CS 381 Data Analytics Mid-Term Review Topics

- Data Science overview
  - Datawarehouse vs Transaction based database (OLTP vs OLAP), Entity relationship, Relational data modeling
  - Dataflow (ETL), Pipeline, different data types, unstructured vs structured data, different job functions of data engineers vs data scientists
- SQL
  - join tables, various aggregate functions
- Know your Statistics
  - Sampling, Sample means and standard deviation vs population means and standard deviation, Parameters vs Statistics
  - Central Limit Theorem, Hypothesis Testing, p-values
  - Type I vs Type II errors
  - Mean, Standard Deviation, Skews and Kurtosis
  - Correlation, Pearson correlation vs Spearman correlation
  - Correlation and Causation

CS 381 Data Analytics Mid-Term Review Topics

- Exploratory Data Analysis
  - Goals and typical steps, different ways to deal with missing values and outliers (bad data) removal
  - Various form of basis, Systematic Bias, Survival Bias
  - Different forms of mis-use of data visualization
- General Machine Learning principle
  - In-sample data vs out-of-sample data, training vs testing dataset
  - K-fold cross validation
  - Bias vs Variance, overfit vs underfit
- Linear Regression
  - R-squared, Adjusted-R squared, MSE, Cost function, Gradient descent
- Logistics Regression
  - Confusion matrix, Precision vs Recall, True Positive Rate, False Positive Rate
  - Sigmoid function, Odd vs Probability, Principle of maximum Likelihood