Steps in Data Science Projects

- Business Understanding
  - Determine Business Objectives
  - Assess Situation
  - Determine Data Mining Goals
  - Produce Project Plan
- Data Understanding
  - Collect Initial Data
  - Describe Data
  - Explore Data
  - Verify Data Quality
- Data Preparation
  - Integrate Data
    - Data Cutting
    - Data Discretization
    - Pre-calculated Analysis
    - Data Compression
    - Data Sorting
    - Data Merging
    - Data Grouping
    - Data Aggregation
    - ...
  - Data Cleaning
    - Noisy Data Handling
      - Business Understanding
      - Expert
    - Outliers Handling
      - Detection
        - Visualization
          - Scatter Plot
          - Joint Plot
          - Histogram
          - ...
          - **Uncertainty (For Data Exploration)**
        - Statistical
          - Z-score (Data Standardization)
          - Modified Z-Score (Using Robust Median)
          - Boxplot (Inter-Quartile Range)
          - Log Transform
          - Dixon Test
          - **Certain (Capable of Data Exploitation)**
        - Model Base
          - DBSCAN (Clustering)
          - LOF
          - Mahalanobis Metric
          - ABOD
          - Isolation Forest
          - **Can handle multi variable outliers**
      - Management
        - Delete
        - Replace
        - Imputation
        - Talk to Expert
    - Missing Data Handling
      - Mean & Median — Binary data are NOT ALLOWED
      - Mode — Binary data are NOT ALLOWED
      - Interpolate — Mean of the previous and next data of the same feature
      - Moving Average — For example: average of same 10 feature around the missing data
      - Forward Fill & Backward Fill
      - Constant Number — Given by expert
      - Zero replacement — Last Choice
      - Deleting some cases — Last choice / When one or some cases are full of missing data
      - Deleting feature — Last Choice / When almost more than 50% of data of one feature are missed
      - **Statistical**
      - KNN
      - Linear Regression
      - Logistic Regression — Categorical
      - Time Series Techniques
      - **Algorithm Base**
    - Data Duplication
  - Feature Engineering
    - Feature Selection
      - Filter
        - Correlation Coefficient Scores (Correlation Matrix)
        - Pearson correlation
        - Spearman
        - Kendall
        - Chi-squared Score
        - ANOVA
        - Mutual Information
        - Fisher Score
        - Information Gain (IG)
        - **To find any correlation and relation between features**
      - Wrapper
        - Heuristic Search
          - Sequential Forward Selection (SFS)
          - Sequential Backward Selection (SBS)
          - Sequential Forward Floating Selection (SFFS)
          - Sequential Backward Floating Selection (SBFS)
          - Breadth First Search (BFS)
          - BDS
          - **Greedy Algorithm**
        - Metaheuristic Search
          - Genetic Algorithm
          - Gray Wolf
          - Ant Colony
          - Artificial Bee Colony
          - Particle swarm optimization
          - Migrating birds optimization
          - The whale
          - ...
          - **Optimization Algorithm**
      - Embedded
        - LASSO & Ridge Regression
        - Tree-Based
    - Feature Extraction
      - PCA (Principle Component Analysis)
      - ICA (Independent Component Analysis)
      - LDA (Linear Discriminated Analysis)
      - t-SNE t-distributed Stochastic Neighbor Embedding
    - Feature Encoding
      - Binary Encoding
      - Label Encoding (Ordinally Encoding)
      - One-Hot Encoding
      - Dummy Encoding
      - **Low Ordinality**
      - Group Encoding
      - Target Encoding
      - Probability Encoding
      - Frequency Encoding
      - Hash Encoding
      - **High Ordinality**
      - Mean Encoding
      - Catboost Encoding — For Ensemble Algorithms
      - Hierarchical Encoding — Mostly used in images
    - Feature transformation & Scaling
      - Standardization
      - Normalization (Min-max Scaling)
      - Feature Discretization
      - Feature Filling & Binning
  - Unbalanced Data Handling
    - Under Sampling
    - Over Sampling
    - Weighing
    - ...
    - **Not Practical in industry**
  - Data Sampling
    - Sampling & Evaluation Techniques
      - Out of Sample Prediction
        - 80% (train) - 20% (test) --> Random
        - 60% (train) - 20% (evaluate) - 20% (test) --> Random
          Suitable for TON OF DATA
        - Unbalanced Data — Stratify Method
      - Train-Test Random Cross Validation
      - K-Fold Cross Validation --> Not Random
    - Performance Measure
    - Problems
      - Over Fitting
      - Under Fitting
  - Format Data
- Modeling
  - Select Modeling Technique
  - Generate Test Design
  - Build Model
  - Assess Model
- Evaluation
  - Evaluate Results
  - Review Process
  - Determine Next Steps
- Deployment
  - Plan Deployment
  - Plan Monitoring & Maintenance
  - Produce Final Report
  - Review Project