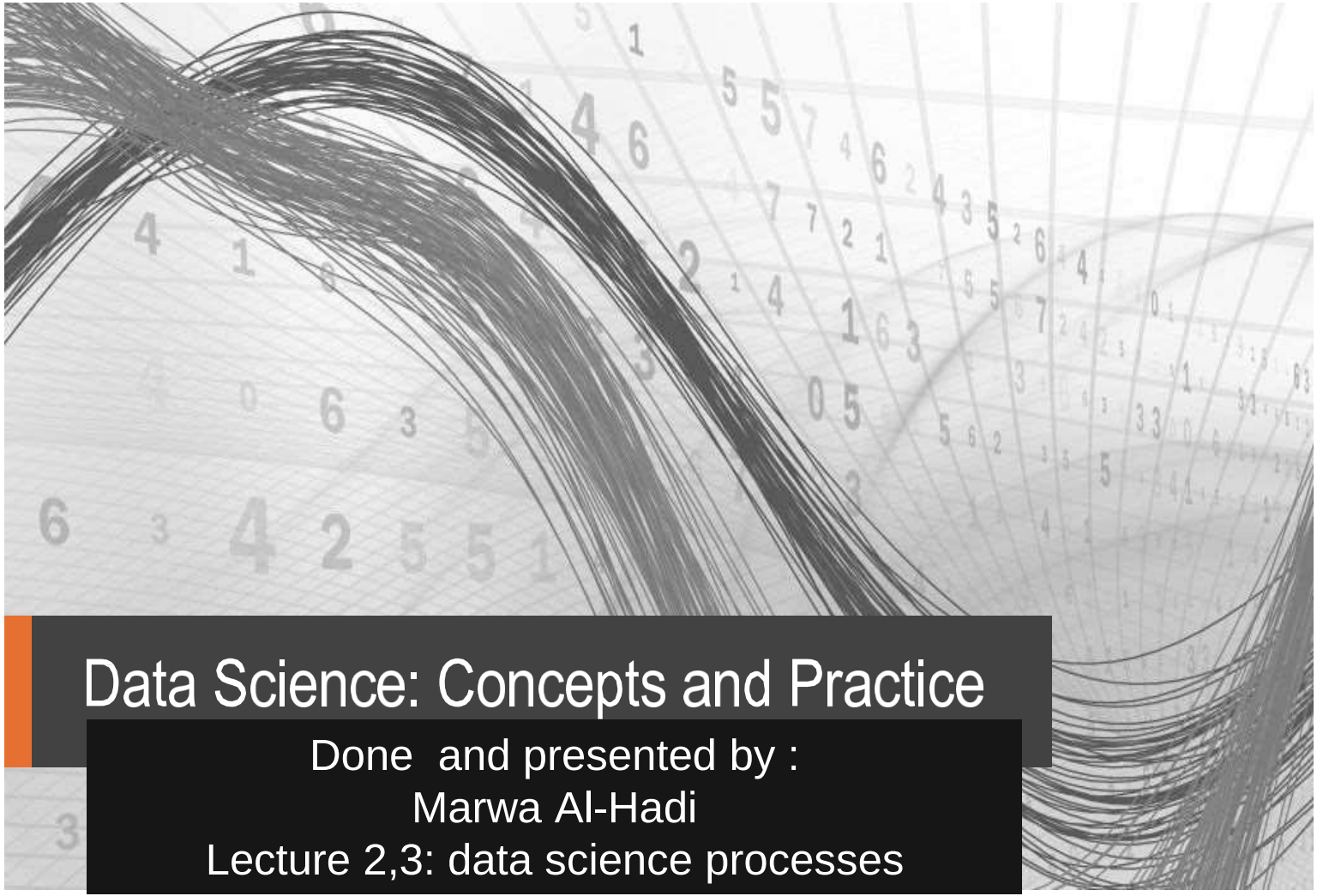


Data Science: Concepts and Practice

Done and presented by :
Marwa Al-Hadi

Lecture 2,3: data science processes

Data mining and data science

- **Data Mining**: Focuses on **extracting patterns, trends**, and useful information from large datasets **using** mostly **machine learning techniques.** It is often considered a subset of Data Science.
- **Data Science**: A broader interdisciplinary field that involves **extracting knowledge and actionable perceptions** from data. It includes data mining, but also includes data cleaning, data visualization, predictive modeling, deployment, and communication of results

Data mining and data science

2. Scope

Aspect	Data Mining	Data Science
Objective	Find hidden patterns and trends in data.	Solve complex problems using a complete data pipeline.
Scope	Focuses primarily on pattern extraction.	Covers data processing, analysis, and implementation.
End-to-End Process	Typically focuses on analysis.	Includes data collection, analysis, and model deployment.

Data mining and data science

3. Tools and Techniques

Aspect	Data Mining	Data Science
Techniques	Clustering, association rules, anomaly detection, classification, regression.	Machine learning, deep learning, data wrangling, statistical analysis.
Tools	WEKA, RapidMiner, Orange, SQL.	Python, R, TensorFlow, Spark, Hadoop, Tableau.

Data mining and data science

4. Key Differences

Area	Data Mining	Data Science
Focus	Identifying patterns and relationships in data.	Building, deploying, and refining predictive models.
Programming	Minimal coding, relies on pre-built tools.	Heavy use of programming (e.g., Python, R).
Big Data	Less emphasis on handling Big Data.	Strong focus on Big Data tools and frameworks.
Interdisciplinary Nature	Primarily statistical and algorithmic.	Combines statistics, machine learning, domain expertise, and communication.
Output	Insights, patterns, and trends.	Comprehensive solutions and actionable insights.

Data mining and data science

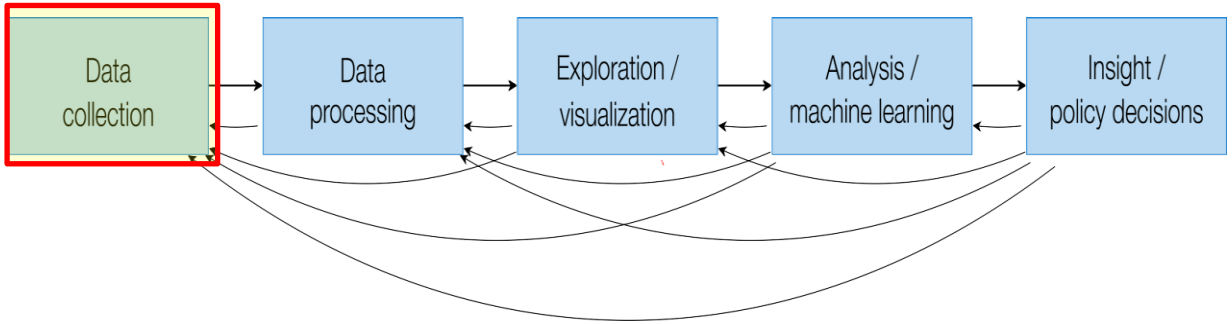
Data Mining Workflow:	**Data Science Workflow:**
1.Define the problem. 2.Collect and preprocess data. 3.Apply algorithms to find patterns (e.g., clustering, classification). 4.Evaluate and interpret results.	1.Define the business problem. 2.Collect and preprocess data (data wrangling and cleaning). 3.Conduct exploratory data analysis (EDA). 4.Develop machine learning or statistical models. 5.Validate and fine-tune models. 6.Deploy the solution (e.g., web apps, APIs). 7.Communicate insights to stakeholders.

Data mining and data science

6. Use Cases

Use Case	Data Mining	Data Science
Retail	Market basket analysis to find product associations.	Build a recommendation system.
Healthcare	Identify patterns in patient records.	Predict disease progression or outcomes.
Finance	Fraud detection using rule-based systems.	Develop predictive algorithms for stock prices.

Data Science



First :Data Collection and Understanding Data

- Types of Data Sources
- The Importance of Data Collection
- Methods of Data Collection
- Understanding Data
- Data Quality and Issues

Types of Data Sources

- **Primary Data:**
  - Data collected directly from the source.
  - Examples: Surveys, experiments, IoT sensors.
- **Secondary Data:**
  - Pre-existing data collected by others.
  - Examples: Public datasets, web data, company records.

Importance of Data Collection

- - Provides the **foundation for data analysis** and modeling.
- - Ensures **accurate and reliable decision-making**.
- - Critical for **understanding patterns, trends, and insights.**
- - Examples:
- * Customer purchase history for personalized marketing.
- * Sensor data for predictive maintenance.

Methods of Data Collection

- **Manual Methods:**
  - Surveys and interviews
  - Observations
- **Automated Methods:**
  - APIs for real-time data
  - Web scraping
  - IoT device integration

Tools and Techniques for Data Collection

- Manual Tools:
  - Google Forms
  - Excel for manual entry
- Automated Tools:
  - Python libraries (e.g., requests, BeautifulSoup)
  - IoT platforms (e.g., Arduino, Raspberry Pi)
  - Data APIs (e.g., Twitter API, OpenWeather API)

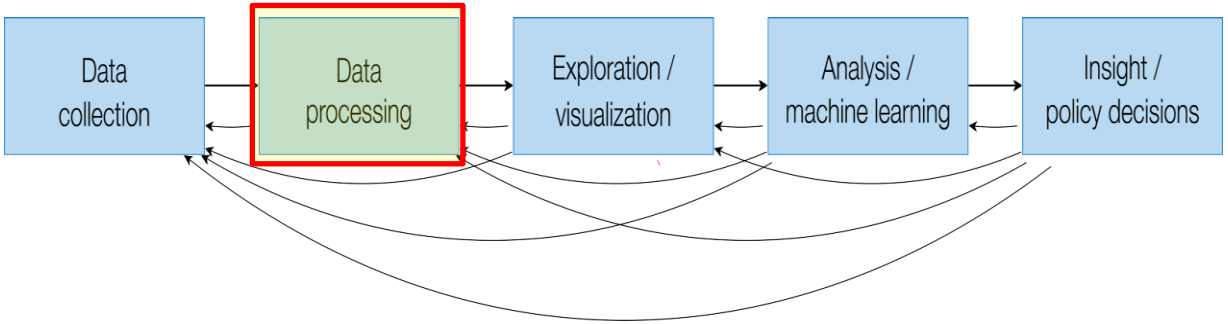
Data Understanding

- **Definition**
  - The process of exploring and summarizing data to understand its structure and patterns.
- - **Steps**:
  - Identifying data types and formats
  - Checking data quality
  - Visualizing basic statistics and distributions
  - **Tools**: Pandas, Matplotlib, Seaborn in Python

Data Quality and issues

- **Key Dimensions of Quality:**
  - Accuracy
  - Completeness
  - Consistency
  - Timeliness
- **Issues:**
  - Missing or incomplete data
  - Noisy or irrelevant data
  - Ethical and privacy concerns

Data Science



Second : Data Preprocessing and Cleaning

- Data Preprocessing Techniques
  - Data Cleaning
  - Data Transformation
  - Feature Engineering
- Handling Missing Data
- Data Normalization and Standardization

Data Preprocessing

- **Definition:**
  - The process of preparing raw data for analysis by cleaning and transformation via :
    - **Data Cleaning:** Handling missing values, correcting errors.
    - **Data Transformation:** Scaling, encoding categorical variables.
    - **Feature Engineering:** Creating new features that improve model performance.

Data Cleaning

1. Handling Missing Data

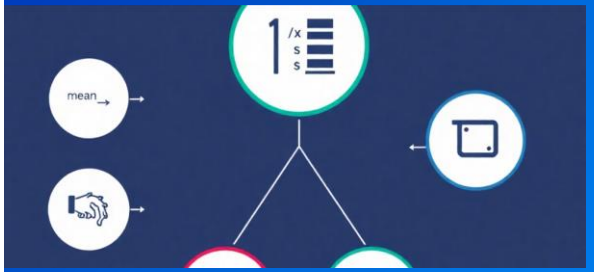
-Imputation: Filling missing values with mean, median, or mode.

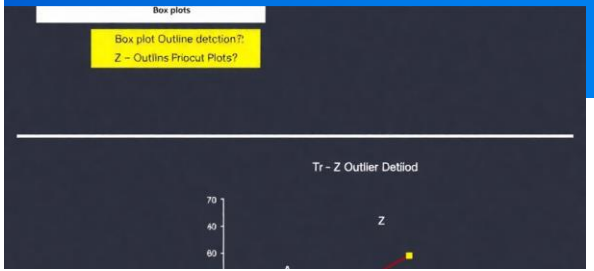
-Deletion: Removing rows with missing data (small)

2. Handling Outliers

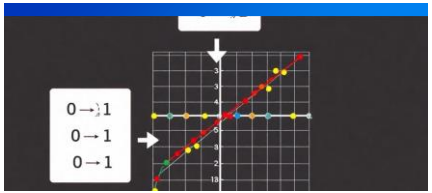
-IQR Method: Identifying and removing outliers based on interquartile range.

-Z-score: Removing data points that out of range



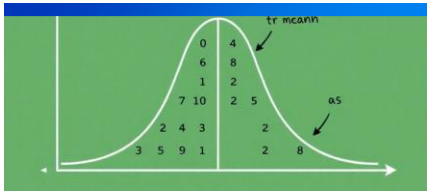


Data Transformation



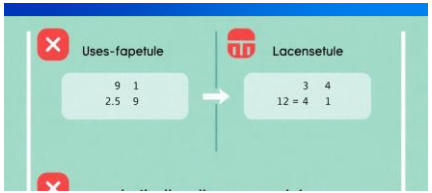
Normalization

Rescaling data to a specific range
0 ,.g.e) to 1)



Standardization

Centering data **around the mean** tinu ,naem orez)
.(ecnairav



Encoding Categorical Data

-**One-Hot Encoding** :**Converting categorical variables into binary** columns.

-**Label Encoding** gningissA :
.seirogetac ot sregetni euqinu

Feature Engineering

Definition

- Creating new features that improve model performance .
  - **Polynomial Features:** Generating higher-degree features.
  - **Binning:** Grouping continuous values into discrete intervals.
  - **Feature Selection:**-eert ,OSSAL ekil seuqinhcet gnisu serutaef tnaveler tsom eht gnitceleS .sdohtem desab

Handling Missing Data

- **Mean/Median Imputation:** Filling missing values with the **mean** ro **median**.
  - **Prediction Models:** Using a **machine learning model** to **predict missing**.seulav
  - **Multiple Imputation:** Imputing or generating missing values multiple times and averaging results.

Data Normalization and Standardization

Normalization

Useful when features have **different**.selacs ro stinu

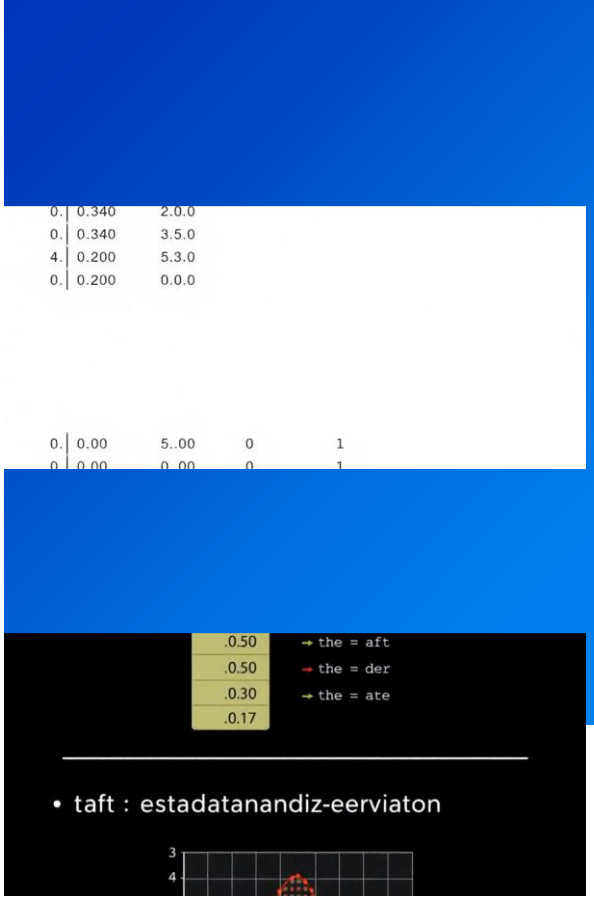
Example.:

Converting feature values into a range of[1 ,0]

Standardization

Useful for algorithms that assume the data is centered and scaled.

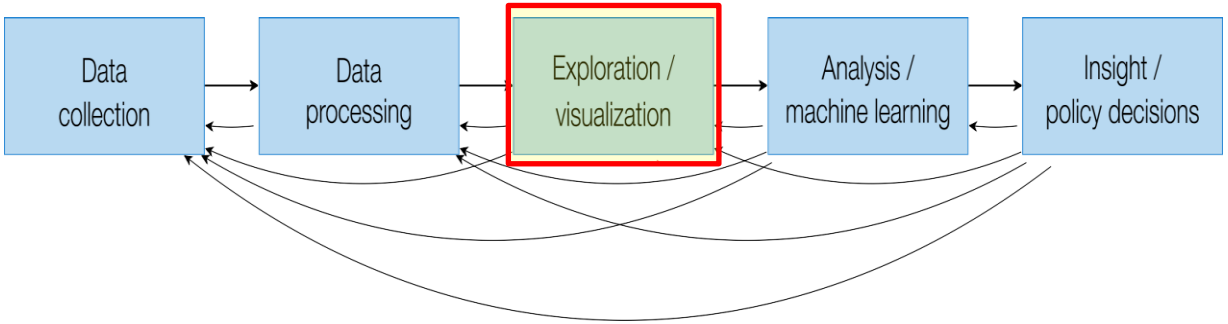
Example: Subtracting the mean and dividing by the standard deviation.



Importance of Data Preprocessing

- - Ensures data quality and consistency.
- - Helps models perform better and reduces errors.
- - Converts raw data into a usable format.

Data Science



Third: Exploring and visualization

- What is Exploratory Data Analysis (EDA)?
- Importance of EDA in Data Science
- Key Steps in EDA
- Visualizations for EDA
- Tools for EDA

What is (Exploratory Data Analysis)EDA?

- is a crucial step in the DS process
- It involves **summarizing the main characteristics of the data**, often using techniques.
- This **presentation will explore the significance** of EDA in unlocking insights and guiding further analysis.
- It focuses on visualization and statical techniques to uncover patterns , trends ,and anomalies ,concentrating on the way for **deeper analysis.**

Con..

- EDA Focuses on
  - Detecting patterns and relationships.
  - Identifying anomalies and outliers.
  - Often involves both visual and quantitative approaches.

Importance of EDA

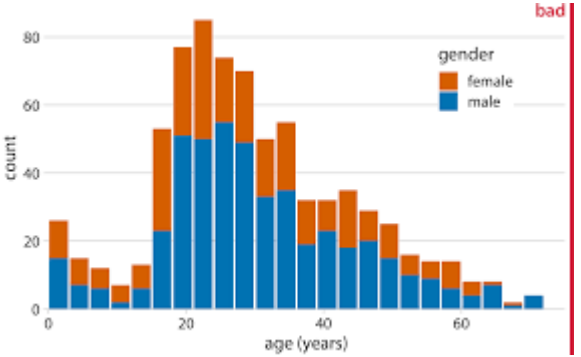
- Data Understanding: Reveals data trends and distributions.
- Problem Definition: Helps refine research questions.
- Preparation: Guides data cleaning and preprocessing.
- Decision Making: Informs feature selection and modeling strategies.

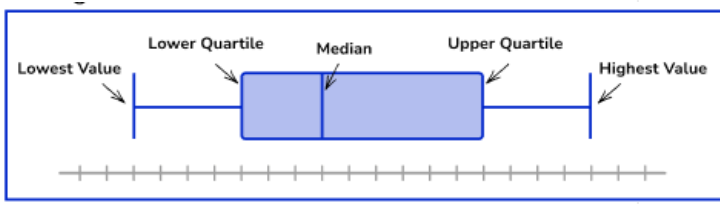
Key Steps in EDA

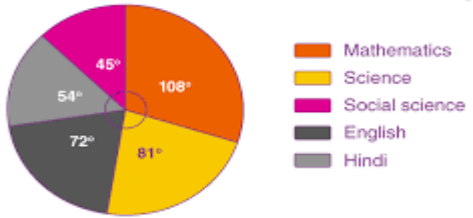
1. Data Summarization: Descriptive statistics (mean, median, standard deviation).
2. Data Visualization: Graphical representation of data.
3. Correlation Analysis: Checking relationships between variables.
4. Outlier Detection: Identifying unusual data points.
5. Missing Value Analysis: Assessing and addressing missing data.

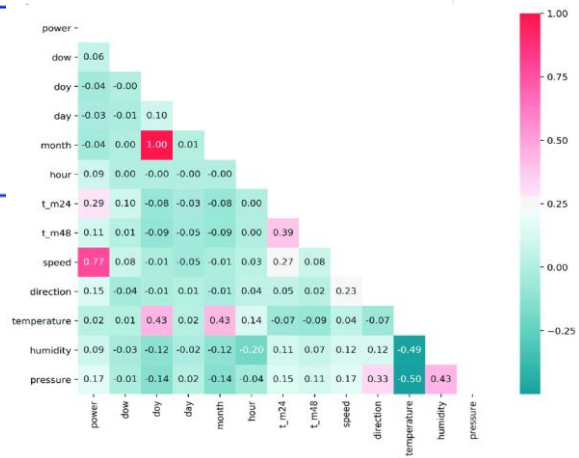
Common Visualizations for EDA

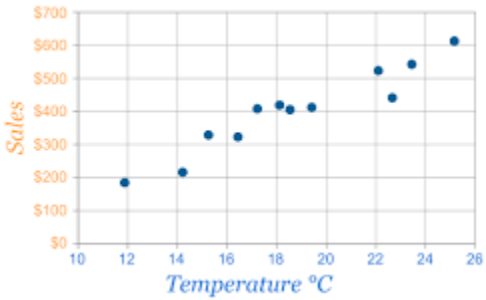
- Histograms: Distributions of numeric data.
- Scatter Plots: Relationships between two variables.
- Box Plots: Spread and outliers in data.
- Correlation Heatmaps: Relationships among features.
- Pie Charts: Proportions in categorical data.







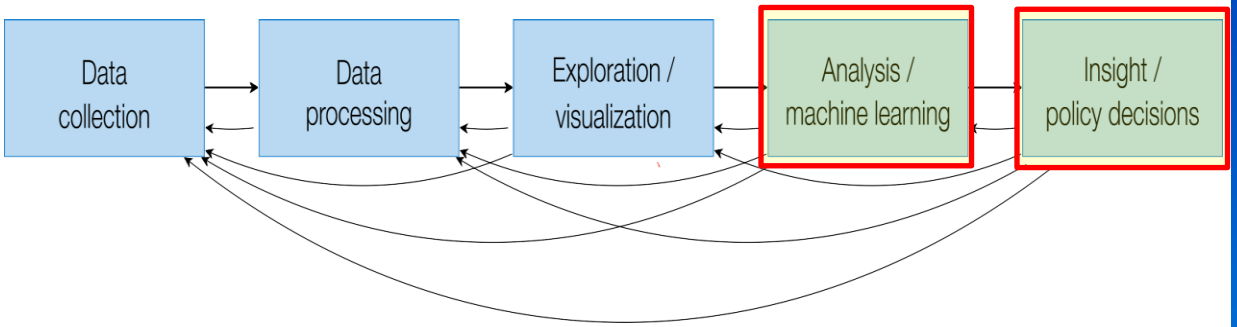




Tools for EDA

1. Python Libraries:
   - Pandas: Data manipulation and summaries.
   - Matplotlib/Seaborn: Data visualization
2. .R Programming:
   - gplot2 for complex visualizations.
3. Power BI/Tableau:
   - Interactive and dynamic dashboards.
4. Excel:
   - Simple summaries and charts

Data Science



Forth: using machine learning/deep learning

- Machine learning algorithms and deep learning algorithms.

Fifth: decision

- Figure out the result from algorithm using

Practical example

1. https://www.kaggle.com/code/shivan118/employee-attrition-analysis-visualisation
2. HR Analysis, Prediction and Visualization

Employee attrition, Analysis & Visualisation

Notebook Input Output Logs Comments (1)

1. Importing Library

In [1]:

```
### importing important library

import numpy as np ## Matrix Multiplication
import pandas as pd  ## Exploratory Data Analysis
import matplotlib.pyplot as plt   ## Visualization
import seaborn as sns   ## Visualization

%matplotlib inline
sns.set_style("whitegrid")
plt.style.use("fivethirtyeight")
```

2. Loading the Dataset

In [2]:

```
#### Loading the Dataset

pd.set_option('display.max_columns', None)
data = pd.read_csv("../input/ibm-hr-analytics-attrition-dataset/WA_Fn-UseC_-HR-Employee-Attrition.csv")
data.head(10)   ### Print the top 10 rows
```

Dataset

Out[2]:

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	Emplo
0	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	1	1
1	49	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences	1	2
2	37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other	1	4
3	33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sciences	1	5
4	27	No	Travel_Rarely	591	Research & Development	2	1	Medical	1	7
5	32	No	Travel_Frequently	1005	Research & Development	2	2	Life Sciences	1	8
6	59	No	Travel_Rarely	1324	Research & Development	3	3	Medical	1	10
7	30	No	Travel_Rarely	1358	Research & Development	24	1	Life Sciences	1	11
8	38	No	Travel_Frequently	216	Research & Development	23	3	Life Sciences	1	12
9	36	No	Travel_Rarely	1299	Research & Development	27	3	Medical	1	13

Checking the Null Values

In [7]:
```
data.isnull().sum()  ## Checking the Null Values using isnull finction
```

Out[7]:

Age	0
Attrition	0
BusinessTravel	0
DailyRate	0
Department	0
DistanceFromHome	0
Education	0
EducationField	0
EmployeeCount	0
EmployeeNumber	0
EnvironmentSatisfaction	0
Gender	0
HourlyRate	0
JobInvolvement	0
JobLevel	0
JobRole	0
JobSatisfaction	0
MaritalStatus	0
MonthlyIncome	0
MonthlyRate	0
NumCompaniesWorked	0
Over18	0
OverTime	0

```
### Pairplot using seaborn library
sns.pairplot(data1)
```

:[12]:

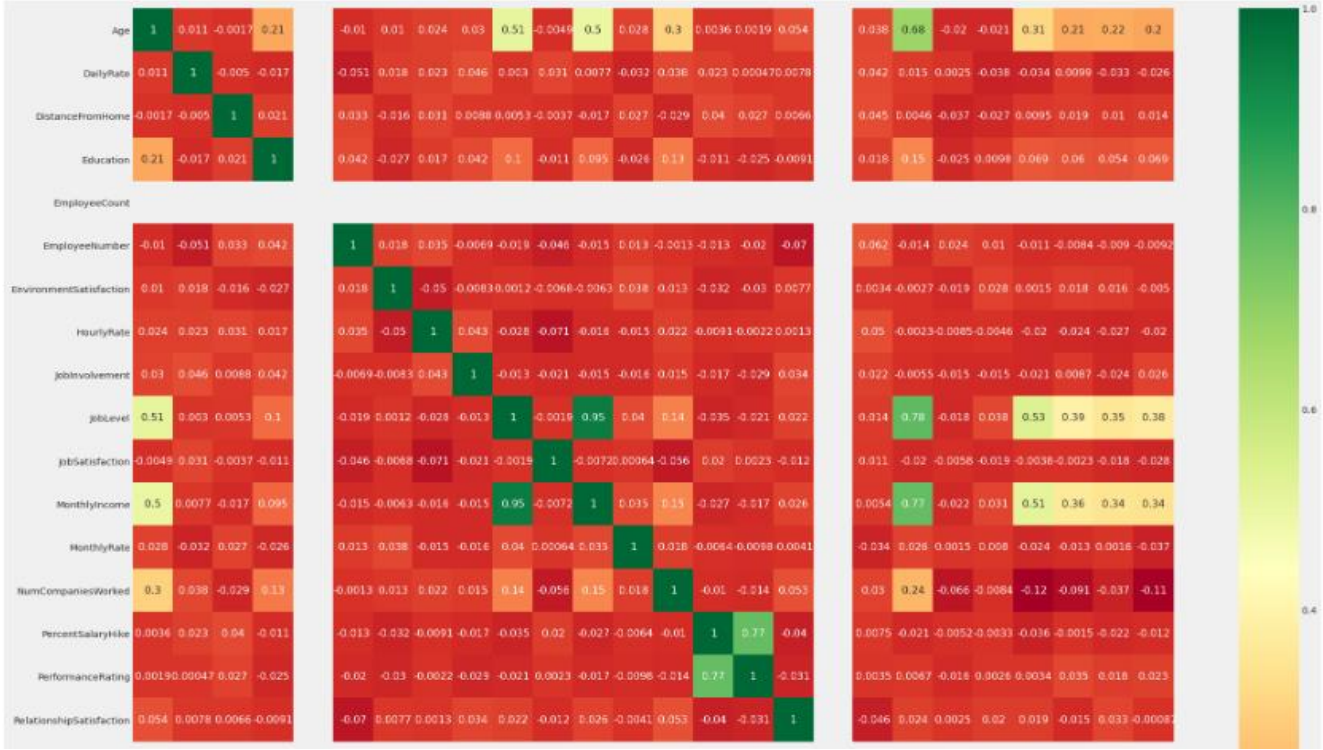
```
<seaborn.axisgrid.PairGrid at 0x7f501e340410>
```

In [13]:

```
#### Heatmap using seabron libaray
plt.figure(figsize=(30, 30))
sns.heatmap(data.corr(), annot=True, cmap="RdYlGn", annot_kws={"size":15})
```

Out[13]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f5015a2a250>
```



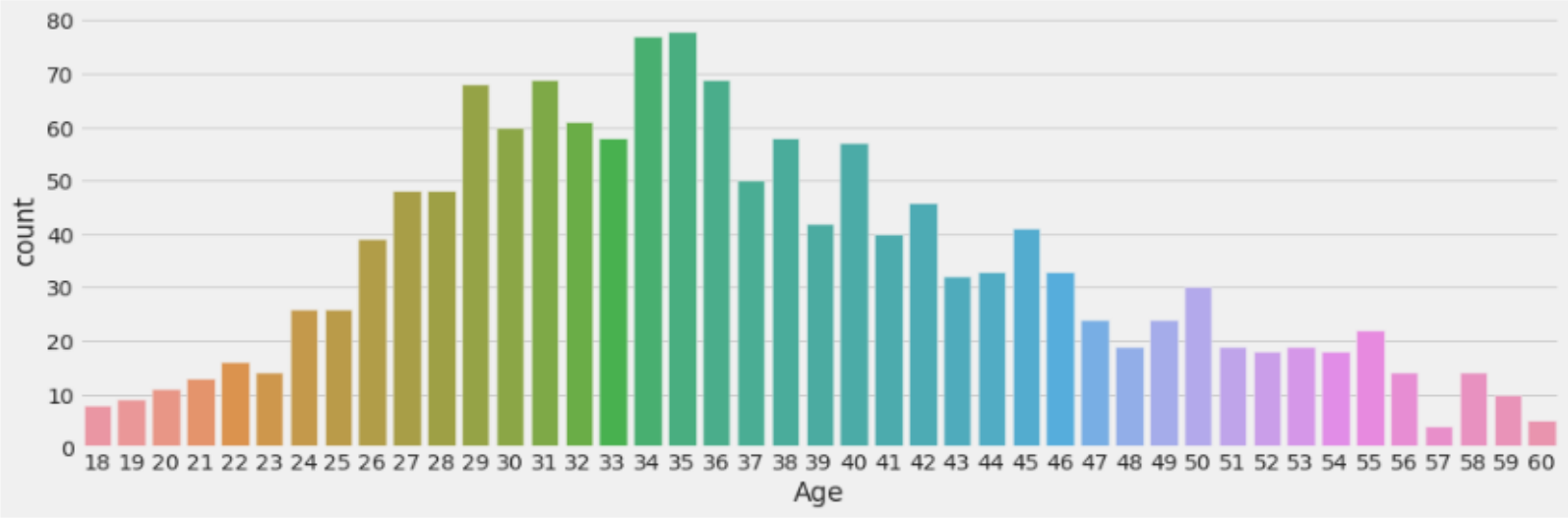
In [14]:

```
### Countplot

plt.subplots(figsize=(15,5))
sns.countplot(data.Age)
```

Out[14]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f50153e9210>
```

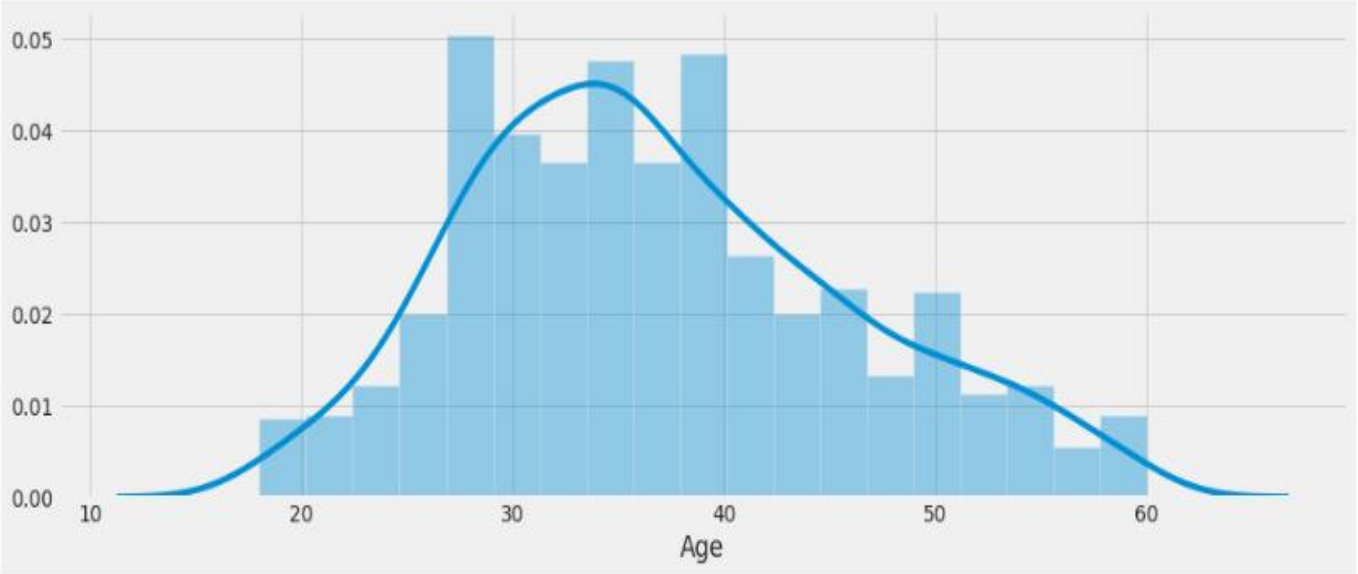


In [15]:

```
plt.subplots(figsize=(15,5))
sns.distplot(data['Age'])
```

Out[15]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f50158c6e90>
```



Output

HR Analysis, Prediction and Visualization

Summary and Conclusion

Data Science combines various disciplines to analyze and make predictions from data. IoT and Deep Learning are important parts of Data Science that enable real-time data analysis and complex decision-making. In the next lectures, we will delve deeper into each aspect of Data Science, explore tools, and work with datasets.