


Awarding
Great British
Qualifications



Introduction to Data Science and Big Data


Topic 3: Lecture 1

Understanding Data & Exploration

Understanding Data & Exploration Topic 3 - 3.2

Unit Syllabus

- Data Science and Big Data Fundamentals
- Introduction to Data
- **Understanding Data & Exploration**
- Data Pre-Processing
- Data Processing
- Model Selection and Evaluation
- Data Visualization
- Business Intelligence and Tools
- Data Science Ethical and Privacy Issues
- Unit Summary




Understanding Data & Exploration Topic 3 - 3.3

Scope and Coverage

This topic will cover:

- What is Descriptive Statistic?
- What is Inferential Statistic?
- What is EDA? Numerical or Graphical methods
- Aims of EDA
- Exploratory vs Confirmatory Data Analysis
- Numerical Methods of EDA: Central Tendency, Measurement of Variability
- Graphical Methods of EDA: histogram, box plot etc.




Understanding Data & Exploration Topic 3 - 3.4

Learning Outcomes

By the end of this topic students will be able to:

- Demonstrate a systematic awareness of the theoretical foundations of data processing.
- Understand and apply descriptive statistics to summarise and describe data.
- Utilise numerical and graphical methods in Exploratory Data Analysis (EDA) to gain insights from data.
- Differentiate between exploratory and confirmatory data analysis approaches and recognize their objectives.



Understanding Data & Exploration Topic 3 - 3.5

Review Quiz

1. How big is considered "Big Data"?
 - a. Data that exceeds the storage capacity of a personal computer
 - b. Data that is too large to fit into an Excel spreadsheet
 - c. Data that exceeds the processing capabilities of traditional databases **Correct Answer**
 - d. Any dataset larger than 1 GB



Understanding Data & Exploration Topic 3 - 3.6

Review Quiz

2. Which of the following is an example of unstructured data?
 - a. Excel spreadsheet
 - b. Customer reviews on social media **Correct Answer**
 - c. SQL database
 - d. Stock market prices



Understanding Data & Exploration Topic 3 - 3.7

Review Quiz

3. In the context of Big Data, what does the term "Velocity" refer to?
 - a. The speed at which data is generated and processed **Correct Answer**
 - b. The size of the dataset
 - c. The variety of data types
 - d. The accuracy of data analysis algorithms



Understanding Data & Exploration Topic 3 - 3.8

Review Quiz

4. What type of data is binary data?
 - a. Categorical Data **Correct Answer**
 - b. Text Data
 - c. Semi-Structured Data
 - d. Time Series Data

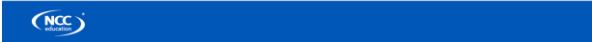


Understanding Data & Exploration Topic 3 - 3.9

Review Topic 2

- Introduction to Data:
- Big data refers to any large and complex collection of data.
 - Normal architectures fail to work with a such a large volume of data.
 - There are different types of data types such as Numerical, Categorical, Textual, Temporal, Multimedia, Derived.
 - Understanding the data type is crucial for selecting the right analytical techniques, interpreting results accurately, and effectively deriving insights from the data.

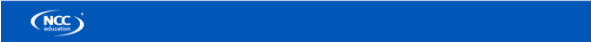
Today, we will go through the next step of data science process and understand **data and different data exploration techniques.**



Understanding Data & Exploration Topic 3 - 3.10

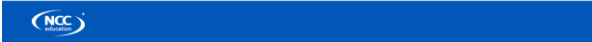
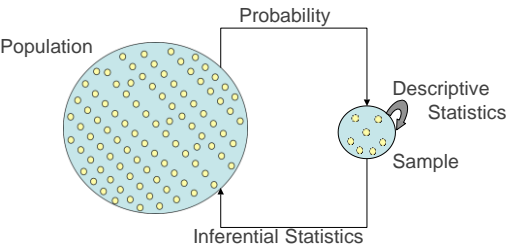
What is Statistics?

Statistics is the area of applied maths that deals with the collection, organization, analysis, interpretation, and presentation of data.



Understanding Data & Exploration Topic 3 - 3.11

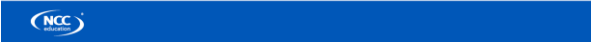
“Central Dogma” of Statistics



Understanding Data & Exploration Topic 3 - 3.12

Population

- The **entire group from which you wish to draw data.**
- In day-to-day life, the word is often used to describe groups of people (such as the population of a country).
- In statistics, it can apply to **any group from which you will collect information.**
 - Can be people, cities of the world, animals, objects, plants, colours, and so on.



Understanding Data & Exploration Topic 3 - 3.13

Sample

- **Random sampling** from **representative groups** allows us to draw broad conclusions about an overall population.
- Commonly used in opinion polling. E.g:
 - Pollsters ask a **representative small group** of people about their views on certain topics.
 - Can use this information to make informed judgments about what the larger population thinks.
 - Saves time and expense of extracting data from an entire population.



Understanding Data & Exploration Topic 3 - 3.14

Dimensionality of Data Sets

- **Univariate:** Measurement made on **one** variable per subject.
- **Bivariate:** Measurement made on **two** variables per subject.
- **Multivariate:** Measurement made on **many** variables per subject.



Understanding Data & Exploration Topic 3 - 3.15

Analysing and Summarising Data

Different approaches used in analysing and summarising data.

- **Descriptive Statistics**
- **Inferential Statistics**
- **Confirmatory Data Analysis (CDA)**
- **Exploratory Data Analysis (EDA)**



Understanding Data & Exploration Topic 3 - 3.16

Descriptive Statistics

- Descriptive statistics involve the **analysis and summary of data** to provide meaningful descriptions of its main features.
- It aims to describe the data set in a concise and understandable manner.

In a nutshell, **descriptive statistics focus on describing the visible characteristics of a dataset.**



Understanding Data & Exploration Topic 3 - 3.17

Descriptive Statistics

Descriptive statistics techniques include:

- **Measures of central tendency** (mean, median, mode) that provide information about the typical value in the data.
- **Measures of dispersion** (variance, standard deviation, range) that indicate the spread or variability of the data.
- **Measures of shape** (skewness, kurtosis) that describe the distribution's characteristics.
- Graphical techniques such as histograms, box plots, and scatter plots are used to visually represent the data.



Understanding Data & Exploration Topic 3 - 3.18

Inferential Statistics

- Inferential statistics involves making inferences and generalizations about a population based on a sample of data.
- It aims to draw conclusions or make predictions about a larger group or population using a smaller representative sample.

Inferential statistics focus on **making predictions or generalizations about a larger dataset, based on a sample of those data.**



Understanding Data & Exploration Topic 3 - 3.19

Inferential Statistics

- Use techniques such as:
 - **Hypothesis testing**
 - **Confidence intervals**
 - **Regression analysis**
 to analyse the data and make statistical inferences.
- Help **determine** the likelihood of observed differences or relationships being due to **chance or if they represent true population characteristics.**
- Account for uncertainty and provide **estimates of the reliability or confidence** of the conclusions drawn from the sample.



Understanding Data & Exploration Topic 3 - 3.20

Confirmatory Data Analysis (CDA)

- A **specific application of inferential statistics** that focuses on confirming or rejecting pre-specified hypotheses,
- Involves testing predetermined hypotheses or models.
- Inferential statistics more broadly encompasses techniques used to make inferences and draw conclusions from sample data.
- Inferential statistics can also be used in exploratory analyses to generate new hypotheses or discover patterns in the data.



Exploratory Data Analysis (EDA)

- Exploratory Data Analysis is an approach for analyzing and visualizing data to gain insights, identify patterns, and generate hypotheses.
- EDA is focused on understanding the data itself and exploring its main characteristics, rather than making formal statistical inferences.



Exploratory Data Analysis (EDA)

- Involves techniques such as:
 - **Data visualization** (e.g., histograms, scatter plots, box plots).
 - **Summary statistics** (e.g., mean, median, standard deviation), and identifying relationships or correlations between variables.
- Often performed at the initial stages of data analysis to discover interesting patterns, detect outliers, and generate hypotheses for further investigation.
- Helps understanding the structure and nature of the data and guides subsequent analysis and modeling decisions.



Goals of EDA

- **Data Understanding:**
 - Helps in getting familiar with the dataset, understanding its structure, and identifying any potential issues or limitations.
 - Involves examining the variables, their types, and the overall data organisation.
- **Data Quality Assessment:**
 - Helps in assessing the quality of the data, identifying missing values, outliers, or inconsistencies.
 - Involves checking for data completeness, accuracy, and reliability.



Goals of EDA

- **Summary Statistics:** EDA involves calculating and visualizing descriptive statistics such as measures of central tendency (mean, median) and measures of dispersion (variance, standard deviation) to understand the distribution and spread of the variables.
- **Visualization:** EDA utilises various graphical techniques to visually represent the data. This includes histograms, box plots, scatter plots, line plots, and heatmaps, among others. Visualizations help in identifying patterns, trends, and relationships between variables.



Understanding Data & Exploration Topic 3 - 3.25

Goals of EDA

- **Data Exploration:** EDA encourages the exploration of relationships between variables to uncover potential correlations or dependencies. This involves computing and visualizing cross-tabulations, correlations, and other statistical measures to reveal insights.
- **Hypothesis Generation:** EDA aids in formulating hypotheses and research questions based on the observed patterns and relationships in the data. These hypotheses can guide further analysis and modeling.



Understanding Data & Exploration Topic 3 - 3.26

Comparison Between Approaches

- **Descriptive statistics** summarise the main features of the data.
 - **Inferential statistics** make inferences about populations based on sample data.
 - **CDA:** CDA is a type of inferential statistics that specifically involves testing predetermined hypotheses or models.
 - **EDA:** EDA is a broader approach to data analysis that involves exploring the data, identifying patterns, and generating hypotheses.
- These approaches serve different purposes but are often used together in the data analysis process.

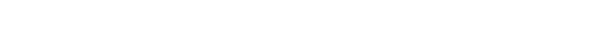


Understanding Data & Exploration Topic 3 - 3.27

Class Activity 1

Students should form small groups (3-4 students per group).

Identify some practical example of descriptive statistics, inferential statistics, and exploratory data analysis in various fields such as business, healthcare, and engineering.



Understanding Data & Exploration Topic 3 - 3.28

Class Activity 1: Real-life examples

- Business**
- **Descriptive Statistics:** A company wants to understand the **average sales and variability of its products** across different regions. They calculate the **mean, median, and standard deviation** of sales data for each region to gain insights.
 - **Inferential Statistics:** A business is interested in determining **if there is a significant difference** in customer satisfaction levels between two versions of a product. They conduct a t-test to compare the means of satisfaction ratings from two groups of customers who used different product versions.



Understanding Data & Exploration Topic 3 - 3.29

Class Activity 1: Real-life examples

Business

- **Exploratory Data Analysis (EDA):** A retail company explores customer purchase behavior to identify customer segments. They analyse customer demographics, purchase history, and use clustering techniques such as k-means or hierarchical clustering to identify distinct groups of customers with similar buying patterns.



Understanding Data & Exploration Topic 3 - 3.30

Class Activity 1: Real-life examples

Healthcare

- **Descriptive Statistics:** A hospital analyses patient wait times in the emergency department and calculates median and interquartile range to understand the waiting period and the spread of wait times.
- **Inferential Statistics:** A pharmaceutical company conducts a clinical trial to evaluate the effectiveness of a new drug. They use hypothesis testing to determine if the drug has a significant impact on patient recovery rates compared to a control group.
- **Exploratory Data Analysis (EDA):** Epidemiologists examine a dataset on disease outbreak patterns. They create spatiotemporal maps, box plots, and time series plots to identify clusters of cases, trends over time, and potential risk factors associated with the disease outbreak.



Understanding Data & Exploration Topic 3 - 3.31

Class Activity 1: Real-life examples

Engineering

- **Descriptive Statistics:** An engineering team measures the tensile strength of a material and calculates the mean and standard deviation to summarize the material's strength characteristics.
- **Inferential Statistics:** An automotive company wants to assess whether a new manufacturing process improves fuel efficiency. They conduct a hypothesis test to compare the mean fuel efficiency of vehicles produced before and after implementing the new process.
- **Exploratory Data Analysis (EDA):** Civil engineers analyse sensor data from a bridge to identify potential structural anomalies. They use data visualization techniques such as line plots, scatter plots, and spectral analysis to uncover patterns, correlations, or unusual behavior in the sensor data.




Understanding Data & Exploration Topic 3 - 3.32


Checkpoint Summary

- Statistics is the base of data processing.
- Four different approaches used in analysing and summarising data namely:
 - ✓ Descriptive statistics
 - ✓ Inferential statistics
 - ✓ CDA
 - ✓ EDA





Awarding
Great British
Qualifications



Introduction to Data Science and Big Data

Topic 3: Lecture 2


Methods of EDA

Understanding Data & Exploration Topic 3 - 3.34

Methods of EDA

EDA is usually performed:


- **Numerical Methods:** Central Tendency, Measurement of Variability
- **Graphical Methods:** Histogram, Box Plot, etc. More will be covered in later Visualization section.



Understanding Data & Exploration Topic 3 - 3.35

Numerical Methods of EDA

- **Central Tendency measures.** They are computed to give a “center” around which the measurements in the data are distributed.
- **Variation or Variability measures.** They describe “data spread” or how far away the measurements are from the center.




Understanding Data & Exploration Topic 3 - 3.36

Measures of Central Tendency

A measure of central tendency is a descriptive statistic that describes the average, or typical value of a set of scores

There are three common measures of central tendency:

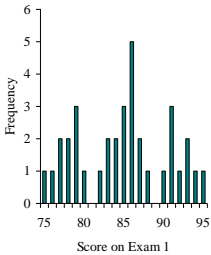
- the mode
- the median
- the mean



Understanding Data & Exploration Topic 3 - 3.37

Central Tendency: Mode

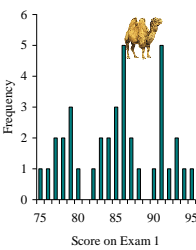
The mode is the score that occurs **most frequently** in a set of data.



Understanding Data & Exploration Topic 3 - 3.38

Bimodal Distribution

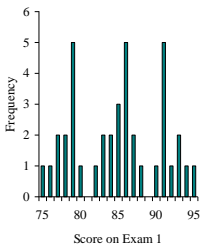
When a distribution has two “modes,” it is called **bimodal**.



Understanding Data & Exploration Topic 3 - 3.39

Multimodal Distribution

If a distribution has more than 2 “modes,” it is called **multimodal**.

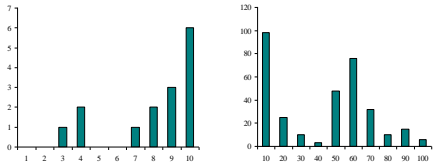


Understanding Data & Exploration Topic 3 - 3.40

When To Use the Mode

The mode is **not a very useful** measure of central tendency.

- It is insensitive to large changes in the data set.
- That is, two data sets that are very different from each other can have the same mode.



Understanding Data & Exploration Topic 3 - 3.41

When To Use the Mode

The mode is primarily used with **nominally** scaled data.

- It is the only measure of central tendency that is appropriate for nominally scaled data.



Understanding Data & Exploration Topic 3 - 3.42

Central Tendency: Mean

To calculate the average \bar{x} of a set of observations, add their value and divide by the number of observations:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$



Understanding Data & Exploration Topic 3 - 3.43

Calculating the Mean

Calculate the mean of the following data:
1 5 4 3 2
Sum the scores ($\sum x$):
 $1 + 5 + 4 + 3 + 2 = 15$
Divide the sum ($\sum x = 15$) by the number of scores ($N = 5$):
 $15 / 5 = 3$

Mean = $\bar{X} = 3$



Understanding Data & Exploration Topic 3 - 3.44

When To Use the Mean

You should use the mean when:

- the data are interval or ratio scaled
 - Many people will use the mean with ordinally scaled data too
- and the data are not skewed

The mean is preferred because it is sensitive to every score

- If you change one score in the data set, the mean will change



Central Tendency: Median

Median – the exact middle value

Calculation:

- If there are an odd number of observations, find the middle value
- If there are an even number of observations, find the middle two values and average them

Example

Age of participants: 17 19 21 22 23 23 23 38

Median = $(22+23)/2 = 22.5$



How To Calculate Median

Conceptually, it is easy to calculate the median

- There are many minor problems that can occur; it is best to let a computer do it.

1. Sort the data from highest to lowest

2. Find the score in the middle

$middle = (N + 1) / 2$

If N, the number of scores, is even the median is the average of the middle two scores



Median Example – Class Activity

What is the median of the following scores:

10 8 14 15 7 3 3 8 12 10 9

Sort the scores:

15 14 12 10 10 9 8 8 7 3 3

Determine the middle score:

$middle = (N + 1) / 2 = (11 + 1) / 2 = 6$

Middle score = median = 9



Median Example – Class Activity

What is the median of the following scores:

24 18 19 42 16 12

Sort the scores:

42 24 19 18 16 12

Determine the middle score:

$middle = (N + 1) / 2 = (6 + 1) / 2 = 3.5$

Median = average of 3rd and 4th scores:

$(19 + 18) / 2 = 18.5$



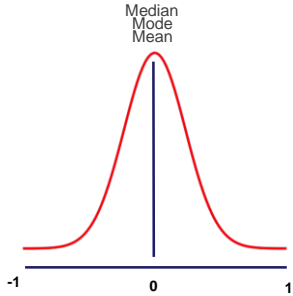
When To Use the Median

The median is often used when the distribution of scores is either positively or negatively skewed.

- The few really large scores (positively skewed) or really small scores (negatively skewed) will not overly influence the median.

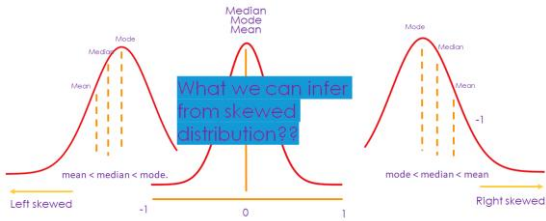
Relations Between the Measures of Central

Normal Distribution



Relations Between the Measures of Central

Skewness



Central Tendency: Mean, Median, Mode

Which one to look at?

	Mean	Median	Mode
Data	Continuous	Ordinal	Nominal
Example	Temperature Reaction time Age	Income bracket	Voting Ethnicity Viewership

Understanding Data & Exploration Topic 3 - 3.53

Measure of Variability: Range

Measure of Variability: How data points are distributed from its mean.

Range

Difference between maximum and minimum value



Insensitive to distribution



Understanding Data & Exploration Topic 3 - 3.54

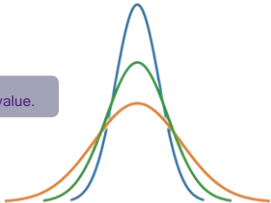
Measure of Variability: Variance

Variance

How far numbers are set apart

Average of squared differences from mean value.

Building block of many statistical tests



Understanding Data & Exploration Topic 3 - 3.55

Measure of Variability: Standard Deviation

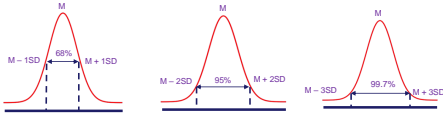
Standard Deviation

Where most of the data lie within distribution

Square root of variance

Works on uniform distribution

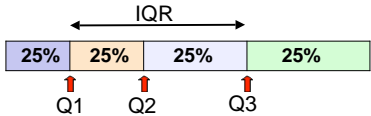
Empirical rule: 68, 95, 99.7%



Understanding Data & Exploration Topic 3 - 3.56

Measure of Variability: Interquartile Range (IQR)

- The first quartile, Q1, is the value for which 25% of the observations are smaller and 75% are larger
- Q2 is the same as the median (50% are smaller, 50% are larger).
- Only 25% of the observations are greater than the third quartile.



Understanding Data & Exploration Topic 3 - 3.57

Graphical Method of EDA

A (Good) Picture Is Worth A 1,000 Words



Understanding Data & Exploration Topic 3 - 3.58

Univariate Data:
Histograms and Bar Plots

What's the difference between a histogram and bar plot?

Bar plot

- Used for categorical variables to show frequency or proportion in each category.
- Translate the data from frequency tables into a pictorial representation.

Histogram

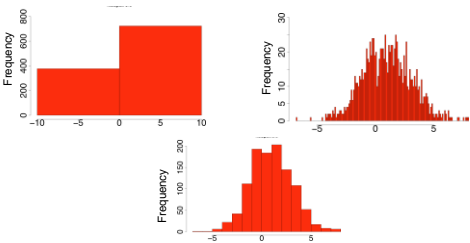
- Used to visualize distribution (shape, center, range, variation) of continuous variables.
- "Bin size" important!



Understanding Data & Exploration Topic 3 - 3.59

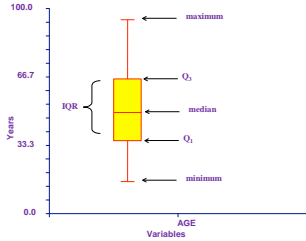
Effect of Bin Size on Histogram

Simulated 1000 $N(0,1)$ and 500 $N(1,1)$



Understanding Data & Exploration Topic 3 - 3.60

Box Plots



Bivariate Data

Variable 1	Variable 2	Display
Categorical	Categorical	Crosstabs Stacked Box Plot
Categorical	Continuous	Boxplot
Continuous	Continuous	Scatterplot Stacked Box Plot

Multivariate Data

Clustering

- Organize units into clusters
- Descriptive, not inferential
- Many approaches
- “Clusters” always produced

Data Reduction Approaches (PCA)

- Reduce n-dimensional dataset into much smaller number
- Finds a new (smaller) set of variables that retains most of the information in the total sample
- Effective way to visualize multivariate data

How to Make a Bad Graph

The aim of good data graphics:

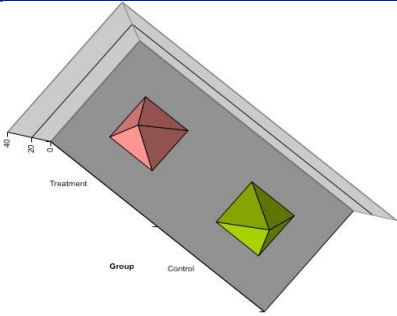
- Display data accurately and clearly

Some rules for displaying data badly:

- Display as little information as possible
- Obscure what you do show (with chart junk)
- Use pseudo-3d and color gratuitously
- Make a pie chart (preferably in color and 3d)
- Use a poorly chosen scale

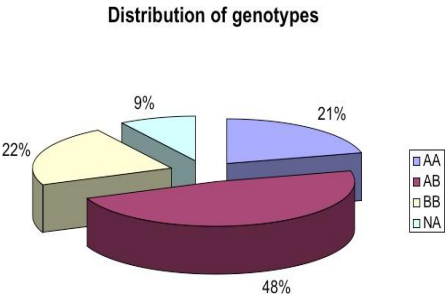
From Karl Broman: <http://www.biostat.wisc.edu/~kbroman/>

Example 1



Understanding Data & Exploration Topic 3 - 3.65

Example 2



Understanding Data & Exploration Topic 3 - 3.66

Checkpoint Summary

- EDA is usually performed:
- Numerical Methods: Central Tendency, Measurement of Variability
 - Graphical Methods: Histogram, Box Plot, etc.
 - Central tendency measures includes Mode, Median, Mean
 - Measurement of Variability includes Range, Variance, Standard Deviation, IQR

Understanding Data & Exploration Topic 3 - 3.67

Discussion Session

Let's say we have a dataset of exam scores for a class of students: [70, 75, 80, 85, 90, 95, 100].

How does the choice of measure of central tendency (mean, median, mode) and measure of variability (range, variance, standard deviation) impact our understanding of a dataset?

Provide examples to support your explanation

Understanding Data & Exploration Topic 3 - 3.68

Discussion Session

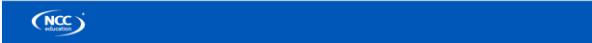
Let's say we have a dataset of exam scores for a class of students: [70, 75, 80, 85, 90, 95, 100].

- **Mean:** The mean is calculated by summing up all the scores and dividing by the total number of scores. For this dataset, the mean would be $(70 + 75 + 80 + 85 + 90 + 95 + 100) / 7 = 85.71$. The mean represents the average score of the class.
- **Median:** To find the median, we arrange the scores in ascending order and identify the middle value. In this case, the median is 85. Since there is an odd number of scores, there is a single middle value. The median gives us the score that separates the higher and lower scores equally.
- **Mode:** The mode is the value that occurs most frequently. In this dataset, there is no mode as all the scores appear only once.

Understanding Data & Exploration Topic 3 - 3.69

Discussion Session

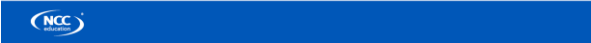
- Moving on to measures of variability:
- Range: The range is determined by finding the difference between the highest and lowest scores. In this example, the range would be $100 - 70 = 30$. The range indicates the span of scores from the lowest to the highest, providing a basic understanding of the spread.
 - Variance and Standard Deviation (SD): Variance and SD measure the dispersion or spread of the scores around the mean. Calculating these values requires more detailed calculations, but they give us a quantitative measure of how the scores deviate from the average.
- We might find that the variance is 87.76 and the SD is approximately 9.37. These values indicate the degree of spread or dispersion of scores from the mean, giving us insights into the variability of performance within the class.



Understanding Data & Exploration Topic 3 - 3.70

Discussion Session

- In summary,
- Choice of measures such as mean, median, mode, range, variance, and standard deviation provide different perspectives on the dataset.
 - By using these measures, we can understand the central tendency, identify common patterns, and quantify the variability within the data, helping us gain a deeper understanding of the exam scores and the performance of the class.



Understanding Data & Exploration Topic 3 - 3.71

Topic Summary

- Statistics is the base of data processing.
 - Four different approaches used in analyzing and summarizing data namely
 - Descriptive statistics
 - Inferential statistics
 - CDA
 - EDA
- EDA is usually performed:**
- Numerical Methods: Central Tendency, Measurement of Variability
 - Graphical Methods: Histogram, Box Plot, etc.



Understanding Data & Exploration Topic 3 - 3.72

Next Topic 4: Data Pre-Processing

- What is Data Pre-processing?
- Data Pre-processing Importance?
- Data Pre-processing Steps
- Data Pre-processing examples?
- Data Pre-processing Methods
- Missing Values
- Categorical Encoding
- Feature engineering



References

- Grus, J. (2019). Data Science from Scratch. O'Reilly Media.
- Holmes, D. E. (2013). Big Data: A Very Short Introduction. Oxford University Press.
- Kotu, V., & Deshpande, B. (2014). Predictive analytics and data mining: concepts and practice with RapidMiner. Morgan Kaufmann.
- Marz, N., & Warren, J. (2015). Big Data: Principles and best practices of scalable real-time data systems. Manning Publications.
- Provost, F., & Fawcett, T. (2013). Data Science for Business. O'Reilly Media.



Topic 3 – Understand Data and Exploration

Any Questions?