Exploratory Data Analysis (EDA)

- *Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as*
- *to discover patterns,*
- *to spot anomalies,*
- *to test hypothesis*
- *to check assumptions with the help of summary statistics and graphical representations.*

- Exploratory data analysis (EDA) is the first step in the data analysis process.
- Researchers and data analysts use EDA to understand and summarize the contents of a dataset, typically with a specific question in mind, or to prepare for more advanced statistical modeling in future stages of data analysis.
- EDA relies on data visualizations that enable researchers to identify and define patterns and characteristics in the dataset that they otherwise would not have known to look for.

EDA entails the examination of patterns, trends, outliers, and unexpected results in existing survey data, and using visual and quantitative methods to highlight the narrative that the data is telling.

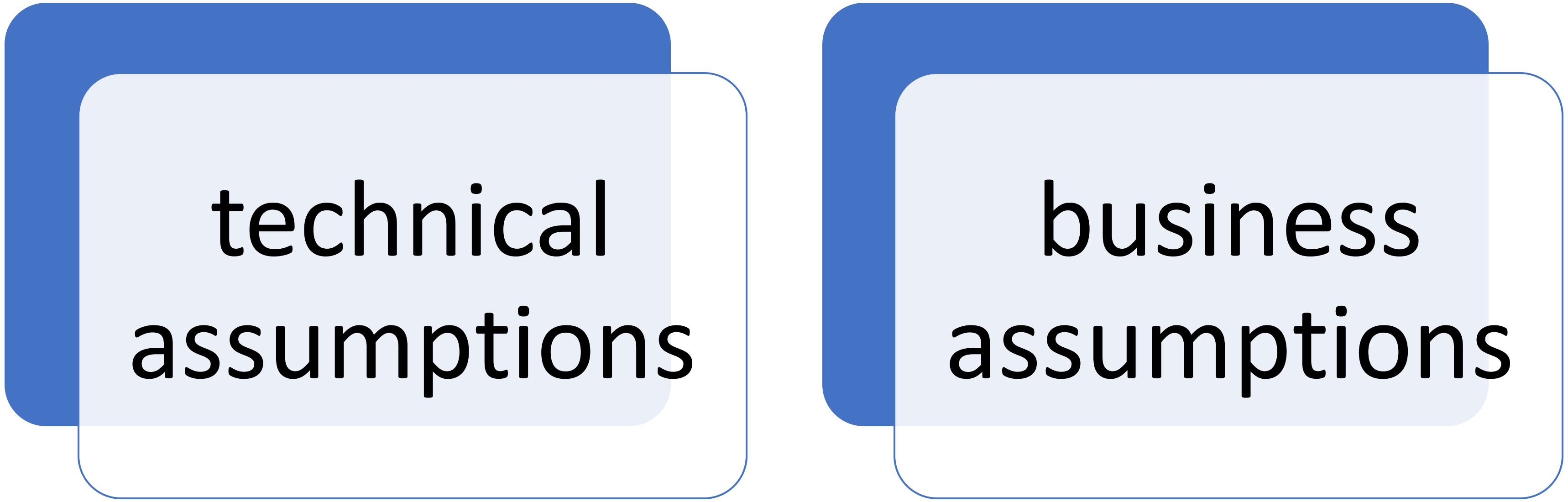
Researchers that conduct exploratory data analysis are able to:

- Identify mistakes that have been made during data collection, and areas where data might be missing.
- Map out the underlying structure of the data.
- Identify the most influential variables in the dataset.
- List and highlight anomalies and outliers.
- Test previously proposed hypotheses.
- Establish a parsimonious model.
- Estimate parameters, determine confidence intervals, and define margins of error.

The Purpose of Exploratory Data Analysis

- The primary purpose of EDA is to examine a dataset without making any assumptions about what it might contain.
- By leaving assumptions at the door, researchers and data analysts can recognize patterns and potential causes for observed behaviors.
- This ultimately helps to answer a particular question of interest or to inform decisions about which statistical model would be best to use in later stages of data analysis.

- Exploratory data analysis is used to validate technical and business assumptions, and to identify patterns



technical assumptions

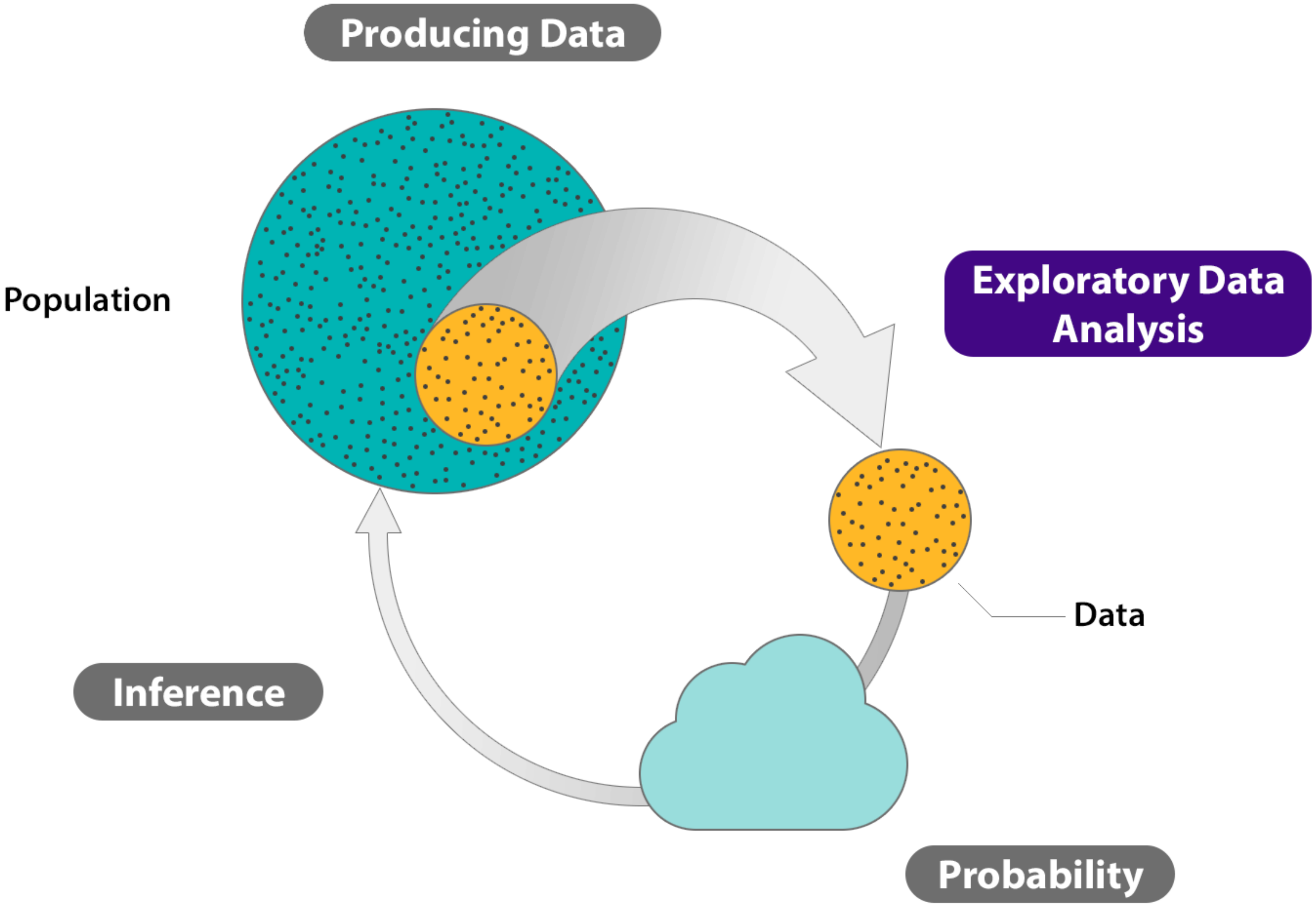
business assumptions

For example, the technical assumption that no data is missing from the dataset, or that no data is corrupted in any way, must be correct so that the insights derived from statistical analysis later on hold true.

- In order to validate and confirm the accuracy of technical and business assumptions, data scientists must systematically drill into the contents of each data field, and examine its interactions with other variables.
- By creating data visualizations, and strategically investigating those visualizations one next to the other, researchers are able to leverage the human mind’s natural skill of pattern recognition.
- Pattern recognition allows these analysts to identify potential causes of a particular behavior, highlight problematic data points, and form hypotheses that they can test to inform the decision making process when it comes to choosing a statistical model to use in future analysis of the data

- Once exploratory data analysis has been thoroughly executed, R enables researchers to perform various statistical functions, including but not limited to:
- Cluster analysis
- Univariate visualization of and summary statistics for each field in the original dataset
- Bivariate visualization and summary statistics that enable researchers to examine and assess the relationship between each of the variables in the dataset and a specific variable of interest
- Multivariate visualizations that enable researchers to uncover insight into the interactions between different fields in the data
- L-means clustering
- Predictive models, such as linear regression

- Carrying out these statistical functions allows researchers and data analysts to validate previously established assumptions and highlight patterns that will then help them to better understand the problem at hand and select a predictive model accordingly.
- By doing so, researchers can ensure high quality data analysis, and can confirm that the data has been collected and organized in the way that was expected.

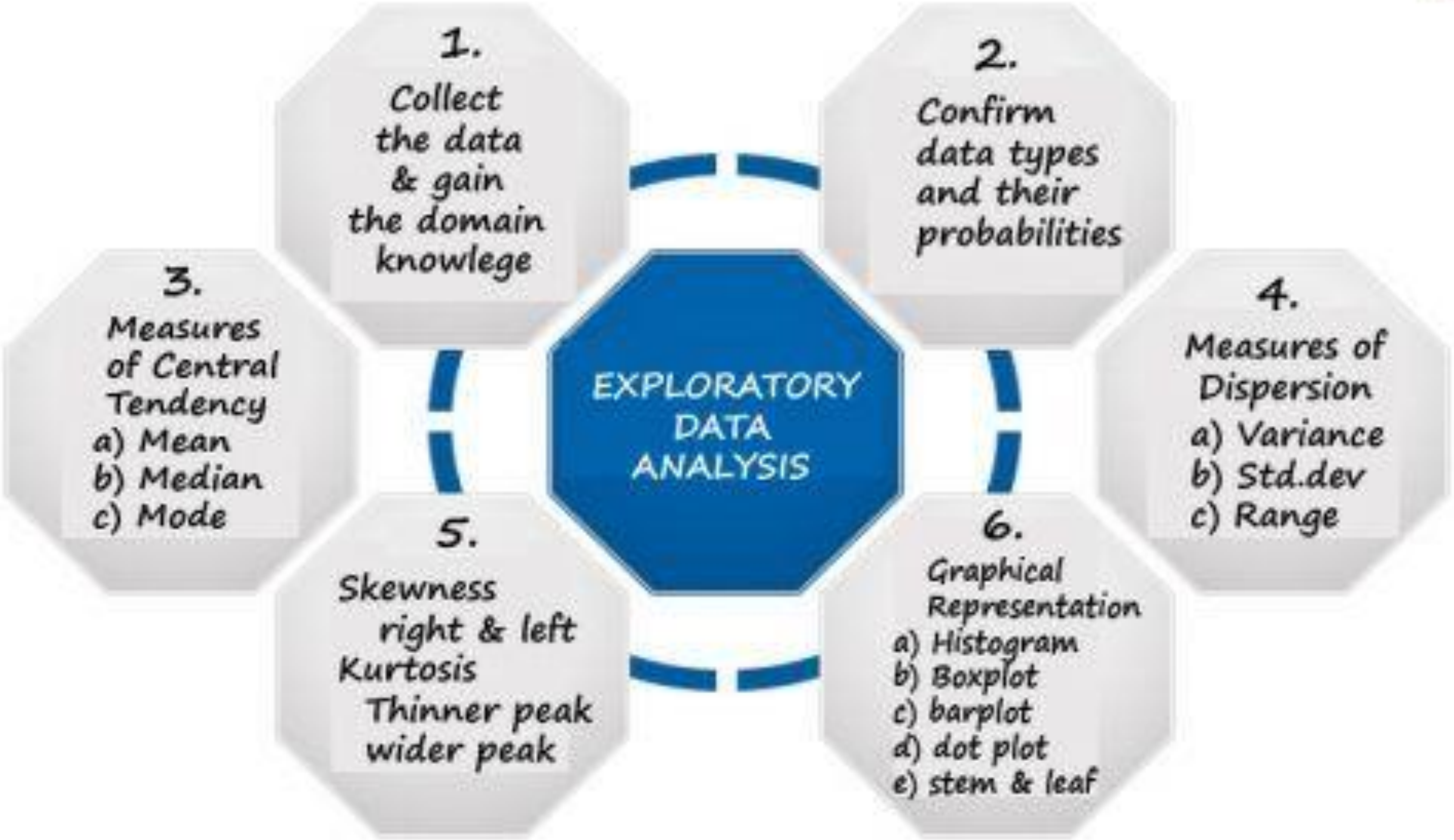

Producing Data
Population
Exploratory Data Analysis
Data
Inference
Probability

Exploratory Data Analysis (EDA) is how we make sense of the data by converting them from their raw form to a more informative one.

In particular, **EDA consists of:**
organizing and summarizing the raw data,
discovering important features and patterns in the data and any striking deviations from those patterns, and then
interpreting our findings in the context of the problem
And can be useful for:
describing the distribution of a single variable (center, spread, shape, outliers)
checking data (for errors or other problems)
checking assumptions to more complex statistical analyses
investigating relationships between variables

EDA

EXCELR
Raising Excellence



EXPLORATORY DATA ANALYSIS

1. Collect the data & gain the domain knowlege

2. Confirm data types and their probabilities

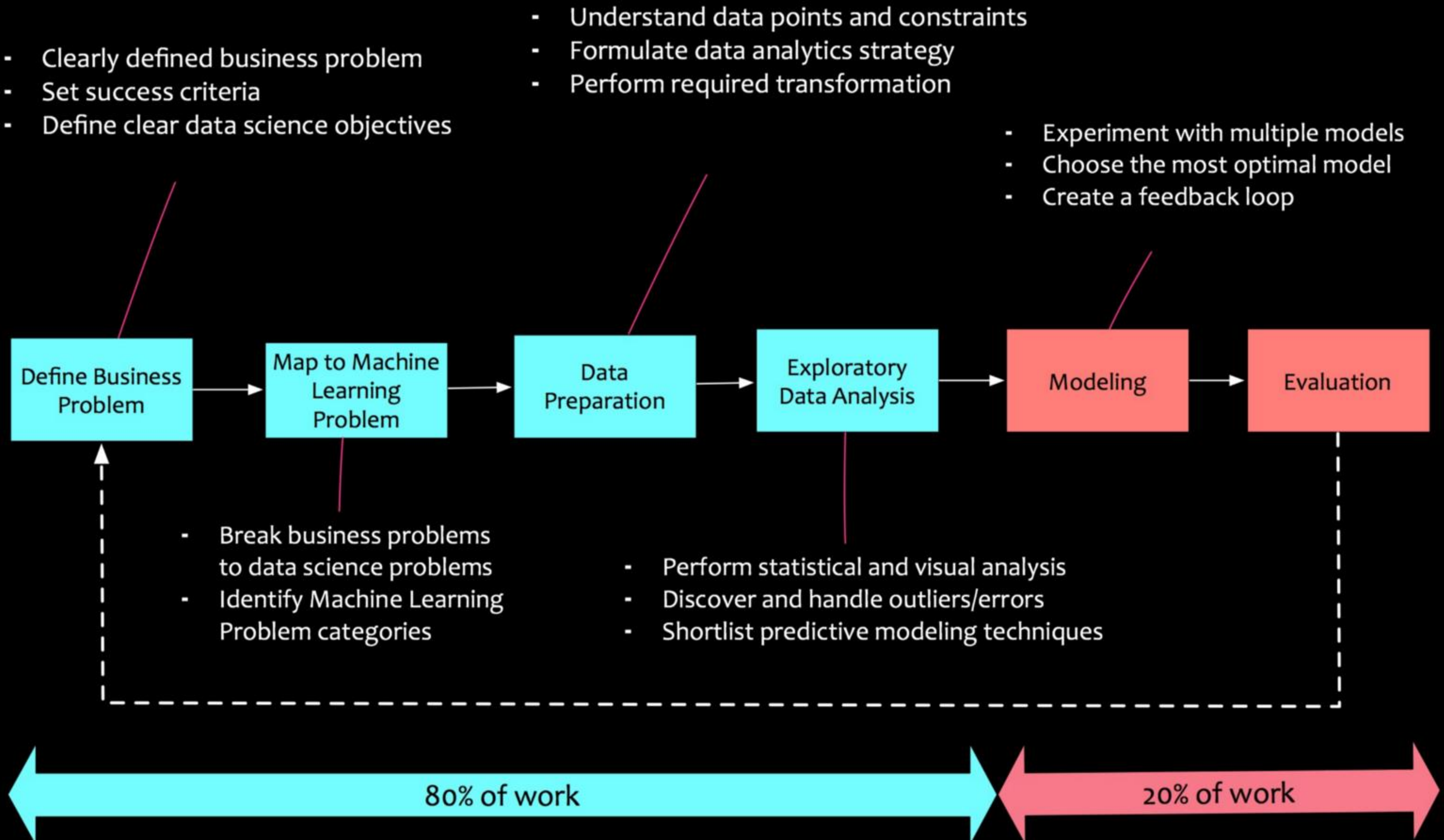
3. Measures of Central Tendency
   a) Mean
   b) Median
   c) Mode

4. Measures of Dispersion
   a) Variance
   b) Std.dev
   c) Range

5. Skewness
   right & left
   Kurtosis
   Thinner peak
   wider peak

6. Graphical Representation
   a) Histogram
   b) Boxplot
   c) barplot
   d) dot plot
   e) stem & leaf

to know more : https://www.excelr.com/blogs/



Define Business Problem

- Clearly defined business problem
- Set success criteria
- Define clear data science objectives

Map to Machine Learning Problem

- Break business problems to data science problems
- Identify Machine Learning Problem categories

Data Preparation

- Understand data points and constraints
- Formulate data analytics strategy
- Perform required transformation

Exploratory Data Analysis

- Perform statistical and visual analysis
- Discover and handle outliers/errors
- Shortlist predictive modeling techniques

Modeling

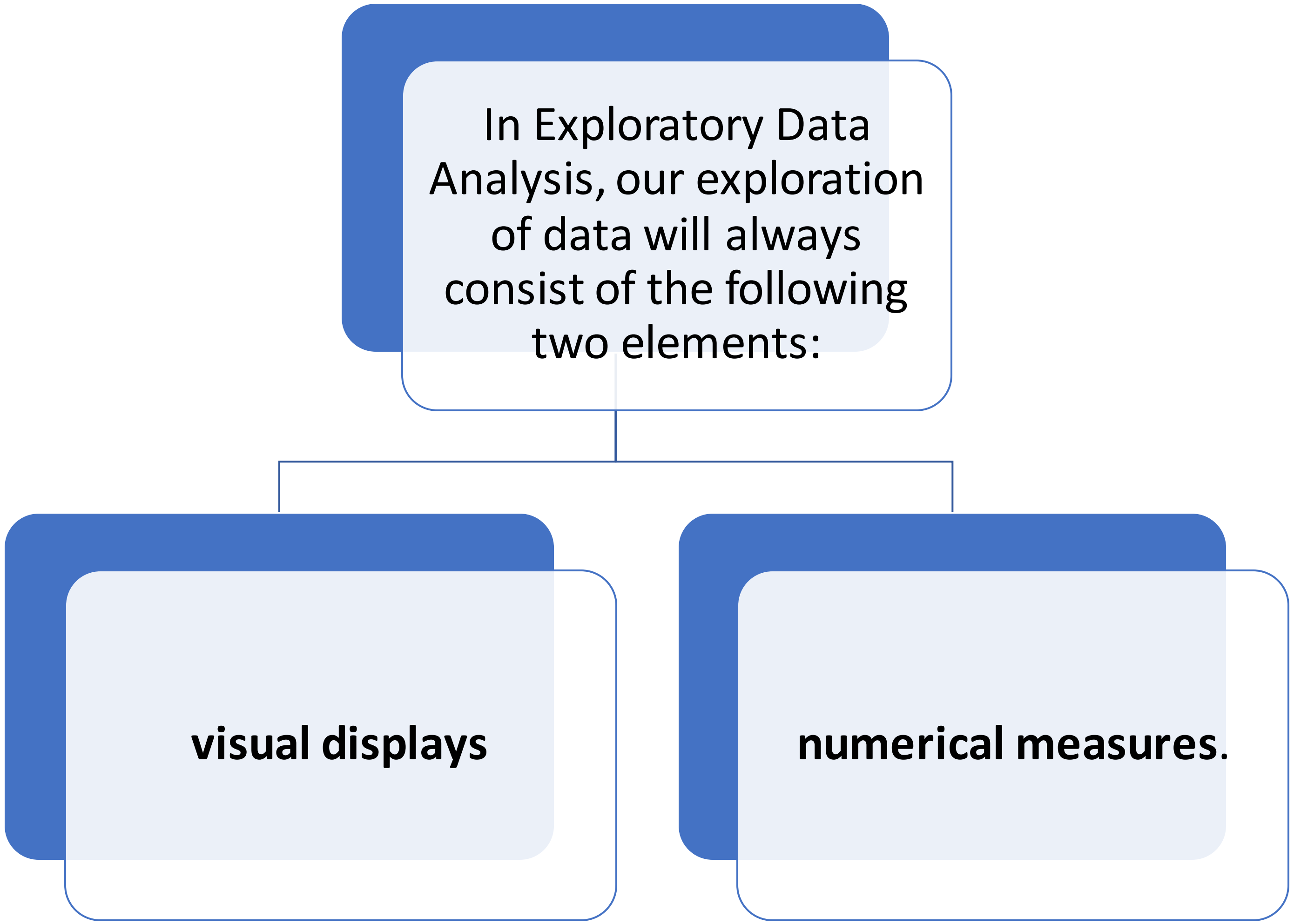
- Experiment with multiple models
- Choose the most optimal model
- Create a feedback loop

Evaluation

80% of work

20% of work

Exploratory data analysis (EDA) methods are often called **Descriptive Statistics** due to the fact that they simply describe, or provide estimates based on, the data at hand.

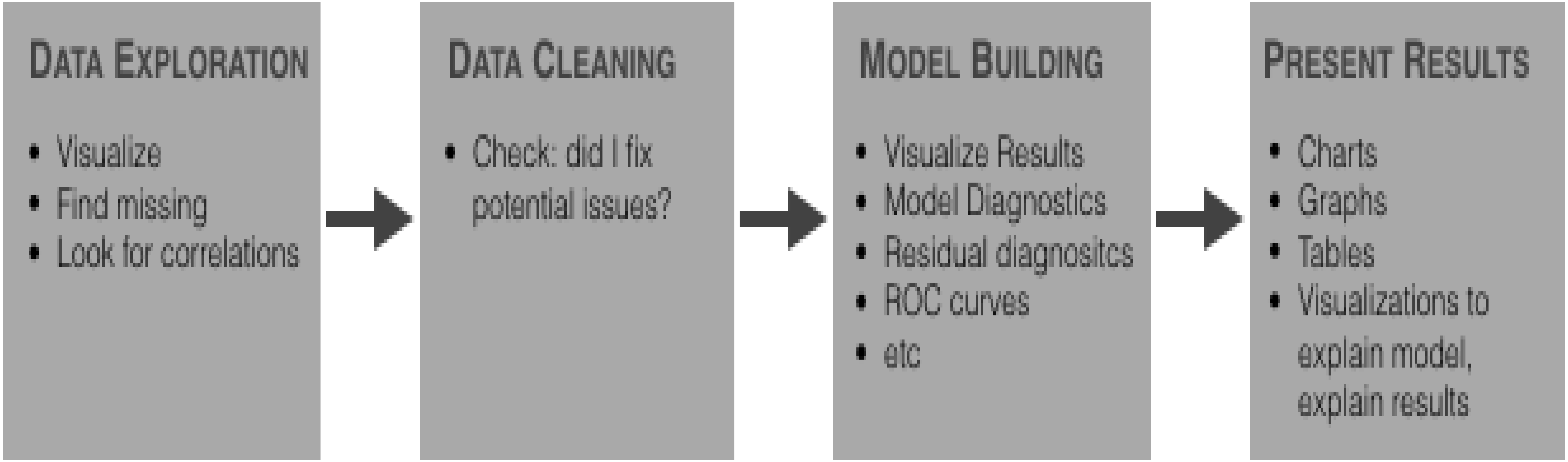


In Exploratory Data Analysis, our exploration of data will always consist of the following two elements:

visual displays

numerical measures.

We use data analysis and visualization at every step of the machine learning process



Data Exploration

- Visualize
- Find missing
- Look for correlations

Data Cleaning

- Check: did I fix potential issues?

Model Building

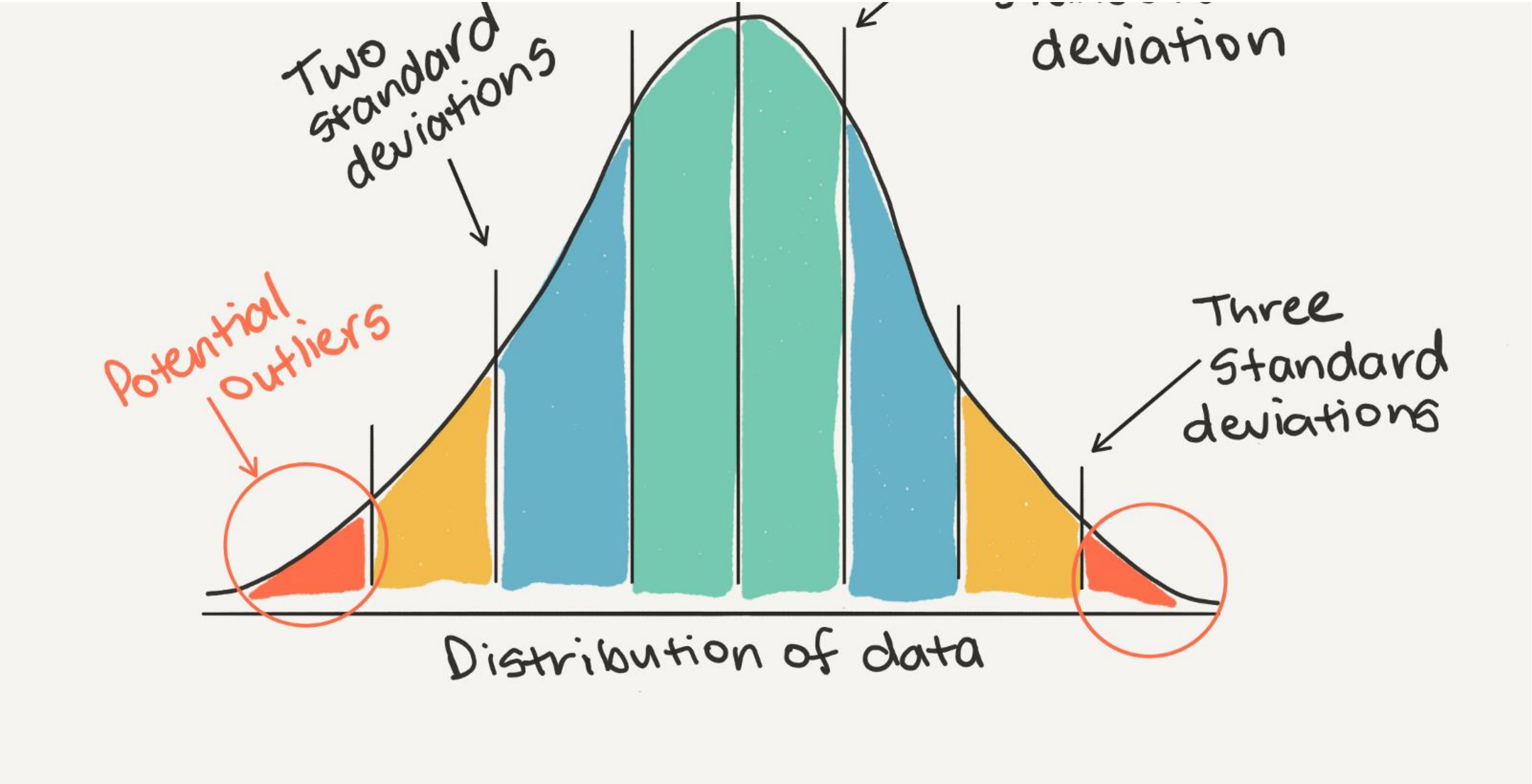
- Visualize Results
- Model Diagnostics
- Residual diagnositcs
- ROC curves
- etc

Present Results

- Charts
- Graphs
- Tables
- Visualizations to explain model, explain results

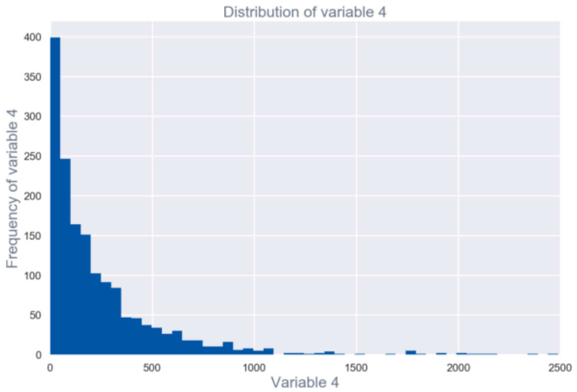
An EDA checklist

- 1. What question(s) are you trying to solve (or prove wrong)?
  2. What kind of data do you have and how do you treat different types?
  3. What’s missing from the data and how do you deal with it?
  4. Where are the outliers and why should you care about them?
  5. How can you add, change or remove features to get more out of your data?

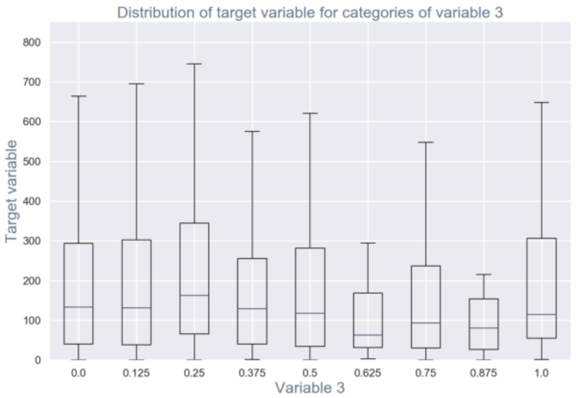

Two standard deviations
deviation
Potential outliers
Three standard deviations
Distribution of data

EDA usually involves a combination of the following methods:

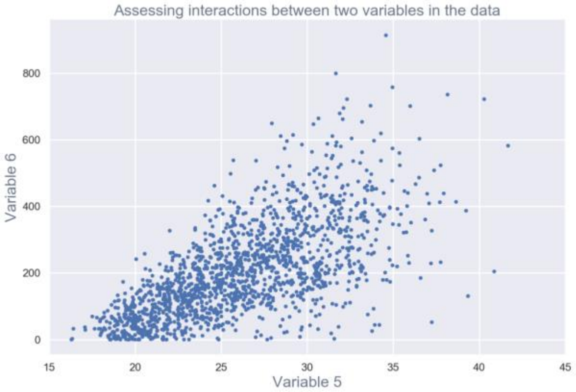
Univariate visualization of and summary statistics for each field in the raw dataset (see figure 1)



Bivariate visualization and summary statistics for assessing the relationship between each variable in the dataset and the target variable of interest (e.g. time until churn, spend) (see figure 2)

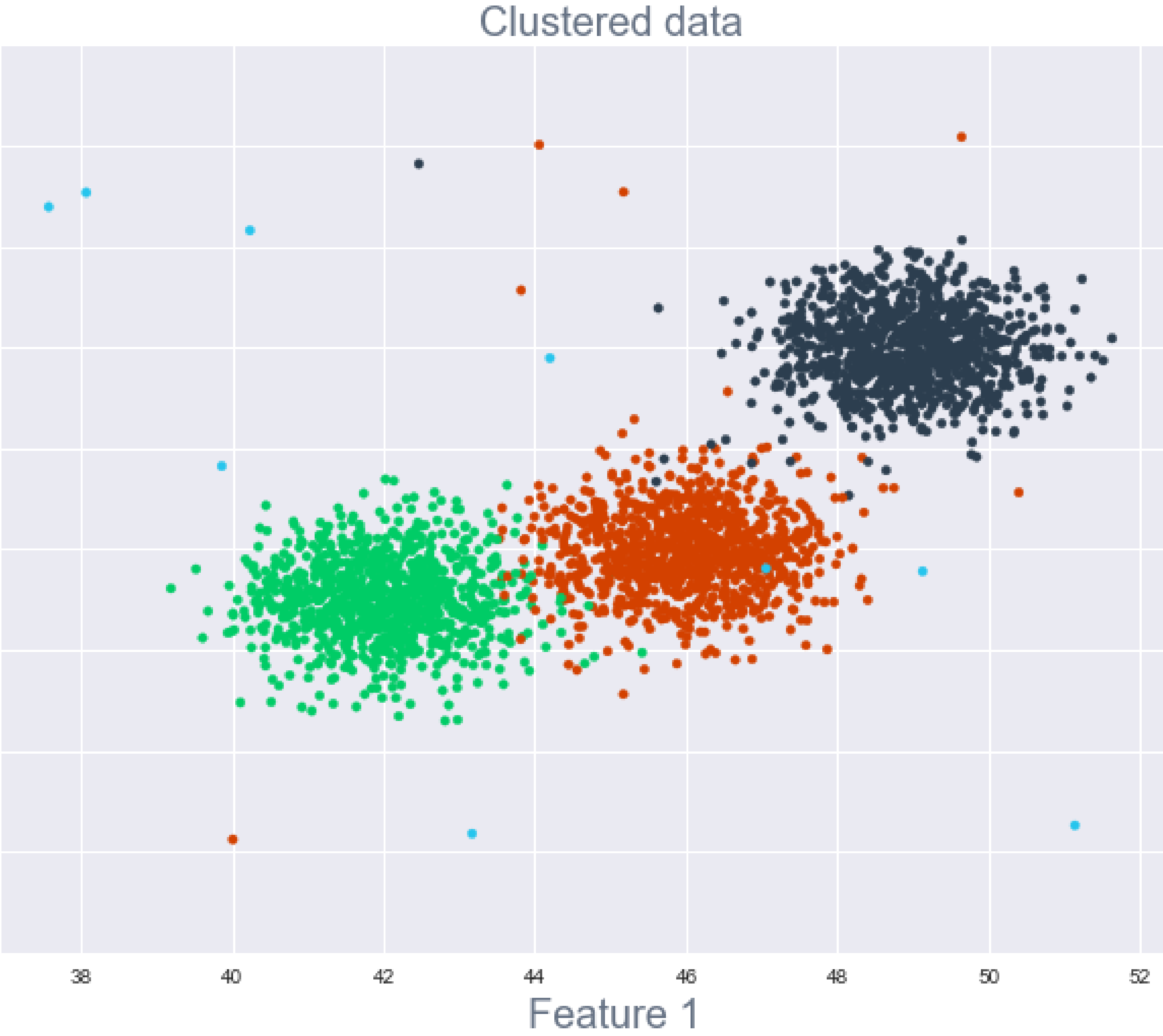


Multivariate visualizations to understand interactions between different fields in the data (see figure 3).



Dimensionality reduction to understand the fields in the data that account for the most variance between observations and allow for the processing of a reduced volume of data

- Clustering of similar observations in the dataset into differentiated groupings, which by collapsing the data into a few small data points, patterns of behavior can be more easily identified (see figure 4)



- Through these methods, the data scientist validates assumptions and identifies patterns that will inform the understanding of the problem and model selection, builds an intuition for the data to ensure high quality analysis, and validates that the data has been generated in the way it was expected to.

EDA is used for:

- Catching mistakes and anomalies
- Gaining new insights into data
- Detecting outliers in data
- Testing assumptions
- Identifying important factors in the data
- Understanding relationships
- And perhaps, most importantly, EDA is used to help figure out our next steps with respect to the data. For instance, we might have new questions we need answered or new research we need to conduct.

EDA Methods

Descriptive statistics

Central tendency of a distribution

Mean	$\frac{\text{Sum of all values}}{\text{Total number of values}}$
Median	Middle value(when data are arranged in order
Mode	Most common value

Measure of Variation

Variance	**how far a set of numbers are spread out from mean**
Interquartile range	divides a data set into quartiles.
Standard deviation	dispersion of a set of data from mean

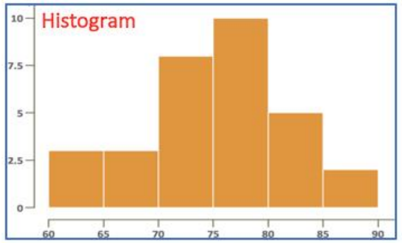
Skewness & Kurtosis

Skewness	**Measure of symmetry**
Kurtosis	**Kurtosis is a measure of “peakedness” relative to a Gaussian shape**

Visualizations

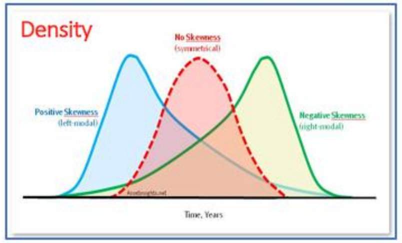
1-dimension

Few data points



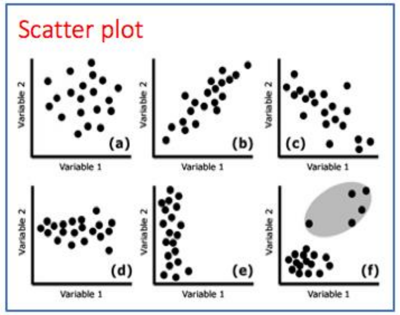
Histogram

Many data points



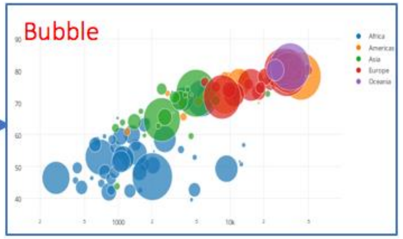
Density

2-dimension



Scatter plot

3-dimension



Bubble

Other exploratory data analysis methods include:

- **Dimensionality reduction:** reduces the number of variables to a few interpretable linear combinations of the data making it easier to understand the fields in the data that account for the most variance between observations and allow for the processing of a reduced volume of data.
- **Cluster analysis:** organizes observed data into similar observations in the dataset into differentiated clusters (groups, which allows for easy identification of patterns of behavior.

- Specific statistical functions and techniques you can perform with these tools include:
- **Clustering and dimension reduction techniques**, which help you to 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**, creating “centers” for each cluster based on the nearest mean.
- **Predictive models**, for example, linear regression.