# Principal Component Analysis in scikit-learn

## Introduction

Now that you've seen the curse of dimensionality, it's time to take a look at a dimensionality reduction technique! This will help you overcome the challenges of the curse of dimensionality (amongst other things). Essentially, PCA, or Principal Component Analysis, attempts to capture as much information from the dataset as possible while reducing the overall number of features.

## Objectives

You will be able to:

- Explain at a high level how PCA works
- Explain use cases for PCA
- Implement PCA using the scikit-learn library
- Determine the optimal number of n components when performing PCA by observing the explained variance

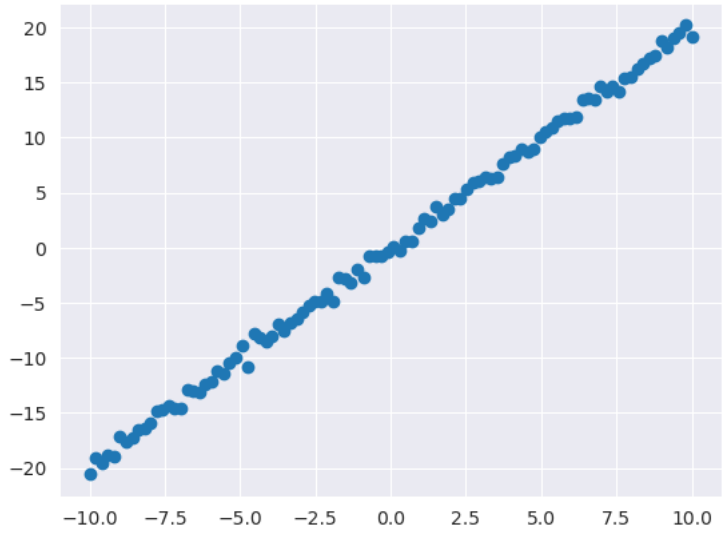
## Generate some data

First, you need some data to perform PCA on. With that, here's a quick dataset you can generate using NumPy:

In [1]:
```
import numpy as np

x1 = np.linspace(-10, 10, 100)
# A linear relationship, plus a little noise
x2 = np.array([xi*2 + np.random.normal(loc=0, scale=0.5) for xi in x1])

X = np.matrix(list(zip(x1, x2)))
```

Let's also generate a quick plot of this simple dataset to further orient ourselves:

In [2]:
```
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline

sns.set_style('darkgrid')

plt.scatter(x1, x2);
```




## PCA with scikit-learn

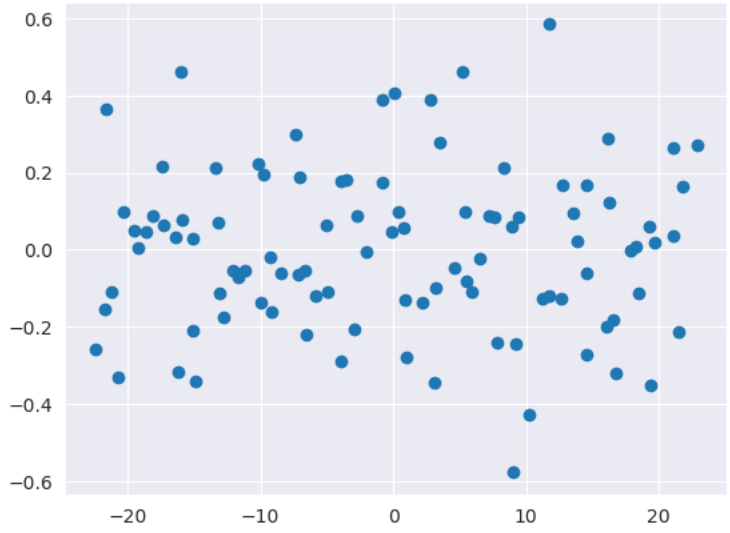
Now onto PCA. First, take a look at how simple it is to implement PCA with scikit-learn:

In [4]:
```
from sklearn.decomposition import PCA

pca = PCA()
transformed = pca.fit_transform(X)
```

```
/opt/saturncloud/envs/saturn/lib/python3.10/site-packages/sklearn/utils/validation.py:727: FutureWarning: np.matrix usage is deprecated in 1.0 and will raise a TypeError in 1.2. Please convert to a numpy array with np.asarray. For more information see: https://numpy.org/doc/stable/reference/generated/numpy.matrix.html (https://numpy.org/doc/stable/reference/generated/numpy.matrix.html)
  warnings.warn(
```

And you can once again plot the updated dataset:

In [5]:
```
plt.scatter(transformed[:,0], transformed[:,1]);
```




In [6]:
```
pca.components_
```

Out[6]:
```
array([[-0.4470939 , -0.89448703],
       [ 0.89448703, -0.4470939 ]])
```

In [7]:
```
pca.mean_
```

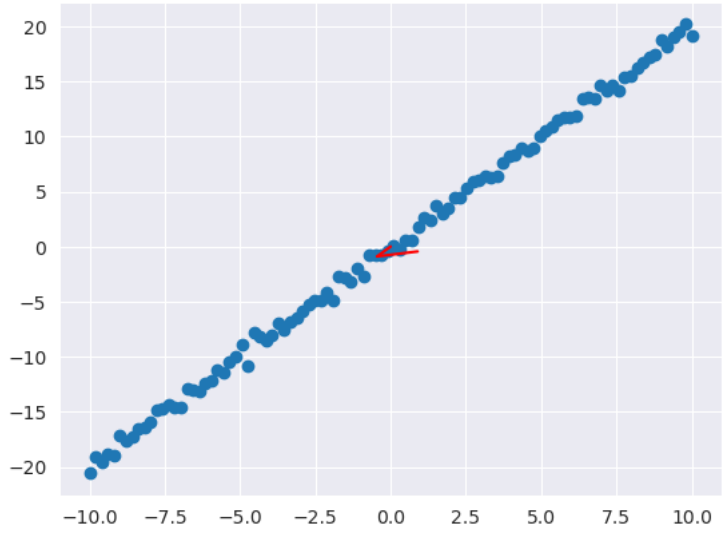
Out[7]:
```
array([7.10542736e-17, 6.77182210e-03])
```

## Interpret Results

Let's take a look at what went on here. PCA transforms the dataset along principal axes. The first of these axes is designed to capture the maximum variance within the data. From here, additional axes are constructed which are orthogonal to the previous axes and continue to account for as much of the remaining variance as possible.

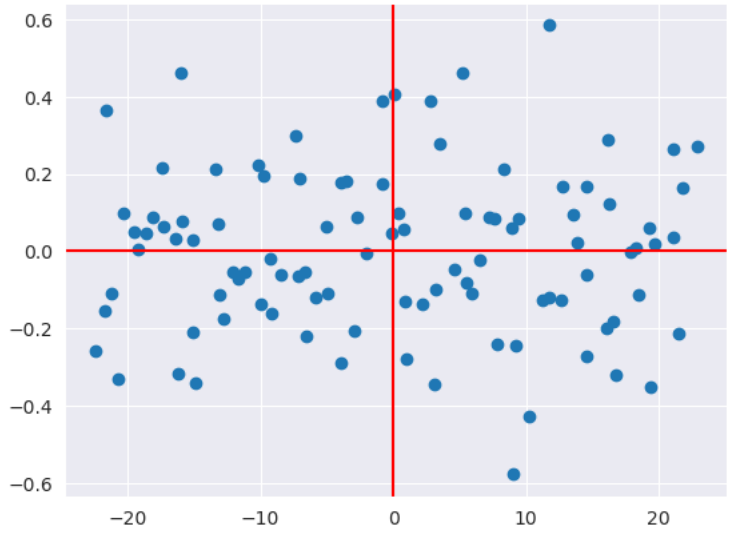
For the current 2-d case, the axes which the data were projected onto look like this:

In [8]:
```
plt.scatter(x1, x2);
ax1, ay1 = pca.mean_[0], pca.mean_[1]
ax2, ay2 = pca.mean_[0] + pca.components_[0][0], pca.mean_[1] + pca.components_[0][1]
ax3, ay3 = pca.mean_[0] + pca.components_[1][0], pca.mean_[1] + pca.components_[1][1]
plt.plot([ax1, ax2], [ay1, ay2], color='red')
plt.plot([ax2, ax3], [ay2, ay3], color='red');
```



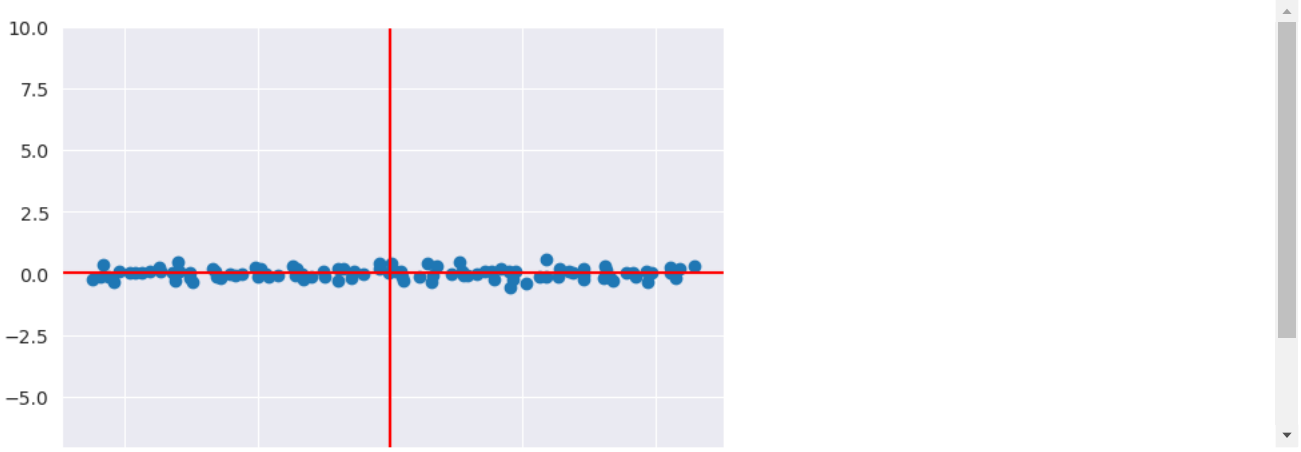

So, the updated graph you saw is the same dataset rotated onto these red axes:

In [9]:
```
plt.scatter(transformed[:,0], transformed[:,1])
plt.axhline(color='red')
plt.axvline(color='red');
```

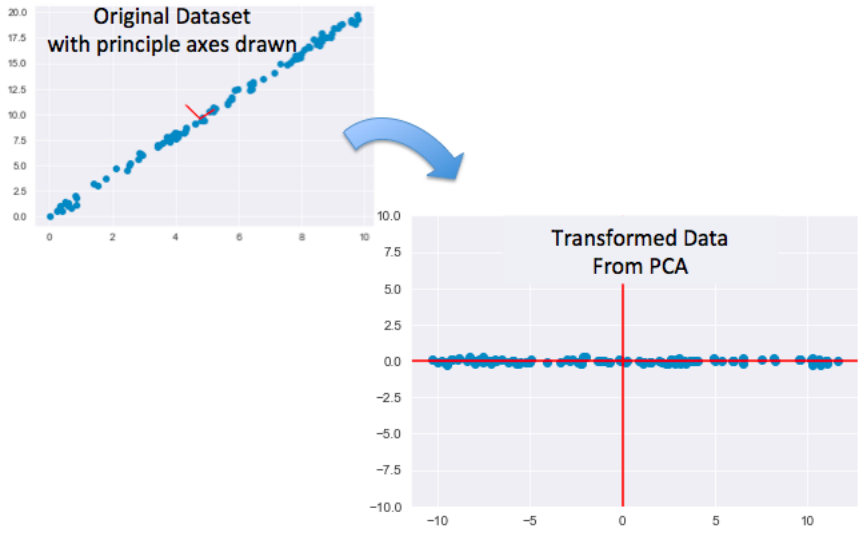



Note the small scale of the y-axis. You can also plot the transformed dataset on the new axes with a scale similar to what you saw before:

In [10]:
```
plt.scatter(transformed[:,0], transformed[:,1])
plt.axhline(color='red')
plt.axvline(color='red')
plt.ylim(-10,10);
```




Again, this is the geographical interpretation of what just happened:




## Determine the Explained Variance

Typically, one would use PCA to actually reduce the number of dimensions. In this case, you've simply re-re-parametrized the dataset along new axes. That said, if you look at the first of these primary axes, you can see the patterns encapsulated by the principal component. Moreover, scikit-learn also lets you quickly determine the overall variance in the dataset accounted for in each of the principal components.

In [10]:
```
pca.explained_variance_ratio_
```

Out[10]:
```
array([9.99717770e-01, 2.82229957e-04])
```

Keep in mind that these quantities are cumulative: principal component 2 attempts to account for the variance not accounted for in the primary component. You can view the total variance using `np.cumsum()` :

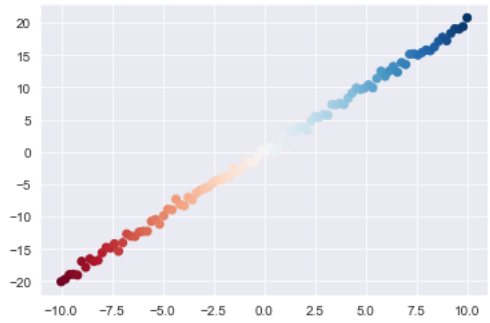
In [11]:
```
np.cumsum(pca.explained_variance_ratio_)
```

Out[11]:
```
array([0.99971777, 1.        ])
```

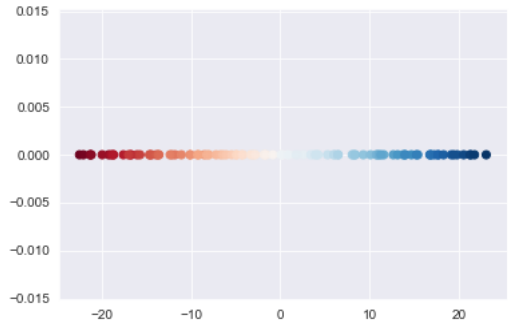
## Visualize the Principal Component

To help demonstrate the structure captured by the first principal component, observe the impact of coloring the dataset and then visualizing the first component.

In [12]:
```
plt.scatter(x1,x2, c=sns.color_palette('RdBu', n_colors=100));
```




In [13]:
```
plt.scatter(transformed[:,0], [0 for i in range(100)] , c=sns.color_palette('RdBu', n_colors=100));
```




## Steps for Performing PCA

The theory behind PCA rests upon many foundational concepts of linear algebra. After all, PCA is re-encoding a dataset into an alternative basis (the axes). Here are the exact steps:

1. Recenter each feature of the dataset by subtracting that feature's mean from the feature vector
2. Calculate the covariance matrix for your centered dataset
3. Calculate the eigenvectors of the covariance matrix
    A. You'll further investigate the concept of eigenvectors in the upcoming lesson
4. Project the dataset into the new feature space: Multiply the eigenvectors by the mean-centered features

You can see some of these intermediate steps from the `pca` instance object itself.

In [11]:
```
# Pulling up the original feature means which were used to center the data
pca.mean_
```

Out[11]:
```
array([7.10542736e-17, 6.77182210e-03])
```

In [12]:
```
# Pulling up the covariance matrix of the mean centered data
pca.get_covariance()
```

Out[12]:
```
array([[ 34.35023637,  68.63313472],
       [ 68.63313472, 137.35735283]])
```

In [13]:
```
# Pulling up the eigenvectors of the covariance matrix
pca.components_
```

Out[13]:
```
array([[-0.4470939 , -0.89448703],
       [ 0.89448703, -0.4470939 ]])
```

## Summary

In this lesson, you looked at implementing PCA with scikit-learn and the geometric interpretations of principal components. From here, you'll get a chance to practice implementing PCA yourself before going on to code some of the underlying components implemented by scikit-learn using NumPy.