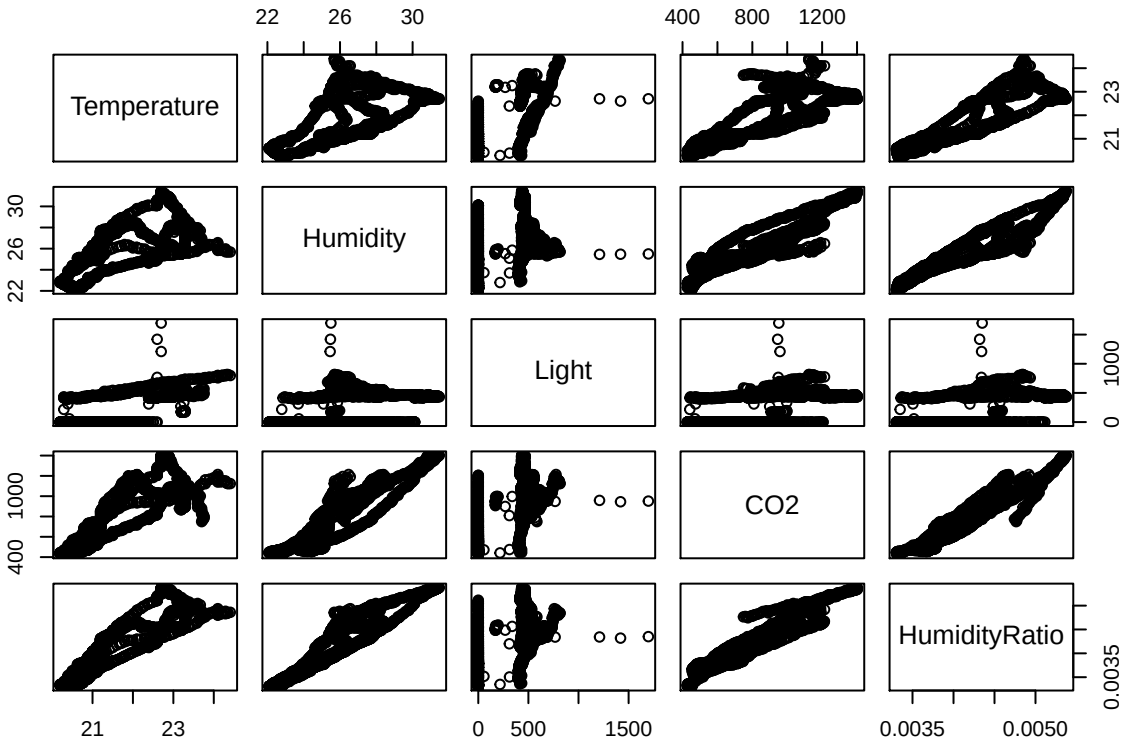
day_2_oct_30_2016_occupancy_pca

I still need to read this for a quick review of PCA: https://tgmstat.wordpress.com/2013/11/21/introduction-to-principal-component-analysis-pca/

Today I will try to interpret the PCA components and see if it yields any insight.

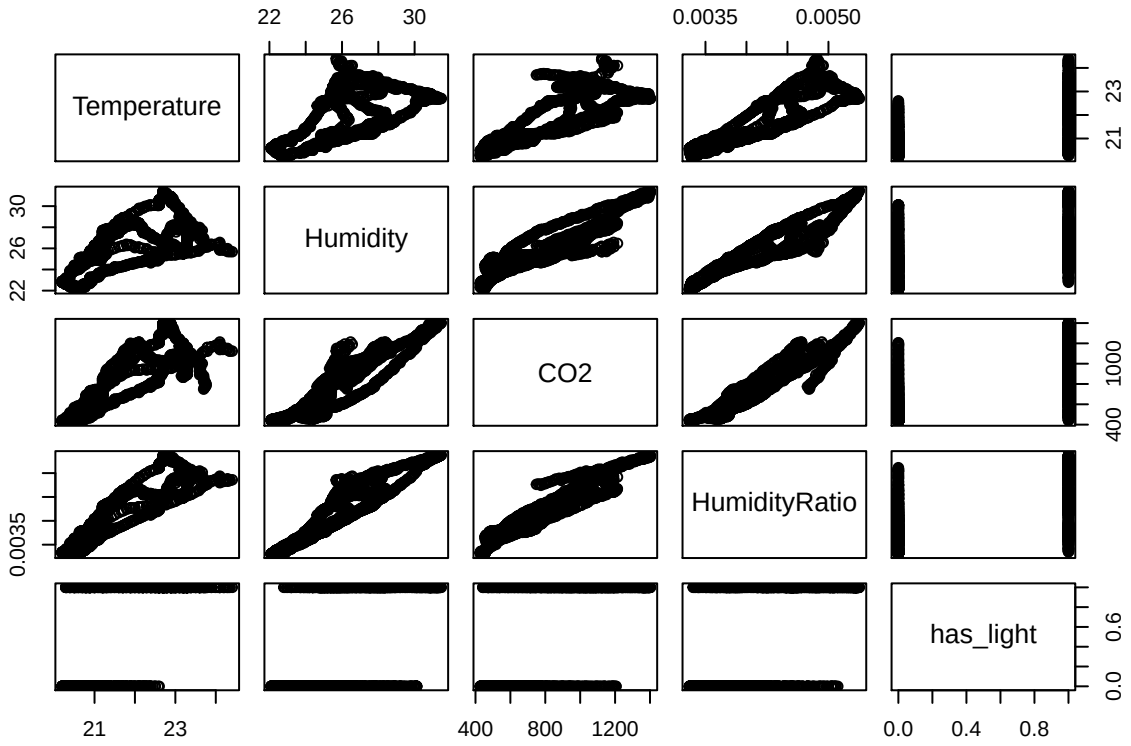
```
X <- df[,c(-1,-7)] # remove the date and occupancy variables
plot(X)
```



After removing all non-numerical predictors, I try to visualize the distributions of the variables.

Plots with light on one of the axis is usually very bimodal. However, this is expected as we saw yesterday that light is almost like a binary variable.

```
X$has_light <- (X$Light > 0) * 1
X <- X[,-3]
plot(X
)
```



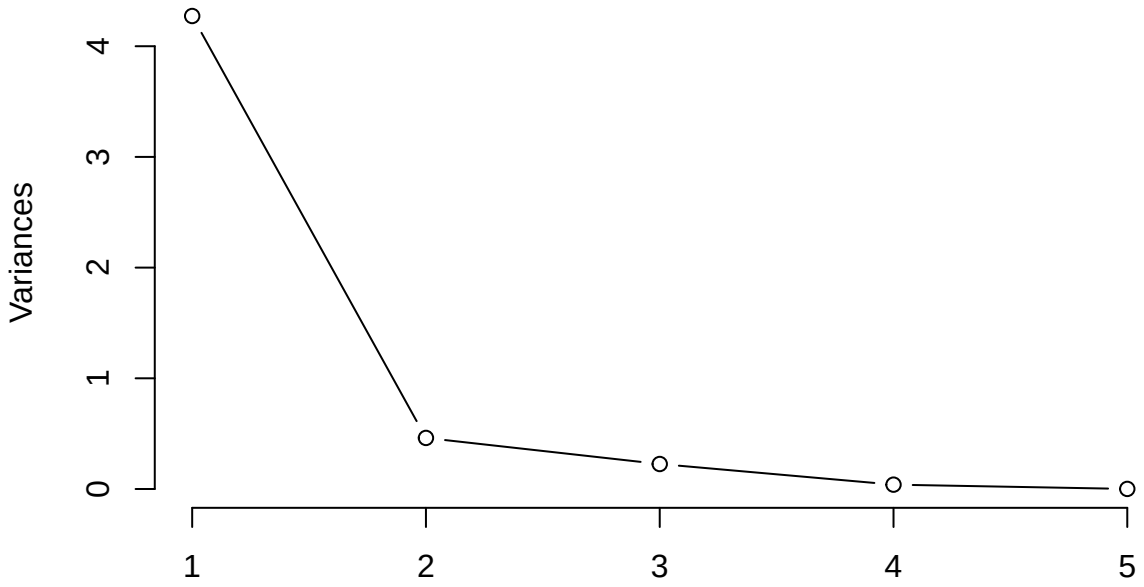
Going to remove the light variable, and change the feature to simply if there was light or not. We will probably lose some information, but at the moment, the information loss should be fairly trivial.

The data seems fairly standardized. So I am going to apply PCA now. I will probably try box and cox to remove any skew later.

```
occ.pca <- prcomp(X, center=TRUE, scale=TRUE)
print(occ.pca)
```

```
## Standard deviations:
## [1] 2.06707259 0.67917859 0.47517060 0.19834528 0.02827395
##
## Rotation:
##                       PC1         PC2         PC3        PC4          PC5
## Temperature    -0.4431792  0.20105018  0.79076961 -0.1482178  0.340420267
## Humidity       -0.4404313 -0.53457479 -0.38990959 -0.3500893  0.495639054
## CO2            -0.4751997 -0.08572054 -0.09297102  0.8703280  0.026885155
## HumidityRatio -0.4752353 -0.25492235  0.09543841 -0.2497235 -0.798558468
## has_light      -0.3974281  0.77554832 -0.45266084 -0.1887759 -0.006126248
```

```
plot(occ.pca, type = "l")
```



From the scree plot, we see that first 2, maybe 3 PCs are super important while 4 to 6 can be ignored.

```
summary(occ.pca)
```

```
## Importance of components:
##                           PC1     PC2     PC3     PC4     PC5
## Standard deviation     2.0671 0.67918 0.47517 0.19835 0.02827
## Proportion of Variance 0.8546 0.09226 0.04516 0.00787 0.00016
## Cumulative Proportion  0.8546 0.94681 0.99197 0.99984 1.00000
```

PC1 and PC2 describe about 95% of the variance.

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 3.2.4
```

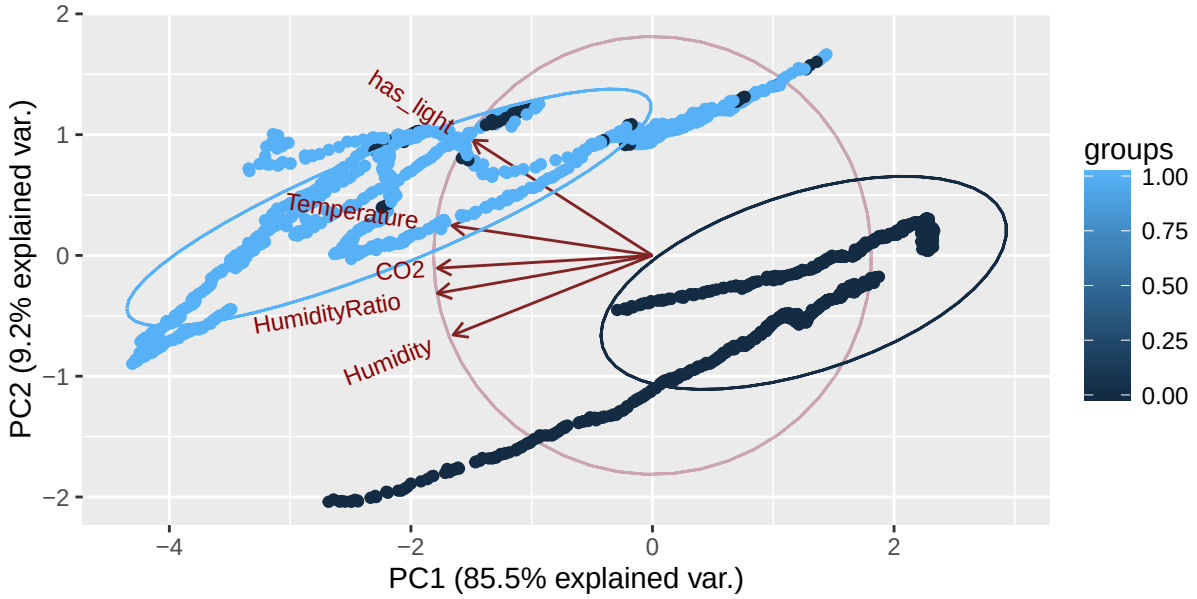
```
## Loading required package: plyr
```

```
## Warning: package 'plyr' was built under R version 3.2.5
```

```
## Loading required package: scales
```

```
## Warning: package 'scales' was built under R version 3.2.3
```

```
## Loading required package: grid
```



The black and blue circles are the 68% (plus minus 1SD) contours for each class.

The projection onto the first 2 PCs make the data almost linearly separable. Just by eye-balling it, I think I can get a 95%+ accuracy just via the first 2 PCs.

Looking at the first 2 principal components:

- PC1 seems like a straight up negative scaling of everything
- PC2's temperature, humidity, CO2 and humidity ratio basis are different