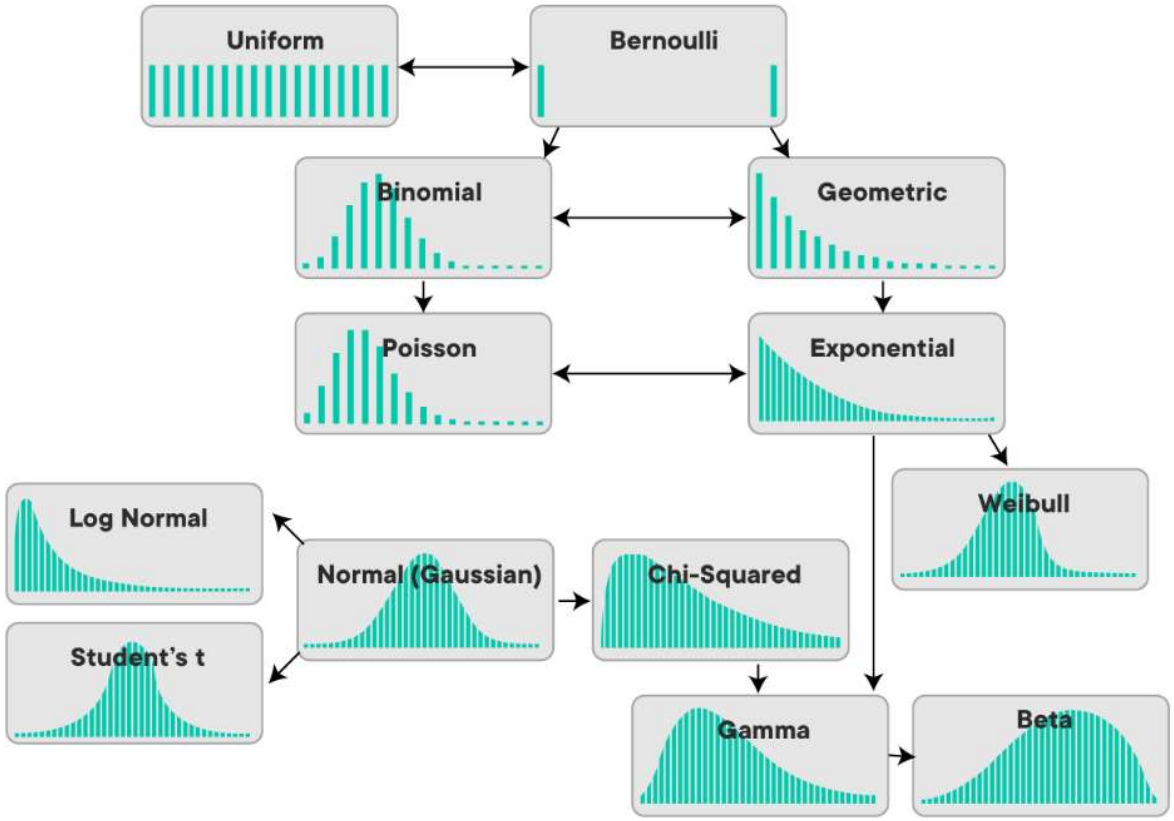
Standardization *(also called **z-score normalization**)* transforms your data such that the resulting distribution has a mean of 0 and a standard deviation of 1.

$$x' = \frac{x - x_{mean}}{\sigma}$$

where x is the original feature vector, x_{mean} is the mean of that feature vector, and σ is its standard deviation.


Original Data
Normalized data
0.0
0.2
0.4
0.6
0.8
0.0
2.5
5.0
7.5
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4

Different Types of Distributions



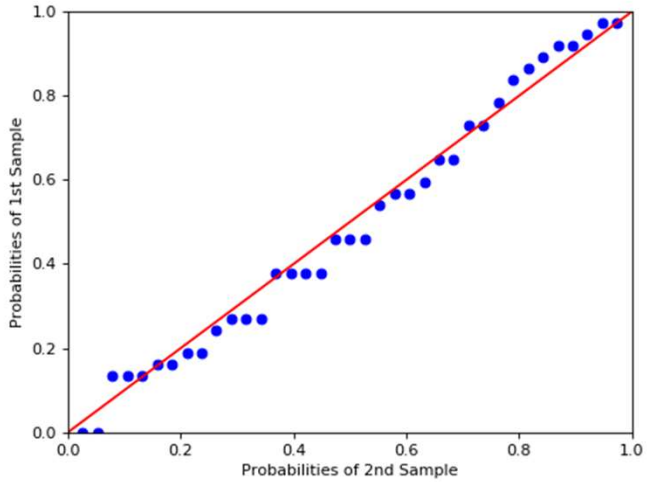
Quantile-Quantile or Q-Q Plots

When the quantiles of two variables are plotted against each other, then the plot obtained is known as quantile – quantile plot or qqplot.

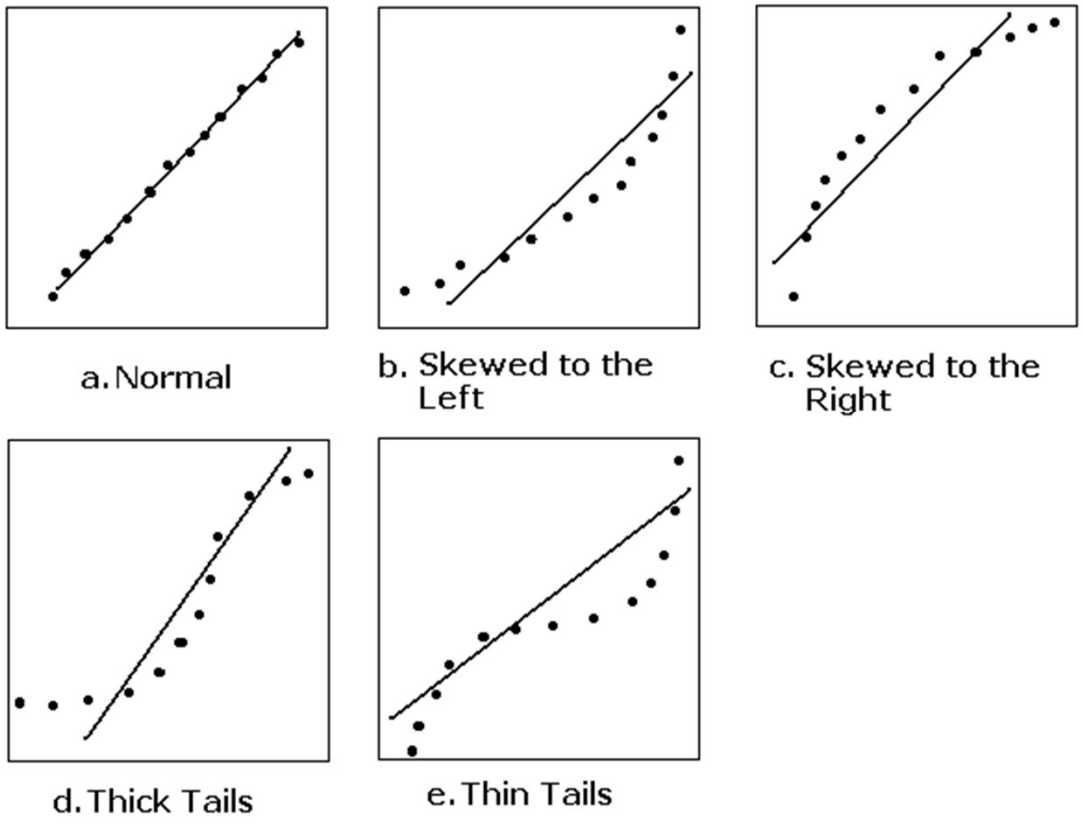
This plot provides a summary of whether the distributions of two variables are similar or not with respect to the locations.

```
import statsmodels.api as sm
from matplotlib import pyplot as plt

sm.qqplot(data_points, fit=True, line ='45')
py.show()
```



Understanding Q-Q Plots



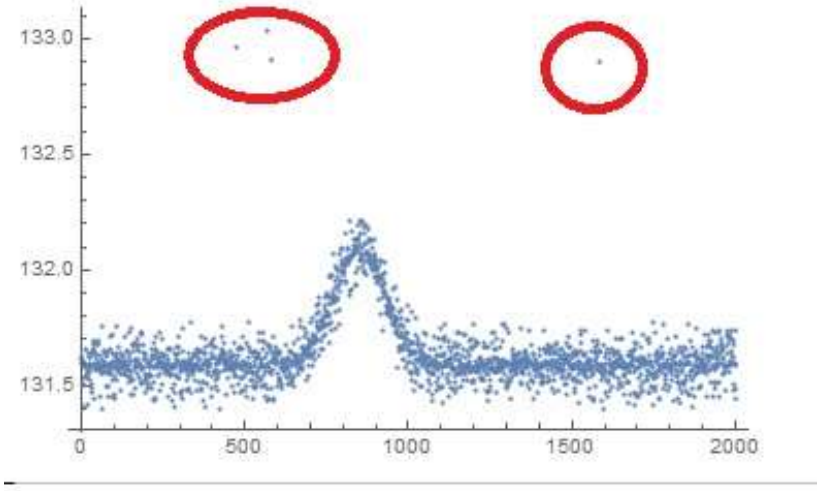
Outliers in Data

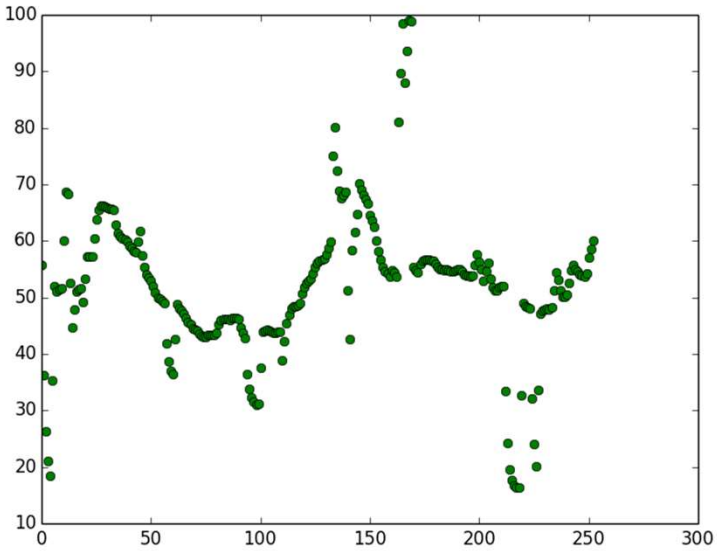
A measurement error or data entry error, correct the error if possible. If you can't fix it, remove that observation because you know it's incorrect.

Not a part of the population you are studying (i.e., unusual properties or conditions), you can legitimately remove the outlier.

Sometimes it's best to keep outliers in your data:

- A natural part of the population you are studying, you should not remove it.
- They can capture valuable information that is part of your study area.


133.0
132.5
132.0
131.5
0
500
1000
1500
2000


100
90
80
70
60
50
40
30
20
10
0
50
100
150
200
250
300

Outliers Affect Statistics