ADA2 hw6, due April 25 Thursday

Problem 1: NM County-level Poverty Data

In this example we'll use NM county-level poverty data to understand how counties differ by living conditions, and how those living conditions vary together. We hope to reduce our 13-dimensional dataset to the vital few components that explain about 75% of the variability.

Here is a description of the codebook for this data.

```
NM county-level poverty data from S16 student:
Nathan Dobie, Student Technical Specialist, Bureau of Business Economic Research, UNM
Thanks, Nathan!

Data combined from:
http://bber.unm.edu/county-profiles                                          (poverty)
http://factfinder.census.gov/bkmk/table/1.0/en/ACS/14_5YR/DP04/0400000US35 (other values)
http://www2.census.gov/geo/docs/reference/codes/files/national_county.txt  (county names)

DATA COLUMNS:
 1 area
 2 county
 3 periodyear (2014)
   -Vacancy Status %
 4   Homeowner vacancy rate
 5   Rental vacancy rate
   -Occupancy Status %
 6   Owner-occupied
 7   Renter-occupied
   -Main source of heating (% of homes)
 8   Utility gas
 9   Electricity
10   Wood
11 Lacking complete plumbing facilities %
12 No telephone service available %
13 rentover35        (gross rent as a percentage of household income (grapi))
   -Poverty
14   est_percent     (Estimated percent of people of all ages in poverty)
15   child_percent   (Estimate of people age 0-17 in poverty)
16   fam_percent     (Estimated percent of related children age 5-17 in families in poverty)
```

```
fn.data <- "http://statacumen.com/teach/ADA2/homework/ADA2_HW_11_PCA_NMCensusPovertyHousingCharacteristi
nm.census <- read.csv(fn.data, header=TRUE, skip=1, as.is = TRUE)
# remove state average, use county-level
nm.census <- subset(nm.census, area != 0)

# Shorter column names
colnames(nm.census) <- c("Area", "County", "Year"
                       , "VacantH", "VacantR"
                       , "Owner", "Renter"
                       , "HeatG", "HeatE", "HeatW"
                       , "NoPlumb", "NoPhone", "Rent35"
                       , "PovAll", "PovChild", "PovFam")
```

```r
str(nm.census)
```

```
'data.frame':   33 obs. of  16 variables:
 $ Area    : int  1 3 5 6 7 9 11 13 15 17 ...
 $ County  : chr  "Bernalillo" "Catron" "Chaves" "Cibola" ...
 $ Year    : int  2014 2014 2014 2014 2014 2014 2014 2014 2014 2014 ...
 $ VacantH : num  1.7 14.8 2.3 1.6 7.4 3.9 11.4 2.1 0.4 3.2 ...
 $ VacantR : num  6.9 7.5 7.8 6.8 20.4 7 8.5 7.4 7.5 8.1 ...
 $ Owner   : num  62.4 87.2 65.4 74.8 67.6 59.4 82.7 64.7 73.5 75.6 ...
 $ Renter  : num  37.6 12.8 34.6 25.2 32.4 40.6 17.3 35.3 26.5 24.4 ...
 $ HeatG   : num  81.7 3.8 50.1 49.1 50.2 47 46.6 70.4 53.5 51.4 ...
 $ HeatE   : num  13 2.8 42.6 10 15.4 46.4 19.1 15.6 38.7 18.6 ...
 $ HeatW   : num  2 51.2 1.9 21.7 13.7 1.1 9 1.6 1 10.6 ...
 $ NoPlumb : num  0.5 0.9 0.5 5.2 0.1 0.1 0 0.7 0.7 1.2 ...
 $ NoPhone : num  3 2.4 2.8 3.5 4.4 3.2 4.5 3.1 2.2 3 ...
 $ Rent35  : num  43.8 51.7 36.7 45.1 38 42 0 46.9 31.4 41.9 ...
 $ PovAll  : num  18.7 22.2 23.4 28.8 20.5 19.2 20.6 27.9 14.1 19.1 ...
 $ PovChild: num  24.5 42.8 32.4 37.6 30.6 27.3 32.1 39.4 18.5 27.8 ...
 $ PovFam  : num  22.6 40.1 28.7 35.9 27.2 26.7 31.6 36 17.3 25.3 ...
```

```r
head(nm.census, 3)
```

```
  Area     County Year VacantH VacantR Owner Renter HeatG HeatE HeatW
2    1 Bernalillo 2014     1.7     6.9  62.4   37.6  81.7  13.0   2.0
3    3     Catron 2014    14.8     7.5  87.2   12.8   3.8   2.8  51.2
4    5     Chaves 2014     2.3     7.8  65.4   34.6  50.1  42.6   1.9
  NoPlumb NoPhone Rent35 PovAll PovChild PovFam
2     0.5     3.0   43.8   18.7     24.5   22.6
3     0.9     2.4   51.7   22.2     42.8   40.1
4     0.5     2.8   36.7   23.4     32.4   28.7
```

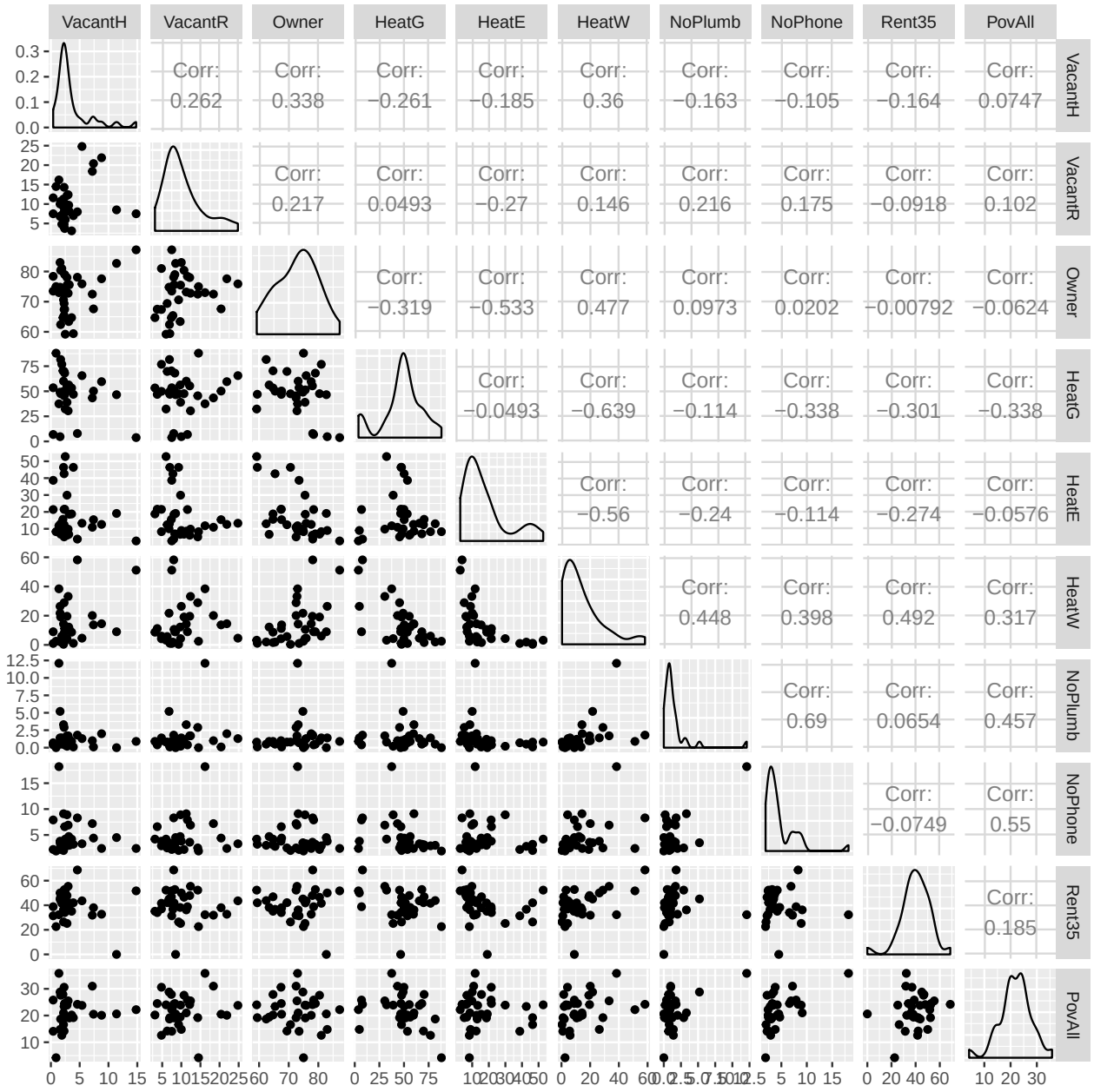
```r
nrow(nm.census)  #33 counties
```

```
[1] 33
```

```r
# columns to use for analysis
ind.col <- c(4:6, 8:14)
```

(1.) Below is the scatterplot matrix of variables of interest, describe qualitatively what you see.

```r
# Scatterplot matrix
library(ggplot2)
library(GGally)
p <- ggpairs(subset(nm.census, select = ind.col))
print(p)
```



(2.) One county has the least plumbing and phone service, in the following analysis, remove that county and redo the scatterplot matrix, describe qualitatively what you see.

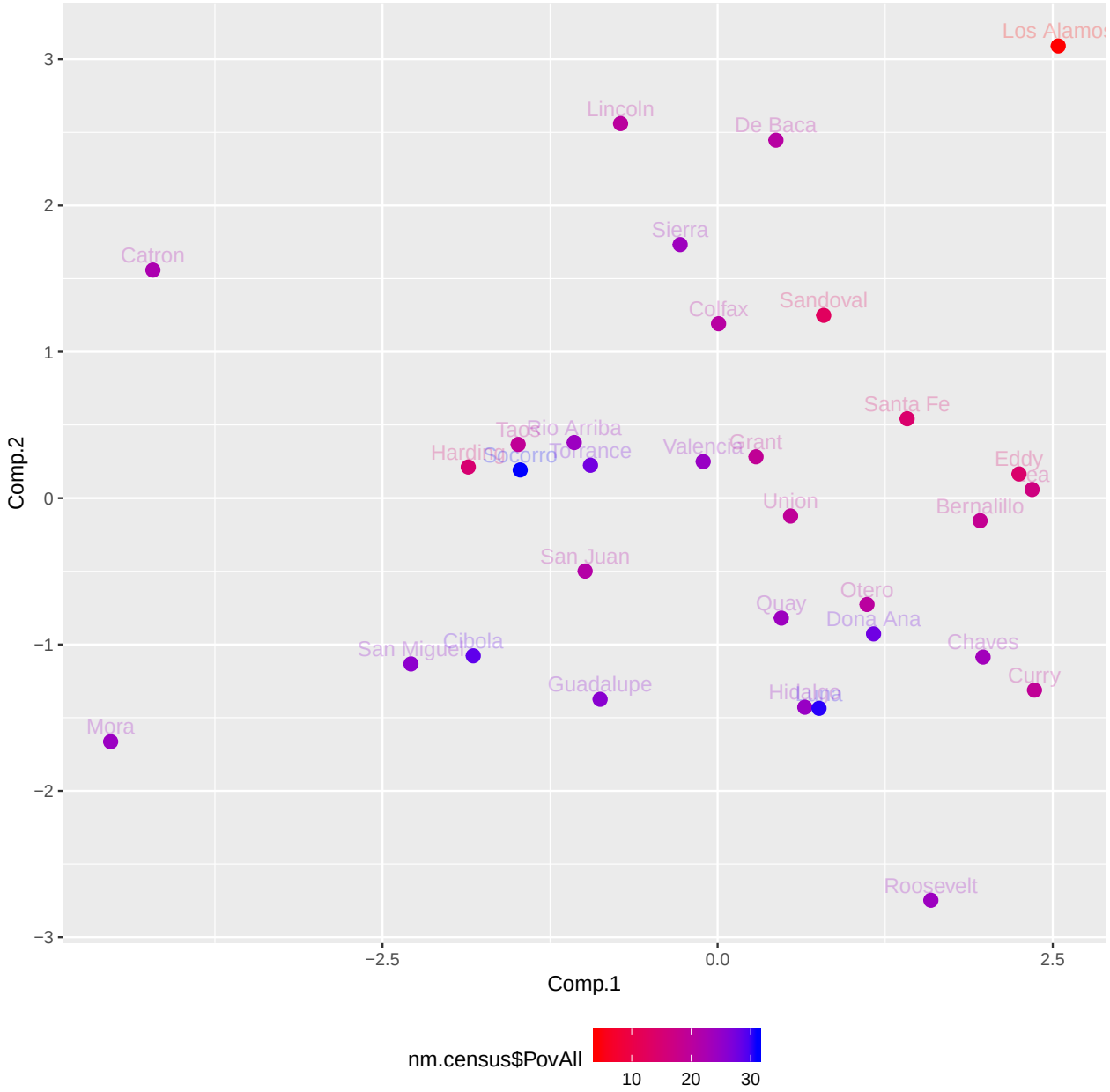
In the following analysis, let's continue with data without the extreme county. Now perform PCA analysis using correlation matrix, answer the following questions.

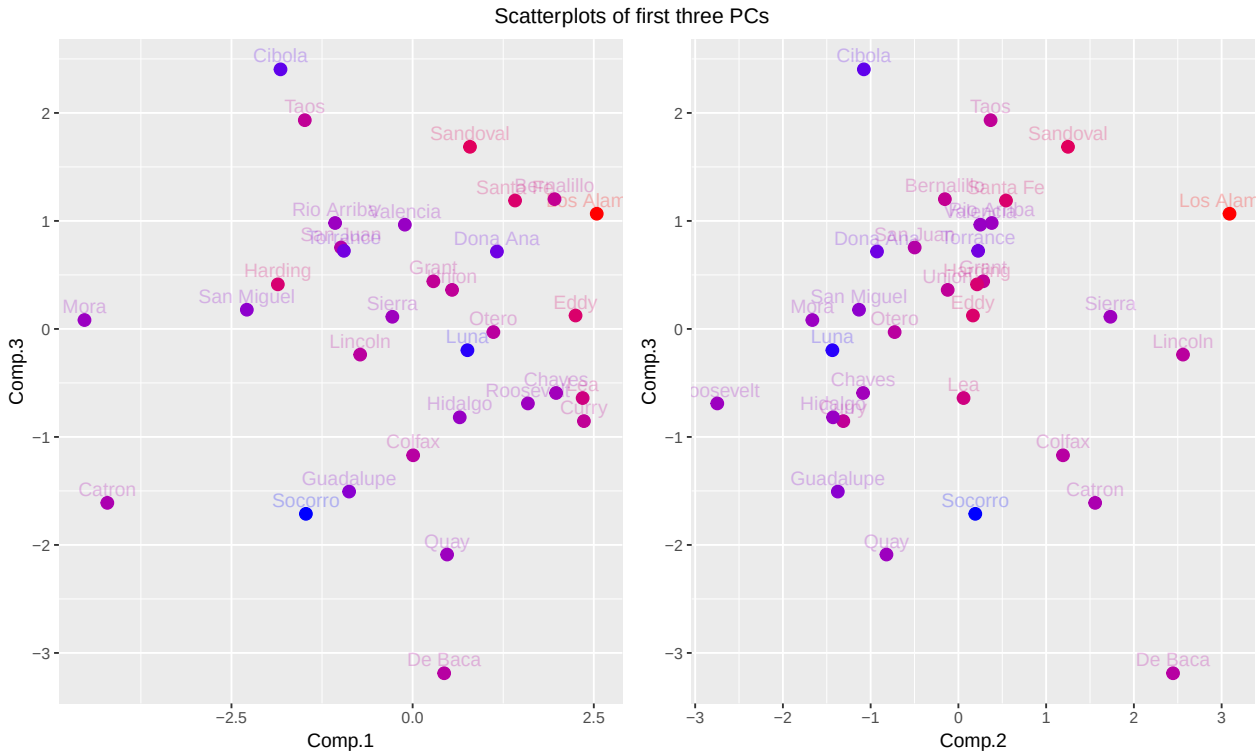
(3.) How many principal components would you retain to explain about 3/4ths of the total variability? How much variability is actually retained?

(4.) Interpret the principal components you retained in the previous step.

(5.) Below are two-dimensional plots of PC1 against PC2, PC1 against PC3, and PC2 against PC3. The points and labels are colored by poverty proportion. Using your interpretations of PC1 and PC2 etc, describe these three counties:

Bernalillo, Mora, and Roosevelt.





As an example, here's a description for Los Alamos.

Los Alamos has large PC1 and large PC2, this indicates

(both) there is very low poverty,

(PC1) dwellings heat with gas and electric, and

(PC2) there tends to be high dwelling vacancy.

The characteristics roughly match that description:

```
subset(nm.census, County == "Los Alamos")
```

```
   Area      County Year VacantH VacantR Owner Renter HeatG HeatE HeatW
17   28 Los Alamos 2014     0.9    14.5    75     25    88   8.3   2.5
   NoPlumb NoPhone Rent35 PovAll PovChild PovFam
17       0     1.9   22.6    4.2      4.6    3.8
```

Problem 2: Parkinson's disease data

You will use a dataset used to try to understand Parkinson's disease. The dataset includes 31 people, 23 of whom have Parkinson's. There are multiple observations per patient, which means that the observations are not independent. In the first column, a value like phon_R01_S13_3 indicates that this is the third observation from subject 13. A more thorough description of the data can be found from http://archive.ics.uci.edu/ml/datasets/parkinsons. Note that except for the first column, which gives the subject, and the column for status, all other variables are quantitative. For the status variable, a 1 indicates that the subject has Parkinson's disease, and a 0 indicates otherwise.

To read in the data, you can download it from the class webpage as parkinsons.csv. Or

```
parkinsons <- read.csv("https://math.unm.edu/~luyan/ADA219/parkinsons.csv")
nrow(parkinsons)
```

```
[1] 195
```

```
head(parkinsons)
```

```
            name MDVP.Fo.Hz. MDVP.Fhi.Hz. MDVP.Flo.Hz. MDVP.Jitter...
1 phon_R01_S01_1     119.992      157.302       74.997        0.00784
2 phon_R01_S01_2     122.400      148.650      113.819        0.00968
3 phon_R01_S01_3     116.682      131.111      111.555        0.01050
4 phon_R01_S01_4     116.676      137.871      111.366        0.00997
5 phon_R01_S01_5     116.014      141.781      110.655        0.01284
6 phon_R01_S01_6     120.552      131.162      113.787        0.00968
  MDVP.Jitter.Abs. MDVP.RAP MDVP.PPQ Jitter.DDP MDVP.Shimmer
1          0.00007  0.00370  0.00554    0.01109      0.04374
2          0.00008  0.00465  0.00696    0.01394      0.06134
3          0.00009  0.00544  0.00781    0.01633      0.05233
4          0.00009  0.00502  0.00698    0.01505      0.05492
5          0.00011  0.00655  0.00908    0.01966      0.06425
6          0.00008  0.00463  0.00750    0.01388      0.04701
  MDVP.Shimmer.dB. Shimmer.APQ3 Shimmer.APQ5 MDVP.APQ Shimmer.DDA     NHR
1            0.426      0.02182      0.03130  0.02971     0.06545 0.02211
2            0.626      0.03134      0.04518  0.04368     0.09403 0.01929
3            0.482      0.02757      0.03858  0.03590     0.08270 0.01309
4            0.517      0.02924      0.04005  0.03772     0.08771 0.01353
5            0.584      0.03490      0.04825  0.04465     0.10470 0.01767
6            0.456      0.02328      0.03526  0.03243     0.06985 0.01222
     HNR status     RPDE      DFA   spread1  spread2       D2      PPE
1 21.033      1 0.414783 0.815285 -4.813031 0.266482 2.301442 0.284654
2 19.085      1 0.458359 0.819521 -4.075192 0.335590 2.486855 0.368674
3 20.651      1 0.429895 0.825288 -4.443179 0.311173 2.342259 0.332634
4 20.644      1 0.434969 0.819235 -4.117501 0.334147 2.405554 0.368975
5 19.649      1 0.417356 0.823484 -3.747787 0.234513 2.332180 0.410335
6 21.378      1 0.415564 0.825069 -4.242867 0.299111 2.187560 0.357775
```

Note that R will modify some of the variable names, for example changing underscores to periods. Just let R do this automatically.

___(1.)___Now cluster the observations using only the variables MDVP.Fo.Hz., MDVP.Fhi.Hz., and MDVP.Flo.Hz.. Make dendrograms of your clustering using average linkage, single linkage, and complete linkage. Describe similarities or differences between the different clustering methods.

___(2.)___Comment on the dendrograms. Do observations within the same individual tend to cluster together or not?

___(3.)___Do cases with Parkinson's tend to cluster together or not?