Data Science
MD2201

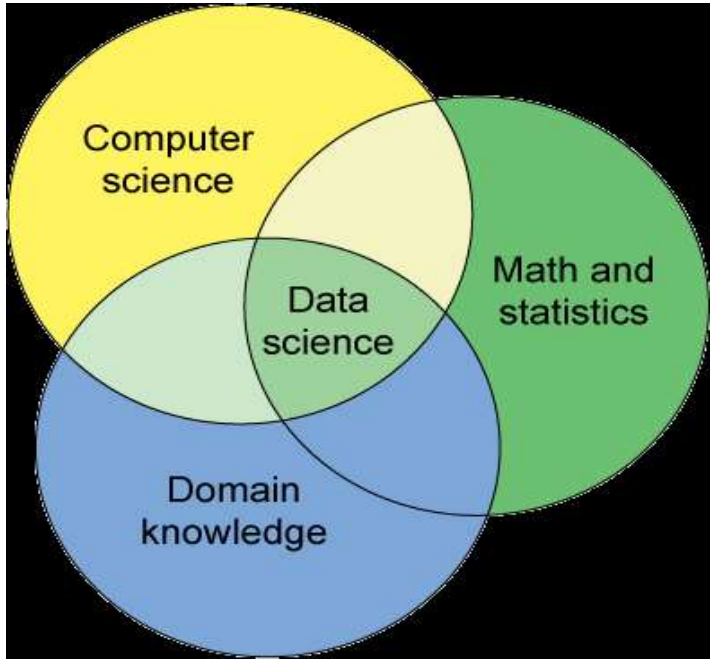
Why to study Data Science?

Lots of data everywhere around you.

- Social Media
- Banking, Railways
- Economic and Finance
- Sports
- E-commerce
- Conversational Agents
- Transport
- Health Care
- Industries
- Google,Yahoo

What is Data Science?

- It is a combination of Mathematics and statistics, Domain Knowledge and Computer Science



What is Data?

- Data are the values of qualitative or quantitative **variables** belonging to a **set of items**.
- Set of items may be the population we are interested in.
- **Variables** are the measurement or characteristics of the item. They might be measured in qualitative terms or quantitative terms.
- Qualitative terms could be such as country of origin, medical treatment , gender etc. Quantitative terms could be height, weight , blood pressure etc.

Example of Data with different variables

- The data set shown contains information about the drivers of different countries, their weight, gender, speed of vehicle(car) they drive.
- The data can be used to analyze the following :

i) What is the average speed of indian drivers?

ii) What is the general speed of female drivers from Italy?

iii) Whether Male drivers drive fast or female?

S.No	Name of driver	Country	Age	Avg Speed of vehicle	Gender
1	John	USA	28	80Km/Hr	M
2	Lisa	Italy	35	65 Km/Hr	F
...	…….	…….	……..	……	…..
100	Rahul	India	48	70Km/Hr	M

Set of items

Set of items may be population or the set of items we are interested in.

For example: To give the count of Indian census, the population of Indian people should be considered , population from other country will not make sense. Thus Indian citizen becomes our set of interests

Definition of Data Science

Data Science is the application of computational and statistical techniques to gain insight into a real world problem expressed using data.

Computational because data science typically involves some sort of algorithm methods written in a code.

Statistical because inferences based on statistics help us to build predictions that we make.

Raw Data

- It is also called as source data or atomic data.
- It is unprocessed data.
- This information may be stored in file or it can be mere collection of numbers , characters etc.
- This data can be entered in computer or it can be generated by the computer.
- It is hard to parse or analyze.
- This data is then cleaned and processed.
- The processing of raw data has to be carried out more than once, this record should be maintained.

Processed Data

- Some kind of cleaning, transformation is performed on raw data to get processed data which can be analyzed and visualized.
- It is the data which is ready for analysis.
- Processing can include merging, subsetting , transforming etc.
- It is important that all the steps should be recorded.

- *For example: Voltage signal from a microphone with each data point is raw data. It is then filtered to remove the noise is the processed data.*

Raw data to Processed Data

https://www.futurelearn.com/courses/technical-report-writing-for-engineers/0/steps/40143

Sources of Raw data

- Binary file generated by a measurement machine.
- Unformatted excel file.
- JSON from Twitter API.
- Hand entered numbers(readings) you collected.

Typical Attributes of Raw data

- No software is run on the data.
- No manipulation of any of the numbers in the data has taken place.
- No data is removed from the data set.
- The data set is not summarized in any possible way.

Expected Attributes of Processed data

- Each variable to be measured should be in one column.
- Each different observation of that variable should be in a different row.
- There should be one table for each kind of variable e.g data collected from Twitter should be kept in a separate table and the data from Facebook in another table etc.
- If you end up with multiple tables , they should include a column in the table that allows them to be linked. This is important for merging the dataset.

Minor Attributes:

- Include a row at the top of each file with variable names.
- Assign ’easily perceivable’ variables names – e.g. If a table is reporting User Experience as excellent, good, fair, poor etc. The variable name may be ‘user_exp’ instead of ‘ux’ etc.

Code Book or Meta data

It should contain following information about the variables:

- Units- whether the unit of column is in Rs in lacs or thousands.
- Summary choices-whether mean or median is used.
- Information about resource of data- which data base is used or structured survey etc is used.
- Valid link to the data base should be given.
- For structured survey- what was the population, whether it was observational or experimental design , selection of samples , confounding variables, information biases if any , mathematical formulation should be mentioned in the code book.

Case Study

- “Growth in a Time of Debt”, Reinhart C. and Rogoff K. , American Economic Review: Papers and Proceedings , Vol-100.
- Another Economist , Thomas Herndon got hold of raw excel file and metadata and proved that **coding errors**, **selective exclusion of available data** and **unconventional weighting of summary statistics** lead to serious errors.
- Hence the representation of relationship between public debt and GDP is inaccurate.
- “Does High public debt consistently stifle economic growth? A critique of Reinhart and Rogoff” , Herndon T., Ash M., Pollin R., PERI working paper series , no.32, April 2013.
- Case study highlights the importance of metadata in data processing and more importantly ethics in data science.

Extraction of data

- Real life situations offer data in a far from easier format.
- Data present in a file as line would be embedded with headers, strings, start , stop instructions etc. We should get the raw files, figure out their structure and extract the relevant bits.
- Another challenge is data may be well organized but may be in some different format which is difficult to analyze. Hence it should be first converted to the format which can be easily analyzed- API of Twitter is formatted in JSON format. We would want to reorganize the data in such a way that it is easy to analyze.

Extraction of data Contd….

- Sometimes data is available in the form of free text. This data may be pliable with human intelligence but can be a challenge as a task.—Doctor’s Prescription.
- Sometimes data would be given by two different sources for combining and joint processing.
- MySQL and MongoDB are popular free databases from where sources may provide data.
- Hence actually getting to know the raw data itself is a challenge in reality. Therefore, the knowledge of salient features of different types of formats is also necessary.
- Later on, based on the programming platforms, syntax can be set up to get the raw data

Data Cleaning

- It is defined as the process to ensure the correctness , consistency and usability of the data.
- For any type of data, data quality is very important.

Data quality criteria

1. ***Validity***: The degree to which the data conform to defined business rules or constraints. e.g dates should fall in typical range, certain columns cannot be empty etc.
2. ***Accuracy***: The degree to which the data is close to the true values. e.g the address of a street is given in valid format but is not true i.e the address does not exist.
3. ***Completeness***: The degree to which all required data is known.
4. ***Consistency***: The degree to which data is consistent. Inconsistency occurs when two values in the data set contradict with each other e.g Boy with age 10 years is defined as senior citizen.
5. ***Uniformity***: The degree to which data is specified using same unit e.g currency of one country is different from other.

Steps of data cleaning

- Monitor errors
- Standardize processes
- Validate accuracy
- Scrub for duplicate data

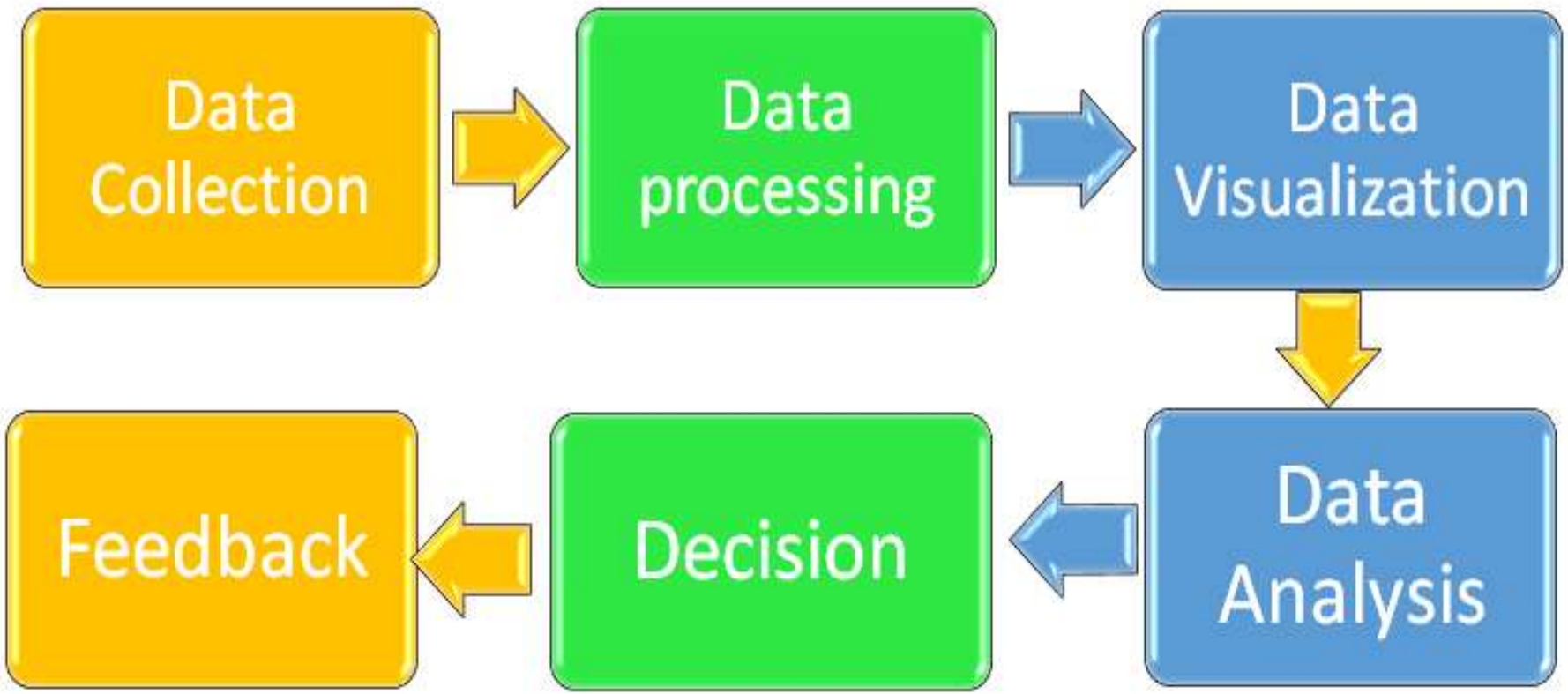
Data Cleaning

1. **Removing Irrelevant Data:** e.g analyzing health related data phone numbers not required.
2. **Removing Duplicates**: It happens when data are combined from different sources or say user has submitted submit button twice or request to online booking submitted twice etc.
3. **Type Conversion**: Ensure numbers are stored as numeric data type and not as string. If the value cannot be converted to any specific type then it should be converted to NA value and warning to be issued.
4. **Syntax errors**: Remove white spaces at the beginning or end e.g " Hello World " instead it should be "Hello World".

Data Cleaning Contd…

5. **Padding**: Some times to adjust the width of a number zeros are required to be padded e.g 456 to convert it to 6 digits for data uniformity zeros can be padded at the beginning i.e 000456.
6. **Fix Typos**: let gender variable is actually having two classes as female and male. But if used as female, fem,fe_male,male,M can be considered as different variable . This should be removed.
7. **Standardize**: To recognize the typo and to out them in standard format.For example for all strings use only one format i.e either upper case or lower case.
8. **Missing values**: The data is not recorded. It can be ignored or impute i.e finding that missing data from the available data or copying values from another similar records.

Data Science Pipeline



Data Collection

- All available datasets are gathered from structured/unstructured data sources such as internet , internal/external data bases/third party sources etc.
- Then their data is extracted into a usable format such as CSV,JSON,etc
- Typical skills required are

➢ Distributed storage: Hadoop , Apache spark

➢ Database Management- MySQL, MongoDB

➢ Handling relational data base queries.

➢ Retrieving unstructred data- text, audio ,video etc.

Data Processing

- Time consuming and laborious.
- The steps which we discussed for data cleaning, to be performed as per the requirement.

- Skills required are
  - Coding Language- R, Python
  - Data Modification Tools-NumPy, R, Pandas
  - Distributed Processing-Hadoop, Map Reduce

Data Exploration/Visualization

- Finds the patterns of the data.
- Different types of visualization and statistical techniques are used.
- Data reveals the trend through different graphs , charts and analysis.
- For understanding the visualization domain expertise is required.
- Skills required are
  - Python utilities-Numpy, Matplotlib , Pandas, Scipy
  - R utilities- GGplot2, Dplyr
  - Statistics- Inferential Statistics
  - Data Visualization- Tableau

Data Analysis/Machine Learning

- Its objective is to do the in-depth analytics, mainly the creation of relevant machine learning models such as predictive model or algorithm development.
- Second objective is it evaluates and refine the model. This involves multiple sessions of evaluation and optimization cycles.
- The accuracy of algorithm to be increased by training it with fresh ingestion of data, minimizing losses etc.
- Skill required are
  - Machine learning-supervised and unsupervised algorithms
  - Mathematics- Algebra, calculus

Decision

- Interpreting the data is more like communicating the decision to interested parties.
- The objective of this step is to first identify the business insight and correlate it to the decision.
- Involve domain experts in correlating the decisions with business problems
- Domain experts help in visualizing the decisions according to the business dimensions which also assists in communicating facts to a non technical audiences.
- Skill required are

➢ Business domain knowledge

➢ Data visualization tools- Tableau,D3.js,seaborn

➢ Communication-Presentation, speaking, reporting, writing

Feedback

- It is important to revisit and update the model on a periodic basis, depending on the frequency of new data generation.
- The more the data that is received , the more frequent the feedback is.
- Thus, data processing pipeline must connect ,collect ,integrate ,cleanse , prepare, relate ,protect and deliver trusted data at scale and at the speed of business.

Types of Variable

- Numerical (Quantitative)- Takes numerical values , one can add , subtract , take averages with these values .
  - ➢Continuous
  - ➢Discrete
- Categorical (Qualitative)- Takes on a limited number of distinct categories. The categories can be identified with numbers as well , but it is not sensible to do arithmetic operations, e.g code 0 can be used for male and 1 for female as a gender variable . But doing arithmetic operations such as averaging, adding say number of girls and boys would not yield meaningful outcomes.
  - ➢Ordinal
  - ➢Regular

Numerical Variable

- ***Continuous Variable*** : Takes any of the value in given range e.g height. (5.4”).
- ***Discrete Variable***: Takes one of the specific set of numerical values e.g number of two wheelers ,no. of houses a person has.(3,4).
- Height is a continuous variable but we tend to report it as discrete variable by rounding its value. Thus rounding of continuous variable make them appear as discrete.

Categorical Variable

- ***Ordinal categorical variable***: The categorical variables having ordered level are ordinal variables e.g customer satisfaction survey for a service can be very satisfactory, satisfactory , neutral , unsatisfactory , very unsatisfactory or speed if low, medium , high etc.
- ***Regular categorical variable***: Does not have any order. E.g which chocolate do you like dark or white, gender is female or male etc.

Case Study : Google Transparency Report

Country	Cr_req	Cr_comply	Ud_req	Ud_comply	Hemisphere	hdi
Argentina	21	100	134	32	Southern	Very high
Australia	10	40	361	73	Southern	Very high
Belgium	6	100	90	67	Northern	Very high
Brazil	224	67	703	82	Southern	high
USA	92	63	5950	93	Northern	Very high

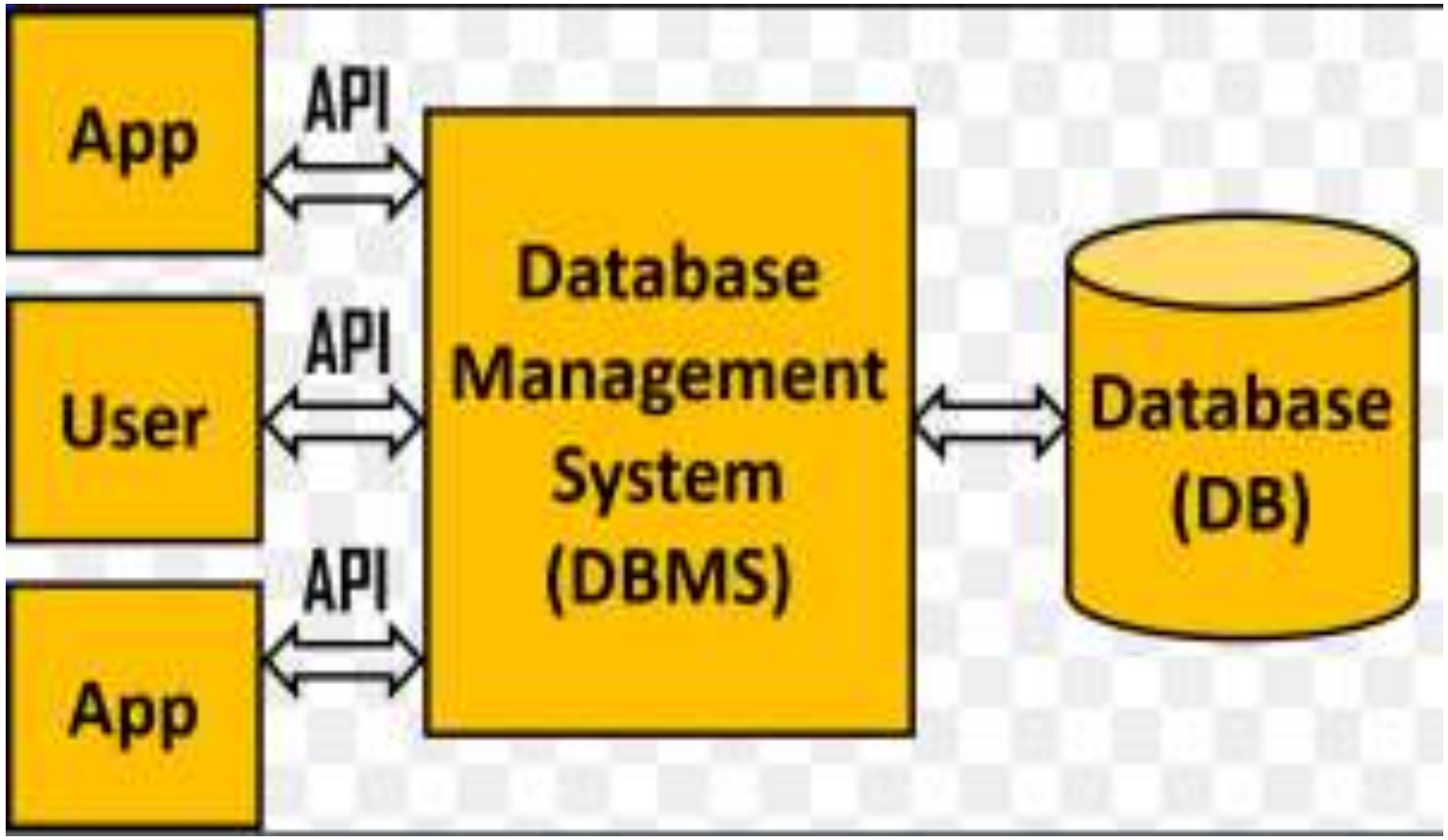
Description of variables

- Country: Identifier variable, indicating name of the country for which data are gathered.
- Cr_req: No.of content removal requests made by the respective country. It is Discrete numerical variable.
- Cr_comply: Percentage of content removal requests that Google has complied with. It is Continuous numerical variable.
- Ud_req: No.of user data requests by the country as a part of criminal investigation. It is discrete numerical variable.
- Ud_comply: Percentage of user data requests that Google has complied with. It is continuous numerical data variable.
- Hemisphere: Whether the country is in southern or northern hemisphere. It is regular categorical variable.
- Hdi: human development index: It combines indicators of life expectancy, educational attainment and income and is released by United Nations(UN). It is ordinal categorical.

Data base and DBMS

- It is a collection of logically related data. Data base management system is a software (DBMS) which manages this data
- It is a layer between programs and data
- It makes it possible for user to create , read, update and delete data
- It provides features for data like data retrieval, data integrity , data modification , data security

DBMS


App
User
App
API
API
API
Database Management System (DBMS)
Database (DB)

Advantages of DBMS

- Reducing data redundancy: Being a single entity any change in it is reflected immediately. Because of this , there is no chance of encountering multiple data or duplicate data.
- Sharing of Data: In a database, users of the database can share data amongst themselves. There are various levels of authorization to access the data, and consequently the data can only be shared based on correct authorization protocols followed. Remote users can also simultaneously access and share the data between themselves.
- Data Integrity: It mean data is accurate and consistent.
- Data Security: Only authorized users are allowed to access the data.
- Backup and recovery: DBMS takes care of backup and recovery. It also restores the database after a crash or system failure to its previous condition.

Structured Query Language

- SQL is a structured language for accessing and manipulating databases.
- SQL can execute queries against a database. It can retrieve data from a database.
- It can insert, update or delete records from a database
- SQL can create new databases, or creates new tables in a database.

Univariate , Bivariate , Multivariate Data analysis

- Univariate Data: Contains only one variable.

e.g: Height of applicants in army recruitment camp measured in cms.

Height	164	168	170	172	174	178	180	….

- For the description of such type of data measures of central tendency like mean ,median , mode etc or dispersion or spread of data such as range , minimum ,maximum ,quartiles , variance etc can be used.
- Also frequency distribution tables ,pie charts ,bar charts , box plot , histogram etc can be used.

Bivariate Data

- Bivariate contains two different variables.
- Deals with cause and consequence relationships and the analysis is done to find out the relationship among two variables.

Temp in deg	20	25	35	43
Ice cream sales	2000	2500	5000	7800

- It shows relationship between temperature and sales is directly proportional.

Bivariate Contd…

- It involves comparisons ,relationships ,causes and explanations.
- These are generally plotted on X and Y axis on the graph for better understanding.
- One of the variable is independent (such as Temperature) while the other one is dependent (such as ice cream sales).

Multivariate Data

- It contains three or more variables.
- For example research work of a doctor who has collected data on eating habits of the participants such as meat consumption , dairy products and chocolate consumed per week and their respective cholesterol ,blood pressure and weight data.
- Techniques used are regression analysis , factor analysis ,analysis of variance i.e ANOVA etc.

Case study 1- Display of categorical data

- A Survey was conducted with a group of 20 persons. They were asked to inform about their hair and eye color.
- A 2 way contingency table is formed as under-

Hair color	Eye 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**

Questions

1. How many people have Brown eye color?

   10

2. How many people have Blonde hair?

   6

3. How many Brown haired people have Black eyes?

   2

4. What is the percentage of people with Green eyes?

   10

5. What percentage of people have red hair and Blue eyes?

   5

Case Study 2- Display of numerical data

- Following data shows the experiment data to study the effect of different amounts of water on the germination of seeds.
- For each amount of water,4 identical boxes were sown with 100 seeds each and the number of seeds having germinated after 2 weeks was recorded.
- The experiment was repeated with the boxes covered to slow the evaporation.
- There were six levels of watering, coded 1 to 6 with higher codes corresponding to more water.

Case Study 2- Display of numerical data Contd…

1. Uncovered boxes

	Amount of water					
	1	2	3	4	5	6
Number of seeds germinated per box	22	41	66	82	79	0
	25	46	72	73	68	0
	27	59	51	73	74	0
	23	38	78	84	70	0

Case Study 2- Display of numerical data Contd…

2. Covered boxes

	Amount of water					
Number of seeds germinated per box	1	2	3	4	5	6
	45	65	81	55	31	0
	41	80	73	51	36	0
	42	79	74	40	45	0
	43	77	76	62	39	0

Questions

1. What is the average number of seeds germinated for the uncovered boxes with level of watering equal to 4?

   78

2. What is the median value for the data covered boxes?

   45

3. Establish conclusions on the basis of available data:

   a. Association of levels of watering with the number of germinating seeds in case of covered boxes as well as uncovered boxes.

   b. Association of number of germinating seeds with the fact that the boxes were covered or uncovered.

Case study 3- Numerical data display.

- The data gives the production of wheat in years 2015 and 2016 for top 12 wheat producing states in India:
- ..\RData sets\wheat box.xlsx
- Question:

  Plot the box plots for the wheat produce as applicable for 2015 and 2016.

To be carried out in Lab

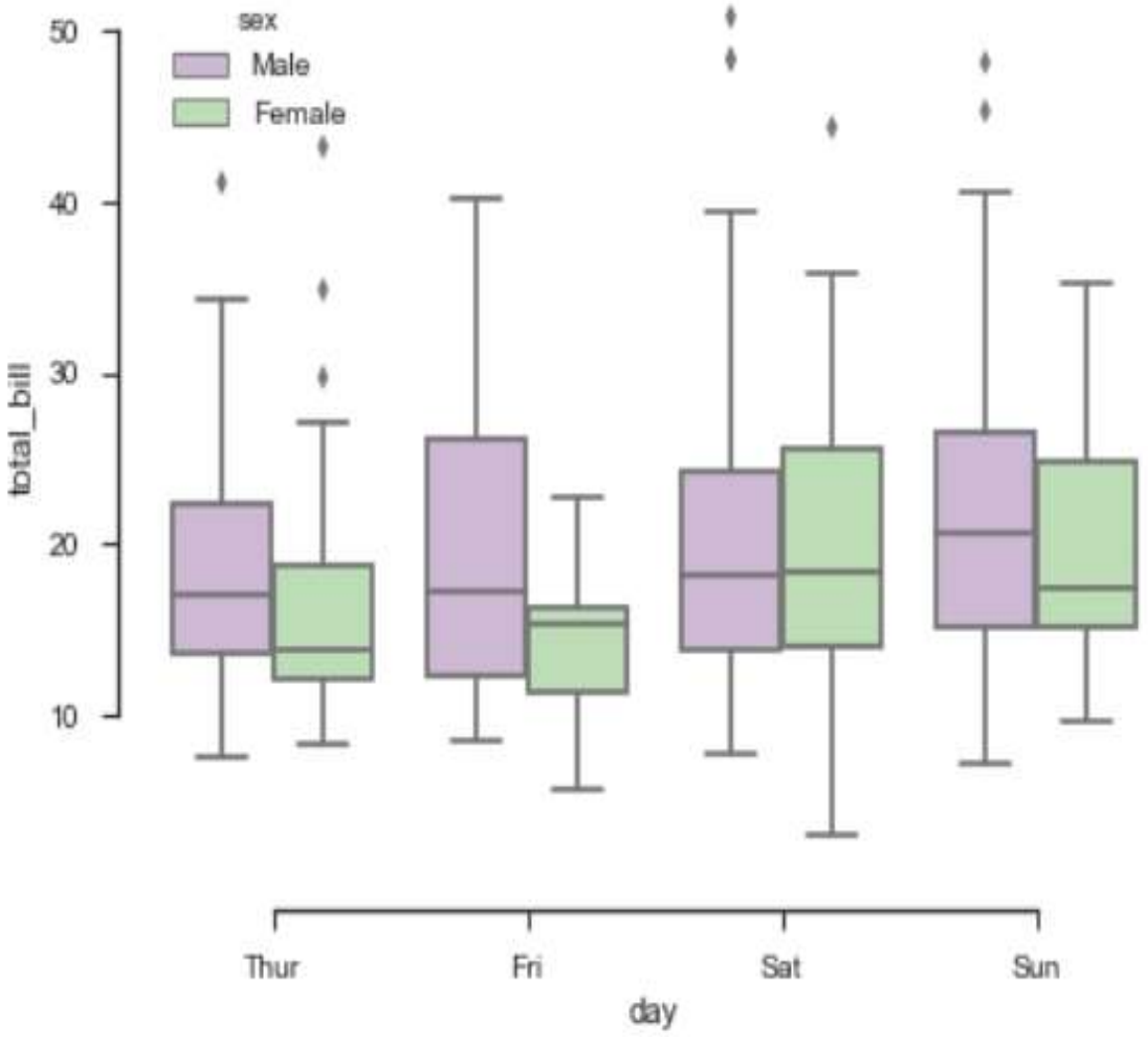
Case study 4

- The data set is a part of a larger data set. The small dataset we consider , gives measurements of two flower parts sepal length and sepal width in cms on 50 specimens of each of two species of the iris flower , namely Iris setosa and Iris Virginica.
- ..\RData sets\iris data.xlsx
- Question:

  Plot a scatter plot with sepal length on X-axis and Sepal width on Y axis. The scatter plot should be a common single plot.

To be carried out in Lab

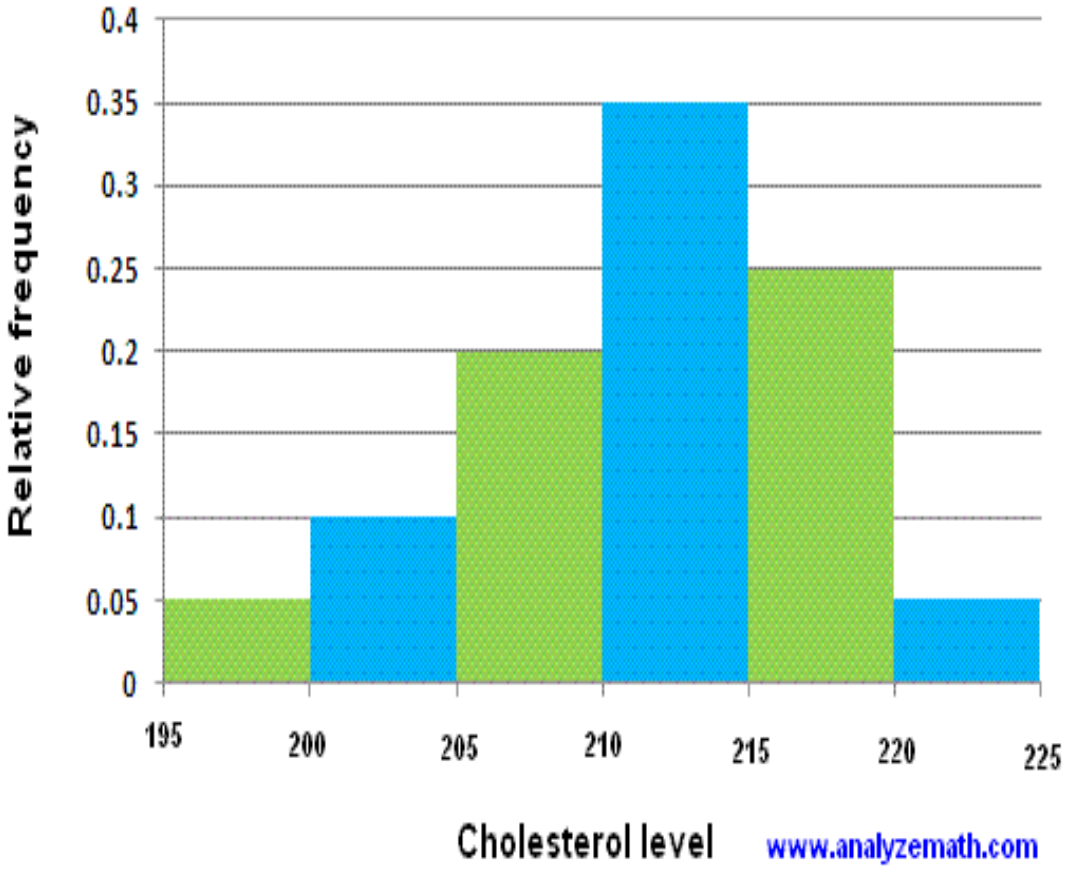
Box plot showing the amount of bill spend



- Males and females spend on an average different amount on bills.
- Whiskers shows for men the bill amount variability is more as compared to females except Saturday night.
- E.g Friday night males , 75% of data is below $25 i.e 25% spend bill greater than $25.

https://magoosh.com/statistics/reading-interpreting-box-plots/

Level of cholesterol (in mg per dl) of 200 people



- How many people have a level of cholesterol between 205 and 210 mg per dl?
- 40
- How many people have a level of cholesterol less than 205 mg per dl?
- 30
- What percentage of people have a level of cholesterol more than 215 mg per dl?
- 30%
- How many people have a level of cholesterol between 205 and 220 mg per dl?
- 160

Draw Histogram for following data

Customer Wait Time in Seconds (n=20)	
43.1	42.2
35.6	45.5
37.6	30.3
36.5	31.4
45.3	35.6
43.5	45.2
40.3	54.1
50.2	45.6
47.3	36.5
31.2	43.1