General

Collapse all

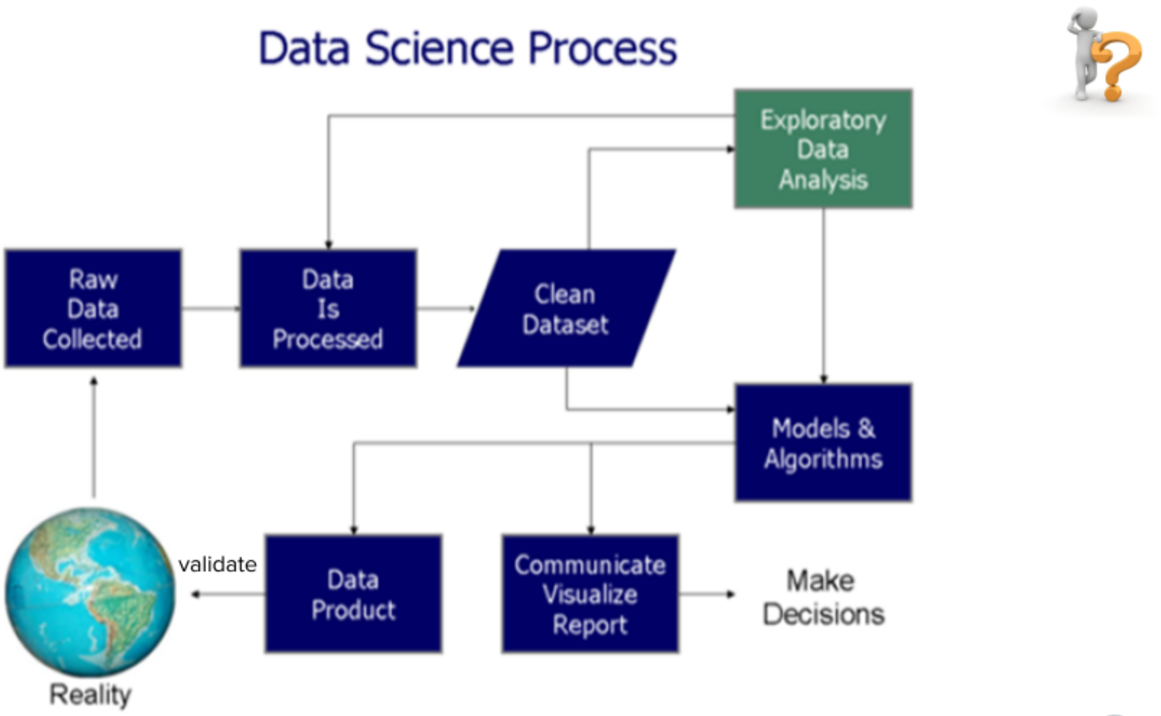
FORUM
Announcements

Course Logistics

Course Logistics

- *Teacher*: Dr. Mahesh Mohan M R
- Ph.D Teaching Assistants: Ashraf, Shubhajit
- *References*
  - Data Science Design Manual, Steven S Skiena, Springer (available online)
- *Theory*
  - Monday - 10:00 -10:55 AM, Wednesday – 8:00-9:55 AM, Thursday – 10:00-10:55 AM
  - Evaluation: Class Test-1, Midsem, Class Test-2, and Endsem (10%+20%+10%+30%)
- *Programming*
  - Programming Assignments: 15%
  - Mini Project: 10% (Team of three, even marking)
- Class Participation: 5% (Attendance compulsory to avoid deregistration.)
- Office Hours: Fridays 5-6 PM based on appointments (mail to mahesh.mohan@cai.iitkgp.ac.in)
- Plagiarism: No tolerance policy. Binary marking (both parties).

An Overview of the Course



FILE
Course Logistics

Introduction to AI and Data Science

Introduction to AI and Data Science

- When AI looks like a hammer, everything becomes a nail? True of False.
- Is AI a Gen Z Invention? Why Gen Z got the AI boom?
- Which kind of problems are AI effective? (Rule-based approach vs Learning based approach)
- What all subjects come together to form Data Science? (CS, Math or Statistics, and Common Sense 😀)
- Difference between Data Science and Computer Science?
- Moto of Data Scientist: The purpose of computing is insight, not numbers
- Big Data and Three Fundamental Questions
- Difference between Hypothesis Driven and Data Driven Science?

FILE
Intro to DS

Reading Exercise: Chapter 1 (The Data Science Design Manual)

Tutorial: From The Data Science Manual; Qns 1-1, 1-3, 1-4,1-6, 1-7, 1-12, 1-13, 1-14, 1-15

Probability and Random Variables

1. Uncertainity in Data Science
2. Probability Vs Statistics
3. Axioms of Probability
4. Random Variable (Discrete and Continuous and Random variables)
5. Probability Functions (Probability Mass Function and Probability Distribution Function)
6. Multivariate Random variable
7. Joint Probability, Marginal Probability and Conditional Probability
8. Baye's Theorem and Independence
9. Expected Value, Variance, Covariance, and Covariance matrix
10. Standard Probability Distribution

FILE
Probability

Reading Exercise: Chapter 2, Sec 2.1 (The Data Science Design Manual)

Tutorial: From The Data Science Manual; Qns 2-1, 2-2, 2-3, 2-5

Vectors, Matrices and their Manipulations

- Scalar Vs Vector Vs Matrix
- Scalar Vector product
- Scalar Matrix Product
- Matrix Multiplications (Inner product and Outer product interpretation)
- Linear Network explained
- Inner product - Explain individual single output node
- Outer product - Explain full output nodes at once

FILE
Linear_algebra

Reading Exercise: Chapter 8, Sec 8.1-8.2 (The Data Science Design Manual)

Tutorial: From The Data Science Manual; Qns 8-1, 8-2, 8-3, 8-4, 8-10

ASSIGNMENT
Program Assignment 1

Opened: Saturday, 8 February 2025, 12:00 AM
Due: Tuesday, 11 March 2025, 12:00 AM

Chapter 8, Qn 8-18

Basic Predictive Models

Linear Perceptron

- Perceptron
- History
- Pre-inner product Interpretation
- Post-inner product Interpretation
- Optimization
  - Gradient descent
    - Batch Gradient Descent
    - Stochastic Gradient Descent
    - Minibatch Gradient Descent

K Nearest Neigbour

- Instance based vs Model based Learning
- Local vs Global Approximation of Target

  - K Nearest Neighbor for Classification and Regression
  - Effect of K (less K -- > sensitive to noise and large K --> sensitive to possibly irrelevant inputs)
  - Weighted KNN for Classification and Regression
  - Issues of KNN
    - Measurement Scales (Solution: min-max or z-score normalization)
    - Curse of Dimensionality
    - Expensive and Storage Need

FILE
Linear Regression

Reading Exercise: Chapter 9: Sec 9.1, 9.2 (The Data Science Design Manual)

Tutorial: From The Data Science Manual; Qns 9.1, 9.2, 9.15, 9.16, 9.17

FILE
K Nearest Neighbor

Reading Exercise: Chapter 10: 10.1, 10.2 (The Data Science Design Manual)

Tutorial: From The Data Science Manual; 10.1, 10.2, 10.3, 10.4. 10.7, 10.8, 10.9, 10.10, 10.11, 10.27, 10.30, 10.31

Statistical Analysis of Data (Central tendency, Variance, and Correlation)

1. Centrality measures -- Mean, Median, Geometric Mean, Mode
2. Merits and Demerits of Individual Centrality measures
3. Variability measures -- Variance
4. Effect of Outliers in Centrality and Variability Measures
5. Interpreting Variance
6. Correlation Analysis -- Pearson Correlation Coefficient
7. Correlation does not imply Causation

Reading Exercise: Chapter 2, Sec 2.2-2.3 (The Data Science Design Manual)

Tutorial: From The Data Science Manual; Qns 2-6, 2-8, 2-9, 2-10, 2-11

Tutorial Discussion

FILE
tutorial_discussion

Languages of Data Science

Computers and Programming

- Hardware Basics
- Programming Language
  1. Machine Language
  2. Assembly Language --> Assembler
  3. High Level Language --> Compiler Vs Interpreter
- Programming Languages for Data Science
  1. Python
  2. C++/Java Script
  3. Perl
  4. R
  5. Matlab/Octave
  6. Excel
- Python IDEs (Jupyter Notebook) and Importance of Notebook Environments
- Introduction to Python for Data Science (Practical)
  1. Numpy
  2. Pandas
  3. Matplotlib
  4. Scikit Learn

Different Data and Standard Data Formats

- Kinds of Databases: Unstructuresd, Semistructured, Structured
- Useful properties of Data formats
- Standard Formats
  1. Tabular: CSV, TSV, XML, JSON, SQL, Protocol buffers
  2. Image Data: Sampling and Representation (Gray scale and Color); TIFF, Bitmap, JPEG, GIF, PNG, EPS, RAW image files
  3. Audio Data: Sampling and Representation (lossy vs lossless compression); WAV, AIFF, ALAC, FLAC, MP3, AAC, WMA, OGG
  4. Text Data: DOC, PAGES, DOCX, RTF, TXT, LOG, ASC, IPYNB, etc

FILE
Languages of Data Science (Part 1): Programming Language

FILE
Python slides and Code Links

FILE
Languages of Data Science (Part 2): Types of Data

ASSIGNMENT
Program Assignment 2: Predictive Models

Opened: Sunday, 16 March 2025, 12:00 AM
Due: Tuesday, 25 March 2025, 12:00 AM

?

Based on KNN Notebook (shared), design a Linear Classifier. You are supposed to do an analysis of Curse of Dimensionality (dropping some features, classify again, and analysing the results -- as done for KNN Notebook). Also, compare the performance with KNN. Pl upload a Pynb file for evaluation.

Reading Exercise: Chapter 3, Secs 3.1 (The Data Science Design Manual)

Tutorial: From The Data Science Manual; Qns 3-2 (use Python, and an arbitrary dataset with standard data format); 3-4 (use Python and create the file in CSV)

Collecting Data

Different Kinds of Data

1. Scale of measurement: Nominal, Ordinal, Interval, Ratio scales
2. Simple and Composite variables
3. Objective and Subjective measurements
4. Continuous and Discrete Measurements
5. Primary and Secondary Data

Data Collection

1. Data Collection Tools
2. Data Collection: Secondary -- Hunting
3. Data Collection: Primary -- Scraping and Logging
4. Data Collection: Primary -- Crowd-sourcing

FILE
Collecting Data

Reading Exercise: Chapter 3, Secs 3.2, 3.5 (The Data Science Design Manual)

Tutorial: From The Data Science Manual; Qns 3-5, 3-6, 3-14

Cleaning Data

Data Cleaning

1. Measures for data quality: Accuracy, Completeness, Consistency, Timeliness, Believability, Interpretability
2. Why to Clean Data?
3. Errors Vs Artifacts

Data Compatibility

1. Unit Conversion
2. Numerical Representations
3. Name Unifcation
4. Time/Date Unification
5. Financial Unification

Missing Data and Types

1. Extent of Missing Data
2. Longitudinal Vs Cross-sectional Missing data
3. Observational Vs Randomized (Missing by design, Missing completely at random, Missing at random, Missing Not at Random

Dealing with Missing Data (and when to use what)

1. Heuristic-based Imputation
2. Mean value Imputation
3. Random Value Imputation
4. Imputation by Nearest Neighbour
5. Imputation by Interpolation

Outliers and Types

1. What are outliers?
2. Impact of Outliers in Machine Learning
3. Types: Global outlier, Contextual outlier, Collective Outliers
4. Outlier Cues

Dealing with Outliers (and when to use what)

1. Statistical Method
2. Proximity-Based Method
3. Clustering-Based Method

FILE
Cleaning Data

ASSIGNMENT
Program Assignment 3: Data Cleaning

Opened: Wednesday, 2 April 2025, 12:00 AM
Due: Saturday, 12 April 2025, 12:00 AM

Perform Outlier and Missing Value Study on Prima India Diabetes dataset

https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

Reading Exercise: Chapter 3, Sec 3.3 (The Data Science Design Manual)

Tutorial: From The Data Science Manual; Qns 3-8, 3-9, 3-15, 3-16,3-17,3-18

Visualizing Data

Data Visualization Roles

1. Why to Visualize?
2. What to do when confronting a new Data Set

Data Visualization Roles

1. Showing change over time
2. Showing a part-to-whole composition
3. Depicting flows and processes
4. Looking at how data is distributed
5. Comparing values between groups
6. Observing relationships between variables
7. Looking at geographical data
8. Raw Numbers

Developing a Visualization Aesthetic

1. Maximize data-ink ratio
2. Minimizing the lie factor

    3. Minimizing Chart Junk
    4. Proper Scaling and Labelling
    5. Effective use of Color and Shading
    6. The power of repetition

Reading Graphs

    1. The obscured distribution
    2. Overinterpreting variance
    3. Interactive Visualization

FILE
visualizing data

Reading Exercise: Chapter 6, (The Data Science Design Manual)

Tutorial: From The Data Science Manual; Qns 6-3, 6-4, 6-11

Scoring Predictive Models and Statistical Significance

Why we need Scores?
Training Data Vs Validation Data vs Testing Data
When is a network doing well based on Training and Validation Accuracy?
Classification Scores -- Where to use what?

    1. Concept of TP, FP, TN, FN, and confusion matrix
    2. Percentage Accuracy
    3. Precision
    4. Recall

Regression Scores -- Where to use what?

    1. MBE
    2. MAE
    3. MSE
    4. RMSE
    5. Huber

Statistical Significance

    1. Sampling Strategies
    2. Statistical Scoring: Mean and Std-deviation based
    3. Checking the Statistical significance: T test and P values

FILE
Scoring Predictive Models

Reading Exercise: Chapter 5: 5.2, 5.3.2 (The Data Science Design Manual)

Mini-Project

An open-ended project (with your allocated team of three members):

- Your chosen data (via hunting/scraping/logging)

- Your chosen problem to solve with the data
- Your chosen statistical analysis and EDA (with appropriate visualizations)
- Your chosen case studies (with conclusions)
- Your chosen cleaning strategies (unification, outliers, missing values, etc)
- Your chosen Predictive Model(s) for your chosen problem with appropriate analysis (with and without data cleaning)
- Your chosen scoring of model(s)

Deadline: 04-05-2025
Submission: A single Python Notebook with the results and analysis. Upload in a google drive, and share that drive link in a Google form (see below).

Step 1: Team Formation Inputs: Click here

FILE
Team Allocation

Submission Link (only one submission per team): Click here

Evaluation Reports and Statistics

FILE
Class Test 1-2 Evaluations

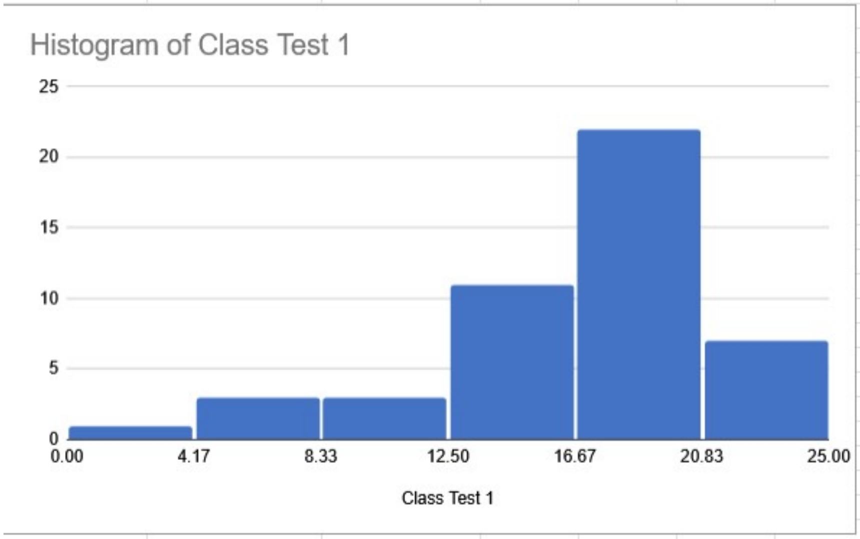
FILE
Midsem Evaluations

FILE
Program Assignments 1-3

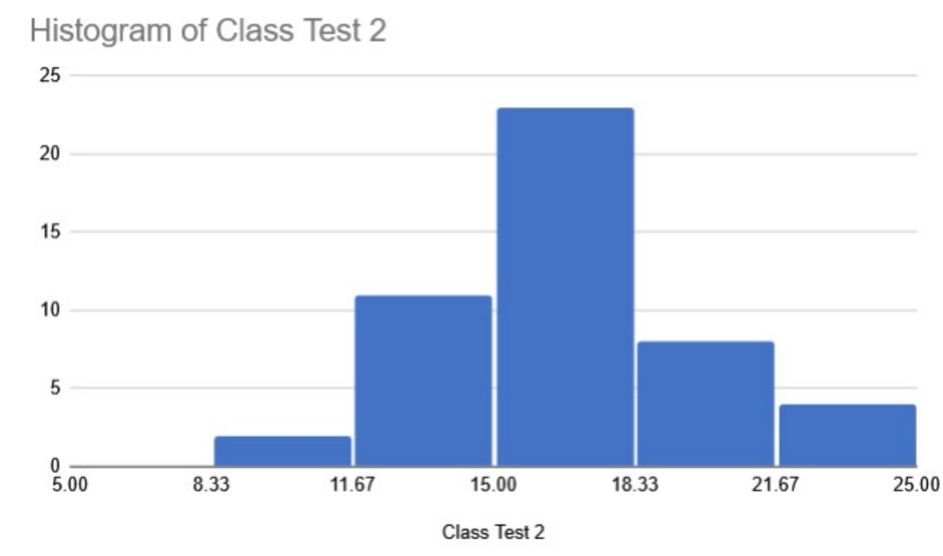
FILE
MiniProject Marks

Statistics of the Course Evaluation

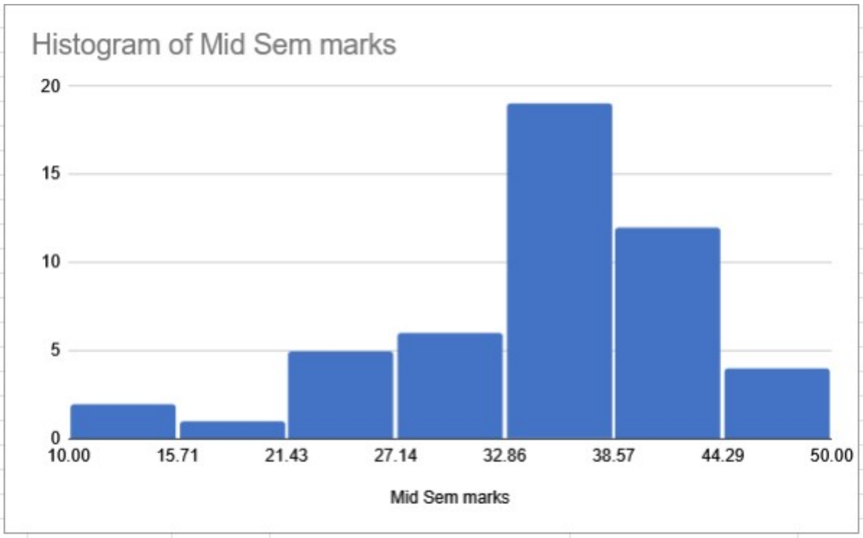
Class Test 1 (statistics)



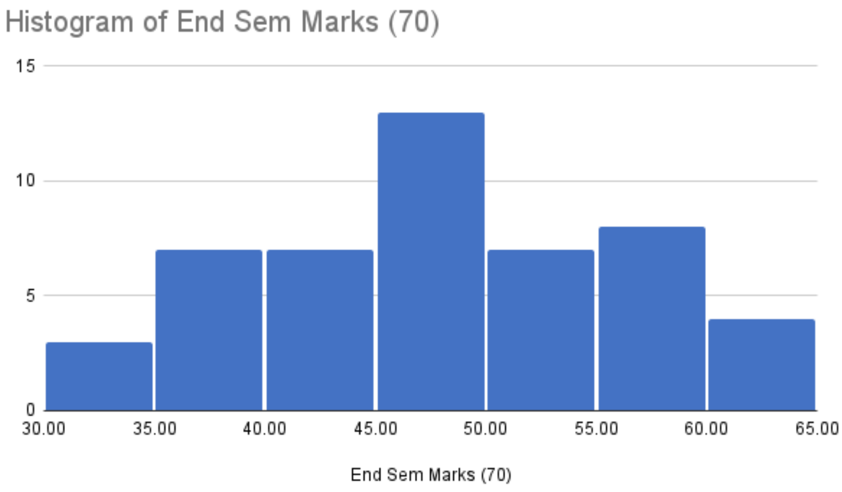
Class Test 2 (statistics)



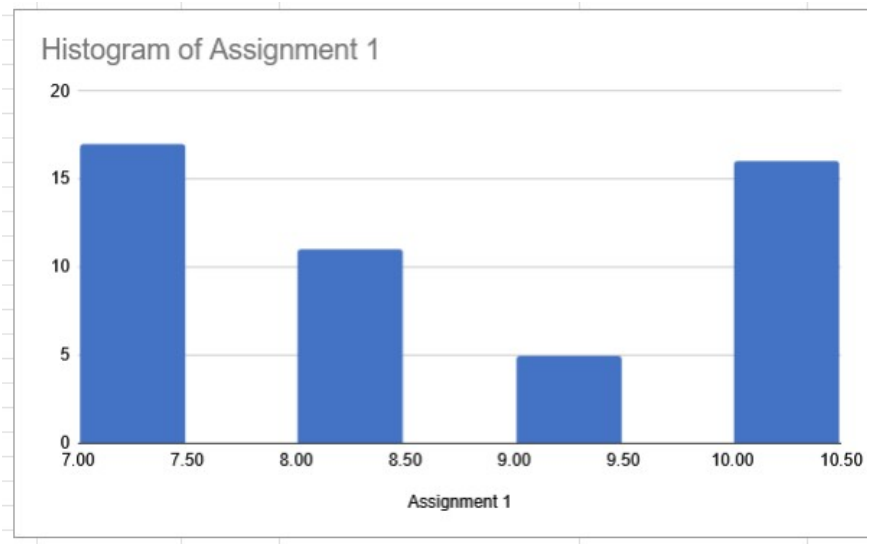
Midsem (statistics)



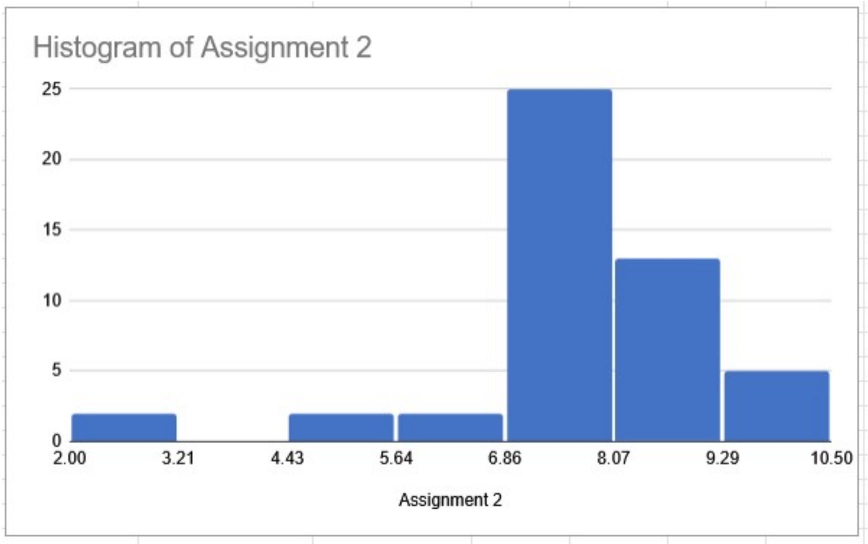
Endsem (statistics)



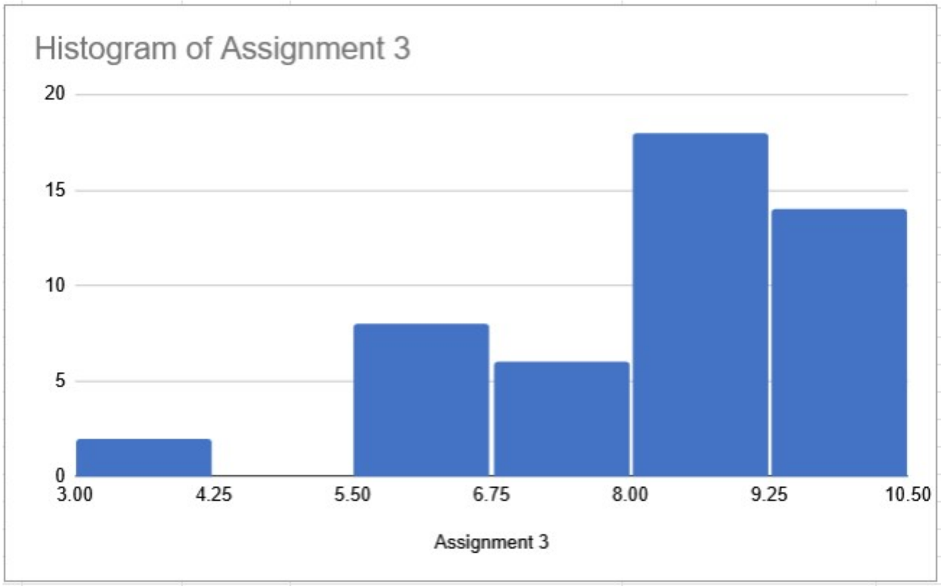
Program Assignment 1 (statistics)



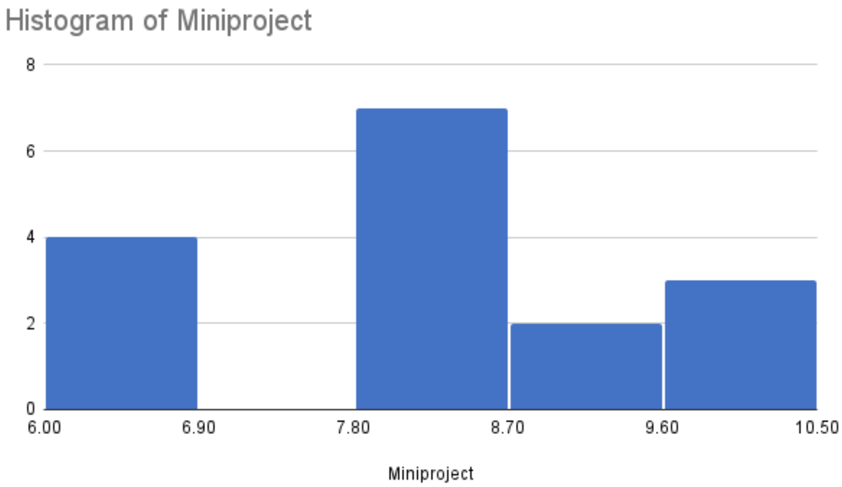
Program Assignment 2 (statistics)



Program Assignment 3 (statistics)



Mini-project (statistics - based on Teams)


Histogram of Miniproject
Miniproject