318 Term Project Step by Step

1. Perform a PCA

Main ideas of PCA:

* Given a set of features (columns) with rows of data.
* If we plot features against one another, we can see their relationship (if any).
* When we have too many features to visualize effectively using graphs, we use PCA plots.
* PCA plot convers the correlations into a 2D graph.
  + Cells that are highly correlated cluster together
  + **Axes are ranked in order of importance:**
    - **Differences along the first principal component axis (x axis) are more important than differences along the second principal component axis (y axis).**
      * This means that clusters of equidistance on the x axis are more different than clusters of equidistance on the y axis. A picture containing graphical user interface

        Description automatically generated

A principal component is a linear combination of variables

Once you have all of the principal components figured out, you can use the eigenvalues (sum of squared distances) to determine the proportion of variation that each PC accounts for.

A scree plot can show us the degree of variation in the data explained by the principal components

PCA in R video

prcomp() … the goal is to draw a graph that shows how the features are related (or not related) to each other. Features should be columns and samples (or records) should be rows.

Prcomp() returns three things: 1) x 2) sdev 3) rotation

Choosing Features:

PC1 explains .4086 of the variance of the data

* And has features with largest magnitude:
  + Global active power
  + Global intensity
  + Sub\_metering\_3

PC2 explains .1426 of the variance of the data

* And has features with largest magnitude:
  + Global reactive power
  + Sub\_metering\_3

Creating Values for Comparison Methodology: Multiply the loading values by the PCx’s proportion of variance

PC1

Global Active Power coefficient \* PC1 Proportion of Var = -0.4687 \* .4086 = -0.1915

Global intensity coefficient \* PC1 Proportion of Var = -0.5596 \* .4086 = -0.2286

Sub\_metering\_3 coefficient \* PC1 Proportion of Var = -0.3874 \* .4086 = -0.1583

Voltage coefficient \* PC1 Proportion of Var = 0.3305 \* .4086 = 0.1350

PC2

Global Active Power coefficient \* PC2 Proportion of Var = -0.1345 \* .1426 = -0.019

Global intensity coefficient \* PC2 Proportion of Var = 0.0190 \* .1426 = 0.0027

Sub\_metering\_3 coefficient \* PC2 Proportion of Var = 0.4719 \* .1426 = 0.0673

Voltage coefficient \* PC2 Proportion of Var = - 0.1331 \* .1426 = -0.0190

PC3

Global Active Power coefficient \* PC2 Proportion of Var = -0.0874 \* .1343 = -0.0117

Global intensity coefficient \* PC2 Proportion of Var = 0.0012\* .1343 = 0.0027

Sub\_metering\_3 coefficient \* PC2 Proportion of Var = -0.0943 \* .1343 = 0.0673

Voltage coefficient \* PC2 Proportion of Var = -0.0349\* .1343 = -0.0190