Kari Palmier DePaul University Coursework Completed

- CSC 555 - Mining Big Data
  - Hadoop Apache Hive, Apache Pig, Hadoop Streaming, Apache Mahout, Apache HBase, Apache Spark, and Python used
  - Creation of multi-node cluster using Amazon AWS instances
  - Installation of Hadoop HDFS, Hive, Pig, Mahout, HBase, and Spark on multi-node cluster
  - Table creation, population, and querying (including transformations) in Hive and Pig
  - Performing MapReduce operations such as queries and joins using Hadoop Streaming with mapper and reducer code written in Python
  - Table creation and population in HBase
  - Kmeans clustering using synthetic control in Mahout
  - Matrix factorization and prediction using test and training recommender data in Mahout
  - File loading, mapping, and reducing in Spark using PySpark
- CSC 425 – Time Series Analysis
  - R used
  - Normality (skewness, kurtosis, Jacque-Bera), autocorrelation and independence (Ljung Box), and stationarity (augmented Dickey-Fuller) testing
  - Autoregressive, moving average, ARMA, and ARIMA modeling and prediction
  - Data transformation using sampling, differencing and/or return calculations to provide stationarity
  - GARCH volatility modeling
- CSC 478 - Programming Machine Learning Techniques
  - Python numpy, pandas, sklearn, and matplotlib libraries used
  - Preprocessing data with sklearn (missing value replacement, normalization, test/train splitting) and exploratory analysis including plotting and correlation
  - Sklearn decision tree, k nearest neighbors, naïve bayes, and linear discriminant analysis classification and linear, Ridge, Lasso and Stochastic Gradient Decent regression, all with cross-validation and test/train split evaluation and feature and model selection
  - Sklearn k means clustering and principal component analysis
  - Apriori association rule discovery
  - Collaborative and content based filtering recommender systems
  - Matrix factorization and singular value decomposition
  - Text classification and cluster modelling
  - Random forest and Adaboost ensemble classification
- CSC 465 - Data Visualization
  - R and Tableau used
  - Line plots, scatter plots, bar charts, treemaps, hierarchical bar charts, stacked bar and area plots, pie charts, polar plots, horizon plots, level plots, boxplots, heatmaps, map chloropleths, cartograms, mosaic plots, network graphs, forced direct graphs
- CSC 424 - Advanced Data Analysis
  - SPSS used
  - Multivariate regression, principal component analysis, factor analysis, linear discriminant analysis, K means and hierarchical clustering, canonical correlation
- IS 467 - Fundamentals of Data Science
  - SPSS used
  - Data cleaning, normalization, binning, sampling, aggregation, filtering, cross-validation, test/train split evaluation, decision tree and k nearest neighbor classification, k means and hierarchical clustering
- CSC 455 - Database Processing for Large-Scale Analytics
  - Oracle SQL and SQLite through Python sqlite3 API used
  - Relational database structure, table creation, table modification, constraints, queries, joins, views, schema normal form decompositions
- CSC 423 - Data Analysis and Regression

    - R used
    - Exploratory analysis through scatterplots, histograms, bar charts, and boxplots, correlation analysis, multiple linear regression, logistic regression feature selection
    - Regression analysis includes R squared and adjusted R squared, F test, coefficient significance, residual analysis, outlier, multicollinearity, and influential point analysis
- CSC 412 - Tool and Techniques for Computational Analytics
    - Matlab used
    - Linear algebra covering reduced row echelon form, linear systems, LU decomposition, orthogonalization, eigen vectors and values, singular value decomposition, least squares
- IT 403 - Statistics and Data Analysis
    - SPSS used
    - Descriptive statistics (max, min, standard deviation, quartiles), boxplots, histograms, normal plots, normal distribution, z scores, quartile plots, correlation analysis, linear regression analysis (including residual analysis), probability, hypothesis testing
- CSC 401 - Introduction to Programming
    - Python used
    - Data types and their methods (string, integer, float, list, dictionary, set, tuple), conditional statements, loops, functions and namespaces, try/except statements, user inputs, print statements and formatting, constructors, file I/O, recursion, classes