CS 381/780 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, data analysts and data scientists
- SQL
  - join tables, group by, subqueries and 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, Median, Standard Deviation, Skews and Kurtosis
  - Correlation, Pearson correlation vs Spearman correlation
  - Correlation and Causation

CS 381/780 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 bias, 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, MSE,
  - Adjusted-R squared,
  - Cost function, Gradient descent (won't cover in mid-term)
- Classification and Logistics Regression
  - Odd vs Probability

Important concepts

- large Standard Deviation means a lot of heterogeneous data (lots of difference among the data)
- Low standard deviation means data are similar, homogenous data
- High kurtosis (a lot of outliers), Low kurtosis (not much outliers)

- Positive skew means there are more data points on the high end
- Negative skew means there are more data points on the low end

- False Positive means prediction is positive, but reality is negative (Type-1 error)
- False Negative means prediction is negative, but reality is positive (Type-2 error)

- Model is built from picking a training set, it can be tailored made for that particular training data resulting in “overfit”, so it may not generalize well into the testing data set (This is low bias, but high variance case)
- Model can be too simple, resulting in high bias (ie inherently wrong), but it will have low variance when applying to various other testing data set.