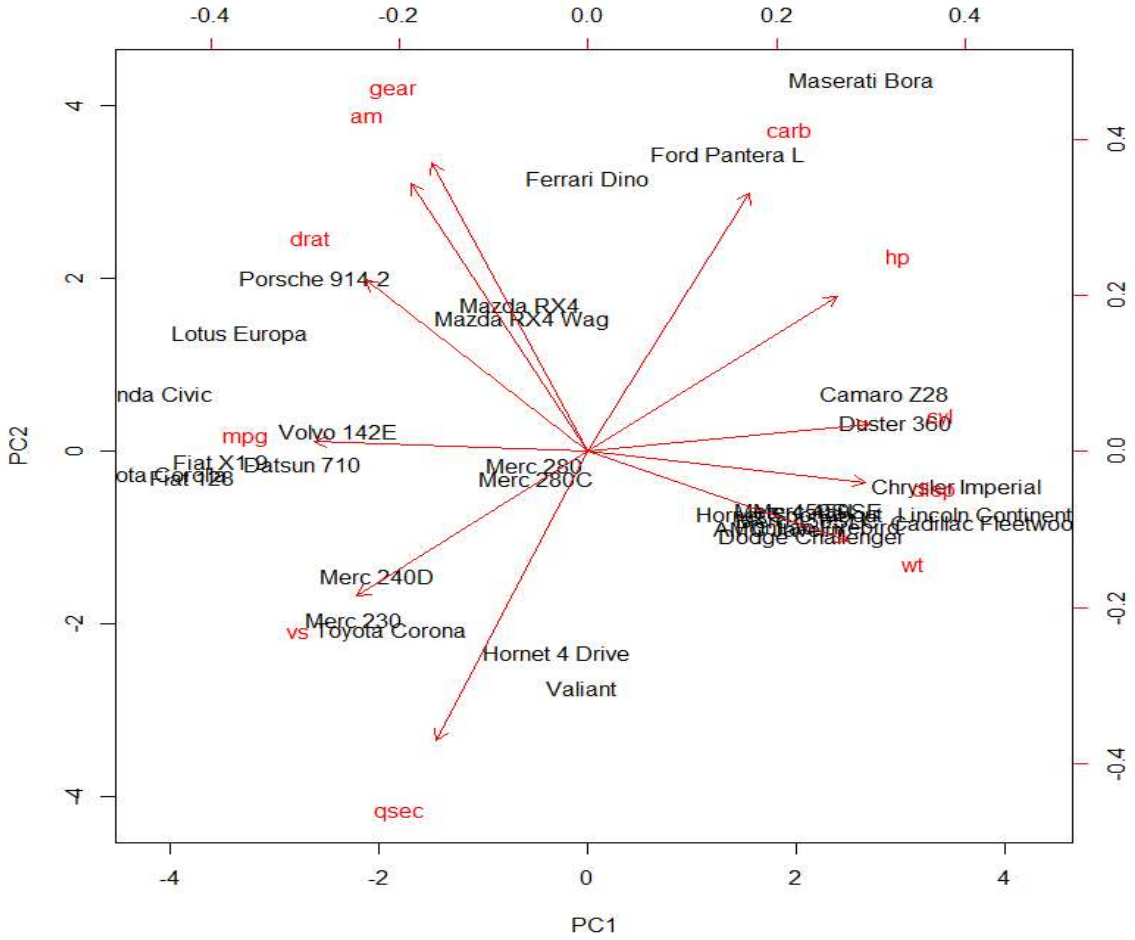
Course: SP18 Data Mining
Assignment: 9
Student: Lina Mi

1. Should a principal components analysis of this data be based on the covariance or the correlation matrix? Explain.
   Ans: the principal components analysis of this data should be based on correlation matrix of all variables except model. Since principal analysis aims to analyze the correlations among variables of dataset, and then extract most important one variable(feature) or more variables(features), since each element in correlation matrix is covariance of each pair of variables, therefore correlation matrix reflect the relationships among all variables. Covariance, however, is a measure between two dimensions(variables), it can't contain all relationships between all different variables. Hence, the principal components analysis of this data should be based on correlation matrix of all variables instead of covariance.

2. Which variables seem to have the strongest relation to the mileage?
   Ans: the obtained plot from PCA analysis is shown as following:



It can be seen from the plot above that variable 'cyl' has strongest relation with 'mpg'(inverse proportional) followed by variable 'disp'.

3. Are Mercedes different from other cars? If so, what characteristic would you say they share?

   Ans: according to the plot of PCA analysis, it can be seen that unlike most of cars which locate in the different quadrant of the plot(some of the characters of these cars are good,high values, some other characters, whereas, are 'bad', low values), Merc 280 and Merc 280c locate at the origin of the plot, it means these two models have best combination of various characters instead of getting some good characters at cost of sacrificing other characters. For example, 'mpg' is inversely proportional to character 'cyl' for most of cars except Mercedes(280 and 280c), it means that for most cars, either have high 'cyl'(more cylinders) and low 'mpg'(low miles per gallon), such as Lincoln continent, Cadillac Fleetwood and Chrysler Imperial, Duster 360, Camaro Z28, or have low 'cyl' and high 'mpg', such as Honda civic, Toyota corala; Merc 280 and Merc 280C, however, have good combination of characters 'cyl' and 'mpg'(moderate value of both 'cyl' and 'mpg')

4. What characteristics separate sports cars from the others?

   Ans: sports cars have distinct characteristics of high 'cyl' and low 'mpg' compared with other cars, namely sports cars have more cylinders and low miles per gallon.

5. Suppose your car gets good mileage. What else is likely to be true about it?

   Ans: according the PCA analysis plot, if the car gets good mileage (high 'mpg'), it would be very likely to have low values of characters 'cyl', 'disp', 'wt','hp' and 'carb', that means the car is likely to have small number of cylinders, low displacement, low weight, low Gross horsepower, and low number of carburetors; it is also very likely to have high values of characters 'drat', 'vs', 'quec', 'am' and 'gear', namely the car with good mileage will have high rear axle ratio, have 1 for v/s character, higher value of 'quec' (1/4 mile time), manual transmission, and higher number of forward gears.

6. Suppose my car gets 20 mpg, has 6 cylinders, a displacement of 425, 200 horsepower, a rear axle ratio of 3.75, weighs 2000 pounds, can go a quarter mile in 16.5 seconds, has v/s (vertical steering?), automatic transmission, 4 gears and 1 carburetor. What are its scores on the first and second principal components? What sort of car, if any, is it most similar to?

   Ans: after the PCA analysis using prcomp() function, I got the loading vectors for PC1 and PC2 as following:

```
            PC1          PC2

mpg  -0.3625305  0.01612440
cyl   0.3739160  0.04374371
disp  0.3681852 -0.04932413
hp    0.3300569  0.24878402
```

```
drat -0.2941514  0.27469408
wt    0.3461033 -0.14303825
qsec -0.2004563 -0.46337482
vs   -0.3065113 -0.23164699
am   -0.2349429  0.42941765
gear -0.2069162  0.46234863
carb  0.2140177  0.41357106
```

since the PCA is performed on the normalized variables, the PC1 and PC2 calculation should also be based on the normalized variables using the loading vectors. The normalization of one observation data is implemented by subtracting mean and then dividing the subtraction result by standard deviation of corresponding variables:

the mean and standard deviation of each variable are provide by prcomp() function(center, scale)

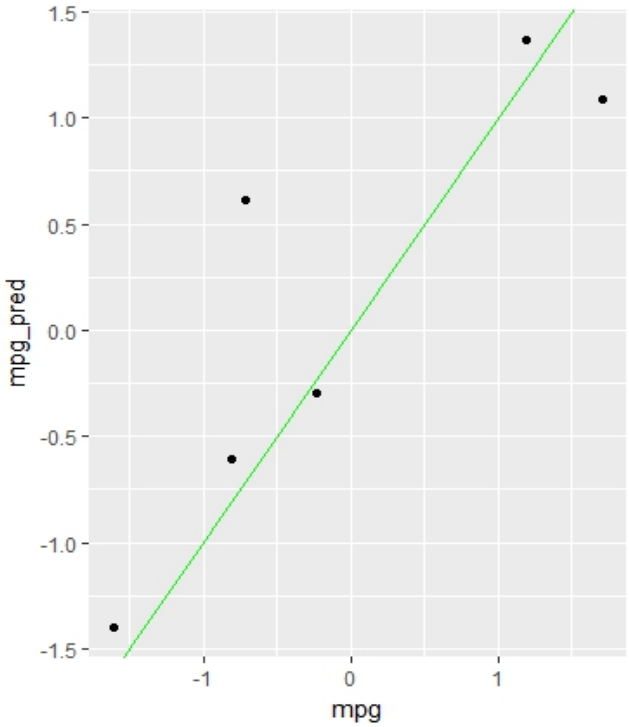
so for the car with described characters(namely mpg=20, cyl=6, disp=425,hp=200,drat=3.75,wt=2.0,qsec=16.5,vs=1,am=0, gear=4, carb=1), the first and second principle components are calculated as following:

```
PC1= -0.04230321, PC2= -0.1585763
```

According to the result of PCA(x), which includes the principle component values of all observations, and the first two principle values of previous car with described characters, this car is most similar to the **Porsche 914-2.**

7. Fit a regression to predict mpg. Evaluate the fit. What can be done to improve it?

Ans: the following plot is a prediction result by regression model using the characters (predictor variables) of dataset.



The obtained RMS is as following:

```
>rms
>0.6145557
```

It can be seen that the prediction made by regression model based on predictor variables of dataset is not very well.

Since The principle components from PCA capture much information with fewer features and be able to explain the data variance much better than predictor variables of dataset, we can use the principle components to replace original variables of dataset as predictor variables in the regression model expecting to get better prediction result.

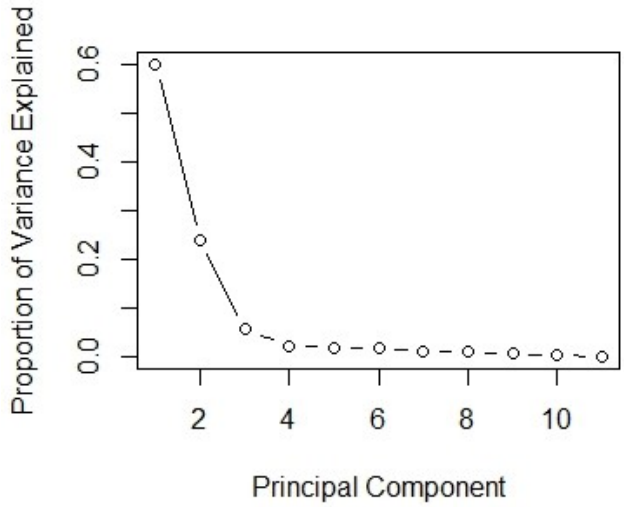
8. Try your suggestion. Did it help? What's the best predictor of mileage?

Ans: 11 principle components are obtained from the PCA, the standard deviation of principle components are also provided by PCA, in order to retain as much information as possible using these components, we need find the components which explain the maximum variance, since higher is explained variance, higher will be the information contained in the components.

```
>pca_stdev<-car_pca$sdev
>prcom_var<-pca_stdev^2
>prcom_var
[1] 6.60840025 2.65046789 0.62719727 0.26959744 0.22345110 0.21159612
[7] 0.13526199 0.12290143 0.07704665 0.05203544 0.02204441
```

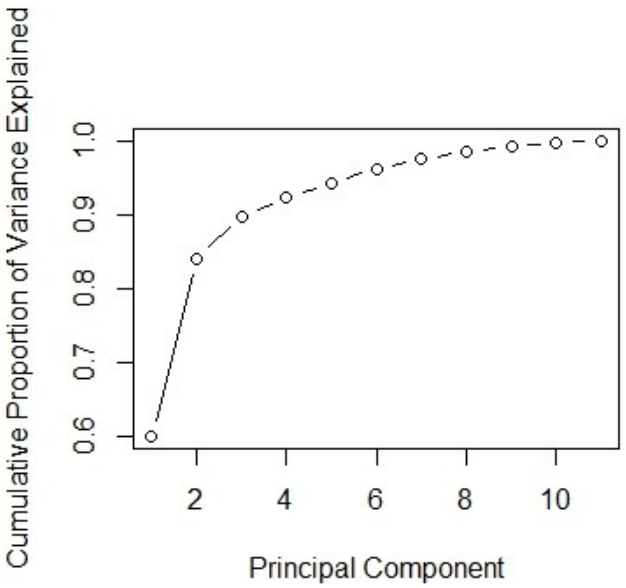
It can be seen from the result above that the first and second principals explained most the variance of the dataset. To demonstrate this, the proportion of variance explained by each principal component is calculated and plot as following:

```
>prop_varex<-prcom_var/sum(prcom_var)
```



The above plot shows that the first 8 principal components explained more than 98% of variance in the data set. This is further explained by the following cumulative variance plot:

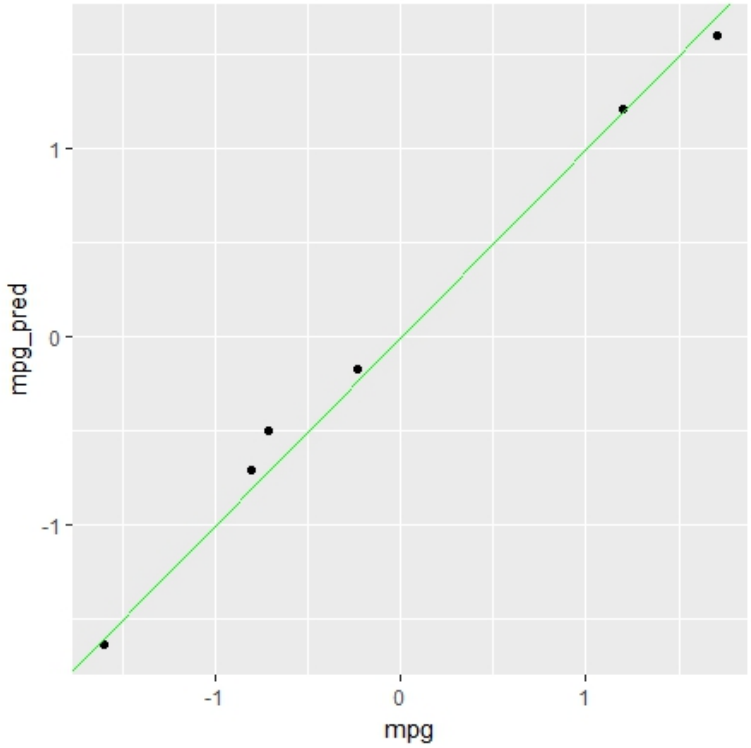
```
>plot(cumsum(prop_varex), xlab="Principal Component", ylab = "Cumulative
Proportion of Variance Explained", type = "b")
```



The above plot shows that first 8 principal components result in variance about 98%, therefore, I select first eight principal components (PC1-PC8)as predictor variables in linear regression model.

```
>prin_car_data<-data.frame(norm_car_data$mpg, car_pca$x[,1:8])
```

the following plot is a prediction result by regression model using the first eight principal components as predictor variables.



the result RMS is

```
>rms_prin
 [1] 0.1075018
```

Compared with the results from regression model using original variables of dataset(which RMS is 0.6145), it can easily be seen that the prediction result using principal components is much better than before. So it can be concluded that PCA is successful, and the best predictor of mileage is first principal component