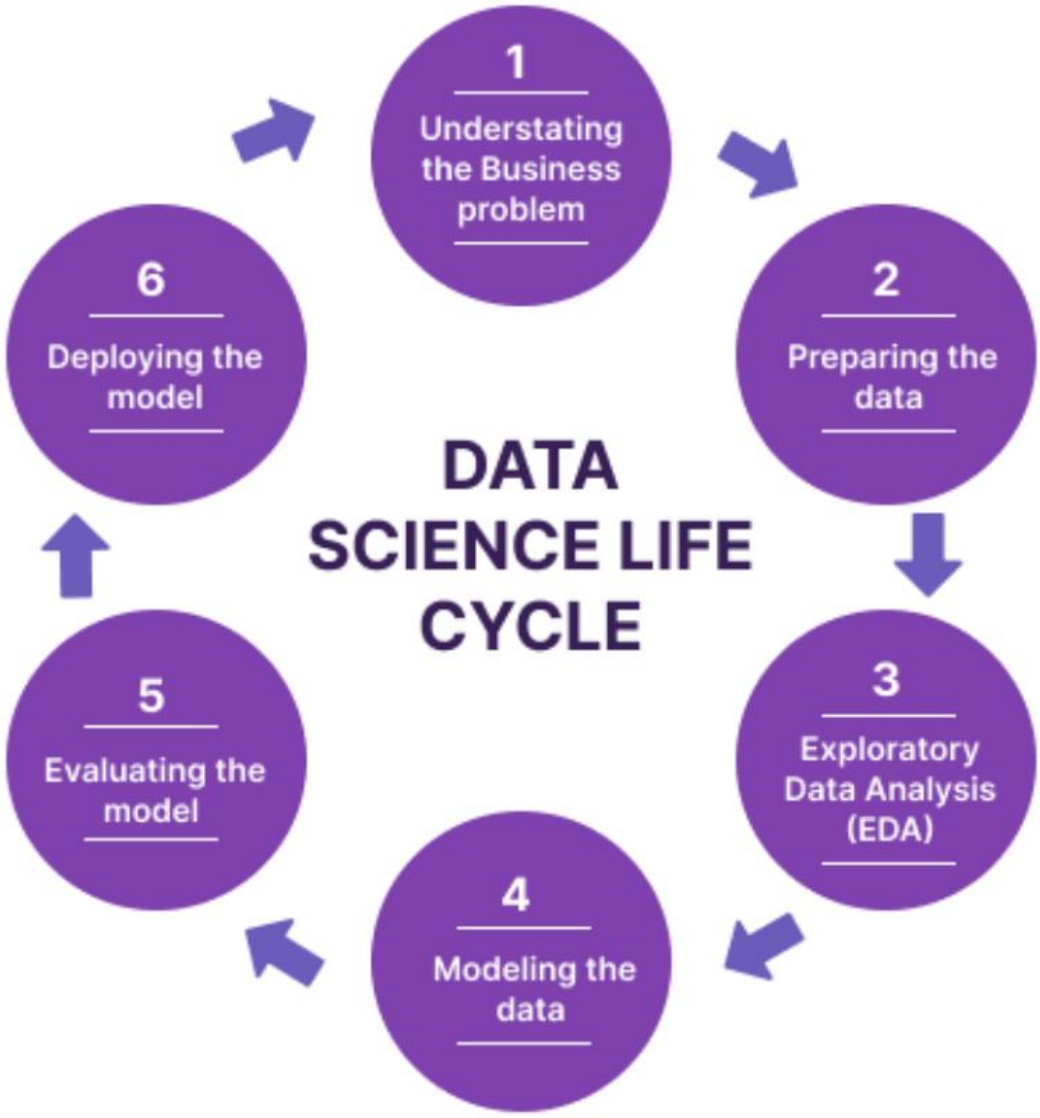
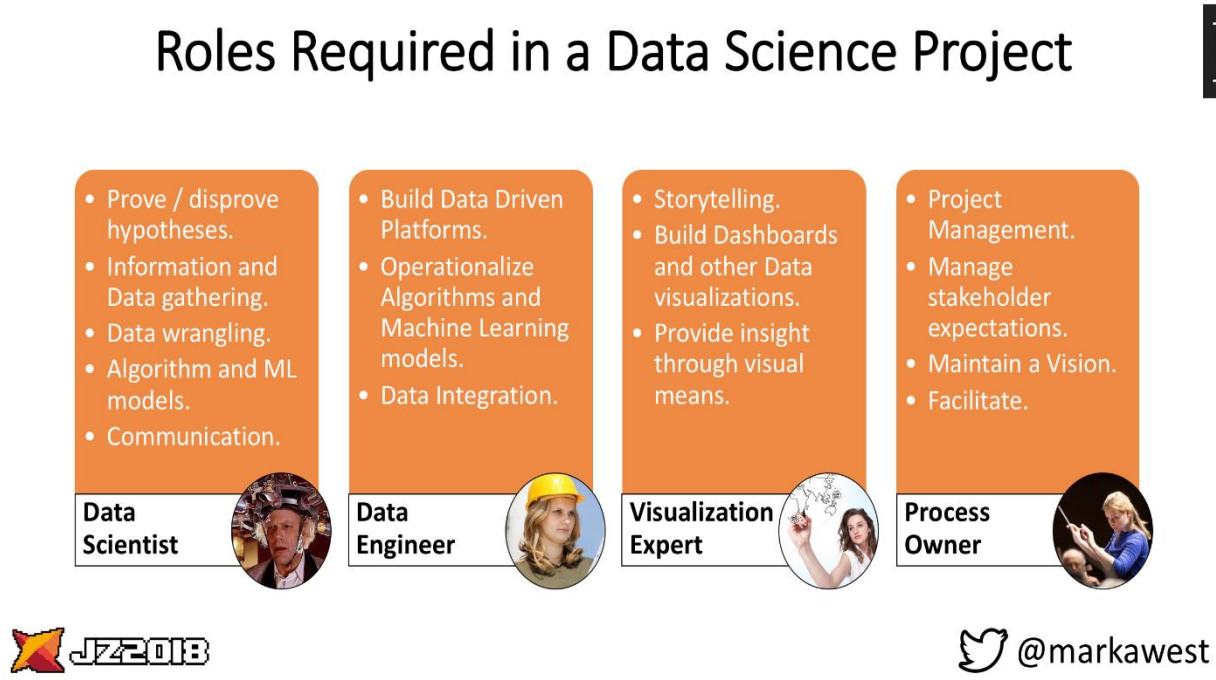
DATA SCIENCE LIFE CYCLE

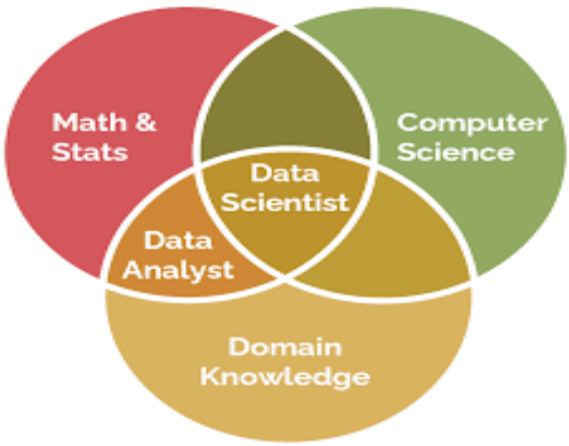
- A data science lifecycle is defined as the iterative set of data science steps required to deliver a project or analysis.
- There are no one-size-fits that define data science projects. Hence you need to determine the one that best fits your business requirements.



ROLES OF A DATA SCIENTIST

- Organize and clean the data
- Collect the dataset
- Data mining to draw the pattern
- Model selection, training and refining the task





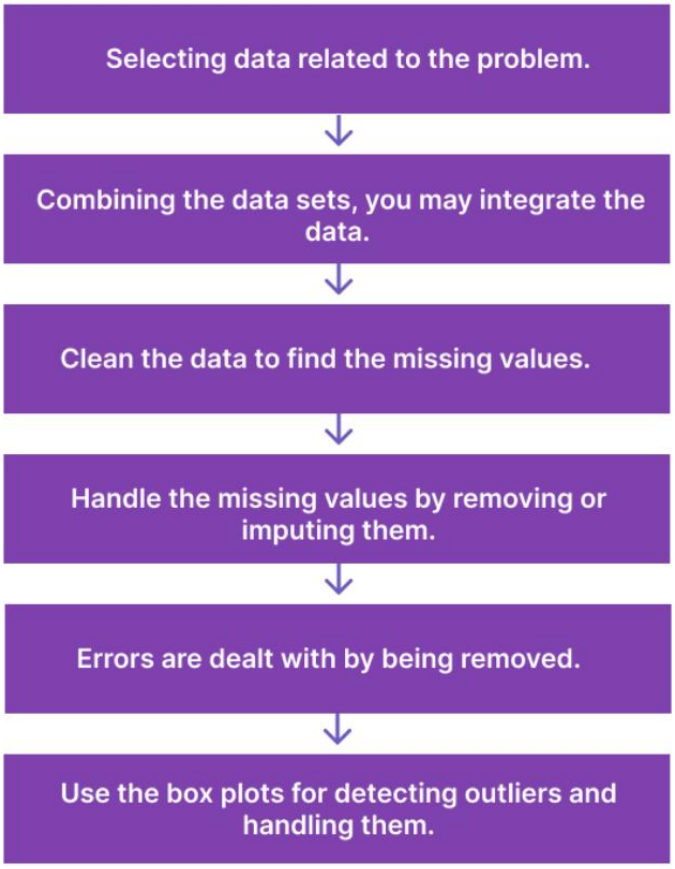
UNDERSTANDING THE BUSINESS PROBLEM

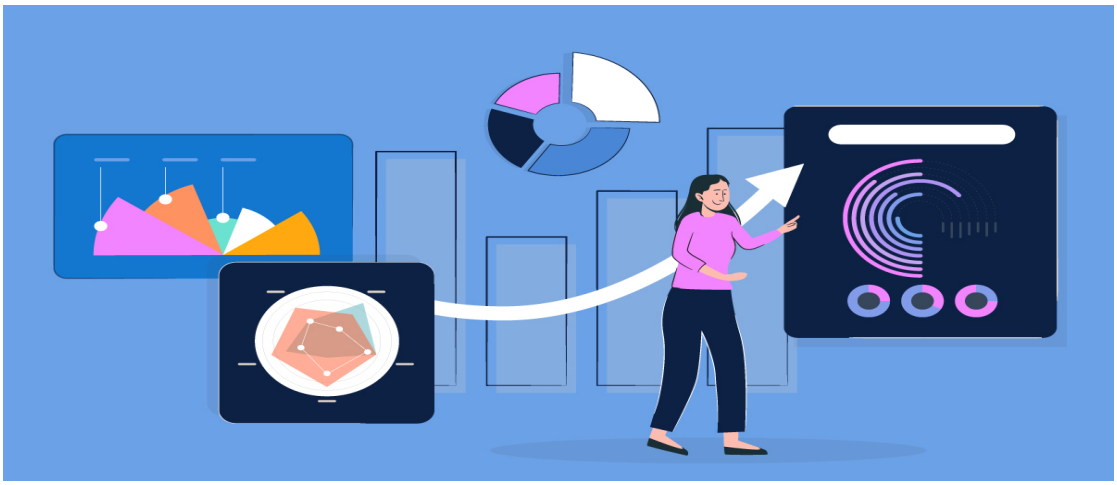
- The business objective must be clearly understood because it will be the analysis's end result.
- The first step involves looking at the business trends, developing case studies of related data analytics conducting market research on that particular industry.
- All members of the team evaluate the internal infrastructure, internal resources, the total amount of time required to complete the project, and the technological requirements for the project.
- Once all of these analysis and evaluations have been completed, the stakeholders begin developing the primary hypothesis on how to resolve all business difficulties based on the current market condition.

Hence, the following points are needed to ensure good understanding:

- List the issue that needs to be resolved.
- Define the project's potential value.
- Determine the project's risks, taking ethical issues into account.
- Create and distribute a flexible, high-level project plan.

DATA PREPARATION



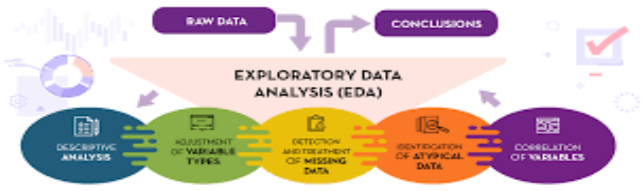


Exploratory Data Analysis (EDA)

- Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods.
- It helps determine how best to manipulate data sources to get the answers, making it easier to discover patterns, spot anomalies, test a hypothesis, or check assumptions.
- The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables.

TOOLS:

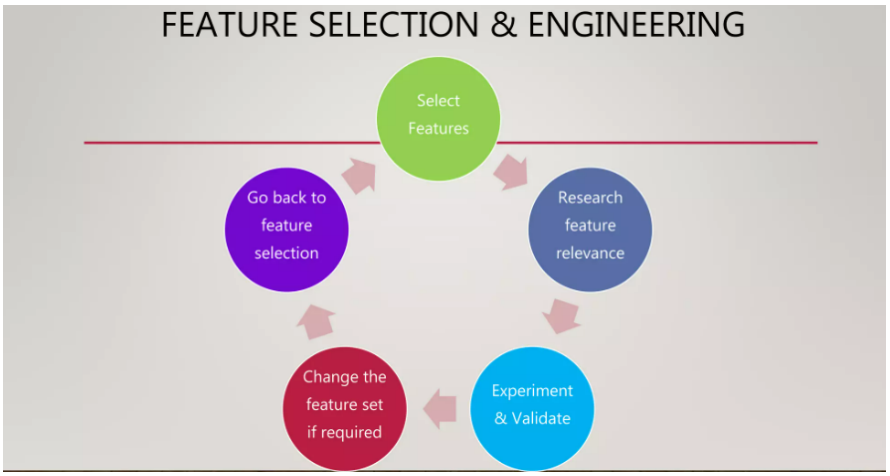
- Clustering and dimension reduction techniques, which help create graphical displays of high-dimensional data containing many variables.
- Univariate visualization of each field in the raw dataset, with summary statistics.
- Bivariate visualizations and summary statistics that allow you to assess the relationship between each variable in the dataset and the target variable you're looking at.
- Multivariate visualizations, for mapping and understanding interactions between different fields in the data.
- K-means Clustering is a clustering method in unsupervised learning where data points are assigned into K groups, i.e. the number of clusters, based on the distance from each group's centroid. The data points closest to a particular centroid will be clustered under the same category. K-means Clustering is commonly used in market segmentation, pattern recognition, and image compression.
- Predictive models, such as linear regression, use statistics and data to predict outcomes.

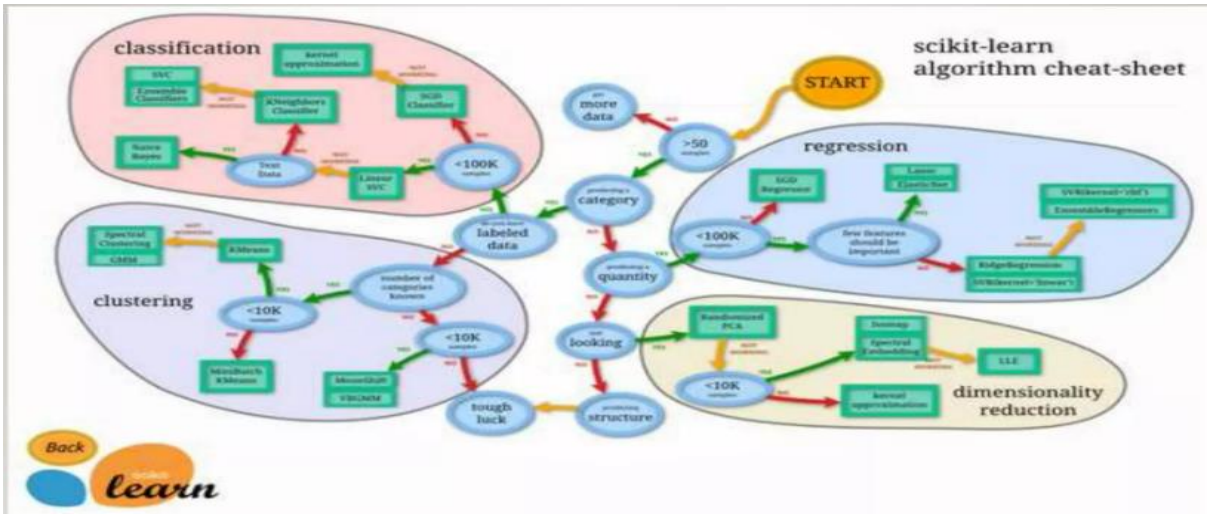


MODELING THE DATA

There are three main tasks addressed in this stage:

- **Feature engineering**: Create data features from raw data to facilitate model training.
- **Model training**: Find the model that answers the question most accurately by comparing the success metrics.
- Determine if the model is **suitable for production**.





EVALUATING THE MODEL

- **Split the input data** randomly for modeling into a training data set and a test data set.
- **Build the models** by using the training data set.
- **Evaluate** the training and test data set. Use a series of competing machine-learning algorithms along with various associated tuning parameters that are geared towards answering the question of interest with the current data.
- **Determine the best solution** to answer the question by comparing the success metrics between alternative methods.

Holdout evaluation is the process of testing a model with data that is distinct from the data it was trained on. This offers a frank assessment of learning effectiveness.

Cross-validation is the process of splitting the data into sets and using them to analyze the performance of the data. In the cross-validation procedure, the initial observation data set is divided into two sets: a training set for the model's training and an independent set for the analyses' evaluation. Both approaches use a test set (unseen by the model) to assess model performance in order to prevent over-fitting.

If the evaluation does not yield a satisfying outcome, we must repeat the modeling procedure in its entirety until the necessary level of metrics is attained.

Metrics that are used to evaluate the models are:

- Classification models:
- Accuracy
- ROC-AUC
- Precision-Recall o Log-Loss
- Regression models:
- MSAE
- MSPE
- R Square
- Adjusted R Square
- Unsupervised Models:
- Mutual Information

DEPLOYMENT

- After we have a set of models that perform well, we can operationalize them for other applications through APIs or other interface to consume various applications such as :

- Online websites
- Spreadsheets
- Dashboards
- Line-of-business applications
- Back-end applications

- The term "data science life cycle" refers to the set of stages or phases involved in a typical data science project. These stages are typically iterative, meaning that they may be repeated several times as new insights are gained or new data becomes available.

  The term "life cycle" is used to describe this process because it involves a series of interconnected steps that are intended to produce a usable data product. The term "cycle" implies that the process is ongoing, with each stage building on the previous one to improve the overall outcome.

  There is no one agreed-upon definition of the data science life cycle, but most commonly it includes stages such as problem definition, data collection and preparation, data analysis and modeling, evaluation, and deployment. By following this life cycle, data scientists can ensure that they are thoroughly analyzing and interpreting data in a systematic way, which can help them to make better-informed decisions and produce more accurate results.