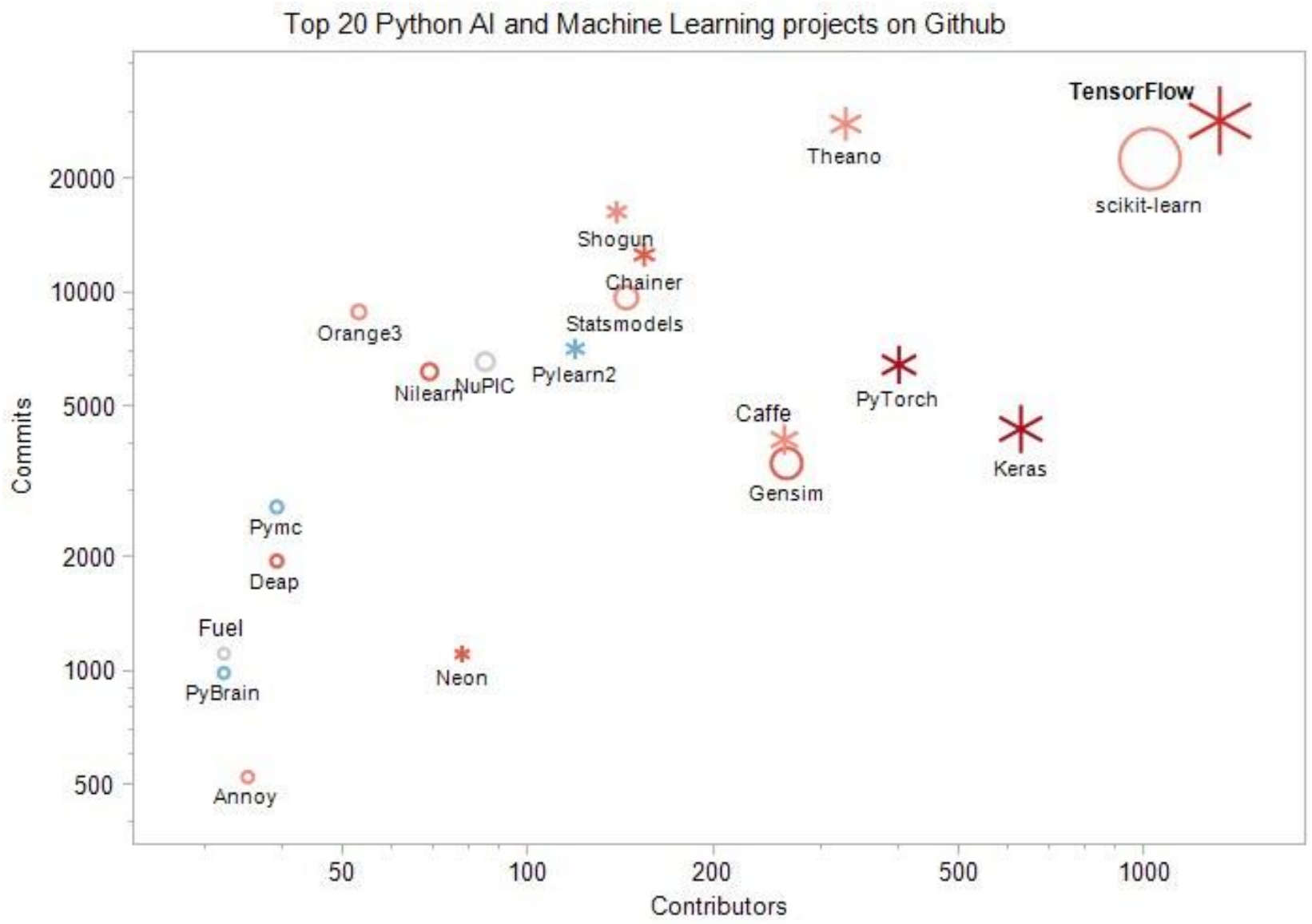
Topic 3
Python Machine Learning

IT8302 APPLIED MACHINE LEARNING

Learning Outcomes

- Use of Python Machine Learning Tools
  - Describe some available python machine learning tools
  - Install and configure Scikit-learn
- Understand Feature Engineering
  - Explain exploratory data analysis
  - Explain feature engineering and feature selection
- Understand Model Evaluation Techniques
  - Explain train set
  - Explain test set
- Understand Evaluation Metrics and Scoring
  - Understand use of evaluation metrics in model selection

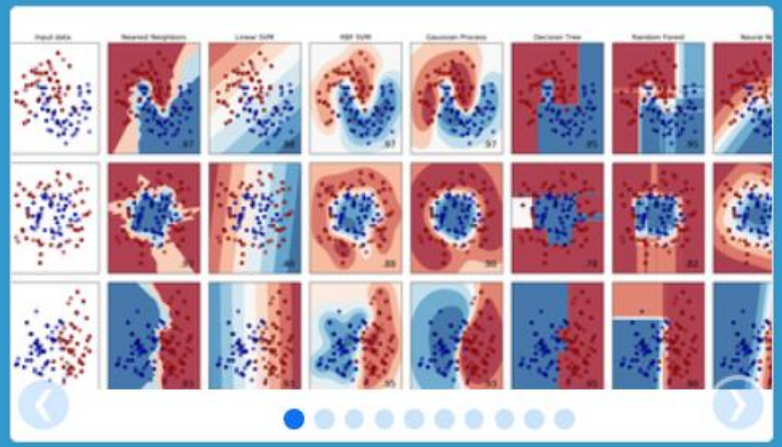
Use of Python Machine Learning Tools



source https://www.kdnuggets.com/2018/02/top-20-python-ai-machine-learning-open-source-projects.html

Open source ML

- **Scikit-learn** is simple and efficient tools for data mining and data analysis, accessible to everybody, and reusable in various context, built on NumPy, SciPy, and matplotlib, open source, commercially usable – BSD license.
  - https://github.com/scikit-learn/scikit-learn
  - http://scikit-learn.org/stable/



scikit-learn

Machine Learning in Python

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

What does scikit-learn do?

Classification

Identifying to which category an object belongs to.

Applications: Spam detection, Image recognition.
Algorithms: SVM, nearest neighbors, random forest, … — Examples

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.
Algorithms: SVR, ridge regression, Lasso, … — Examples

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes
Algorithms: k-Means, spectral clustering, mean-shift, … — Examples

Dimensionality reduction

Reducing the number of random variables to consider.

Applications: Visualization, Increased efficiency
Algorithms: PCA, feature selection, non-negative matrix factorization. — Examples

Model selection

Comparing, validating and choosing parameters and models.

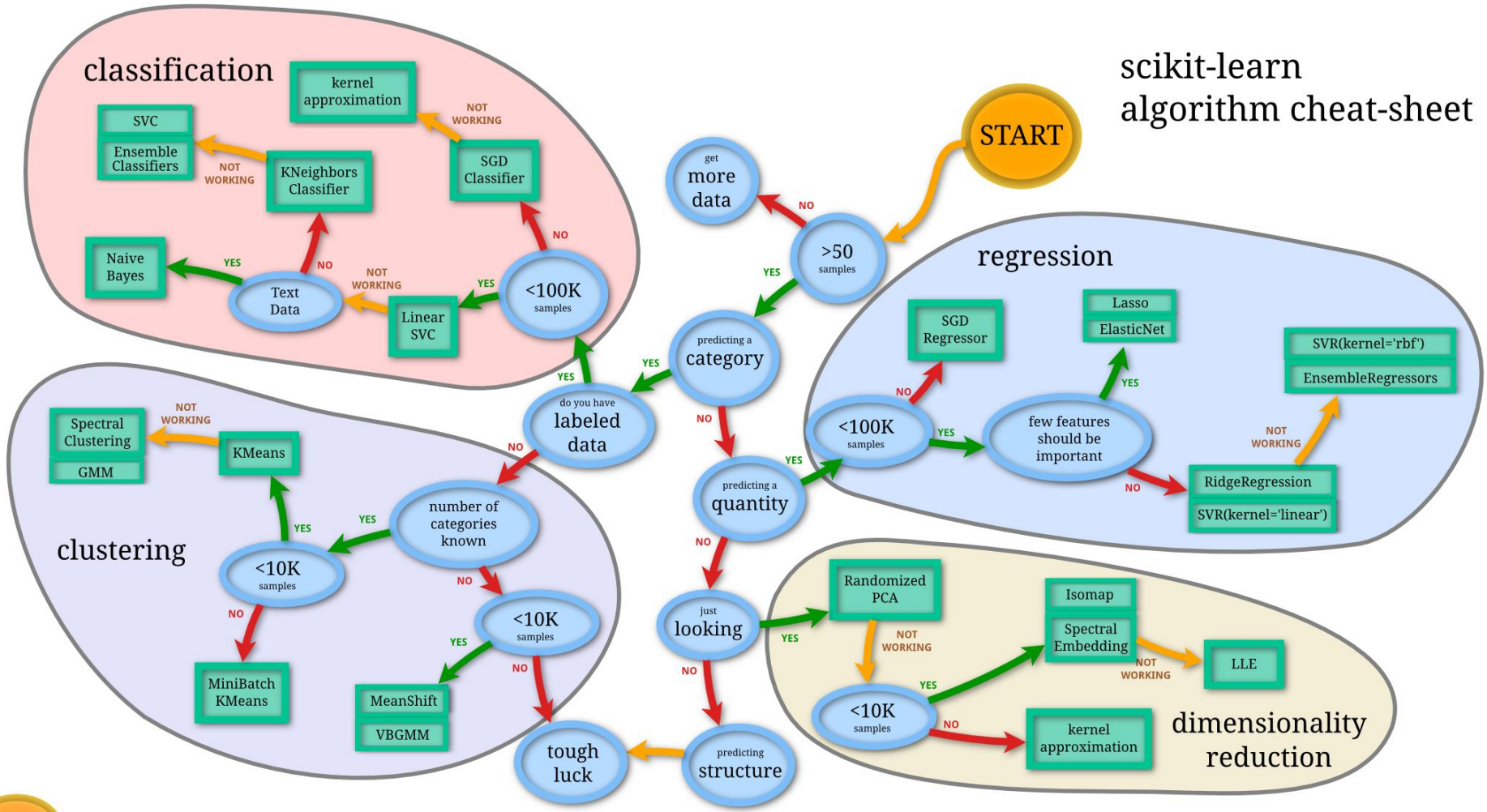
Goal: Improved accuracy via parameter tuning
Modules: grid search, cross validation, metrics. — Examples

Preprocessing

Feature extraction and normalization.

Application: Transforming input data such as text for use with machine learning algorithms.
Modules: preprocessing, feature extraction.

— Examples



source http://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

Scikit-learn installation

http://scikit-learn.org/stable/install.html

Installing the latest release

Scikit-learn requires:

- Python (>= 2.7 or >= 3.3),
- NumPy (>= 1.8.2),
- SciPy (>= 0.13.3).

If you already have a working installation of numpy and scipy, the easiest way to install scikit-learn is using `pip`

```
pip install -U scikit-learn
```

or `conda`:

```
conda install scikit-learn
```

Anaconda ships with a recent version of scikit-learn, in addition to a large set of scientific python library for Windows, Mac OSX and Linux.

Scikit-learn documentation

Scikit-learn is extensive documented. The user guide document is available as a download from:
http://scikit-learn.org/stable/_downloads/scikit-learn-docs.pdf

There is also an online version at:
http://scikit-learn.org/stable/user_guide.html

Scikit-learn datasets

scikit-learn comes with a few small standard datasets that do not require to download any file from some external website.

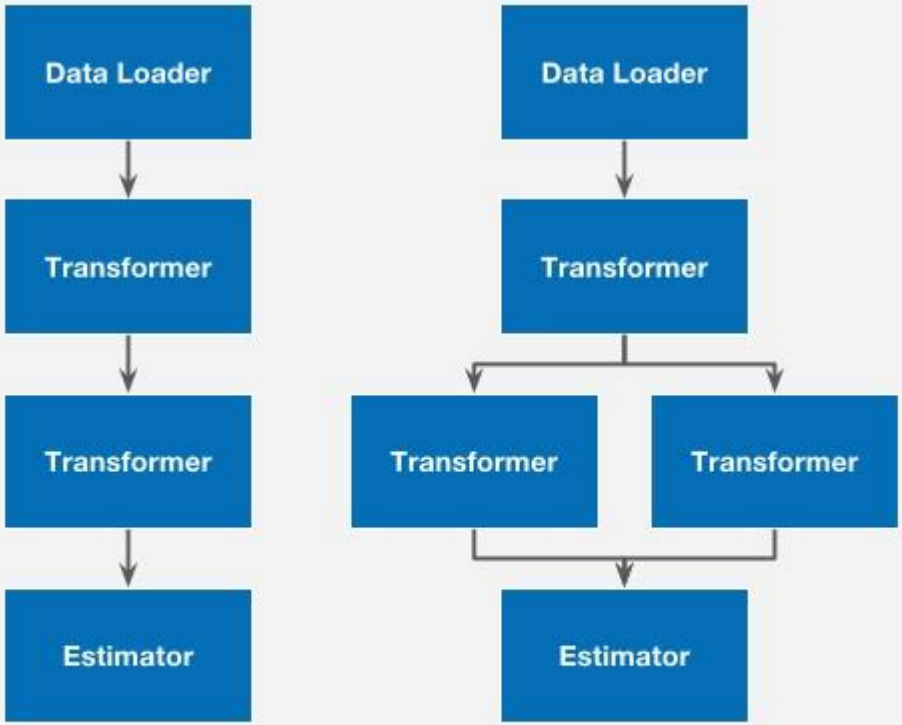
These datasets are useful to quickly illustrate the behavior of the various algorithms implemented in the scikit. They are however often too small to be representative of real world machine learning tasks.

load_boston([return_X_y])	Load and return the boston house-prices dataset (regression).
load_iris([return_X_y])	Load and return the iris dataset (classification).
load_diabetes([return_X_y])	Load and return the diabetes dataset (regression).
load_digits([n_class, return_X_y])	Load and return the digits dataset (classification).
load_linnerud([return_X_y])	Load and return the linnerud dataset (multivariate regression).
load_wine([return_X_y])	Load and return the wine dataset (classification).
load_breast_cancer([return_X_y])	Load and return the breast cancer Wisconsin dataset (classification).

source http://scikit-learn.org/stable/datasets/index.html

Scikit-learn pipeline

- Pipelines are an extremely simple yet very useful tool for managing machine learning workflows.
  - A typical machine learning task generally involves data preparation to varying degrees. Such tasks are known for taking up a large proportion of time spent on any given machine learning task.
  - After a dataset is cleaned up from a potential initial state of massive disarray, however, there are still several less-intensive yet no less-important transformative data pre-processing steps such as feature extraction, feature scaling, and dimensionality reduction, to name just a few.
  - Maybe your pre-processing requires only one of these transformations, such as some form of scaling. But maybe you need to string a number of transformations together, and ultimately finish off with an estimator of some sort.



Scikit-Learn Pipelines: `fit()` and `predict()`

Scikit-learn Pipelines

- Scikit-learn's Pipeline class is designed as a manageable way to apply a series of data transformations followed by the application of an estimator. It is a pipeline of transforms with a final estimator.
- Pipelines is useful for:
  - Convenience in creating a coherent and easy-to-understand workflow
  - Enforcing workflow implementation and the desired order of step applications
  - Reproducibility
  - Value in persistence of entire pipeline objects

Understand Feature Engineering

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 and to check assumptions with the help of summary statistics and graphical representations.

Source: https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15

EDA (Wine Quality Dataset)

Example of white variant of Wine Quality data set which is available on UCI Machine Learning Repository and try to catch hold of as many insights from the data set using EDA.

```
In [2]: df = pd.read_csv('winequality-white.csv',sep=';')
        df.head()
```

Out[2]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.0010	3.00	0.45	8.8	6
1	6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.9940	3.30	0.49	9.5	6
2	8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.9951	3.26	0.44	10.1	6
3	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6
4	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6

Find out total number of rows and columns in the data set using ".shape".

```
In [3]: df.shape
Out[3]: (4898, 12)
```

EDA (Wine Quality Dataset)

```
In [5]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4898 entries, 0 to 4897
Data columns (total 12 columns):
fixed acidity           4898 non-null float64
volatile acidity        4898 non-null float64
citric acid             4898 non-null float64
residual sugar          4898 non-null float64
chlorides               4898 non-null float64
free sulfur dioxide     4898 non-null float64
total sulfur dioxide    4898 non-null float64
density                 4898 non-null float64
pH                      4898 non-null float64
sulphates               4898 non-null float64
alcohol                 4898 non-null float64
quality                 4898 non-null int64
dtypes: float64(11), int64(1)
memory usage: 459.3 KB
```

Data has only float and integer values.
No variable column has null/missing values.

EDA (Wine Quality Dataset)

```
In [6]: df.describe()
```

Out[6]:

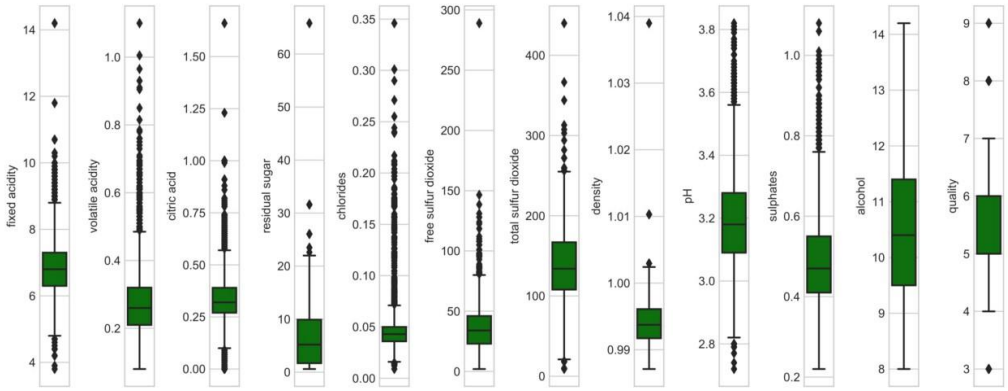
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
count	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000
mean	6.854788	0.278241	0.334192	6.391415	0.045772	35.308085	138.360657	0.994027	3.188267	0.489847	10.514267
std	0.843868	0.100795	0.121020	5.072058	0.021848	17.007137	42.498065	0.002991	0.151001	0.114126	1.230621
min	3.800000	0.080000	0.000000	0.600000	0.009000	2.000000	9.000000	0.987110	2.720000	0.220000	8.000000
25%	6.300000	0.210000	0.270000	1.700000	0.036000	23.000000	108.000000	0.991723	3.090000	0.410000	9.500000
50%	6.800000	0.260000	0.320000	5.200000	0.043000	34.000000	134.000000	0.993740	3.180000	0.470000	10.400000
75%	7.300000	0.320000	0.390000	9.900000	0.050000	46.000000	167.000000	0.996100	3.280000	0.550000	11.400000
max	14.200000	1.100000	1.660000	65.800000	0.346000	289.000000	440.000000	1.038980	3.820000	1.080000	14.200000

- Here as you can notice mean value is less than median value of each column which is represented by 50%(50th percentile) in index column.
- There is notably a large difference between 75th %tile and max values of predictors “residual sugar”, ”free sulfur dioxide”, ”total sulfur dioxide”.
- Thus suggesting that there are extreme values — Outliers in our data set.

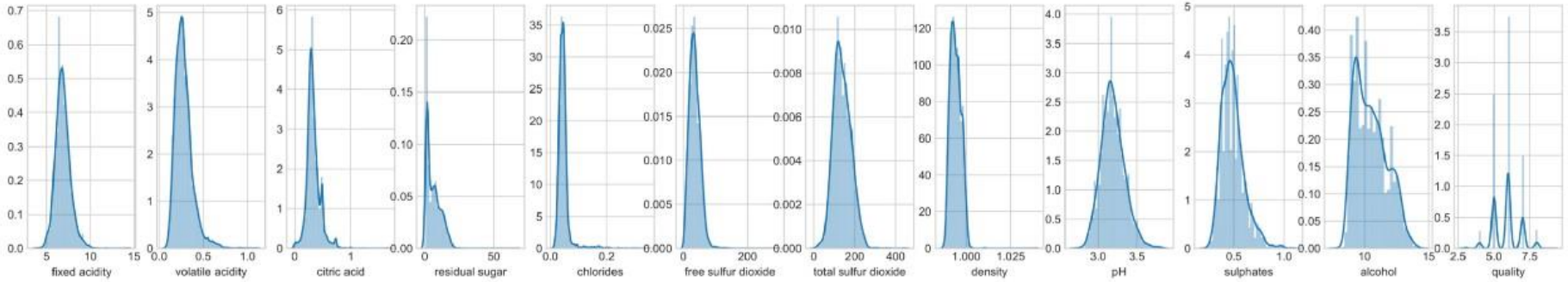
EDA (Wine Quality Dataset)

To check Outliers

In [12]:
```
l = df.columns.values
number_of_columns=12
number_of_rows = len(l)-1/number_of_columns
plt.figure(figsize=(number_of_columns,5*number_of_rows))
for i in range(0,len(l)):
    plt.subplot(number_of_rows + 1,number_of_columns,i+1)
    sns.set_style('whitegrid')
    sns.boxplot(df[l[i]],color='green',orient='v')
    plt.tight_layout()
```



EDA (Wine Quality Dataset)

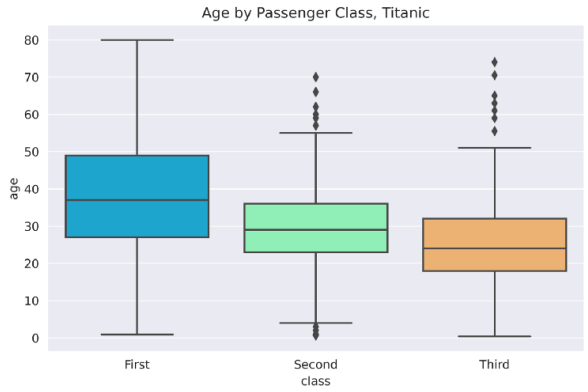


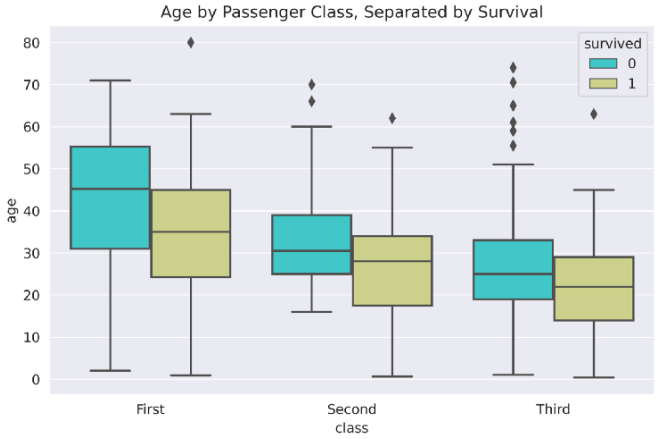
To check distribution-Skewness

```
In [13]: plt.figure(figsize=(2*number_of_columns,5*number_of_rows))
         for i in range(0,len(l)):
             plt.subplot(number_of_rows + 1,number_of_columns,i+1)
             sns.distplot(df[l[i]],kde=True)
```

"pH" column appears to be normally distributed
remaining all independent variables are right skewed/positively skewed.

Exploration with categorical variables





source: https://towardsdatascience.com/a-complete-guide-to-plotting-categorical-variables-with-seaborn-bfe54db66bec#69ab
source: https://towardsdatascience.com/a-major-seaborn-plotting-tip-i-wish-i-had-learned-earlier-d8209ad0a20e

Feature Engineering

Feature engineering is the process of using domain knowledge to extract features from raw data. These features can be used to improve the performance of machine learning algorithms.

The goal of feature engineering is simply to make your data better suited to the problem at hand.

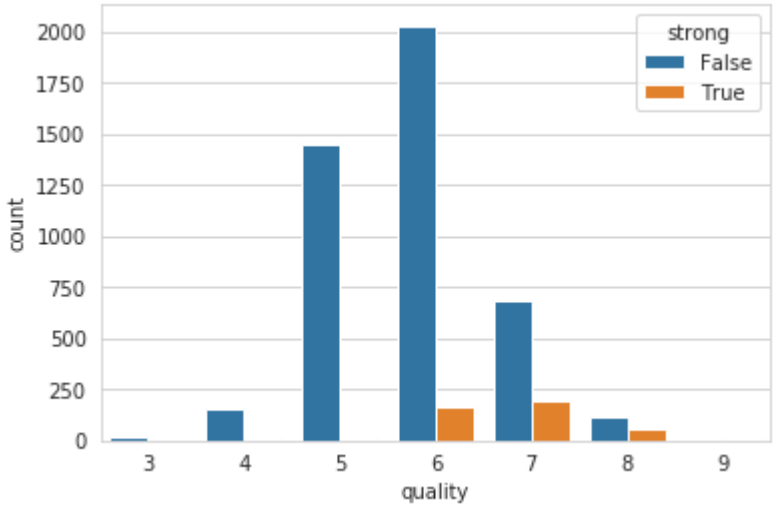
- Consider "apparent temperature" measures like the heat index and the wind chill. These quantities attempt to measure the perceived temperature to humans based on air temperature, humidity, and wind speed, things which we can measure directly.
- You could think of an apparent temperature as the result of a kind of feature engineering, an attempt to make the observed data more relevant to what we actually care about: how it actually feels outside!

Source: https://www.kaggle.com/learn/feature-engineering
Source: https://towardsdatascience.com/feature-engineering-for-machine-learning-3a5e293a5114

EDA (Wine Quality Dataset) with categorical feature

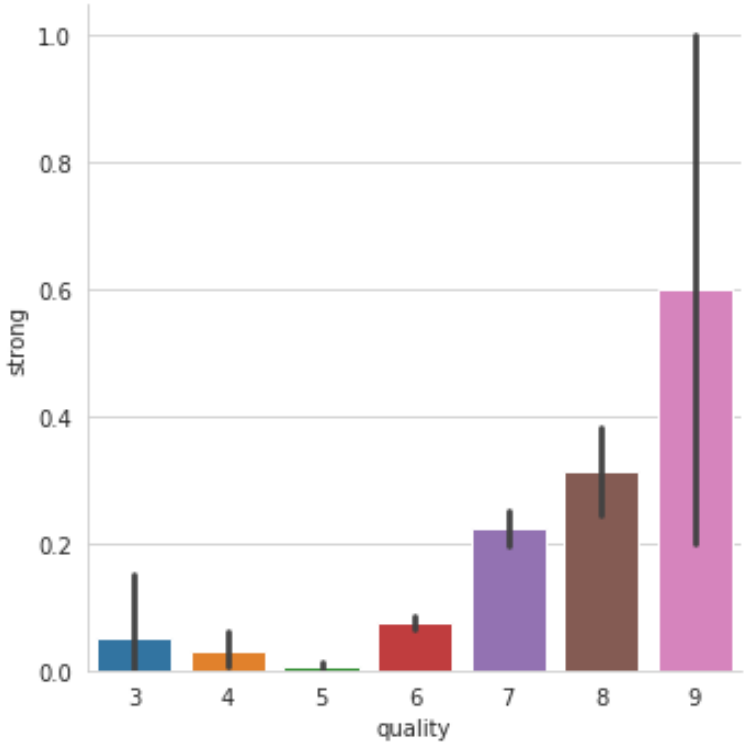
```
[60] 1 # Create a new feature call strong - if alcohol level is >= 12.5
     2 df["strong"] = df["alcohol"] >= 12.5
```

```
[61] 1 # Plot the relationship between different quality and strong alcohol
     2 sns.countplot(x="quality",hue="strong",data=df)
```



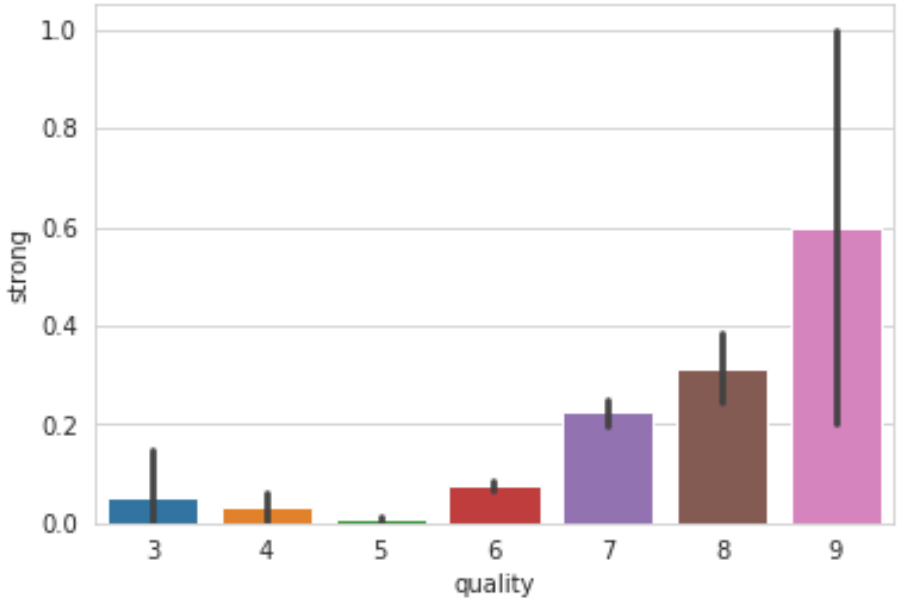
EDA (Wine Quality Dataset) with categorical feature

```
[62]  1 # Plot the relationship between different quality and strong alcohol
      2 sns.catplot(x='quality',y='strong', data=df,kind='bar')
```



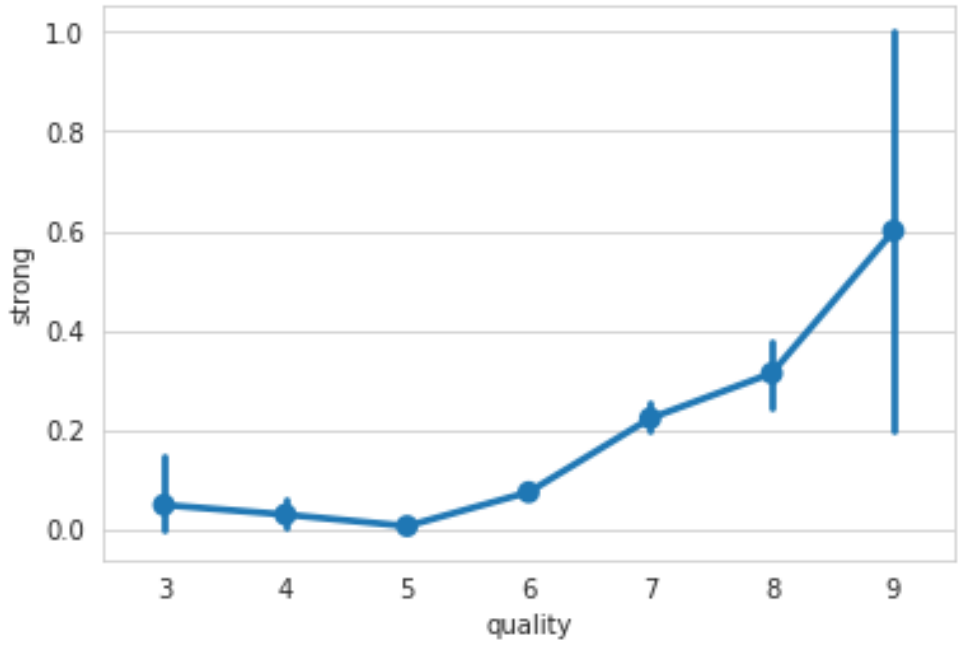
EDA (Wine Quality Dataset) with categorical feature

```
1 # Plot the relationship between different quality and alcohol level
2 sns.barplot(x="quality",y="strong",data=df)
```



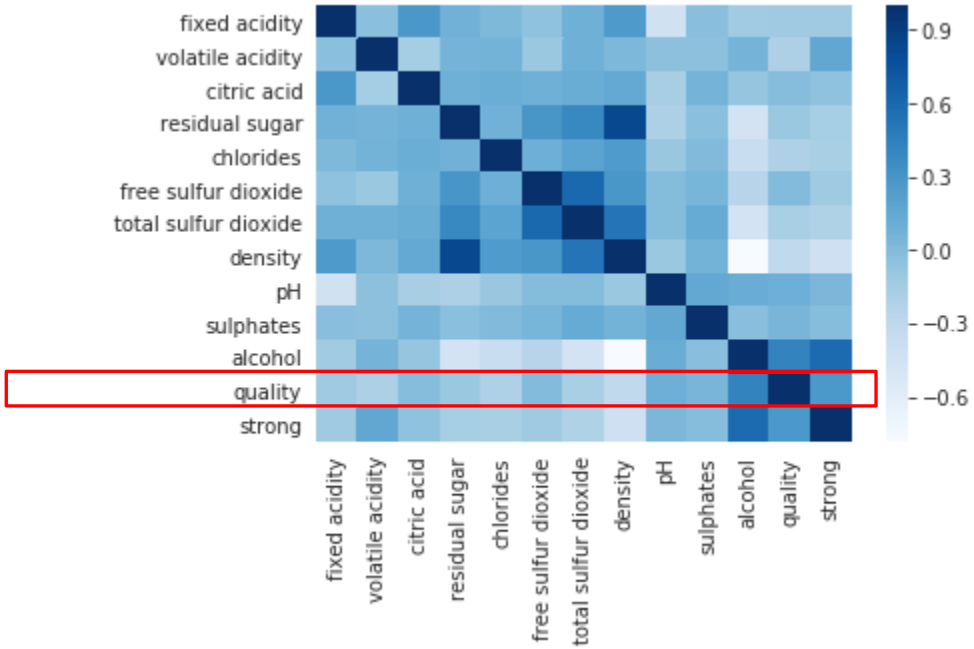
EDA (Wine Quality Dataset) with categorical feature

```
# Plot the relationship between different quality and strong alcohol level
sns.pointplot(x="quality",y="strong",data=df)
```



EDA (Wine Quality Dataset) Heatmap with corr()

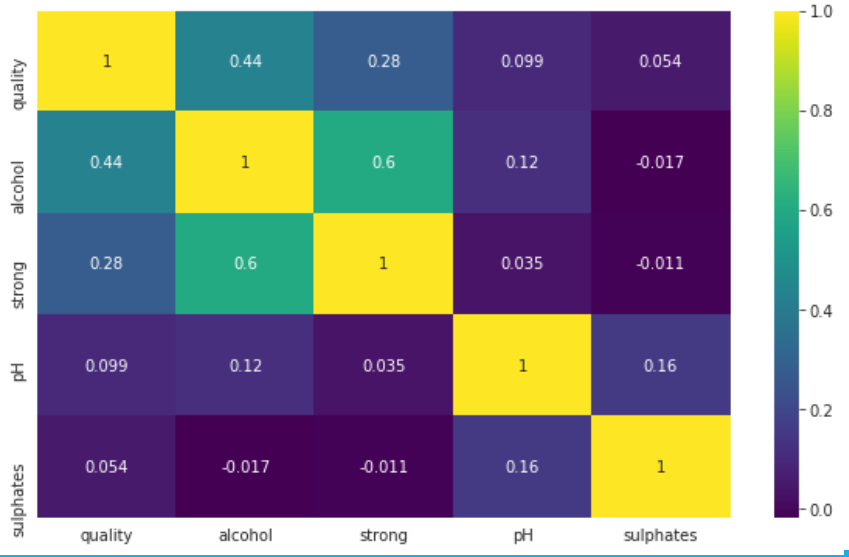
```
1 plt.figure(figsize=(6,4))
2 sns.heatmap(df.corr(),cmap='Blues',annot=False)
```



strong is one of the better Predictors for quality

EDA (Wine Quality Dataset) Heatmap with corr()

```
#Quality correlation matrix
k = 5 # top 5 number of variables for heatmap
cols = df.corr().nlargest(k, 'quality')['quality'].index
cm = df[cols].corr()
plt.figure(figsize=(10,6))
sns.heatmap(cm, annot=True, cmap = 'viridis')
```



Understand Model Evaluation Techniques

Model evaluation: checking if the algorithm/model learnt from the data is good enough to be put to use

- The key idea is that the algorithm/model that has been trained should be able to make good predictions for new data.
- In the absence of new data during development, we set aside some data from the original dataset as the test dataset. The test data stands in as a proxy for future new data.
- The remaining data (training dataset) is used to build/train the model.
- We will calculate the "goodness" of the prediction using metrics (see next section).
- The metrics are calculated for both the training and test datasets.

Understand Evaluation Metrics and Scoring

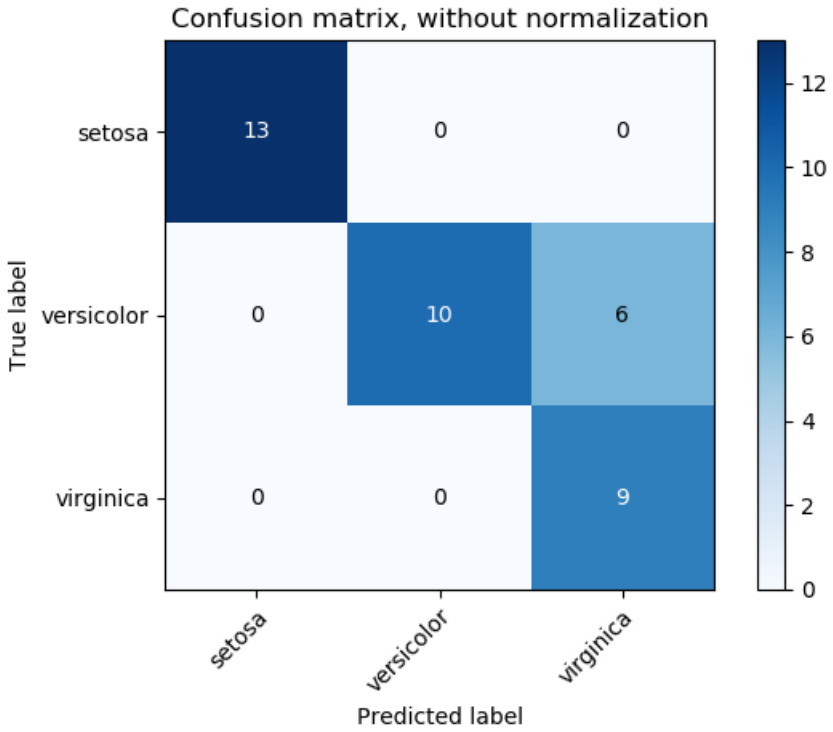
Model evaluation for classifiers

- The metrics used for classification are totally different from those used regression.
- The metrics for classifiers measure that amount of misclassification (i.e. assigning the wrong label to the output) that happens.
- In scikit-learn, the metrics for regression are found at https://scikit-learn.org/stable/modules/model_evaluation.html#classification-metrics
- The accuracy_score function computes the accuracy, either the fraction (default) or the count (normalize=False) of correct predictions.
- The best possible score is 1.0 (100% accurate). The worse is 0.5 which corresponds to random guessing. Score of 0.0 mean there is mislabelling of the correct answer since being 100% wrong is not possible without a certain degree of correct scoring/prediction.

Metrics for Classification: Confusion Matrix

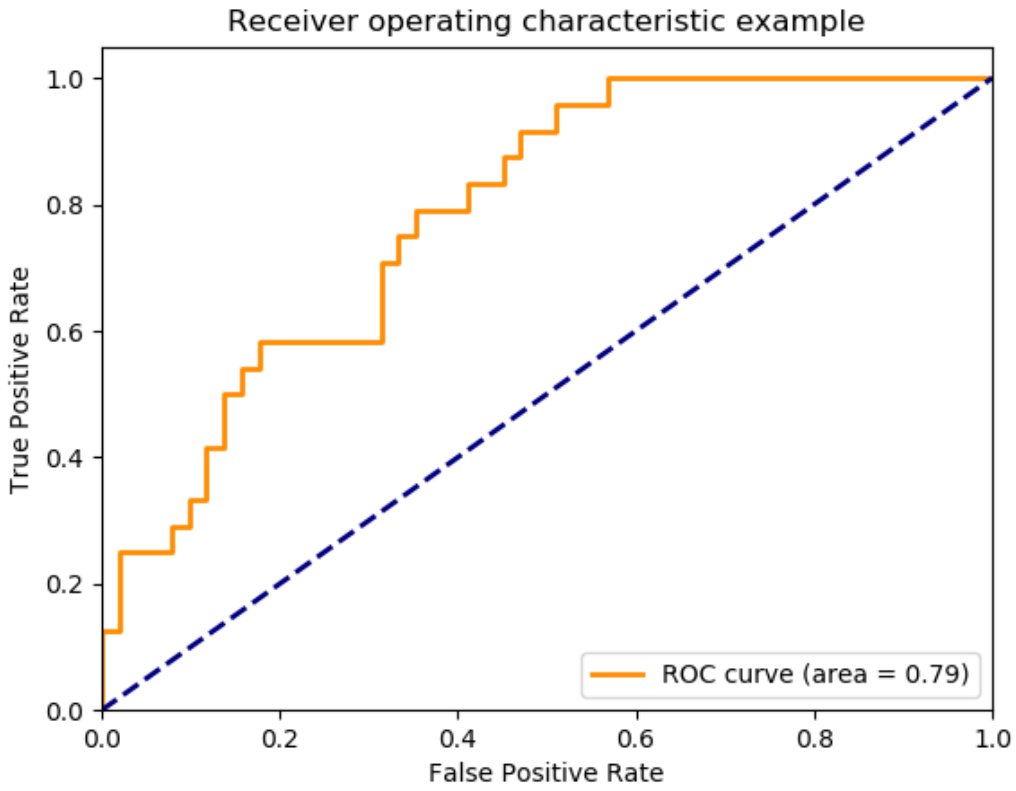
https://scikit-learn.org/stable/modules/model_evaluation.html#confusion-matrix

The confusion_matrix function evaluates classification accuracy by computing the confusion matrix with each row corresponding to the true class



Metrics for two-class/binary classification: ROC curve

https://scikit-learn.org/stable/modules/model_evaluation.html#roc-metrics



Metrics for two-class/binary classification: ROC curve

A receiver operating characteristic (ROC), or simply ROC curve, is a graphical plot which illustrates the performance of a binary classifier system as its discrimination threshold is varied. It is created by plotting the fraction of true positives out of the positives (TPR = true positive rate) vs. the fraction of false positives out of the negatives (FPR = false positive rate), at various threshold settings. TPR is also known as sensitivity, and FPR is one minus the specificity or true negative rate.

Metrics for two-class/binary classifier: Area Under ROC Curve (AUC)

The roc_auc_score function computes the area under the receiver operating characteristic (ROC) curve, which is also denoted by AUC or AUROC. By computing the area under the roc curve, the curve information is summarized in one number.

Best possible AUC is 1.0.

Worst AUC is 0.5 (random guessing)

Model evaluation for regressors

- The metrics used for regression are totally different from those used classification.
- The metrics used for regression basically measure the amount of error in the predictions by comparing the predicted value/score (yhat) with the actual value of the lable (y).
- The Residuals (error) e = y – yhat
- In scikit-learn, the metrics for regression are found at https://scikit-learn.org/stable/modules/model_evaluation.html#regression-metrics
- The r2_score function computes the coefficient of determination, usually denoted as R^2. The best possible r2_score (perfect) is 1.0.
- r2_score can be negative (because the model can be arbitrarily worse). A constant model, disregarding the input features, would get a R^2 score of 0.0.

Metrics for Regression

Mean squared error	MSE	=	$\frac{1}{n}\sum_{t=1}^{n} e_t^2$
Root mean squared error	RMSE	=	$\sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}$
Mean absolute error	MAE	=	$\frac{1}{n}\sum_{t=1}^{n} \lvert e_t \rvert$
Mean absolute percentage error	MAPE	=	$\frac{100\%}{n}\sum_{t=1}^{n} \left\lvert \frac{e_t}{y_t} \right\rvert$

Summary

We have learnt that:

- Scikit-learn is a python library that contains many different implementations of machine learning (ML) algorithms
- Scikit-learn also has many helper/utility routines machine learning such as encoding categorical variables (pre-processing)
- Scikit-learn has routines for evaluating the results of the modelling process (model selection)
- Scikit-learn has some sample data sets included in the library