**SOC 2025 Report**

**24B4222**

**Intro to ML**

**Week 1:**

Github: Complete Git and GitHub Tutorial for Beginners (YT vid)

- Intro to Programming: https://www.kaggle.com/learn/intro-to-programming

- Python: https://www.kaggle.com/learn/python

- Pandas: https://www.kaggle.com/learn/pandas

- Intro to Machine Learning: https://www.kaggle.com/learn/intro-to-machine-learning

Certificates to be uploaded on GitHub.

**Week 2:**

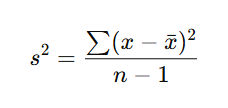
**Statistics Report: Variability, Correlation, Regression & Probability**

**Based on Chapters 4.1–4.6, 6, 7, and 8.7–8.10 from Witte & Witte (11th Ed.)**

**1. Describing Variability**

**Goal**: Measure how much individual data points differ from the average.

* **Range**:
  + Formula: *Max – Min*
  + Quick but overly sensitive to outliers
  + Example: For weights 130 lbs to 180 lbs, range = 50 lbs
* **Variance (s²)**:
  + Average of squared deviations from the mean
  + Sample formula:



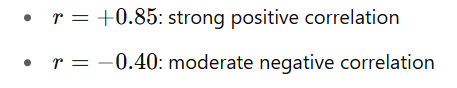
* **Standard Deviation (SD)**:
  + Square root of variance
  + Interpreted in the same units as the data
  + Example: If variance = 16, SD = 4
* **Degrees of Freedom (df)**:
  + Use n−1 instead of n for samples
  + Adjusts for bias in estimating population spread

SD is key in normal curve, error measurement, and inferential statistics.

**2. Correlation – Describing Relationships**

**Goal**: Identify if and how two variables move together.

* **Scatterplot**:
  + Visual graph to detect linear trends
  + Example: Study hours vs. exam score
* **Pearson Correlation Coefficient (r)**:
  + Measures strength & direction of linear relationships
  + Range: -1 (perfect negative) to +1 (perfect positive)
  + Example:

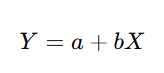


* **Outliers**:
  + Can distort r value significantly
  + Must be examined and reported
* **Other Correlations**:
  + Spearman’s rho for ranked/ordinal data

**Use**: Indicates strength of association; leads to prediction using regression.

**3. Regression – Predicting Scores (Chapter 7)**

**Goal**: Use one variable (X) to predict another (Y)

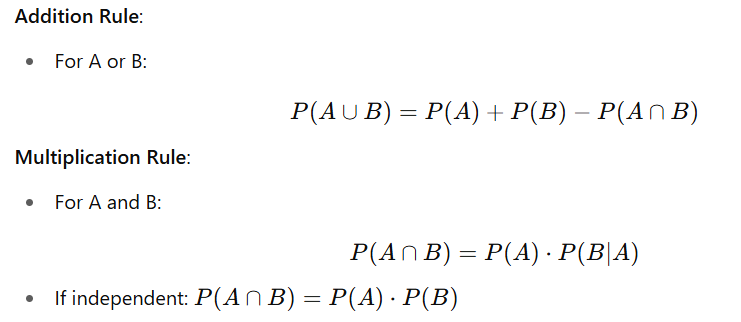
* **Regression Equation**:
  + Formula:
    - *a*: Y-intercept
    - *b*: slope (change in Y per unit of X)
* **Least Squares Method**:
  + Minimizes total squared prediction error
* **Standard Error of Estimate (Sy|x)**:
  + Measures accuracy of prediction
  + Lower Sy|x = better prediction
* **Coefficient of Determination (r²)**:
  + Proportion of variance in Y explained by X
* **Multiple Regression**:
  + Includes multiple predictors (X₁, X₂, …)
* **Regression Toward the Mean**:
  + Extreme scores tend to move toward average on repeated measurement

**Use**: Powerful for forecasting, trend analysis, and effect size estimation.

**4. Probability – Foundations for Inference**

**Goal**: Calculate how likely an event is and use that to test hypotheses.

* **Definition of Probability**:
  + Long-run frequency of an event
  + Range: 0 to 1 (or 0% to 100%)

B:

**Statistical Inference Link**:

* + Uses probability to decide if results are due to chance