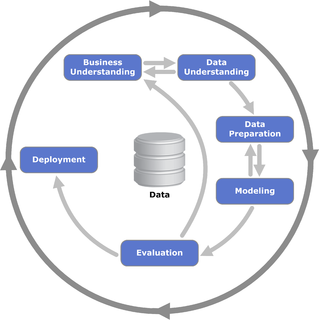
**Project Outline Template**

This is a basic outline for developing your data science portfolio project! The guide loosely follows the CRISP DM model for the data science process, per the diagram below.



1. **Select your dataset**

* This is the data you will be working with throughout your project. Make sure to select a dataset with a substantial number of data entries, preferably 800+, to make sure that your conclusions aren’t skewed by a small sample size
* Feel free to bring in additional datasets to supplement your project! Additional data may be on a related topic, and can bring more depth and clarity to a project if used well

1. **Clean & pre-process your data**

* Cleaning your data may include removing duplicate values, handling missing data values, and detecting outliers
* You may also consider transforming your data’s representation by:
  + encoding values (see: 1-hot encoding)
  + splitting or combining columns (such as dividing a DD/MM/YYYY column into individual day, month, and year columns),
  + scaling numeric columns to be more understandable in your analysis

1. **Exploratory data analysis (EDA)**

* This involves evaluating and visualizing your dataset to understand the types, spread, and relationship between various data columns, and to reveal overall trends and outliers
* Information gathered in this step is also helpful for guiding more detailed analysis
* Three main types of EDA based on number of columns:
  + Univariate analysis
    - Histograms, box plots, and bar charts
    - Summary statistics to describe central tendency and spread
  + Bivariate analysis
    - Scatter plots & correlation coefficient
    - Line graphs, especially for time series data
  + Multivariate analysis
    - Examine relationship between 2 or more feature variables
    - Pair plots (grouped scatterplots)
    - Principal component analysis (PCA) to simplify large, complex datasets
* A good EDA overview: <https://www.geeksforgeeks.org/what-is-exploratory-data-analysis/>

1. **Targeted data analysis & model building**

* This is the key stage of your project! Ideally, developing a handful (3 - 5) guiding questions for your analysis can help maintain focus in your results, and serve as foundations for specific visualizations later on
* Data visualization: <https://www.geeksforgeeks.org/techniques-for-data-visualization-and-reporting/>
  + Data visualization tools
    - Dashboarding: Tableau, PowerBI
    - Libraries: Python ( matplotlib, seaborne), R( ggplot)
* Model building: <https://www.geeksforgeeks.org/model-building-for-data-analytics>

1. **Prepare your presentation!**

* Presentations are open-ended, but can ideally focus on the goal of your analysis, and results achieved in that direction.
* Oftentimes, presentations will be targeted towards a mixed audience of both data scientists and non-technical business professionals. Feel free
* Some points to touch on include:
  + Guiding questions in your analysis process
  + Key decisions you made during model building and analysis
  + Visualizations that communicate your results to a non-technical audience