# Different data types. What to do?

Up to this point, we have been mainly dealing with continuous variables

What do we do when we have more than one data type?

- Continuous data
- Dichotomous data
- Count Data.
- Correlated Data

The standard distance metrics do not to a good job dealing with this.

# Data Transformation

What we need is to transform our data onto another scale so we can use it to cluster:

Ideally the resultant variable will be:

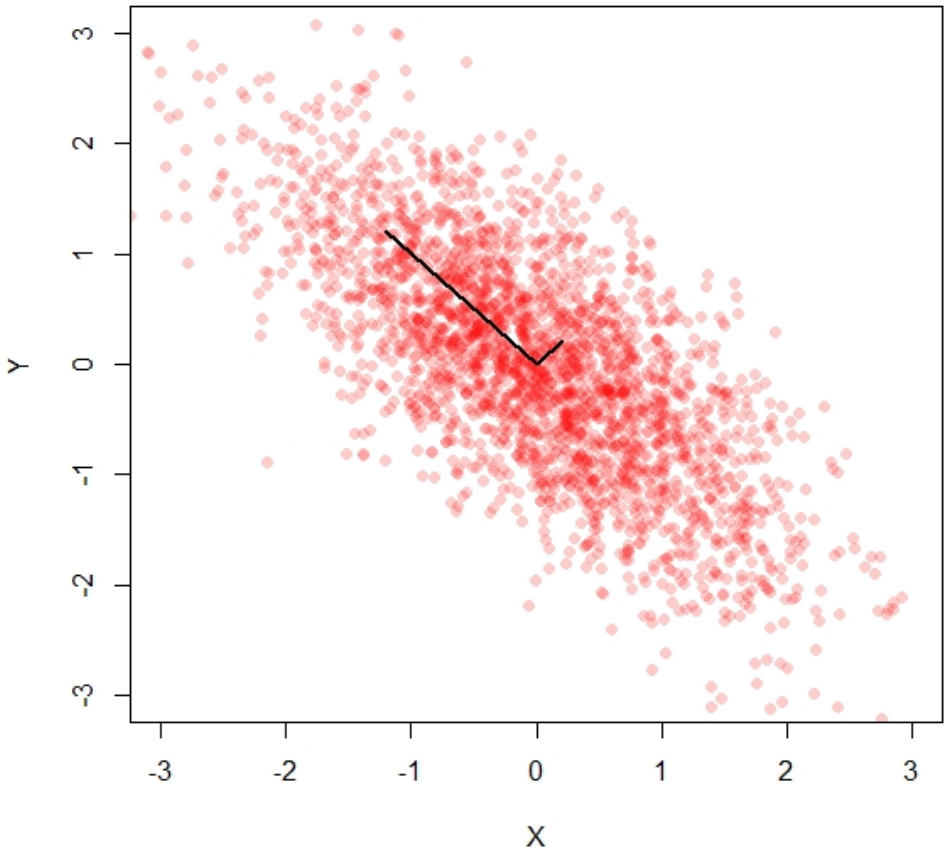
1. Continuous
2. Uncorrelated
3. On the same scale

The method of choice is that of principal component analysis.

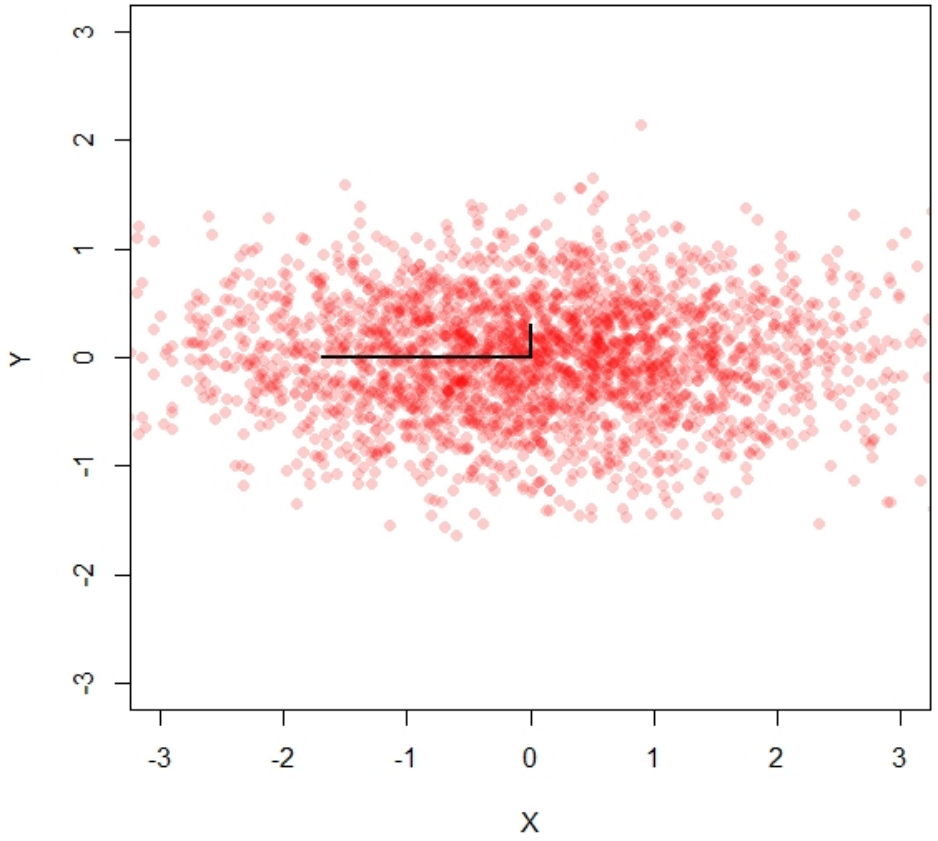
It essentially makes a new set of uncorrelated random variables, by transforming them to a new scale.

Let's review principle component analysis and the Multivariate Normal distribution.

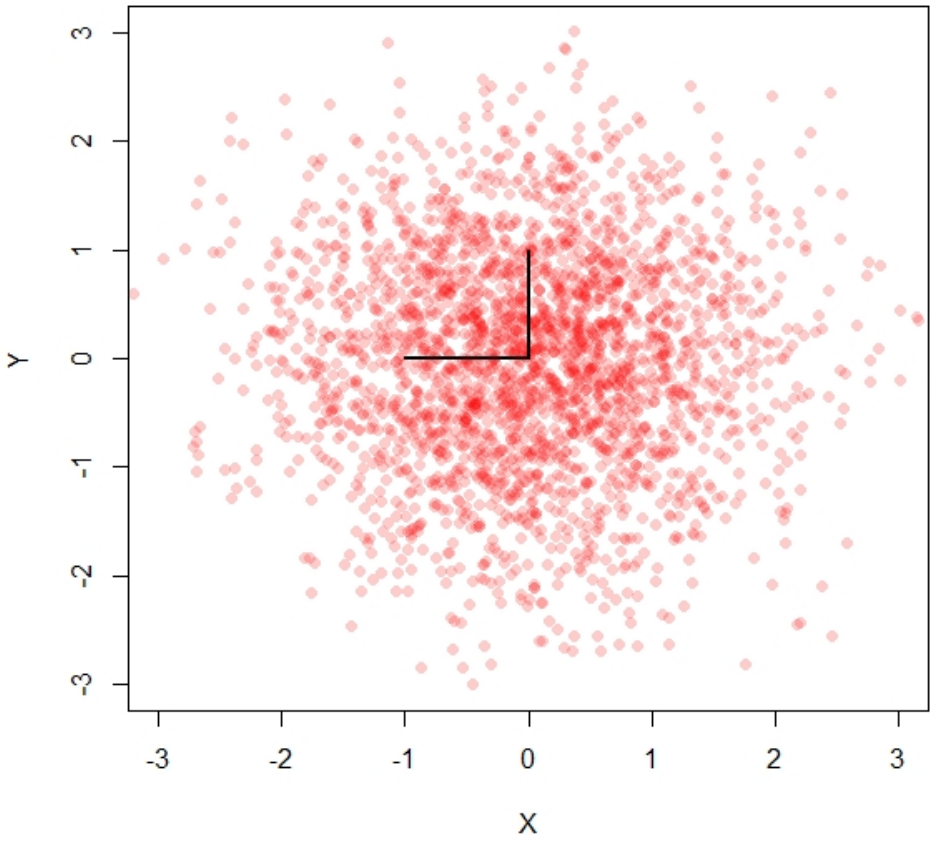
The multivariate normal distribution creates ellipsoid shaped data:



# Principal components firsts tilts the data



# And then standardizes it



# So our plan is to:

1. Take ALL of our data calculate its covariance
2. Compute the Eigen Vectors/Eigen Values of the covariance.
3. Mean center our observations
4. Multiply this centered variable by the Eigen Vector
5. Then standardize it by the Eigen value (or some other method.
6. Cluster on these new variables.

Let’s go to the code

# Group Names

Now that we have our clusters… We need to name them

We also need to name our “Loadings”

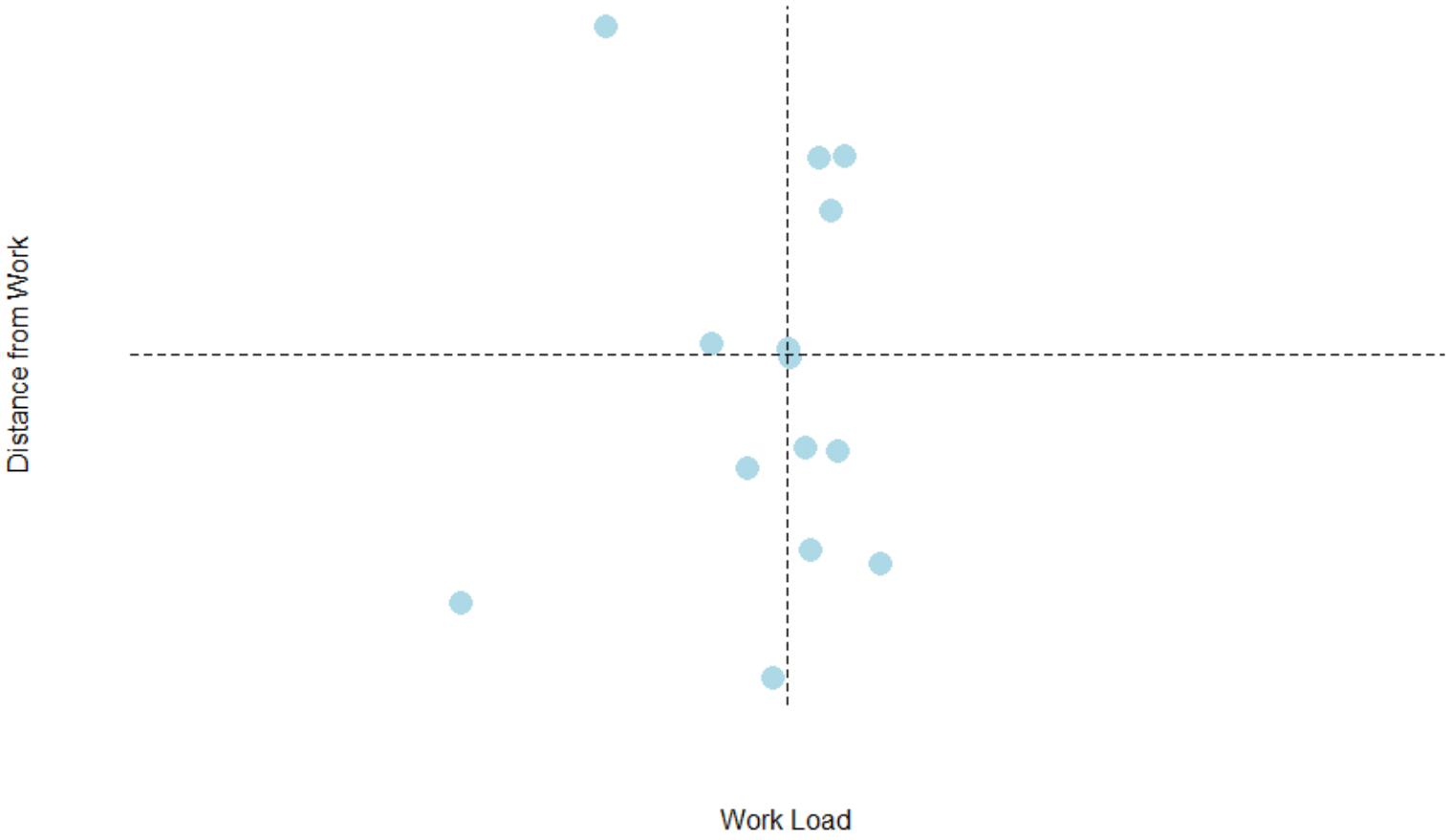
Let’s Go to the spreadsheet.

# New Variables-New Problems:

1. Can't interpret them directly
2. Eigen Vectors load on multiple columns … need to interpret this.
3. Need to work from cluster means for the variables on the original scale to the cluster centroids on this new scale.

Everything we do is based upon this.

# First Two Graphs


Distance from Work
Work Load


Tall,close, Absent over average
Over Weight