Principal Component Analysis

Finds structure in features and aid in visualization

- Find linear combination of variables to create principal components
- Maintain most variance in the data (first component has largest possible variance)
- Principal components are uncorrelated -> orthogonal
  - Resulting vectors are an uncorrelated orthogonal basis set

Same as eigenvalue decomposition of $X^T X$ (covariance matrix) and singular value decomposition of X. Important to normalize data first.

- First principal component is equal to eigenvector with largest eigenvalue

Iris Example

PCA function

```
pca.iris = prcomp(x=iris[-5], scale=T, center=T)
```

Which predictors contribute to which principal components

```
pca.iris$rotation
```

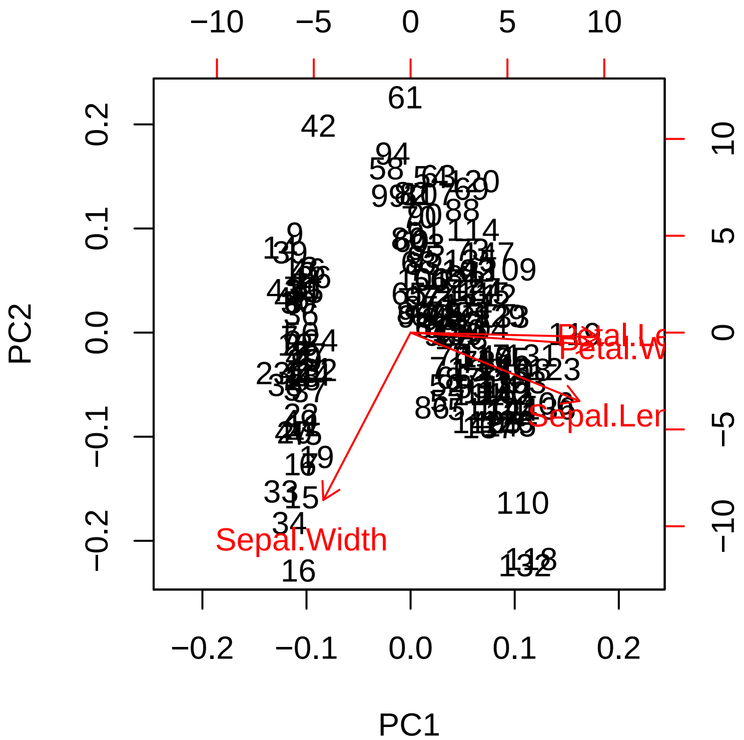
```
##                     PC1         PC2        PC3        PC4
## Sepal.Length  0.5210659 -0.37741762  0.7195664  0.2612863
## Sepal.Width  -0.2693474 -0.92329566 -0.2443818 -0.1235096
## Petal.Length  0.5804131 -0.02449161 -0.1421264 -0.8014492
## Petal.Width   0.5648565 -0.06694199 -0.6342727  0.5235971
```

Sepal.Width has little contribution to PC1 but almost all of PC2

Visualization

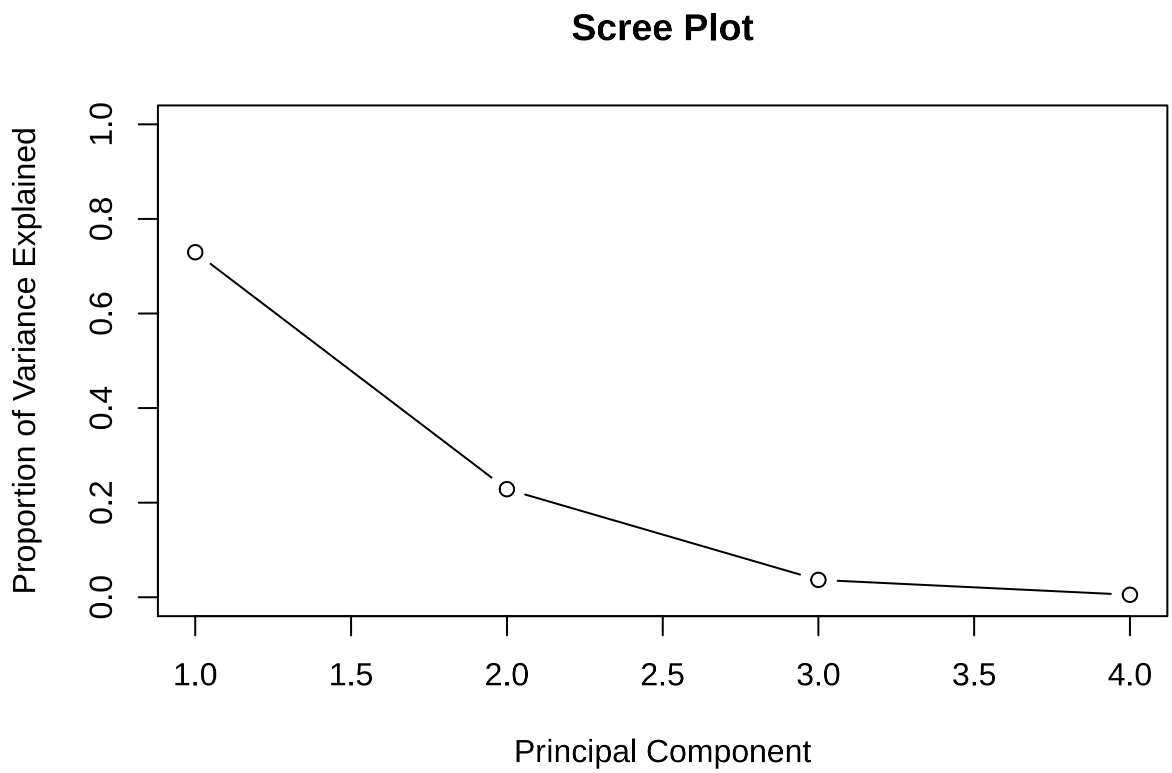
Visualizing first two components

```
biplot(pca.iris)
```



Petal.Width and Petal.Length are in the same direction indicated they are correlated in the original data

```
pca.iris.var = pca.iris$sdev^2
pve = pca.iris.var/sum(pca.iris.var) #proportion of variance explained
plot(pve, xlab = 'Principal Component', ylab = 'Proportion of Variance Explained', main='Scree Plot', y
```



First principal component explains the most variance, explained variance decreases for later principal components