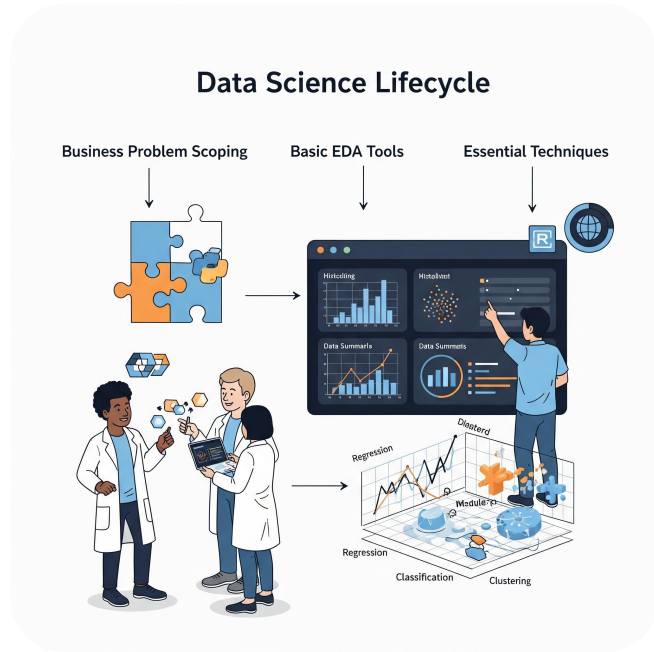




Exploratory Data Analysis

Week 1




Presentation Agenda



- From Business Goals to Data Project
- Essential Tools for Data Exploration
- Applied Exploratory Data Analysis

From Business Goals to Data Project

Why it matters

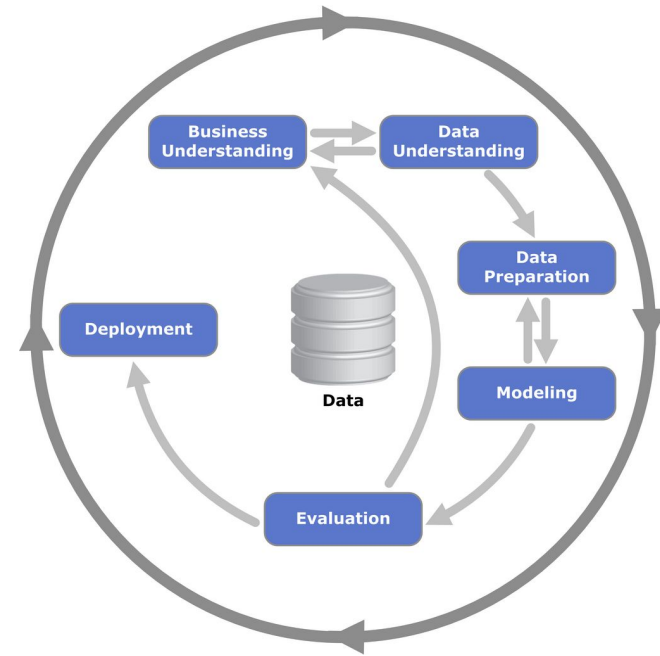
- Imagine you own an online flower shop,
- Sales are dropping, no one knows why!
- You feel like it's a price issue.
- We check the funnel and ...
- There's a big drop off in checkout!
- You fix the checkout UI/UX flaw.
- Sales recover!   

Practice 1

- Pause!
- Go and learn about CRISP-DM for a few minutes.

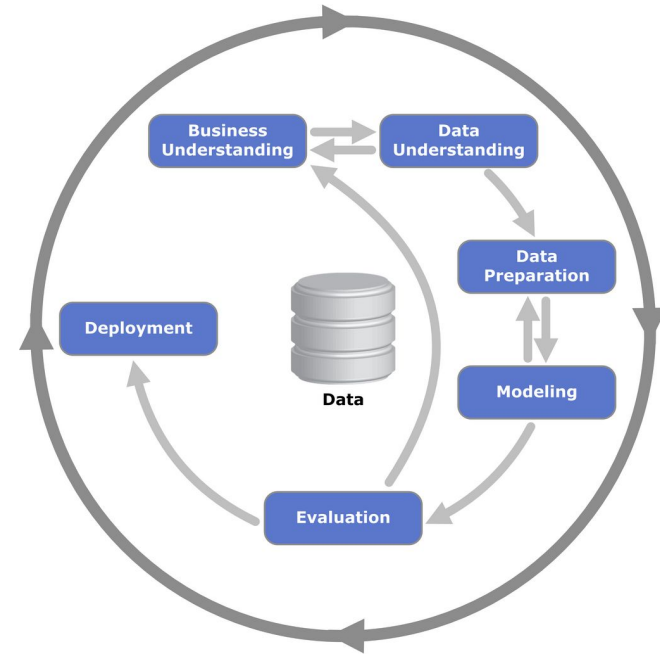
The Data Science Lifecycle

- CRISP - DM
- It's Not Always Linear!
- Works in any industry!
- Helps non-technical teams stay aligned with data teams
- Keeps the project organized and goal-driven
- Other methods: KDD, SEMMA, etc.



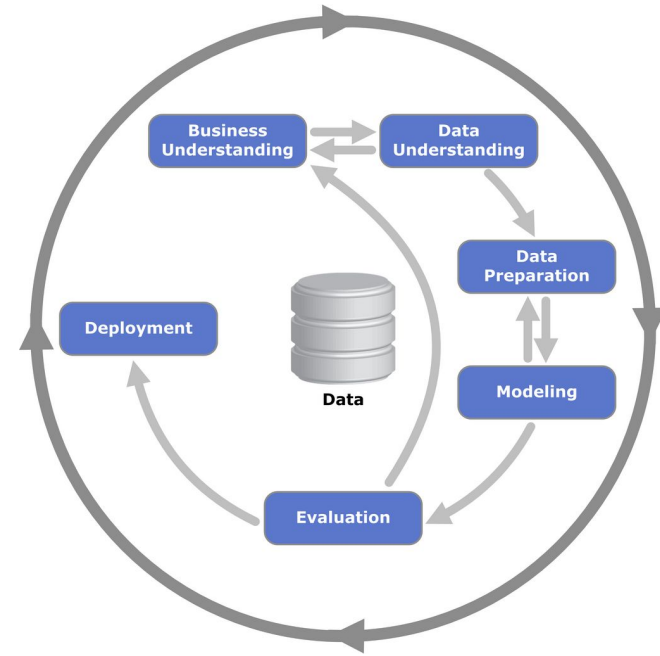
CRISP-DM Stages

- **1. Business Understanding**
- Understand why we are doing this project.



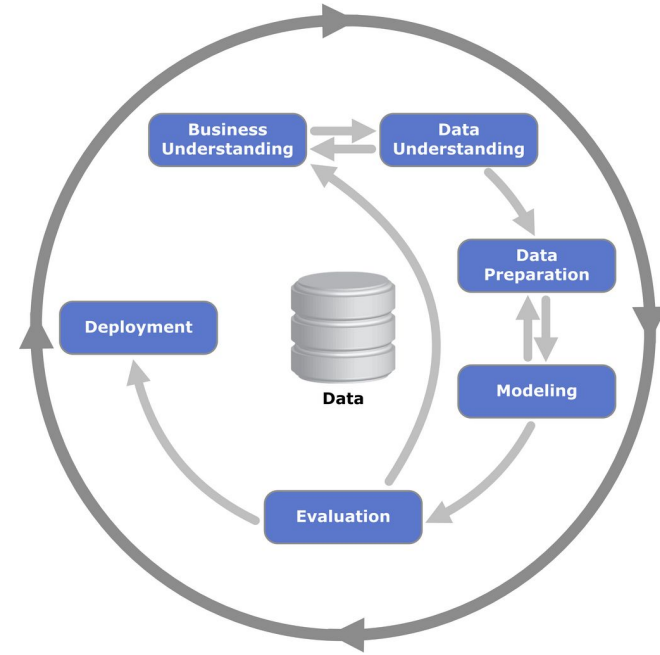
CRISP-DM Stages

- **2. Data Understanding**
- Gather data and get familiar with it.



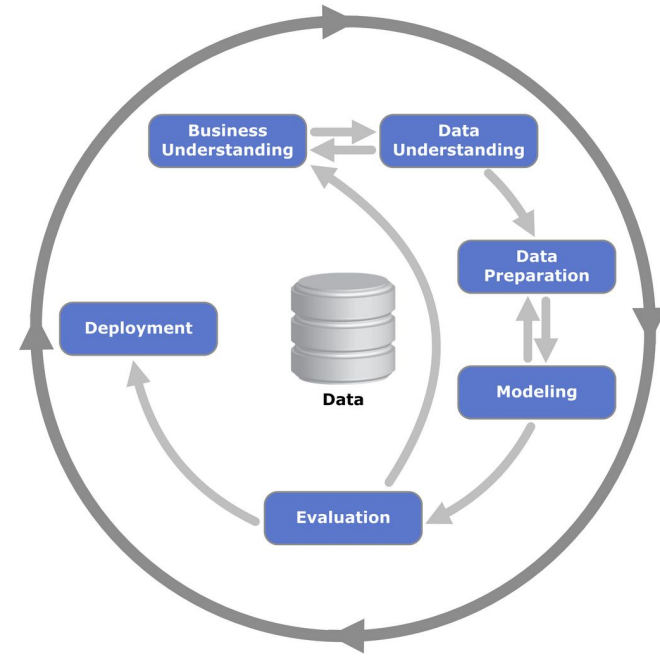
CRISP-DM Stages

- **3. Data Preparation**
- Clean and organize the data for modeling.



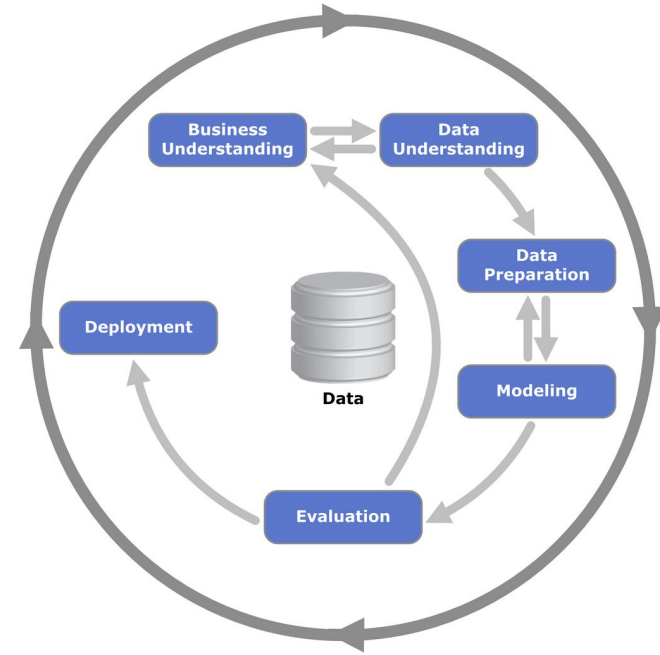
CRISP-DM Stages

- **4. Modeling**
- Model the data to create a solution



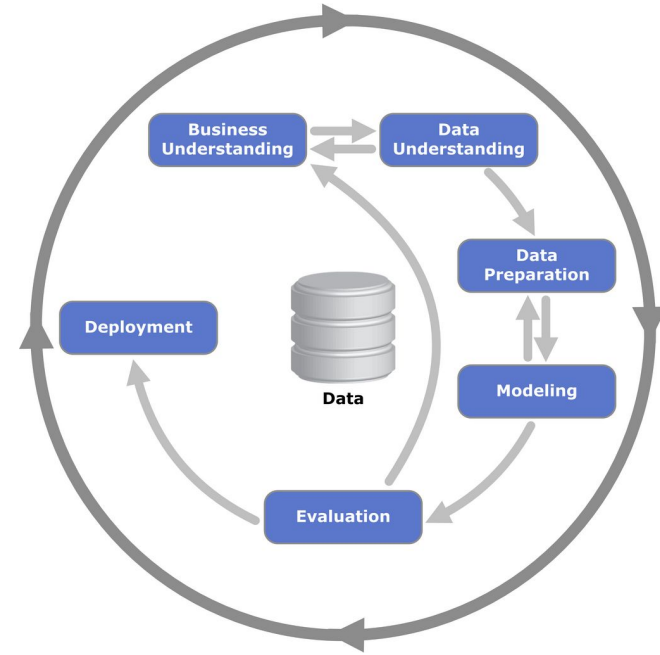
CRISP-DM Stages

- **5. Evaluation**
- Check if the model really solves the business problem.



CRISP-DM Stages

- **6. Deployment**
- Make the results available to the business.



Understanding Business Objectives: Clarity

- Before starting, we need clarity!
- What problem are we trying to solve?
- Who cares about this outcome?
- How do we know if we've succeeded?

Understanding Business Objectives: Business Problem vs. Data Problem

- Business problem: “Why are customers leaving our platform?”
- Data problem: “Can we predict churn based on usage history?”

Understanding Business Objectives: Stakeholders and Their Needs

- Who are they? (e.g., Marketing, Product, Operations, Executives)
- What are they measured on?
- How do they define success?

Understanding Business Objectives: Success Metrics

- “You can’t improve what you don’t measure.”
- Key Performance Indicators aka KPI (Look them up)
 - Conversion Rate
 - Retention Rate
 - CAC (Customer Acquisition Cost)
 - AOV (Average Order Value)
 - Churn Rate

Practice 2

- Go and learn more about metrics from previous slide!

Understanding Business Objectives: SMART Metric Criteria

- Specific
- Measurable
- Achievable
- Relevant
- Time-bound

Practice 3

- Online Fashion Platform
- Increase in product returns
- Rising logistics costs and a lot of refund requests
- CEO wants to "reduce returns" and asks for your help
- How can data help? What steps should we take?

Types of Data Analysis

- Descriptive
- Diagnostic
- Predictive
- Prescriptive

Types of Data Analysis:

Descriptive

- What happened?
- Summarizes past events and patterns
- E.g. Average sales per region or Monthly active users

Types of Data Analysis: Diagnostic

- Why did it happen?
- Investigates causes and relationships behind trends
- Why did churn increase last month?
- Which user segments are returning products more?

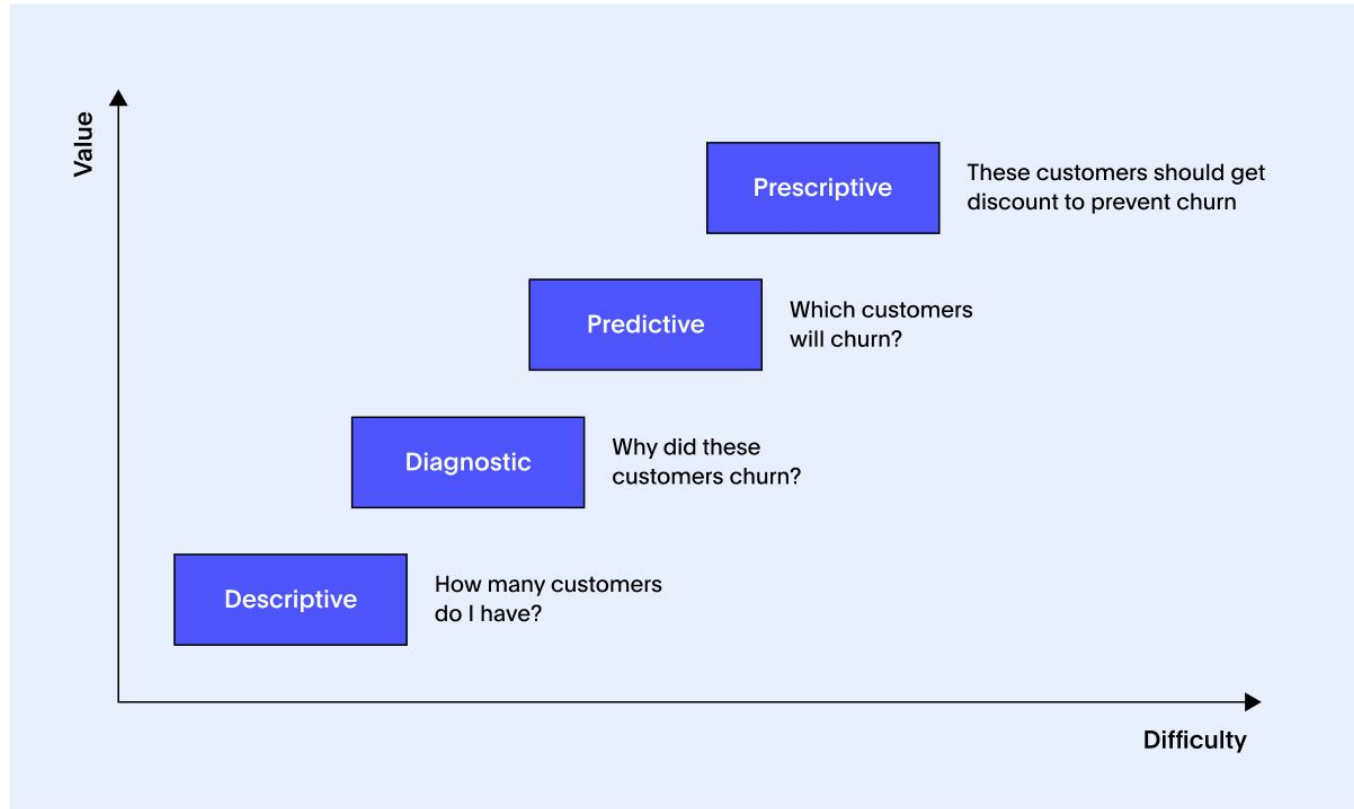
Types of Data Analysis: Predictive

- What is likely to happen?
- Uses historical data to forecast future outcomes
- Will a customer cancel their subscription next month?
- Predicting sales for Black Friday
- Often powered by ML models

Types of Data Analysis: Prescriptive

- What actions should be taken?
- Recommends actions or strategies based on data
- Dynamic pricing suggestions
- Recommending products to users
- Involves optimization, simulations, A/B testing

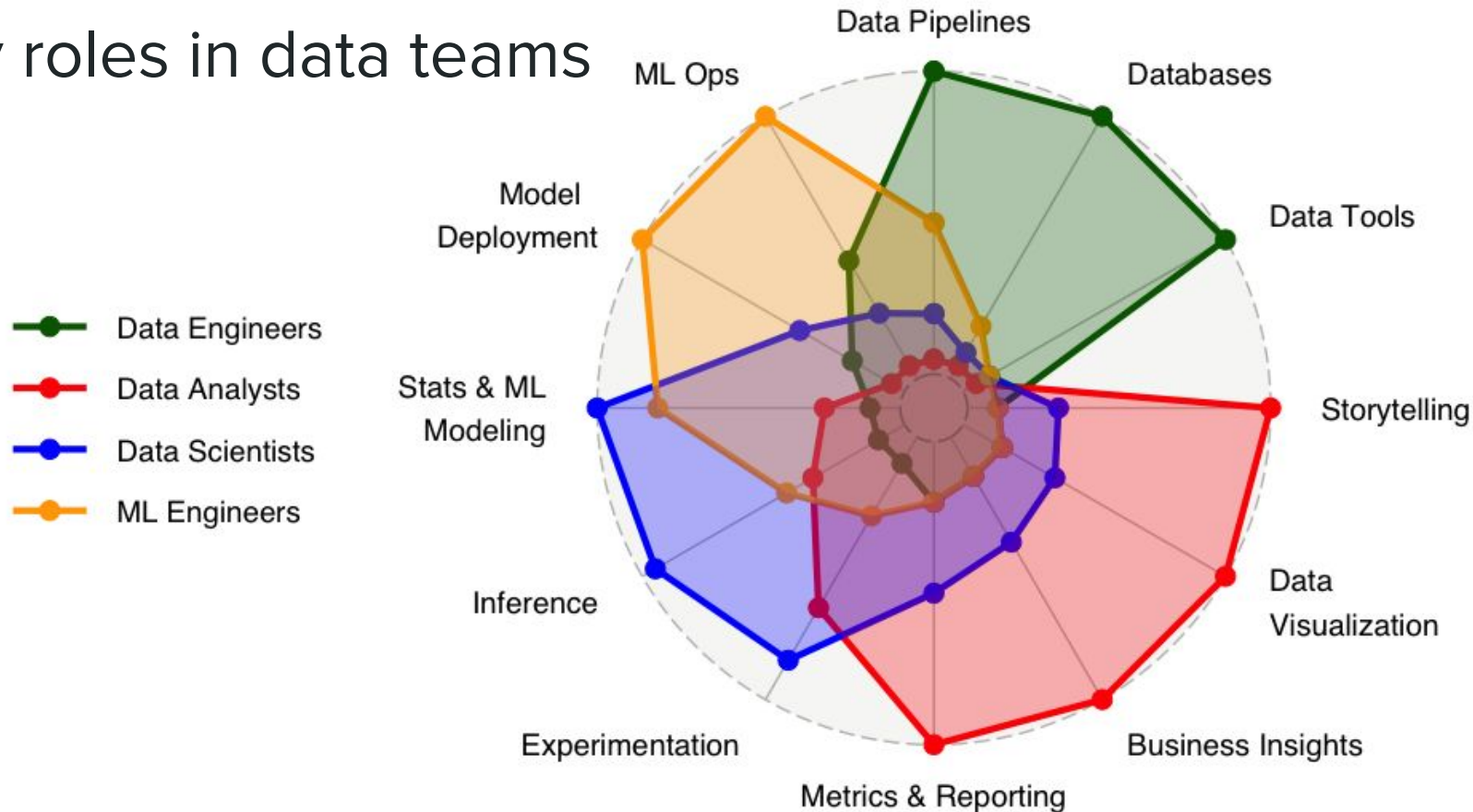
Types of Data Analysis



Key roles in data teams

- Data Engineer – Data pipelines, storage, optimization
- Data Analyst – Exploratory analysis, dashboards, reporting
- Data Scientist – Modeling, experimentation, insights generation
- Business Analyst – Translating business needs into data questions
- Product Manager – Aligning data work with product goals
- Data Architect – Designing data systems and infrastructure
- ML Ops Engineer – Managing model lifecycle in production
- Machine Learning Engineer – Model deployment & scaling
- Domain Expert – Providing business context for analysis

Key roles in data teams



Key roles in data teams

- Confusing? Yeah it is.
- These roles have a lot of overlap.
- Definition and responsibilities differ company by company!

Practice 4

- Which type of analysis is each role mostly doing?
- Explain your reasoning.

Essential Tools for Data Exploration

Jupyter Notebooks & Google Colab

- Fast, visual, interactive coding
- Perfect for quick iteration and prototyping
- Google Colab advantages:
 - No installation required
 - Free access to GPUs and Compute
 - Easy sharing and collaboration



NumPy

- A fundamental package for numerical computing
- Arrays vs. Python lists
- Speed & memory benefits
- Broadcasting & vectorized operations
- No for loops!



Pandas



- Built on top of NumPy
- Data manipulation using Series and DataFrames
- DataFrame is simply a table
- Simplifies working with tabular data
- Loading, Cleaning, Exploring, and Transforming data

Practice 5

- [Week1-Module2.ipynb](#)
- Colab, Numpy and Pandas.

SQL



- Structured Query Language
- A language to interact with relational databases
- SQL helps extract, filter, and join data efficiently
- Not necessary, but very insightful.

Matplotlib, Seaborn, Plotly

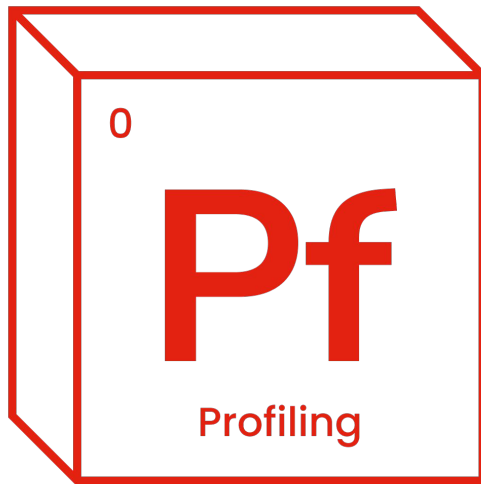


- Python libraries for turning data into graphs and charts.
- Used to spot patterns, explain findings, and tell stories.
- [Matplotlib](#): basic plots, gives full control, not easy!
- [Seaborn](#): focused on stats plots, easy to use.
- [Plotly](#): All plots + 3D + maps + dashboard, Interactive and easy to use.



YData Profiling and Sweetviz

- Auto-generated summary reports
- Single line of code! 🤖
- [YData Profiling Docs](#)
- Downside is that you're bombed with plots, summaries, etc.
- Also check sweetviz!
- [GitHub - fbdesignpro/sweetviz](#)



Kaggle

- Datasets, Courses, Competitions, Others' Notebooks
- Free GPU like Google Colab
- Community based
- Perfect place to start learning

The Kaggle logo, featuring the word "kaggle" in a lowercase, blue, sans-serif font.

Level up with the largest AI & ML community

Join over 25M+ machine learners to share, stress test, and stay up-to-date on all the latest ML techniques and technologies. Discover a huge repository of community-published models, data & code for your next project.

[Register with Google](#)[Register with Email](#)

Who's on Kaggle?

Learners

Dive into Kaggle courses, competitions & forums.



Developers

Leverage Kaggle's models, notebooks & datasets.



Researchers

Advance ML with our pre-trained model hub & competitions.

[See more](#)

Applied Exploratory Data Analysis

What is EDA in Practice?

- The process of using summary statistics and visual methods to understand a dataset before modeling.
- EDA is an Investigation, Not a Checklist
- Ask questions:
 - What does the data say?
 - What's missing?
 - What's strange?
- Go beyond summaries—look for patterns, outliers, surprises

Cleaning the Data

- Before we trust the data, we have to clean it.
- Real-world data is always messy.
- Missing values
 - Drop rows / columns
 - Impute with mean/median/mode
 - Create indicators 'is_missing'
 - Are all missing values the same?
- Incorrect Data Types
 - E.g. numbers stored as strings ('12345' instead of 12345)

Cleaning the Data

- Outliers
 - Can distort statistics and plots
 - Visual detection: boxplots, scatterplots
 - Statistical rules (IQR and Z-Score)
- Duplicates
 - Inflate statistics
 - May bias predictions
 - hide real patterns
 - Exact vs Near duplicates vs Domain-specific

Cleaning the Data

- Fix Inconsistent features
 - Case issues, spacing, typos -> “Tehran” vs “tehran” vs “ Tehran”
- Uniform Units
 - Toman vs Rial

Practice 6

- [Week1-Module3.ipynb](#)
- Go to notebook and practice the “Cleaning the Data” section.

Descriptive Statistics

- Numerical or visual summary
- Describes and simplifies a dataset
- Highlights its main features

Descriptive Statistics

Category	Metric Examples	What It Tells You
Central Tendency	Mean, Median, Mode	What is the “typical” or average value?
Dispersion	Range, Variance, Standard Deviation, IQR	How spread out or variable is the data?
Shape	Skewness, Kurtosis	Is the distribution symmetric or skewed? Are there extreme values?
Counts/Frequencies	Value counts, Percentages	How often do different values appear?

Practice 7

- [Week1-Module3.ipynb](#)
- Go to notebook and practice the “Descriptive Statistics” section.

Univariate Analysis

- Analyze One Variable at a Time
- Understand the basic characteristics of a single variable.
- Identify data quality issues (skewness, strange values)
- Spot potential feature transformations (e.g., log scale).
- Categorical Plots
 - Bar chart, pie chart
- Numeric Plots
 - Histogram, KDE, boxplot

Bivariate Analysis

- Explore Relationships Between Variables
- Understand interactions and correlations.
- Spot predictive power.
- Support or reject early hypotheses.
- A=Numeric, B=Numeric => Scatter plot, correlation matrix
- A=Numeric, B=Categorical => Grouped boxplot, violin plot
- A=Categorical, B=Categorical => Crosstab, stacked bar chart

Practice 8

- [Week1-Module3.ipynb](#)
- Go to notebook and practice the “Univariate vs Bivariate Analysis” section.

Practice 9 - Optional

- [Week1-Optional.ipynb](#)
- This practice is optional.