Anomaly Detection

- Motivation and Introduction
- Supervised Methods
- Semisupervised Methods
- Unsupervised Methods:
  - Graphical and Statistical approaches
  - Nearest neighbor based approaches
  - Clustering based approaches
- Evaluation

Unsupervised Methods

Unsupervised Methods →

Training cases include anomalies and they are not labelled.

Tid	**SrcIP**	**Start time**	**Dest IP**	**Dest Port**	**Number of bytes**	**Attack**
1	206.135.38.95	11:07:20	160.94.179.223	139	192	No
2	206.163.37.95	11:13:56	160.94.179.219	139	195	No
3	206.163.37.95	11:14:29	160.94.179.217	139	180	No
4	206.163.37.95	11:14:30	160.94.179.255	139	199	No
5	206.163.37.95	11:14:32	160.94.179.254	139	19	Yes
6	206.163.37.95	11:14:35	160.94.179.253	139	177	No
7	206.163.37.95	11:14:36	160.94.179.252	139	172	No
8	206.163.37.95	11:14:38	160.94.179.251	139	285	Yes
9	206.163.37.95	11:14:41	160.94.179.250	139	195	No
10	206.163.37.95	11:14:44	160.94.179.249	139	163	Yes

Unsupervised Methods

Graphical approaches:

Given a database D, inspects it visually and determines which points are anomalies

Statistical-based:

Given a database D, and a data point **x** $\in$ D, a statistical test determines whether **x** is an anomaly or not, at a significance level p. These tests assume a latent distribution.

Distance-based:

There's available a distance measure which can be applied to any pair of data instances and is able to discriminate between the anomalies and normal instances well enough.

> Nearest neighbor based approaches

> Cluster based approaches

Anomaly Detection

- Motivation and Introduction
- Supervised Methods
- Semisupervised Methods
- Unsupervised Methods:
  - Graphical and Statistical approaches
  - Nearest neighbor based approaches
  - Clustering based approaches
- Evaluation

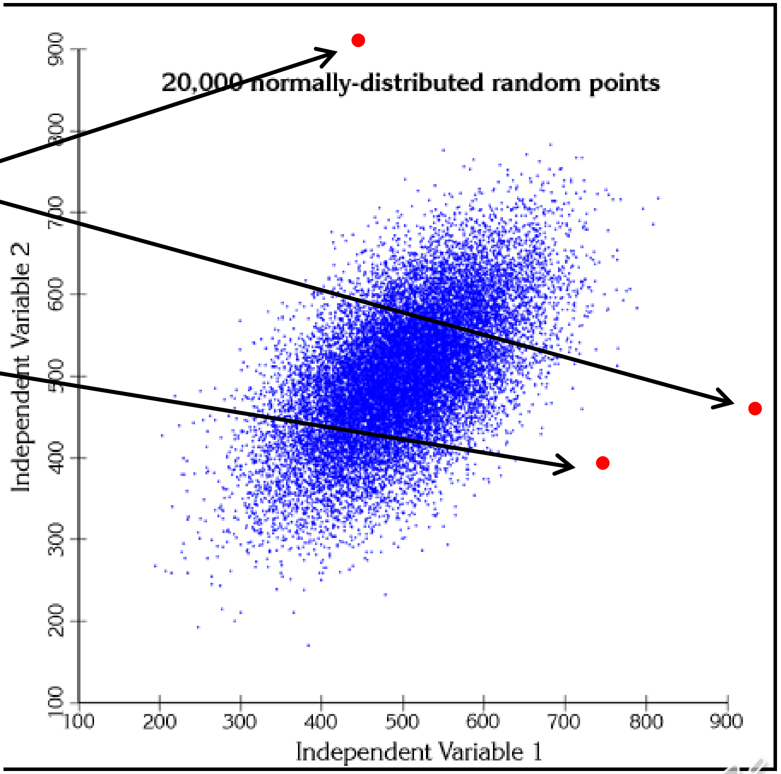
Unsupervised Methods

Graphical approaches: Scatterplot

Graphical approaches:
Given a database D, inspects it (visually) and determines which points are anomalies

What’s an outlier in *2*-dimensions?

- A data with an extreme value in some attribute(s)
- A data with an abnormal combination of common (non-extreme) attribute values



3-dimensions → Cube

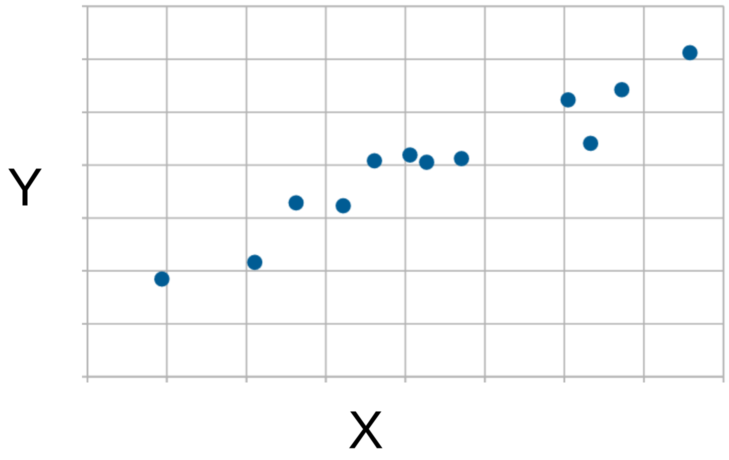
More than 3 dimensions?

Unsupervised Methods

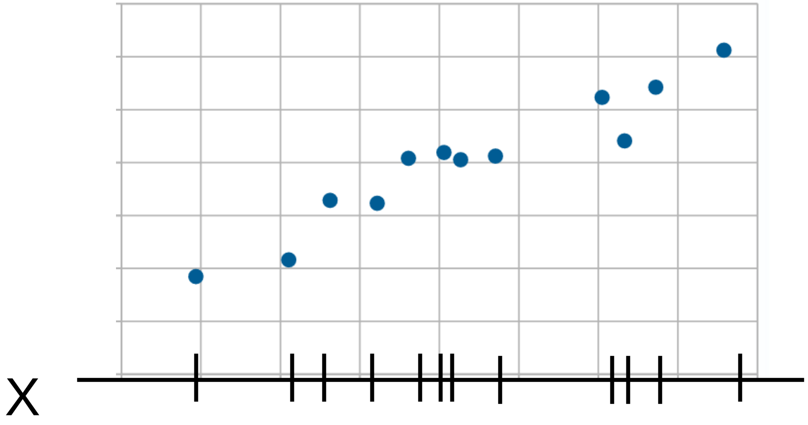
How to resume the information given by several attributes into two or three dimensions, so we can plot them?

Let’s consider two variables and the following scatterplot.

If we want to remove one variable. Which one should be selected?



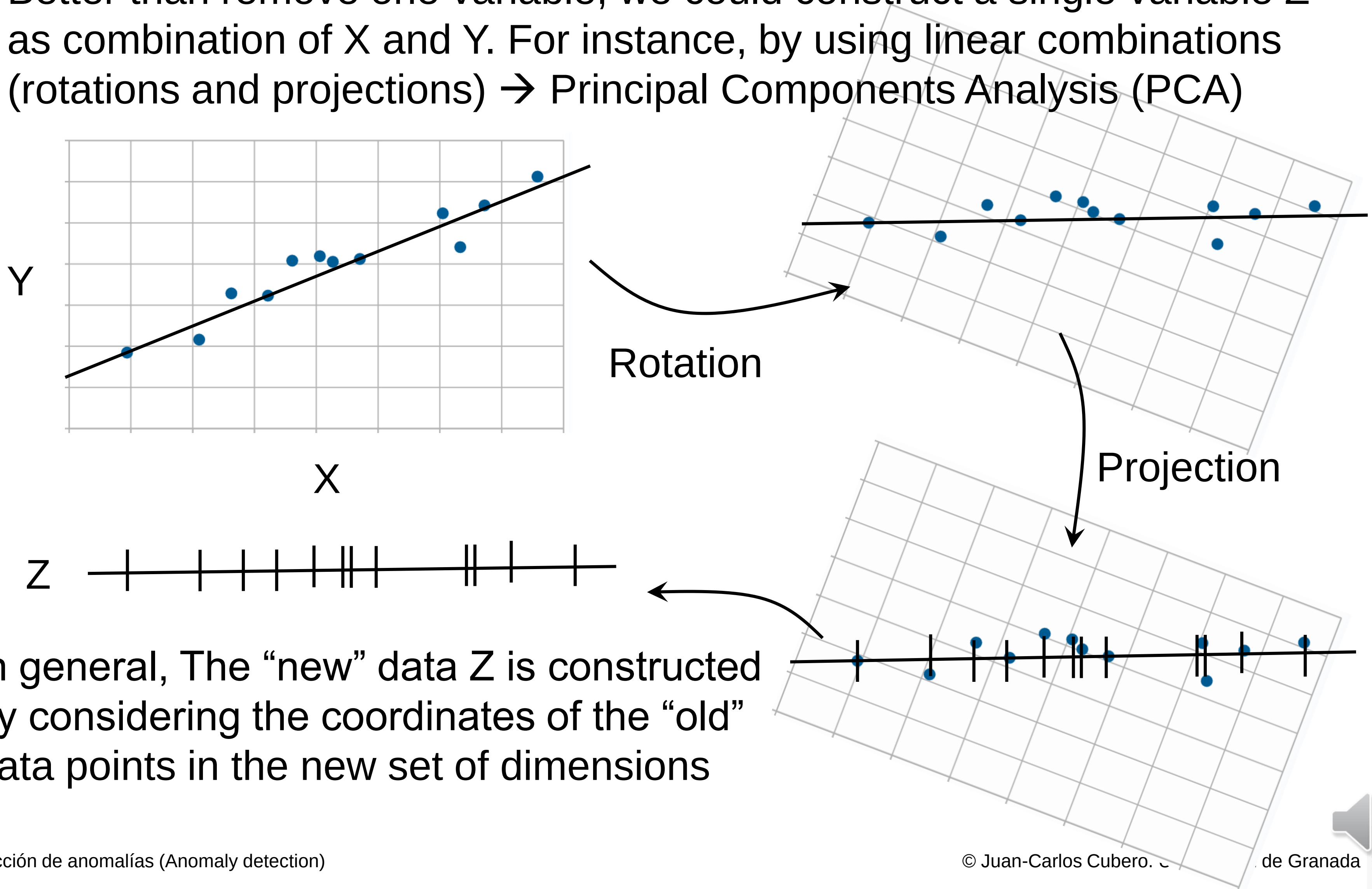
The best solution is to remove the variable with less variance, because it provides less information. We should remove Y and keep X



Unsupervised Methods

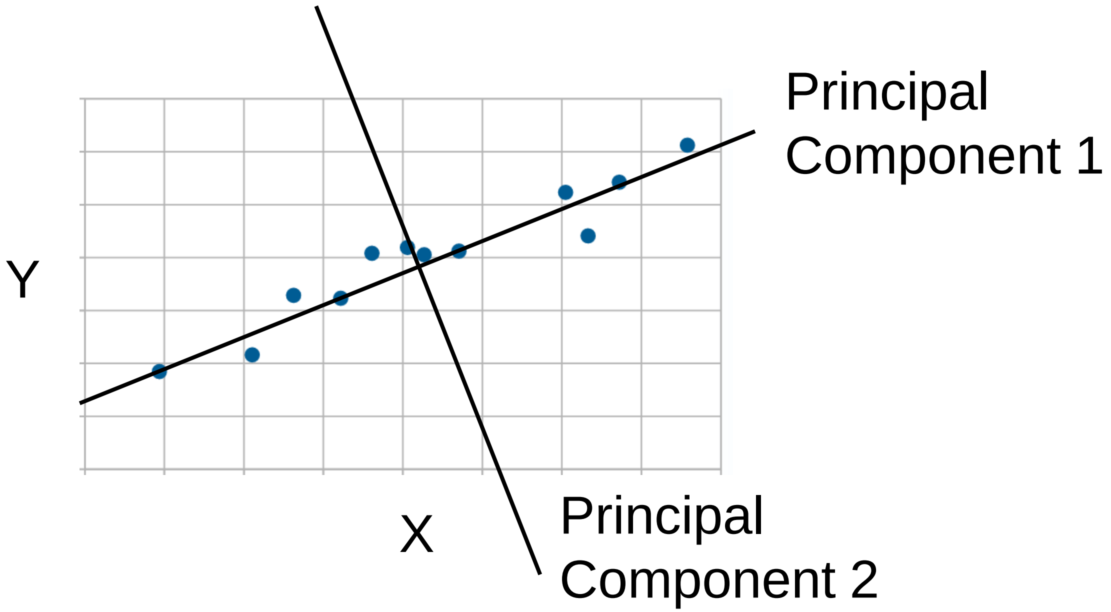
Graphical approaches: Biplot

Better than remove one variable, we could construct a single variable Z as combination of X and Y. For instance, by using linear combinations (rotations and projections) → Principal Components Analysis (PCA)



In general, The “new” data Z is constructed by considering the coordinates of the “old” data points in the new set of dimensions

Unsupervised Methods



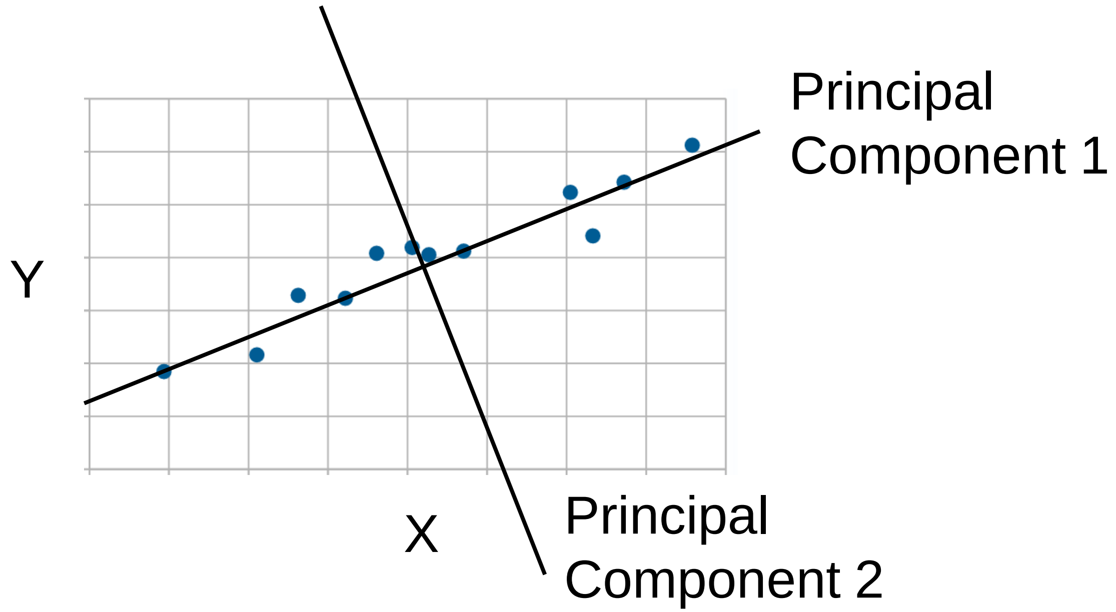
If we use PC1 and PC2 instead of X and Y, there’s no information lost.

If we use PC1 instead of X and Y, there’s some information lost.

If we use X instead of X and Y, there’s more information lost.

If we use Y instead of X and Y, there’s a lot of information lost.

Unsupervised Methods



In general, the PCA method constructs a set of “Principal Components”: as many as the number of variables. Each Principal Component explains an amount of variability and they are ordered, beginning with 1.

In practice, in order to visualize the results, the dataset variables are replaced by two or three principal components. The results can be considered good enough if PC1 + PC2 explains at least 70% of variability

Unsupervised Methods

Graphical approaches: Biplot

In general, a **biplot** performs a PCA projection into 2 dimensions and, in addition, add arrows for each original attribute.

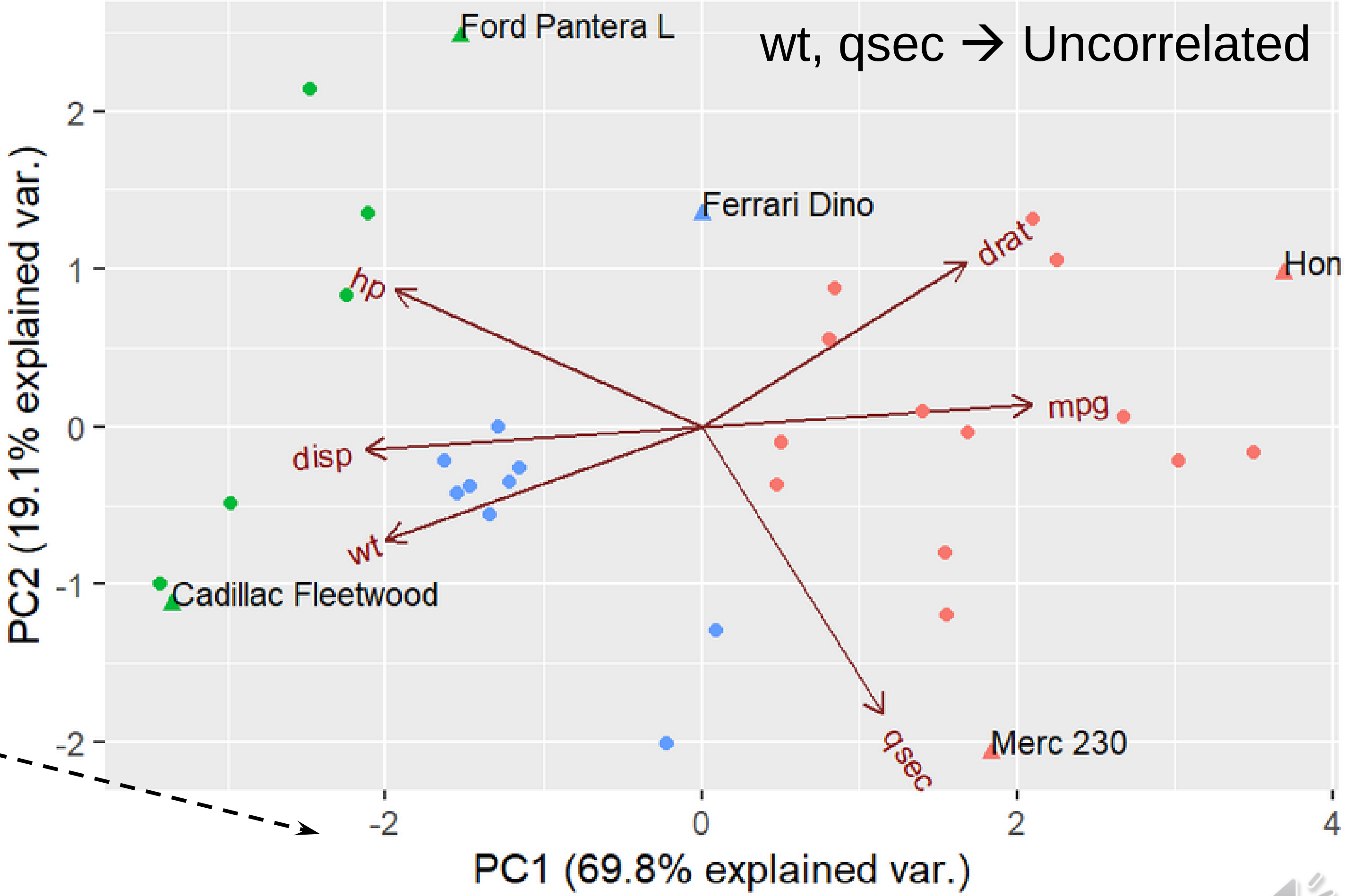
69.8+19.1=

89.2% variance explained

disp, wt → Positively correlated

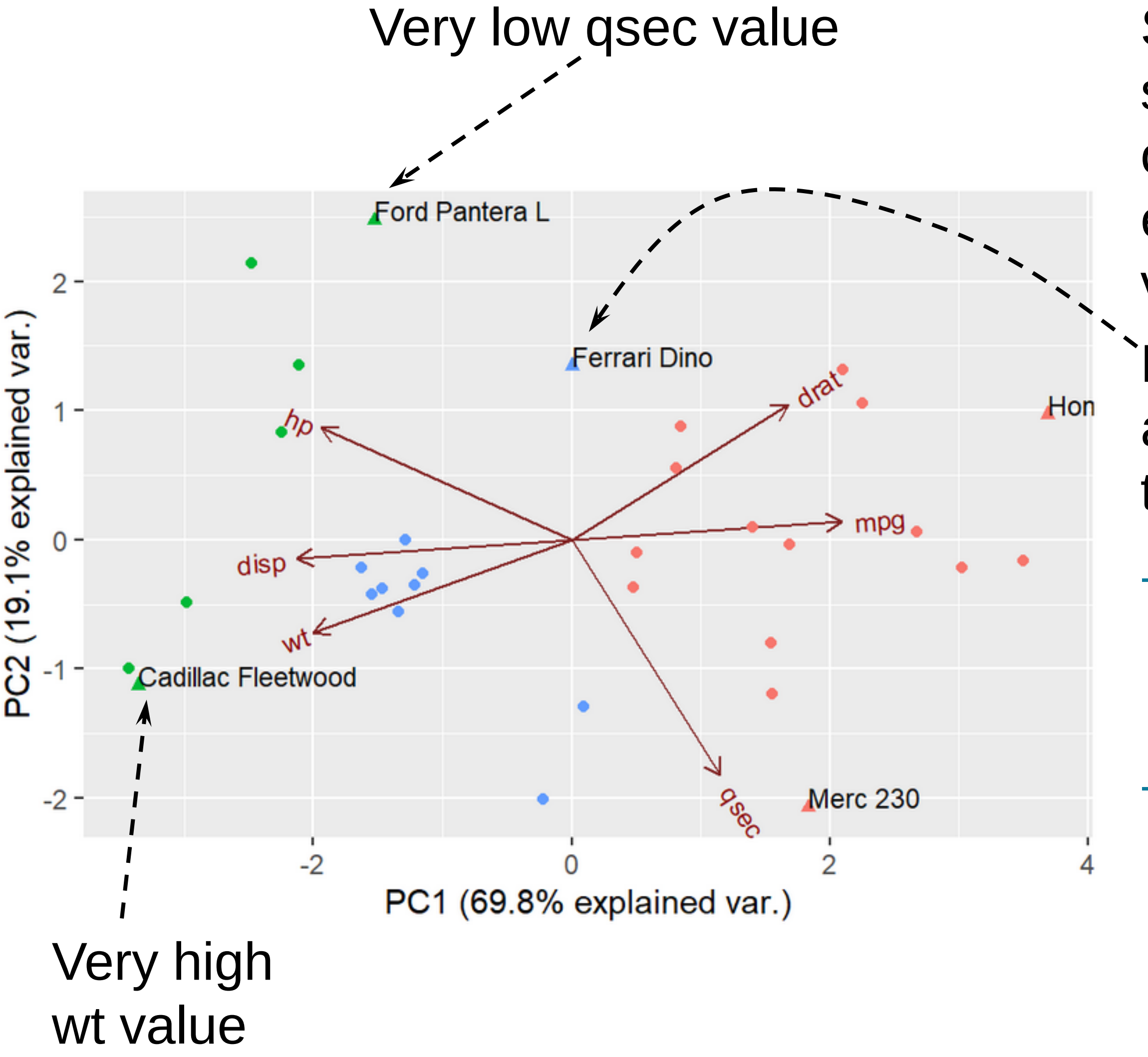
disp, mpg → Negatively correlated

wt, qsec → Uncorrelated



mtcars

Unsupervised Methods



Some outliers are easily seen in the biplot. These outliers usually have some extreme value in one variable.

But, some outliers do not appear in the outer part of the points cloud:

- Because of the information lost by the projection
- Because they may have some abnormal combination of attribute values. They may be interesting and deserve deeper inspection