- Principal component analysis to select the components that account for the variation in the dataset
- Feature importances using ensembled or tree classifiers

Statistical Tests to Select the Best *k* Features Using the SelectKBest Module

The following list is a selection of statistical tests to use with **SelectKBest**. The choice depends if the dataset target variable is numerical or categorical:

- ANOVA F-value, **f_classif** (classification)
- Chi-squared stats of non-negative features, **chi2** (classification)
- F-value, **f_regression** (regression)
- Mutual information for a continuous target, **mutual_info_regression** (regression)

Let's see an example using chi-squared test to select the best variables.

```
# import packages
from sklearn import datasets
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2

# load dataset
data = datasets.load_iris()

# separate features and target
X = data.data
y = data.target

# display first 5 rows
X[0:5,:]

# feature engineering. Let's see the best 3 features by setting k = 3
kBest_chi = SelectKBest(score_func=chi2, k=3)
fit_test = kBest_chi.fit(X, y)
```

```
# print test scores
fit_test.scores_
'Output': array([ 10.81782088,   3.59449902, 116.16984746,  67.24482759])
```

From the test scores, the top 3 important features in the dataset are ranked from feature 3 to 4 to 1 and to 2 in order. The data scientist can choose to drop the second column and observe the effect on the model performance.

We can transform the dataset to subset only the important features.

```
adjusted_features = fit_test.transform(X)
adjusted_features[0:5,:]
'Output':
array([[5.1, 1.4, 0.2],
       [4.9, 1.4, 0.2],
       [4.7, 1.3, 0.2],
       [4.6, 1.5, 0.2],
       [5. , 1.4, 0.2]])
```

The result drops the second column of the dataset.

Recursive Feature Elimination (RFE)

RFE is used together with a learning model to recursively select the desired number of top performing features.

Let's use RFE with **LinearRegression**.

```
# import packages
from sklearn.feature_selection import RFE
from sklearn.linear_model import LinearRegression
from sklearn import datasets

# load dataset
data = datasets.load_boston()

# separate features and target
X = data.data
y = data.target
```