2022 JPMC Intro to Data Science Agenda

Duration: 3 days

Prerequisites:

- Familiarity with Excel, basic knowledge of Python and data tables
- Pre-assessment consisting of 15-20 questions that will test capability of using Python (e.g. knowledge of if statements, looping, functions, data types, string methods, etc.)

Day 1 – Intro to Data Analysis and Visualization

1. **Introduction to Data Analysis and Manipulation with Pandas**
   - Brief refresher on basics of Python and using the proposed IDE (Spyder, Jupyter or Athena)
   - Intro to Python packages
   - Intro to Pandas package (importing data, cleaning & exploring, slicing data, sorting & filtering)
   - Advanced uses of Pandas (calculated fields, apply & rolling methods, data merging)

2. **Basic Statistics and Visualization**
   - Basic statistics terminology and functions (measures of central tendency, percentiles, population vs sample, etc.)
   - Visualizations with pandas, matplotlib and seaborn
   - Histograms and boxplots with seaborn
   - Detection and removal of outliers

Day 2 – Linear and Logistic Regressions

3. **Regression Analysis with OLS of statsmodels**
   - Overview of simple regression models and OLS
   - Explanation of correlation, R-squared, p-tests and error terms
   - Multiple regression and multicollinearity
   - Case study with single factor and multiple factors regressions (e.g. CAPM model, Fama French 3 Factor Model)

4. **Logistic Regression with sklearn**
   - Overview and validation of logistic regression models
   - Multicollinearity in logistic regressions
   - Individual impact of variables
   - Confusion matrix
   - Case study (e.g. credit card approval or investor classifier)

Day 3 – Decision Trees and Model Selection

5. **Decision Trees with sklearn**
    - Overview of decision trees and key terms (segmentation, entropy, information gain)
    - Building and validating decision trees
    - Pruning, fine tuning and prediction
    - Decision trees for regression vs. classifiers
    - Case study (similar case studies to Linear and Logistic Regression to show difference in model predictions)

6. **Model Selection and Cross Validation**
    - How to validate and determine best model
    - Discussion of overfitting/underfitting data
    - Types of errors
    - Training and tuning models with hyperparameters
    - Splitting data into train/test
    - Cross validation
    - Case study (validating previous models covered in training program)