DSC 424 Assignment #3

1. For my group, I ran a Principal Component Analysis (PCA) on our dataset. Due to our dataset having 30 independent variables, we would need 30 different components to account for 100% of variance in our dataset. The parallel analysis suggested using 6 components. There are also 6 components with eigenvalues greater than one, and according the scree plot, the ‘knee’ appears to indicate that 3 or 4 components may be appropriate. I ran the PCA with 6 factors. The reason I ran the PCA with 6 components is because the cumulative variance in the factors continued to increase by a significant margin with all 6 of these factors. The cumulative variance of the first component is 35%, the second component adds an additional 17.6%, bringing the cumulative variance to 5.26%. The third and fourth components add 13.6% and 7.9% respectively bringing the cumulative variance to 74.1%. And the final two components add 7.7% and 6.7% respectively bringing the cumulative variance to 88.4% for our 6 components. The first component is made up of the variables representing the mean, standard error, and worst or largest measurements of the radius, perimeter, and area of the cell nuclei in addition to the number of concave points and severity of concave points. Because the first component focuses on the measurements of the different aspects of the cell nuclei, measurement summary could be a possible name for the first component. The second component is made up of the standard error of the concave points and severity of concave points, compactness and fractal dimensions. Because all of these variables are standard error measurements, standard error of concavity could be an appropriate name for this component. The third component is composed of the mean for smoothness and worst or largest smoothness measurements as well the worst or largest measurement for fractal dimension. This component could be referred to as smoothness and fractal dimension extreme. The fourth component is the standard error for texture and smoothness, texture error could be a possible name. The fifth component is made up of the mean and worst or largest measurement for texture, so texture measurements could be a possible name for this component. The sixth component is made up of the mean, standard error and worst value of symmetry, so symmetry or symmetry measurements could be a possible name for the sixth component. The equation for the first component is:

$$
\begin{aligned}
RC1 = 0.96\mathrm{radius}_{\mathrm{mean}} + {} & .96\, perimeter_{mean} + 0.97\, area_{mean} + 0.66\, concavity_{mean} \\
& + .82\, concave.points_{mean} + 0.83\, radius_{se} + 0.82 perimeter_{se} + 0.87 area_{se} \\
& + 0.95 radius_{worst} + 0.95 perimeter_{worst} + 0.96 area_{worst} \\
& + 0.7 concave.points_{worst}
\end{aligned}
$$

R Code:

```
library(Hmisc) #Describe Function
library(psych) #Multiple Functions for Statistics and Multivariate Analysis
library(GGally) #ggpairs Function
library(ggplot2) #ggplot2 Functions
library(vioplot) #Violin Plot Function
library(corrplot) #Plot Correlations
library(REdaS) #Bartlett's Test of Sphericity
library(psych) #PCA/FA functions
library(factoextra) #PCA Visualizations
library("FactoMineR") #PCA functions
library(ade4) #PCA Visualizations
library(car)
install.packages('lm.beta')
library(lm.beta)

install.packages('fastDummies')
library('fastDummies')

#Read in IBreast Cancer Data
setwd("/Users/danielobrien/desktop")

bcancer = read.csv("data.csv", header = TRUE, sep = ",")
head(bcancer)

bcancer2 = subset(bcancer, select = -c(X))

head(bcancer2)

#Identify all zero values as NA and delete all values

bcancer2[bcancer2==0] <- NA

sum(is.na(bcancer2))

#Listwise Deletion
bcancer3 <- na.omit(bcancer2)

head(bcancer3)

sum(is.na(bcancer3))

#Create dummy variable for response variables, and remove all unnecessary variables
install.packages('fastDummies')
library('fastDummies')

bcancer4 <- dummy_cols(bcancer3, select_columns = c('diagnosis'), remove_selected_columns = TRUE)

head(bcancer4)
```

```
bcancer5 = subset(bcancer4, select = -c(id, diagnosis_B))

head(bcancer5)


library(psych)
describe(bcancer5)

cor.bcancer5 = cor(bcancer5)
cor.bcancer5
corrplot(cor.bcancer5, method="ellipse")
corrplot(cor.bcancer5, method="number")
corrplot(cor.bcancer5, method="circle",col=c("yellow", "red","blue","green"))

#Full Model
attr(bcancer5,'variable.labels')
fullFit = lm(diagnosis_M~ ., data=bcancer5)
summary(fullFit)

lm.beta.fullfit <- lm.beta(fullFit)
lm.beta.fullfit

vif(fullFit)

pcacancer = subset(bcancer5, select = -c(diagnosis_M))

#PCA Code
PCA_Plot = function(pcaData)
{
  library(ggplot2)

  theta = seq(0,2*pi,length.out = 100)
  circle = data.frame(x = cos(theta), y = sin(theta))
  p = ggplot(circle,aes(x,y)) + geom_path()

  loadings = data.frame(pcaData$rotation, .names = row.names(pcaData$rotation))
  p + geom_text(data=loadings, mapping=aes(x = PC1, y = PC2, label = .names, colour = .names, 
fontface="bold")) +
    coord_fixed(ratio=1) + labs(x = "PC1", y = "PC2")
}

PCA_Plot_Secondary = function(pcaData)
{
  library(ggplot2)
  theta = seq(0,2*pi,length.out = 100)
  circle = data.frame(x = cos(theta), y = sin(theta))
  p = ggplot(circle,aes(x,y)) + geom_path()

  loadings = data.frame(pcaData$rotation, .names = row.names(pcaData$rotation))
  p + geom_text(data=loadings, mapping=aes(x = PC3, y = PC4, label = .names, colour = .names,
fontface="bold")) +
    coord_fixed(ratio=1) + labs(x = "PC3", y = "PC4")
}

PCA_Plot_Psyc = function(pcaData)
{
  library(ggplot2)

  theta = seq(0,2*pi,length.out = 100)
  circle = data.frame(x = cos(theta), y = sin(theta))
  p = ggplot(circle,aes(x,y)) + geom_path()

  loadings = as.data.frame(unclass(pcaData$loadings))
  s = rep(0, ncol(loadings))
  for (i in 1:ncol(loadings))
  {
    s[i] = 0
    for (j in 1:nrow(loadings))
      s[i] = s[i] + loadings[j, i]^2
    s[i] = sqrt(s[i])
  }

  for (i in 1:ncol(loadings))
    loadings[, i] = loadings[, i] / s[i]

  loadings$.names = row.names(loadings)

  p + geom_text(data=loadings, mapping=aes(x = PC1, y = PC2, label = .names, colour = .names,
fontface="bold")) +
    coord_fixed(ratio=1) + labs(x = "PC1", y = "PC2")
}

PCA_Plot_Psyc_Secondary = function(pcaData)
{
  library(ggplot2)

  theta = seq(0,2*pi,length.out = 100)
  circle = data.frame(x = cos(theta), y = sin(theta))
  p = ggplot(circle,aes(x,y)) + geom_path()

  loadings = as.data.frame(unclass(pcaData$loadings))
  s = rep(0, ncol(loadings))
  for (i in 1:ncol(loadings))
  {
    s[i] = 0
    for (j in 1:nrow(loadings))
      s[i] = s[i] + loadings[j, i]^2
    s[i] = sqrt(s[i])
  }

  for (i in 1:ncol(loadings))
    loadings[, i] = loadings[, i] / s[i]

  loadings$.names = row.names(loadings)

  print(loadings)
  p + geom_text(data=loadings, mapping=aes(x = PC3, y = PC4, label = .names, colour = .names,
fontface="bold")) +
    coord_fixed(ratio=1) + labs(x = "PC3", y = "PC4")
}

#Test KMO Sampling Adequacy

library(psych)
KMO(pcacancer)
#Overall MSA =  0.83

#Test Bartlett's Test of Sphericity
library(REdaS)
bart_spher(pcacancer)
#p-value < 2.22e-16 (Very Small Number)


#Test for Reliability Analysis using Cronbach's Alpha
library(psych)
alpha(pcacancer,check.keys=TRUE)
#raw_alpha = 0.94

library(psych)
comp <- fa.parallel(pcacancer)
comp
```

```
p = prcomp(pcacancer, center=T, scale=T)
p

#Scree Plot
plot(p)
abline(1, 0)

#Check PCA Summary Information
summary(p)
print(p)


p2 = psych::principal(pcacancer, rotate="varimax", nfactors=6, scores=TRUE)
p2
print(p2$loadings, cutoff=.6, sort=T)
```

2. a. The data is suitable for canonical correlation. One indication that this data is suitable for canonical correlation is that they are comparing two different groups of information. The region of Beijing's logistical capacity and Beijing's regional economic development are both being used as factors in this study, and because we have two distinct groups of data that will be compared and examined, canonical correlation is certainly appropriate in this setting.
b. The two groups being examining in this canonical correlation are regional economic development and system of regional logistic capabilities. The regional economic development is composed of our x-values. Those x-values are GDP, Industrial GDP, Tertiary Industrial GDP, the total retail sales of consumer goods, GDP per capita, and consumption level of residents. The system of regional logistic capabilities is made up of the y-values. The y-variables are freight volume, rotation volume of freight transport, business total of Posts and telecommunication, infrastructure mileage, and the logistic industry output value. All of the variables appear to be continuous variables. The units of the different variables were not all of the same units, so the variables were standardized to allow the researchers to compare the variables.
c. There were 6 x-variables and only 5 y-variables, so 5 canonical correlations are the absolute maximum the researchers could use. However, after examining the significance of each canonical correlation coefficient, they found that only the first two canonical correlation coefficients were significant, due to the p-values being above 0.05 for every canonical correlation coefficient after the first two. Researchers also found that there were some pretty high correlations between x-variables and y-variables, many reaching an almost perfect correlation, particularly with the 2nd and 5th y-variables, showing correlations above 0.97 for all of the x-variables. The researchers also canonical correlations and found that the first and second canonical correlations were above 0.99, the third having a value near 0.85, the fourth with a value of 0.515 and the fifth with a value of 0.472. When researchers ran a redundant analysis and found that regional economic indicators can be explained 87.5% by the first variate(U1), and 10.8% by the second variate(U2). Additionally, 87.2% and 10.8% of regional logistic capabilities are explained by the first and second variates respectively, with the

remaining variates explaining very little. They also found that the typical variables V1, V2, and V3 explained 60.2%, 16.1% and 18.1% of the fluctuations in the regional logistic capability index respectively. And that 59.9%, 15.9% and 13% of the fluctuations in the regional economic indicators can be explained by V1, V2, and V3 respectively. Since U1 and U2 had a much stronger representation than their other variables, U3 and V3, U4 and V4 and U5 and V5 were left off, and the researchers only focused on the first two pairs, U1 and V1 and U2 and V2.
d. They ended up choosing 2 correlates based in the p-values as mentioned above. They attempted to interpret the correlates in terms of their original values in the redundant analysis. Researchers found that the variables x1 and x3 carried a lot of weight for the variable U1, meaning the combination of GDP and tertiary industrial added value contributed a lot in this instance. Additionally, y5 made a big impact for V1, demonstrating logistical industry output value made a significant impact on that specific variate. This was also used to demonstrate that there could be an association between GDP and logistical industry output.
e. There were a few different conclusions drawn for the canonical correlation. One conclusion is that regional economic development and regional logistic capacity in Beijing will become more and more influential. Additionally, GDP and tertiary industry help regional logistics capacity by elevating its role. Also, rotation volume of freight transport, the logistics industry output and infrastructure construction could excel regional logistics capacity.

3.
Part 1:

a. The null hypothesis tells us that the canonical correlations are all equal to zero. When running the Wilk's Lambda test, we can tell that there are 9 significant canonical correlations that all had a p-value less than 0.05. The first 8 canonical correlations had p-values so small that they appeared to be 0 in the Wilk's Lambda test, and the 9th canonical correlation had a p-value of 0.03. After running the yacca package, we can see that we have canonical correlation values greater than 0, so we can reject the null hypothesis. The values of the first 9 canonical correlations are as follows, 0.73, 0.61, 0.52, 0.47, 0.39, 0.37, 0.34, 0.32, and 0.30.
b. There are 9 significant canonical variates. I used the p-values to determine how many canonical variates were significant and 9 canonical correlations had p-values less that 0.05.
c. The first canonical correlation has a value of 0.733. The second canonical correlation has a value of 0.606. From the canonical variate adequacies, we can see that for the x-variables, about 8% of the variance is explained using the first variant, and about 8.5% of the variance is explained by the second canonical variate. For the y-variables, about 16% of the variance is explained by the first canonical variate and about 10% is explained by the second variate. For the first canonical correlation, the loading show that in terms of our x-variables, hobbies and interests, history, poetry reading, art exhibitions, musical instruments, and theater have the highest values, close to 0.5
d. We can conclude that the first canonical variate does not explain the variance in our x and y variables as much as the second canonical correlation.

Part 2:

a. Formula for the first canonical variate for hobbies and interests:

$$\begin{aligned} CV1 = {} & 0.46 History + 0.26\, Psychology + 0.21 Politics + 0.11 Mathematics \\ & + 0.26 Physics - 0.17 internet + 0.02 PC - 0.13 Economy_{management} \\ & + 0.13 Biology + 0.07 Chemistry + 0.50 Reading + 0.20 Geography \\ & + 0.27 Foreign_{language} + 0.15 medicine + 0.05 law - 0.20 cars \\ & + 0.60 art.exhibitions + .034 religion + 0.23 countryside.outdoors \\ & - 0.12 dancing + 0.49 musical.intruments + 0.34 writing \\ & - 0.12 passive.sport - 0.15 active.sport + 0.08 gardening \\ & - 0.38 celebrities - 0.35 shopping + 0.29 science.and.technology \\ & + 0.50 theater - 0.05 fun.with.friends - 0.12 adrenaline.sports \\ & - 0.12 pets \end{aligned}$$

Formula for the first canonical variate for music:

$$\begin{aligned} CV1 = {} & 0.06 music - 0.19 slow.songs.or.fast.songs - 0.39 dance + 0.38 folk \\ & + 0.26 country + 0.76 classical.music + 0.26 musical - 0.44 pop \\ & + 0.36 rock + 0.38 metal.or.hardrock + 0.29 punk \\ & - 0.45 hiphop.rap + 0.11 reggae.ska + 0.54 swing.jazz \\ & + 0.36 rock.n.roll + 0.60 alternative - 0.07 latino \\ & - o.21 techno.trance + 0.66 opera \end{aligned}$$

b. For the first canonical correlation, the loading show that in terms of our x-variables, hobbies and interests, history, poetry reading, art exhibitions, musical instruments, and theater have the highest values, close to 0.5. For the first canonical correlation, the structural correlation loadings of the y-variable, music, classical music, swing/jazz, alternative, and opera have the highest positive values above 0.5, and pop and hip-hop/rap have the highest negative values close to -0.5. For the second canonical correlation, the x-variables, hobbies/interests with the highest structural loadings are dancing, shopping, celebrities and theater, with the largest negative value belonging to PC. For the y-variables, music, the highest structural loadings are musical, Latino, and pop. The highest negative structural loading values belong to metal/hard rock.

c. From the first two canonical correlations we can see that interests and hobbies related to the intellect and the arts, history, poetry reading, art exhibits, musical instruments and theater have an association with classical, swing/jazz, alternative and opera music. We can see that there is a negative association between pop music and hip-hop/rap with the above hobbies and interests. For the second canonical correlation, we can see that there is a positive association with the hobbies of dancing, shopping, celebrity lifestyle and theater have a positive association with interests in musicals, Latin music and pop music with a negative association with metal or hard rock music. We can reasonably assume that people who are interested in arts, history poetry reading, playing musical instruments and theaters would also be interested in classical, swing/jazz, alternative and opera music. And people who are interested in dancing, shopping, celebrity lifestyle and theater may be interested in pop, musicals or Latin music.