UGANDA CHRISTIAN UNIVERSITY
A Centre of Excellence in the Heart of Africa

DSC3108: Big Data Mining and Analytics

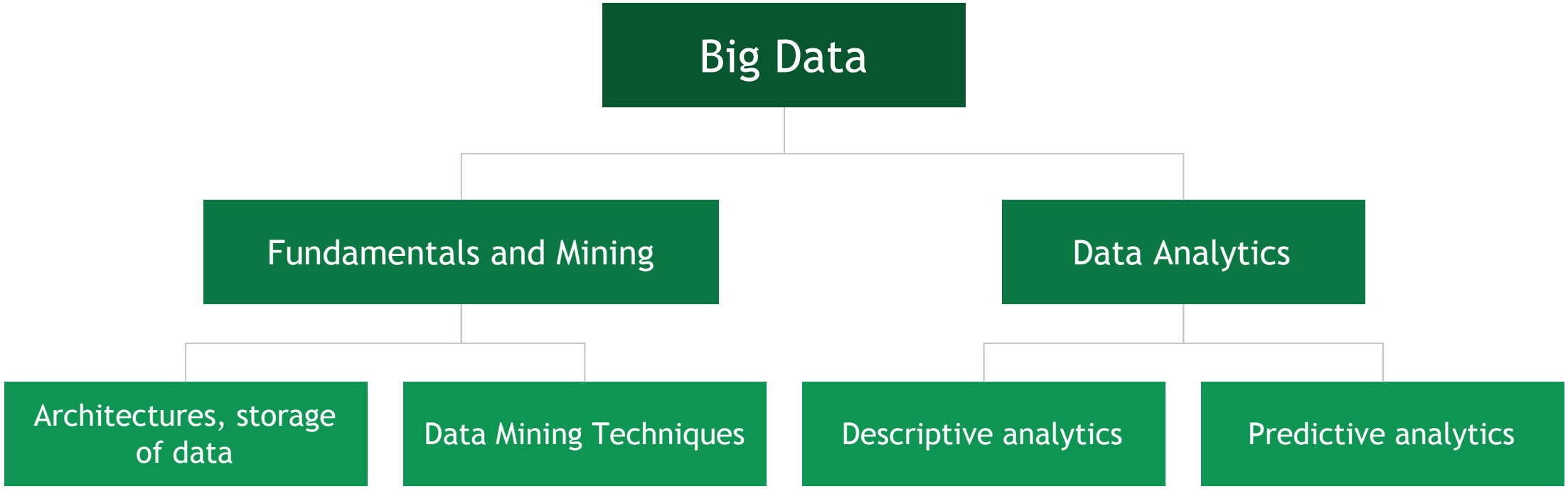
Lecture 05 (BSCS_3:1)

Topic: *Descriptive analytics*

Dr. Daphne Nyachaki Bitalo
Department of Computing & Technology
Faculty of Engineering, Design & Technology

Thur 26th Sept 2024

COURSE OVERVIEW



Lecture Objectives and Learning outcomes

The Objectives of this lecture are :

- Understand the principles of exploratory data analysis.
  - Explore relationships between variables

By the end of this lecture, students should be able to:

- Practically explore relationships between variables
  - Use visuals to exhibit relationships
  - Use statistical tests

Lecture Overview

- Concepts in the EDA cycle
- Data visualisation
- Relationships btn variables
- Hands-on practical

UGANDA CHRISTIAN UNIVERSITY
A Centre of Excellence in the Heart of Africa

DSC1101-Introduction to Data Science

Lecture 7
RECAP

Lesson Recap:

Handled the following under the data flow pipeline

- Handling Outliers
- Handling Missing data
- Handling Duplicates

Lesson recap: EDA

An interactive cycle;

1. Generates questions about your data.
2. Searches for answers by visualizing, transforming, and modeling your data.
3. Uses what you learn to refine your questions and/or generate new questions.

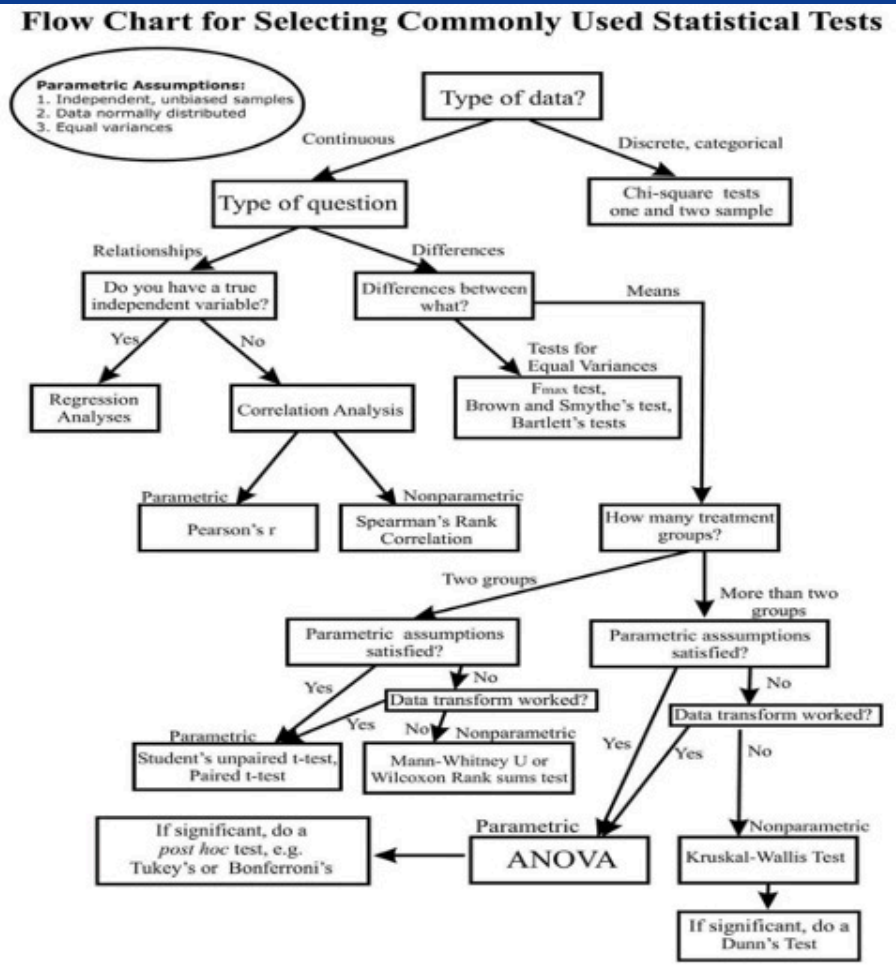
There is no rule about which questions you should ask to guide your research. However, two types of questions will always be useful for making discoveries within your data. You can loosely word these questions as:

1. **What type of variation occurs within my variables? (i.e. variability)**
2. **What type of covariation occurs between my variables? (i.e. central tendency)**

All of the above questions generate **DESCRIPTIVE** and NOT predictive analyses

Exploratory Data Analysis (EDA)

Key statistical tests under EDA



Process of EDA

1. Data types determine how relationships are explored
2. We can use visuals (graphs)
3. E.g. Continuous variables relationship- scatterplot
4. Explore relationships using statistics

Statistical significance

1. p-value adds statistical credence to the tests for null/ alternative hypotheses stated
2. Works for normally distributed data (parametric testing)
3. t-test for continuous variables/data
4. chi-squared test for categorical/qualitative data
5. ANOVA for more in-depth analyses

Statistical significance

6. Stats range from 0-1
7. Thresholds for error usually at 0.05
8. p =<0.05 reject null hypothesis
9. p>= 0.05 fail to reject the null hypothesis

Process of EDA: Statistical investigation

1. Two continuous variables: correlation test (r= -1 to +1)
2. Strong relationship (either >=+0.7, -0.7)
3. Negative correlations exhibit an inverse relationship (i.e. one increases, other decreases)
4. Medium relationship (ranges from 0.3 to 0.69)
5. Less than 0.3 or -0.3 is a weak relationship

Statistical investigation: T test

1. Works for normally distributed data (parametric testing)
2. t-test is suited for continuous variables/ data

Statistical investigation: ANOVA

1. Tests relationship between categorical and continuous variables
2. Hypotheses: Assumption/presumptive statement
3. Null hypothesis: Negative manner (e.g. There's NO relationship between state and revenue)
4. Can be true if; p-value is > 0.05 (No relationship between state and revenue)

Statistical investigation: ANOVA

However;

1. If $p < 0.05$, then we reject the null hypothesis (Relationship between the variables)
2. Therefore accept alternative hypothesis

Statistical investigation: ANOVA

ANOVA test also;

1. Tests the difference in mean values of continuous variable across each categorical variable
2. Null hypothesis: No difference in means across categories

Statistical investigation: Chisquare test

1. Chisquare test (X2)
2. Generates contingency table of the categorical variables
3. Generates a p-value
4. Null Hypothesis: No relationship/no correlation between the variables

In-class assignment

Using the Cassava dataset on Moodle:

1. What is the relationship between two continuous variables?
2. What is the relationship between two categorical variables?
3. Relationship between one categorical and one continuous variable?

Submit your assignment as a jupyter notebook by 27th Sept 2024.

Correlation is NOT covariation

Basis for comparison	Covariance	Correlation
Definition	Covariance is an indicator of the extent to which 2 random variables are dependent on each other. A higher number denotes higher dependency.	Correlation is a statistical measure that indicates how strongly two variables are related.
Values	The value of covariance lies in the range of -∞ and +∞.	Correlation is limited to values between the range -1 and +1
Change in scale	Affects covariance	Does not affect the correlation
Unit-free measure	No	Yes

Uganda Christian University
P.O. Box 4 Mukono, Uganda
Tel: 256-312-350800
https://ucu.ac.ug/ Email: info@ucu.ac.ug.
@ugandachristianuniversity @UCUniversity
@UgandaChristianUniversity

Department of Computing & Technology
FACULTY OF ENGINEERING, DESIGN AND TECHNOLOGY
Tel: +256 (0) 312 350 863 | WhatsApp: +256 (0) 708 114 300
@ucucomputeng @ucu_ComputEng
https://cse.ucu.ac.ug/ Email: dct-info@ucu.ac.ug