جامعة بيرزيت
BIRZEIT UNIVERSITY

Data Science and Analytics
Comp 4381

Ch1: Welcome & Introduction

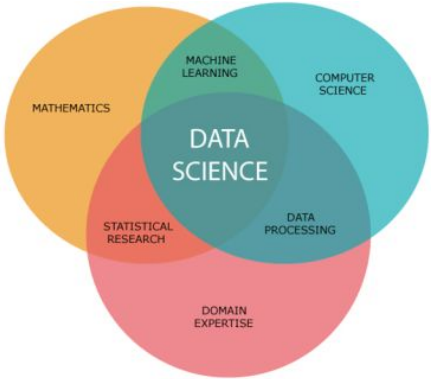
References

- **Books:**
  - Python for Data Analysis 3rd edition - Wes McKinney – O'RIELLY (Ch 2-10)
  - Python data science handbook 2nd edition - Jake VanderPlas – O'RIELLY (Ch 37-40)
  - Statistics unplugged 4th edition – Sally Cardwell - Wadsworth: (Ch 1, 2)
- **Material & Notebooks:**
  - Mr. Hussein Soboh.
- **Additional Resources:**
  - Computational and Inferential Thinking: The Foundations of Data Science 2nd Edition by Ani Adhikari, John DeNero, David Wagner. Link

Lecture 1

Welcome in Data Science

Data Science?



https://medium.com/analytics-vidhya/introduction-to-data-science-28deb32878e7

What is Data Science?

Learning about the world from data using computation

- **Exploration**
  - Identifying patterns in data
  - Uses visualizations
  - Example: Using a heatmap to find high-traffic areas in a city.
- **Inference**
  - Drawing reliable conclusions about the world
  - Uses statistics
  - Example: Determining if a new drug is effective using A/B testing.
- **Prediction**
  - Making informed guesses about unobserved data
  - Uses machine learning
  - Example: Predicting house prices based on features like location and size.

Motivation Example
&
Course Content
&
Questions?

Introduction

- **Data Science Introduction**
- Data Science workflow
- Data Science tools and roles

Data Science

- **A set of methodologies for taking in thousands of forms of data that are available to us today and using them to draw meaningful conclusions**
  - A methodology for analyzing vast amounts of data to uncover insights
- **Data Sources:**
  - Includes likes, clicks, emails, transactions, and social media activity
- **Purpose:**
  - Helps describe the present and predict future trends
- **Impact:**
  - Transforming industries by enabling data-driven decision-making
- **Growing Demand:**
  - Essential in today's digital world with massive data generation

Data-Driven Decision Making

- **Is the process of making business decisions based on data analysis and interpretation**
  - Instead of relying on intuition, personal experience, or just basic observation
  - **Decisions** are guided by concrete, objective data
- **Benefits:**
  - Reduces bias, improves efficiency, and enhances strategic planning
- **Importance:**
  - Essential for modern businesses to stay competitive and informed

Data-Driven Decision Making Phases

- **Ask a Question:**
  - Start by identifying a specific business problem or goal you want to address
  - Example : why sales have dropped and how to find the best way to allocate marketing budget
  - A clear, focused question helps guide the entire process
- **Gather the Correct Data:**
  - Gather the relevant data needed to answer it
  - Involve pulling data from internal sources like sales reports or customer databases
  - From external sources like market research and industry trends

Data-Driven Decision Making Phases

- **Prepare the Data:**
  - Cleaning and organizing the data to make it ready for analysis
- **Conduct Analysis:**
  - Choose suitable analytical techniques
- **Make the Right Decision:**
  - Decision should be guided by the data, aiming to solve the business problem or achieve the goal you set out to address

This approach minimizes risks and maximizes the chances of achieving desired outcomes by relying on factual evidence rather than guesswork. It also enables businesses to be more agile, responsive, and efficient in their operations

We Can Do With Data Science (1/4)

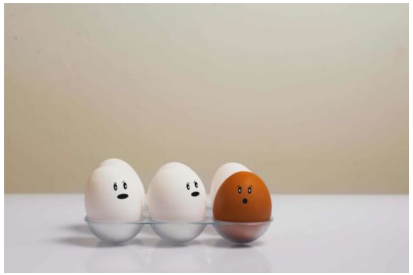
Describe the current state of an organization or process

- **Summarizing historical data helps organizations assess past performance and make informed decisions. Such as:**
  - **Business Intelligence:** Creating dashboards to track key performance indicators (KPIs), sales figures, and operational metrics
  - **Customer Insights:** Analyzing customer demographics, purchasing behavior, and feedback to understand preferences and satisfaction levels
  - **Operational Efficiency:** Monitoring processes such as supply chain operations, manufacturing output, and service delivery to identify areas for improvement

We Can Do With Data Science (2/4)

Aanomly Detection

- **Identify data points, events, or observations that deviate significantly from the norm, which could indicate errors, fraud, or unusual behavior**
  - **Fraud Detection:** Identifying unusual transactions that may indicate credit card fraud, insurance claims fraud, or financial statement manipulation
  - **Network Security:** Monitoring network traffic to detect potential security breaches, cyber attacks, or unusual access patterns
  - **Quality Control:** Detecting defects or anomalies in manufacturing processes to ensure product quality



We Can Do With Data Science (3/4)

Diagnose the causes of events and behaviors

- **Understand why something happened by identifying relationships and causal factors within the data. Such as:**
  - **Root Cause Analysis:** Investigating the underlying reasons for process failures, product defects, or customer churn
  - **Customer Behavior:** Understanding the factors that drive customer decisions, such as why certain products are popular or why customers leave
  - **Healthcare:** Diagnosing the causes of patient symptoms or medical conditions based on historical data and patterns

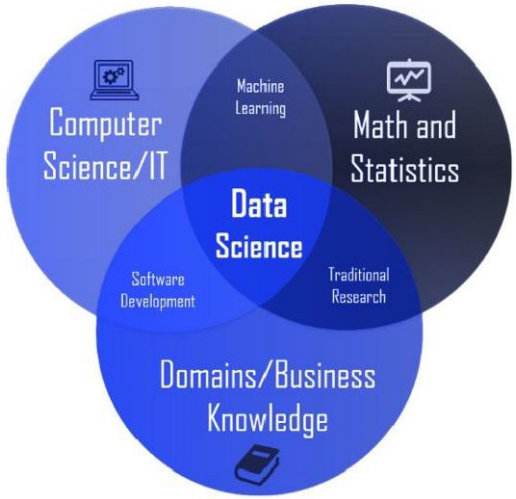
We Can Do With Data Science (4/4)

Predict future events

- **Forecast future outcomes based on historical data and statistical models. Such as:**
  - **Sales Forecasting:** Predicting future sales based on historical sales data, market trends, and economic indicators
  - **Risk Management:** Anticipating potential risks, such as loan defaults, equipment failures, or market volatility, to take preemptive actions
  - **Personalized Marketing:** Predicting customer behavior to tailor marketing campaigns and offers to individual customers, increasing engagement and conversion rates

Data Science Disciplines (1/4)

- **Data science is an interdisciplinary field combining statistics, computer science, and domain expertise to analyze complex data. It involves:**
  - **Algorithms & Data Analysis:** Extracting meaningful insights from raw data
  - **Machine Learning & Predictive Modeling:** Identifying patterns and making data-driven predictions
  - **Strategic Decision-Making:** Leveraging insights to drive business and operational improvements



Data Science Disciplines (2/4)

Math and Statistics

- **Math and statistics is crucial for a data scientist**
  - **Data Analysis & Interpretation**: Use statistical techniques (descriptive statistics : summarize data, probability distributions: extract insights, inferential statistics: informed decision-making)
  - **Data Preprocessing**: Apply techniques like normalization, standardization, and dimensionality reduction (e.g., PCA) to prepare data for analysis and modeling
  - **Building Models**: Develop machine learning models (e.g., linear regression, decision trees, clustering, neural networks) using mathematical foundations such as linear algebra, calculus, and optimization for better performance and fine-tuning

Data Science Disciplines (3/4)

Computer Science

- **Provides the computational foundation and tools necessary to process, analyze, and visualize large and complex data sets:**
  - **Programming Proficiency**: Mastering languages like Python, R, and SQL is crucial for data manipulation, analysis, and visualization
  - **Software Development Practices**: Understanding version control (e.g., GitHub), testing, and modular coding ensures maintainable and scalable data science projects

Data Science Disciplines (4/4)

Domain Expertise

- **Domain expertise refers to the in-depth knowledge and understanding of a specific field or industry. Expertise is crucial in data science because it allows data scientists to:**
  - **Understanding Requirements:** Knowledge of business objectives and data context ensures accurate analysis. For example, healthcare data scientists must understand medical terminology, regulations, and clinical practices
  - **Validating Results:** Domain expertise is essential for interpreting model outputs, identifying anomalies, and ensuring insights are meaningful
  - **Effective Communication:** Helps translate complex data findings into actionable insights that stakeholders can easily understand and applay

Why Data Science is Thriving Now

- **Data Explosion**: The massive growth of digital data from IoT, social media, and e-commerce has fueled the need for data analysis. The global data sphere is projected to reach **175 zettabytes by 2025** (IDC, "The Digitization of the World From Edge to Core").
- **Business Value**: Data-driven organizations are significantly more successful— **23 times more likely to acquire customers**, **6 times as likely to retain customers, and 19 times as likely to be profitable.**(McKinsey).
- **Technological Advancements**: Affordable cloud storage, advanced machine learning, and big data tools have made data science more accessible. Data storage costs have dropped from **$500,000/GB (1980) to $0.02/GB today** (Forbes).
- **Growing Demand for Data Roles**: The need for data scientists and analysts is surging, with **a 37% annual growth rate in data science roles** (LinkedIn) and **16% projected job growth from 2018-2028** (U.S. Bureau of Labor Statistics)

Introduction

- Data Science Introduction
- **Data Science workflow**
- Data Science tools and roles

Data Science workflow

- **Comprehensive process that transforms raw data into actionable insights and predictive models**

Data Collection and Storage

- **Collect data from many sources, such as surveys, web traffic results, geo-tagged social media posts, and financial transactions.**
  - Store that data in a safe and accessible way
- **Storage**
  - **Storage location:** Where we want to store the data
  - **Types of Data Storage:** What kind of data and storage
  - **Retrieval: Data querying:** How we can retrieve the data from storage

Data Collection and Storage (1/3)

Storage location

- **Large Data Storage:** Data science projects often require vast amounts of data that can't be stored on a single computer
- **Distributed Storage:** Data is stored across multiple computers for accessibility and reliability
- **On-Premises Storage:** Large companies use dedicated servers or clusters to store data internally
- **Cloud Storage:** Businesses can use services like Microsoft Azure, AWS, or Google Cloud for scalable and managed storage solutions
- **Beyond Storage**: Cloud providers offer additional services such as data processing, security, and analytics

Data Collection and Storage (2/3)

Data Storage Solutions

- **Unstructured Data: Includes email, text, videos, audio files, web pages, and social media messages.**
  - Typically stored in Document Databases (type of NoSQL database)
- **Structured Data:** Organized in tables, similar to spreadsheets, and
  - Stored in Relational Databases
- Both types of databases are available on cloud storage platforms

Data Collection and Storage (3/3)

Retrieval: Data querying

Once data is stored in either a Document Database or a Relational Database, we need efficient ways to access and analyze it. Common queries include:

- Retrieving specific data, such as "All images created on March 3rd" or "All customer addresses in Montana."
- Performing analytical operations, like summing, counting, or averaging values.

Each type of database has its own query language:

- Relational Databases primarily use SQL (Structured Query Language) for structured data retrieval.
- Document Databases primarily use NoSQL (Not only SQL), which offers flexible querying methods suited for unstructured or semi-structured data.

Extract, Transform, Load (ETL)

- **Process used in data engineering to :**
  - **Extraction:** Retrieving data from various sources like databases, APIs, and flat files (e.g., extracting SQL data from MySQL or fetching weather data from APIs). **Challenges** include managing diverse data formats, large volumes, and data retrieval speed.
  - **Transformation:** Cleaning and structuring data for usability, such as merging datasets, converting formats, and removing irrelevant information (e.g., filtering unnecessary details from API responses).
  - **Loading:** Storing processed data into a target system like a data warehouse (e.g., Amazon Redshift) or relational database (e.g., MySQL) for analysis and reporting.

Data Preparation and Cleaning

- **Clean and well-prepared data ensures more accurate analyses, reliable insights, and better decision-making.**
- **Steps of data cleaning and preparation:**
  - Validating data types
  - Handling duplicates
  - Addressing missing values
  - Resolving data inconsistencies
  - Detecting and handling outliers

Data Preparation and Cleaning (1/5)

Validating data types

- **Data types define how data is stored and processed, such as integers, floats, strings, and dates**
  - Ensuring correct data types is essential for accurate analysis and efficient algorithm performance
- **Key Considerations:**
  - **Dates** should be in the correct format for time-series analysis
  - **Numerical values** must be properly classified as integers or floats
  - **Categorical data** should be stored as strings for classification tasks
- **Handling Issues:**
  - Convert mismatched data types (e.g., strings to dates or numbers)
  - Ensure consistency for accurate insights and reliable machine learning models

Data Preparation and Cleaning (1.1/5)

Validating data types

Befor ?

ID	Name	Age	Weight	Country
0	Sami	“27”	75	”Palestine”
1	Sara	“30”	68	“US”
2	Salwa		65	”Egypt”

After

ID	Name	Age	Weight	Country
0	Sami	27	75	”Palestine”
1	Sara	30	68	“US”
2	Salwa		65	”Egypt”

Data Preparation and Cleaning (2/5)

Handling duplicates

- **Duplicates can skew analysis and lead to misleading conclusions**
- **Causes:**
  - Repeated data entry, dataset merging, or collection errors
- **Example:**
  - Multiple entries for the same customer can inflate customer count
- **Solution:**
  - Identifying and merging duplicates improves data accuracy

Before

ID	Name	Age	Weight	Country
0	Sami	"27"	75	"Palestine"
1	Sara	"30"	68	"US"
2	Salwa		65	"Egypt"
0	Sami	"27"	75	"Palestine"

After

ID	Name	Age	Weight	Country
0	Sami	27	75	"Palestine"
1	Sara	30	68	"US"
2	Salwa		65	"Egypt"

Data Preparation and Cleaning (2.1/5)

Should duplicates always removed?

- **Decision Making:**
  - Remove or consolidate duplicates based on analysis needs
- **Aggregation:**
  - Sometimes, duplicates should be averaged, summed, or grouped
- **Example:**
  - In sales data, duplicate transactions for the same product may need to be summed for accurate total sales
- **Goal:**
  - Ensure data integrity and accuracy in analysis

Data Preparation and Cleaning (3/5)

Addressing missing values

- **Missing data occurs due to incomplete entry, lost records, or varying collection methods**
- **Proper handling is essential for accurate analysis:**
  - **Remove** rows/columns with excessive missing values.
  - **Impute using**: Mean, median, or mode
  - **Advanced techniques** (e.g., regression imputation)
- **Example:**
  - In a medical dataset, missing patient weights can be filled with the average weight of patients in the

Data Preparation and Cleaning (3.1/5)

Advanced techniques (regression imputation)

- **Regression imputation predicts missing values using other variables in the dataset**
  - A regression model is trained on existing data to estimate the missing values
- **Example:**
  - In a housing dataset, if a house's size (sq ft) is missing, we can predict it using number of **bedrooms**, **bathrooms**, and **lot size** by training a regression model

Data Preparation and Cleaning (4/5)

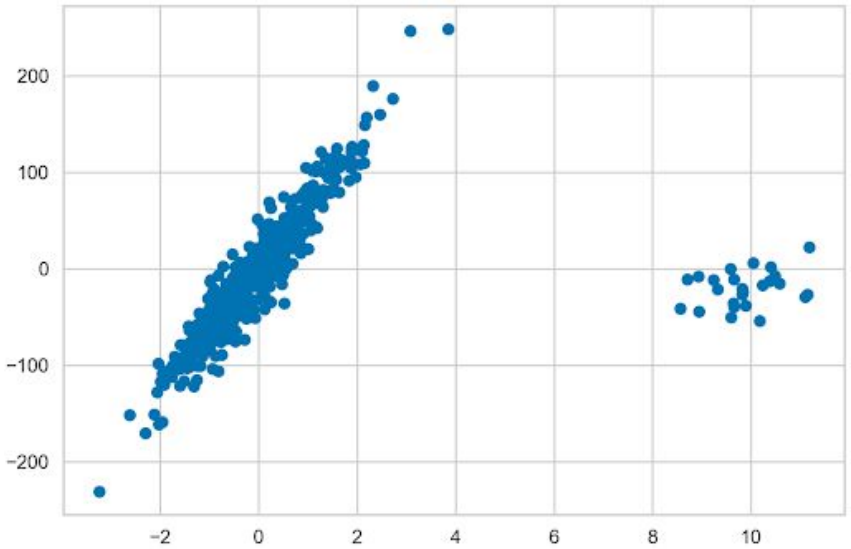
Resolving data inconsistencies

- **Variations in data formats lead to unreliable analysis**
- **Examples:**
  - Spelling differences: **"NYC" vs. "New York City"**
  - Unit inconsistencies: **kg vs. lbs**
  - Mixed capitalization: "**electronics**" vs. "**ELECTRONICS**"
- **Solution:**
  - Convert text to lowercase
  - Standardize date formats
  - Ensure uniform numerical units
- **Example:**
  - In a sales dataset, standardizing *"electronics"*, *"Electronics"*, and *"ELECTRONICS"* ensures accurate sales analysis

Data Preparation and Cleaning (5/5)

Detecting and handling outliers

- **Outliers are data points that significantly deviate from the dataset. They may indicate errors, rare events, or meaningful variations**



Data Preparation and Cleaning (5.1/5)

Detecting and handling outliers

- **Identification Methods:**
  - **Statistical**: Interquartile Range (IQR), Z-score
  - **Visual**: Box plots, scatter plots
- **Handling Strategies:**
  - Remove if it's an error
  - Transform to minimize impact
  - Keep if it holds valuable insights
- **Example:**
  - In a financial dataset, a transaction much higher than average could be a data entry error or a high-value purchase, requiring careful analysis

$$z = \frac{x - \mu}{\sigma}$$

- IQR=Q3-Q1
- Values falling below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR are often considered outliers

Data Exploration and Visualization

Exploratory Data Analysis (EDA)

- **Critical process in data science and statistics used to analyze and summarize the main characteristics of a dataset, including:**
  - Using visual and quantitative methods to understand the data's structure
  - Identify patterns
  - Detect anomalies
  - Test hypotheses
- **EDA helps to reveal underlying trends and relationships that might not be immediately apparent, guiding further analysis and decision-making**

Data Exploration and Visualization

Exploratory Data Analysis (EDA)

- **Why it is important?**
  - Manually finding the counts for each number would be cumbersome and error-prone.

How many zeros are there ?

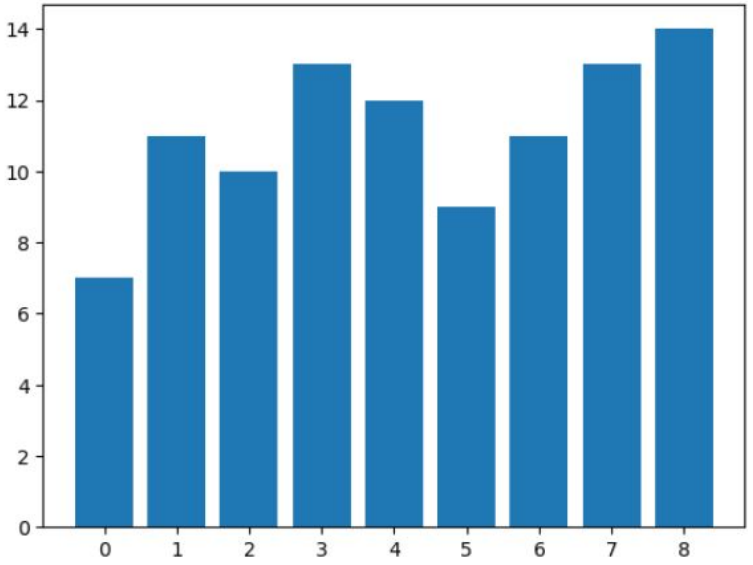
```
[3, 4, 6, 3, 3, 7, 1, 7, 0, 2,
 6, 6, 1, 4, 7, 4, 5, 6, 8, 2,
 5, 1, 1, 4, 4, 2, 1, 8, 4, 7,
 2, 3, 5, 4, 5, 8, 8, 7, 7, 8,
 3, 1, 7, 3, 2, 3, 3, 1, 8, 8,
 8, 3, 6, 5, 2, 4, 8, 6, 1, 7,
 5, 2, 8, 5, 4, 6, 3, 1, 0, 0,
 4, 0, 6, 1, 2, 1, 8, 7, 8, 8,
 7, 0, 4, 6, 8, 7, 3, 2, 0, 5,
 2, 3, 4, 6, 0, 5, 7, 6, 3, 7]
```

Data Exploration and Visualization

Exploratory Data Analysis (EDA)

- **Why it is important?**
  - Reduced the amount of complexity
  - Easy to read and interpret
  - Reveal patterns, trends, and outliers
  - Provides key statistics for quick insights
  - Enhances decision-making.

How many zeros are there ?



Components of EDA (1/2)

Summarization

- **Condenses data into key metrics, highlighting central tendencies, variability, and distribution.**
- **Examples of data summarization:**
  - **Descriptive Statistics:** Calculating basic statistics such as mean, median, mode, variance, and standard deviation
  - Dataset of 15K job salaries (2020–2024), the median salary is $141,300, indicating that half of the salaries are below and half are above this value

	count	mean	std	min	25%	50%	75%	max
Salary (USD)	14,838.0	149,874.7	69,009.2	15,000.0	102,000.0	141,300.0	185,900.0	800,000.0

Components of EDA (2/2)

Summarization

- **Frequency Tables:** Creating tables that show the number of occurrences of different categories or values. For instance, the summary below shows that the majority of jobs are full-time (FT), while part-time (PT) and contract (CT) positions make up the very smaller portion

Employment Type	FT	PT	CT
count	14785	27	26

- **Distribution Metrics:** Summarizing the distribution of the data using **quartiles**, **percentiles**, and **ranges**. For example, we can summarize the test scores by calculating which score that 75% of the scores lies below it (75th percentile)

Announcements

Assignment: Data Distribution Analysis

Each student must submit a **2 to 3-page handwritten report** summarizing key concepts related to data distribution. The report should be **clear, well-structured, and include explanations, equations, and real-world examples** to illustrate each concept

Required Topics:

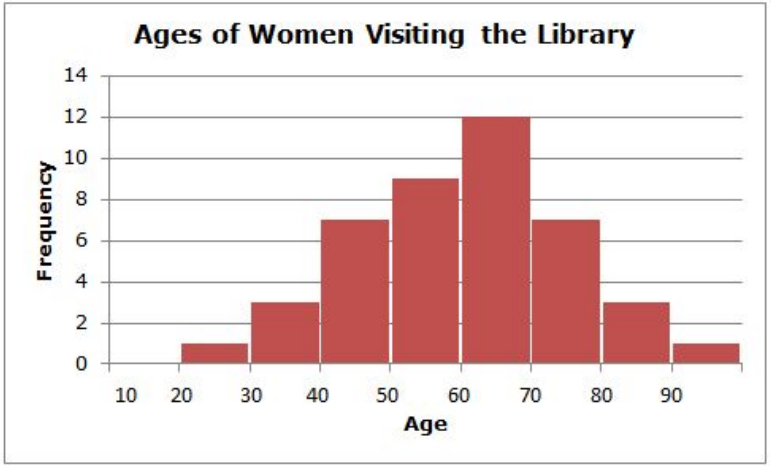
1. **Measures of Central Tendency**: Explanation and calculation of **mean, median, and mode**, with real-world examples
2. **Measures of Dispersion**: Discussion on **standard deviation, variance, range, and interquartile range (IQR)** with calculations
3. **Quartiles and Percentiles**: Explanation of their significance, along with calculations and interpretations
4. **Outliers**: Definition, impact on data analysis, and methods to detect them (e.g., Z-score, IQR method)
5. **Equations & Examples**: Provide formulas and step-by-step calculations, using real-world datasets where possible

Submission Guidelines:

- The report must be **handwritten** to ensure individual understanding and effort
- Use **clear and organized formatting** with headings and subheadings for each section
- Include at least **one real-world dataset or example** to support explanations

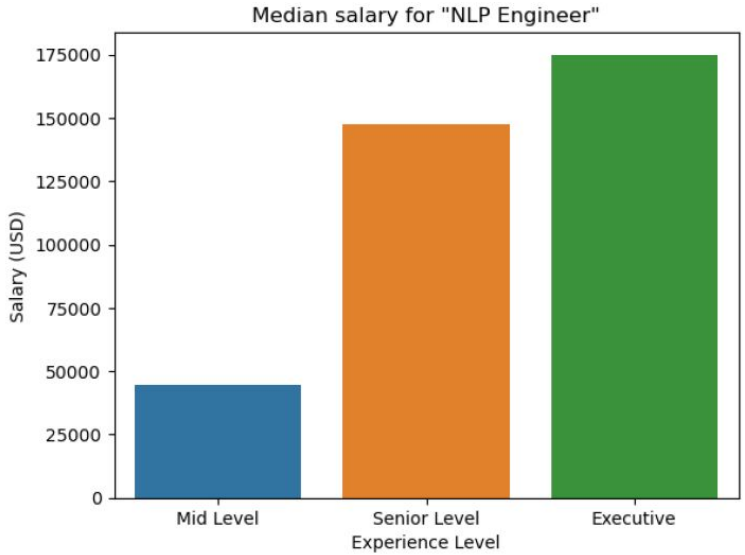
Visualization

- **Transforms data into graphical formats, making it easier for humans to detect patterns, trends, and outliers**
  - histograms, box plots, scatter plots, bar charts, line charts, and others
  - Eg. histogram is a univariate plot graphically representing the distribution of a single variable



Visualization

- **Eg.** Changes in the median salary for "NLP Engineers" based on experience. It indicates that NLP Engineers are predominantly at the Mid or higher levels, as there are no jobs listed at the "Entry Level." Additionally, the plot shows a significant salary increase from $45K to $145K when moving from Mid to Senior level.



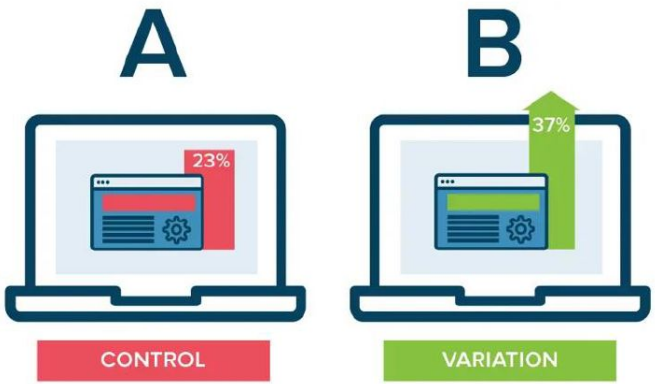
Experimentation and Prediction

- **This final phase** where data truly comes to life, allowing us to test hypotheses, predict future trends, and derive actionable insights
- **Involve:**
  - Applying statistical methods and machine learning techniques
  - To draw conclusions and make forecasts based on the data
- This phase is crucial for validating your assumptions, optimizing processes, and guiding business strategies
- Examples of techniques uses in this stage:
  - A/B Testing
  - Time Series forecasting
  - Machine Learning

Experimentation and Prediction Techniques (1/3)

A/B Testing (split testing)

- Controlled experiment that compares two versions of a variable to determine which one performs better
- Experiments guide decision-making and help form conclusions
- Start with a question and a hypothesis, proceed with data collection, and are followed by a statistical test and its interpretation



Experimentation and Prediction Techniques (1.1/3)

A/B Testing steps

- 1. Start by a question and form hypothesis: what you want to test and what success looks like
- 2. Create the variants by developing two versions of the variable you want to test—Version A (the control) and Version B (the variation/validation)
- 3. Split the audience randomly into two groups. One group will see Version A, and the other will see Version B. The random assignment ensures that the results are unbiased and not influenced by external factors
- 4.Measure the results by collecting data on how each version performs according to the objective you defined
- 5. Analyzing the results by using statistical methods to determine if the differences in performance are significant. If Version B outperforms Version A by a significant margin, you might choose to implement the changes in Version B

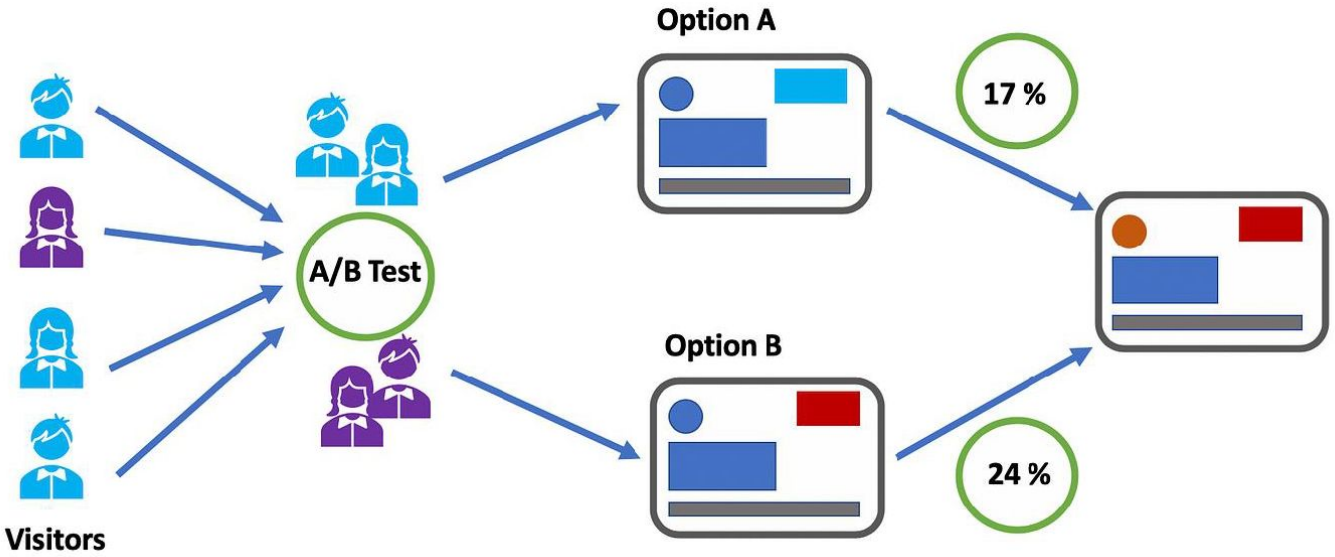
Experimentation and Prediction Techniques (1.2/3)

A/B Testing Example

- **Let's say you're a product manager at an e-commerce company, and you want to increase the number of people who add items to their shopping cart**
  - 1. The **question** you want to answer or in other words, the objective is to increase the "Add to Cart" rate on a product page. The hypothesis could be changing the "Add to Cart" button's color to red will increase the "Add to Cart" rate compared to the current design, where the button is blue.
  - 2. **Build two pages**: The current product page, where the "Add to Cart" button is blue, call this Option A. A redesigned product page, where the "Add to Cart" button is red, call this Option B
  - 3. Split the audience by splitting your website traffic so that 50% of visitors see **Option A** and the other 50% see **Option B.** Over a week, you track how many users in each group click the "Add to Cart" button.
  - 4. Collect the results for example:
    - a.Option A: 1,700 out of 10,000 visitors (17%) click "Add to Cart."
    - B. Option B: 2400 out of 10,000 visitors (24%) click "Add to Cart."
  - 5. Analyze the results by comparing the data collected from the two options. The above result shows that Option B has a higher "Add to Cart" rate than Option A. And by using a statistical test (e.g., a t-test) you can confirm that the difference is statistically significant
- **Therefore, based on the results, you decide to implement the changes in Option B across the website, as it leads to a higher conversion rate.**

Experimentation and Prediction Techniques (1.3/3)

A/B Testing Example



Experimentation and Prediction Techniques (2/3)

Time Series forecasting (Machine Learning)

- **Time series forecasting is a method used to predict future values based on previously observed data points that are ordered in time.**
- **Unlike other types of data, time series data is sequential and temporal, meaning that the order of the data points is crucial to the analysis**

Experimentation and Prediction Techniques (2.1/3)

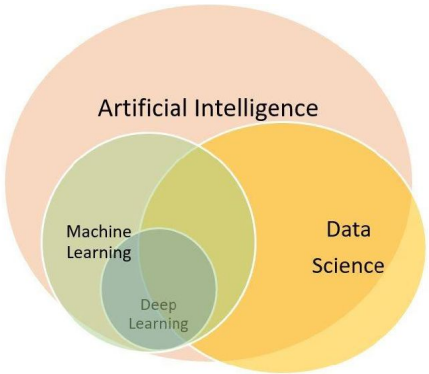
Time Series forecasting (Machine Learning)

- Machines learn from past data to make predictions and decisions without explicit programming
- **Applications:** Fraud detection, healthcare, search engines, recommendations, image & speech recognition.
- ML is a subset of **Artificial Intelligence (AI)**, which aims to create intelligent systems that simulate human thinking.
- **AI Includes:** ML, rule-based systems, robotics, expert systems, and natural language processing (NLP)

Experimentation and Prediction Techniques (2.3/3)

Data Science vs Machine Learning and AI

- **Data Science:** An interdisciplinary field focused on extracting insights from data using techniques from statistics, ML, programming, and data engineering
- **Uses:** Recommendation engines, user behavior prediction, and data-driven decision-making
- **Machine Learning & AI:** ML is a subset of AI that enables machines to learn from data, while AI encompasses broader techniques to simulate human intelligence



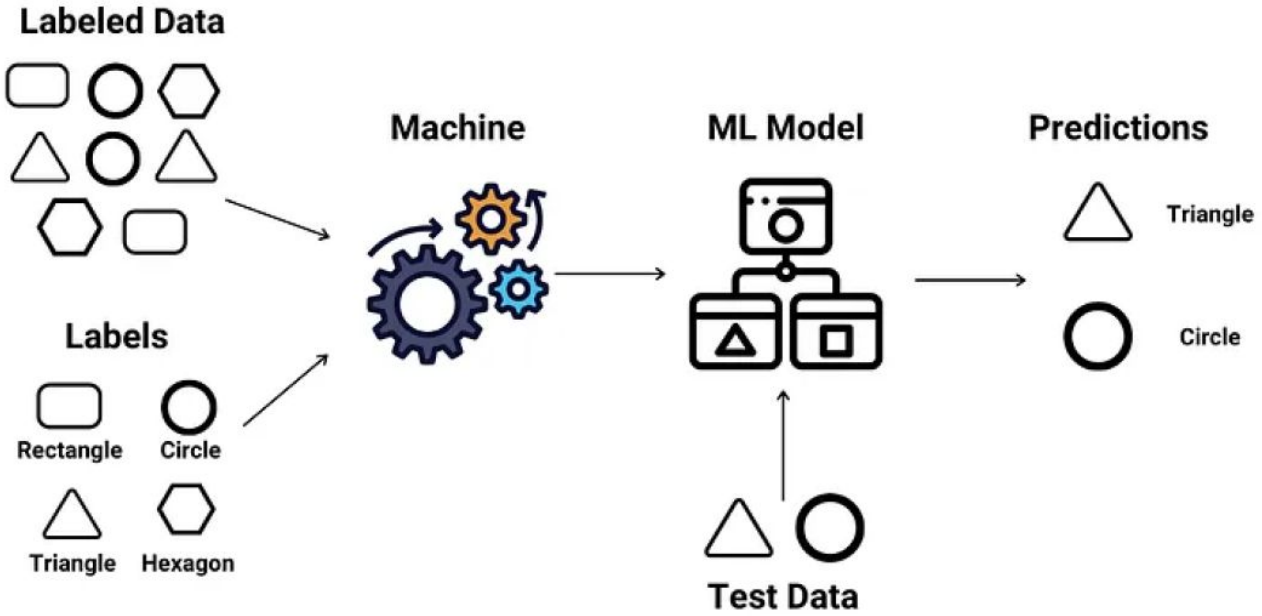
Experimentation and Prediction Techniques (3/3)

Types of machine learning:

- **Supervised Learning:**
  - Uses labeled data to train models for prediction and classification
- **Unsupervised Learning:**
  - Identifies patterns in unlabeled data, such as clustering
- **Reinforcement Learning:**
  - Focuses on sequential decision-making, like robots or game strategies

Experimentation and Prediction Techniques (3.1.1/3)

Supervised machine learning



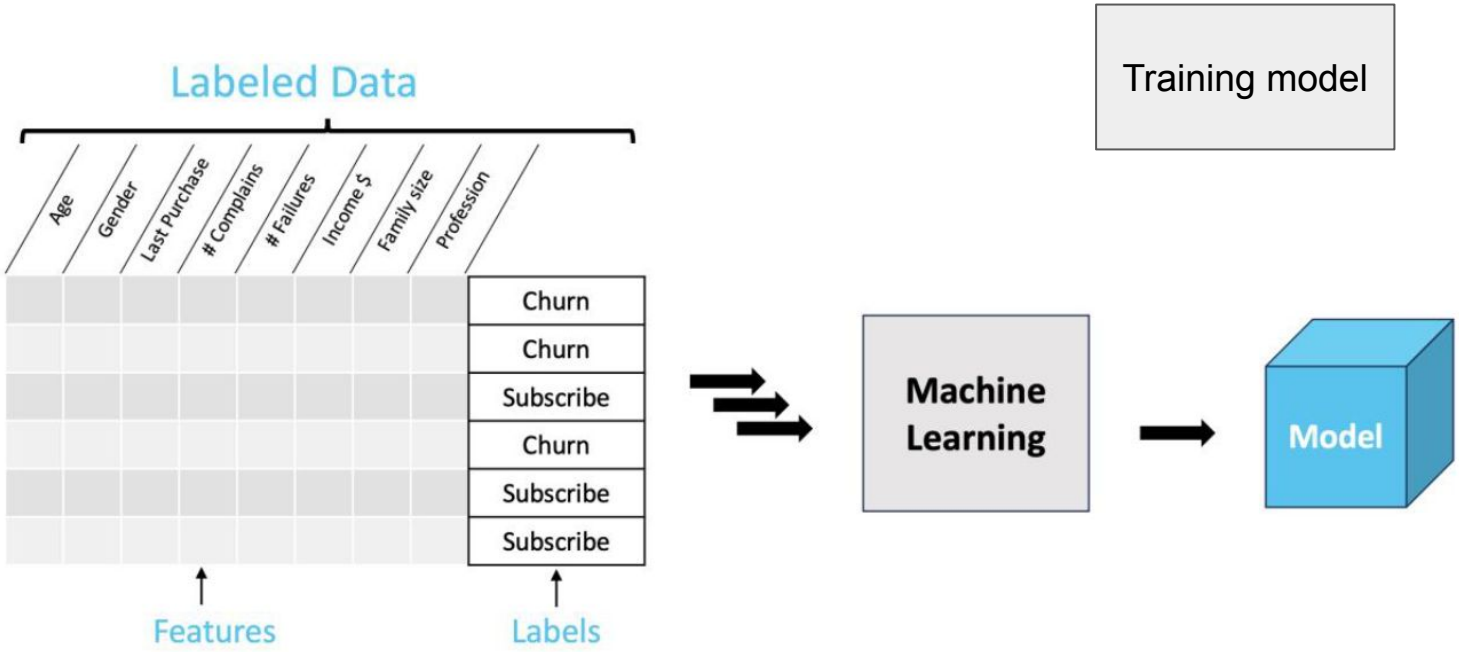
Experimentation and Prediction Techniques (3.1.2/3)

Case Study: Customer Churn Prediction

- Suppose we have a subscription business and want to predict whether a given customer is likely to stay subscribed or cancel their subscription, also known as churn
- Features are different pieces of information about each customer that might affect our label

Experimentation and Prediction Techniques (3.1.3/3)

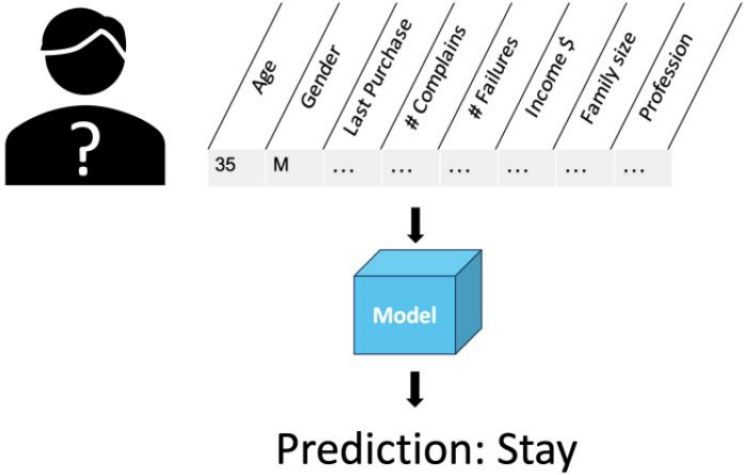
Case Study: Customer Churn Prediction



Experimentation and Prediction Techniques (3.1.4/3)

Case Study: Customer Churn Prediction

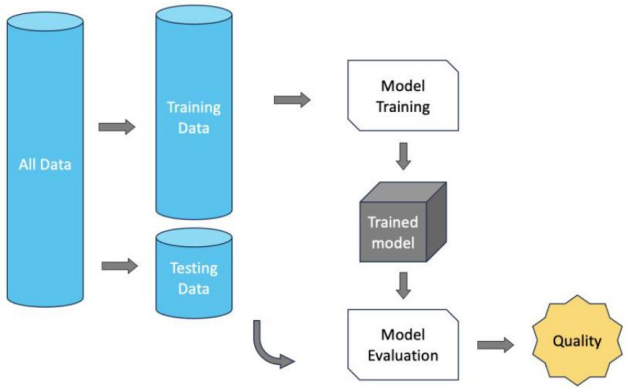
- New customer or unseen sample
- The model not trained on sample before



Experimentation and Prediction Techniques (3.1.5/3)

Evaluation Model

- **Model evaluation: We must ensure that model not only learns from the data but also generalizes well to new, unseen data. Split data into :**
- **Training Set (70-80%)**: Used to teach the model by identifying patterns in the data.
- **Testing Set (20-30%)**: Used to evaluate the model's performance on unseen data.



Experimentation and Prediction Techniques (3.1.6/3)

Evaluation Metrics

- **Accuracy:** Measures the proportion of correctly predicted customers (both churned and non-churned) out of the total customers

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}}$$

- **Precision:** Measures how many of the customers predicted as "churned" actually churned

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

- **Example:** If our model predicted 40 customers as **churned** but only 20 were correct (actual), **precision** = 20/40=50%

Experimentation and Prediction Techniques (3.1.7/3)

Evaluation Metrics

- **Recall:** Measures the proportion of actual churned customers correctly identified by the model

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

- **Example**: Out of 30 actual churned customers, if the model correctly identified 20, **recall** = 20/30=67%
- **F1-Score:** A balance between precision and recall, useful when both false positives and false negatives are costly.

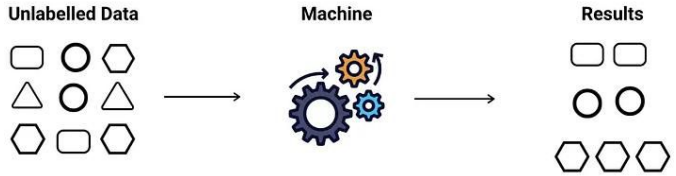
$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **Example**: If precision = 50% and recall = 67%, then **precision** =0.57%
- **This helps evaluate the trade-off between catching more churned customers (recall) versus avoiding false alarms (precision).**

Experimentation and Prediction Techniques (3.2.1/3)

Unsupervised machine learning

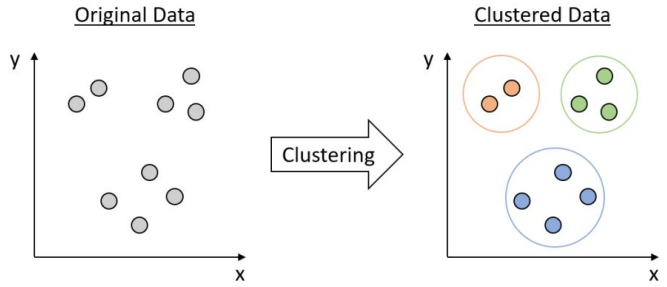
- The model is trained on data without explicit labels or predefined outcomes
- Focuses on identifying patterns, relationships, or structures within the data itself
- This approach is particularly useful when we want to explore the underlying structure of the data or when labeled data is not available
- Common applications of unsupervised learning include clustering, association, dimensionality reduction, and anomaly detection



Experimentation and Prediction Techniques (3.2.2/3)

Clustering

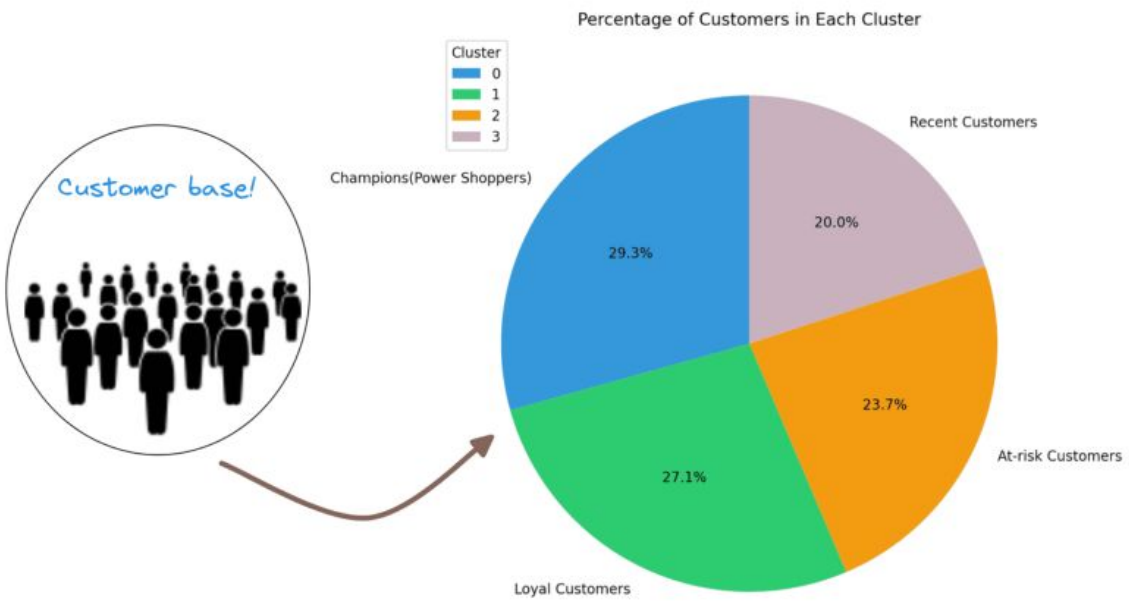
- **Group similar data points together based on their features or characteristics. The goal of clustering is to organize data into clusters, where data points within the same cluster are more similar to each other.**
- **Eg. Segment customers into different groups based on their purchasing behavior, allowing businesses to tailor their strategies to each specific segment**
- **Common clustering algorithms include**
  - K-means, hierarchical clustering, and DBSCAN



Experimentation and Prediction Techniques (3.2.3/3)

Clustering : Use case: Customer segmentation

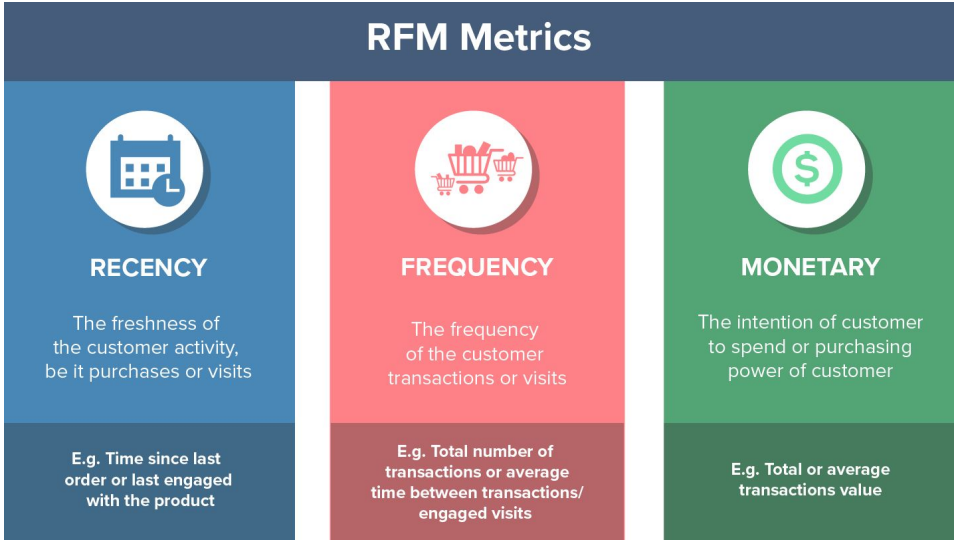
By dividing customers into segments based on specific criteria, such as buying behavior, demographics, or engagement levels, businesses can tailor their marketing efforts, products, and services to meet the unique needs and preferences of each group.



Experimentation and Prediction Techniques (3.2.4/3)

The RFM (Recency, Frequency, Monetary)

- **Method is widely used for customer segmentation**
- **Categorizes customers into segments according to three factors:**
  - How recently they made a purchase
  - How often they purchase
  - How much money they spend



Experimentation and Prediction Techniques (3.2.4/3)

The RFM (Recency, Frequency, Monetary)

- **Recency (R):** Measures how recently a customer made their last purchase. Customers who have purchased recently are more likely to engage with new offers and promotions. The more recent the purchase, the more likely the customer is to return.
- **Frequency (F):** Measures how often a customer makes a purchase within a certain period. Frequent purchasers are typically more loyal and engaged. High-frequency customers are valuable because they consistently contribute to revenue.
- **Monetary (M):** Measures how much money a customer spends on purchases. Customers who spend more are often considered more valuable, and businesses may want to focus on retaining them by offering premium services or rewards.

Experimentation and Prediction Techniques (3.2.5/3)

The RFM (Recency, Frequency, Monetary)

- **Customer segmentation using the RFM method was traditionally done manually with rule-based approaches, which were time-consuming, prone to human error, and often subjective.**
- **By leveraging machine learning, this task can now be automated, allowing for the efficient analysis of large datasets, identification of patterns, and customer segmentation without manual effort.**
- **This approach not only saves time but also ensures greater consistency and accuracy in the segmentation process**

More information：Customer Segmentation in Python: A Practical Approach - KDnuggets

Introduction

- Data Science Introduction
- Data Science workflow
- **Data Science tools and roles**

Data Science Roles and Tools

Data science is a multidisciplinary field that combines various roles

- Each of these roles has distinct responsibilities, requiring specialized skills and expertise, yet they work together to drive data-driven decision-making and innovation in organizations

Data Engineer

Data Analyst

Data Scientist

Machine Learning Scientist

Data Science Roles and Tools (1/4)

Data Engineer : Roles

- Build and maintain data infrastructure for efficient collection, storage, and access
- Design and manage databases, data lakes, and warehouses
- Integrate structured, semi-structured, and unstructured data sources
- Develop ETL (Extract, Transform, Load) pipelines to clean, standardize, and transform data
- Prepare data for analysis and modeling by Data Scientists & Analysts

Data Science Roles and Tools (1.1/4)

Data Engineer : Tools

- **Databases:** Data engineers are proficient in data storage and management tools such as SQL and NoSQL databases which they use to store and organize the data
- **Programming Languages:** They use one or more programing language such as Java, Scala, Python to process the data and building data pipelines. The are also familiar with shell scripting for automating routine tasks, managing data workflows, and interacting with operating systems
- **Big data:** Now more than ever, data engineers must be proficient in handling and processing vast amounts of data using big data frameworks like Hadoop and Apache Spark, as well as leveraging cloud computing platforms such as AWS, Azure, Huawei, and Google Cloud

Data Science Roles and Tools (2/4)

Data Analyst: Roles

- Transform raw data into actionable insights for decision-making
- Analyze, interpret, and visualize data to identify trends and patterns
- Work with cleaned and organized data prepared by Data Engineers
- Create reports, dashboards, and visualizations for stakeholders
- Enable data-driven decisions through meaningful business intelligence

Data Science Roles and Tools (2.1/4)

Data Analyst: Tools

- **SQL**: Query and retrieve data for analysis and reporting
- **Spreadsheets (Excel, Google Sheets)**: Organize, analyze, and visualize small datasets
- **BI Tools (Power BI, Tableau)**: Create interactive dashboards for stakeholders
- **Programming (Python, R)**: Advanced analysis, automation, and handling large datasets

Data Science Roles and Tools (3/4)

Data Scientist: Roles

- **Core Role:** Uses advanced analytics and machine learning to extract insights and solve complex problems
- **Key Skills:** Strong background in statistics and programming
- **Workflow:**
  - Prepares and cleans data from Data Engineers
  - Conducts exploratory data analysis (EDA) to uncover patterns
  - Builds, deploys, and refines machine learning models
  - Collaborates with stakeholders to align insights with business goals

***"A data scientist is better at statistics than any programmer and better at programming than any statistician."* — Josh Wills**

Data Science Roles and Tools (3.1/4)

Data Scientist: Tools

- **Programming languages:** Data scientist must be proficient in at least Python or R during data preparation, exploring, and modeling
    - Data scientists use within these languages, libraries like Pandas, NumPy, Scikit-learn, TensorFlow, and Keras. These libraries contain reusable code for common data science tasks
- **SQL:** Similar to analysts, data scientists have strong skills in SQL to retrieve and query the data

Data Science Roles and Tools (4/4)

Machine Learning Scientist: Roles

- **Main Role: Specializes in machine learning modeling and deployment**
- **Key Responsibilities:**
  - Transforms models created by data scientists into production-ready code
  - Requires strong programming and software development skills
  - Works extensively with deep learning for tasks like image classification and chatbots
  - Focuses on the last three stages of the data science workflow, emphasizing prediction

Data Science Roles and Tools (4.1/4)

Machine Learning Scientist: Tools

- **Programming Languages:**
  - **Python**: Dominant in ML due to its versatility, libraries (TensorFlow, Keras, PyTorch, Scikit-learn), and strong community support
  - **R**: Preferred for statistical analysis, data exploration, and visualization; also used for ML experiments and statistical analysis
- **Cloud Platforms:**
  - **AWS & Microsoft Azure**: Essential for scalable computing resources needed for training deep learning models beyond local machine capabilities