## Chapter 1: Introduction to Data Science

- **Data Science**: The interdisciplinary field that uses scientific methods, algorithms, and systems to extract insights from data.
  - **Components**:
    - **Data**: Raw information collected for analysis.
    - **Computation**: Algorithms to process data.
    - **Visualization**: Graphical representation of data for insight.
  - **Applications**:
    - House price prediction (regression).
    - Fraud detection (classification).
    - Netflix recommendations (collaborative filtering).
- **Data Types**:
  - **Structured Data**: Tabular format (e.g., CSV files with rows and columns).
  - **Semi-structured Data**: Logs or JSON files.
  - **Unstructured Data**: Text, audio, video (1_Data Science - Introd...) (2_Data_Import_Preproces...).
- **Machine Learning (ML)**:
  - **Supervised Learning**: Models are trained on labeled data (e.g., regression, classification).
  - **Unsupervised Learning**: Models identify patterns without labeled data (e.g., clustering).
  - **Common tasks**: Regression, classification, clustering, anomaly detection (4_Intro_to_ML (1)) (1_Data Science - Introd...).

## hapter 2: Data Import and Preprocessing

- **Data Preprocessing Steps**:
  1. **Handling Missing Data**:
     - **Remove instances** (rows) or features (columns) with missing values.
     - **Imputation**: Replace missing values with a constant (e.g., mean, zero, random value).
  2. **Encoding Categorical Variables**:
     - **One-Hot Encoding**: Convert categorical variables into binary columns. For example, a feature "color" with values "red", "green", "blue" becomes three binary features.
     - **Label Encoding**: Assigning integers to categorical values (e.g., "red" = 1, "green" = 2).
  3. **Scaling**:
     - **Min-Max Normalization**: Scales data to a range of [0, 1].

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

     - **Z-Score Normalization**: Centers data around 0 with a standard deviation of 1.

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

     - **Why Normalize?**: Algorithms like KNN and SVM require features to be on the same scale (2_Data_Import_Preproces...) (5_Feature_Selection (1)).

## Chapter 3: Data Visualization

- **Types of Variables**:
  - **Qualitative (Categorical)**: Categories without numerical meaning (e.g., gender, color).
  - **Quantitative (Numerical)**: Data that can be measured and has meaning in terms of magnitude (e.g., house prices, age).
- **1D Plots**:
  - **Bar Plots**: Used for categorical variables (e.g., gender counts).

$$\text{Bar(height} = \text{frequency of the category)}$$

  - **Histograms**: Used for numerical variables to show the frequency distribution.

$$\text{Frequency of values in specified bins}$$

- **2D Plots**:
  - **Scatter Plots**: Displays two quantitative variables.

$$\text{Point(x} = \text{feature 1, y} = \text{feature 2)}$$

  - **Heatmaps**: For categorical x categorical relationships, showing intensity of relationships.
- **3D Plots and Beyond**:
  - **Scatter Matrices**: Pairwise scatter plots for visualizing multi-dimensional data (3_Data_Visualization (1)).

## Chapter 4: Machine Learning Basics

- **Supervised Learning**:
  - **Regression**: Predicts continuous values (e.g., house prices).
    - **Key Algorithm**: Linear Regression (details in Chapter 6).
  - **Classification**: Predicts discrete categories (e.g., loan approval, cancer diagnosis).
    - **Key Algorithms**: Logistic Regression, K-Nearest Neighbors (KNN), Decision Trees (4_Intro_to_ML (1)).
- **Unsupervised Learning**:
  - **Clustering**: Identifies groups in data without predefined labels (e.g., K-means).
    - **K-Means**: Partitions data into $k$ clusters by minimizing the within-cluster variance.
  - **Dimensionality Reduction**: Techniques like PCA (Principal Component Analysis) reduce the number of features by projecting data into lower dimensions (4_Intro_to_ML (1)).

## Chapter 5: Feature Selection

- **Why Feature Selection?**:
  - Improves model performance by removing irrelevant or redundant features.
  - **Common Issues**: Too many features can lead to overfitting or slower training times (5_Feature_Selection (1)).

**Numerical Feature Selection**:

- **Pearson Correlation Coefficient**:

$$r = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

  - Measures the linear relationship between two variables.
  - $r = 1$: Perfect positive correlation, $r = -1$: Perfect negative correlation, $r = 0$: No correlation (5_Feature_Selection (1)).

**Categorical Feature Selection**:

- **Chi-Square Test**:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

  - Compares observed and expected counts. A large chi-square value means there's a relationship between the variables (5_Feature_Selection (1)).

$$\frac{\text{Row total} \times \text{Column total}}{\text{Grand total}} = E$$

## Chapter 6: Linear Regression

- **Key Concept**: Models a linear relationship between a dependent variable $y$ and an independent variable $x$.
  - **Hypothesis Function**:

$$\hat{y} = \theta_1 \cdot x + \theta_2$$

    - $\theta_1$ is the slope, and $\theta_2$ is the intercept (6_Linear_Regression (1)).
- **Cost Function (Mean Squared Error)**:

$$J(\theta_1, \theta_2) = \frac{1}{\text{[redacted]}m} \sum_{i=1}^{m} (\hat{y}^{(i)} - y^{(i)})^2$$

  - Measures how well the line fits the data. Minimizing this function helps find the best $\theta_1$ and $\theta_2$ (6_Linear_Regression (1)).
- **Gradient Descent Algorithm**:
  - **Goal**: Minimize the cost function by iteratively updating the parameters.

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_1, \theta_2)$$

  - $\alpha$: Learning rate (controls step size).
  - **Update Rule**:
    - For $\theta_1$: $\theta_1 := \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} (\hat{y}^{(i)} - y^{(i)}) x^{(i)}$
    - For $\theta_2$: $\theta_2 := \theta_2 - \alpha \frac{1}{m} \sum_{i=1}^{m} (\hat{y}^{(i)} - y^{(i)})$ (6_Linear_Regression (1)).
- **Convergence**: When the changes in $\theta_1$ and $\theta_2$ become very small, indicating that the model has found the optimal parameters (6_Linear_Regression (1)).

### 3. Pearson Correlation Coefficient

**Formula:**

$$r = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

**Explanation:**

- $r$: Pearson correlation coefficient.
- $\text{cov}(X, Y)$: Covariance between $X$ and $Y$.
- $\sigma_X$: Standard deviation of $X$.
- $\sigma_Y$: Standard deviation of $Y$.

**Usage:** Measures the strength of a linear relationship between two variables, with $r$ ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation).

### 7. Cost Function for Linear Regression (Mean Squared Error)

**Formula:**

$$J(\theta_1, \theta_2) = \frac{1}{2m} \sum_{i=1}^{m} (\hat{y}^{(i)} - y^{(i)})^2$$

**Explanation:**

- $J(\theta_1, \theta_2)$: Cost function (error).
- $m$: Number of training examples.
- $\hat{y}^{(i)}$: Predicted value for the $i$-th example.
- $y^{(i)}$: Actual value for the $i$-th example.

**Usage:** Measures how well the regression line fits the data. The goal is to minimize $J(\theta_1, \theta_2)$ by finding the best values for $\theta_1$ and $\theta_2$.

---

### 8. Gradient Descent Update Rule

**Formula:**

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

**Explanation:**

- $\theta_j$: Parameter to be updated.
- $\alpha$: Learning rate (controls the size of the update step).
- $\frac{\partial}{\partial \theta_j} J(\theta)$: Derivative of the cost function with respect to $\theta_j$.

**Usage:** Iterative optimization algorithm to minimize the cost function. Gradient descent adjusts $\theta_j$ in the direction that reduces the cost.

**Formula for Covariance:**

For two datasets $X$ and $Y$, each with $n$ data points:

$$\text{cov}(X, Y) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu_X)(Y_i - \mu_Y)$$

Where:

- $X_i$: The $i$-th value in dataset $X$.
- $Y_i$: The $i$-th value in dataset $Y$.
- $\mu_X$: The mean of the dataset $X$.
- $\mu_Y$: The mean of the dataset $Y$.
- $n$: The number of data points in each dataset (assuming both datasets have the same number of points).

**Formula:**

$$\theta = (X^T X)^{-1} X^T Y$$

**Explanation:**

- $\theta$: Vector of regression coefficients.
- $X$: Matrix of input features.
- $X^T$: Transpose of $X$.
- $(X^T X)^{-1}$: Inverse of $X^T X$.
- $Y$: Vector of actual output values.

**Usage:** Solves for the optimal $\theta$ values in one step (without using gradient descent). Often used when the number of features is small or when computational resources are abundant.

### 5. Linear Regression Hypothesis Function

**Formula:**

$$\hat{y} = \theta_1 \cdot x + \theta_2$$

**Explanation:**

- $\hat{y}$: Predicted value.
- $x$: Input value (independent variable).
- $\theta_1$: Slope (rate of change in $\hat{y}$ as $x$ changes).
- $\theta_2$: Intercept (value of $\hat{y}$ when $x = 0$).

**Usage:** Predicts a continuous output based on a single input. Linear regression finds the best values for $\theta_1$ and $\theta_2$ to minimize the prediction error.

---

### 6. Multiple Linear Regression Hypothesis Function

**Formula:**

$$\hat{y} = \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \cdots + \theta_n \cdot x_n + \theta_0$$

**Explanation:**

- $\hat{y}$: Predicted value.
- $x_1, x_2, \ldots, x_n$: Input features.
- $\theta_1, \theta_2, \ldots, \theta_n$: Coefficients (weights) for each input feature.
- $\theta_0$: Intercept (bias term).

**Usage:** Generalizes linear regression to multiple features. The model predicts $\hat{y}$ as a weighted sum of the input features.

## Cheat Sheet: Data Science and Machine Learning Concepts

### 1. Min-Max Normalization

**Formula:**

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

**Explanation:**

- $x$: The original value.
- $x_{\min}$: The minimum value in the dataset.
- $x_{\max}$: The maximum value in the dataset.
- $x'$: The normalized value, rescaled between 0 and 1.

**Usage:** Rescales data to a range of [0, 1]. Used when features have different ranges and need to be comparable.

---

### 2. Z-Score Normalization (Standardization)

**Formula:**

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

**Explanation:**

- $x$: The original value.
- $\mu$: The mean of the dataset.
- $\sigma$: The standard deviation of the dataset.
- $z$: The standardized value.

**Usage:** Standardizes data to have a mean of 0 and a standard deviation of 1. Useful for algorithms sensitive to feature scales (e.g., SVM, KNN).