Introduction to Data Science with R (Notes)

What is Data ?

Data is the collection of the information(different from what is known) that is produced by people, sensors, machines ...

Data is information that can be used repeatedly but cannot be used alone. It needs to be analyzed to be used.

Data becomes an information that is what we know after it is processed. For Example, the grades of the students are a data but the average of the grades is an information.

The Term of Data

1. Big Data
2. Data Center
3. Database
4. Data Mining
5. Pure Data
6. Structured Data
7. Unstructured Data

What is Data Type ?

Data Types split into 2 group called as Qualitative(Categorical) and Quantitative.

|  |  |
| --- | --- |
| Quantitative | Qualitative |
| Weight | Gender |
| Height | The name of … |
| Number of objects | The places that is your favorite |
| … | … |

Quantitative is split into two groups , Continuous and Discrete

Qualitative is split into two groups too, Ordinal and Nominal (2.2.1 image)

Data Collection Methods

1. Transaction Data
2. Web Data
3. Sensor Data (IoT)

Data Mining

Data Mining the process of analyzing large data sets to find useful information. It combines statistics and AI for this purpose. It can be said the statistical beginnings of data mining starts with discovery of Regression analysis.

Data mining conceptually emerged in the 1960s with the use of computers and computing. As the growth of computer processors, data storage, and technology , data mining became not only more powerful, but also more prolific and important in all kinds of situations. In Today, Data mining nearly impossible without using the use of computer and computing.

What is Database?

A database is an organized collection of structured information, or data, typically stored electronically in a computer system.

Databases includes large data sets, which make data analyses harder. Thus, We need a software to help us to manage datas. It is called as Database Management System. Database Management System handles controlling, accessing, arranging, sorting and more operation for us to manage database. Popular Database Management Systems are MySQL, Oracle…

Data Center

Data Center is a facility that has a physical location to store computing machines and their related hardware equipment(storage component, processor,…). Governments build data centers since they want to keep information securely in their own hands. Thus, They don’t use Cloud System that is a virtual resource that helps businesses to store, organize, and operate data efficiently.

What is Data Science?

Data science is a multidisciplinary area that uses scientific methods, anaylzing methods, algorithm, and systems to obtain information and result from structured and instructed datas. Combining two different types of data and making them meaningful becomes possible with Computer Science. Thus, Data science is a multidisciplinary area and more modern than Statistics. So, Data Science couldn’t continue in subset of Statistics.

History of Data Science

The reason why the history of Data Science is explained is to understand how Data Science becomes multidisciplinary. When statisticians began using computers to analyze data and perform statistical computations, the foundations of Data Science can be traced back to the 1960s and 1970s. In 1990s and 2000s, Companies started to gathering large amounts of Information especially personal Informationg for marketing, the International Federation of Classification Societies became the first conference to specifically feature data science as a topic. The Companies such as Google, Amazon started to be leader companies by using power of Datas. William S. Cleveland laid out plans for training data scientists to meet the needs of the future. He presented an action plan titled, Data Science: An Action Plan for Expanding the Technical Areas of the field of Statistics. In 2002, the International Council for Science: Committee on Data for Science and Technology began publishing the Data Science Journal, a publication focused on issues such as the description of data systems, their publication on the internet, applications and legal issues. Articles for the Data Science Journal are accepted by their editors and must follow specific guidelines. Data Science is said to be one of the most important professions of the future in lates 2000s and early 2010s.

The Relationship of Data Science with other Sciences

The main branches of science in data science are Statistics, Mathematics and Computer Science. Most data control methods, algorithms and terms come from statistics and mathematics. In addition, the use of the computer is crucial for data storage, management, and performing these mathematical and statistical operations.

KVKK and other the Rules of the

Protecting Personal Information(e.g. GDPR from EU)

Data needs to be protected for several critical reasons, such as theft, hacking, scam, stealing, privacy violation, exploiting and more. (In according some researches, The global cyber security landscape has seen increased threats in recent years. Through the pandemic, cyber criminals took advantage of misaligned networks as businesses moved to remote work environments. In 2020, malware attacks increased 358% compared to 2019. It is just one example of ascendence of Cyber Crime. Nearly 1 billion emails were exposed in a single year, affecting 1 in 5 internet users. The criminals cannot make these without data such as Personal Information, Financial Information, Account Passwords, Health Records, …

These rules try to prevent the violation of personal rights as a result of random and unlimited collection, unauthorized access, disclosure or abuse of personal data.

What are Data Science tools?

The actaul thing to learn is to learn which program to use at which stage of Data Science.

Most popular data science programs today are R, SQL, Python, C++, …

What are data science career fields?

There are many career areas of data science, which is used in almost every field. For example, Data Sciencist, Machine Learning Engineer, Data Engineer, Veri Analyst, …

Statistics

Statistical thinking

1. Determining the context of the project
2. Determining of datas
3. Grouping of datas
4. Analyzing of datas
5. Converting datas to information

Statistics are divided into 2 group, Descriptive and Inferential Statistics

Sample

A representative subset of a Population.Using Samples instead of using Population reduces costs, saves time and makes work easier.

The Sample should be equally weighted and randomly selected.

The Sample does not fully reflect the population.

The magnitude of Sample should be selected in a way that is appropriate for cost, time and Population. In this way, an efficient solution can be produced faster and at a lower cost.

Methods of sampling from a population

**Probability Sampling Methods**

1. Simple random sampling
2. Systematic sampling
3. Stratified sampling
4. Clustered sampling

**Non-Probability Sampling Methods**

1. Convenience sampling
2. Quota sampling
3. Judgement (or Purposive) Sampling
4. Snowball sampling

Bias is important term for Sample. My Bias Article can be read

Observation unit

the entity on which information is received and statistics are compiled in the process of collecting statistical data.

Parameter

Parameter is a numerical property related to the population. The numerical information obtained from the population is called a parameter, and the numerical information obtained from the sample is called statistics. Parameter and statistic can have different representation for the same unit.

Variable

Properties that can change from observation to observation, take new values ​​or remain constant are called variables.

Variable Types

1. Nominal - Ordinal
2. Discrete - Continuous
3. Dependent – Independent

Scale Types

* Nominal Scale.
* Ordinal Scale.
* Interval Scale.
* Ratio Scale.

Arithmetic Mean

is the sum of a collection of numbers divided by the count of numbers in the collection.

The Arithmetic Mean of Frequency Table is that the sum of collection of numbers that are multiplied by its frequency divides by the sum of the frequency

The Arithmetic Mean of Grouped Series is that the sum of collection of middle point of the group that are multiplied by its frequency divides by the sum of the frequency.

Median

A median is the middle number in a sorted list of numbers (either ascending or descending).

Mode

is the value that appears most frequently in a data set.

Quartile

a quartile is a type of quantile which divides the number of data points into four parts, or quarters, of more-or-less equal size. Second Quartile is the Median of the Population.

Range

Is the difference of maximum value of data set and minimum value of data set.

Standard Deviation

is a measure of the amount of variation or dispersion of a set of values.

Variance

Is measure of dispersion, meaning it is a measure of how far a set of numbers is spread out from their average value. The variance is also often defined as the square of the standard deviation.

Skew

is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean.

Kurtosis

a measure of the tailedness of a distribution.

Confidence Interval

is a range of estimates for an unknown parameter.

Statistical Hypothesis Testing

is a form of statistical inference that uses data from a sample to draw conclusions about a population parameter or a population probability distribution.

Correlation

is any statistical relationship, whether causal or not, between two random variables or bivariate data.

Regression

is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables).

Linear Algebra

What is Linear Algebra ?

Linear algebra is the field of mathematics that deals with vectors and matrices. Linear algebra is also used in most sciences and fields of engineering, because it allows modeling many natural phenomena, and computing efficiently with such models.

I did not continue to take notes for this part because I have notes on this subject.

R Programming Basics

MAİN SOURCE:

BTK AKADEMİ – Introduction of Data Science with R

Sources to make strengthening the understanding of topic:

https://www.javatpoint.com/

<https://www.dataversity.net/>

<https://aag-it.com/the-latest-cyber-crime-statistics/#:~:text=Nearly%201%20billion%20emails%20were,their%20accounts%20breached%20in%202021>.

<https://www.healthknowledge.org.uk/public-health-textbook/research-methods/1a-epidemiology/methods-of-sampling-population>

<https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Observation_unit#:~:text=An%20observation%20unit%2C%20sometimes%20also,item%20at%20a%20given%20outlet>.

<https://www.investopedia.com/terms/r/regression.asp>

<https://www.wikipedia.org/>