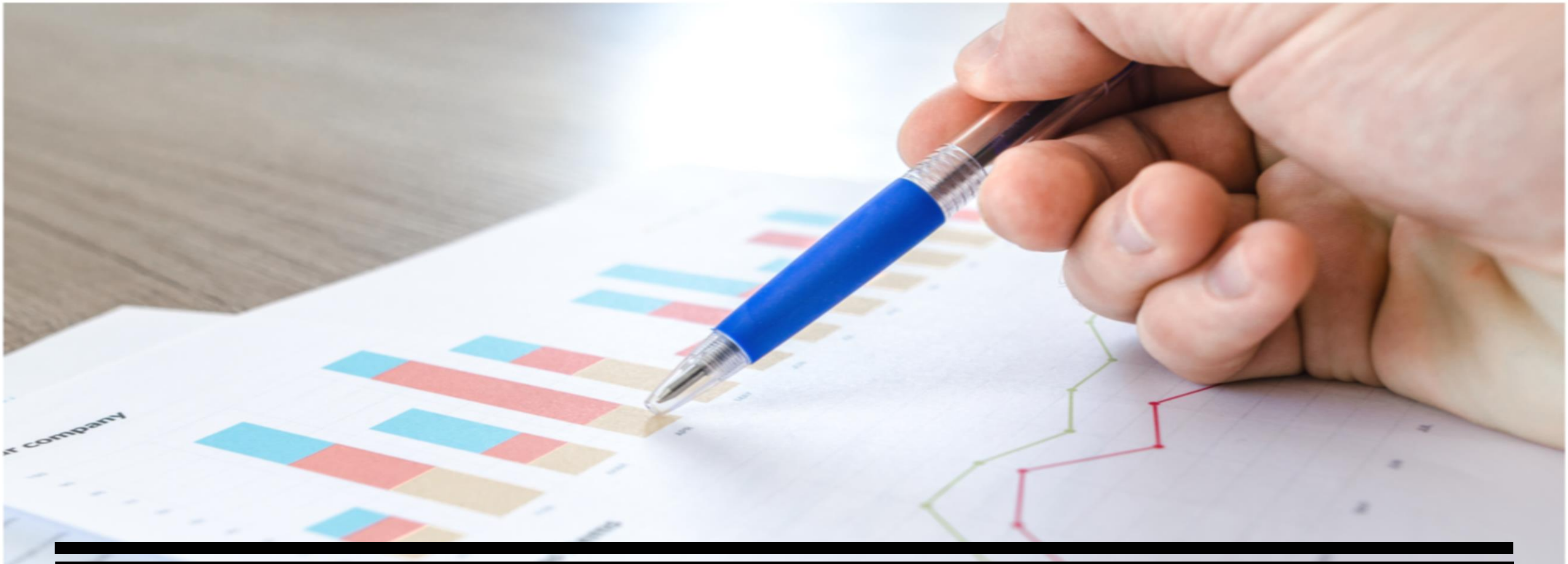

DATA PREPROCESSING AND HYPOTHESIS TESTING

BY ANGELA OGADA

DATA PREPROCESSING

DATA PREPROCESSING

What is Data Preprocessing?

Steps involved in preparation of data for another process, usually data analysis

DATA PREPROCESSING

Importance

- Improve accuracy
- Quality
- Consistency

DATA PREPROCESSING

Concepts

- Data acquisition
- Import Libraries
- Import Data
- Cleaning data
- Encoding
- Feature Selection
- Feature Scaling

Data acquisition

Get data from a source

- Online
- Collect Data
- Client
- Web Scraping

Importing Libraries

- Library – Reusable code that helps optimize tasks
- Choice of libraries to import depends on programming language and tasks to be performed
- R Programming – ggplot2, tidyr
- Python – pandas, numpy

Importing Libraries

```
import pandas as pd
```

Importing Libraries

Pandas

- An open source python library used for data manipulation and analysis

- Pandas represents data in 2 forms;
  1. Series – One dimensional
  2. Data Frame – Two Dimensional (Table-like with rows and columns)

Loading Data

Data can be in different file formats

- json - pd.read_json()
- excel – pd.read_excel()
- csv – pd.read_csv()

Cleaning Data

- Duplicates
- Outliers
- Data Types
- Missing Values

Cleaning Data: Duplicates

Checking for number of Duplicates

```
df.duplicated().sum()
```

Show duplicated rows

```
df[df.duplicated()]
```

Drop Duplicates

```
df.drop_duplicates()
```

Cleaning Data: Data Types

- **Object**

  Text or mixed numeric and non-numeric values

- **Int64**

  Whole numbers

- **Float64**

  Numbers with Decimals

- **Datetime64**

  Date and time values

- **Bool**

  True/False values

- **Check DataType**

  ```
  df.dtypes
  ```

  ```
  df.info()
  ```

- **Type conversion**

  ```
  df['Variable'].astype(' ')
  ```

  There are pandas inbuilt functions to convert to numeric and date

  ```
  to_numeric()
  ```

  ```
  to_datetime()
  ```

Cleaning Data: Outliers

Outliers are values that fall more than three standard deviations from the mean.

Depending on the data being worked on, they can be natural variations in the data or errors during entry or sampling.

Checking for Outliers

- BoxPlot

  Visualizing distribution of data based on a five number summary ("minimum", first quartile [Q1], median, third quartile [Q3] and "maximum")

- Histogram

  Distibution of a numerical variable

- ScatterPlot

  Relationship between two numeric variables

- Z Score

  Relationship between standard deviation of a point and mean of the group

Dropping Outliers

- Find the IQR score (Q3 –Q1) to identify the points to drop
- Drop points that fall outside the not in the range of (Q1 - 1.5 IQR) and (Q3 + 1.5 IQR)

```
df[~((df < (Q1 - 1.5 * IQR)) | (df > (Q3 + 1.5 IQR))).any(axis=1)]
```

Cleaning Data: Missing values

Drop/Fill with a value

- Checking for missing values

  df.isnull()

  df.isna()

- Drop missing observations

  df.dropna()

- Drop rows with missing values

  df.dropna(how = 'all')

- Drop columns with missing values

  df.dropna(axis = 1)

- Fill missing values with zeros or a value

  df.fillna(0)

- Fill missing values forward

  df.fillna(method = 'ffill')

- Fill missing values backward

  df.fillna(method = 'bfill')

- Fill missing values with average value

  df.fillna(df.mean())

  df['col1'].fillna(df['col1'].mean())

- Fill missing values median

  df.fillna(df.median())

  df['col1'].fillna(df['col1'].median())

- Fill missing values mode

  df['col1'].fillna(df['col1'].mode()[0])

Encoding

Categorical data encoding is conversion of categorical variables to numerical dummies.
Most useful before using machine learning algorithms

- One hot Encoding

onehotencoder() scikitlearn method

- Numerical Dummies

getdummies() scikitlearn method

Feature Selection

This is reducing the number of input variables.

It helps reduce cost, time spent and improve accuracy of a model

Feature Selection Methods

- Wrapper methods

  Forward, backward, and stepwise selection

- Filter methods

  ANOVA, Pearson correlation, Variance thresholding

- Embedded methods

  Lasso, Ridge, Decision Tree

Feature Scaling

Feature scaling is a method used to normalize the range of independent variables or features of data so they can be in the same range.

Common Feature Scaling Techniques

- Absolute Maximum Scaling
- Min-Max Scaling
- Normalization
- Standardization
- Robust Scaling

HYPOTHESIS TESTING

HYPOTHESIS TESTING

Testing Statistical Significance of the possibility of an event occurring(Null Hypothesis)

HYPOTHESIS TESTING

Importance

- Prove causation between 2 variables

Null Hypothesis

- Statement that is believed to be true or is used to put forth an argument unless it can be shown to be incorrect by Hypothesis testing.

Alternative Hypothesis

- Claim that is contradictory to the null hypothesis.

Example

Null Hypothesis

- Marital Situation does not affect Education at Elimu Univesity

Alternative Hypothesis

- Marital Situation affects Education at Elimu Univesity

Steps involved in Hypothesis testing

Define Hypothesis	Null Hypothesis Alternative Hypothesis
Test for Assumptions	Shape of distribution - Normality Level of Measurement of the variable Sampling method and Sample size
Identify Test Statistic	Value computed from sample data that to help us decide whether to either accept or reject the null hypothesis
Analyze for evidence against Null Hypothesis	P-value shows evidence against null Hypothesis
Communicate Findings	Accept or fail to accept Null Hypothesis based on significance level

Errors in Hypothesis Testing

Type I Error α

- Probability of rejecting null hypothesis when it is true.
- False Positive

Type II error β

- Probability of falling to reject the null when it is false.
- False Negative

Hypothesis Test

Assuming normality, the choice of the test depends on;

- number of variables
- type of variable

Goal/Number Of Variables	Type of Variable	
	Numerical	Categorical
Compare one group to a Hypothetical value	One sample t test	Chi square
Compare two independent groups	Unpaired t test	Fisher's test
Compare two dependent groups	Paired t test	McNemar's test
Compare three or more unmatched groups	One way ANOVA	Chi square

THANK YOU;

Angela Ogada

Email: angieogada@gmail.com

Linkedin: Angie Ogada