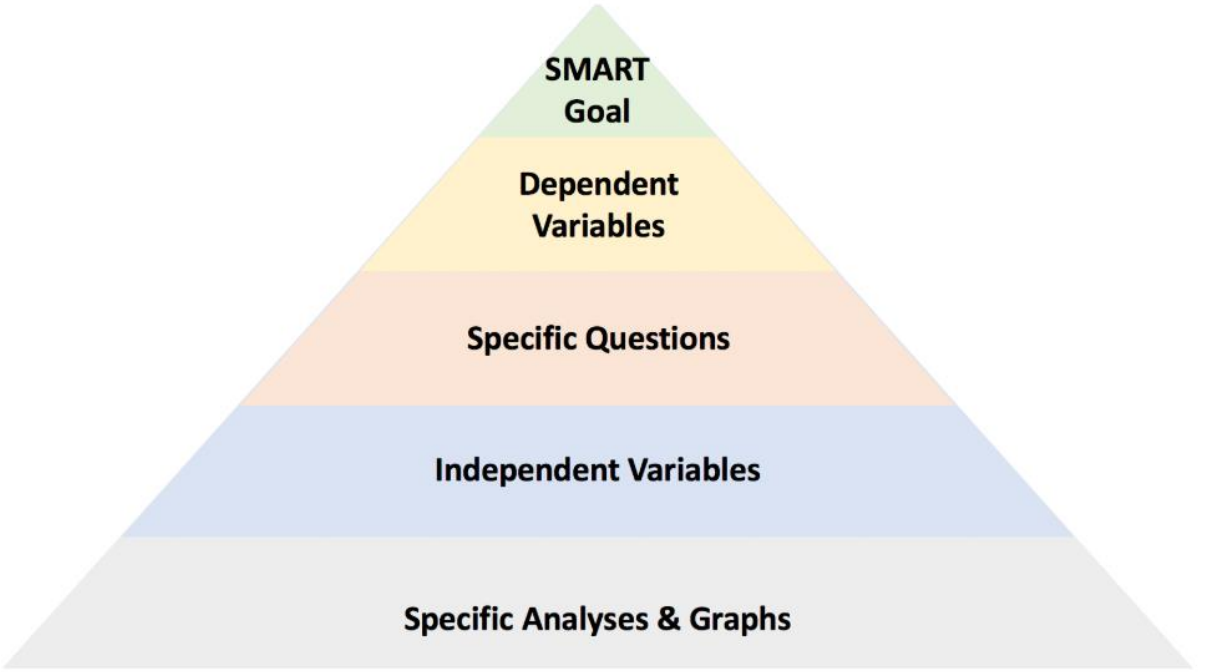
STRUCTURED PYRAMID ANALYSIS PLAN



SMART Goal: It is a goal that is Specific, Measurable, Attainable, Realistic, and Time-Bound.
Dependent Variables: variables that directly pertain to the accomplishment of the S.M.A.R.T. goal.
Specific Questions to Investigate: ideas and intuitions about what parameters may impact the independent variables.
Independent Variables: variables that may help explain the possible connections in the questions, above, to the dependent variables.
Specific Analyses: Statistical analyses and plots of independent versus dependent variables.

SMART GOAL:

SPECIFIC:

- We need to analyse all observations and develop categories on the basis of diagnosis whether the patient is M = malignant, B = benign.
- We need to achieve this by performing data cleaning, exploratory data analysis and various visualizing them through univariate plots.

- We want to predict the outcome of cancer whether is it malignant or benign on the basis of independent variables, this can help us to classify cancer on the basis of existing observations.
- The resources and techniques involved are R and R Studio.

MEASURABLE:

- Measure the project progress weekly on the basis of completion of the task by referring to the checklist and calculating the progress towards the goal.
- Assess the progress to meet the deadlines and achieve the goal.
- Completing and achieving the weekly goals and being on a schedule will make sure that we will achieve our goal.

ACHIEVABLE:

- We can achieve this goal by performing relevant operations on the dataset.

    1. Define Aim
    2. Understand the dataset
    3. Summarize values and data types
    4. Data cleaning
    5. Exploratory Data Analysis and Visualization.
    6. Hypothesis testing
    7. Prepare dataset (Training and Testing)
    8. Model selection
    9. Predictive analysis
    10. Calculate Model performance (accuracy, precision, recall)
    11. Define the best suitable model

RELEVANT:

- This project is done as cancer is one of the most important issues and on the basis of the values of the independent observations and its variations, we can be able to identify the type of cancer in its initial basis and provide treatment for the same.
- This categorization of data in the initial stages would be really useful.

TIME-BOUND:

- Set a deadline to meet your goals.
- Prioritization of the steps and tasks to achieve the goal.
- Plan out short term (weekly goals) and long-term goals (final goal).

DEPENDENT VARIABLES:

- A dependent variable is the variable being tested and measured in a scientific experiment.
- The dependent variable is 'dependent' on the independent variable. As the experimenter changes the independent variable, the effect on the dependent variable is observed and recorded.
- For our dataset, dependent variable is ‘diagnosis’.

SPECIFIC QUESTIONS TO INVESTIGATE:

1. How are the various measurements related?
2. Are there statistically significant differences between Benign and Malignant patients for the mean values of the variables?
3. Do the Benign and Malignant show similar amounts of variation for the variables?

INDEPENDENT VARIABLES:

- An independent variable is the variable that is changed or controlled in a scientific experiment to test the effects on the dependent variable.
- There are 30 independent variables which are all different measurements of cell nucleus size and shape.
- Ten real-valued features are computed for each cell nucleus:

- radius (mean of distances from center to points on the perimeter)
- texture (standard deviation of gray-scale values)
- perimeter
- area
- smoothness (local variation in radius lengths)
- compactness (perimeter^2 / area - 1.0)
- concavity (severity of concave portions of the contour)
- concave points (number of concave portions of the contour)
- symmetry
- fractal dimension ("coastline approximation" - 1)

SPECIFIC ANALYSES & GRAPHS:

- Statistical tests performed:
  T-test

Hotelling T-test
Levene Test

- Data Visualisations
  Bar Plot
  Pie Chart
  Histogram
  Scatter Plot
  Correlation matrix
  Scatter Plot Matrix