(a) To make sure that you have done this correctly, **report the dimensions of the new data frame.**

```
library(caret)
ins = read.csv("Insurance.csv", header = TRUE)
ins$zone = as.factor(ins$zone)
ins$make = as.factor(ins$make)
ins.dv = data.frame(predict(dummyVars("~.", data=ins), newdata = ins))
head(ins.dv)

dim(ins.dv)
# > dim(ins.dv)
# [1] 2182   21
```

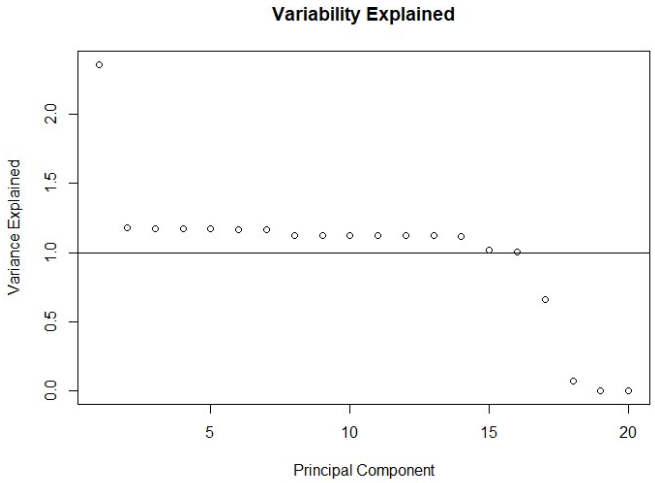
2. Run a principal components analysis on these data, excluding the response variable per, using scale.=TRUE to standardize the data.

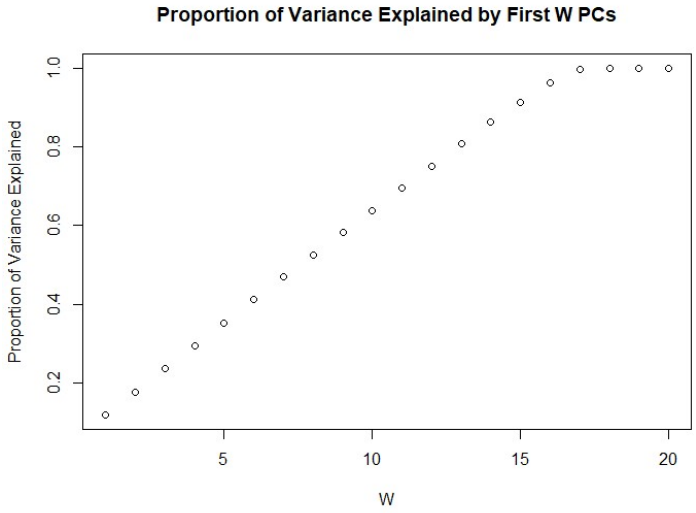
```
data.matrix.raw = model.matrix(per ~ ., data = ins.dv)
data.matrix = data.matrix.raw[,-1]

fit.PCA = prcomp(data.matrix, scale. = T)
print(fit.PCA)


#Scree plot
vars = fit.PCA$sdev^2
plot(1:length(vars), vars, main = "Variability Explained",
     xlab = "Principal Component", ylab = "Variance Explained")
abline(h = 1)

# Cumulative variance plot
c.vars = cumsum(vars)   ### Cumulative variance explained
rel.c.vars = c.vars / max(c.vars)   ### Cumulative proportion of
### variance explained
plot(1:length(rel.c.vars), rel.c.vars,
     main = "Proportion of Variance Explained by First W PCs",
     xlab = "W", ylab = "Proportion of Variance Explained")
```

(a) **How many PC's get chosen** if you use the guideline of choosing all PC's with higher than average contributions to the explained variance?

➔ I might choose 15 variables

(b) Make a scree plot and a cumulative variance plot. **Show these plots.**





(c) Based on these plots, **indicate how many PC's** you think explain "enough" variability while achieving dimension reduction? **Explain very briefly why you chose that number.**

➔ 16 PC's might be enough because until the 17th PC, the proportion of variance explained get bigger linearly.