upGrad

Power of Visualization in Data Science

Snehith Allamraju

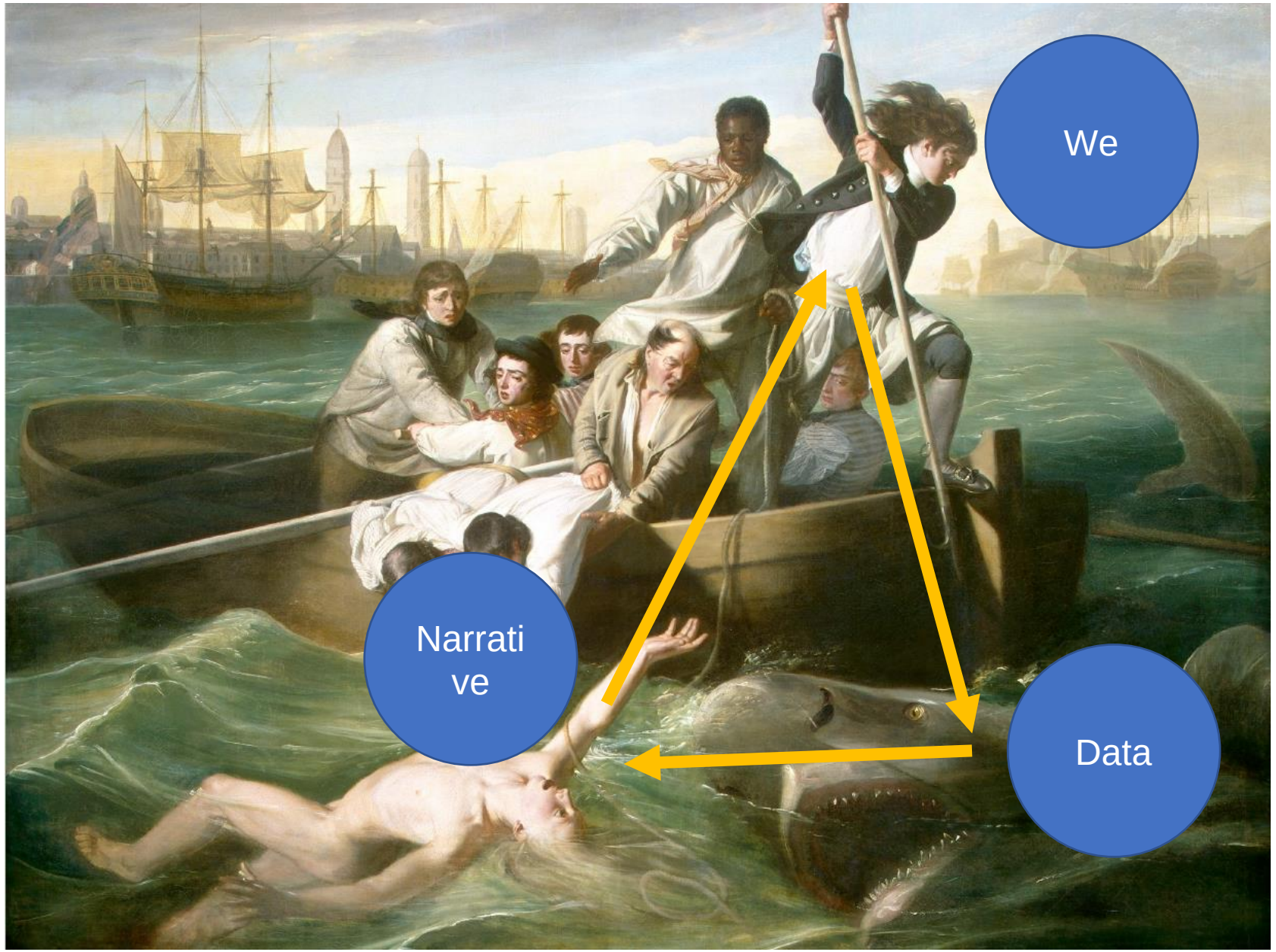
25th May 2022

upGrad

Agenda

- Introduction
- Powering Exploratory Data Analysis
- Visual Aids in Modelling
- Q&A

upGrad



every data tells a story/has a narrative

upGrad

Narrative

/ˈnarətɪv/ (noun): a spoken or written account of connected events; a story.

Origin: Latin word ‘narrare’, “to tell”

Humans are pattern-seeking story-telling animals.

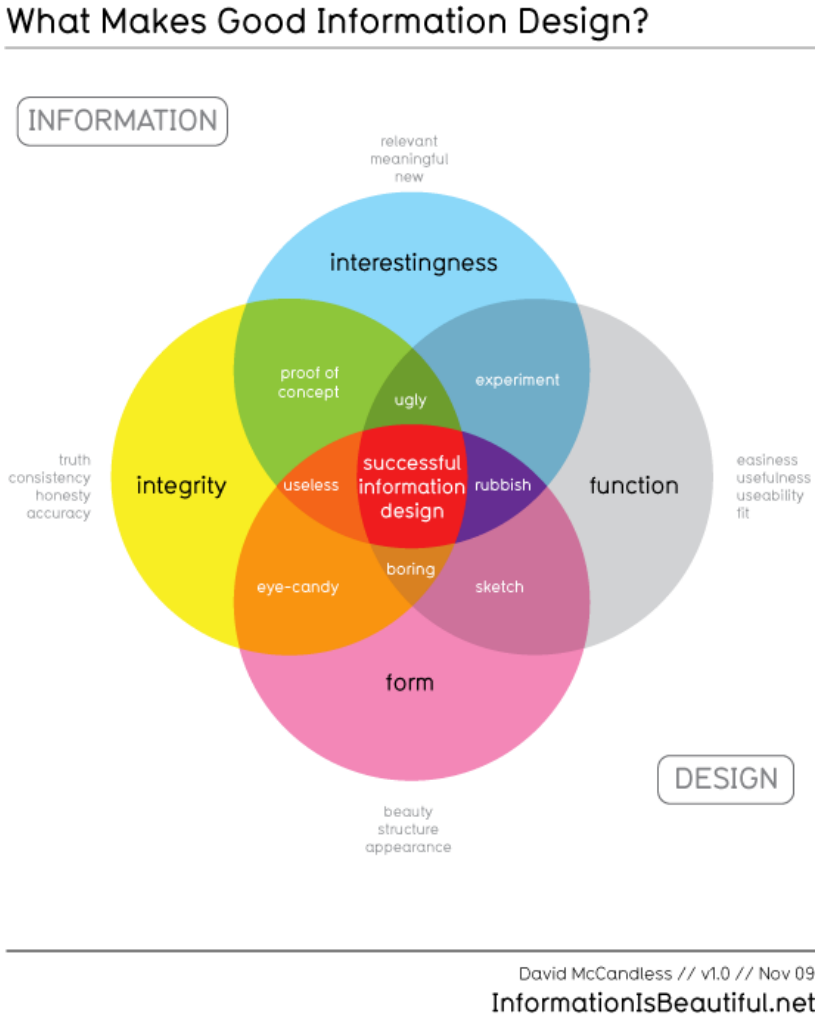
upGrad

Data Visualization effectively is

process of presenting **information** through some **visible means.**

Characteristics of Good Visualization

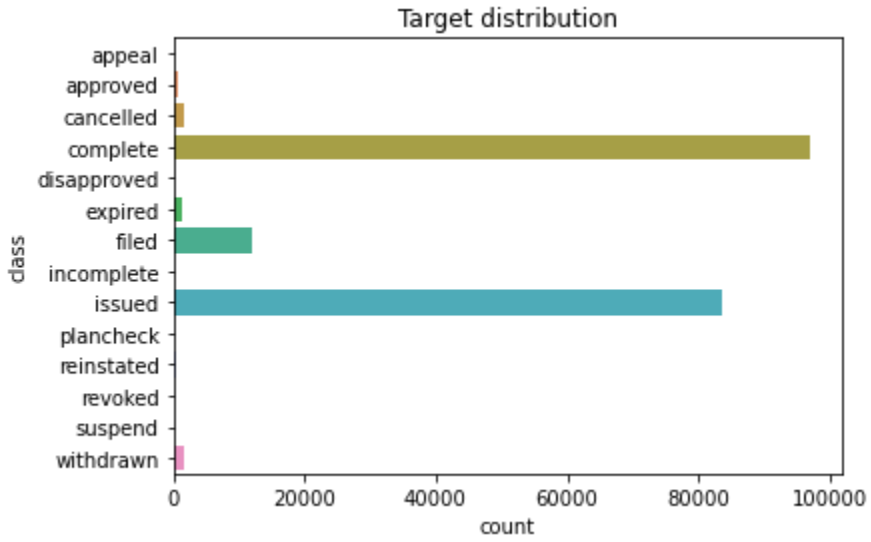
- **Simple & easy to comprehend**
- **Right Information**
- **Meaningful**
- **Usable**
- **Desirable**



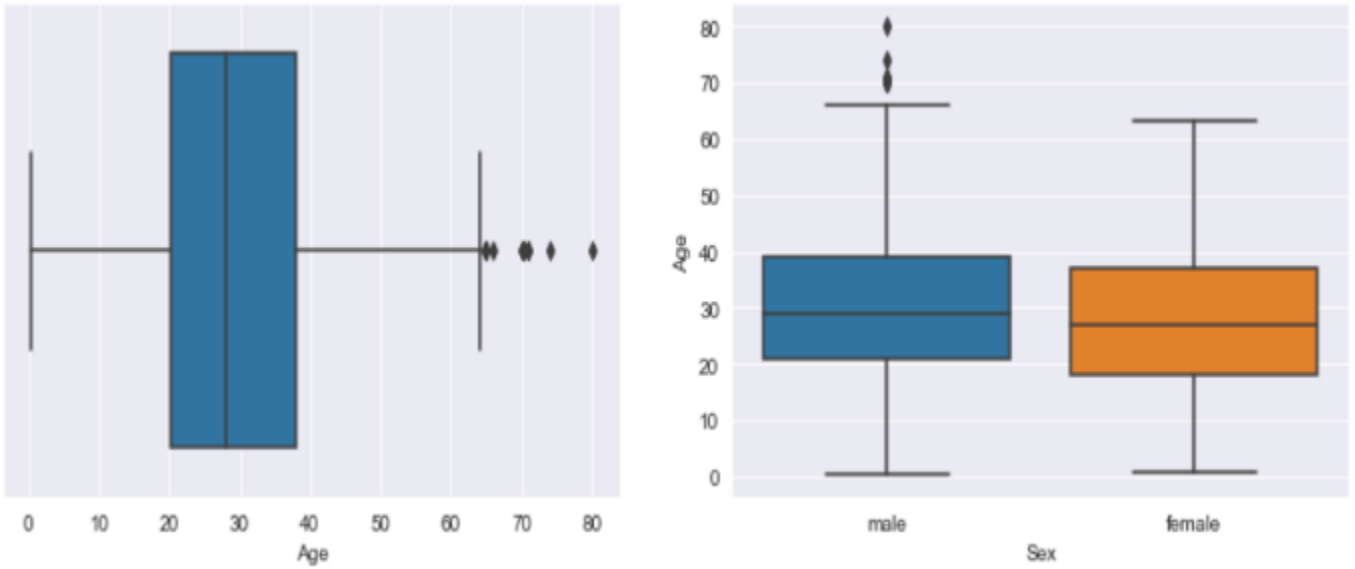
upGrad

Visualization in Data Science

Powering Exploratory Data Analysis

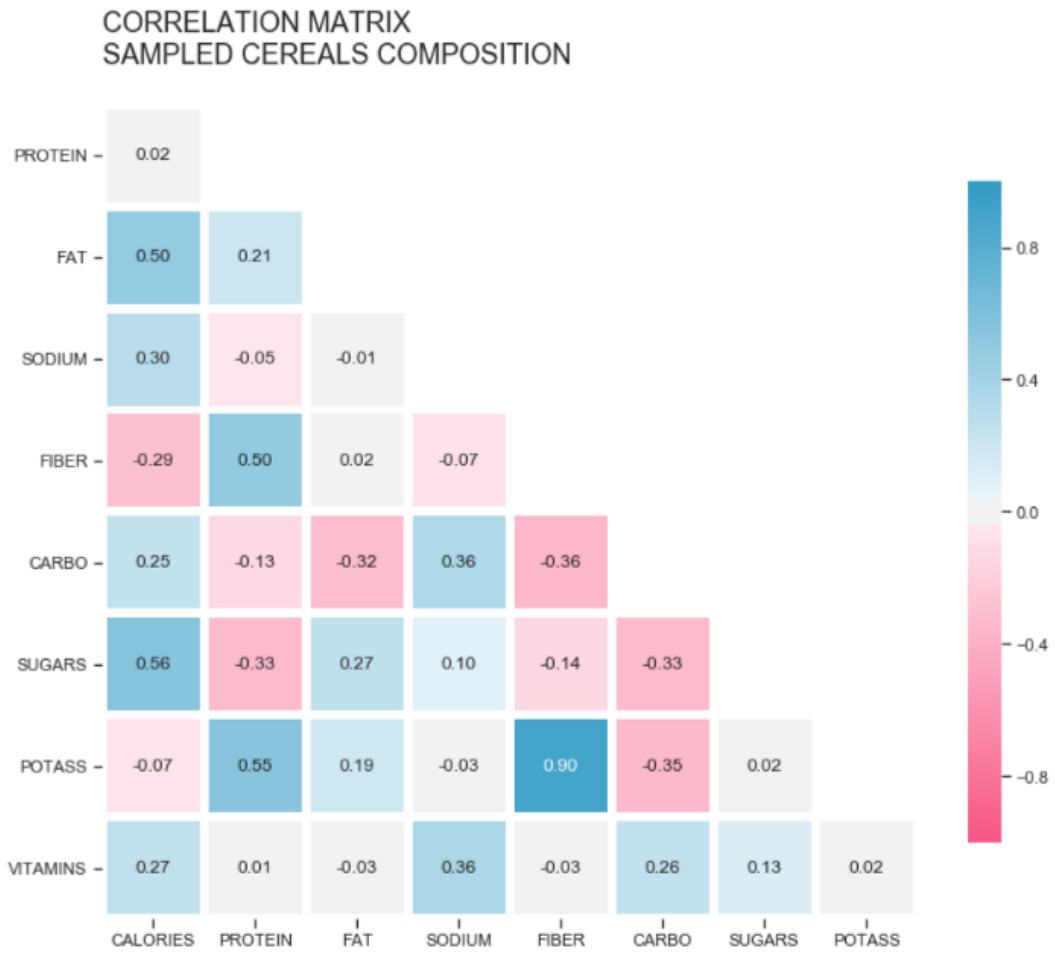


Count Plot

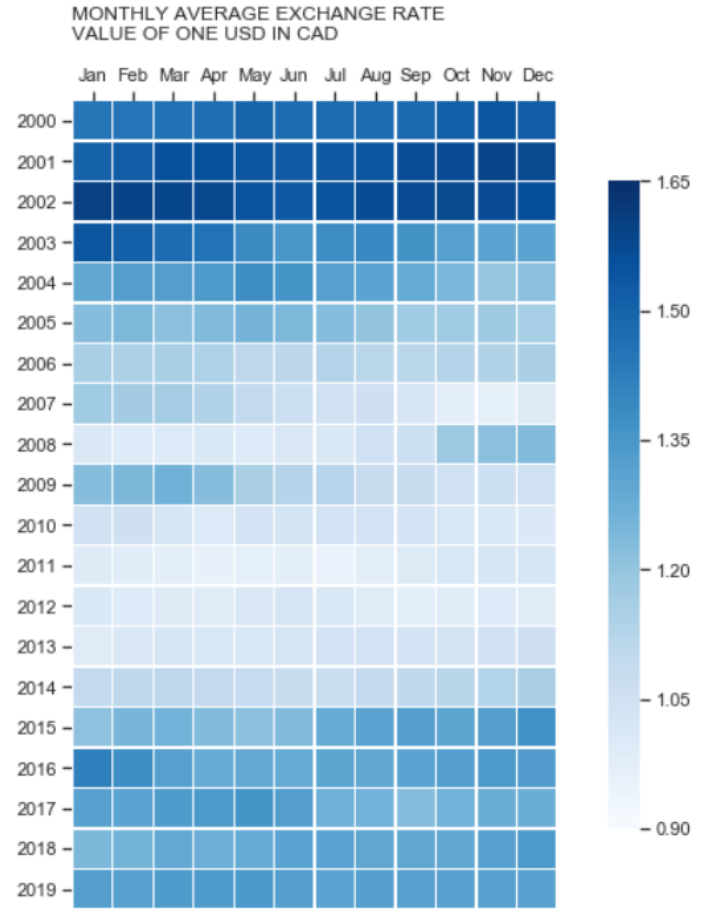


Box Plot

Powering Exploratory Data Analysis



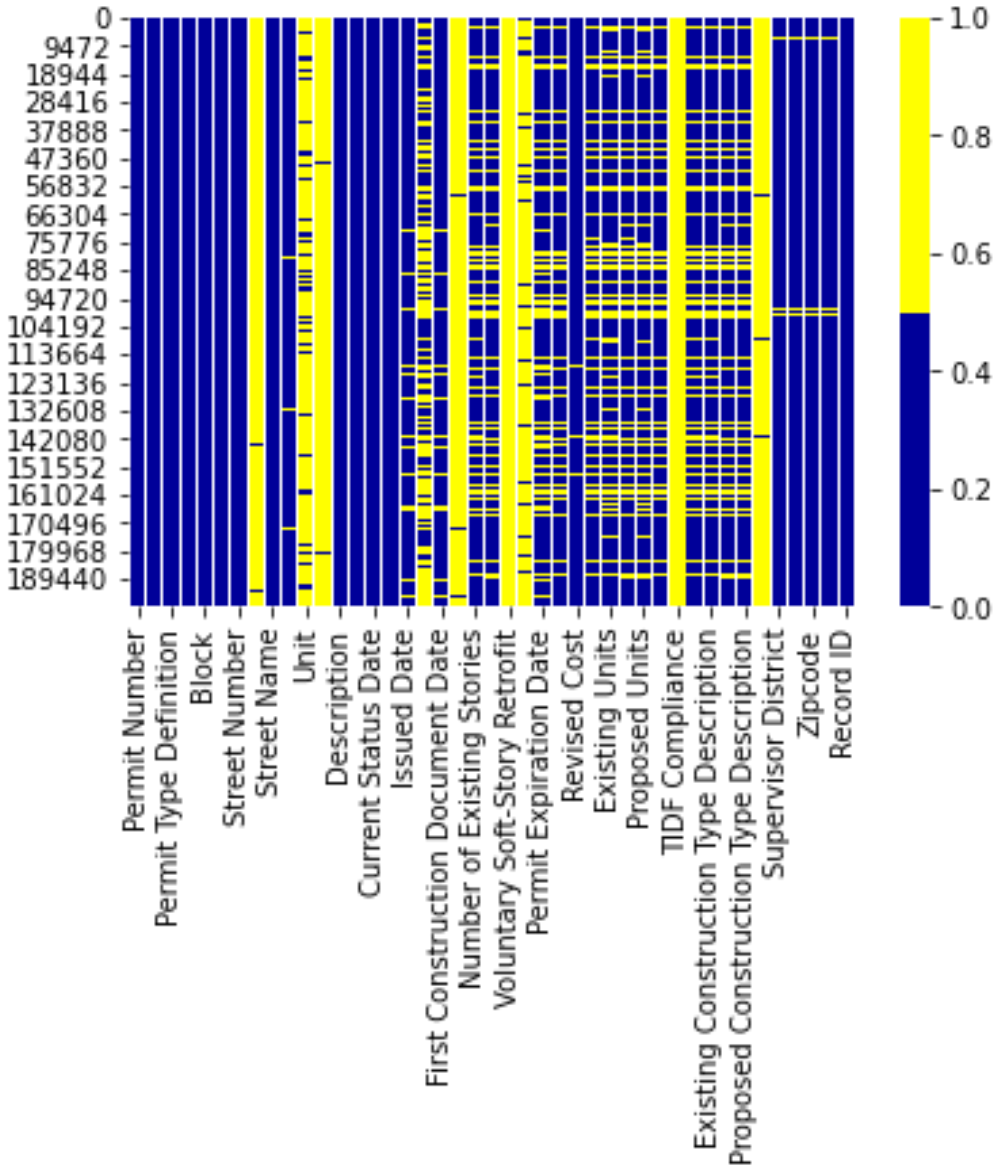
Co-relation matrix



Heatmap

https://towardsdatascience.com/heatmap-basics-with-pythons-seaborn-fb92ea280a6c

Descriptive Statistics

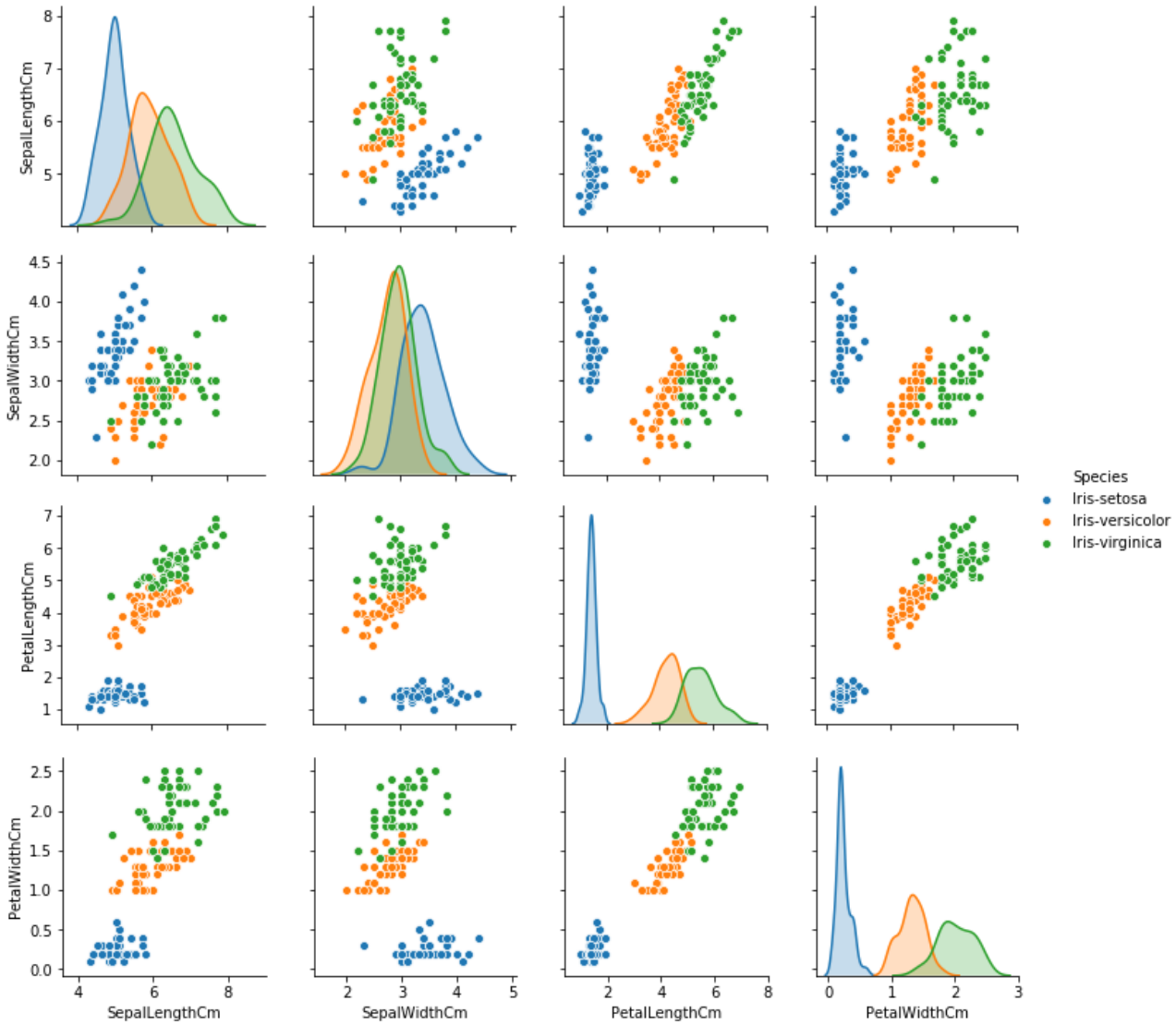


Missing Data Analysis

Heatmaps can also be used to analyze missing data in columns in a dataset

Powering Exploratory Data Analysis

upGrad



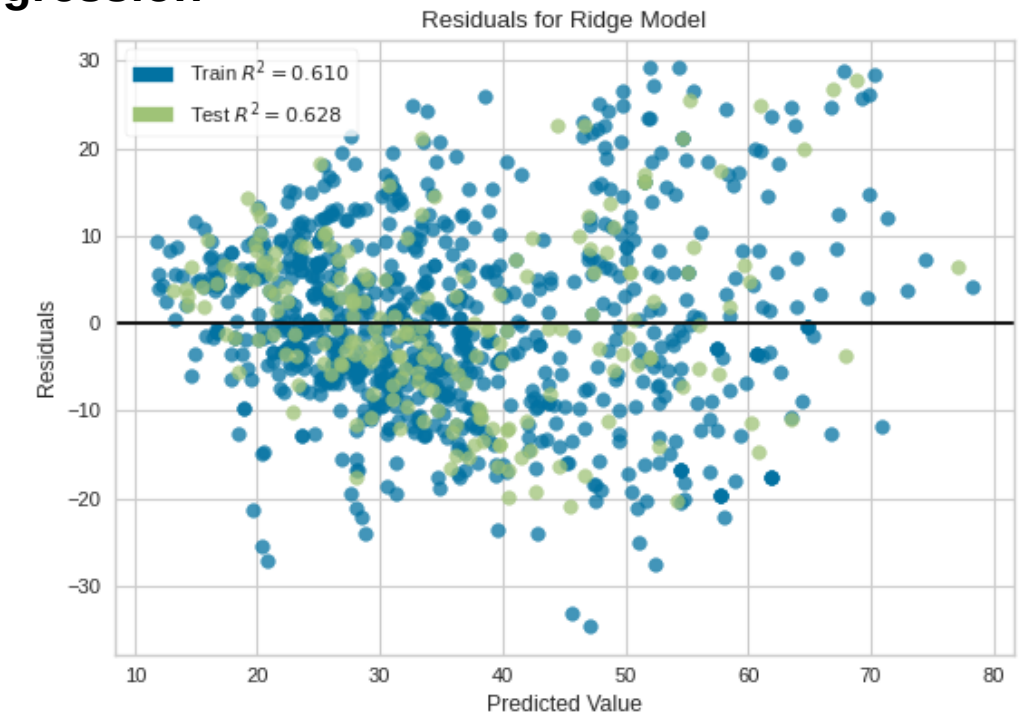
https://www.w3resource.com/machine-learning/scikit-learn/iris/python-machine-learning-scikit-learn-iris-visualization-exercise-13.php

Pair Plot

To understand the best set of features to explain a relationship between two variables

Visual Aids in Modelling

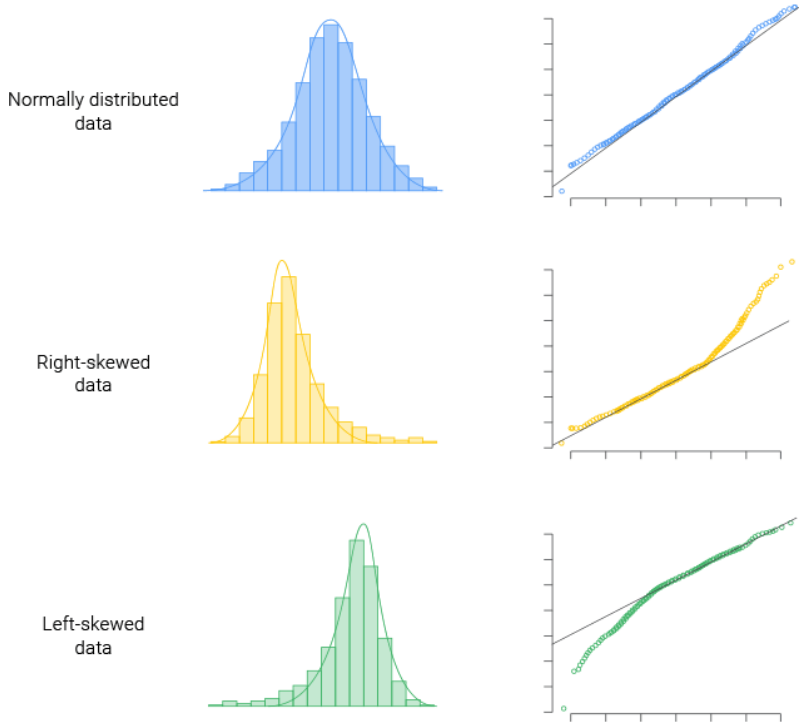
Regression



https://www.scikit-yb.org/en/latest/api/regressor/residuals.html

Residual Plot

- Used to find problems with regression
- Helps to detect regions within the target that may be susceptible to more or less error

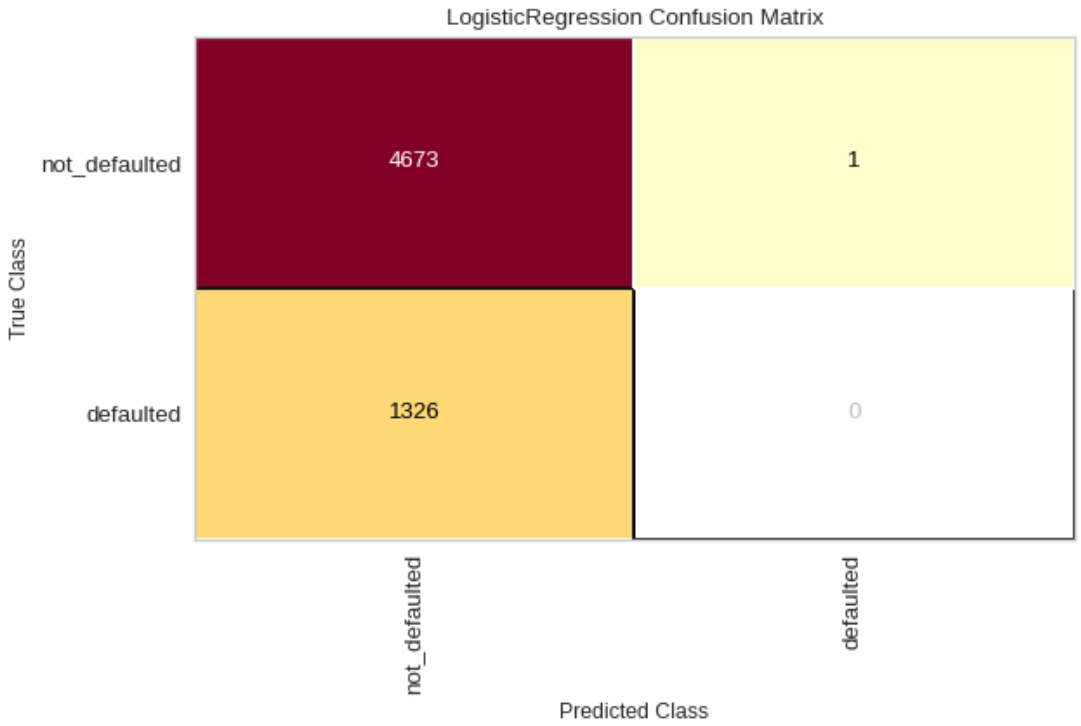


https://www.learnbyexample.org/r-quantile-quantile-qq-plot-base-graph/

QQ Plot

Visual Aids in Modelling

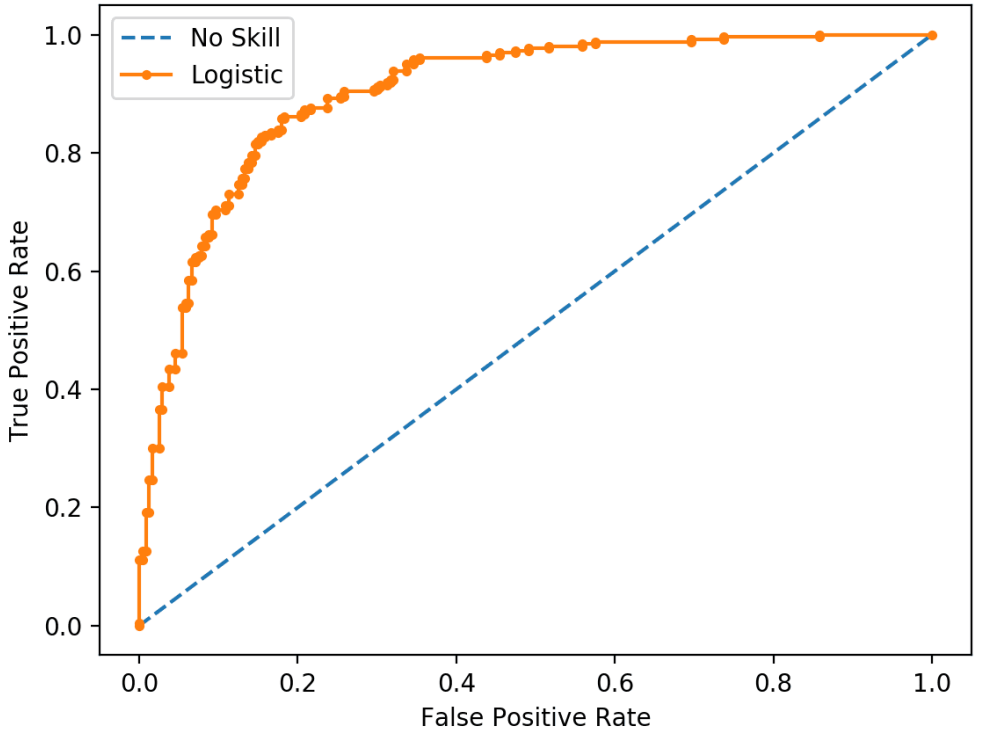
Classification



https://www.scikit-yb.org/en/latest/api/classifier/confusion_matrix.html

Confusion Matrix

- Used to understand classification of individual data points
- Helps to detect Type I and Type II errors



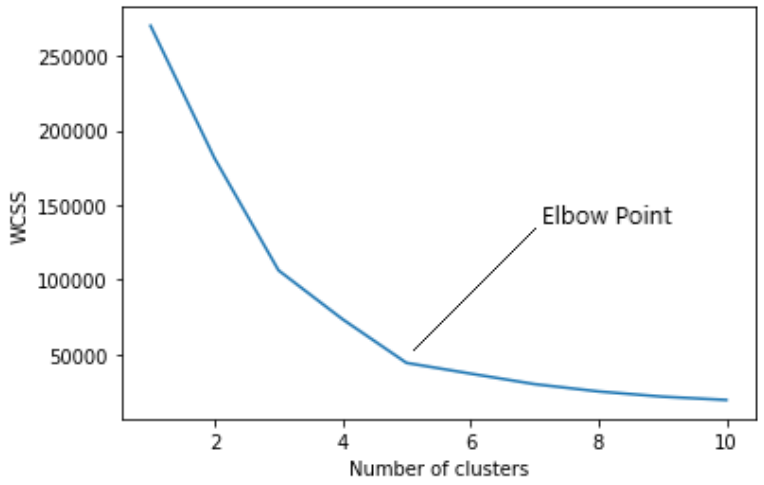
https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-imbalanced-classification/

ROC Curve

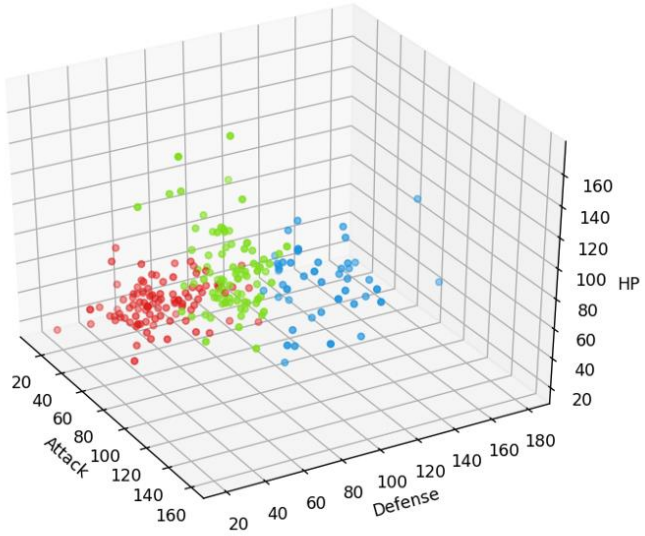
- Plots the false alarm rate vs the hit rate
- Describes how good the model is at predicting the positive class when the actual outcome is positive

Visual Aids in Modelling

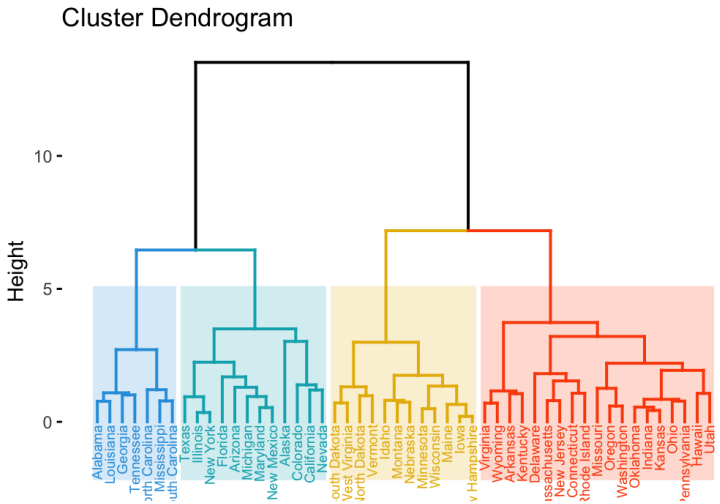
Clustering



https://www.scikit-yb.org/en/latest/api/classifier/confusion_matrix.html



https://towardsdatascience.com/visualizing-clusters-with-pythons-matplolib-35ae03d87489



Elbow Method

- Determine the optimal number of clusters into which the data may be clustered

3D Scatter Plots

- Helps to analyze Clusters in a multidimensional scenario

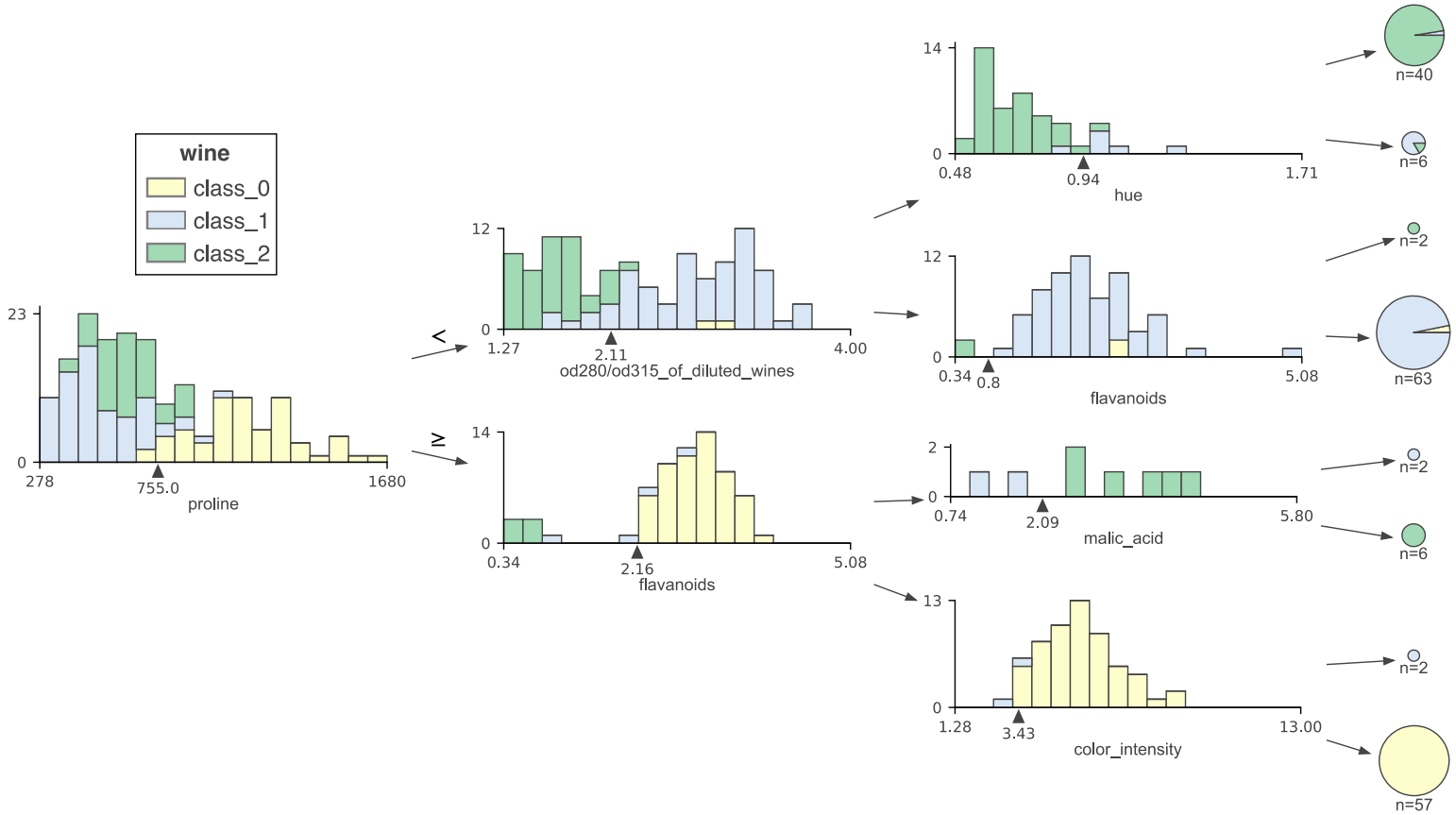
Cluster Dendrogram

- Helps to visualize segmentations/clusters and the distance between the clusters

Visual Aids in Modelling

Decision Trees

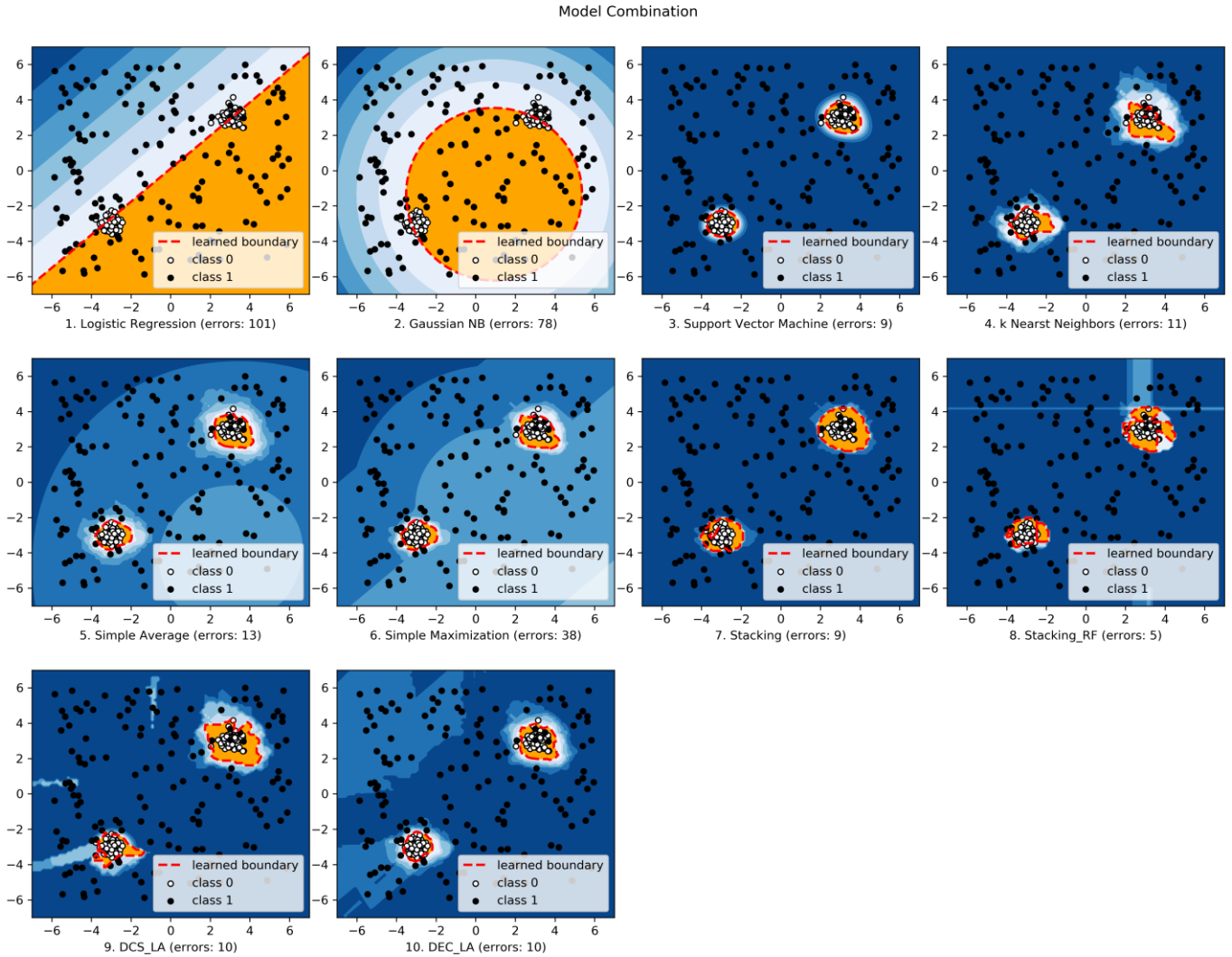
- How a specific feature vector is run down the tree to a leaf. This helps explain why a particular feature vector gets the prediction it does



https://explained.ai/decision-tree-viz/images/samples/iris-TD-5.svg

Visual Aids in Modelling

Model Comparison/ Ensemble modeling



https://github.com/yzhao062/combo

upGrad

Q&A