Principal Component Analysis (PCA)

PCA

- Unsupervised Machine Learning technique to reduce Dimensionality

- PCA aims to extract ***p*** features from the total ***f*** features (p≤m) of a dataset
  - ➢ Which feature is more valuable to cluster the data
  - ➢ To explain the most variance of the dataset
  - ➢ Regardless of dependent variable (**ŷ**)
  - ➢ Visualisation becomes better and more informative

- PCA performed on a correlation matrix
  - Numeric Dataset
  - Standardized dataset *

Typical problems with a huge dataset

- Unwanted features
- Features may exhibit multicollinearity
- Features may exhibit singularity
- Indecisiveness + may lead to building a bad model *

+

  - Right set of features
  - Right algorithm

*

  - Overfit
  - Underfit
  - Poor Accuracy

- Formula for the Number of (scatter) plots on a given dataset = **$p(p-1) / 2$**

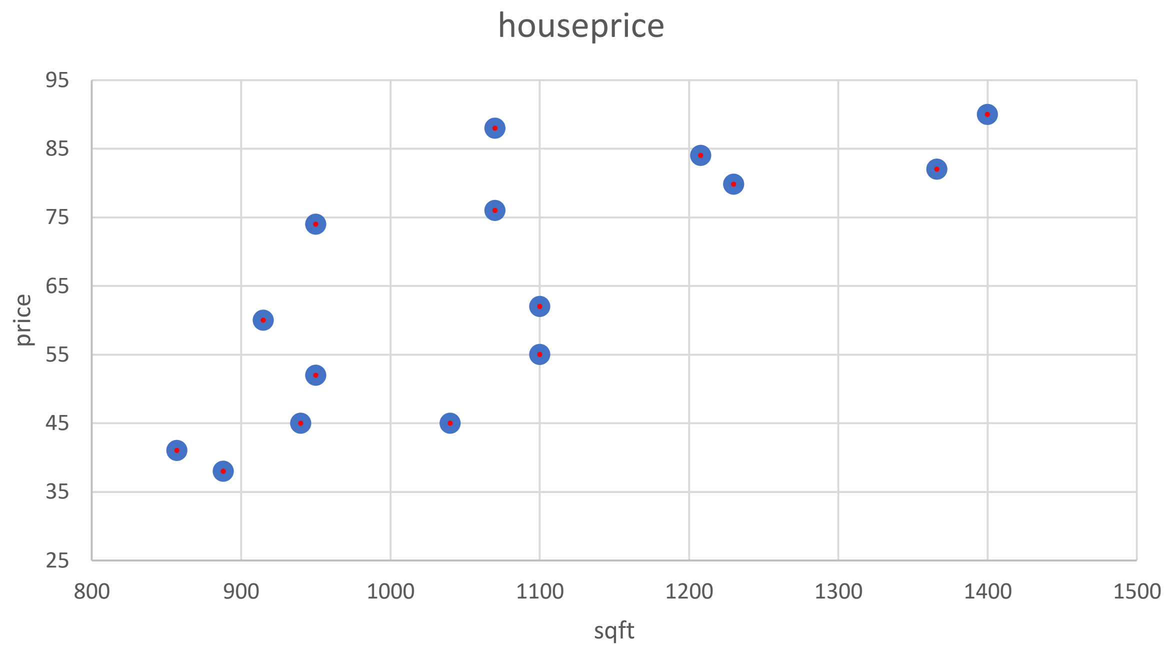
p = number of features / predictors / independent variables

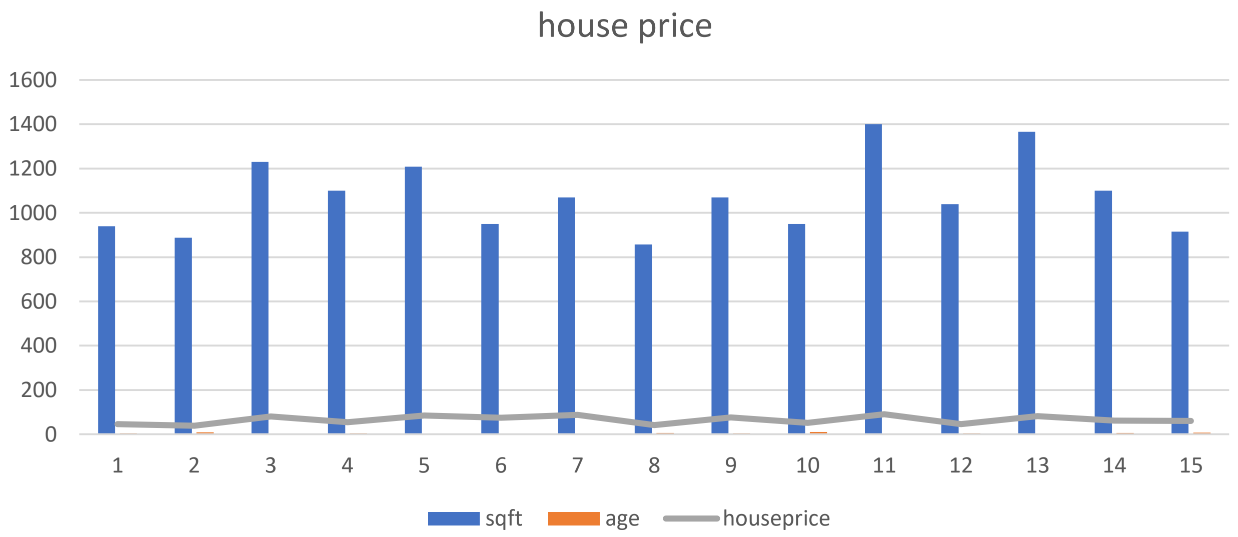
- Greater the value of p (i.e. more features), greater will the plots
  - Difficult and tedious to perform analysis

Principal Components

- A Principal Component is a linear combination of normalized features from the original dataset
- PC's can be
  - **Z1**: First Principal Component
    - ✓ Captures the maximum variance
    - ✓ Captures the highest variability
  - Z2: Second Principal Component
    - ✓ Captures the remaining variance
    - ✓ Uncorrelated to Z1
  - **Z3, Z4 ….**
- The number of PC's that can be constructed for a **nxp** dataset is:
  - ✓ **min(n-1, p)**

sqft	age	houseprice
940	5	45
888	9	38
1230	2	79.8
1100	4	55
1208	1	84
950	3	74
1070	2	88
857	6	41
1070	4	76
950	10	52
1400	1	90
1040	4	45
1366	3	82
1100	6	62
915	8	60



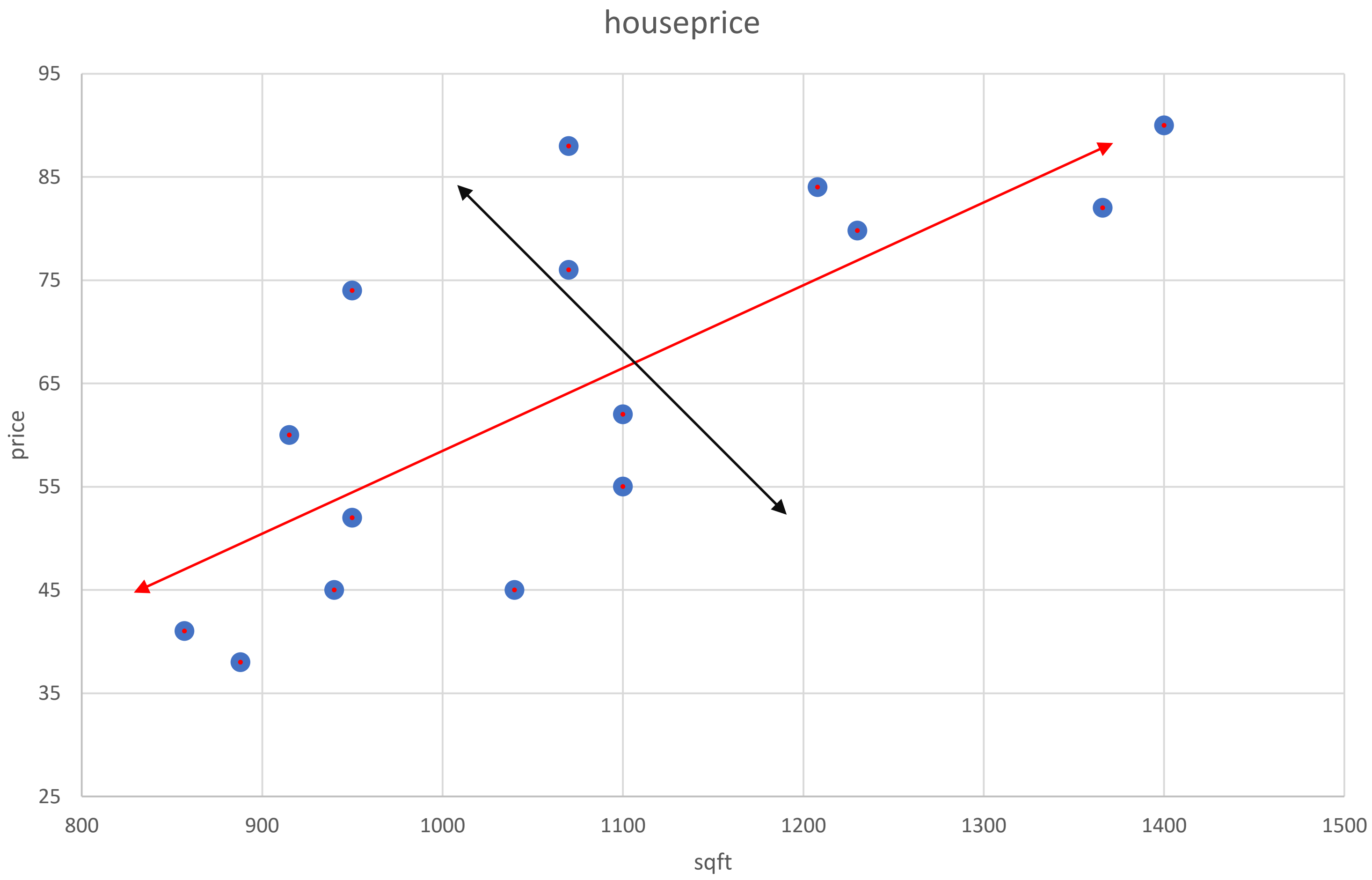


- With increase in dimensions, the chart becomes more difficult to plot

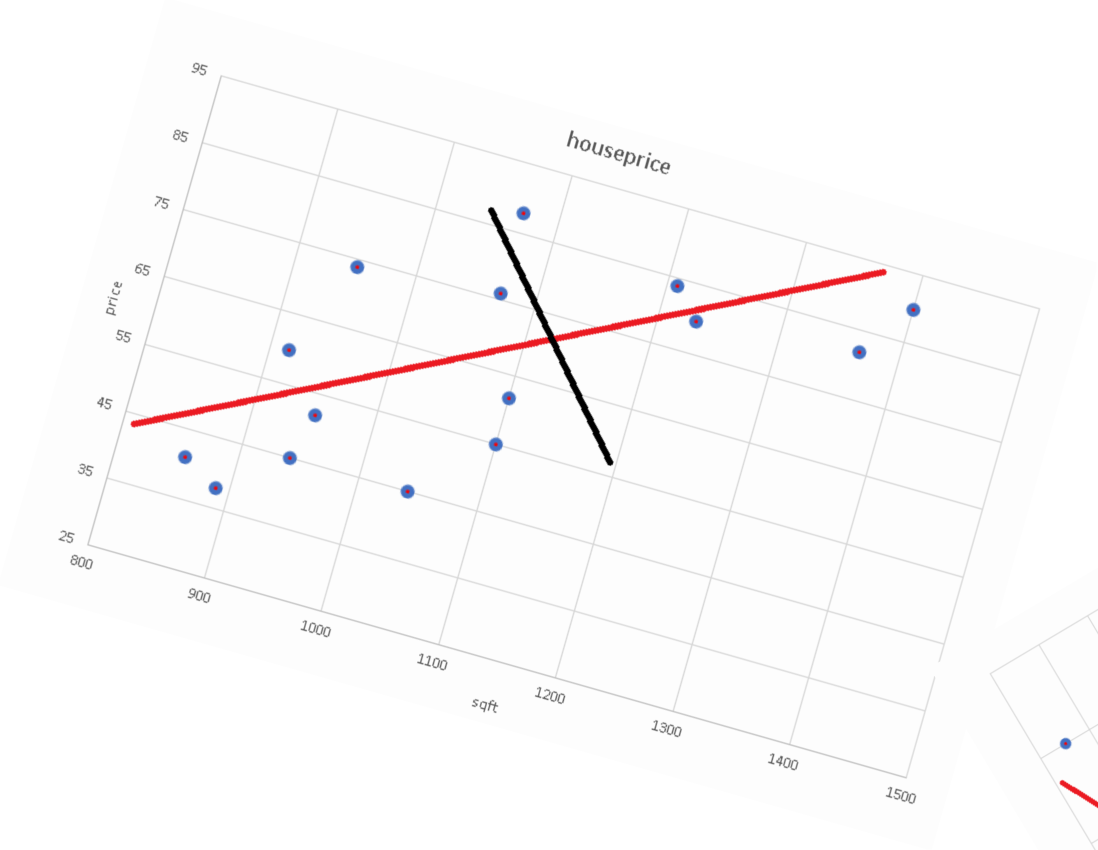
- Are all dimensions relevant / important ?

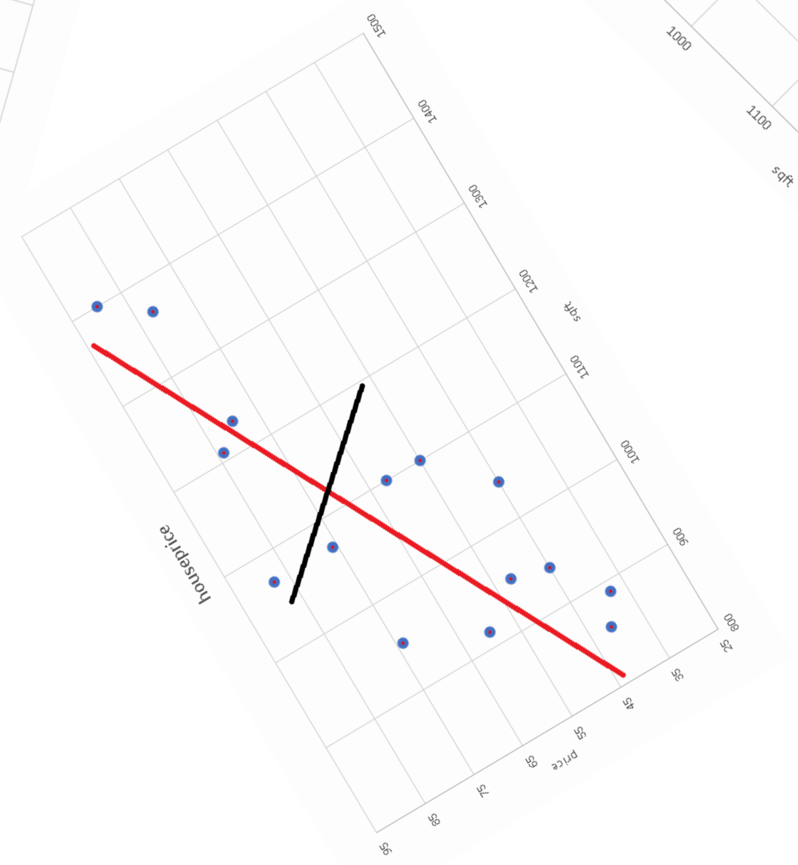
- 2-dimensional vs 3-dimensional movies
  - ➢ less information loss in 2-D

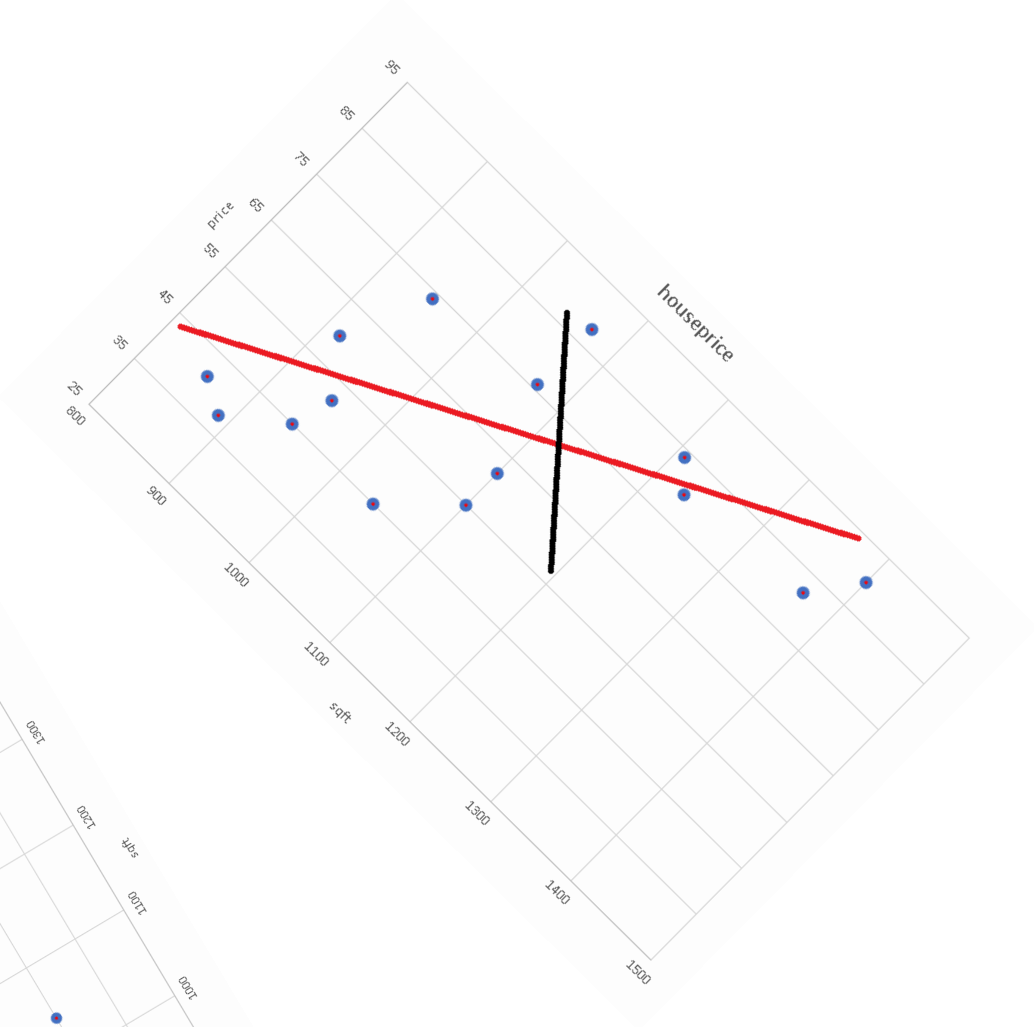
- PCA flattens dimensions



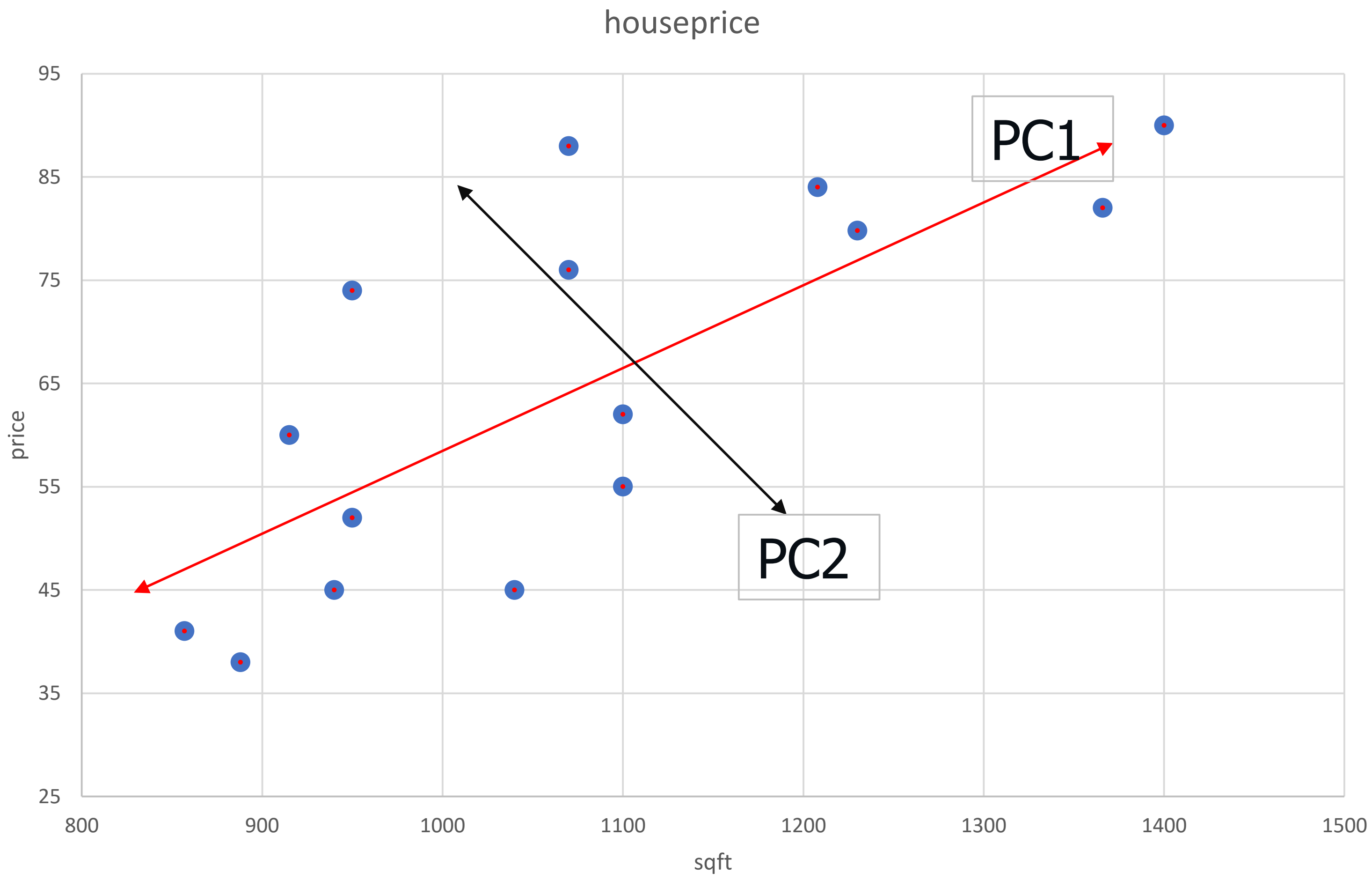
- Maximum variation
  - ➢ Left – Right (Red line)
- 2^{nd} most Variation
  - ➢ Top-bottom (Black line)







Rotation of the graph maintains the distances



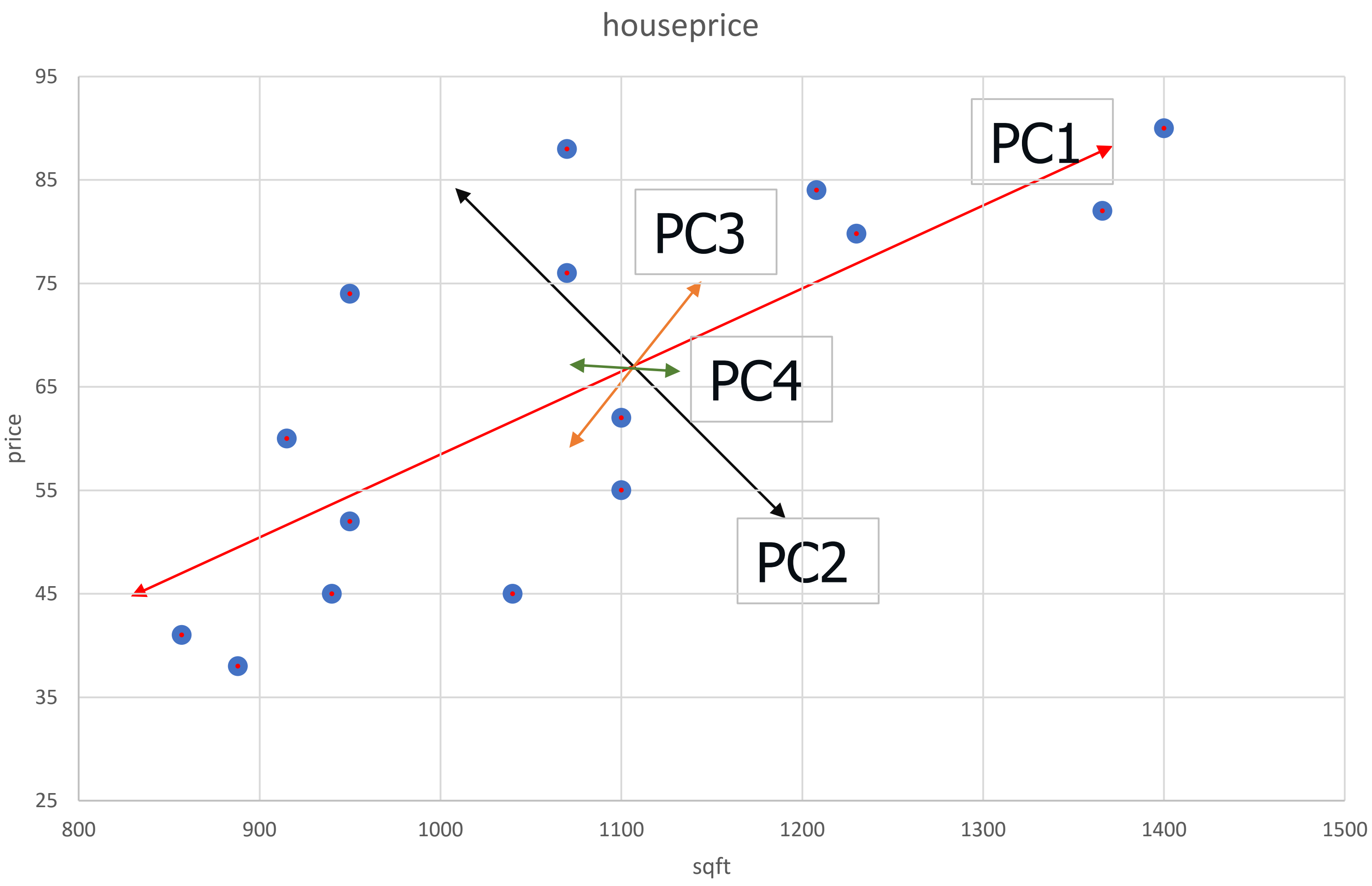
- Maximum variation
  - Left – Right (Red line)
  - PC1

- 2nd most Variation
  - Top-bottom (Black line)
  - PC2

Maximum variation is captured with these 2 lines

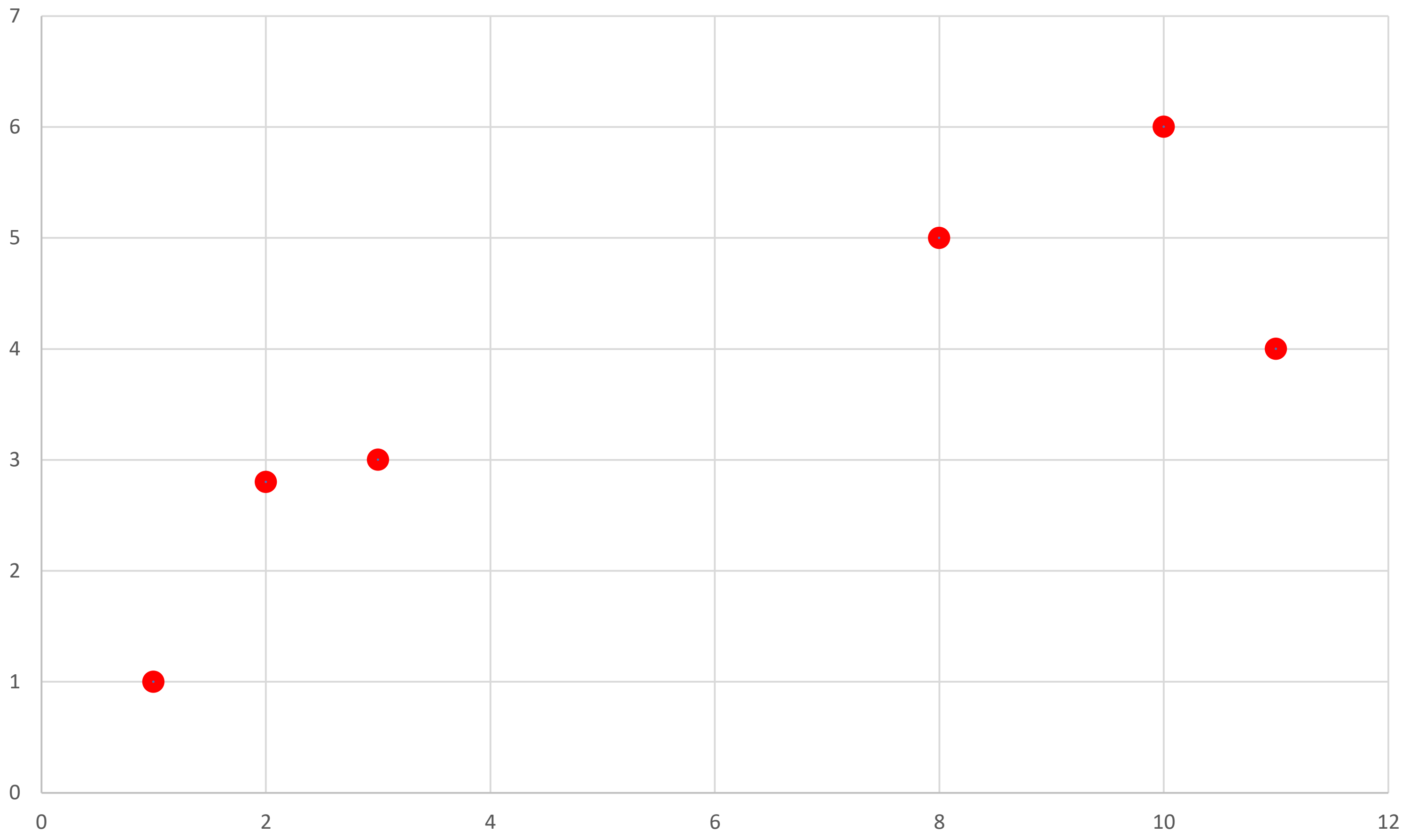
No real need for a diagonal line to capture more variation

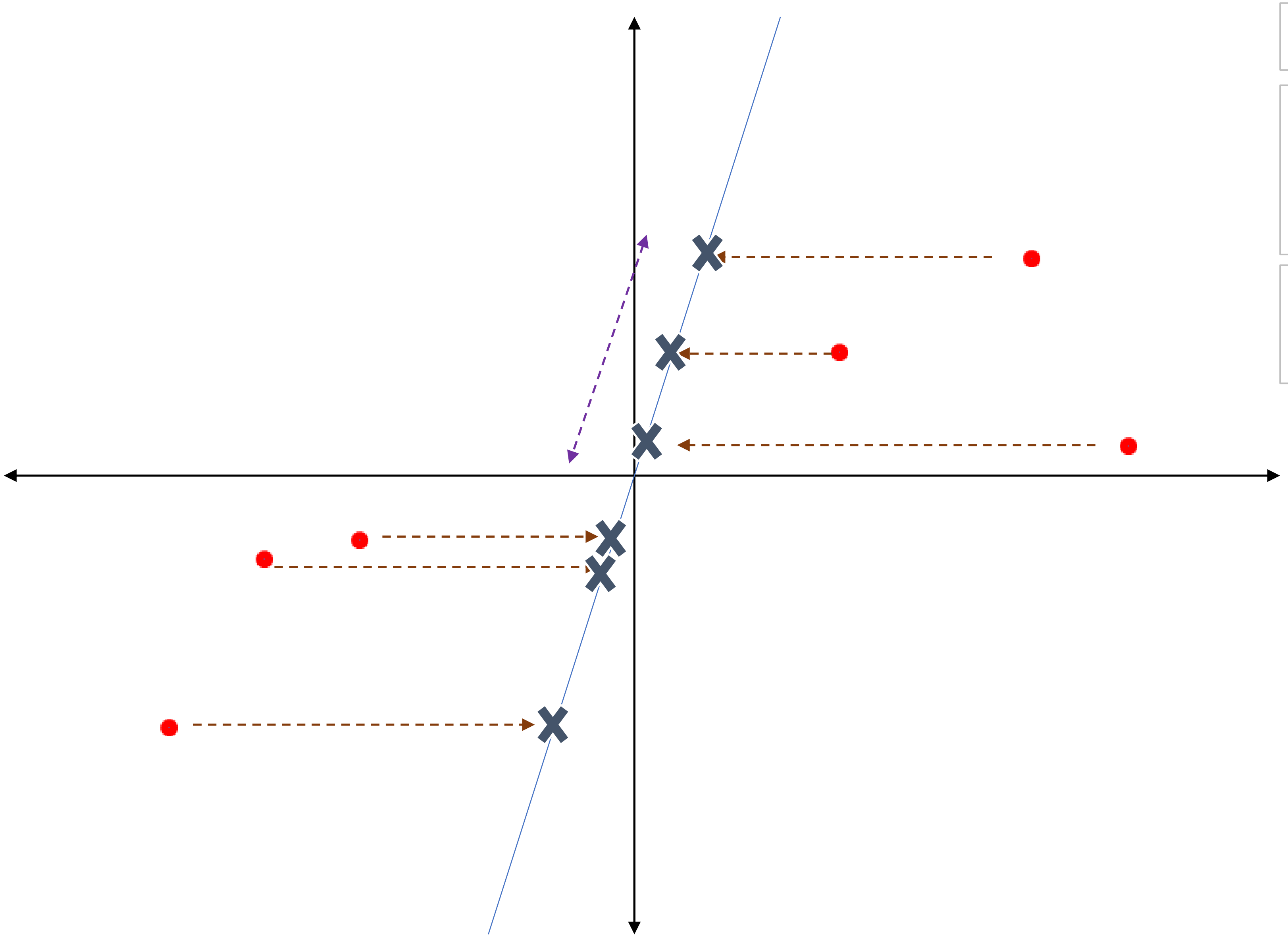
When there are multiple dimensions



- PC1
  - Maximum variation
- PC2
  - 2nd most variation
- PC3
  - 3rd most variation
- PC4
  - 4th most variation

Score 1	Score 2
10	6
11	4
8	5
3	3
2	2.8
1	1

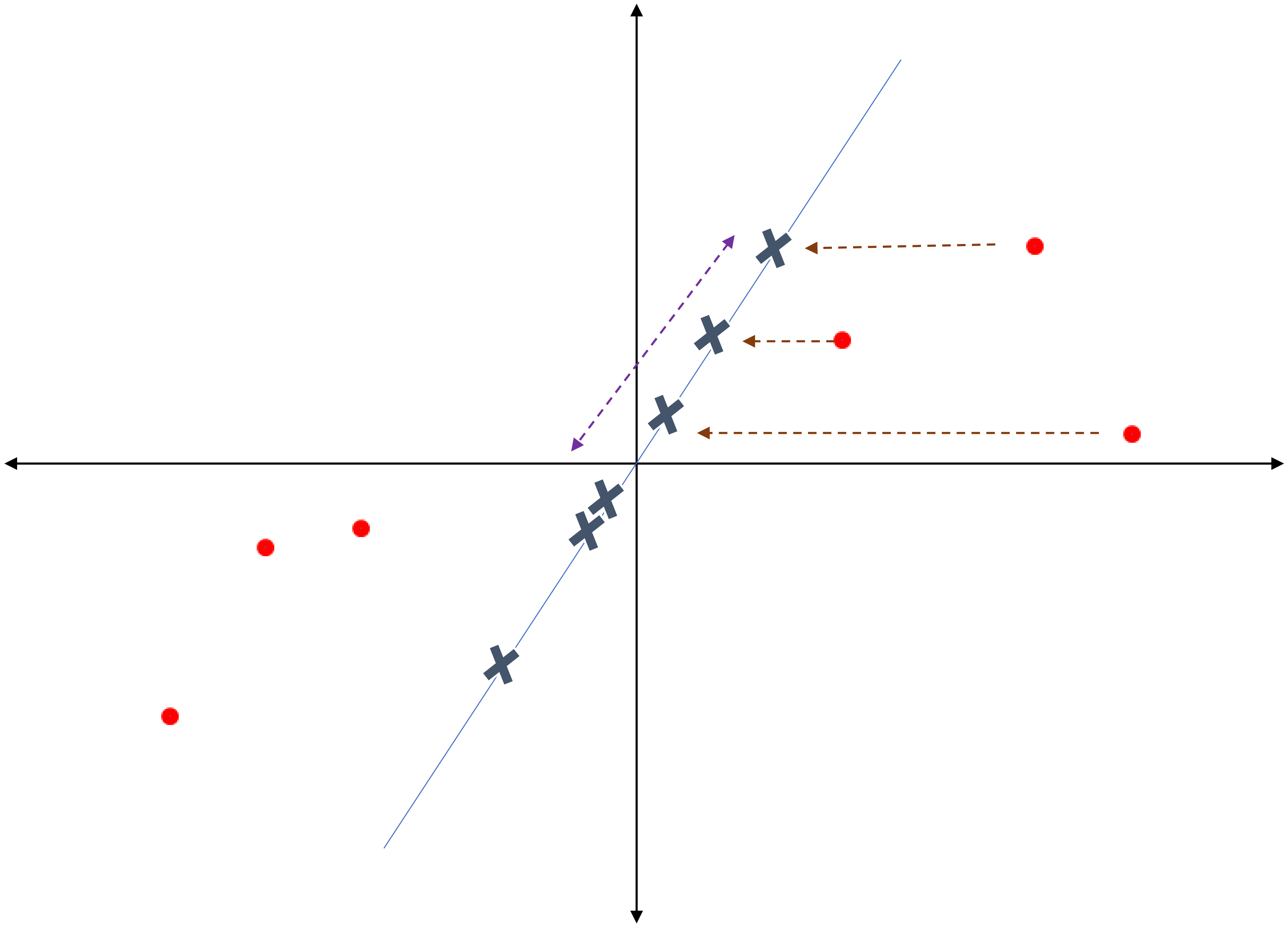


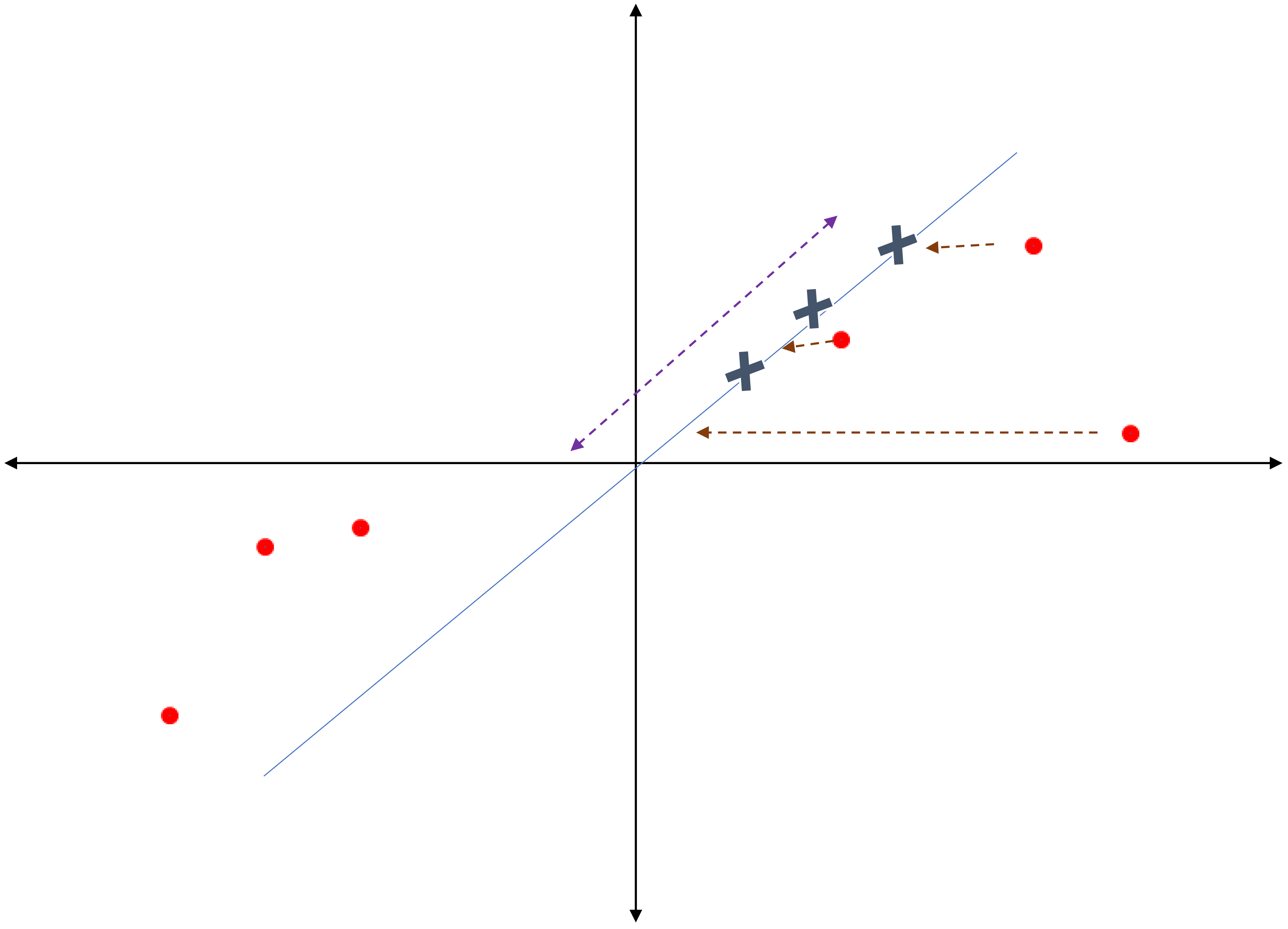


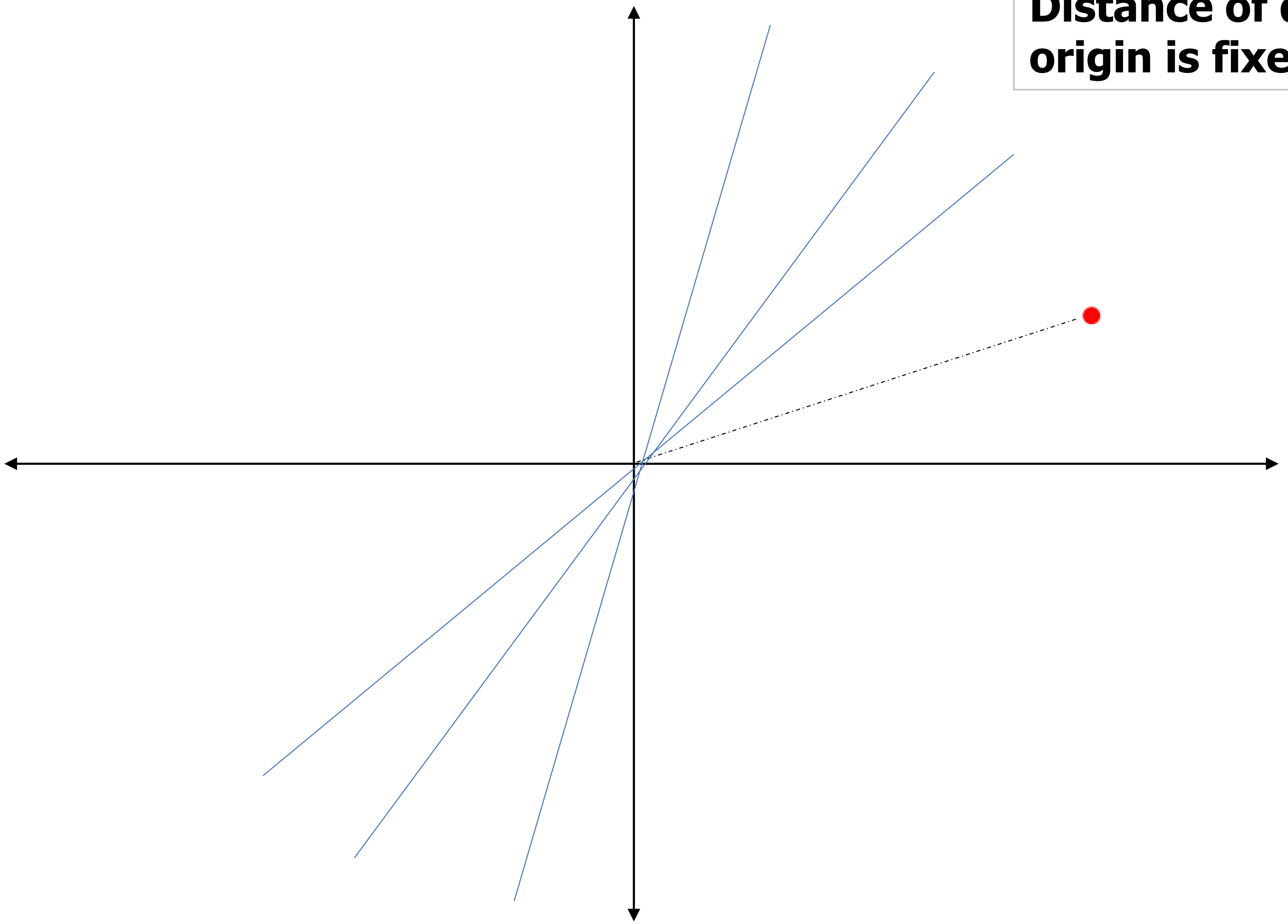
How good is this line ?

- Project Data on the line
- Measure distance of data to line
- Minimise this distance

- Maximise the distance of projected data from origin

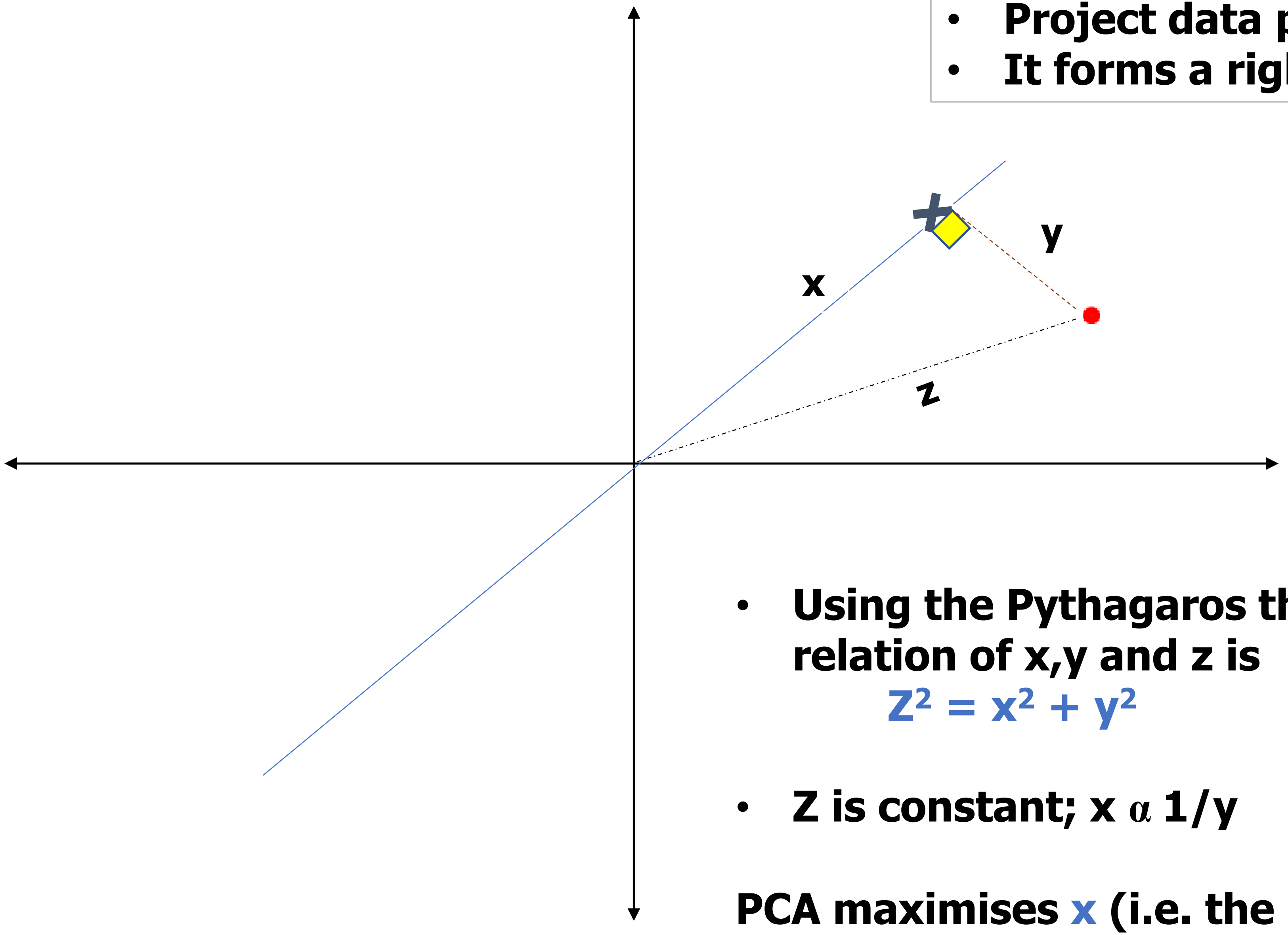






Distance of data point from origin is fixed

- Project data point on the line
- It forms a right angle

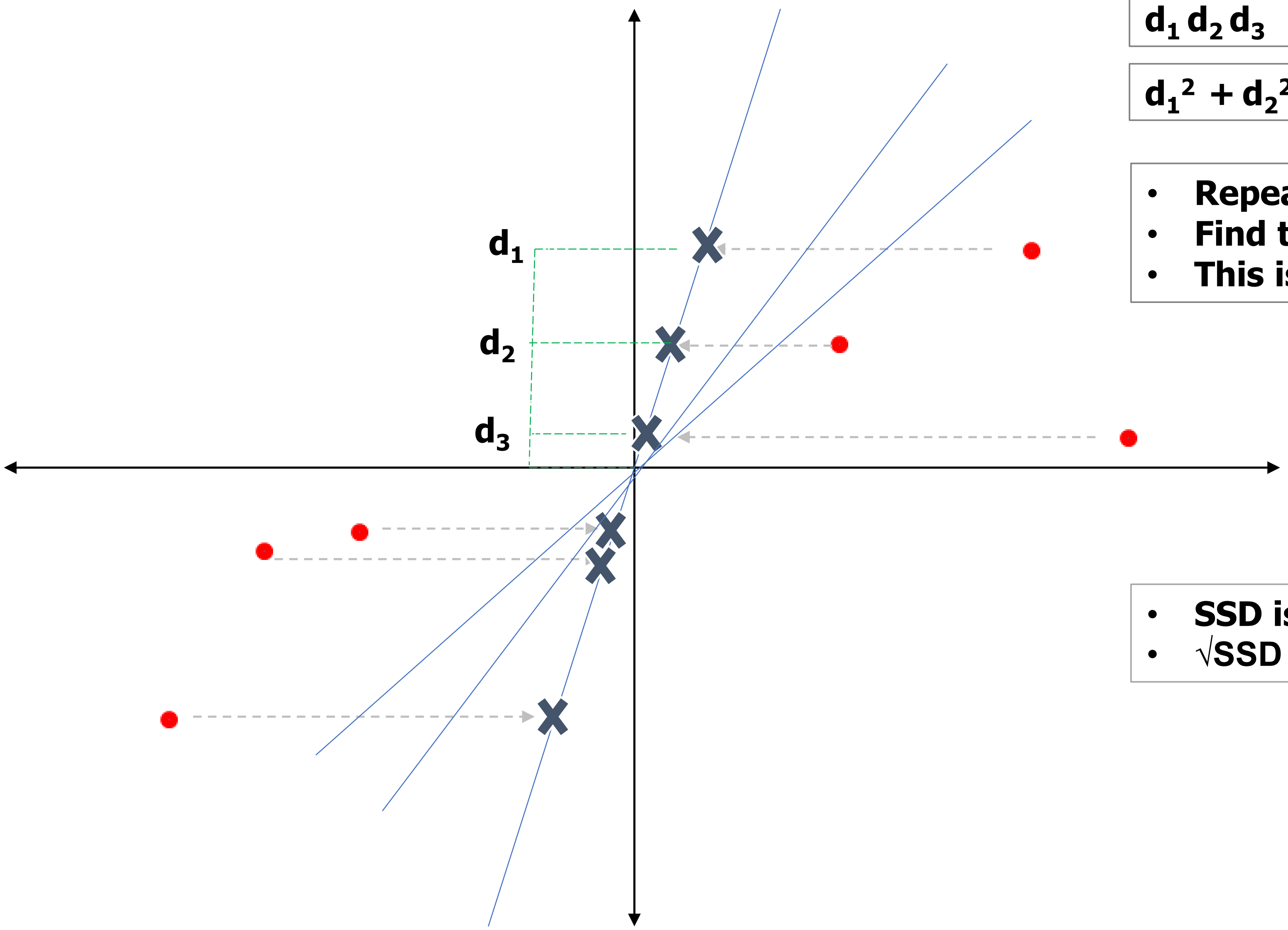


- Using the Pythagaros theorem, the relation of x,y and z is

  $$Z^2 = x^2 + y^2$$

- Z is constant; x α 1/y

PCA maximises x (i.e. the distance of projected data points from the origin)

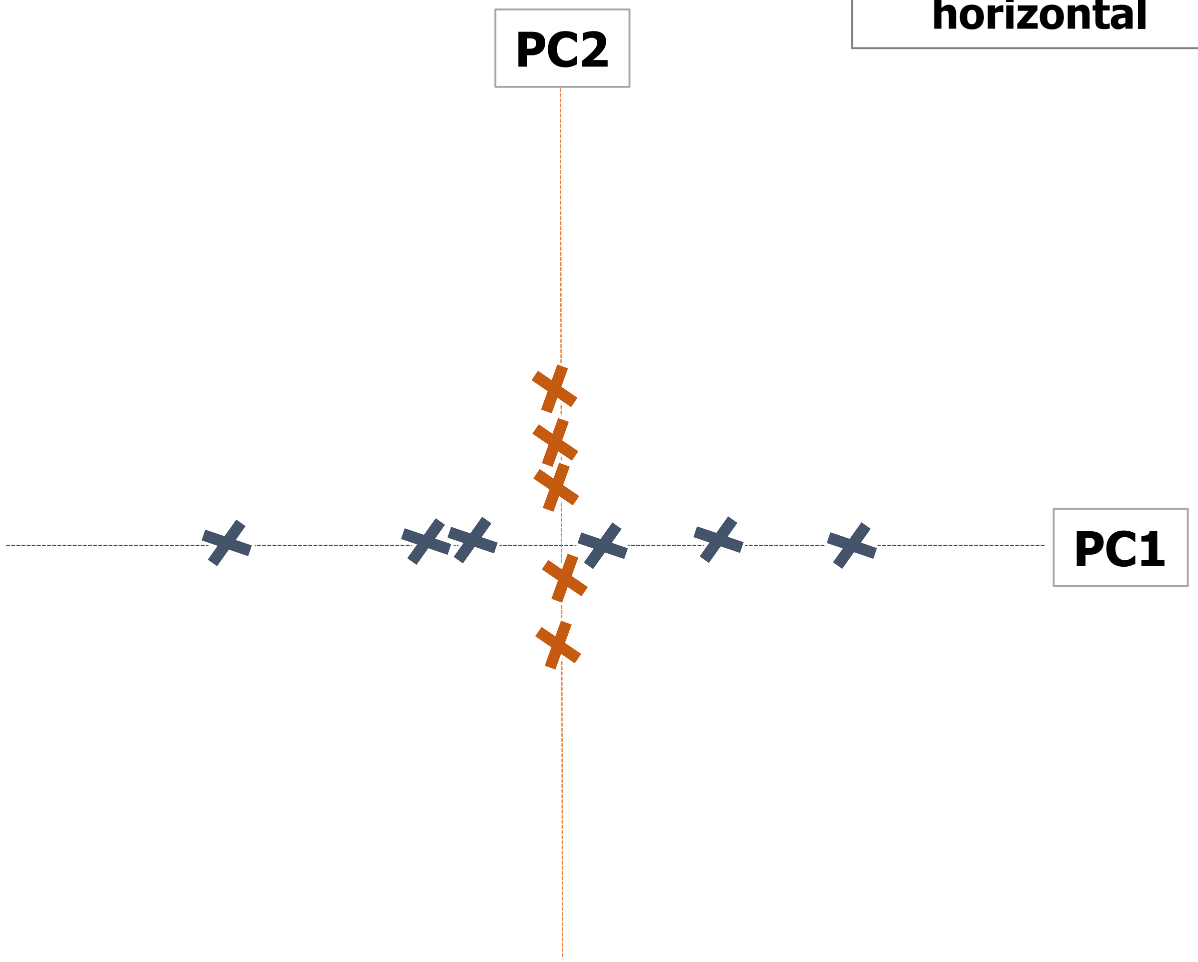


d_1 d_2 d_3	
$d_1^2 + d_2^2 + d_3^2$	**Sum of squared distances**

- **Repeat process of each projected line**
- **Find the largest SSD**
- **This is PC1**

- **SSD is referred to as Eigenvalue for PC1**
- **√SSD is referred to as Singular value for PC1**

- **Rotate everything such that PC1 lies horizontal**



Scree Plot

- **Graphical representation of percentages of variation of every Principal Component**

Principal Components – finer points

The first principal component results in a line which is closest to the data i.e. it minimizes the sum of squared distance between a data point and the line.

No other component can have variability higher than first principal component.

If the two components are uncorrelated, their directions should be orthogonal (90 degrees)

The principal components are supplied with normalized version of original predictors. This is because, the original predictors may have different scales. For example: Imagine a data set with variables' measuring units as gallons, kilometers, light years etc. It is definite that the scale of variances in these variables will be large.