MATHEMATICS FOR COMPUTER SCIENCE ENGINEERS

Unit 1: Web Scraping

Mamatha.H.R

Department of Computer Science and Engineering

PES UNIVERSITY
CELEBRATING 50 YEARS

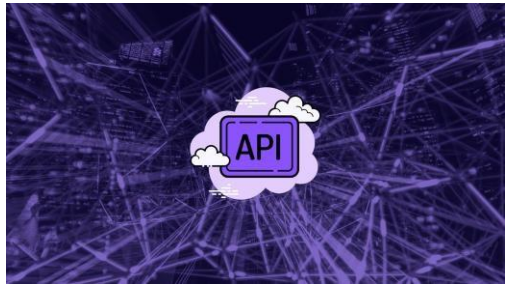
MATHEMATICS FOR COMPUTER SCIENCE ENGINEERS

Getting and Analyzing Data: Scraping the Web

Mamatha H R
Department of Computer Science and Engineering

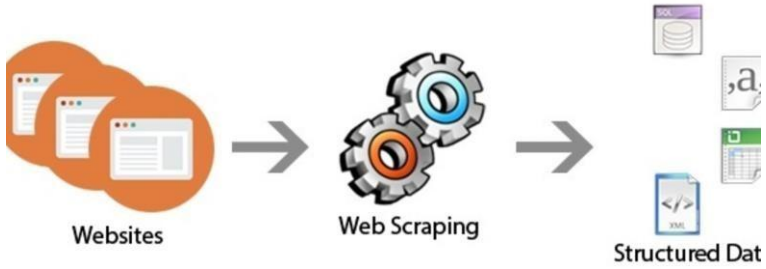
Web Scraping

How to extract Data from Websites?



- There are mainly two methods to get data from websites namely:

1. Using the website's API (If exists)
2. **Using Web scraping**

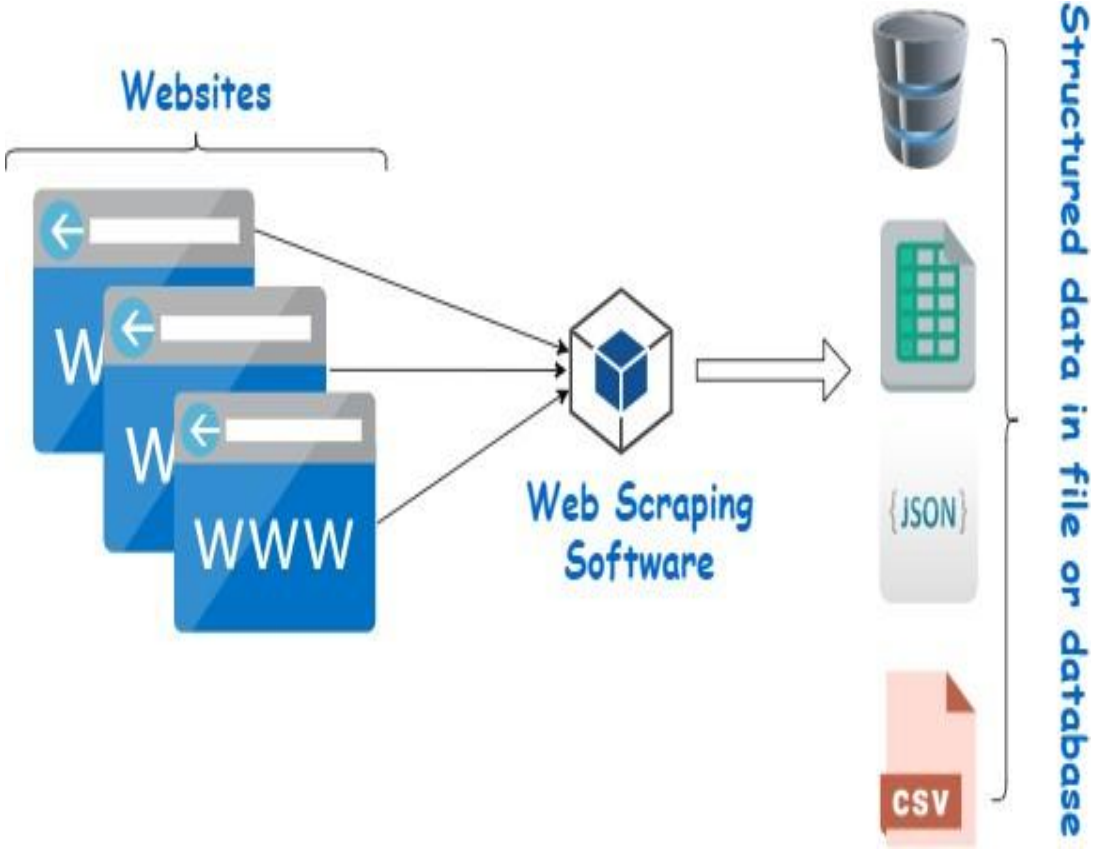


Web Scraping

What is Web Scraping?

- Web scraping is a term used to describe the use of a program or algorithm to extract and process large amounts of data from the web.
- Some popular python libraries for web scraping in Python are:
  - Beautiful Soup
  - Requests
  - Scrapy

and so on…



Web Scraping

Why Web Scraping?

Web Scraping is the **technique of automating** this **process.**

Web Scraping

How vital is Web Scraping

- The Internet would be far less useful and terribly small without Web Scraping.
- The lack of availability of “real integration” through APIs has turned Web Scraping into a massive industry with trillions of dollars in impact on the Internet economy.
- *There is an enormous amount of data “available” on the Internet but it is hardly “accessible”.*
- **Web scraping makes this data accessible** to all kinds of applications and uses

PES UNIVERSITY
CELEBRATING 50 YEARS

MATHEMATICS FOR COMPUTER SCIENCE ENGINEERS

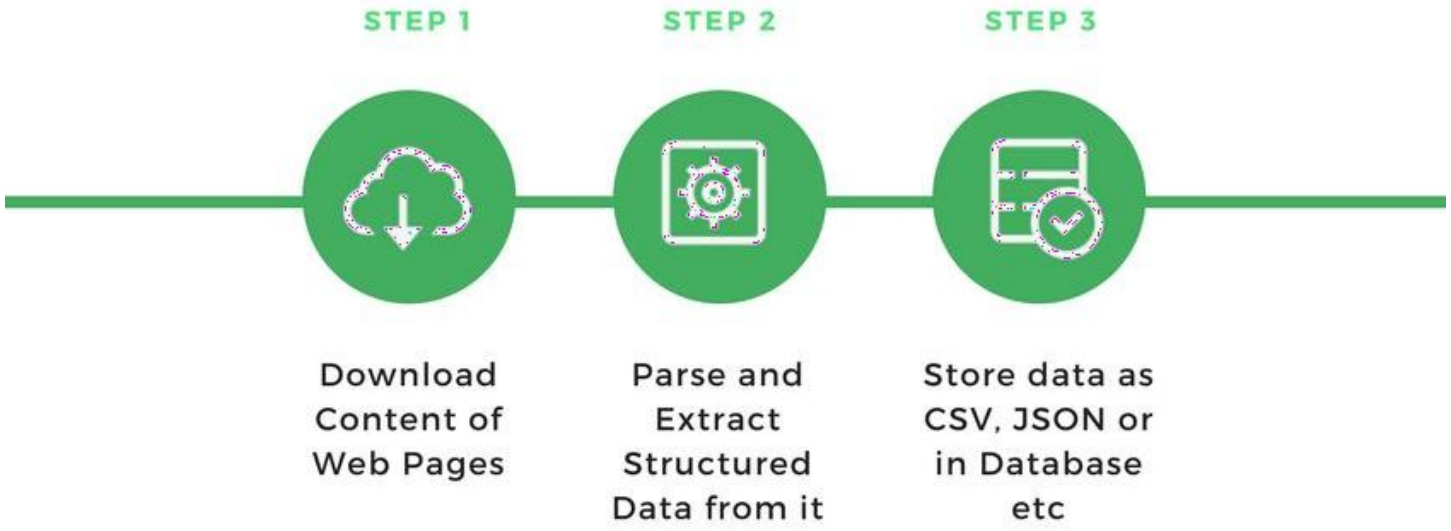
Web Crawling vs. Web Scraping

- Web Crawling mostly refers to downloading and storing the contents of a large number of websites, by following links in web pages.
- Search Engines depend heavily on web crawlers.
- **Googlebot** is an example of a web crawler.

Web Scraping and web crawling

How does a web scraper work?

- A web scraper is a software program or script that is used to download the contents (usually text based and formatted as HTML) of multiple web pages and then extract data from it.



Steps involved in Scraping

The web scraping process follows the below 3 steps.

1. Request-Response
2. Parse and Extract
3. Transform the data

Third-party python library called **Beautiful Soup** is used for pulling data out of HTML and XML files.

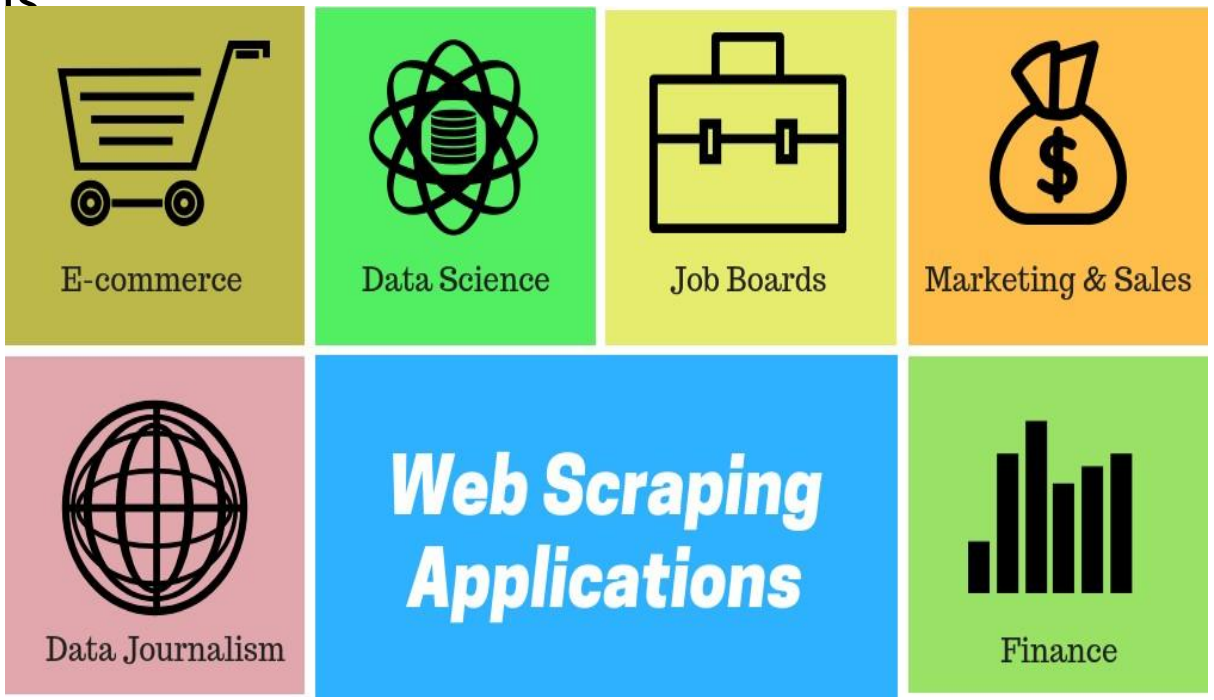
Web Scraping

What Web Scraping is used for?

- Web Scraping has multiple applications across various industries such as:

1. Price and Competition monitoring
2. Market Research (such as Lead Generation for Marketing)
3. Data Analysis
4. E-Commerce
5. Academic Research

and so on…..



Web Scraping

Is Web Scraping legal?

- Some websites allow web scraping and some don't
- Scraping data without the permission of the owner is illegal
- To know whether a website allows web scraping or not, you can look at the website's "robots.txt" file.
- Example:

  So, to see the "robots.txt" file for flipkart.com the following URL is used: www.flipkart.com/robots.txt.

Web Scraping

Basic Steps for Web Scraping

- To extract data using web scraping with python, you need to follow these basic steps:

1.Find the URL that you want to scrape 2.Inspecting the Page

3. Find the data you want to extract
4. Write the code

5.Run the code and extract the data

6.Store the data in the required

Do's and Don'ts of Web Scraping

- Do's:

1. Inspect robots.txt before web scraping
2. Identify yourself

- Don'ts:

1. Don't overburden a website
2. Don't violate copyrights
3. Don't use illegal methods to get what you want

MATHEMATICS FOR COMPUTER SCIENCE ENGINEERS

PES UNIVERSITY
CELEBRATING 50 YEARS

Reading Files

MATHEMATICS FOR COMPUTER SCIENCE ENGINEERS

How to read csv files in Python?

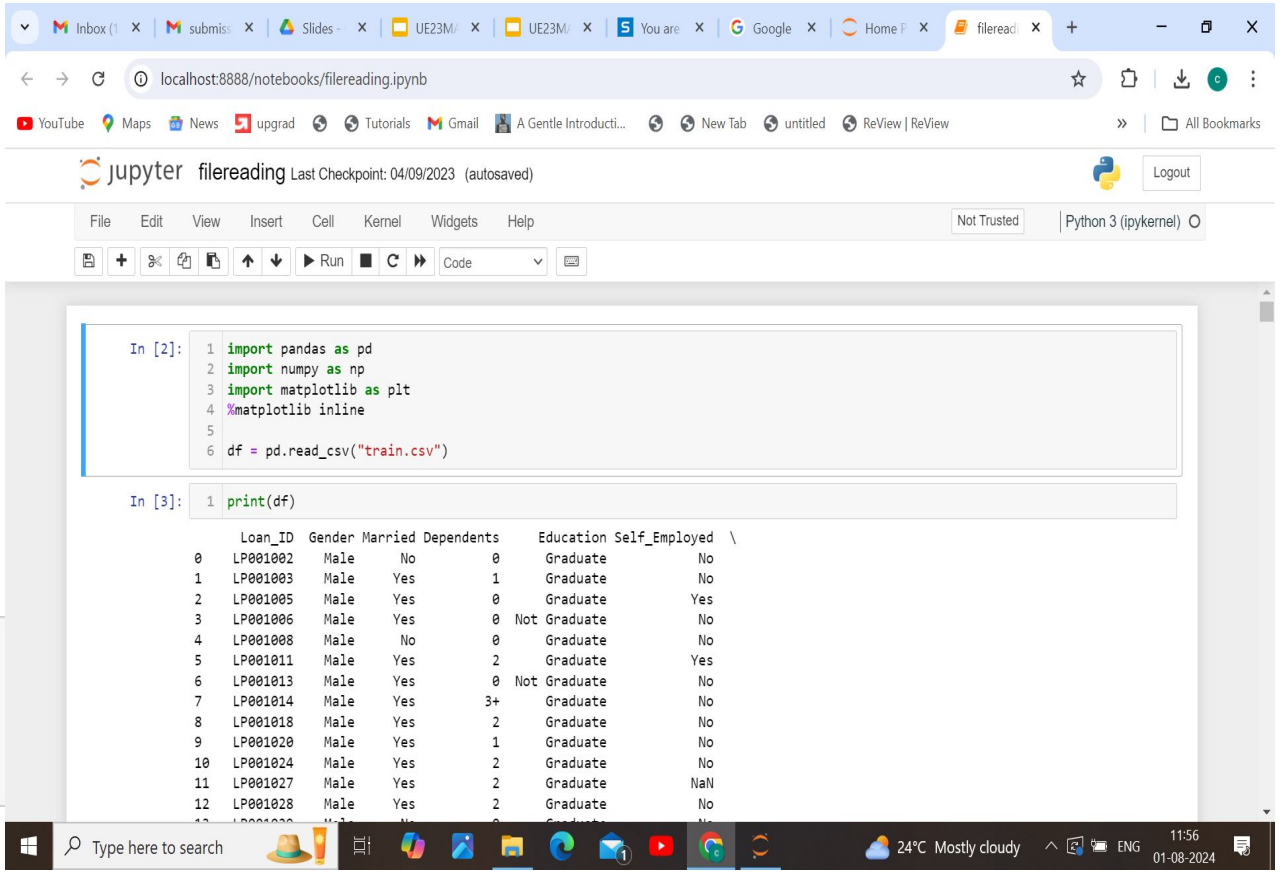
- There are multiple libraries to read .csv files in python such as:

1. Pandas
2. csv library etc

- Using Pandas a csv file can be easily read as a pandas Dataframe using very few lines of code as shown in the below code segment

```
import pandas as pd
df = pd.read_csv('path to csv file')
print(df)
```

Sample Output for reading a CSV File

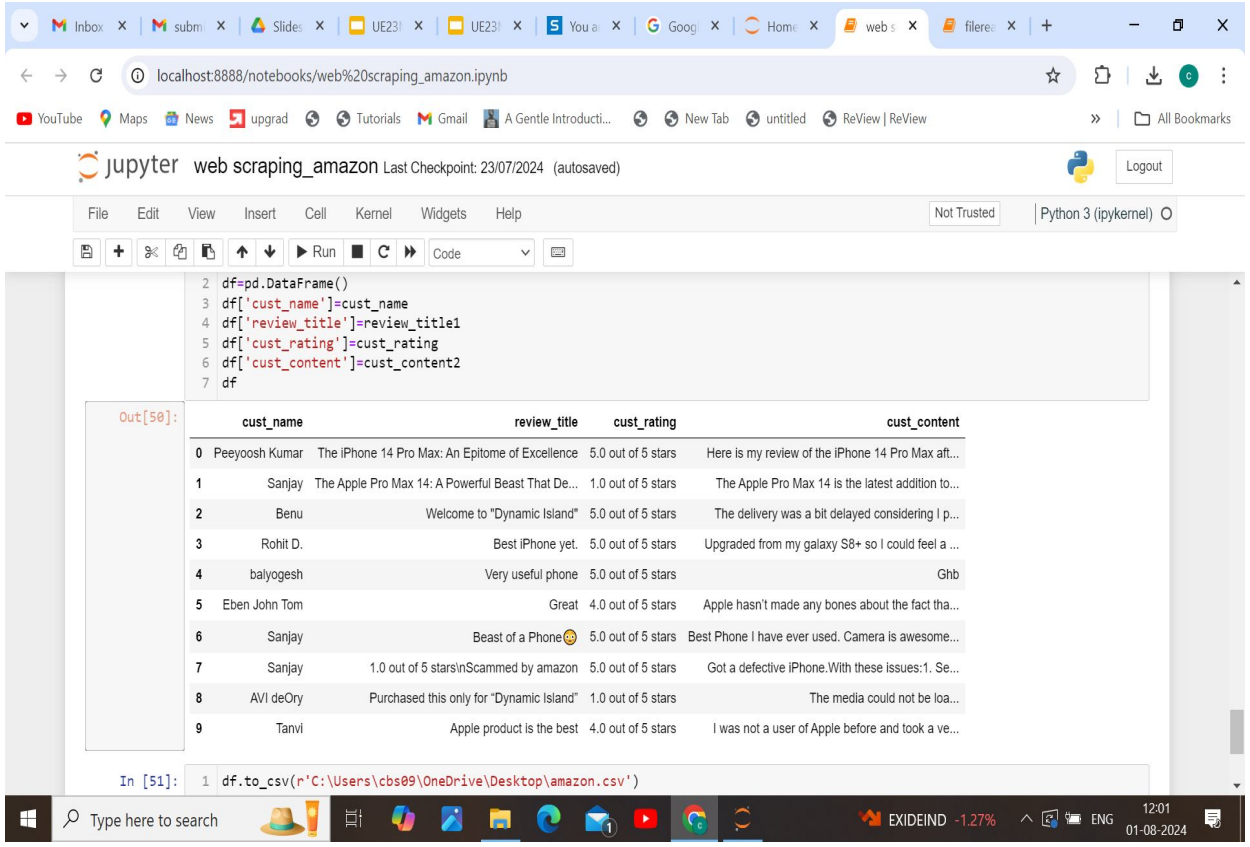


How to write csv files in Python?

- There are multiple libraries to read .csv files in python such as:

1. Pandas
2. csv library etc

- Using Pandas a csv file can be easily read as a pandas Dataframe using very few lines of code as shown in the below code segment



THANK YOU

Dr.Mamatha H R

Professor, Department of Computer Science

mamathahr@pes.edu

+91 80 2672 1983 Extn 834

MATHEMATICS FOR COMPUTER SCIENCE ENGINEER

Data Cleaning

D. Uma

Department of Computer Science and Engineering

MATHEMATICS FOR COMPUTER SCIENCE ENGINEER

Data Cleaning

D. Uma
Department of Computer Science and Engineering

Data Quality

Suppose you have a dataset/database sitting in front of you, and I ask

"**Is it a good quality dataset/database?**"

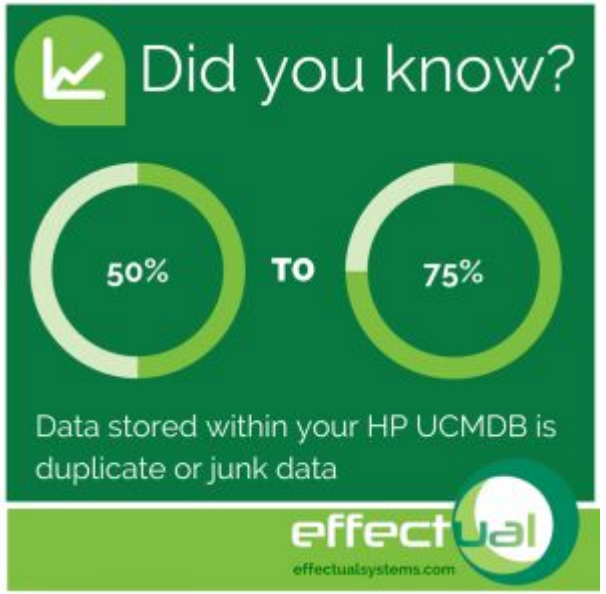
This is **about the Data** themselves, not the system in use to access it.

Sources: www.youtube.com, aibook.in

Data Quality

Data in the Real World is Dirty:

Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, transmission error.



Incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data.

e.g., *Occupation*=“ ” (missing data)

Noisy: containing noise, errors, or outliers

e.g., *Salary*=“−10” (an error)

Sources: Trustinsights, Qgate, Lotame.com

Data Quality

Inconsistent: containing discrepancies in codes or names, e.g.,

Age="42", *Birthday*="03/07/2010"

Was rating "1, 2, 3", now rating "A, B, C"

discrepancy between duplicate records

Intentional (e.g., *disguised missing* data)

Jan. 1 as everyone's birthday?

Sources: Trustinsights, QGate

MATHEMATICS FOR COMPUTER SCIENCE ENGINEER

Data Quality

Sources: Trustinsights, QGate

Data Quality

Improved data quality leads to **better decision making** across an organization.

The **more high-quality data** you have, the **more confidence** you can have in **your decisions**

Good data **decreases risk** and can result in **consistent improvements** in **results.**

Source: www.youtube.com

MATHEMATICS FOR COMPUTER SCIENCE ENGINEER

Data Cleaning

Data cleaning or cleansing is the **process** of **detecting and correcting** (or removing) corrupt or inaccurate records from a record set, table, or database.

It also refers to **identifying incomplete**, **incorrect, inaccurate or irrelevant parts** of the data and then replacing, modifying, or deleting the dirty or coarse data.

Source: towardsdatascience.com

Data Cleaning

- Makes the **data fit** for **purpose/plausible**
- **Reduces** the **negative impact of errors**
- **Improves** the **data quality**
- **Improves** the **quality of the outputs**

Source: www.youtube.com

Data Cleaning

PROCESS OF CLEANING DATA

- Detect
- Resolve
- Treat

Source:

Data Quality

- **Identify erroneous or suspicious data**

  - Graph or sort data - look at outliers

  - I have a **student who throws ten dice** and records the number of sixes. They recorded:

  (2, 0, 3, 12, 2, 0, 1, 1, 3, 1, 4).

  - What is wrong?
  - What do you think is the cause of it?

Source:
www.youtube.com

MATHEMATICS FOR COMPUTER SCIENCE ENGINEER

Data Cleaning

- Consider the data points
  - 3, 4, 7, 4, 8, 3, 9, 5, 7, 6, **92**
  - **"92" is suspicious** - an **outlier**

- Outliers:
  - are potentially legitimate (correct)
  - can be data or model glitches
  - can be a data miners dream, for example, a highly profitable customer

- **Outlier - "departure from the expected"**

Source:

Data Cleaning

RESOLVE

- Deciding if **erroneous or suspicious data** should be **corrected** or amended
- Deciding on the action to "**treat**" the data

Source:

Data Cleaning

WHAT TO LOOK FOR ?

- **Non-response**
  - an item non response
  - Eg. missing data
- **Erroneous data**
  - Can negatively affect data and resulting quality
- **Suspicious data**

Source:
www.youtube.com

MATHEMATICS FOR COMPUTER SCIENCE ENGINEER

Data Cleaning

- **Missing Data**
- **Irregular Data (Outliers)**
- **Unnecessary Data** — Repetitive Data, Duplicates and more
- **Inconsistent Data** — Capitalization, Addresses and more

Source: towardsdatascience.com

MATHEMATICS FOR COMPUTER SCIENCE ENGINEER

Handling Missing Data

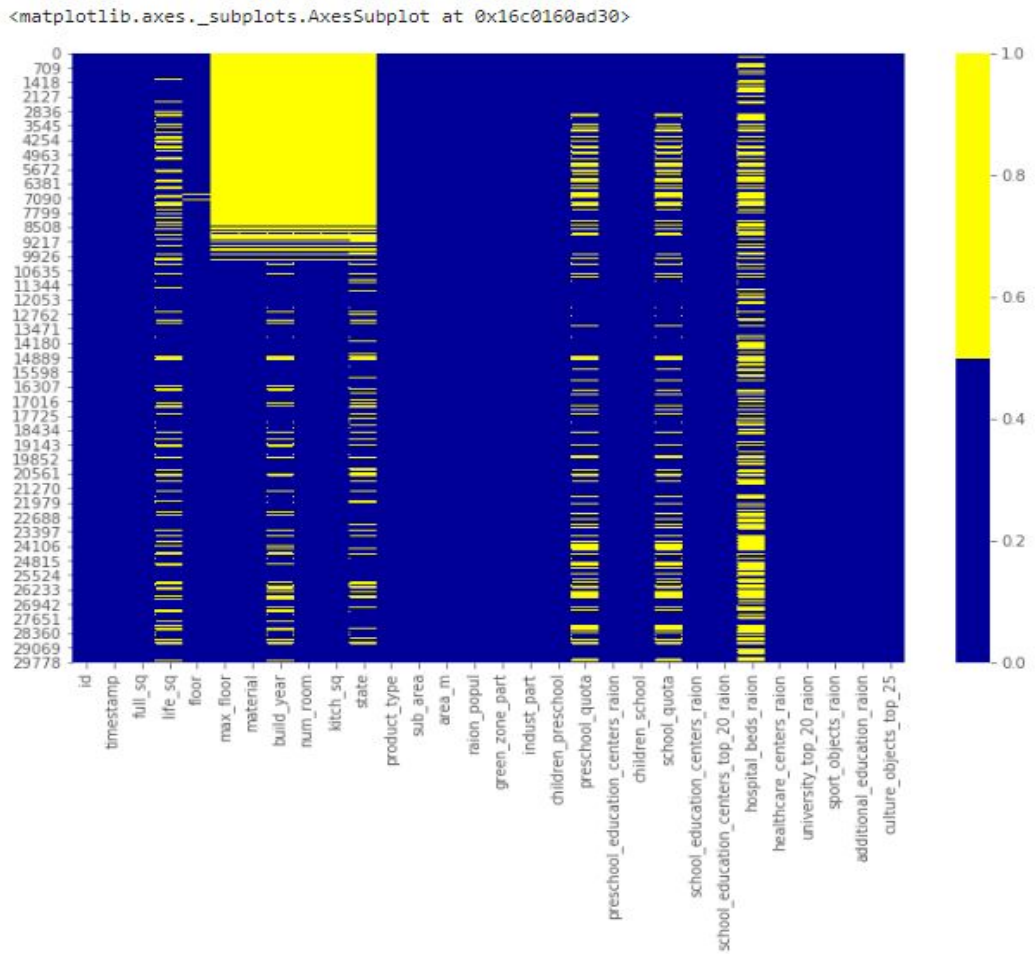
PES UNIVERSITY
CELEBRATING 50 YEARS

Technique : Missing Data Heat map

The chart demonstrates the missing data patterns of the **first 30 features**.

The horizontal axis shows the feature name; the vertical axis shows the number of observations/rows.

The **yellow color** represents the missing data while the blue color otherwise.



Source: towardsdatascience.com

Handling Missing Data

Technique : Missing Data Percentage List

When there are **many features** in the dataset, we can make a list of **missing data %** for each feature.

This produces a list showing the percentage of missing values for each of the features.

```
id - 0.0%
timestamp - 0.0%
full_sq - 0.0%
life_sq - 21.0%
floor - 1.0%
max_floor - 31.0%
material - 31.0%
build_year - 45.0%
num_room - 31.0%
kitch_sq - 31.0%
state - 44.0%
product_type - 0.0%
sub_area - 0.0%
area_m - 0.0%
raion_popul - 0.0%
green_zone_part - 0.0%
indust_part - 0.0%
children_preschool - 0.0%
preschool_quota - 22.0%
preschool_education_centers_raion - 0.0%
children_school - 0.0%
school_quota - 22.0%
school_education_centers_raion - 0.0%
school_education_centers_top_20_raion - 0.0%
hospital_beds_raion - 47.0%
healthcare_centers_raion - 0.0%
university_top_20_raion - 0.0%
sport_objects_raion - 0.0%
additional_education_raion - 0.0%
culture_objects_top_25 - 0.0%
```

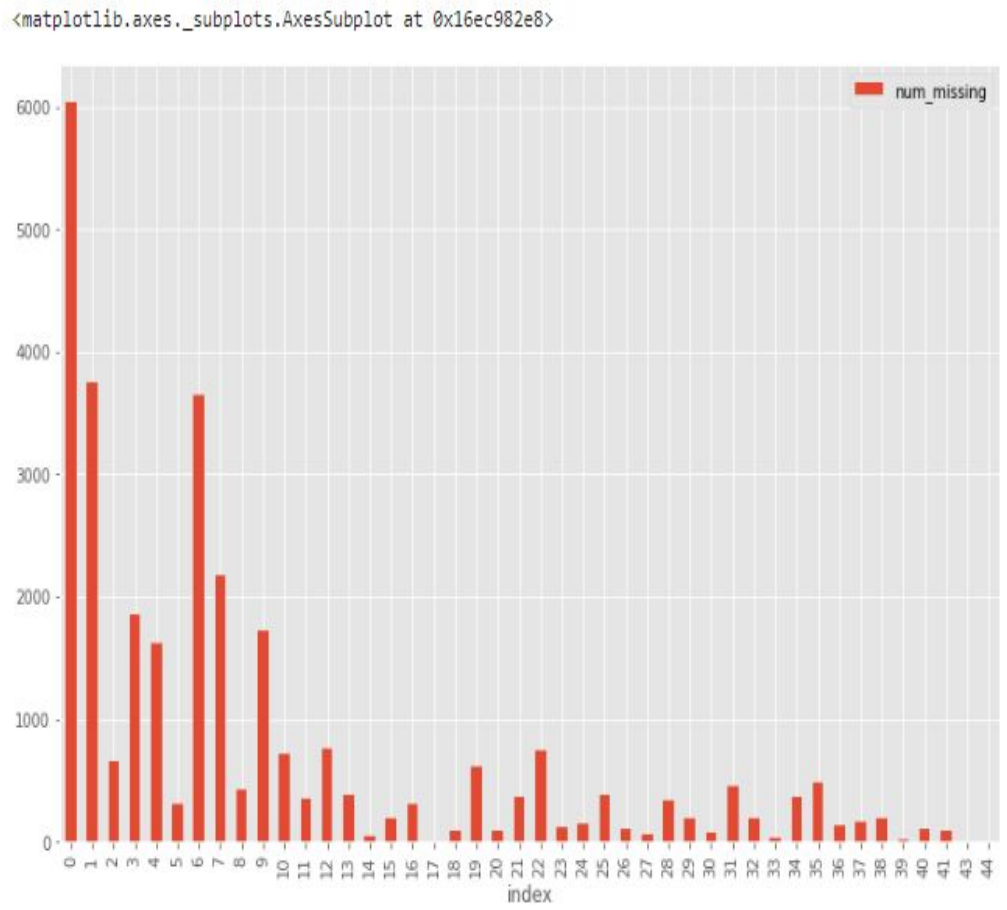
Source: towardsdatascience.com

MATHEMATICS FOR COMPUTER SCIENCE ENGINEER

Handling Missing Data

Technique : Missing Data Histogram

Missing data histogram is also a technique for when we have many features.



Source: towardsdatascience.com

Handling Missing Data

Solution : Drop the Observation

In statistics, this method is called **the *listwise* deletion technique**.

In this solution, we drop the entire observation as long as it contains a missing value.

Only if we are sure that the **missing data** is **not informative**, we perform this. Otherwise, we should consider other solutions.

Source: towardsdatascience.com

Handling Missing Data

Solution : Drop the Feature

Similar to previous one, we only do this when we are confident that this feature doesn't provide useful information.

For example, from the missing data % list, we notice that hospital_beds_raion has a high missing value percentage of 47%. We may drop the entire feature.

Index	Age	Sex	Income
1	NA	M	NA
2	39	NA	75000
3	NA	NA	NA
4	28	F	50000
...	...	...	...
10000	18	F	NA

Source: towardsdatascience.com, Medium

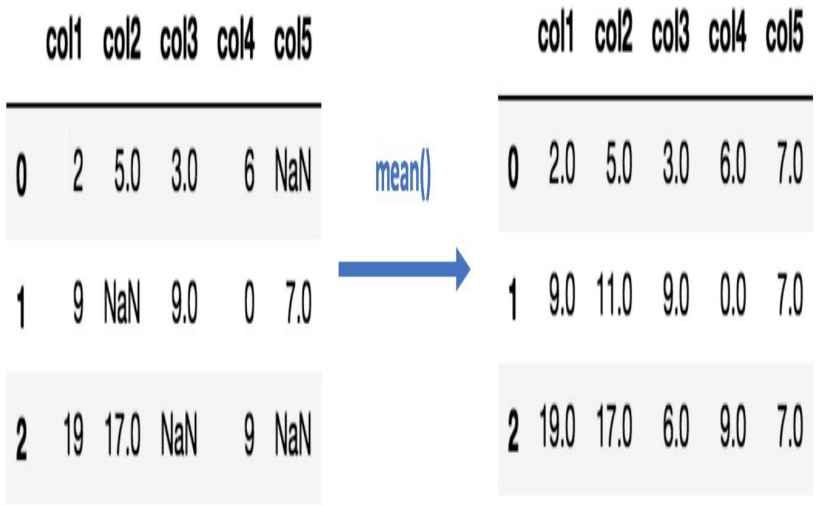
Handling Missing Data

Solution : Impute the Missing

When the feature is a **numeric variable**, we can conduct missing data imputation.

We replace the missing values with the **average** or **median** value from the data of the same feature that is not missing.

When the feature is a **categorical variable**, we may impute the missing data by the **mode** (the most frequent value)



Source: towardsdatascience.com

Handling Missing Data

Solution : Replace the Missing

For **categorical features**, we can add a new category with a value such as "**_MISSING_**".

For **numerical features**, we can replace it with a particular value such as -**999**.

This way, we are still keeping the missing values as valuable information.

Source: towardsdatascience.com, Shutterstock.com

Outliers

Irregular data (Outliers)

Outliers are data that is **distinctively different** from other observations.

They could be **real outliers** or **mistakes**.

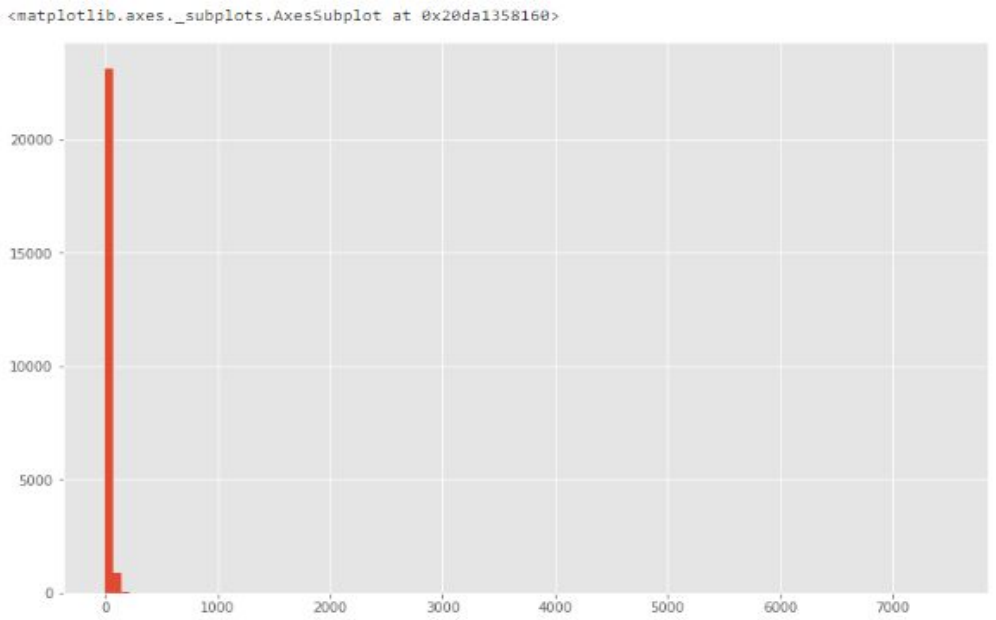
Source: towardsdatascience.com

Outliers

Technique : Histogram/Box Plot

When the feature is numeric, we can use a histogram and box plot to detect outliers.

The data looks highly skewed with the possible existence of outliers.

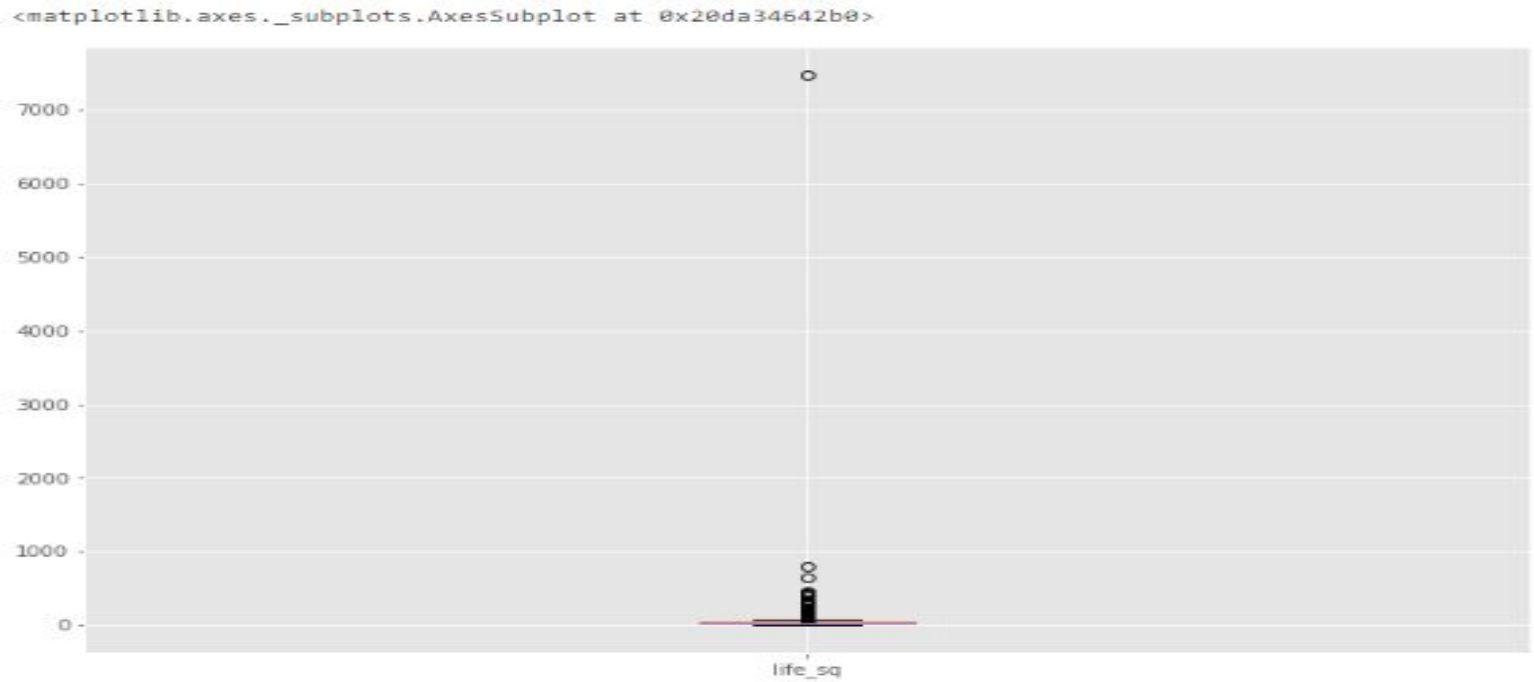


Source: towardsdatascience.com

Outliers

To study the feature closer, let's make a **box plot.**

In this plot, we can see there is an outlier at a **value of over 7000**.



Source: towardsdatascience.com

Outliers

Technique : Descriptive Statistics

For **numeric features**, the outliers could be too distinct that the **box plot can't visualize** them.

Instead, we can look at their **descriptive statistics.**

For example, for the feature *life_sq* again, we can see that the **maximum value is 7478**, while the **75% quartile is only 43**.

The 7478 value is an **outlier**.

```
count    24088.000000
mean        34.403271
std         52.285733
min          0.000000
25%         20.000000
50%         30.000000
75%         43.000000
max       7478.000000
Name: life_sq, dtype: float64
```

Source: towardsdatascience.com

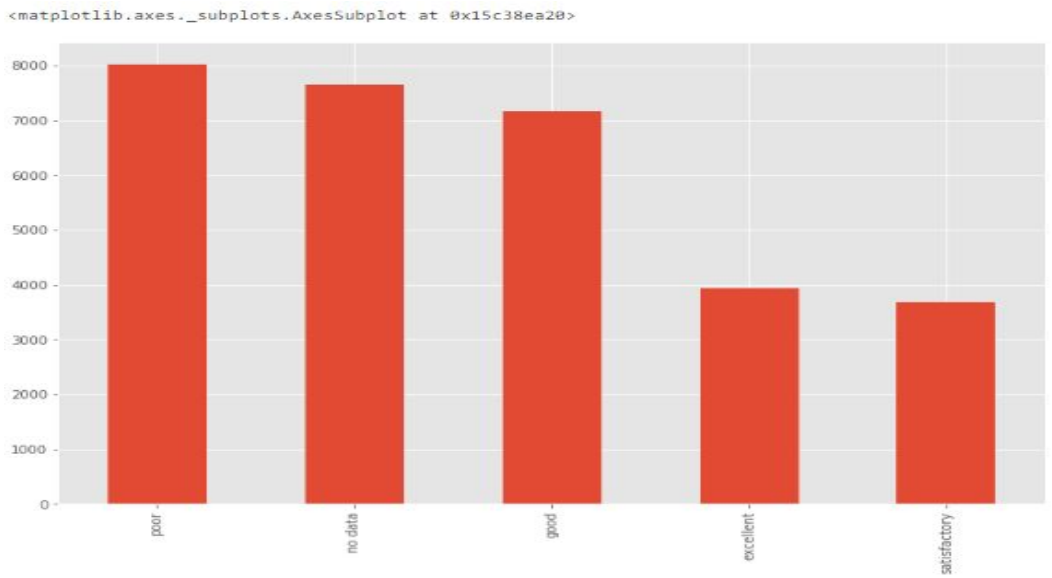
Outliers

Technique : Bar Chart

When the feature is **categorical,** we can use a bar chart to learn about its categories and distribution.

For example, the feature *ecology* has a reasonable distribution.

But if there is a category with **only one value** called “other”, then that would be an **outlier**.



Source: towardsdatascience.com

Unnecessary Data

After all the hard work done for missing data and outliers, let's look at **unnecessary data**, which is more straightforward.

The unnecessary data is when the data **doesn't add value**.

We cover three main types of unnecessary data due to different reasons.

MATHEMATICS FOR COMPUTER SCIENCE ENGINEER

Unnecessary Data

Unnecessary type : Uninformative / Repetitive

Sometimes one feature is uninformative because it has too many rows being the same value.

How to find out?

We can create a list of features with a high percentage of the same value.

For example, we specify below to show features with over 95% rows being the same value.

```
oil_chemistry_raion: 99.02858%
no      30175
yes       296
Name: oil_chemistry_raion, dtype: int64

railroad_terminal_raion: 96.27187%
no      29335
yes      1136
Name: railroad_terminal_raion, dtype: int64

nuclear_reactor_raion: 97.16780%
no      29608
yes       863
Name: nuclear_reactor_raion, dtype: int64

big_road1_1line: 97.43691%
no      29690
yes       781
Name: big_road1_1line, dtype: int64

railroad_1line: 97.06934%
no      29578
yes       893
Name: railroad_1line, dtype: int64

cafe_count_500_price_high: 97.25641%
0     29635
1       787
2        38
3        11
Name: cafe_count_500_price_high, dtype: int64

mosque_count_500: 99.51101%
0     30322
1       149
Name: mosque_count_500, dtype: int64

cafe_count_1000_price_high: 95.52689%
0     29108
1      1104
2       145
3        51
4        39
5        15
6         8
7         1
Name: cafe_count_1000_price_high, dtype: int64

mosque_count_1000: 98.08342%
0     29887
1       584
Name: mosque_count_1000, dtype: int64

mosque_count_1500: 96.21936%
0     29319
1      1152
Name: mosque_count_1500, dtype: int64
```

Source: towardsdatascience.com

MATHEMATICS FOR COMPUTER SCIENCE ENGINEER

Unnecessary Data

Unnecessary Type :

- **Irrelevant**
- **Duplicates**

timestamp	full_sq	life_sq	floor	build_year	num_room	price_doc	
2014-12-09	40	-999.0	17.0	-999.0	1.0	4607265	2
2014-04-15	134	134.0	1.0	0.0	3.0	5798496	2
2013-08-30	40	-999.0	12.0	-999.0	1.0	4462000	2
2012-09-05	43	-999.0	21.0	-999.0	-999.0	6229540	2
2013-12-05	40	-999.0	5.0	-999.0	1.0	4414080	2
2014-12-17	62	-999.0	9.0	-999.0	2.0	6552000	2
2013-05-22	68	-999.0	2.0	-999.0	-999.0	5406690	2
2012-08-27	59	-999.0	6.0	-999.0	-999.0	4506800	2
2013-04-03	42	-999.0	2.0	-999.0	-999.0	3444000	2
2015-03-14	62	-999.0	2.0	-999.0	2.0	6520500	2
2014-01-22	46	28.0	1.0	1968.0	2.0	3000000	2
2012-10-22	61	-999.0	18.0	-999.0	-999.0	8248500	2
2013-09-23	85	-999.0	14.0	-999.0	3.0	7725974	2
2013-06-24	40	-999.0	12.0	-999.0	-999.0	4112800	2
2015-03-30	41	41.0	11.0	2016.0	1.0	4114580	2
2013-12-18	39	-999.0	6.0	-999.0	1.0	3700946	2
2013-08-29	58	58.0	13.0	2013.0	2.0	5764128	1
	50	33.0	2.0	1972.0	2.0	8150000	1
	52	30.0	9.0	2006.0	2.0	10000000	1
2013-08-30	38	17.0	15.0	2004.0	1.0	6400000	1

Name: id, dtype: int64

There are 16 duplicates based on this set of key features.

Inconsistent Data

Inconsistent : Capitalization

What to do?

To avoid this, we can put all letters to lower cases (or upper cases).

```
Poselenie Sosenskoe                    1776
Nekrasovka                             1611
Poselenie Vnukovskoe                   1372
Poselenie Moskovskij                    925
Poselenie Voskresenskoe                 713
                                        ...
Molzhaninovskoe                           3
Poselenie Kievskij                        2
Poselenie Shhapovskoe                     2
Poselenie Mihajlovo-Jarcevskoe            1
Poselenie Klenovskoe                      1
Name: sub_area, Length: 146, dtype: int64
```

Source: towardsdatascience.com

MATHEMATICS FOR COMPUTER SCIENCE ENGINEER

Inconsistent Data

Inconsistent Type : Formats

Another standardization we need to perform is the data formats.

One example is to convert the feature from string to DateTime format.

How to find out?

The feature *timestamp* is in string format while it represents dates.

	id	timestamp	full_sq	life_sq	floor	max_floor	material	build_year	num_room	kitch_sq	...	cafe_count_5000_price_high	big_church_count_5000	church_count_5000	mosque_count_5000	leisure_count_5000	sport_count_5000	market_count_5000	price_doc	sub_area_lower	ecology_new
0	1	2011-08-20	43	27.0	4.0	NaN	NaN	NaN	NaN	NaN	...	0	13	22	1	0	52	4	5850000	bibirevo	good_or_better
1	2	2011-08-23	34	19.0	3.0	NaN	NaN	NaN	NaN	NaN	...	0	15	29	1	10	66	14	6000000	nagatinskij zaton	good_or_better
2	3	2011-08-27	43	29.0	2.0	NaN	NaN	NaN	NaN	NaN	...	0	11	27	0	4	67	10	5700000	tekstil'shhiki	poor
3	4	2011-09-01	89	50.0	9.0	NaN	NaN	NaN	NaN	NaN	...	1	4	4	0	0	26	3	13100000	mitino	good_or_better
4	5	2011-09-05	77	77.0	4.0	NaN	NaN	NaN	NaN	NaN	...	17	135	236	2	91	195	14	16331452	basmannoe	good_or_better
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
30466	30469	2015-06-30	44	27.0	7.0	9.0	1.0	1975.0	2.0	6.0	...	0	15	26	1	2	84	6	7400000	otradnoe	good_or_better
30467	30470	2015-06-30	86	59.0	3.0	9.0	2.0	1935.0	4.0	10.0	...	24	98	182	1	82	171	15	25000000	tverskoe	poor
30468	30471	2015-06-30	45	NaN	10.0	20.0	1.0	NaN	1.0	1.0	...	0	2	12	0	1	11	1	6970959	poselenie vnukovskoe	no data
30469	30472	2015-06-30	64	32.0	5.0	15.0	1.0	2003.0	2.0	11.0	...	1	6	31	1	4	65	7	13500000	obruchevskoe	satisfactory
30470	30473	2015-06-30	43	28.0	1.0	9.0	1.0	1968.0	2.0	6.0	...	0	7	16	0	9	54	10	5600000	novogireevo	poor

30471 rows × 294 columns

Source: towardsdatascience.com

MATHEMATICS FOR COMPUTER SCIENCE ENGINEER

Inconsistent Data

What to do?

We can convert it and extract the date or time values. After this, it's easier to analyze the transaction volume group by either year or month.

```
2014    13662
2013     7978
2012     4839
2015     3239
2011      753
Name: year, dtype: int64

12    3400
4     3191
3     2972
11    2970
10    2736
6     2570
5     2496
9     2346
2     2275
7     1875
8     1831
1     1809
Name: month, dtype: int64
```

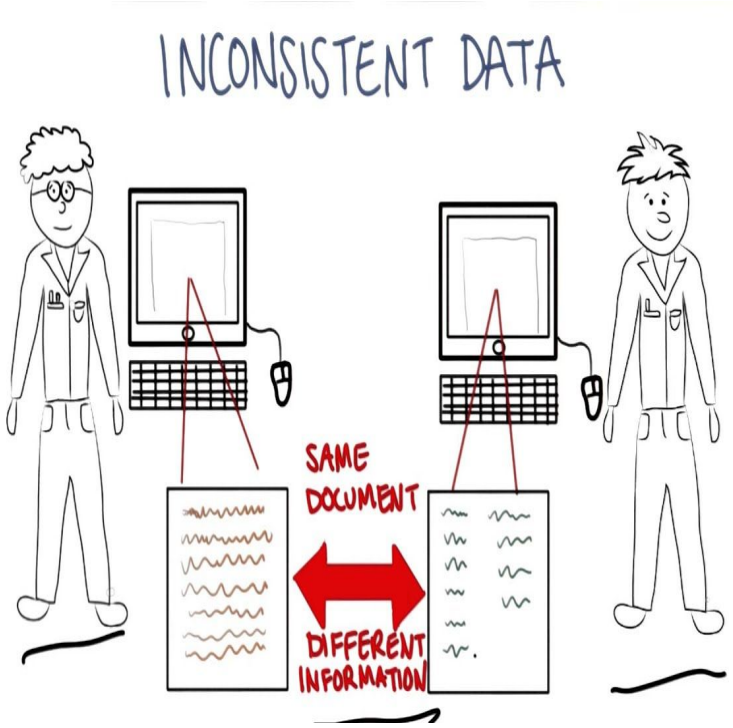
Source: towardsdatascience.com

Inconsistent Data

Inconsistent Type : Categorical Values

Inconsistent categorical values are the last inconsistent type we cover.

A categorical feature has a limited number of values. Sometimes there may be other values due to reasons such as typos.



How to find out?

For instance, the value of *city* was typed by mistakes as “torontoo” and “tronto”. But they both refer to the correct value “toronto”.

Source: towardsdatascience.com

Inconsistent Data

Inconsistent Type : Addresses

The address feature could be a headache for many of us. Because people entering the data into the database often *don't* follow a standard format.

	address
0	123 MAIN St Apartment 15
1	123 Main Street Apt 12
2	543 FirSt Av
3	876 FIRst Ave.

How to find out?

We can find messy address data by looking at it. Even though sometimes we can't spot any issues, we can still run code to standardize them.

	address	address_std
0	123 MAIN St Apartment 15	123 main st apt 15
1	123 Main Street Apt 12	123 main st apt 12
2	543 FirSt Av	543 first ave
3	876 FIRst Ave.	876 first ave

Source: towardsdatascience.com

THANK YOU

D. Uma
Department of Computer Science and Engineering
umaprabha@pes.edu

MATHEMATICS FOR COMPUTER SCIENCE ENGINEERS

Data Visualization and Interpretation

Dr. Mamatha H.R

Department of Computer Science and Engineering
mamathahr@pes.edu

PES UNIVERSITY
CELEBRATING 50 YEARS

MATHEMATICS FOR COMPUTER SCIENCE ENGINEERS

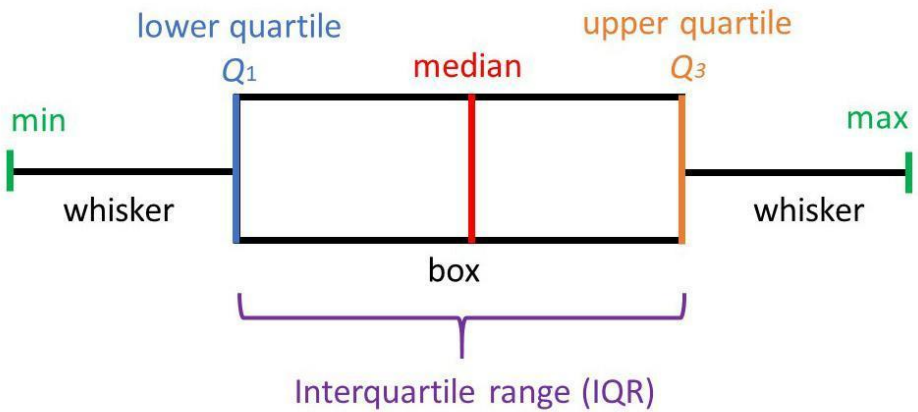
Data Visualization and Interpretation
Boxplot

Dr. Mamatha H.R
Department of Computer Science and Engineering

Box and Whisker Plot

A **box and whisker plot** is a way of summarizing a set of data measured on an **interval scale**.

It shows the distribution of a set of data along a number line, dividing the data into four parts using the median and quartiles.



Source: Simply Psycl

Slide Courtesy: Dr. Uma

Why Boxplot?

It **does not show a distribution** in as much detail as **a stem and leaf plot** or **histogram** does, but is especially useful for indicating whether a distribution is skewed and whether there are potential unusual observations (outliers) in the data set.

Box and whisker plots are also very useful when large numbers of observations are involved and when two or more data sets are being compared.

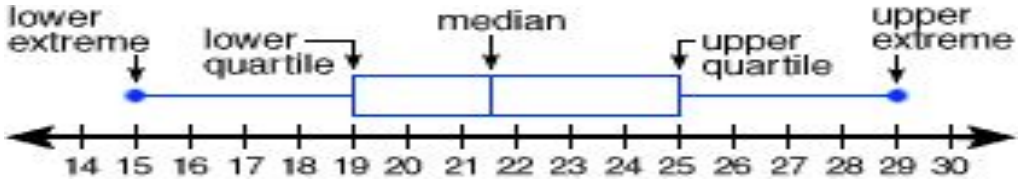
Shows how the values in the data are spread out.

Source: www.150.statcan.gc.ca

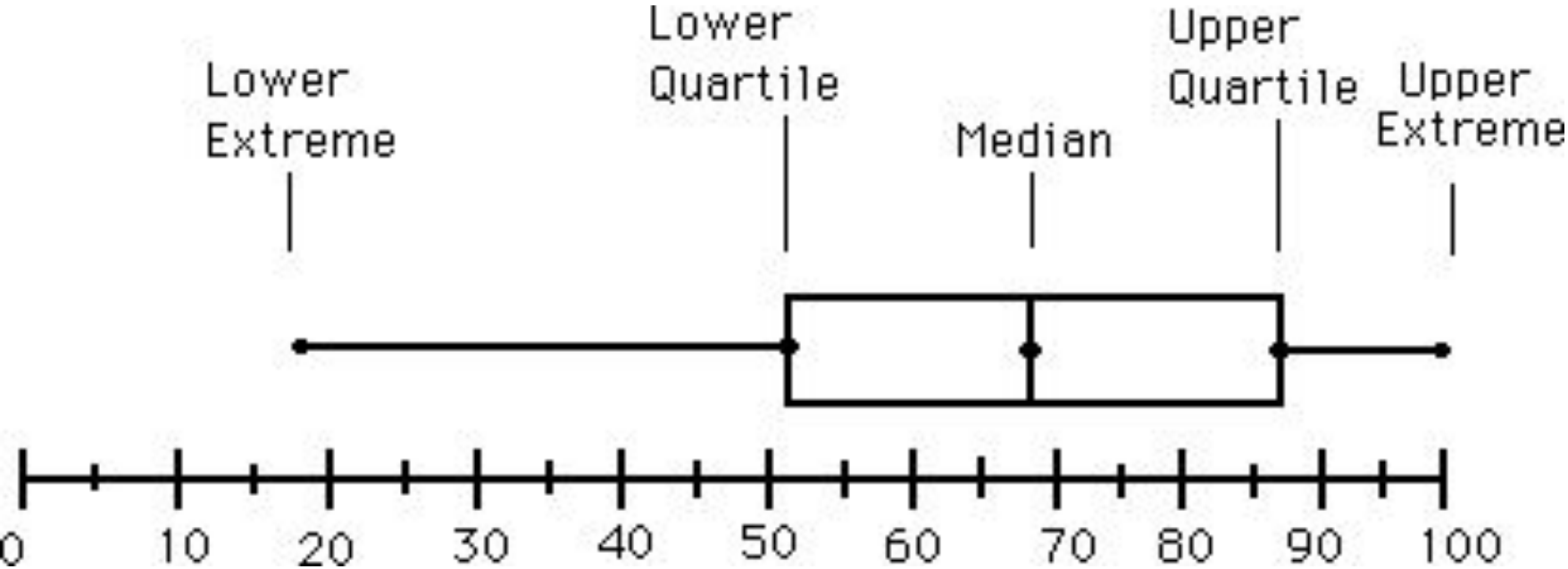
Slide Courtesy: Dr. Uma

Box and Whisker Plot Analysis

- **Five-number summary** of a distribution
  - Minimum, Q1, Median, Q3, Maximum
- Boxplot
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
  - The median is marked by a line within the box
  - **Whiskers**: two lines outside the box extended to Minimum and Maximum
  - **Outliers**: points beyond a specified outlier threshold, plotted individually

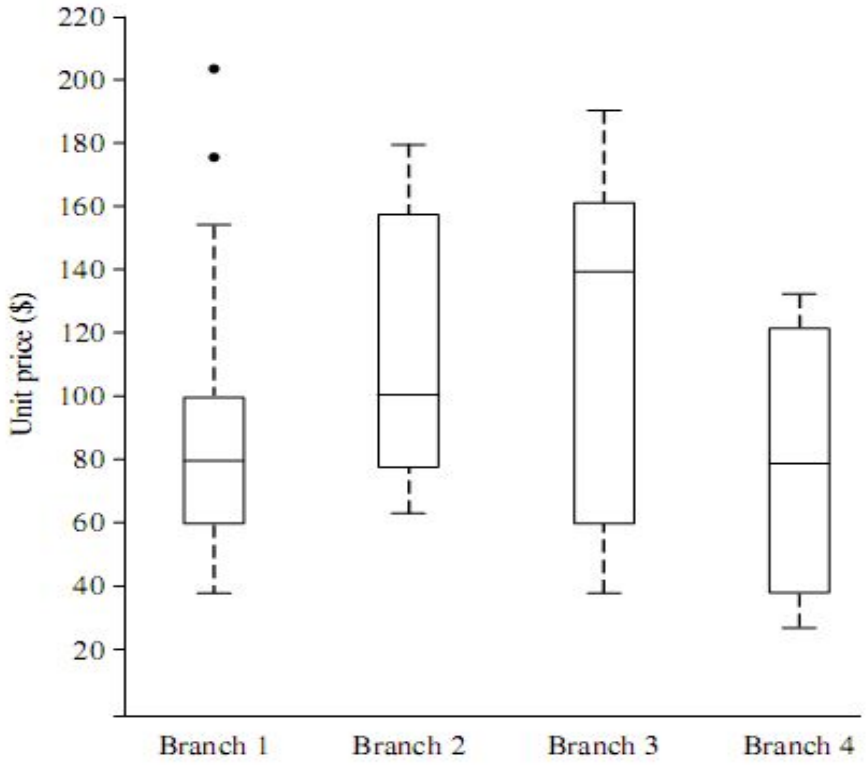


Box and Whisker Plot Analysis - Horizontal



MATHEMATICS FOR COMPUTER SCIENCE ENGINEERS

Box and Whisker Plot Analysis - Vertical



Box and Whisker Plot Analysis

- Boxplot is a method for graphically depicting groups of numerical data through their quartiles.

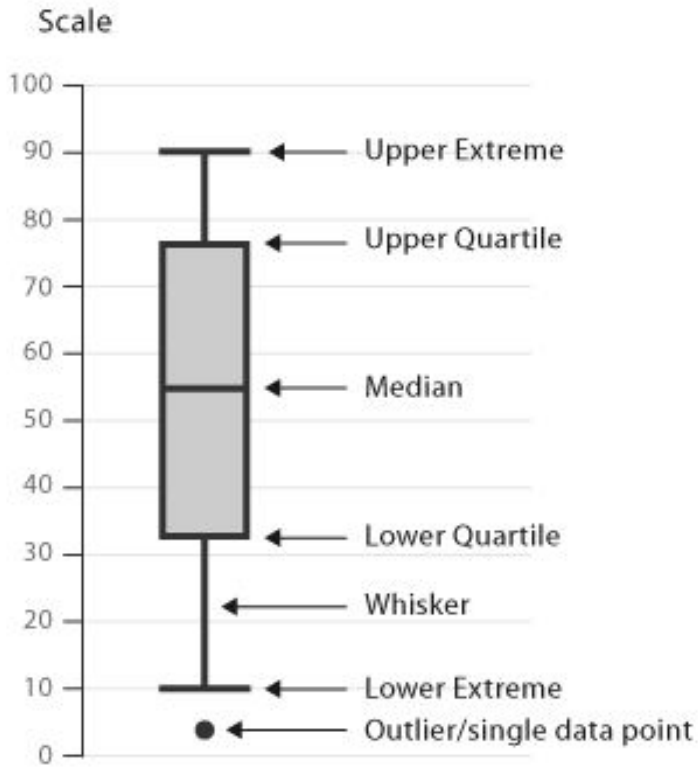
- It's probably the best method to identify the **outliers** in the data.
- The whole data is divided into 4 quartiles, 1st Quartile 2nd Quartile, 3rd Quartile, 4th Quartile.
- 1st Quartile = Q1 – Lower Extreme
- 2nd Quartile = Q2 – Q1
- 3rd Quartile = Q3 – Q2
- 4th Quartile = Upper Extreme – Q3

Box and Whisker Plot Analysis

- Roughly speaking:
- The **"25th percentile"** is the number such that 25% of the data points fall below the number.
- The **"median" or "50th percentile"** is the number such that half of the data points fall below the number.
- The **"75th percentile"** is the number such that 75% of the data points fall below the number.
- **"Whiskers"** are drawn to the most extreme data points that are not more than 1.5 times the length of the box beyond either quartile.
  - Whiskers are useful for identifying outliers.
- **"Outliers,"** or extreme observations, are denoted by asterisks.
  - Generally, data points falling beyond the whiskers are considered outliers.

MATHEMATICS FOR COMPUTER SCIENCE ENGINEERS

Box and Whisker Plot Analysis



Box and Whisker Plot Key Terms

Upper extreme - min (Q3 + 1.5 times IQR, max value in data)

Upper quartile – Q3

Median – Q2

Lower Quartile - Q1

Lower extreme - max (Q1 - 1.5 times IQR, min value in data)

IQR – is interquartile range = Q3 – Q1

Note: A normal distribution has Q1 – Q2 = Q2 – Q3, which means there's equal amount of data spread between Q1 to Q2 and Q2 to Q3. The mean also coincides with Q2 as it is the median of the data.

Five Number Summary - Example

10	11	12	25	25	27	31	33
34	34	35	36	43	50	59	

Arrange the data in order n= 15

Position of the 1st Quartile = 0.25(n+1)=4 ; **First Quartile** = **25**

Position of the Median = $\frac{n+1}{2}$ or 0.50(n+1)= 8 ; **Median** = **33**

Position of the 3rd Quartile = 0.75(n+1)=12 ; **Third Quartile** =**36**

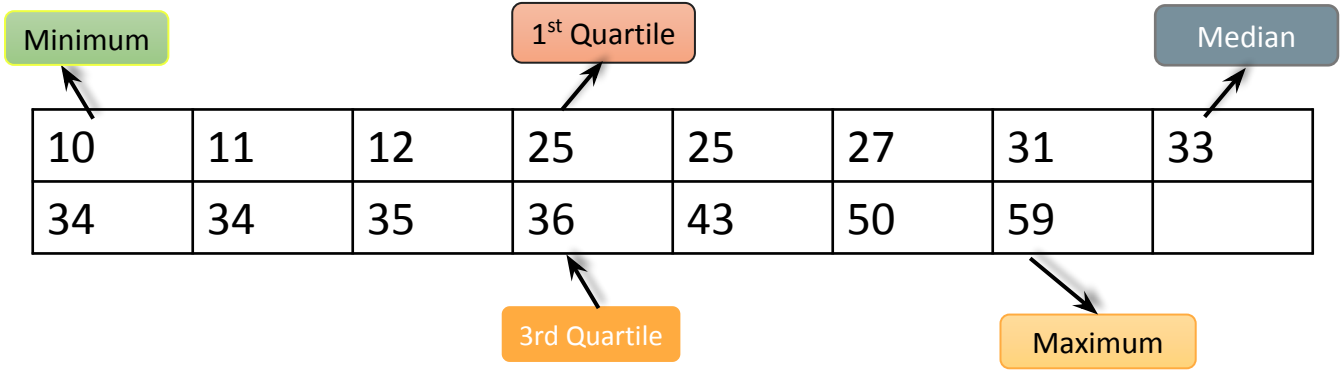
Minimum = **10** **Maximum** = **59**

Steps to Construct Boxplot

Step 1: Order the data from smallest to largest.
Step 2: Find the median.
Step 3: Find the quartiles.
Step 4: Complete the five-number summary by finding the min and the max.
Step 5: Making a boxplot.

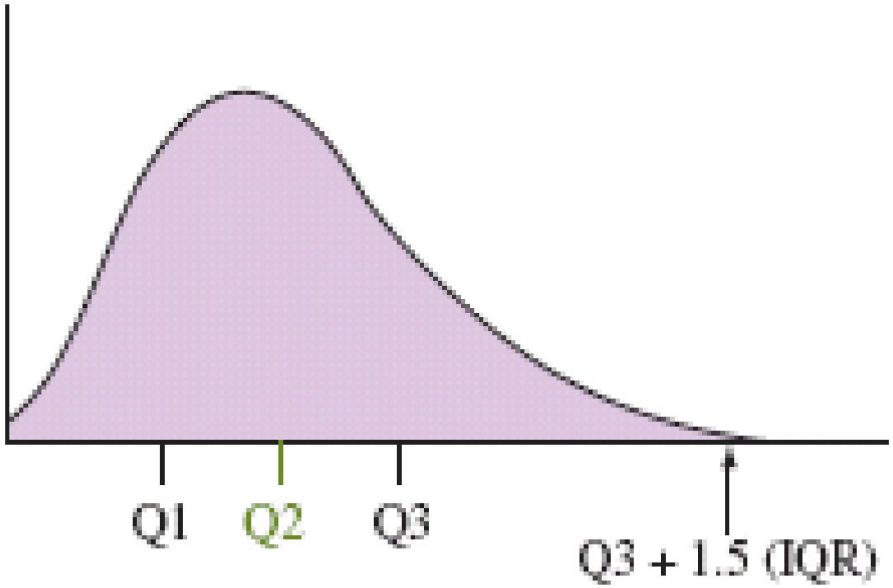
- **Step 1:** Scale and label an axis that fits the five-number summary.
- **Step 2:** Make solid dots against Q1, Q2 and Q3 values above the number line.
- **Step 3:** Draw vertical lines from number line to those 3 points and complete the box .
- **Step 4:** Draw two horizontal lines from outside box(either side) to till the minimum and maximum value respectively. These lines are called whiskers.

Five Number Summary - Example



Criteria for Identifying Outlier

An observation is a potential outlier if it falls **more than 1.5 x IQR below** the first or **more than 1.5 x IQR above** the third quartile.



MATHEMATICS FOR COMPUTER SCIENCE ENGINEERS

Boxplots- Application in Machine Learning

Machine Learning is a highly data driven field. The quality of the model will depend very much on the data fed to train it.

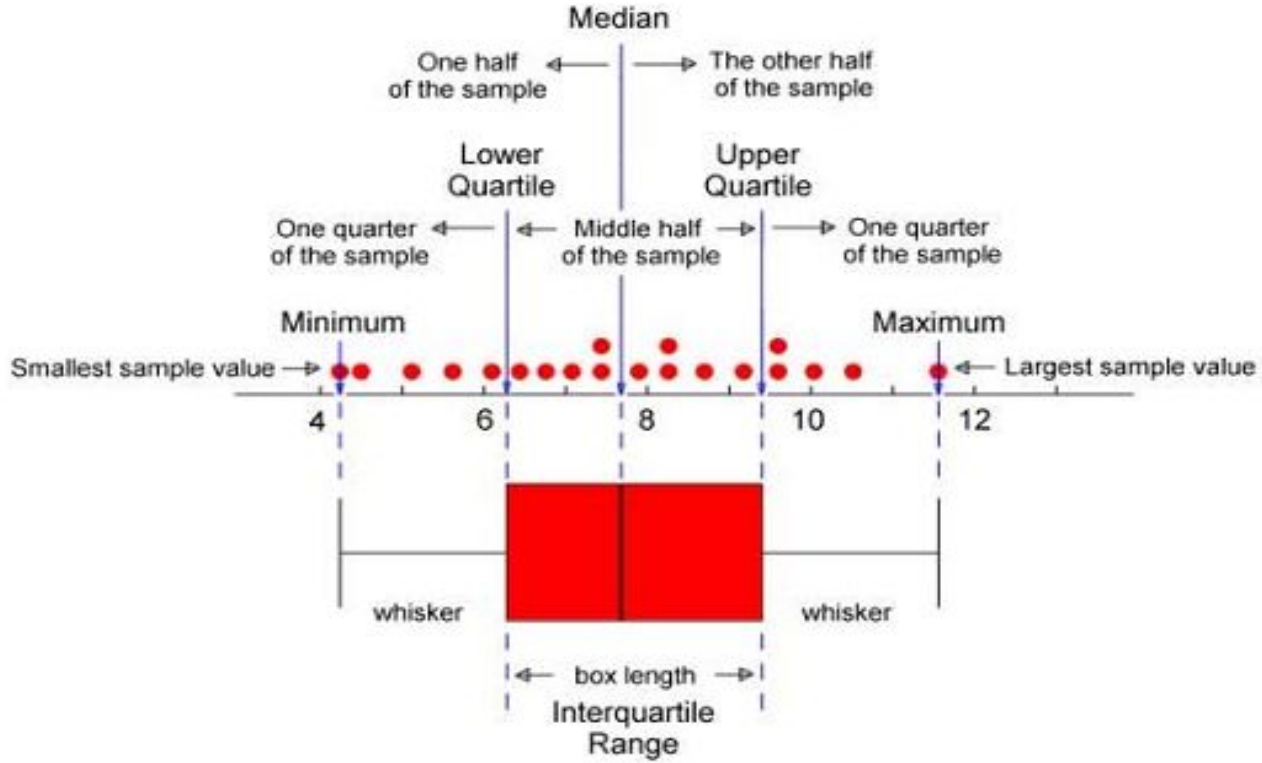
Presence of outliers can skew the training data, which may lead to overestimated or underestimated values. This can cause great harm in important predictions/models such as patient medical diagnosis, Disaster prediction and so on.

Boxplots are one of the simplest methods to find outliers present in training data. Taking care of these outliers will lead to higher classification accuracy.

Source:https://www.mdpi.com/2227-7390/8/5/662/htm

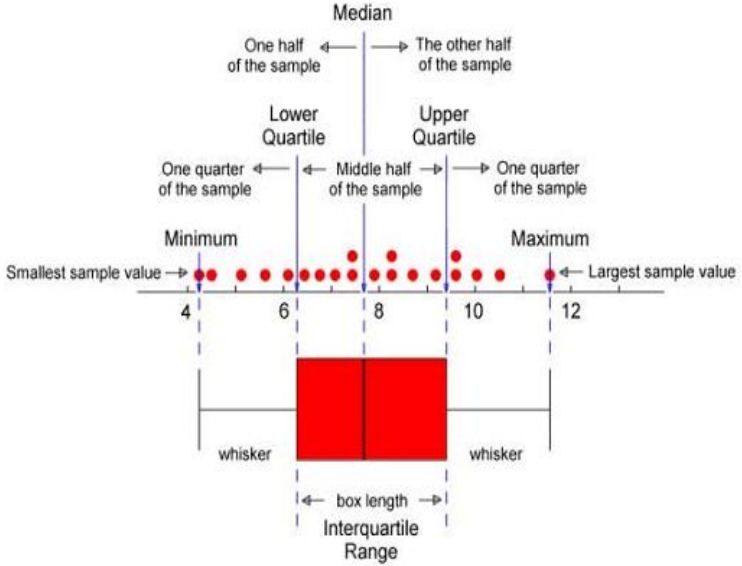
MATHEMATICS FOR COMPUTER SCIENCE ENGINEERS

Boxplots- Interpretations

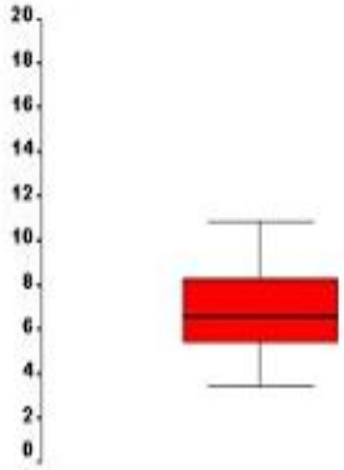


Boxplots- Interpretations

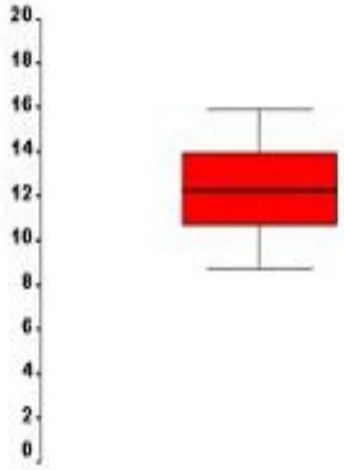
- The box length gives an indication of the sample variability and the line across the box shows where the sample is centered.
- The position of the box in its whiskers and the position of the line in the box also tells us whether the sample is symmetric or skewed, either to the right or left.



Boxplots- The Boxplot as an Indicator of Centrality

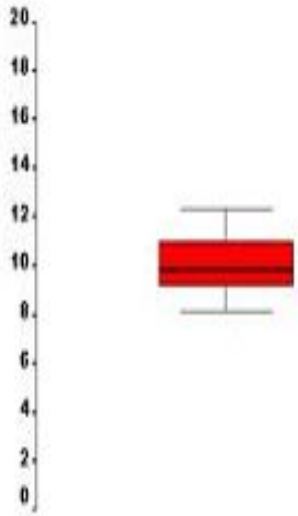


The boxplot of a sample of 20 points from a population centred on 7.

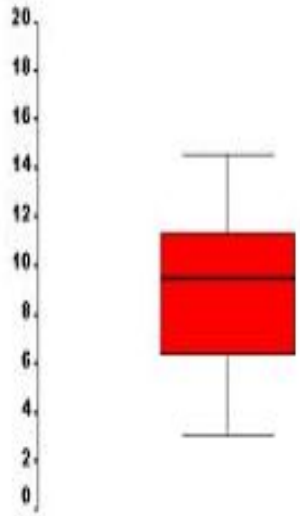


The boxplot of a sample of 20 points from a population centred on 12.

Boxplots- The Boxplot as an Indicator of Spread



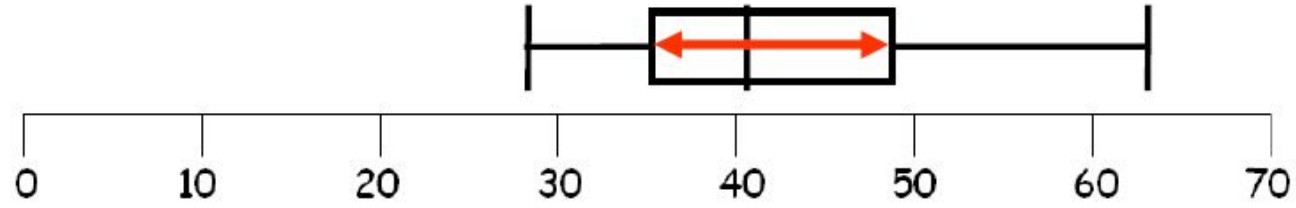
The boxplot of a sample of 20 points from a population centred on 10 with standard deviation 1.



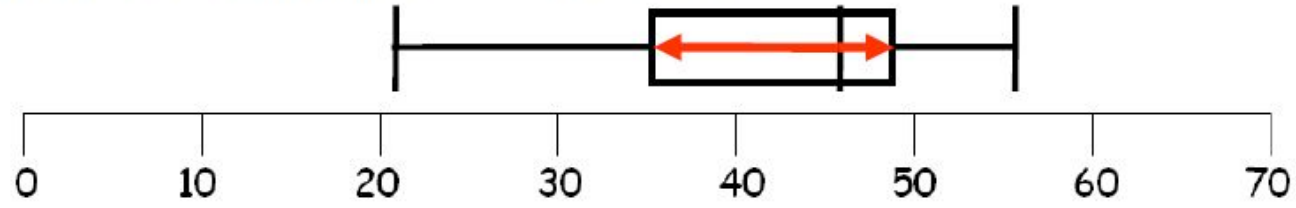
The boxplot of a sample of 20 points from a population centred on 10 with standard deviation 3.

Boxplots- The Boxplot as an Indicator of Symmetry

Positive skew: median closer to LQ than UQ

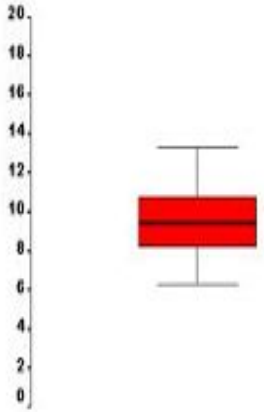


Negative skew: median closer to UQ than LQ

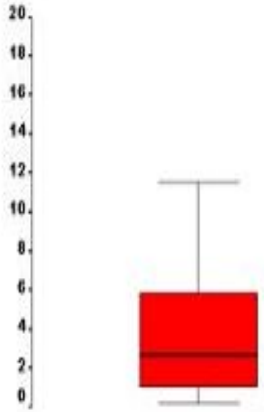


Symmetrical distribution

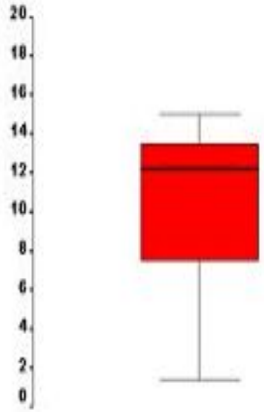
Boxplots- The Boxplot as an Indicator of Symmetry -Cont.



The boxplot of a sample of 20 points from a symmetric population. The line is close to the centre of the box and the whisker lengths are the same.

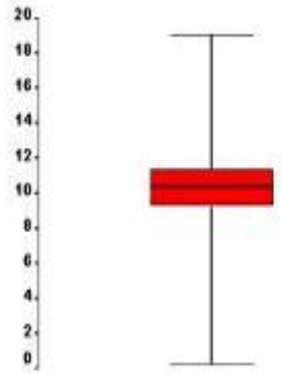


The boxplot of a sample of 20 points from a population which is skewed to the right. The top whisker is much longer than the bottom whisker and the line is gravitating towards the bottom of the box.

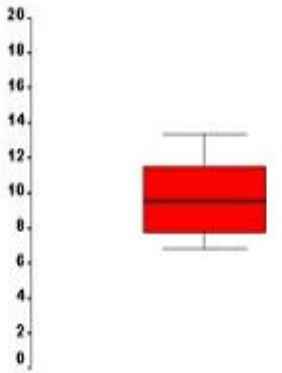


The boxplot of a sample of 20 points from a population which is skewed to the left. The bottom whisker is much longer than the top whisker and the line is rising to the top of the box.

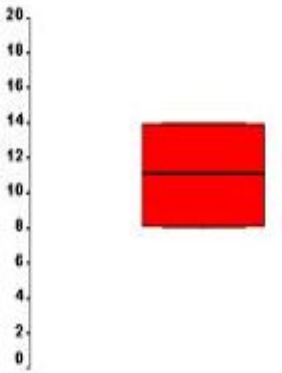
Boxplots- The Boxplot as an Indicator of Tail Length



The boxplot of a sample of 20 points from a population with long tails. The length of the whiskers far exceeds the length of the box. (A well proportioned tail would give rise to whiskers about the same length as the box, or maybe slightly longer.)

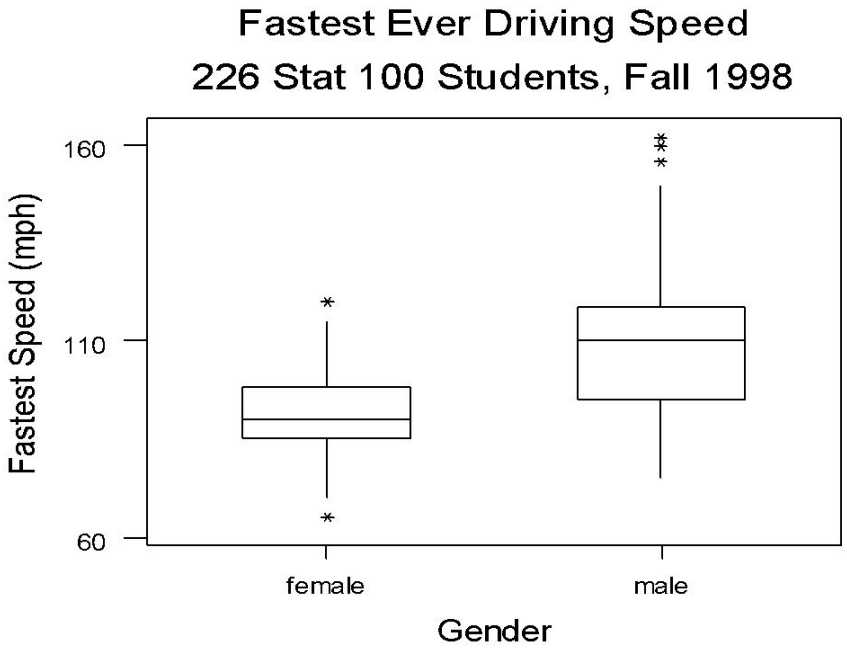


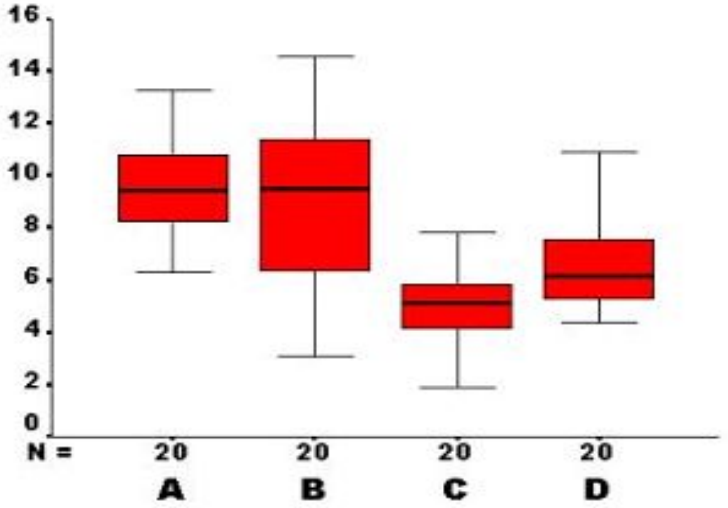
The boxplot of a sample of 20 points from a population with short tails. The length of the whiskers is shorter than the length of the box.



The boxplot of a sample of 20 points from a population with extremely short tails (actually a U-shaped population, with a dip in the middle rather than a hump). The whiskers are absent.

Boxplots- The Boxplot for Comparison





Interpreting Box plots in general

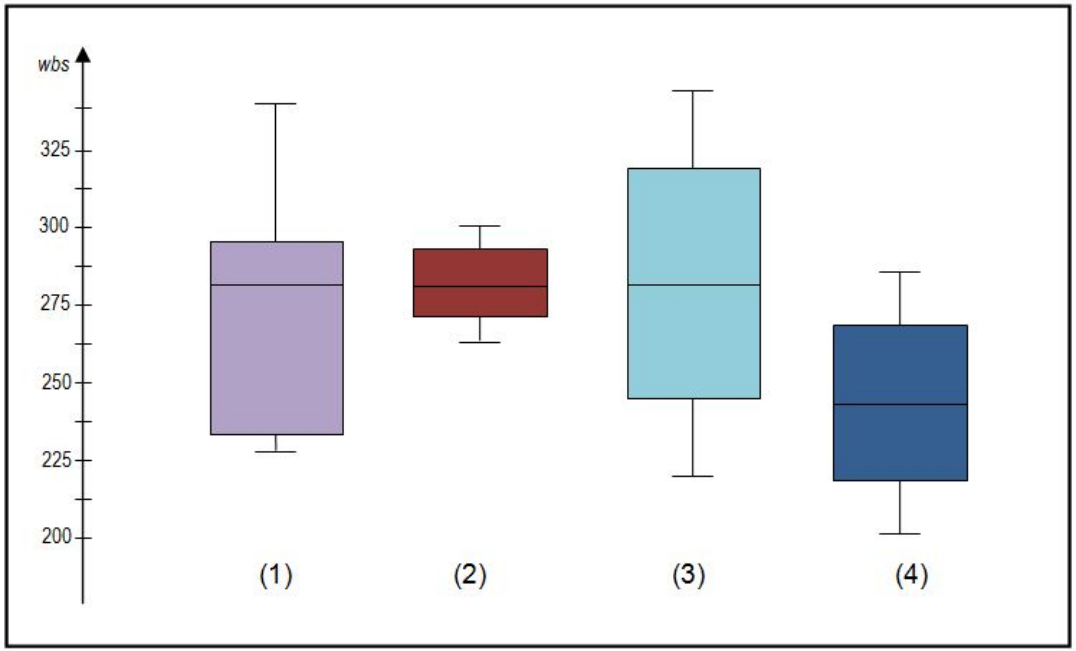
Box plots are used to show overall patterns of response for a group.
They provide a useful way to visualize the range and other characteristics of responses for a large group.

Example:
Box plots are drawn for groups of a school scale scores. They enable us to study the distributional characteristics of a group of scores as well as the level of the scores.

Interpreting Box plots in general

Example:

Box plots are drawn for groups of a school scale scores. They enable us to study the distributional characteristics of a group of scores as well as the level of the scores.



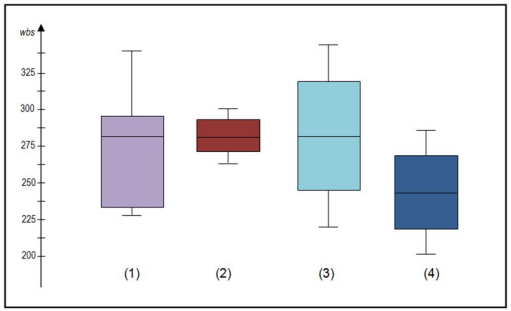
Interpreting Box plots in general

- **Some general observations about box plots**
- The box plot is **comparatively short** – see example (2). This suggests that overall students have a high level of agreement with each other.

- The box plot is **comparatively tall** – see examples (1) and (3). This suggests students hold quite different opinions about this aspect or sub-aspect.

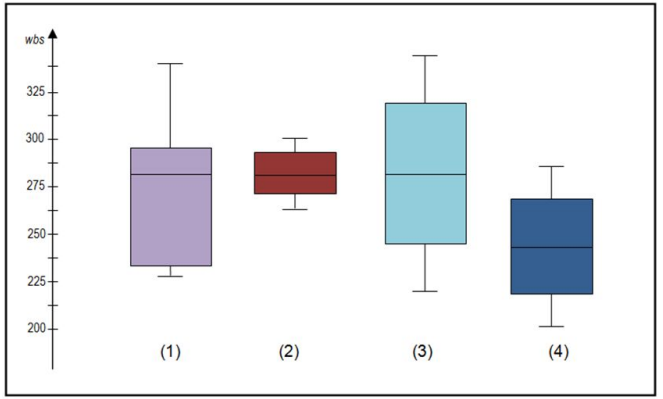
- **One box plot is much higher or lower than another** – compare (3) and (4) – This could suggest a difference between groups. For example, the box plot for boys may be lower or higher than the equivalent plot for girls.

- **Obvious differences between box plots** – see examples (1) and (2), (1) and (3), or (2) and (4). Any obvious difference between box plots for comparative groups is worthy of further investigation



Interpreting Box plots in general

The 4 sections of the box plot are uneven in size – See example (1). This shows that many students have similar views at certain parts of the scale, but in other parts of the scale students are more variable in their views. The long upper whisker in the example means that students views are varied amongst the most positive quartile group, and very similar for the least positive quartile group.

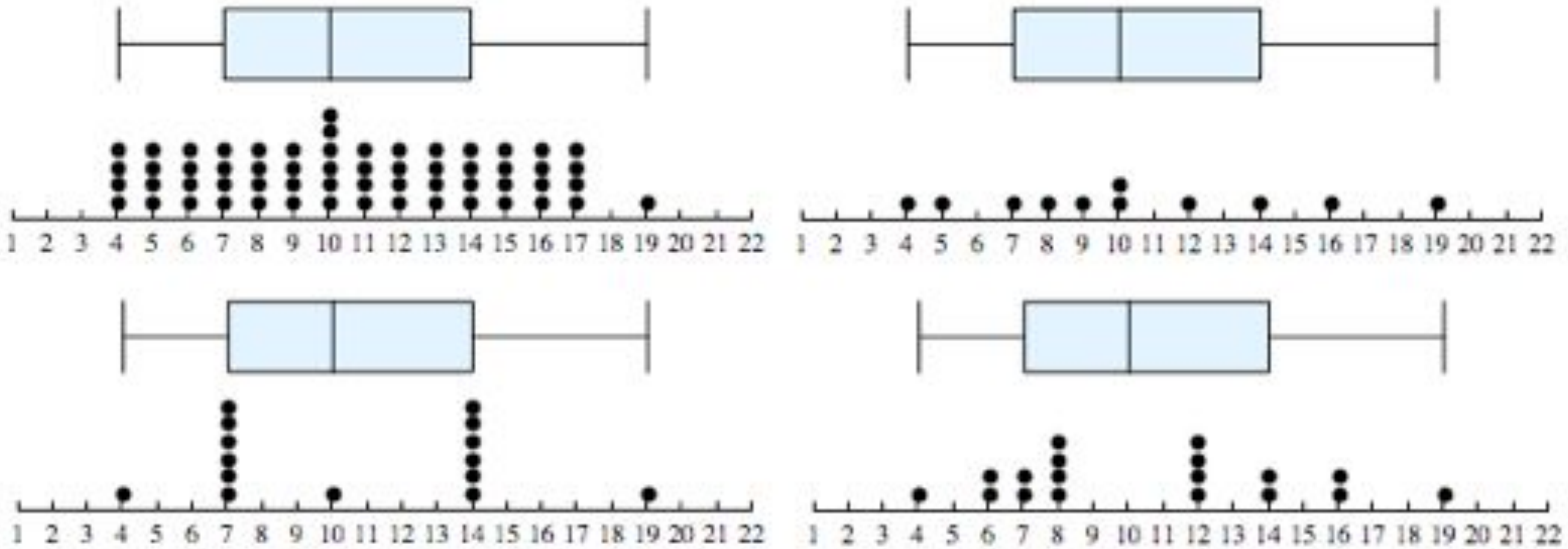


Same median, different distribution – See examples (1), (2), and (3). The medians (which generally will be close to the average) are all at the same level. However the box plots in these examples show very different distributions of views.

It always important to consider the pattern of the whole distribution of responses in a box plot.

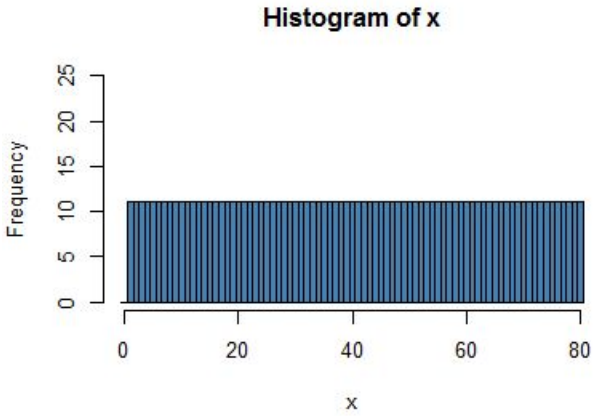
Box Plots and Histograms

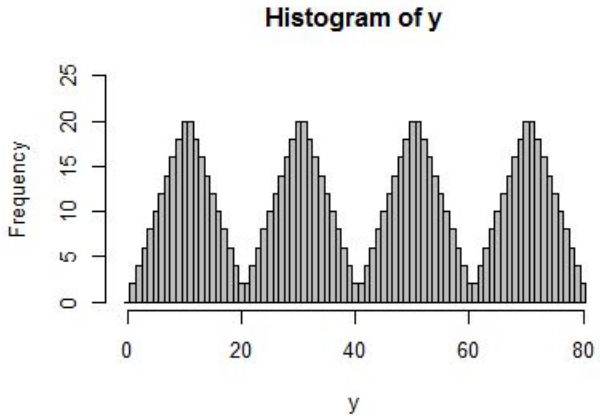
Take a look at these images. What observation can you make?

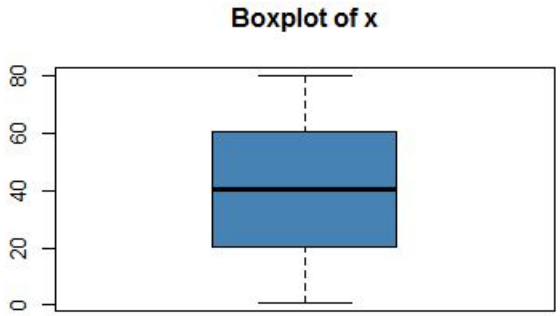


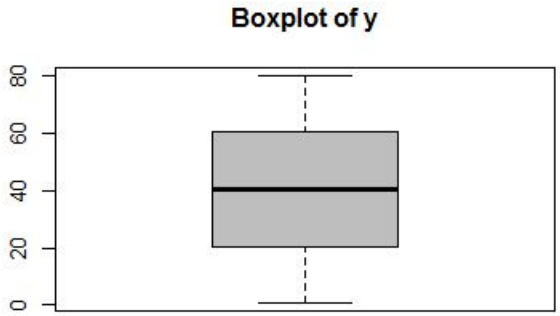
MATHEMATICS FOR COMPUTER SCIENCE ENGINEERS

Box Plots and Histograms





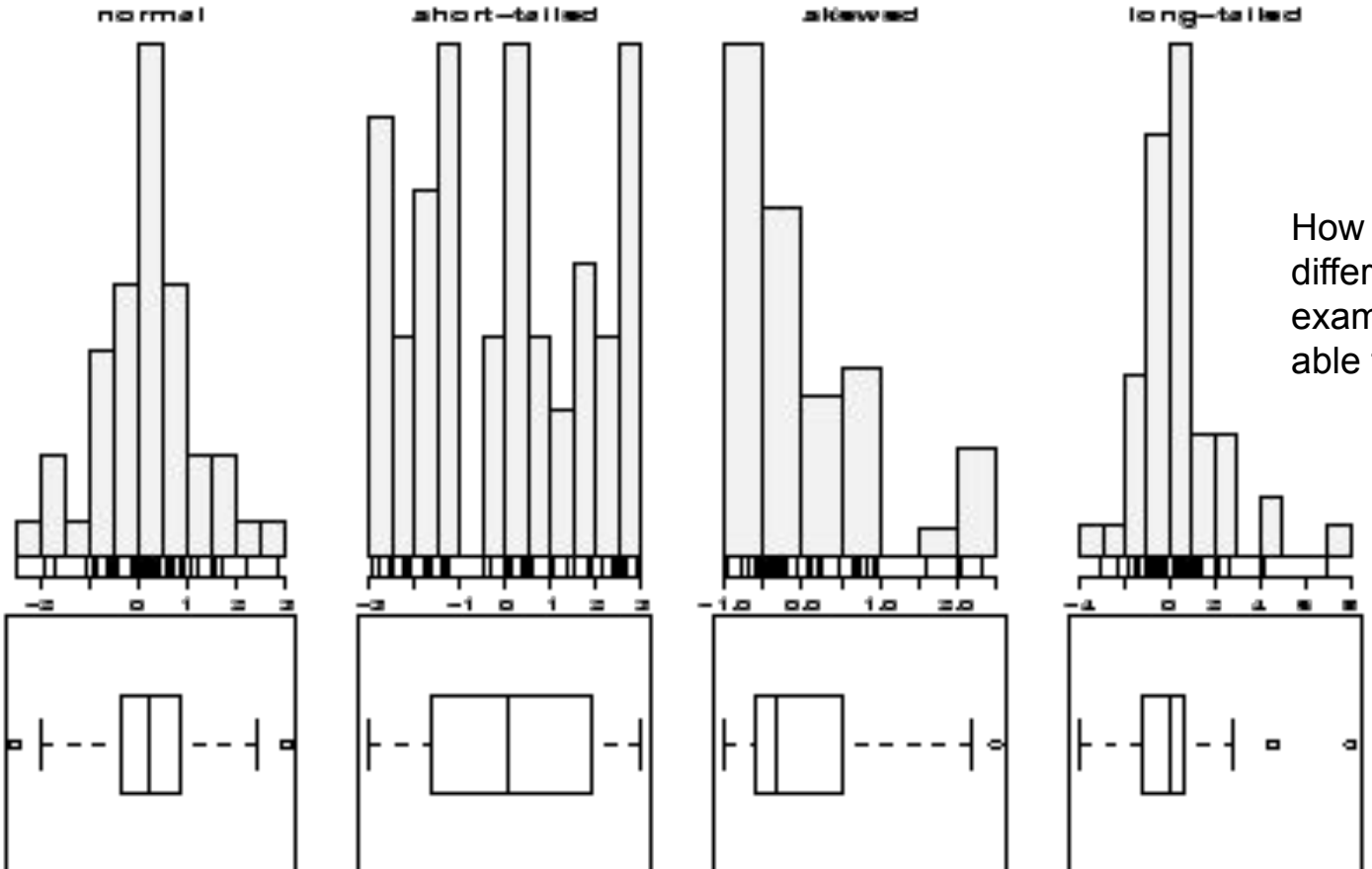




What is the observation in this scenario? How does the shape of the histogram affect the boxplot?

MATHEMATICS FOR COMPUTER SCIENCE ENGINEERS

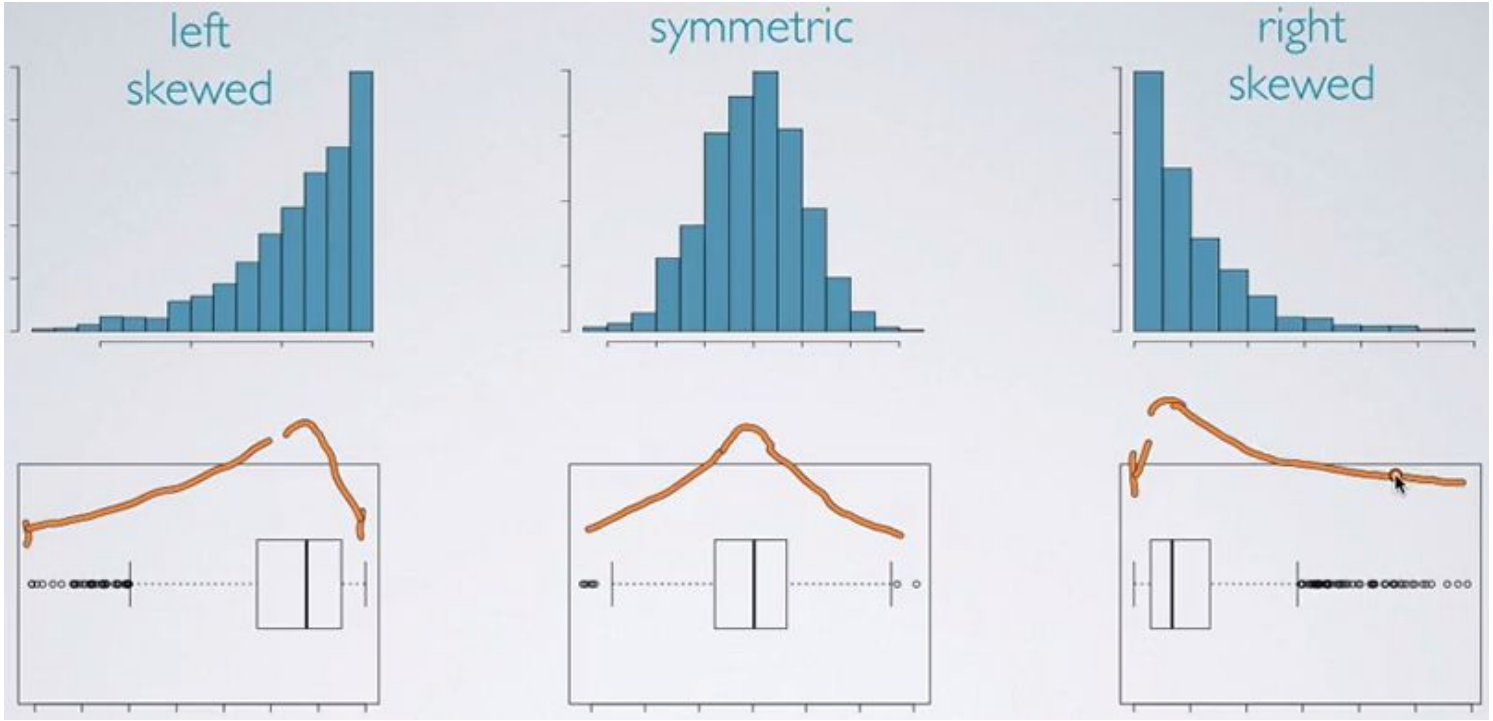
Box Plots and Histograms



How are these images different from previous examples? What are you able to conclude?

MATHEMATICS FOR COMPUTER SCIENCE ENGINEERS

Box Plots and Histograms

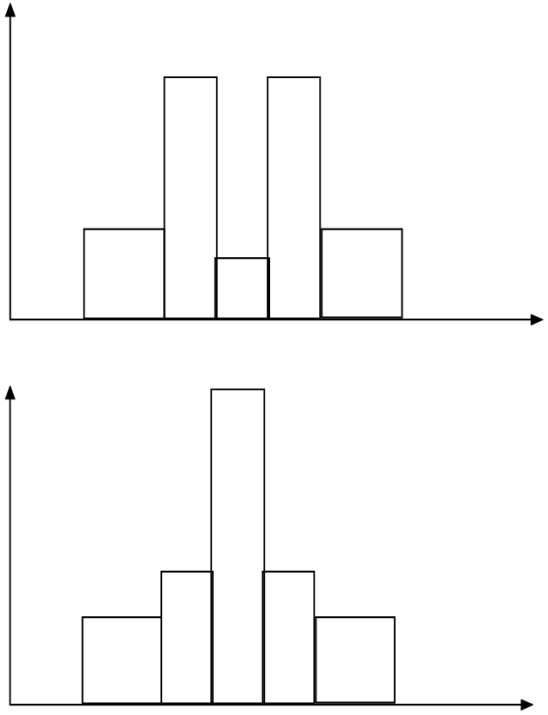


Box Plots and Histograms-Which graph to use when?

- Dot plots are good for small data sets, while histograms and box plots are good for large data sets.
- Boxplots and dotplots are good for comparing two groups.
- Boxplots are good for identifying outliers.
- Histograms and boxplots are good for identifying "shape" of data.

Box Plots and Histograms- Histograms Often Tell More than Boxplots

- The two histograms shown in the left may have the same boxplot representation
- The same values for: min, Q1, median, Q3, max
- But they have rather different data distributions



Reference

Text Book:

Statistics for Engineers and Scientists, William Navidi.

THANK YOU

Dr. Mamatha H R

Department of Computer Science and Engineering
mamathahr@pes.edu

MATHEMATICS FOR COMPUTER SCIENCE ENGINEERS

Data Visualization and Interpretation

Mamatha H R

Department of Computer Science and Engineering
mamathahr@pes.edu

PES UNIVERSITY
CELEBRATING 50 YEARS

MATHEMATICS FOR COMPUTER SCIENCE ENGINEERS

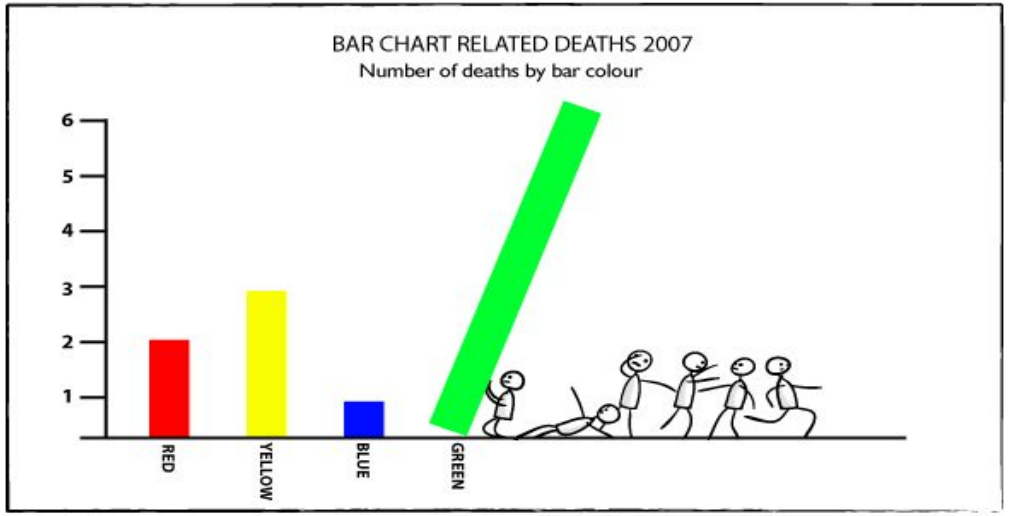
Data Visualization and Interpretation - Bar Charts

Mamatha H R
Department of Computer Science and Engineering

Data Visualization: Bar Charts

- Often known as the “King of Charts”, Bar Charts are one of the most commonly used charts in the field of Data Science.

- The advantage of bar plots (or “bar charts”, “column charts”) over other chart types is that the human eye has evolved a refined ability to compare the **length of objects**, as opposed to **angle or area.**

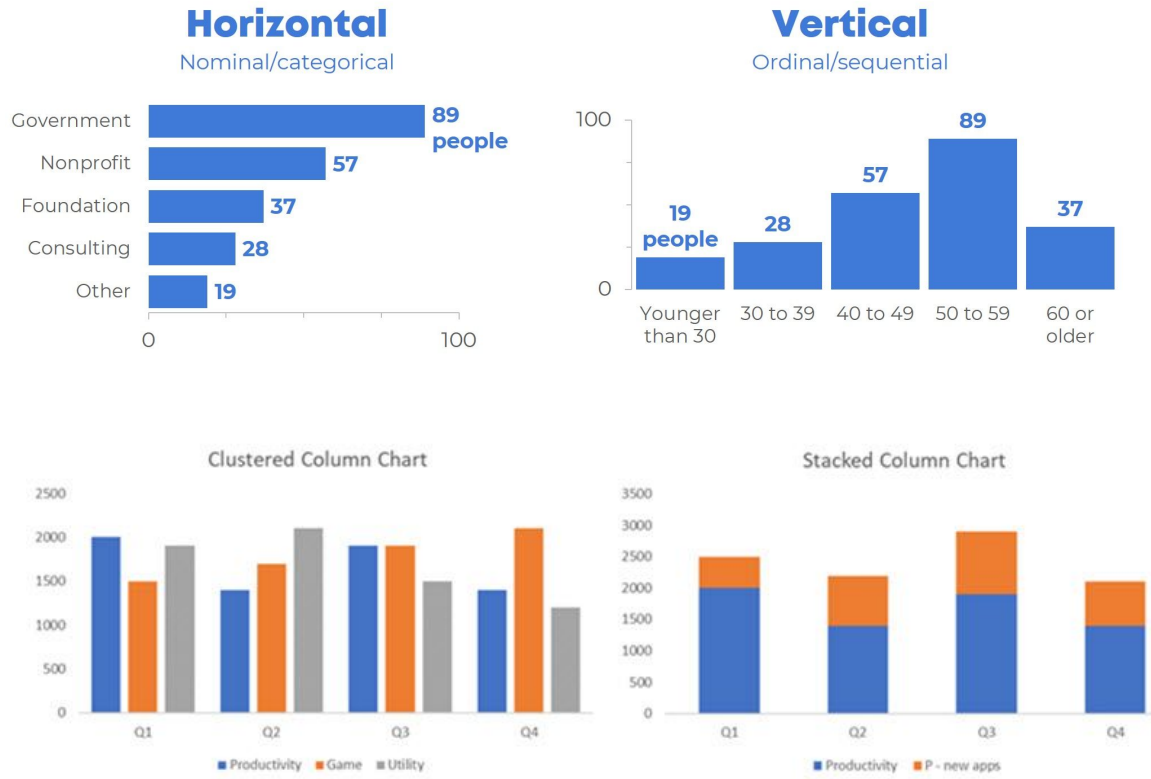


Data Visualization: Bar Charts

- Summarizes categorical data.
- Horizontal axis represents categories, while vertical axis represents either counts (“frequencies”) or percentages (“relative frequencies”).
- Used to illustrate the differences in percentages (or counts) between categories.
- The graph represents categories on one axis and a discrete value in the other. The goal is to show the relationship between the two axes
- Bar charts can also show big changes in data over time

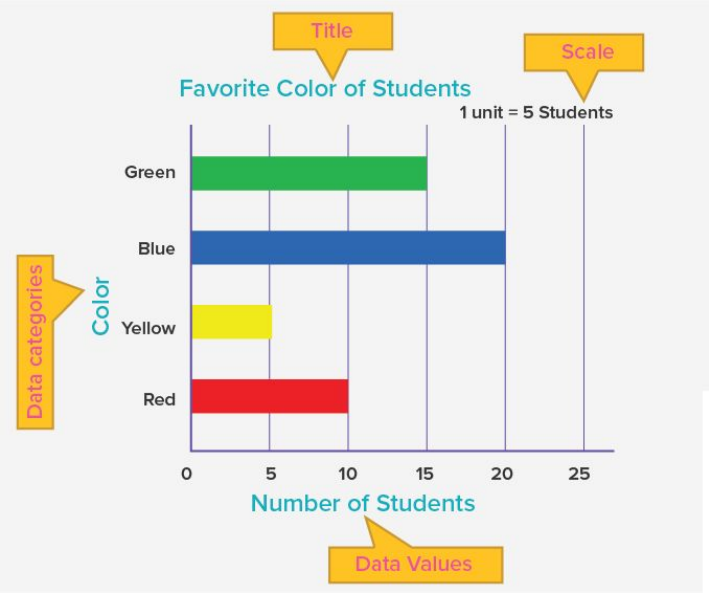
MATHEMATICS FOR COMPUTER SCIENCE ENGINEERS

Data Visualization: Types of Bar Graphs



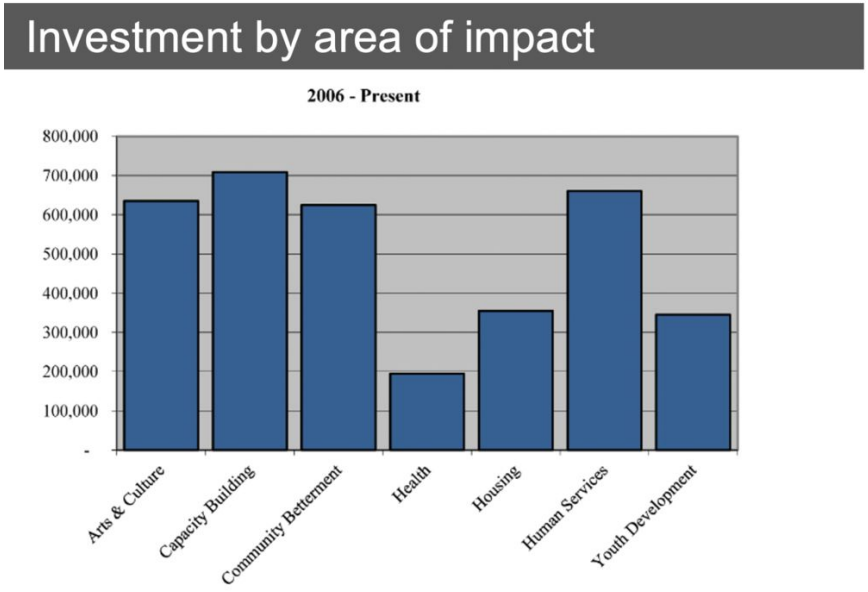
Data Visualization: Types of Bar Graphs - Horizontal

- **Horizontal Bar Graphs:** The classes are displayed on the y-axis, and the values(scores) of those classes are displayed on the x-axis. Useful only when comparing one set of data.



Data Visualization: Types of Bar Graphs - Vertical

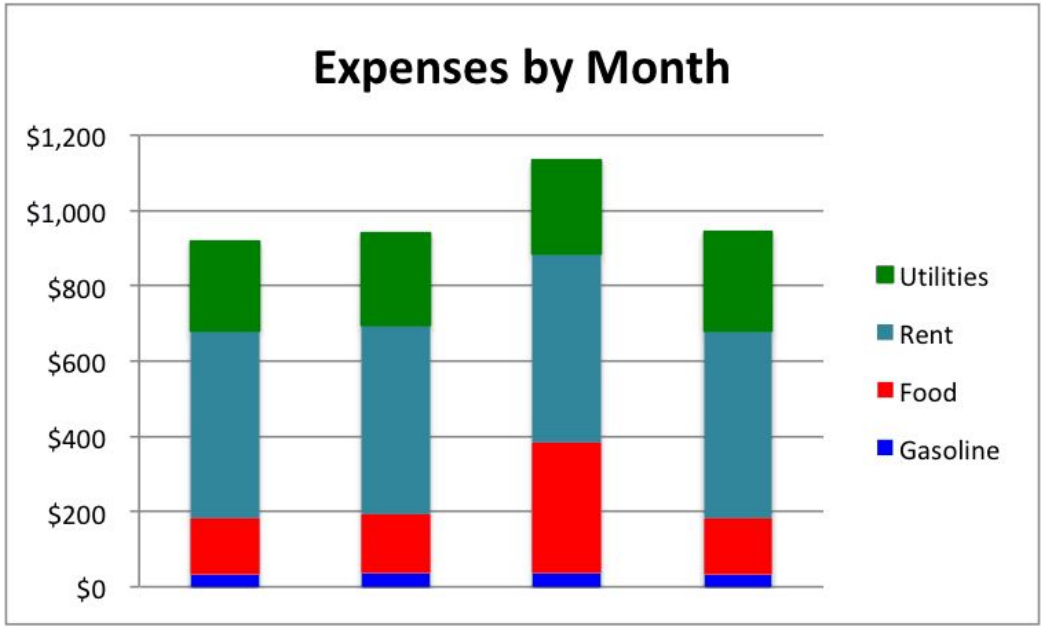
- **Vertical Bar Graphs :** The **classes** are displayed on the **x-axis**, and the **values(scores)** of those classes are displayed on the **y-axis**. Useful only when comparing one set of data.



Data Visualization: Types of Bar Graphs - Stacked

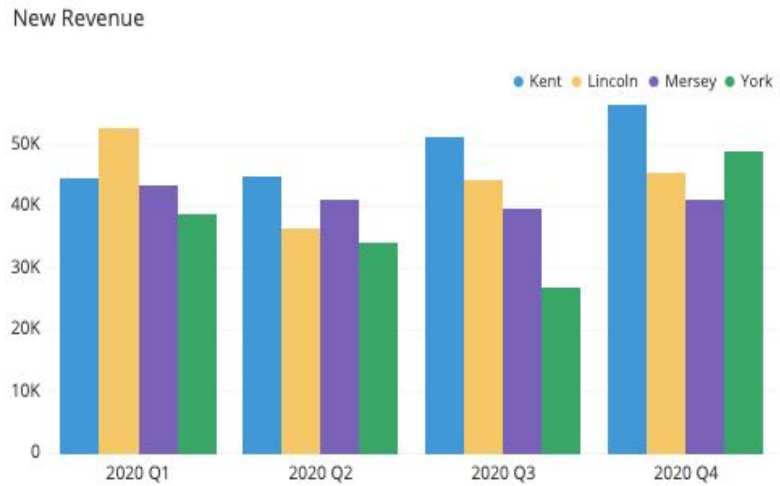
- **Stacked Bar Graphs :** Each bar has multiple datasets to be compared, each set of values belonging to the class of different datasets are stacked over one other.

- Useful when comparing multiple datasets but having same set of classes



Data Visualization: Types of Bar Graphs - Grouped

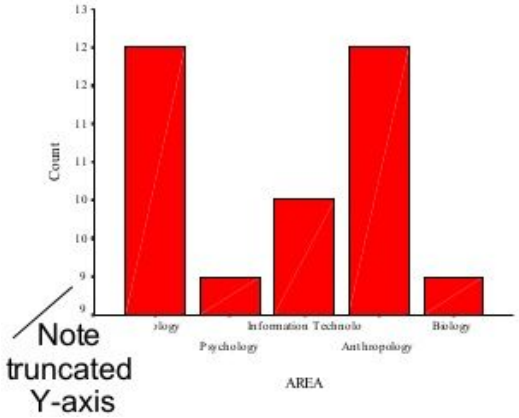
- **Grouped Bar Graphs :** Grouped bar charts are Bar charts in which multiple sets of data items are compared, with a single color used to denote a specific series across all sets.
- A grouped or clustered bar graph is used to represent discrete values for more than one item that share the same category.
- Grouped bar charts are a way of showing information about different sub-groups of the main categories.
- But care needs to be taken to ensure that the chart does not contain too much information making it complicated to read and interpret.

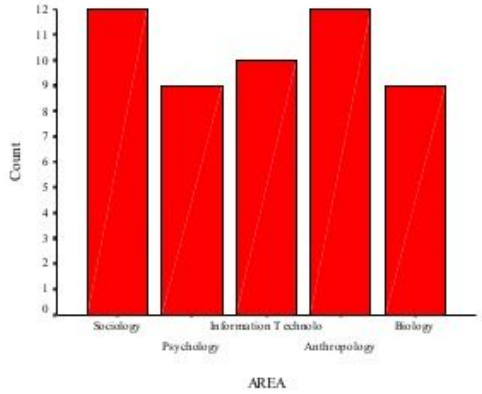


Data Visualization: Bar Charts

Bar chart (Bar graph)

- Allows comparison of heights of bars
- X-axis: Collapse if too many categories
- Y-axis: Count/Frequency or % - truncation exaggerates differences
- Can add data labels (data values for each bar)



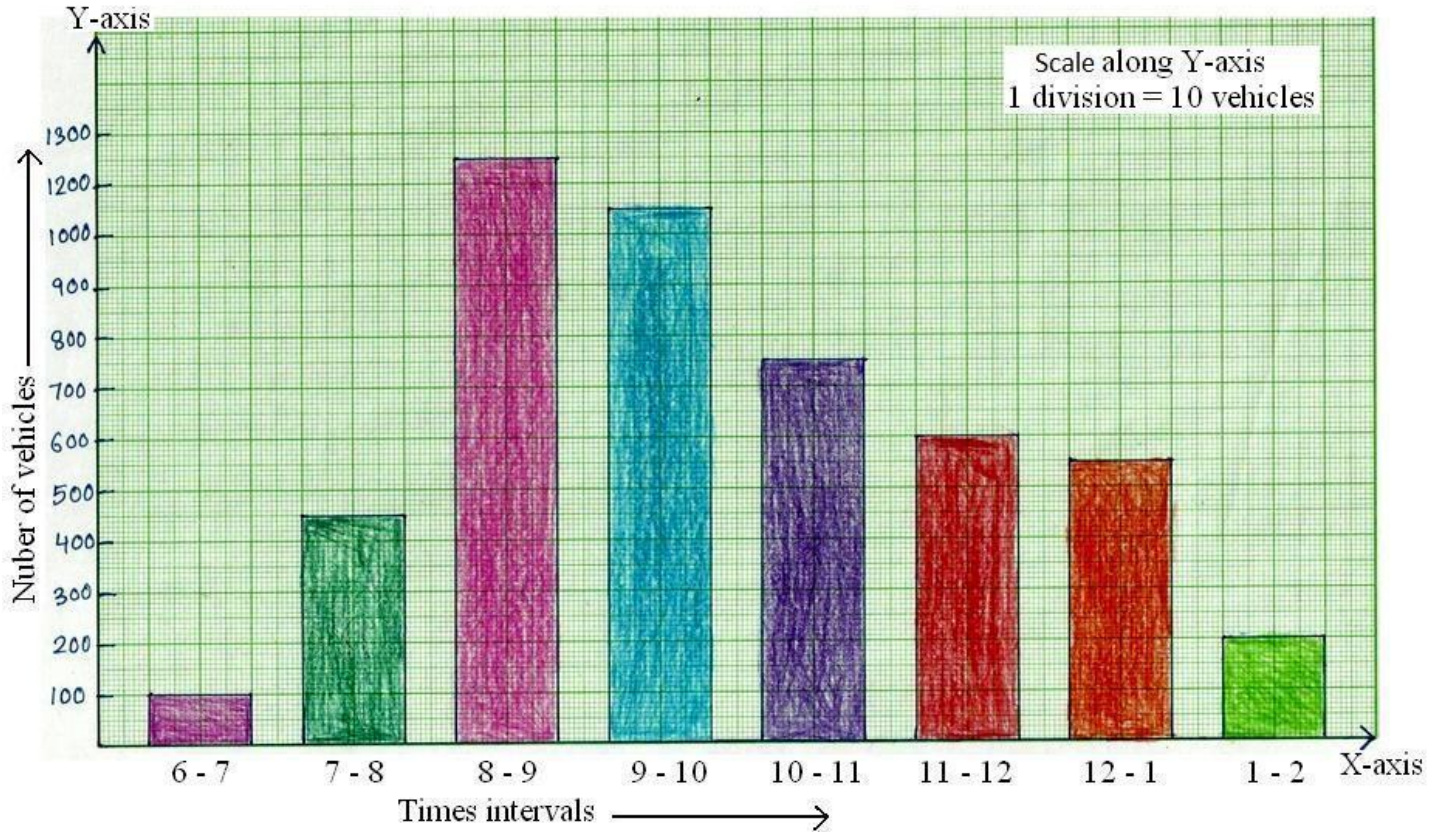


Data Visualization: Bar Charts Examples

The vehicular traffic at a busy road crossing in a particular place was recorded on a particular day from 6am to 2 pm and the data was rounded off to the nearest tens. Construct a Bar Chart.

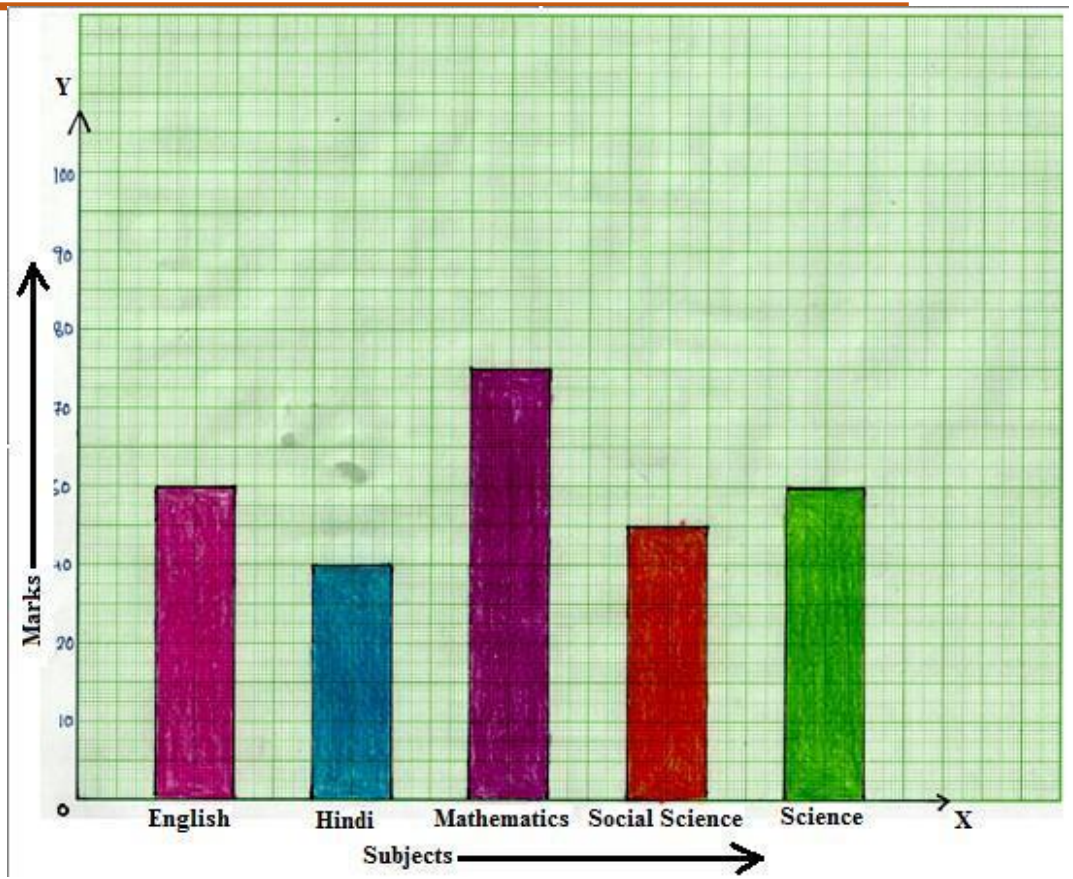
Time in Hours	6 - 7	7 - 8	8 - 9	9 - 10	10 - 11	11 - 12	12 - 1	1 - 2
Number of Vehicles	100	450	1250	1050	750	600	550	200

Data Visualization: Bar Charts Examples



Data Visualization: Bar Charts Examples

Look at the graph given



Data Visualization: Bar Charts Examples

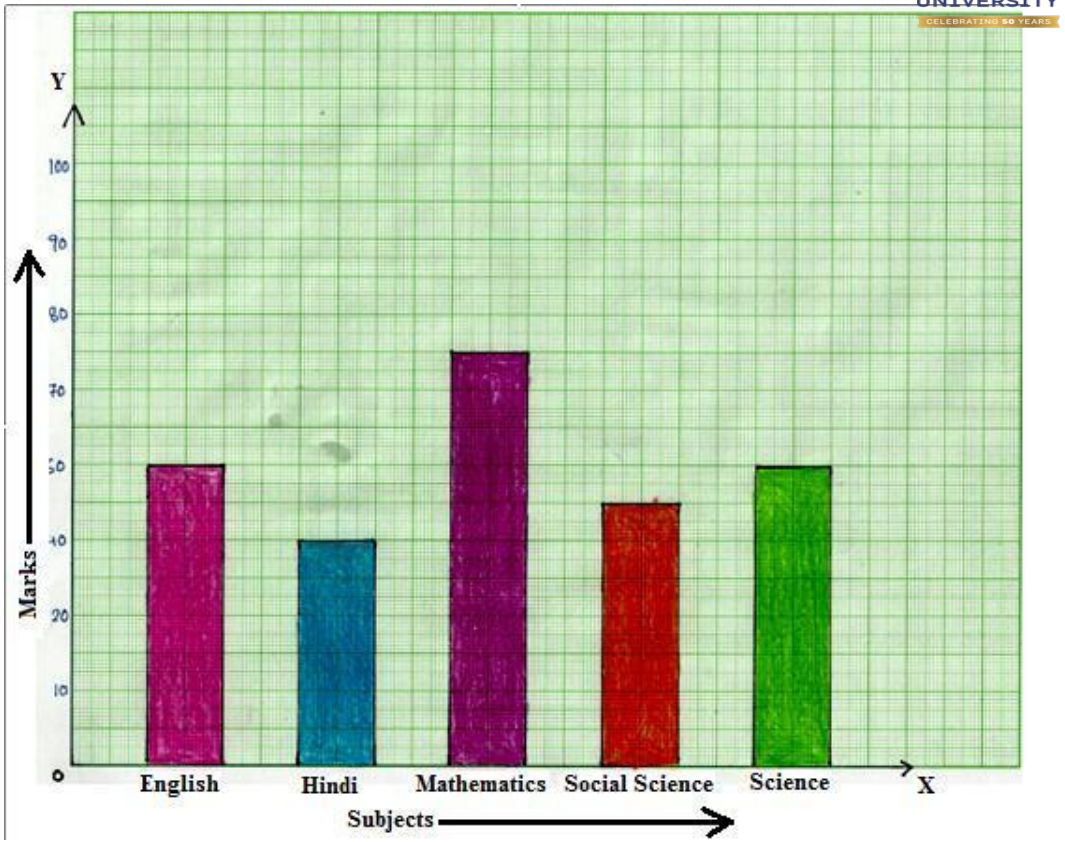
Read it carefully and answer the following questions.

(i) What information does the bar graph give?

(ii) In which subject is the student very good

(iii) In which subject is he poor?

(iv) What are the average of his marks?



Data Visualization: Bar Charts Examples

(i) It shows the marks obtained by a student in five subjects

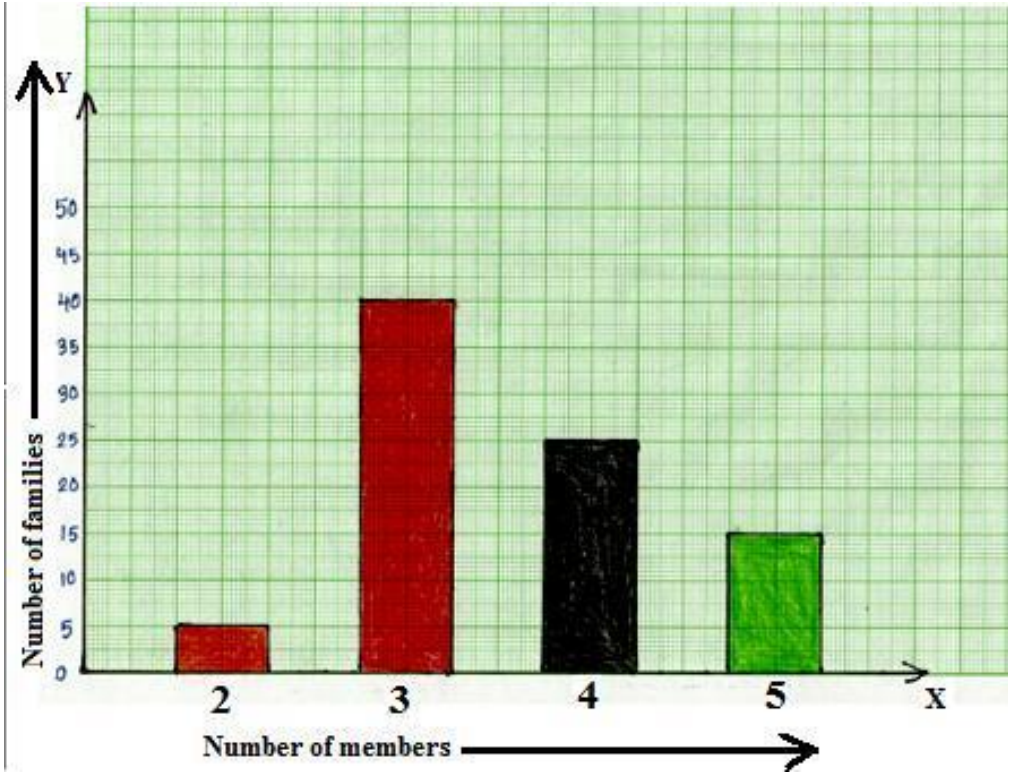
(ii) Mathematics

(iii) Hindi

(iv) 56

Data Visualization: Bar Charts Examples

In a survey of 85 families of a colony, the number of members in each family was recorded, and the data has been represented by the following bar graph.



Data Visualization: Bar Charts Examples

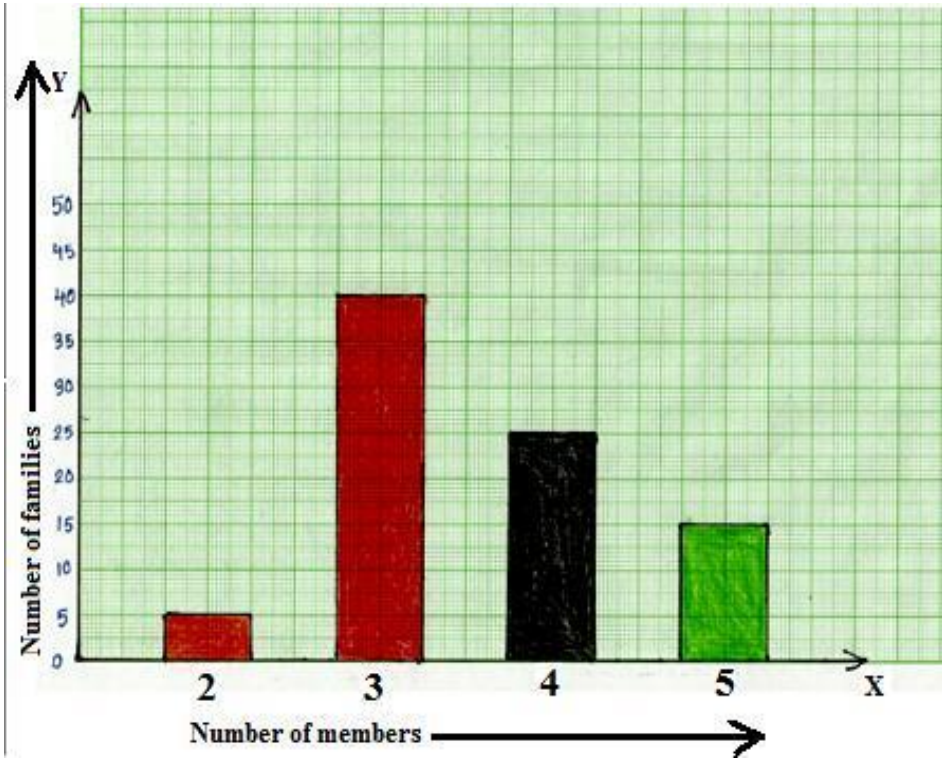
Read the bar graph carefully and answer the following questions:

(i) What information does the bar graph give?

(ii) How many families have 3 members?

(iii) How many people live alone?

(iv) Which type of family is the most common? How many members are there in each family of this kind?



Data Visualization: Bar Charts Examples

(i)It gives the number of families containing 2, 3, 4, 5 members each.

(ii) 40

(iii) none

(iv) Family having 3 members, 3 members.

MATHEMATICS FOR COMPUTER SCIENCE ENGINEERS

Data Visualization:Bar Chart

Determine the discrete range

- Examine your data to find the bar with the largest value. This will help you determine the range of the vertical axis and the size of each increment.

Determine the number of bars

Examine your data to find how many bars your chart will contain. Use this number to draw and label the horizontal axis

Source: www.slideshare.net

Slide Courtesy: Dr.Uma

MATHEMATICS FOR COMPUTER SCIENCE ENGINEERS

Data Visualization:Bar Chart

Determine the order of the bars

Bars may be arranged in any order. (A bar chart arranged from highest to lowest incidence is called a Pareto chart)

Draw the bars

If you are preparing a grouped bar graph, remember to present the information in the same order in each grouping

Source: www.slideshare.net

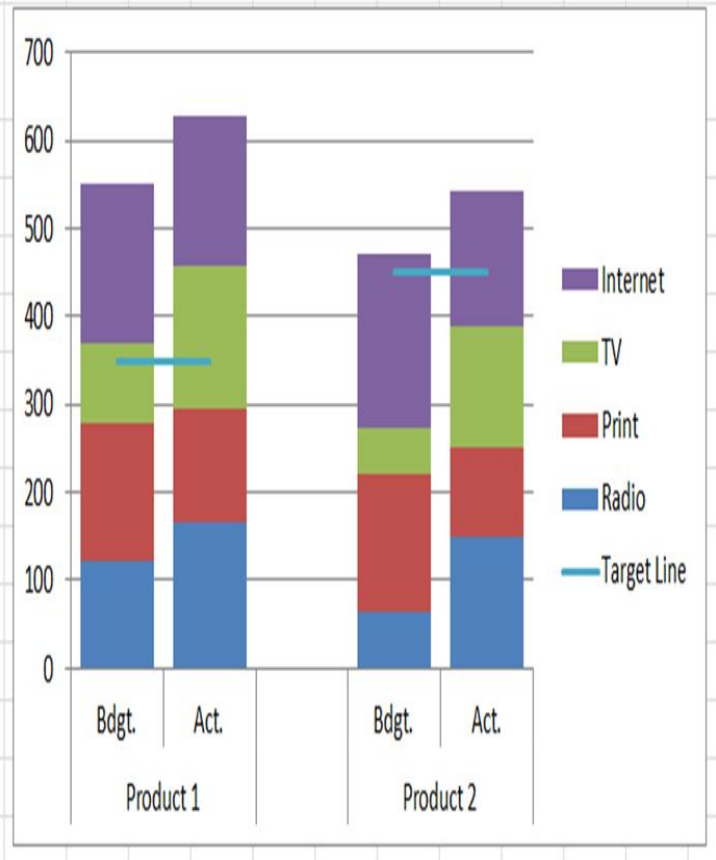
Slide Courtesy: Dr.Uma

MATHEMATICS FOR COMPUTER SCIENCE ENGINEERS

Data Visualization:Bar Chart

Bar chart statistics

- **Mean** — the average height of all the bars.
- **Maximum** — the maximum value (tallest bar) in the series.
- **Minimum** — the minimum value (shortest bar) in the series.
- **Sample Size** — the number of values (bars) in the series.
- **Range** — the maximum value minus the minimum value.
- **Standard Deviation** — Indicates how widely data is spread around the mean.



Source: www.slideshare.net

Slide Courtesy: Dr.Uma

Difference between Bar and Histogram

Bar

Type of Data

In bar graphs are usually used to display "**categorical data**", that is data that fits into categories.

Histogram

Type of Data

Used to present "**continuous data**", that is data that represents measured quantity where, at least in theory, the numbers can take on any value in a certain range.

Slide Courtesy: Dr.Uma

THANK YOU

Mamatha H R

Department of Computer Science

mamathahr@pes.edu

MATHEMATICS FOR COMPUTER SCIENCE ENGINEERS

Data Visualization and Interpretation

Mamatha H R

Department of Computer Science and Engineering

mamathahr@pes.edu

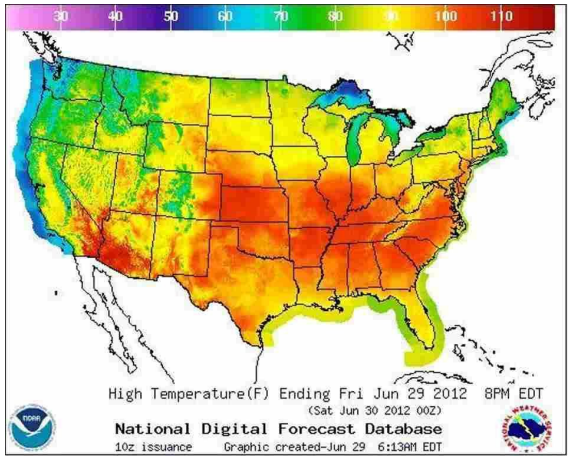
MATHEMATICS FOR COMPUTER SCIENCE ENGINEERS

Data Visualization and Interpretation – Heat Map

Mamatha H R
Department of Computer Science and Engineering

Data Visualization: Heat Map

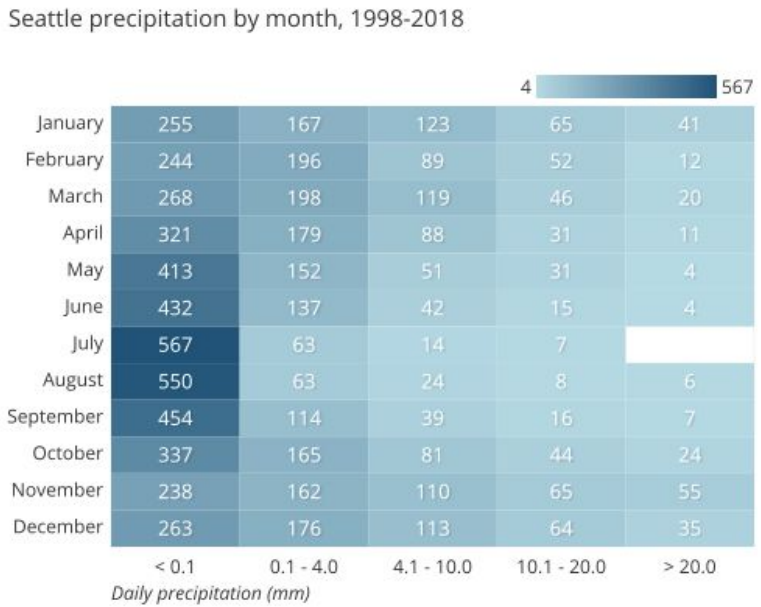
A **heat map** (or **heatmap**) is a data visualization technique that shows magnitude of a phenomenon as color in two dimensions. The variation in color may be by hue or intensity, giving obvious visual cues to the reader about how the phenomenon is clustered or varies over space.



Slide Courtesy:Ms.Yousha Mahamuni

Data Visualization: Heat Map

A heatmap (aka heat map) depicts values for a main variable of interest across two axis variables as a grid of colored squares. The axis variables are divided into ranges like a bar chart or histogram, and each cell's color indicates the value of the main variable in the corresponding cell range.



https://chartio.com/learn/charts/heatmap-complete-guide/

Data Visualization: Heat Map

The term heatmap is also used in a more general sense, where data is not constrained to a grid. For example, tracking tools for websites can be set up to see how users interact with the site, like studying where a user clicks, or how far down a page readers tend to scroll.

Data Visualization: Heat Map

Every click (or other tracking event) is associated with a position, which radiates a small amount of numeric value around its location.

These values are totaled together across all events and then plotted with an associated colormap.

The visual language of these tools' output, associating value with color, is similar to the type of heatmap defined at the top, just without a grid-based structure.

Heatmaps of this type are sometimes also known as 2-d density plots.

Data Visualization: When to use Heat Map?

Heatmaps are used to show relationships between two variables, one plotted on each axis.

By observing how cell colors change across each axis, you can observe if there are any patterns in value for one or both variables.

The variables plotted on each axis can be of any type, whether they take on categorical labels or numeric values.

In the latter case, the numeric value must be binned like in a histogram in order to form the grid cells where colors associated with the main variable of interest will be plotted.

Data Visualization: When to use Heat Map?

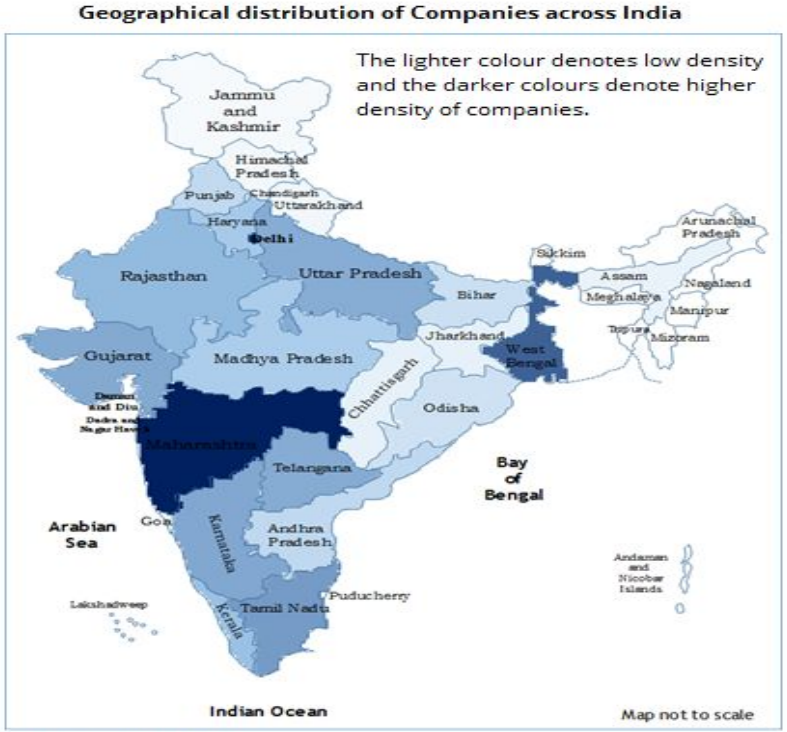
Cell colorings can correspond to all manner of metrics, like a frequency count of points in each bin, or summary statistics like mean or median for a third variable.

One way of thinking of the construction of a heatmap is as a table or matrix, with color encoding on top of the cells.

In certain applications, it is also possible for cells to be colored based on non-numeric values (e.g. general qualitative levels of low, medium, high).

Why are Heat Maps Used?

- Visually appealing
- Provides a more generalized interpretation of numerical values
- Useful when dealing with large data, colour coded data is easier to interpret
- Generally self explanatory
- Draw attention to trends



Source : www.tofler.in

Slide Courtesy:Ms.Yousha Mahamuni

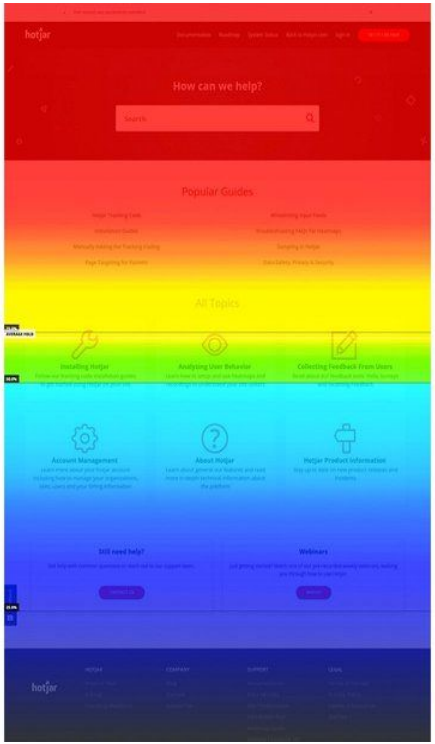
Where are Heat Maps used in Computer Science/Business ?

- Heat Maps are widely used in Search Engine Optimization (SEO)
- **Search engine optimization** is the process of improving the quality and quantity of website traffic to a website or a web page from search engines.
- SEO Heatmaps help track cursor movements, and click events of users across a given webpage.
- This data helps business/website owner to find optimal layout for advertisements, learn more about user intent, and marketing techniques.

Source:https://www.searchenginejournal.com/use-heatmaps-seo-strategy/366268/#close Slide Courtesy:Ms.Yousha Mahamuni

MATHEMATICS FOR COMPUTER SCIENCE ENGINEERS

Where are Heat Maps used in Computer Science/Business ?



The given figure is known as a **Scroll Map**.

A Scroll Map depicts how many users (%) scroll down to any point in the page.

The **redder the area**, the larger percentage of users.

Can you think of how this information can be useful?

Source:https://www.searchenginejournal.com/use-heatmaps-seo-strategy/366268/#close

Slide Courtesy:Ms.Yousha Mahamuni

MATHEMATICS FOR COMPUTER SCIENCE ENGINEERS

Where are Heat Maps used in Computer Science/Business ?

Fig 1

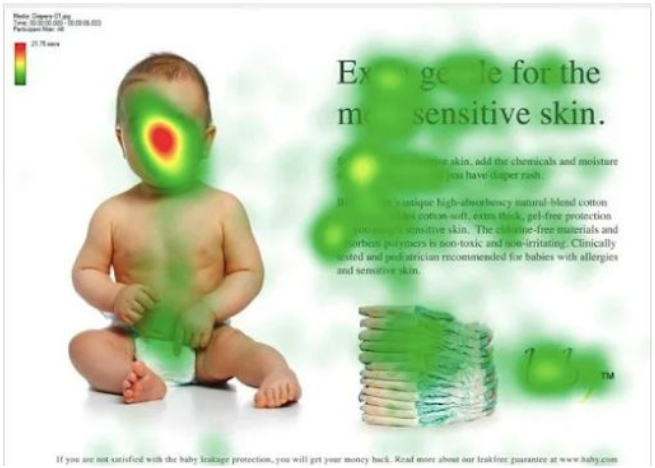
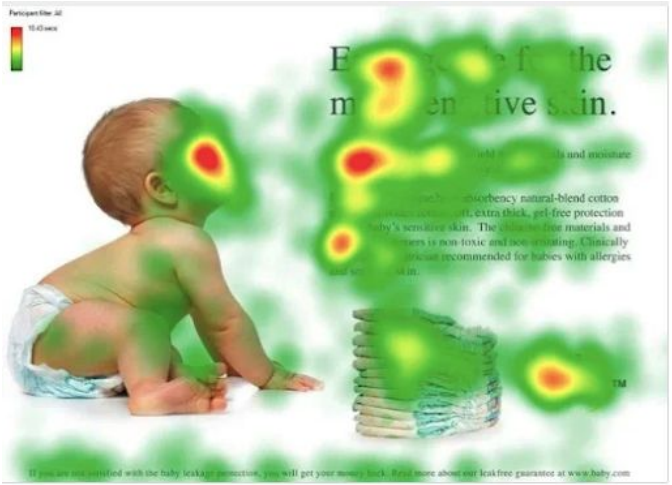


Fig 2



Which Figure is more beneficial to use as a business ?

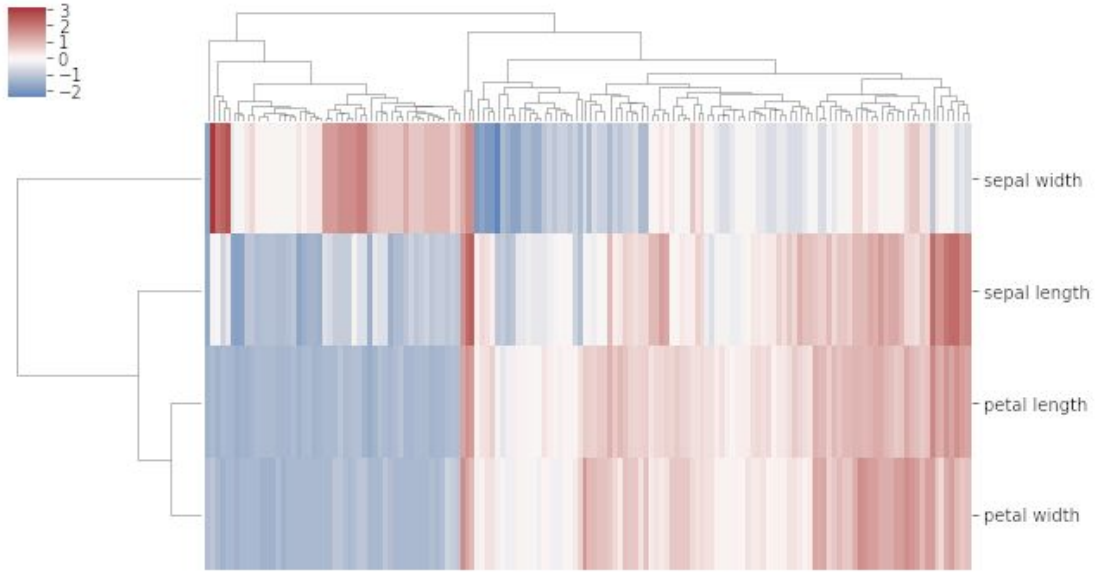
Figure 2 as more focus is on the message rather than the face of the character

Source:https://www.searchenginejournal.com/use-heatmaps-seo-strategy/366268/#close

Slide Courtesy:Ms.Yousha Mahamuni

Data Visualization: Clustered Heat Map

build associations between both the data points and their features.



In the above clustered heatmap, each column represents an individual flower specimen, and each row a measurement from that specimen.

Data Visualization: Clustered Heat Map-Applications

When we want to see which individuals are similar or different from each other, with a similar objective for variables.

Analysis tools that construct this type of heatmap will usually implement clustering as part of their process.

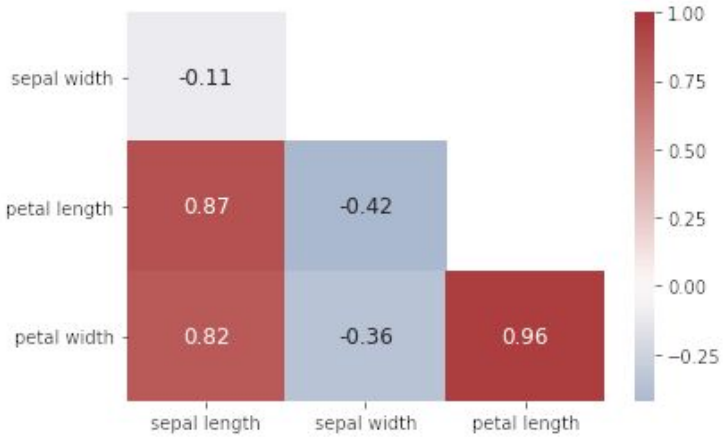
This use case is found in areas like the biological sciences, such as when studying similarities in gene expression across individuals.

Data Visualization: Correlogram

A correlogram is a variant of the heatmap that replaces each of the variables on the two axes with a list of numeric variables in the dataset.

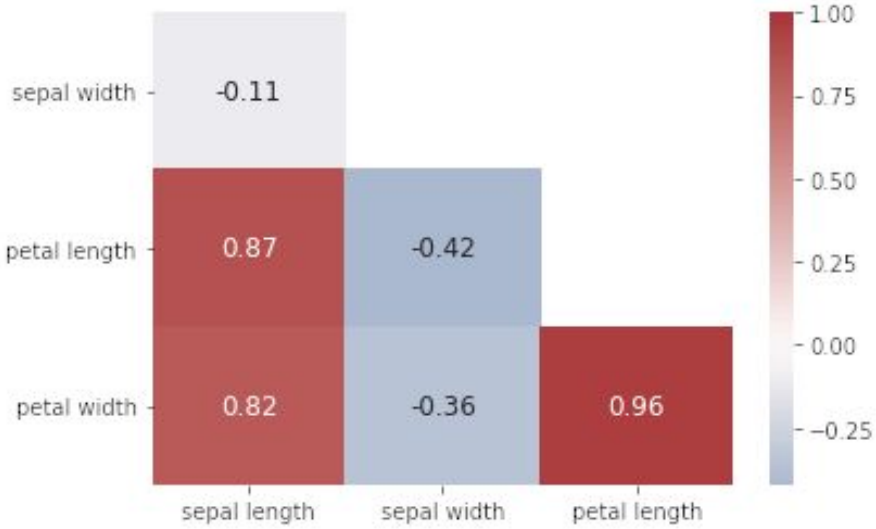
Each cell depicts the relationship between the intersecting variables, such as a linear correlation.

Sometimes, these simple correlations are replaced with more complex representations of relationship, like scatter plots.



Data Visualization: Correlogram

Correlograms are often seen in an exploratory role, helping analysts understand relationships between variables in service of building descriptive or predictive statistical models.



Petal length is highly correlated with petal width and sepal length; sepal width is negatively correlated with the other three variables.

Reference

https://chartio.com/learn/charts/heatmap-complete-guide/

THANK YOU

Mamatha H R

Department of Computer Science

mamathahr@pes.edu

MATHS FOR COMPUTER SCIENCE ENGINEER

Data Visualization and Interpretation

Mamatha H R

Department of Computer Science and Engineering
mamathahr@pes.edu

PES UNIVERSITY
CELEBRATING 50 YEARS

MATHS FOR COMPUTER SCIENCE ENGINEER

Data Visualization and Interpretation – Histogram

Dr. Mamatha H R
Department of Computer Science and Engineering

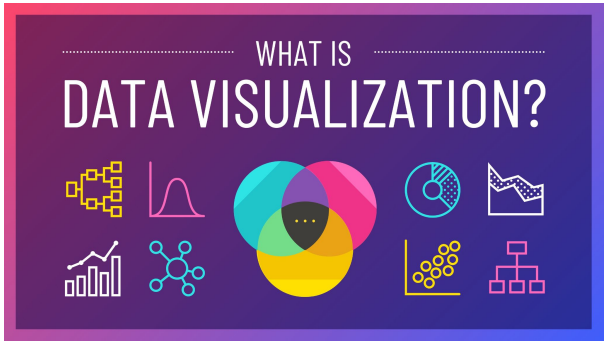
MATHS FOR COMPUTER SCIENCE ENGINEER

Data Visualisation

- **Data visualization** is the practice of translating information into a **visual context**, such as a map or graph, to make data easier for the human brain to understand and pull insights from.

- As the “age of Big Data” kicks into high-gear, visualization is an increasingly key tool to make sense of the trillions of rows of data generated every day.

- Good or Effective Data Visualisation must balance between form and function.



Source: venngage.com

Source: https://www.tableau.com/learn/articles/data-visualization

Graphic Displays of Basic Statistical Descriptions

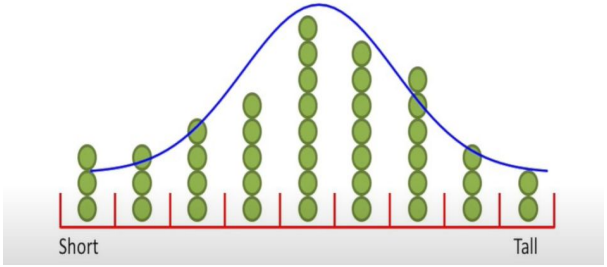
- **Boxplot**: graphic display of five-number summary
- **Histogram**: x-axis are values, y-axis repres. frequencies
- **Quantile plot**: each value x_i is paired with f_i indicating that approximately 100 f_i % of data are $\leq x_i$
- **Quantile-quantile (q-q) plot**: graphs the quantiles of one univariant distribution against the corresponding quantiles of another
- **Scatter plot**: each pair of values is a pair of coordinates and plotted as points in the plane

Which Graph to Use?

- Depends on **type of data**
- Depends on **what you want to illustrate**
- Depends on available **statistical software**

Histograms

- **Definition**: A diagram consisting of rectangles whose **area** is proportional to the **frequency** of a variable and whose **width** is equal to the **class interval**.
- Can only be used when the data is **continuous**, this is huge drawback of histograms. But are very powerful tools when representing the continuous data
- Histograms are similar to Vertical Bar graphs, however the key difference is Histograms **do not have gaps** between the bars.
- The first step in creating a histogram is to divide the entire value range into a series of intervals called "bins" and then to "drop" the individual values into the bins that they belong to.

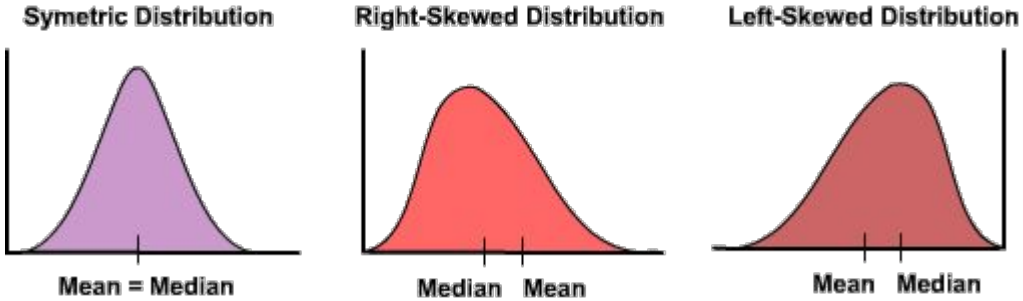


Source:StatQuest

Histograms

- The width of each bin may or may not be equal. If they're equal then, the height of bins represents the frequency of data points in that range. Else bin sizes can be made equal by calculating the density of bin heights.

- These become too handy when visualizing continuous data, as it can be used to show if the data is normalized, standardized, skewed.



Source: SaiGayathri Madali -Medium

MATHS FOR COMPUTER SCIENCE ENGINEER

Example: Histograms in Image Processing

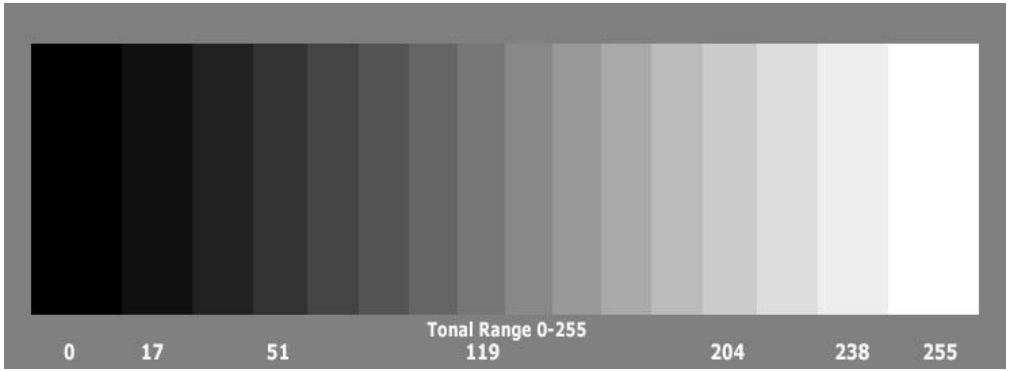
- Which fields in Computer Science make use of Histograms?
  - Image Processing
  - Computer Vision
  - Natural Language Processing

- In Image Processing, Histograms used to observe distribution of values taken by the pixels of an image.
- For example, plotting a histogram with grayscale (or degree of “grayness”) on the X axis for an Image. Frequency of different pixels having gray degree from 0 to 255 plotted on the histogram.

Source: https://www.eyesweb.org/introduction-to-computer-vision-stretching-and-histogram-equalization/

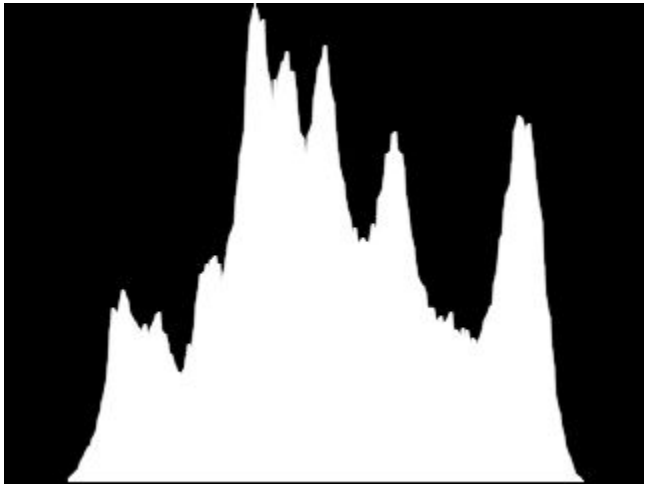
Example: Histograms in Image Processing

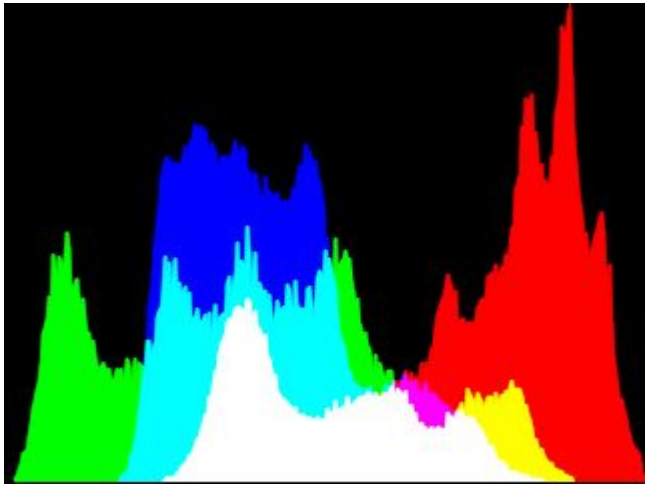
- Since the image is mostly very dark, the corresponding histogram has “shrunken” to the left, indicating grey level of most pixels are closer to black. *What kind of Skew is seen in the histogram below?*



Source: https://www.eyesweb.org/introduction-to-computer-vision-stretching-and-histogram-equalization/

Example: Histograms in Image Processing





Cumulated histogram

MATHS FOR COMPUTER SCIENCE ENGINEER

Histogram

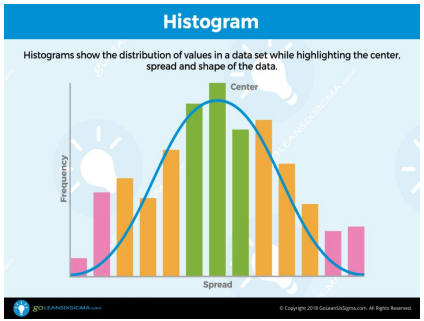
Histograms show the **distribution of data values** in a data set while highlighting the **center, spread and shape** of the data.

Histograms, also known as **Frequency Plots**, are a visual displays of **how much variation exists** in a process.

They highlight the **center of the data** measured as the **mean, median and mode.**

They highlight the **distribution of the data** measured as the **range and standard deviation.**

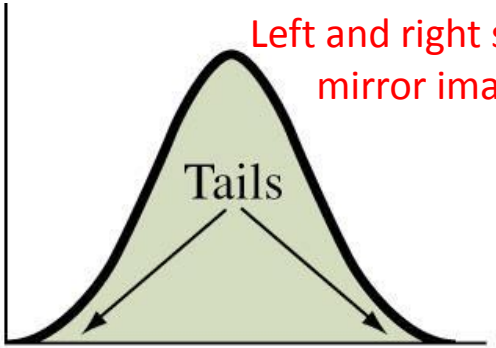
The **shape of a Histogram** indicates whether the distribution is **normal, bi-modal, or skewed.**



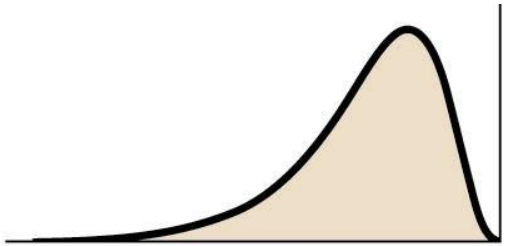
Source: goLEANSIXSIGMA.com

Interpreting Histograms

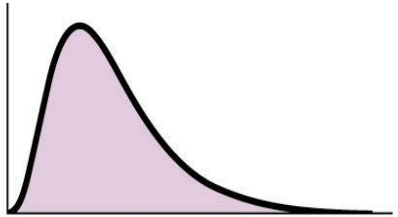
- Assess where a distribution is **centered** by finding the median
- Assess the **spread** of a distribution
- **Shape** of a distribution: roughly symmetric, skewed to the right, or skewed to the left



Symmetric



Skewed to the left



Skewed to the right

Distributions of a Histogram

A normal unimodal distribution:

In a normal distribution, points on one side of the average are as likely to occur as on the other side of the average.

Example: Weights and heights (when you look at genders **individually**) often follow this pattern.

Source: corporatefinanceinstitute.com

Slide Courtesy: Dr. Uma

Distributions of a Histogram

A bimodal distribution: In a bimodal distribution, there are two peaks.

In a bimodal distribution, the data should be separated and analyzed as separate normal distributions.

Example: If we were to plot weights or heights of men and women on the same Histogram, we would see two distinct peaks.

Source: corporatefinanceinstitute.com

Slide Courtesy: Dr. Uma

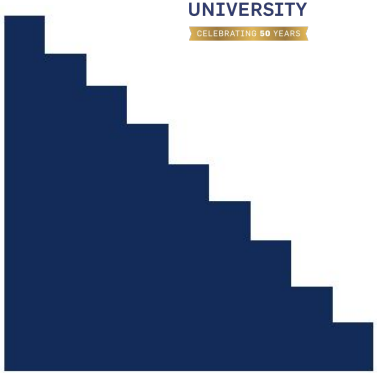
Distributions of a Histogram

A right-skewed distribution: A right-skewed distribution is also called a positively skewed distribution.

In a right-skewed distribution, a large number of data values occur on the left side with a fewer number of data values on the right side.

A right-skewed distribution usually occurs when the data has a range boundary on the left-hand side of the histogram. For example, a boundary of 0.

Example: U.S household incomes (or Income in general) are often right skewed as relatively few households are extremely rich.



Source: corporatefinanceinstitute.com

Slide Courtesy: Dr. Uma

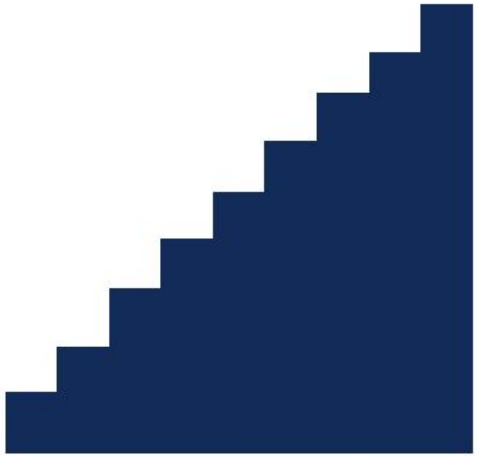
Distributions of a Histogram

A left-skewed distribution: A left-skewed distribution is also called a negatively skewed distribution.

In a left-skewed distribution, a large number of data values occur on the right side with a fewer number of data values on the left side.

A left-skewed distribution usually occurs when the data has a range boundary on the right-hand side of the histogram. For example, a boundary such as 100.

Example: Refer to the Image processing example in previous slides. In what case would there be a left skewed distribution?



Source: corporatefinanceinstitute.com

Distributions of a Histogram

A random distribution:

A random distribution lacks an apparent pattern and has several peaks.

In a random distribution histogram, it can be the case that different data properties were combined.

Therefore, the data should be separated and analyzed separately.

Source: corporatefinanceinstitute.com

Slide Courtesy: Dr. Uma

Test Your Understanding

What would be the type of distribution for the following cases?

- NBA Team Salaries (including star players)
- Histogram of when customers tend to enter a BBQ Nation restaurant in a given day.
- Marks obtained by students in an extremely easy examination
- Histogram of IQ scores of students in PES University

Test Your Understanding

What would be the type of distribution for the following cases?

- NBA Team Salaries (including star players)
- Histogram of when customers tend to enter a BBQ Nation restaurant in a given day.
- Marks obtained by students in an extremely easy examination
- Histogram of IQ scores of students in PES University

1-Right Skew
2-Bimodal (Lunch and Dinner)
3- Left Skew
4-Normal

Histograms- Frequency and Bins

- Histograms are based on area, not height of bars
- In a histogram, it is the area of the bar that indicates the frequency of occurrences for each bin.
- This means that the height of the bar does not necessarily indicate how many occurrences of scores there were within each individual bin.
- It is the product of height multiplied by the width of the bin that indicates the frequency of occurrences within that bin.

Histograms- Frequency and Bins

- One of the reasons that the height of the bars is often incorrectly assessed as indicating frequency and not the area of the bar is due to the fact that a lot of histograms often have equally spaced bars (bins), and under these circumstances, the height of the bin **does** reflect the frequency

- The **number of bins k** can be assigned directly or can be calculated from a suggested bin width h as:
  - k=(max-min)/h ---h is bin width
  - k=sqrt(n) ----used in Excel

Histograms- Frequency and Bins

- In statistics, the Freedman – Diaconis rule can be used to select the **size of the bins** to be used in a histogram:

$$\text{Bin size} = 2\,\frac{\text{IQR}(x)}{\sqrt[3]{n}}$$

Histogram -Frequency and Relative Frequency

- "**Frequency** " presents the numbers of data points that fall into each of the class intervals.

- "**Relative Frequency**" presents the frequencies divided by the total number of data points

- The relative frequency of a class interval is the proportion of data points that fall into the interval

- Note that since every data point is in exactly one class interval, the relative frequencies must sum to 1

Histogram -Examples

- "**Frequency Density**" presents the frequency divided by the class width
- "**Density**" presents the **relative** frequency divided by the class width.

- Note that when the classes are of equal width, the frequencies, relative frequencies, and densities are proportional to one another.

- When the class intervals are of unequal widths, the heights of the rectangles must be set equal to the densities. The areas of the rectangles will then be the relative frequencies.

Example 1– Construct a Histogram

The weather in Los Angeles is dry most of the time, but it can be quite rainy in the winter. The rainiest month of the year is February. The following table presents the annual rainfall in Los Angeles, in inches, for each February from 1965 to 2006.

0.2	3.7	1.2	13.7	1.5	0.2	1.7
0.6	0.1	8.9	1.9	5.5	0.5	3.1
3.1	8.9	8.0	12.7	4.1	0.3	2.6
1.5	8.0	4.6	0.7	0.7	6.6	4.9
0.1	4.4	3.2	11.0	7.9	0.0	1.3
2.4	0.1	2.8	4.9	3.5	6.1	0.1

Step:1 – Prepare the Data

Arrange the values in ascending order (number of data points (n) = 42)

0.0	0.1	0.1	0.1	0.1	0.2	0.2
0.3	0.5	0.6	0.7	0.7	1.2	1.3
1.5	1.5	1.7	1.9	2.4	2.6	2.8
3.1	3.1	3.2	3.5	3.7	4.1	4.4
4.6	4.9	4.9	5.5	6.1	6.6	7.9
8.0	8.0	8.9	8.9	11.0	12.7	13.7

Step: 2 Identify the Bin Widths

By using the Freedman – Diaconis , the bin width / class intervals can be found.

$$\text{Bin Width} = 2\frac{\text{IQR}(x)}{\sqrt[3]{n}}$$

Find the IQR (InterQuartile Range)

$$\text{IQR} = Q_3 - Q_1$$

Quartile 1, $Q_1 = 0.25\,(n+1) = 0.25\,(43) = 10.75$

0.0	0.1	0.1	0.1	0.1	0.2	0.2
0.3	0.5	0.6	0.7	0.7	1.2	1.3

$$\frac{0.6 + 0.7}{2} = 0.65$$

Step: 2 Conti..

Quartile 3, $Q_3 = 0.75\ (n+1) = 0.75\ (43) = 32.25$

3.1	3.1	3.2	3.5	3.7	4.1	4.4
4.6	4.9	4.9	5.5	6.1	6.6	7.9

$$\frac{5.5 + 6.1}{2} = 5.8$$

IQR = 5.8 – 0.65 = 5.15

Substitute in the formula, lets find the Bin width

$$\frac{2 * 5.15}{\sqrt[3]{42}} = 2.9 = (\sim 3)$$

Step: 3 Find the number of Bins / Buckets

$$\text{Number of bins / buckets} = \frac{\text{Max} - \text{Min}}{\text{Bin Width}}$$

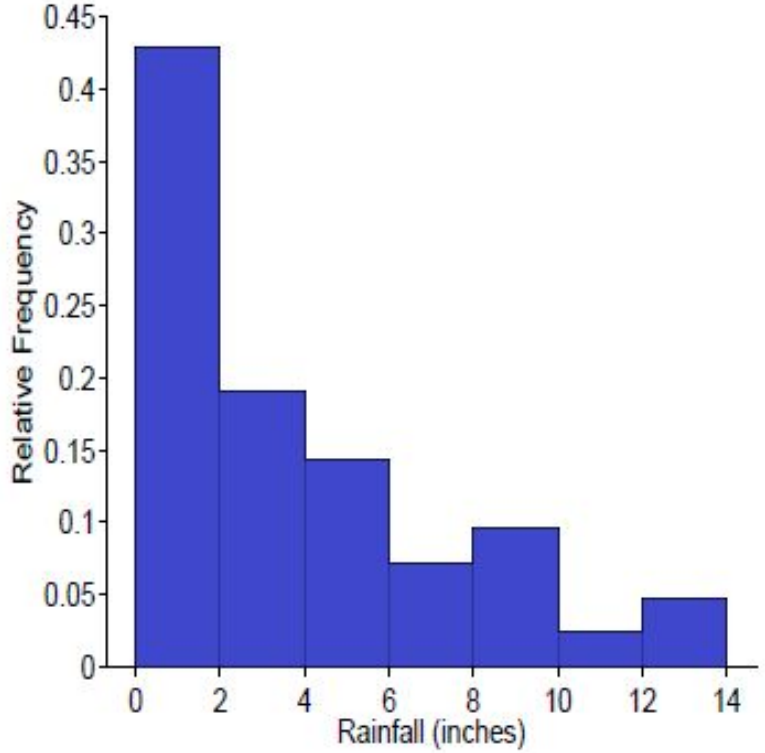
$$\frac{15 - 0}{3} = 5$$

Step: 4 Build the Frequency Distribution Table

Class	Frequency	Relative Frequency	Density
0 – 3	21	0.5	0.1667
3 – 6	11	0.2619	0.0873
6 – 9	7	0.1667	0.0555
9 – 12	1	0.0238	0.00793
12 - 15	2	0.0476	0.0159
	Sum = 42	Sum = 1	

MATHS FOR COMPUTER SCIENCE ENGINEER

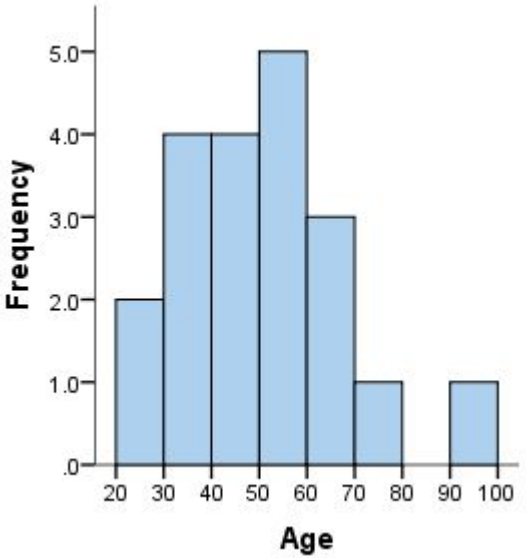
Step: 5 Plot the Histogram

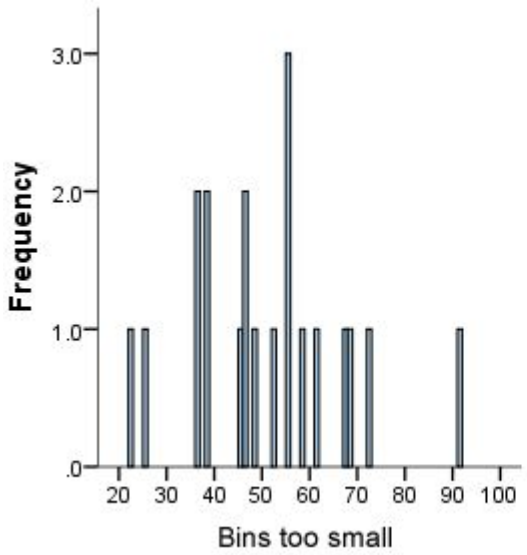


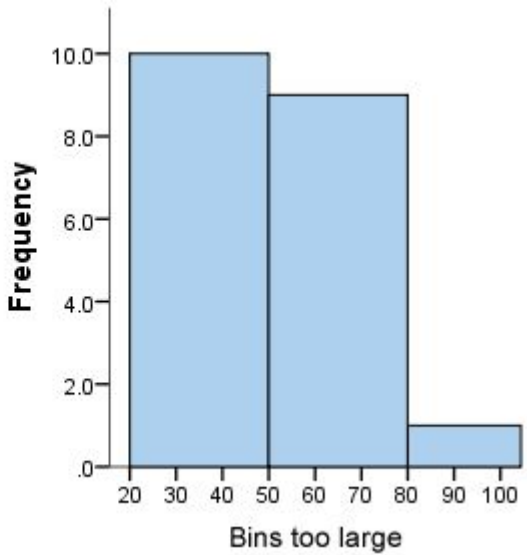
MATHS FOR COMPUTER SCIENCE ENGINEER

Histogram - Good Vs. Bad Visualisations

- Look at the images below. Which of the following graphs are a good representation and bad representation? Why?







Source: chttps://statistics.laerd.com/statistical-guides/understanding-histograms.php

Histogram - Good Vs. Bad Visualisations

- In order to have a Good Histogram Visualization, one of the key aspects is to ensure bin widths are appropriate. There is no right or wrong answer as to how wide a bin should be, but there are rules of thumb. You need to make sure that the bins are not too small or too large

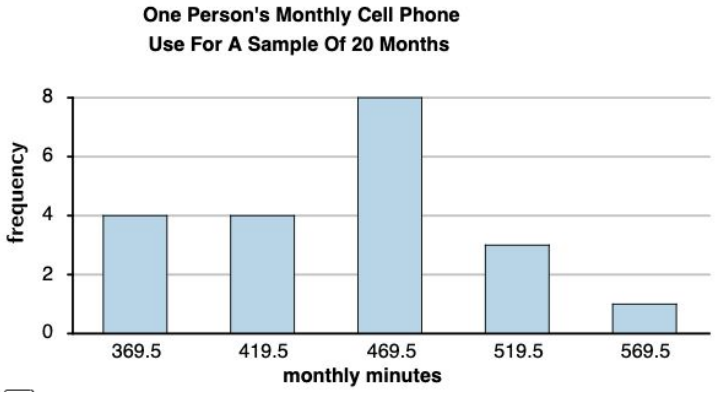
- We can see from the histogram in the center that the bin width is too small because it shows too much individual data and does not allow the underlying pattern (frequency distribution) of the data to be easily seen.

- At the other end of the scale is the diagram on the far right, where the bins are too large, and again, we are unable to find the underlying trend in the data.

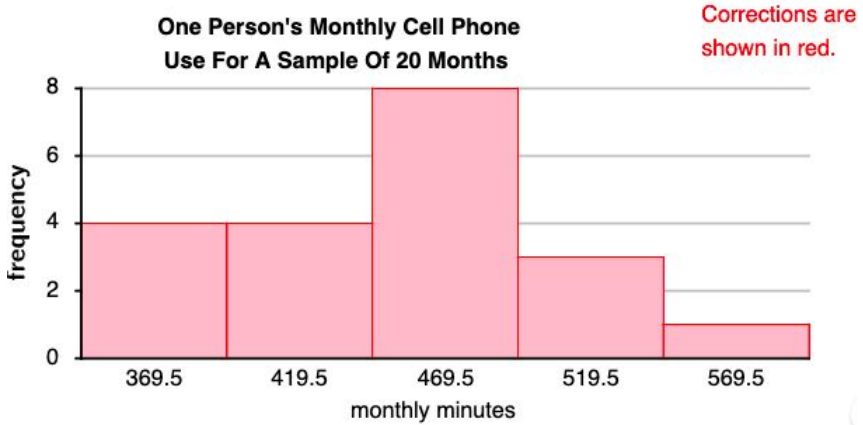
Source: chttps://statistics.laerd.com/statistical-guides/understanding-histograms.php

Histogram - Good Vs. Bad Visualisations

- For Histograms, ensure that there are no gaps between the bars. Refer to source link for further examples of bad visualisations in histograms.



Correction



Source: https://www.geogebra.org/m/u5qbmgfm

THANK YOU

Dr. Mamatha H R
Department of Computer Science and Engineering
mamathahr@pes.edu