Mini Project 5: Report

Raheel Ahmed

rsa170130

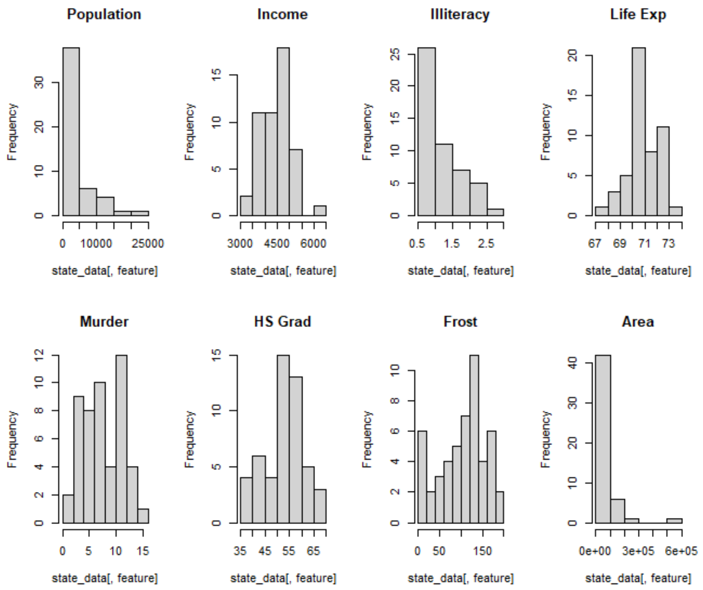
Section 1

Question 1

a. **Fit a PCR model with M chosen optimally via LOOCV. Compute the test MSE of the model.**
   The M chosen via LOOCV was 1 component. The test MSE of the model was 0.7606115.
b. **Fit a PLS model with M chosen optimally via LOOCV. Compute the test MSE of the model.**
   The M chosen via LOOCV was 1 component. The test MSE of the model was 0.7800945.
c. **Compare the models in (a) and (b) and also with the model you recommended in Mini Project 4. Which model(s) would you recommend now?**
   The models in a and b were similar in that the number of components they chose were 1 and that their test MSEs were also very close (approximately 0.76 vs. 0.78). In Mini Project 4, however, a lasso model with an estimated MSE of 0.5522568 was found, which outperformed both a and b, provided an optimal lambda value. I would recommend this lasso model, or even the ridge model from Mini Project 4 as a solid alternative.

Question 2

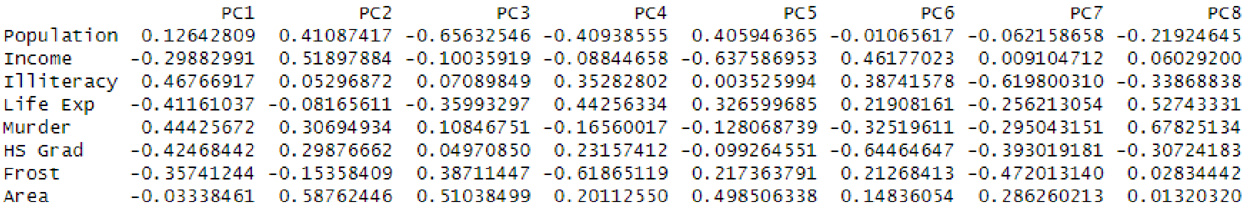
a. **Perform an exploratory analysis of the data.**
   The state data is a dataset with 50 observations and 8 features. Some of the features have fairly strong correlations, like Illiteracy and Murder, Life expectancy and Murder, and Illiteracy and HS Graduation. The standard deviations of some features are quite large, due to the sheer range of these features (e.g. Income, Population, and Area). Other features had different range values and smaller standard deviations.
b. **Do you think standardizing the variables before performing the analysis would be a good idea?**
   I think standardizing the variables would be a very good idea due to, as mentioned before, the difference in ranges of the features and the fact that some features have very high standard deviations. Also, the distributions of the features differ quite a bit.



c. **Regardless of your answer in (b), standardize the variables, and perform a principal components analysis (PCA) of the data.**

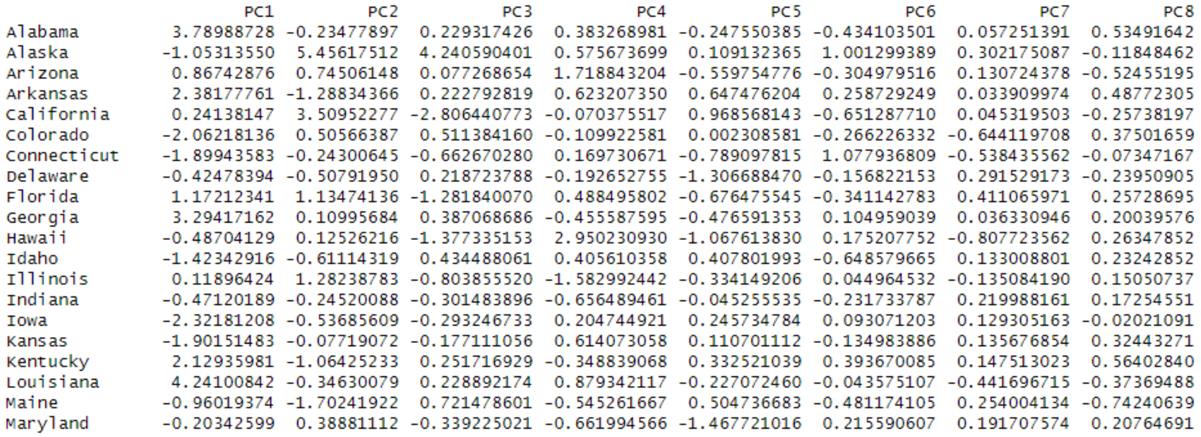
Summarize the results using appropriate tables and graphs. How many PCs would you recommend?

Eigenvector values:



```
                 PC1         PC2         PC3         PC4          PC5         PC6          PC7         PC8
Population  0.12642809  0.41087417 -0.65632546 -0.40938555  0.405946365 -0.01065617 -0.062158658 -0.21924645
Income     -0.29882991  0.51897884 -0.10035919 -0.08844658 -0.637586953  0.46177023  0.009104712  0.06029200
Illiteracy  0.46766917  0.05296872  0.07089849  0.35282802  0.003525994  0.38741578 -0.619800310 -0.33868838
Life Exp   -0.41161037 -0.08165611 -0.35993297  0.44256334  0.326599685  0.21908161 -0.256213054  0.52743331
Murder      0.44425672  0.30694934  0.10846751 -0.16560017 -0.128068739 -0.32519611 -0.295043151  0.67825134
HS Grad    -0.42468442  0.29876662  0.04970850  0.23157412 -0.099264551 -0.64464647 -0.393019181 -0.30724183
Frost      -0.35741244 -0.15358409  0.38711447 -0.61865119  0.217363791  0.21268413 -0.472013140  0.02834442
Area       -0.03338461  0.58762446  0.51038499  0.20112550  0.498506338  0.14836054  0.286260213  0.01320320
```

Some scores:

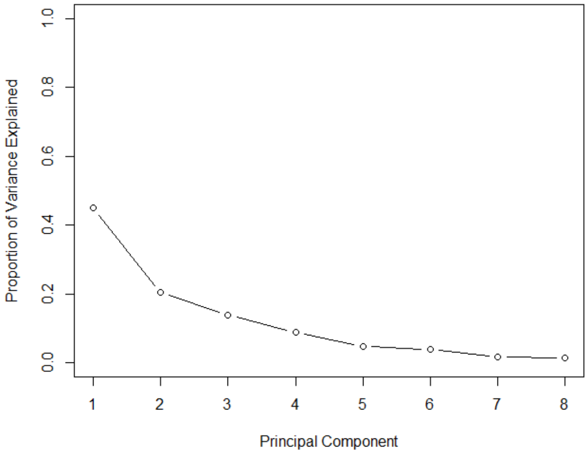


```
                    PC1         PC2          PC3          PC4          PC5          PC6          PC7         PC8
Alabama      3.78988728 -0.23477897  0.229317426  0.383268981 -0.247550385 -0.434103501  0.057251391  0.53491642
Alaska      -1.05313550  5.45617512  4.240590401  0.575673699  0.109132365  1.001299389  0.302175087 -0.11848462
Arizona      0.86742876  0.74506148  0.077268654  1.718843204 -0.559754776 -0.304979516  0.130724378 -0.52455195
Arkansas     2.38177761 -1.28834366  0.222792819  0.623207350  0.647476204  0.258729249  0.033909974  0.48772305
California   0.24138147  3.50952277 -2.806440773 -0.070375517  0.968568143 -0.651287710  0.045319503 -0.25738197
Colorado    -2.06218136  0.50566387  0.511384160 -0.109922581  0.002308581 -0.266226332 -0.644119708  0.37501659
Connecticut -1.89943583 -0.24300645 -0.662670280  0.169730671 -0.789097815  1.077936809 -0.538435562 -0.07347167
Delaware    -0.42478394 -0.50791950  0.218723788 -0.192652755 -1.306688470 -0.156822153  0.291529173 -0.23950905
Florida      1.17212341  1.13474136 -1.281840070  0.488495802 -0.676475545 -0.341142783  0.411065971  0.25728695
Georgia      3.29417162  0.10995684  0.387068686 -0.455587595 -0.476591353  0.104959039  0.036330946  0.20039576
Hawaii      -0.48704129  0.12526216 -1.377335153  2.950230930 -1.067613830  0.175207752 -0.807723562  0.26347852
Idaho       -1.42342916 -0.61114319  0.434488061  0.405610358  0.407801993 -0.648579665  0.133008801  0.23242852
Illinois     0.11896424  1.28238783 -0.803855520 -1.582992442 -0.334149206  0.044964532 -0.135084190  0.15050737
Indiana     -0.47120189 -0.24520088 -0.301483896 -0.656489461 -0.045255535 -0.231733787  0.219988161  0.17254551
Iowa        -2.32181208 -0.53685609 -0.293246733  0.204744921  0.245734784  0.093071203  0.129305163 -0.02021091
Kansas      -1.90151483 -0.07719072 -0.177111056  0.614073058  0.110701112 -0.134983886  0.135676854  0.32443271
Kentucky     2.12935981 -1.06425233  0.251716929 -0.348839068  0.332521039  0.393670085  0.147513023  0.56402840
Louisiana    4.24100842 -0.34630079  0.228892174  0.879342117 -0.227072460 -0.043575107 -0.441696715 -0.37369488
Maine       -0.96019374 -1.70241922  0.721478601 -0.545261667  0.504736683 -0.481174105  0.254004134 -0.74240639
Maryland    -0.20342599  0.38881112 -0.339225021 -0.661994566 -1.467721016  0.215590607  0.191707574  0.20764691
```

Cumulative sum of the PVE:

0.4498619 0.6538519 0.7928445 0.8812825 0.9293627 0.9677954 0.9858515 1.0000000

Scree Plot of the PVE values:

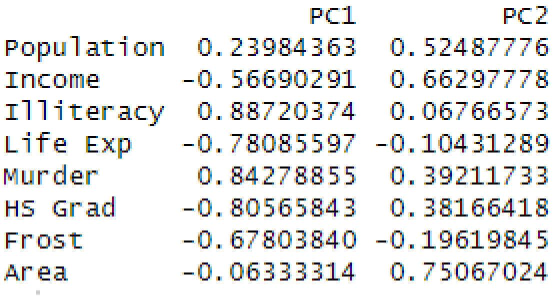


Since the scree plot levels off around 5, we can state that 5 PCs is an appropriate amount of PCs to select. However, numerically, there is an argument to be made for 4 PCs since the increase in value from 4 PCs to 5 PCs is essentially as minimal as from 5 PCs to 6 PCs.

d. **Focus on the first two PCs obtained in (c). Prepare a table showing the correlations of the standardized variables with the components and the cumulative percentage of the total variability explained by the two components. Also, display the scores on the two components**

and the loadings on them using a biplot. Interpret the results. Can you identify, for example, a “southern” component?

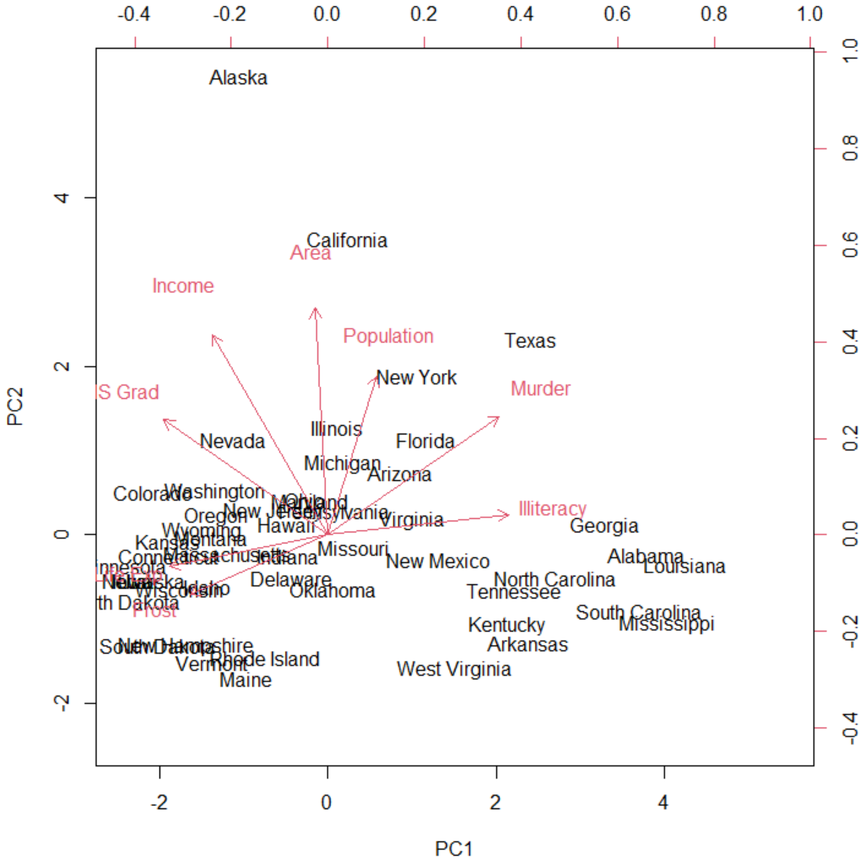
Correlations between features and PCs:



```
                  PC1         PC2
Population  0.23984363  0.52487776
Income     -0.56690291  0.66297778
Illiteracy  0.88720374  0.06766573
Life Exp   -0.78085597 -0.10431289
Murder      0.84278855  0.39211733
HS Grad    -0.80565843  0.38166418
Frost      -0.67803840 -0.19619845
Area       -0.06333314  0.75067024
```

Cumulative PVE explained by the two components: 0.6538519

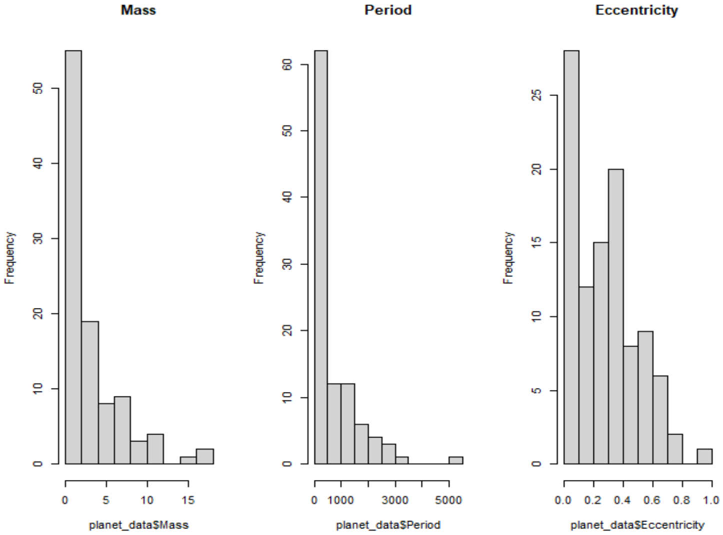
Biplot display of scores and loadings:



There’s lots of patterns that can be seen from this biplot display. Southern states tend to be placed in the further direction of the Illiteracy and Murder vectors. It appears Northern states tend to have higher levels of Frost and HS Graduates. Additionally, the biggest states are located in the upper portions of the graph, with population and area being large for these observations as indicated by the fact that the population and area have vectors in a similar general direction.

Question 3

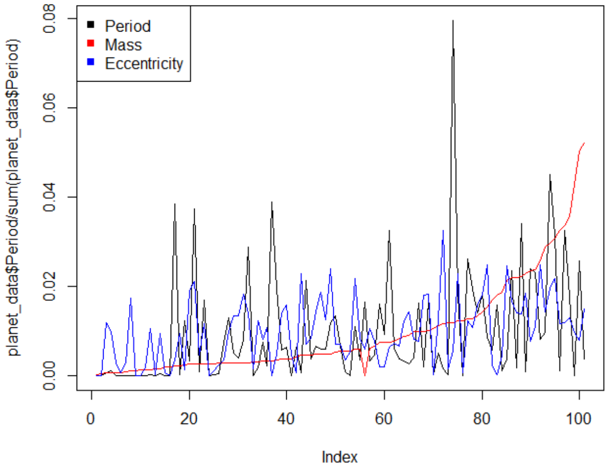
a. **Perform an exploratory analysis of the data. Be sure to examine the univariate distributions of the variables and their bivariate relationships using appropriate plots and summary statistics.** The planets dataset has 101 observations and 3 features: Mass, Period, and Eccentricity. The univariate distributions are similar for Mass and Period, but Eccentricity has a more uniform of a distribution than the highly skewed distributions of the other two features, as shown below:



Looking at the correlations and bivariate relationships via the plot function, we can see that these features have fairly low correlation; no feature pair exceeds even 0.33.

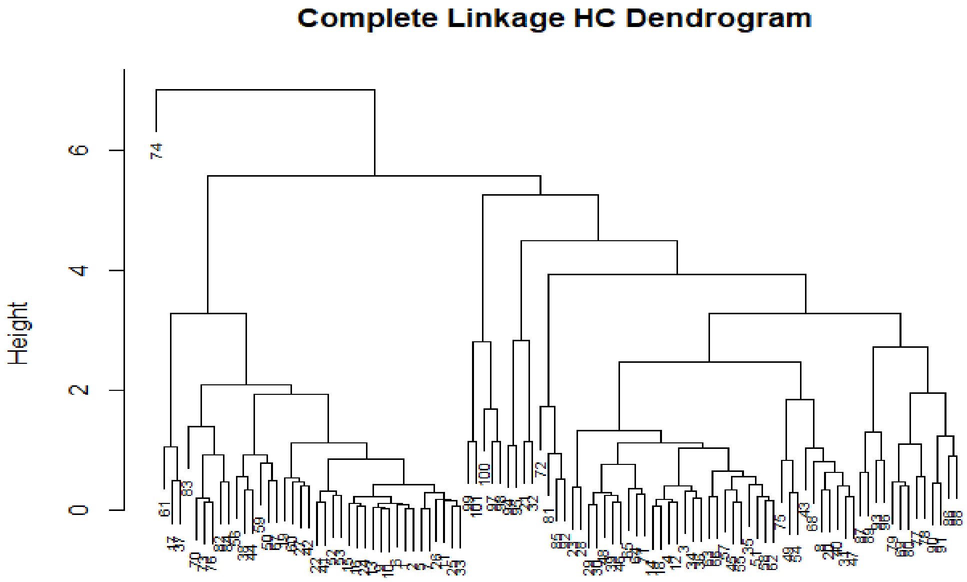
b. **Do you think standardizing the variables before clustering would be a good idea?** Yes, for the same reasons as before, it would be a good idea. Again, the standard deviations are high, the underlying distributions are unique at least in terms of Eccentricity, and the ranges of the feature values are high.

c. **Would you use metric-based or correlation-based distance to cluster the exoplanets?** It is important to note that upon standardizing our features, clustering based on correlation or metric distance is the same. However, looking at the correlation information and a variable index based on normalized data, we can tell that perhaps metric based would be better (for the normalized data, at least).



d. **Display the results using a dendogram. Cut the dendogram at a height that results in three distinct clusters. Summarize the cluster-specific means of the three variables. Also, make pairwise scatterplots of the three clusters. What do you observe?**

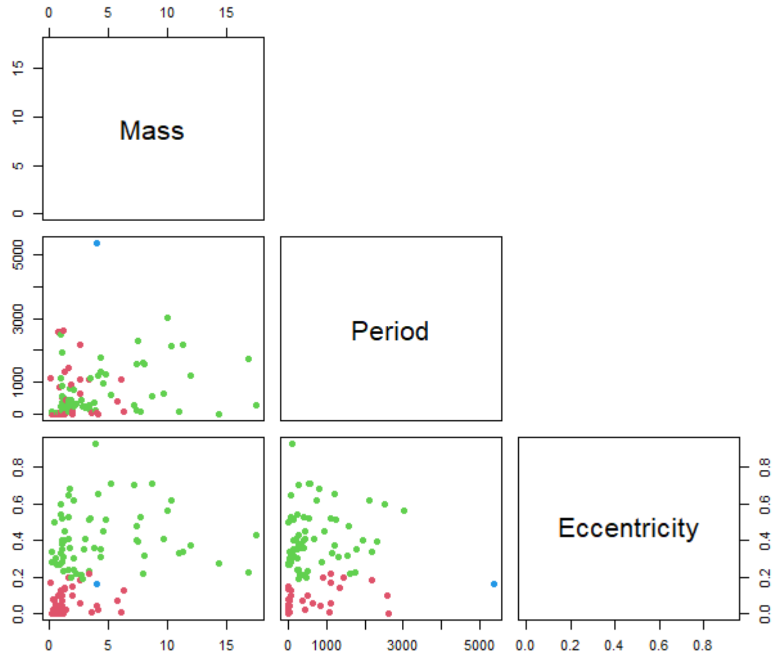
Complete link Dendrogram:



Cluster specific means:

	Mass	Period	Eccentricity
Cluster 1	1.70331579	488.74792711	0.07077105
Cluster 2	4.3117742	699.7941161	0.4126935
Cluster 3	4.00	5360.00	0.16

Pairwise Scatterplots:

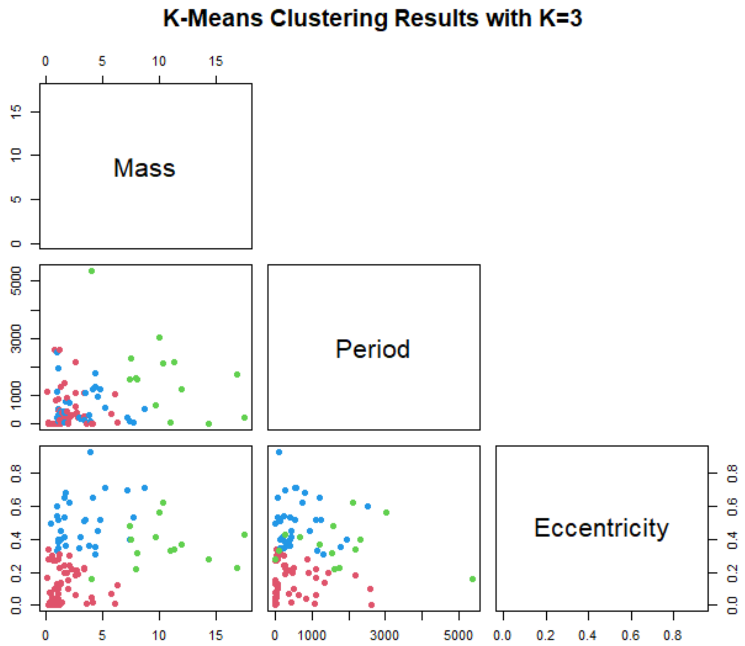


From these scatterplots, we can see that Mass and Eccentricity and Period and Eccentricity, separately, are bivariate criteria that can be used to reasonably differentiate observations that belong to Cluster 1 and Cluster 2. It should be noted that even though we have a cluster 3, it only has one observation and is very distant from all other observations.

e. **Repeat (d) using K-means clustering with K = 3. (Of course, you won't have a dendogram in this case.)**

Cluster-specific means:

```
         Mass      Period Eccentricity
1 -0.45734032 -0.28145496   -0.7374511
2  1.96722503  1.17498249    0.4027263
3 -0.08620525 -0.03791475    1.0358846
```



I observe that only one scatterplot (Mass vs. Eccentricity) has a reasonable interclass dissimilarity. The other two plots have observations from different clusters deeply mixed. The Mass and Eccentricity plot does show that observations in one cluster have low mass and low eccentricity, whereas another cluster has high mass and low eccentricity, and the final cluster has high eccentricity and low mass.

f. **Compare the results from the two clustering algorithms. Which clustering method would you recommend?**

This one was fairly hard for me to answer. Hierarchical clustering provided more options to cluster with, but K-means did not have one cluster that was practically an outlier, only having one subject. I will say K-means might be better since it should prevent a scientist from trying to categorize observations based off of period and eccentricity, but I do believe other linkage methods should be tried and perhaps limiting K to 2 would yield better clustering results.

Section 2

Question 1

```r
library(pls)

# Read in pc data and remove subject column
pc_data <- read.csv("prostate_cancer.csv")
pc_data <- pc_data[,-1]

# Treat vesinv as a qualitative variable
pc_data$vesinv <- factor(pc_data$vesinv, order=F, levels = c(0, 1))

# Conduct a natural log transformation on the response
# to adjust it's distribution to something more appropriate.
pc_data[, 1] <- log(pc_data[, 1])
hist(pc_data$psa)

# Calculate LOOCV estimate of MSE via pcr regression on prostate cancer data
prostate_pcr <- pcr(psa ~ ., data = pc_data, scale = TRUE, validation="LOO")
# Summary will provide more information
summary(prostate_pcr)

# Store MSE information for all components
pcr_MSE <- MSEP(prostate_pcr)$val[1, 1,]

# Find number of components corresponding to minimum MSE
which.min(pcr_MSE)
# Confirmation of minimum MSE point
validationplot(prostate_pcr, val.type = "MSEP")

# Test MSE
min(pcr_MSE)

# Train MSE
prostate_pcr_pred <- predict(prostate_pcr, pc_data, ncomp = 1)
mean((prostate_pcr_pred - pc_data$psa)^2)

# Calculate LOOCV estimate of MSE via a PLS regression on the prostate cancer
data
prostate_pls <- plsr(psa ~ ., data = pc_data, scale = TRUE, validation="LOO")

# Store MSE information for all components for the PLS regression
pls_MSE <- MSEP(prostate_pls)$val[1, 1,]

# Find number of components corresponding to minimum MSE for the PLS
regression
which.min(pls_MSE)
# Confirm minimum MSE point for the PLS regression
validationplot(prostate_pls, val.type="MSEP")

# Test MSE
min(pls_MSE)
```

Question 2

```
library(datasets)

# Read in state data
state_data <- state.x77

# See dimensions of data
nrow(state_data)
ncol(state_data)

# correlation matrix of data
cor(state_data)

#Standard deviation of population, income, and area
sd(state_data[,1])
sd(state_data[,2])
sd(state_data[,8])

# Display histograms of all features
par(mfrow=c(4,2))
for (feature in 1:ncol(state_data)) {
  hist(state_data[,feature])
}

# Scale state data for PCA
state_data <- scale(state_data)
state_pca <- prcomp((state_data), center = T, scale = T)

# Eigenvectors, scores, and variances provided via rotation, x, and stdev
squared
state_pca$rotation
state_pca$x
state_pca_var <- state_pca$sdev^2

# Find proportion of variance explained by normalizing variances
state_pve <- state_pca_var / sum(state_pca_var)

# Prove that the cumulative sum of the PVEs is 1.
cumsum(state_pve)

par(mfrow=c(1,1))
# Plot PVE on a scree plot to find number of principal components to use
plot(state_pve,
     xlab = "Principal Component", ylab = "Proportion of Variance Explained",
     ylim = c(0,1), type = 'b')
    #ylim is 0 to 1 because it's a probability

# Plotting cumulative sum can also help you to make the same decision
plot(cumsum(state_pve),
     xlab = "Principal Component", ylab = "Proportion of Variance Explained",
     ylim = c(0,1), type = 'b')
    #ylim is 0 to 1 because it's a probability

# Correlation between variable X and PC y is equivalent to (loading value for
that X and y) / (standard deviation of that feature X).
```

```
# Due to standardization we can see that standard deviations of each feature
is 1.
for(i in 1:ncol(state_data)){
  print(paste0("standard deviation of feature X", i, ": ",
sd(state_data[,i])))
}

# Correlation between all features and PCs 1 and 2 can thus be shown via this
cropping of the rotation matrix
# multiplied by the corresponding standard deviation associated with that
principal component

state_corr <- cbind(data.frame(state_pca$rotation[,1]*state_pca$sdev[1]),
state_pca$rotation[,2]*state_pca$sdev[2])
colnames(state_corr) <- c("PC1", "PC2")


# Proportion of variance explained summed for the first two Principal
Components.
cumsum(state_pve)[1:2]

# Biplot of scores and loadings
biplot(state_pca, scale=0)
```

Question 3

```
# Read in planet data from planet.csv
planet_data <- read.csv("planet.csv")

# Number of observations and features
nrow(planet_data)
ncol(planet_data)

# The standard deviation of planet data features can be large and vary
greatly
sd(planet_data$Period)
# Range via visual inspection of values is also quite different.
max(planet_data$Mass) - min(planet_data$Mass)
max(planet_data$Period) - min(planet_data$Period)
max(planet_data$Eccentricity) - min(planet_data$Eccentricity)
# As such I decided it would be a good idea to standardize the variables

# Univariate distributions
hist(planet_data$Mass)
hist(planet_data$Period)
hist(planet_data$Eccentricity)

# Bivariate relations
cor(planet_data)
plot(planet_data$Mass, planet_data$Period)
plot(planet_data$Mass, planet_data$Eccentricity)
plot(planet_data$Eccentricity, planet_data$Period)

# Variable index of normalized planet data
```

```r
plot(planet_data$Period/sum(planet_data$Period), type="l")
lines(planet_data$Mass/sum(planet_data$Mass), col="red")
lines(planet_data$Eccentricity/sum(planet_data$Eccentricity), col = "blue")
legend("topleft",legend=c("Period","Mass","Eccentricity"),
       pch=15, col=c("black", "red","blue"))

# Variable index of scaled planet data
scaled_planet_data <- data.frame(scale(planet_data))
plot(scaled_planet_data$Period, type="l")
lines(scaled_planet_data$Mass, col="red")
lines(scaled_planet_data$Eccentricity, col = "blue")
legend("topleft",legend=c("Period","Mass","Eccentricity"),
       pch=15, col=c("black", "red","blue"))


# Hierarchical clustering of scaled planet data using metric (euclidean)
distance
# and complete link
hc.complete <- hclust(dist(scaled_planet_data), method = "complete")
plot(hc.complete, main = "Complete Linkage HC Dendogram",
     xlab = "", sub = "", cex = 0.6)

# Get cluster assignments of all observations
three_cut <- cutree(hc.complete, k = 3)

# Find cluster-specific means by selecting rows matching cutree results
clust1 <- planet_data[which(three_cut == 1),]
# Then applying mean function over columns to figure out feature means
clust1_means <- apply(clust1, 2, mean)

# Repeat for cluster 2 and cluster 3 observations respectively
clust2 <- planet_data[which(three_cut == 2),]
clust2_means <- apply(clust2, 2, mean)

clust3 <- planet_data[which(three_cut == 3),]
clust3_means <- apply(clust3, 2, mean)

# Show pairwise plot of original planet data with colors based off of cutree
results
# to indicate which observation is in which cluster
pairs(planet_data, pch=19, col=(three_cut+1), upper.panel = NULL)

# Repeat the process but with K-means where K = 3
km.out <- kmeans(scaled_planet_data, 3, nstart = 20)
# Print out the centers values to get cluster-specific means
km.out$centers

# Display pairwise plot of original planet data with
# colors based off of K-means cluster assignments
plot(planet_data, col = (km.out$cluster + 1),
     main = "K-Means Clustering Results with K=3",
     pch = 19, cex = 1, upper.panel = NULL)
```