Ass 5: Dimension Reduction

Chloe Hall

2022-11-03

Dimension Reduction

Research Question

What are the most relevant predictive variable categories for housing price in Nashville?

Data

The data used in this analysis is the Nashville Housing data set which comprises the home value data of properties sold in Nashville from 2013-16. The data set contains information about the different properties included in the reported housing sale including: land use, property address, sale date, sale price, whether the property was sold as vacant or part of a parcel, tax district, neighborhood, land value, building value, total value, foundation type, year built, number of bedrooms, and finished area for each of 56,000 parcel ids sold in Nashville from 2013-16.

I will be using the variables Finished Area, Land Value, Total Value, Acreage, Bedrooms, Full Bath & Half Bath in order in the beginning and then seeing what variables this could simply down to.

The data used in this assignment is available at https://www.kaggle.com/datasets/tmthyjames/nashville-housing-data

Data Wrangling

In order to organize the data successfully the following steps were completed. 1. Called in the dataframe 2. Filtered the dataframe to the variables I would be using in the regression. (add the real steps!!)

Read and wrangle data.

```
#Loading the necessary libraries
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
#importing the data set & Filtering to relevant variables and filtering out NAs
nash <- read_csv('~/Downloads/DS 3000/Assignments/Nashville_housing_data_2013_2016.csv')%>%
  na_if("") %>% #convert empty cells to NA
```

```
  dplyr::select(`Sale Price`, `Acreage`, `Finished Area`, `Land Value`, `Building Value`, `Total Value`
  na.omit() #we lose about 1/2 of the data here.
```

```
## New names:
## Rows: 56636 Columns: 31
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (17): Parcel ID, Land Use, Property Address, Suite/ Condo #, Property... dbl
## (13): ...1, Unnamed: 0, Sale Price, Acreage, Neighborhood, Land Value, ... date
## (1): Sale Date
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

The PCA

Checking for multicollinearity (r > .899)

```
str(nash)
```

```
## tibble [24,014 x 9] (S3: tbl_df/tbl/data.frame)
##  $ Sale Price    : num [1:24014] 191500 202000 32000 102000 93736 ...
##  $ Acreage       : num [1:24014] 0.17 0.11 0.17 0.34 0.17 0.2 0.2 0.4 0.34 0.23 ...
##  $ Finished Area : num [1:24014] 1149 2091 2146 1969 1037 ...
##  $ Land Value    : num [1:24014] 32000 34000 25000 25000 25000 16000 16000 25000 25000 21500 ...
##  $ Building Value: num [1:24014] 134400 157800 243700 138100 86100 ...
##  $ Total Value   : num [1:24014] 168300 191800 268700 164800 113300 ...
##  $ Bedrooms      : num [1:24014] 2 3 4 2 2 2 2 2 2 3 ...
##  $ Full Bath     : num [1:24014] 1 2 2 1 1 1 1 1 1 1 ...
##  $ Half Bath     : num [1:24014] 0 1 0 0 0 0 0 0 0 1 ...
##  - attr(*, "na.action")= 'omit' Named int [1:32622] 1 7 18 19 26 27 29 30 31 32 ...
##   ..- attr(*, "names")= chr [1:32622] "1" "7" "18" "19" ...
```

```
nash_tib_cor <- nash[,2:9]

corr_nash <- cor(nash_tib_cor)

corr_nash
```

```
##                   Acreage Finished Area Land Value Building Value Total Value
## Acreage        1.00000000     0.3352372  0.3247902      0.2231078   0.2878823
## Finished Area  0.33523718     1.0000000  0.6532044      0.8380655   0.8519757
## Land Value     0.32479024     0.6532044  1.0000000      0.6160393   0.8234128
## Building Value 0.22310784     0.8380655  0.6160393      1.0000000   0.9537241
## Total Value    0.28788226     0.8519757  0.8234128      0.9537241   1.0000000
## Bedrooms       0.15511693     0.5828939  0.3469898      0.4474114   0.4537437
## Full Bath      0.22058335     0.7525905  0.5418612      0.6465790   0.6711824
## Half Bath      0.05455697     0.3518423  0.1878543      0.3789126   0.3433749
##                  Bedrooms  Full Bath  Half Bath
## Acreage        0.1551169 0.22058335 0.05455697
## Finished Area  0.5828939 0.75259054 0.35184235
## Land Value     0.3469898 0.54186119 0.18785435
## Building Value 0.4474114 0.64657898 0.37891263
## Total Value    0.4537437 0.67118235 0.34337490
## Bedrooms       1.0000000 0.61281820 0.19367261
```

```
## Full Bath      0.6128182 1.00000000 0.08884099
## Half Bath      0.1936726 0.08884099 1.00000000
```

Building Value and Total Value have a 0.9537241 correlation so we will need to remove the less predictive option

```
cor(nash) #This shows building value is the less predictive option on the sale price so it will be remo
```

```
##                Sale Price    Acreage Finished Area Land Value Building Value
## Sale Price      1.0000000 0.28607106     0.7041204  0.7441424      0.7290494
## Acreage         0.2860711 1.00000000     0.3352372  0.3247902      0.2231078
## Finished Area   0.7041204 0.33523718     1.0000000  0.6532044      0.8380655
## Land Value      0.7441424 0.32479024     0.6532044  1.0000000      0.6160393
## Building Value  0.7290494 0.22310784     0.8380655  0.6160393      1.0000000
## Total Value     0.8097018 0.28788226     0.8519757  0.8234128      0.9537241
## Bedrooms        0.3744258 0.15511693     0.5828939  0.3469898      0.4474114
## Full Bath       0.5506111 0.22058335     0.7525905  0.5418612      0.6465790
## Half Bath       0.2523105 0.05455697     0.3518423  0.1878543      0.3789126
##                Total Value  Bedrooms  Full Bath  Half Bath
## Sale Price       0.8097018 0.3744258 0.55061110 0.25231053
## Acreage          0.2878823 0.1551169 0.22058335 0.05455697
## Finished Area    0.8519757 0.5828939 0.75259054 0.35184235
## Land Value       0.8234128 0.3469898 0.54186119 0.18785435
## Building Value   0.9537241 0.4474114 0.64657898 0.37891263
## Total Value      1.0000000 0.4537437 0.67118235 0.34337490
## Bedrooms         0.4537437 1.0000000 0.61281820 0.19367261
## Full Bath        0.6711824 0.6128182 1.00000000 0.08884099
## Half Bath        0.3433749 0.1936726 0.08884099 1.00000000
```

```
nash_tib <- nash[,c(2:4,6:9)]
```

Scaling all the variables

```
library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
scaled_data_pca <- nash_tib %>%
  mutate_at(c(1:7), ~(scale(.) %>% as.vector))

str(scaled_data_pca)
```

```
## tibble [24,014 x 7] (S3: tbl_df/tbl/data.frame)
##  $ Acreage      : num [1:24014] -0.36 -0.436 -0.36 -0.145 -0.36 ...
##  $ Finished Area: num [1:24014] -0.6941 0.1593 0.2089 0.0489 -0.7956 ...
##  $ Land Value   : num [1:24014] -0.363 -0.343 -0.431 -0.431 -0.431 ...
##  $ Total Value  : num [1:24014] -0.2788 -0.1922 0.0914 -0.2917 -0.4816 ...
##  $ Bedrooms     : num [1:24014] -1.281 -0.107 1.067 -1.281 -1.281 ...
##  $ Full Bath    : num [1:24014] -0.935 0.112 0.112 -0.935 -0.935 ...
##  $ Half Bath    : num [1:24014] -0.584 1.462 -0.584 -0.584 -0.584 ...
##  - attr(*, "na.action")= 'omit' Named int [1:32622] 1 7 18 19 26 27 29 30 31 32 ...
##   ..- attr(*, "names")= chr [1:32622] "1" "7" "18" "19" ...
```

```
psych::describe(scaled_data_pca)
```

```
##               vars     n mean sd median trimmed  mad   min   max range  skew
## Acreage          1 24014    0  1  -0.23   -0.15 0.19 -0.52 59.56 60.09 20.77
## Finished Area    2 24014    0  1  -0.26   -0.17 0.62 -1.33 16.14 17.47  3.06
## Land Value       3 24014    0  1  -0.38   -0.22 0.16 -0.68 17.67 18.34  4.42
## Total Value      4 24014    0  1  -0.33   -0.20 0.33 -0.85 22.71 23.56  5.26
## Bedrooms         5 24014    0  1  -0.11   -0.06 0.00 -3.63  9.28 12.91  0.91
## Full Bath        6 24014    0  1   0.11   -0.14 1.55 -1.98  8.49 10.48  1.41
## Half Bath        7 24014    0  1  -0.58   -0.15 0.00 -0.58  5.55  6.14  1.44
##               kurtosis   se
## Acreage         858.36 0.01
## Finished Area    18.42 0.01
## Land Value       33.91 0.01
## Total Value      54.84 0.01
## Bedrooms          3.02 0.01
## Full Bath         3.52 0.01
## Half Bath         1.40 0.01
```

Visualizing the PCA

```
library(factoextra) #extract and visualize the output of multivariate data analyses, including 'PCA'
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

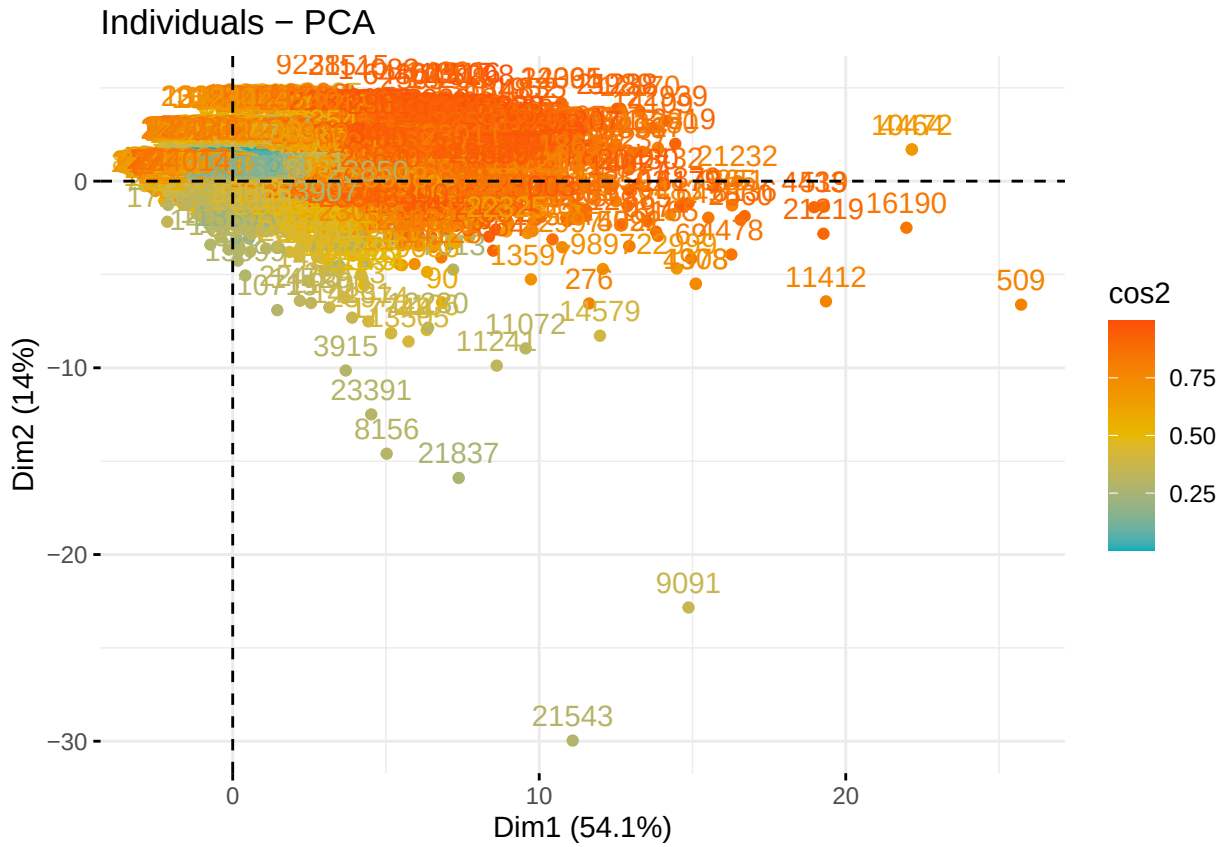
```
viz_pca <- prcomp(scaled_data_pca, center = TRUE,scale. = TRUE)

summary(viz_pca) #show the proportion of variance explained by all possible components along with cumul
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     1.9453 0.9891 0.9560 0.83524 0.58478 0.44346 0.29523
## Proportion of Variance 0.5406 0.1398 0.1306 0.09966 0.04885 0.02809 0.01245
## Cumulative Proportion  0.5406 0.6804 0.8109 0.91060 0.95946 0.98755 1.00000
```

```
viz_pca$rotation #show the loadings for each component by variable
```

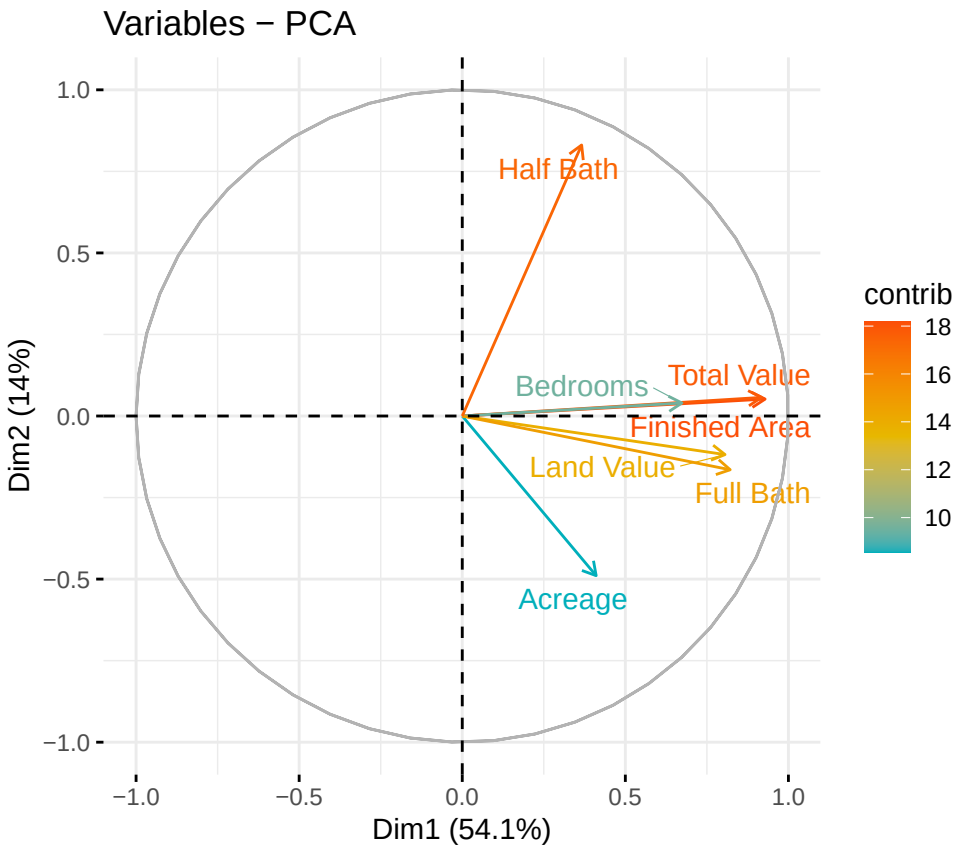
```
##                     PC1         PC2         PC3          PC4          PC5
## Acreage       0.2109916 -0.49479787  0.68293644 -0.485510517  0.030268656
## Finished Area 0.4773221  0.05246374 -0.02612534 -0.005957179  0.325965572
## Land Value    0.4141392 -0.11954678  0.17979354  0.507817027 -0.518404870
## Total Value   0.4717714  0.05766484  0.08540920  0.341416587 -0.001178923
## Bedrooms      0.3462100  0.04033799 -0.46916551 -0.581742416 -0.557644918
## Full Bath     0.4222686 -0.16653058 -0.36533133 -0.055258089  0.556252031
## Half Bath     0.1881344  0.83990855  0.37378012 -0.219824786  0.060836487
##                      PC6         PC7
## Acreage       0.05251216  0.06972267
## Finished Area -0.60740112 -0.54175959
## Land Value     0.34955675 -0.36477377
## Total Value   -0.29966210  0.74863059
## Bedrooms      -0.06193915  0.07187963
## Full Bath      0.58782864  0.04977511
## Half Bath      0.25872785 -0.02198600
```

```
#Graph of observations
fviz_pca_ind(viz_pca,
             c = "point", #point
             col.ind = "cos2", # Color by the quality of representation,
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"), #color gradient
             repel = FALSE     # Avoid overlapping numbers, which is not important, so set as false
             )
```

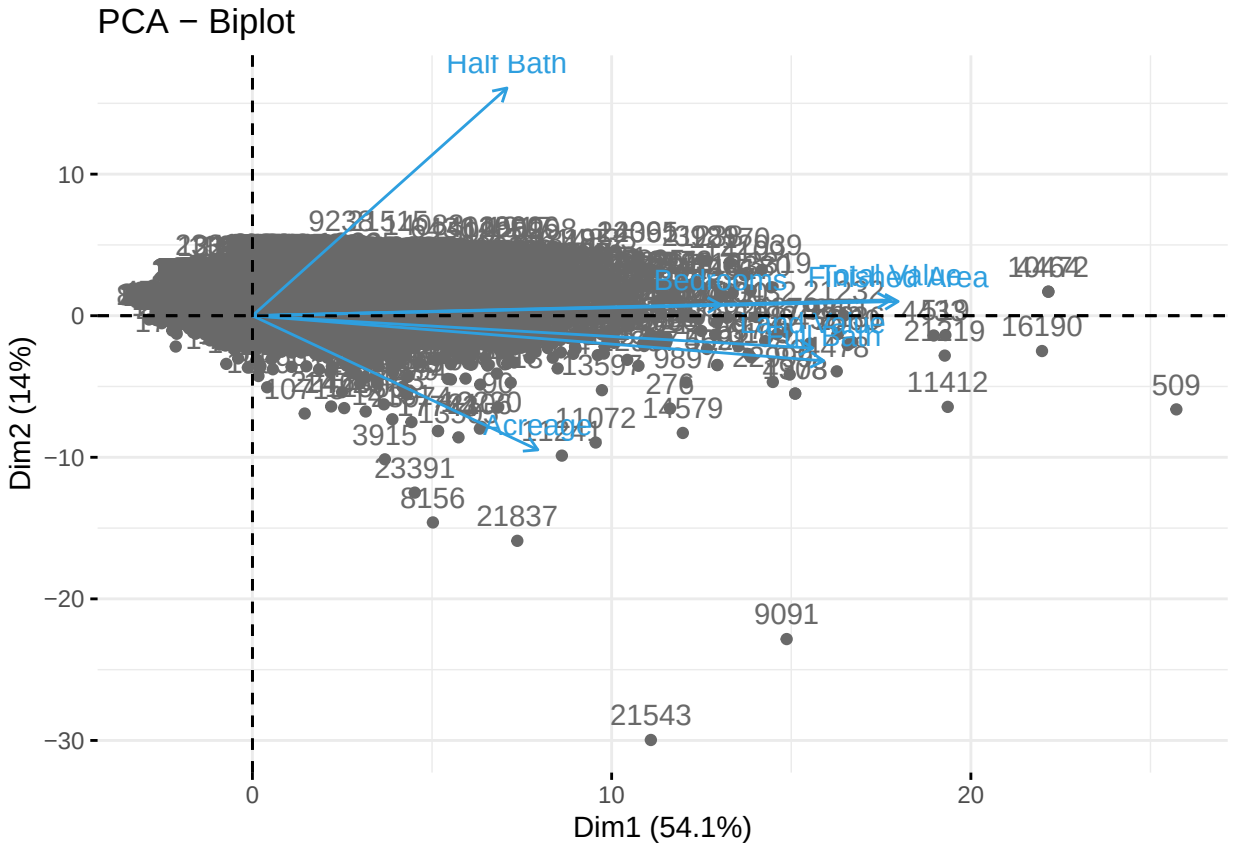


This is an interesting distribution because it shows the shape is not a circle but has many extreme outliers in one quadrant of the graph.

```
#Graph of variables. Positive correlated variables point to the same side of the plot. Negative correla
fviz_pca_var(viz_pca,
             col.var = "contrib", # Color by contributions to the PC
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE #Avoid overlapping text if possible
             )
```




```
#Biplot together.
fviz_pca_biplot(viz_pca, repel = FALSE, #Had to turn off repel or else I got the error "ggrepel: 24002 
                col.var = "#2E9FDF", # Variables color
                col.ind = "#696969"  # Individuals color
                )
```



This seems to indicate that acreage is the predictor variable that results in such extreme outliers.

Bartlett's test including sample size

```
cortest.bartlett(scaled_data_pca, 24014)
```

```
## R was not square, finding R from data
```

```
## $chisq
## [1] 102771.3
##
## $p.value
## [1] 0
##
## $df
## [1] 21
```

KMO on the data (look for variables below .5 and remove)

```
KMO(scaled_data_pca)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = scaled_data_pca)
## Overall MSA =  0.77
## MSA for each item =
##       Acreage Finished Area    Land Value   Total Value      Bedrooms
##          0.77          0.78          0.75          0.72          0.85
##     Full Bath     Half Bath
```

```
##          0.82          0.61
```

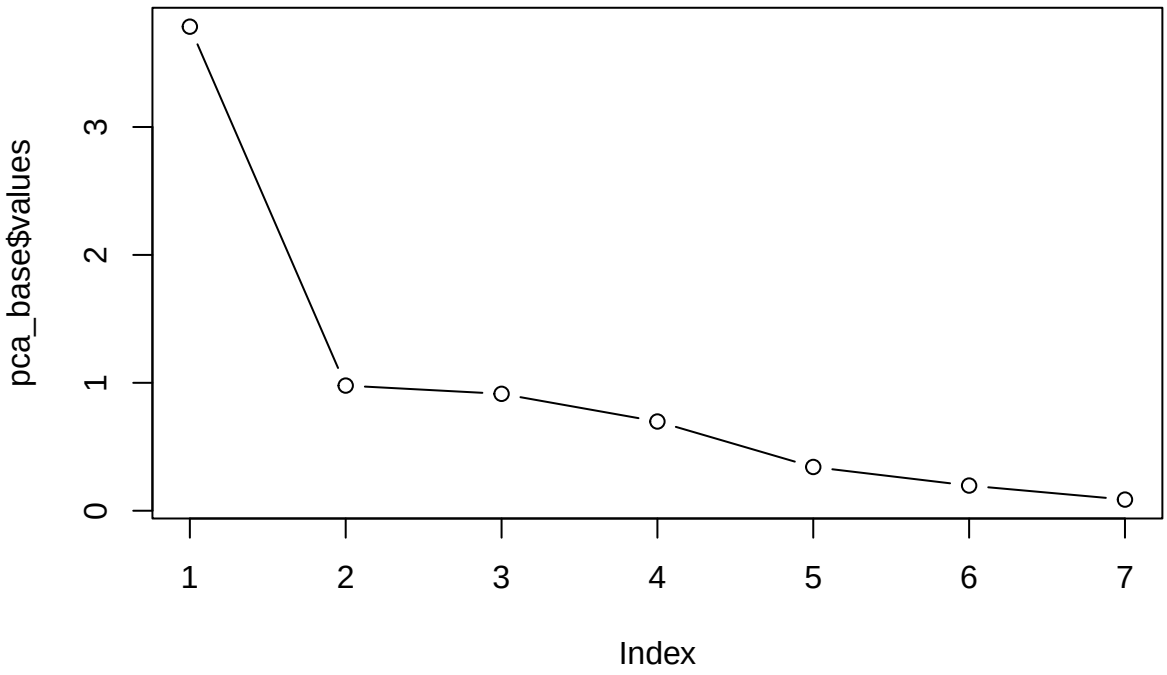
All are above .5 so nothing to remove!

Baseline PCA to check scree plot, SS loadings above 1, and normal distribution of variables

```
pca_base <- principal(scaled_data_pca, nfactors = 7, rotate = "none")

pca_base #results
```

```
## Principal Components Analysis
## Call: principal(r = scaled_data_pca, nfactors = 7, rotate = "none")
## Standardized loadings (pattern matrix) based upon correlation matrix
##                PC1   PC2   PC3   PC4   PC5   PC6   PC7 h2       u2 com
## Acreage       0.41 -0.49  0.65  0.41 -0.02  0.02  0.02  1 -8.9e-16 3.4
## Finished Area 0.93  0.05 -0.02  0.00 -0.19 -0.27 -0.16  1  0.0e+00 1.3
## Land Value    0.81 -0.12  0.17 -0.42  0.30  0.16 -0.11  1  3.3e-16 2.2
## Total Value   0.92  0.06  0.08 -0.29  0.00 -0.13  0.22  1  6.7e-16 1.4
## Bedrooms      0.67  0.04 -0.45  0.49  0.33 -0.03  0.02  1  1.6e-15 3.2
## Full Bath     0.82 -0.16 -0.35  0.05 -0.33  0.26  0.01  1  6.7e-16 2.1
## Half Bath     0.37  0.83  0.36  0.18 -0.04  0.11 -0.01  1  1.2e-15 2.0
##
##                        PC1  PC2  PC3  PC4  PC5  PC6  PC7
## SS loadings           3.78 0.98 0.91 0.70 0.34 0.20 0.09
## Proportion Var        0.54 0.14 0.13 0.10 0.05 0.03 0.01
## Cumulative Var        0.54 0.68 0.81 0.91 0.96 0.99 1.00
## Proportion Explained  0.54 0.14 0.13 0.10 0.05 0.03 0.01
## Cumulative Proportion 0.54 0.68 0.81 0.91 0.96 0.99 1.00
##
## Mean item complexity =  2.2
## Test of the hypothesis that 7 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0
##  with the empirical chi square  0  with prob <  NA
##
## Fit based upon off diagonal values = 1
```

```
#scree plot using eigen values stored in pca_1$values
plot(pca_base$values, type = "b")
```



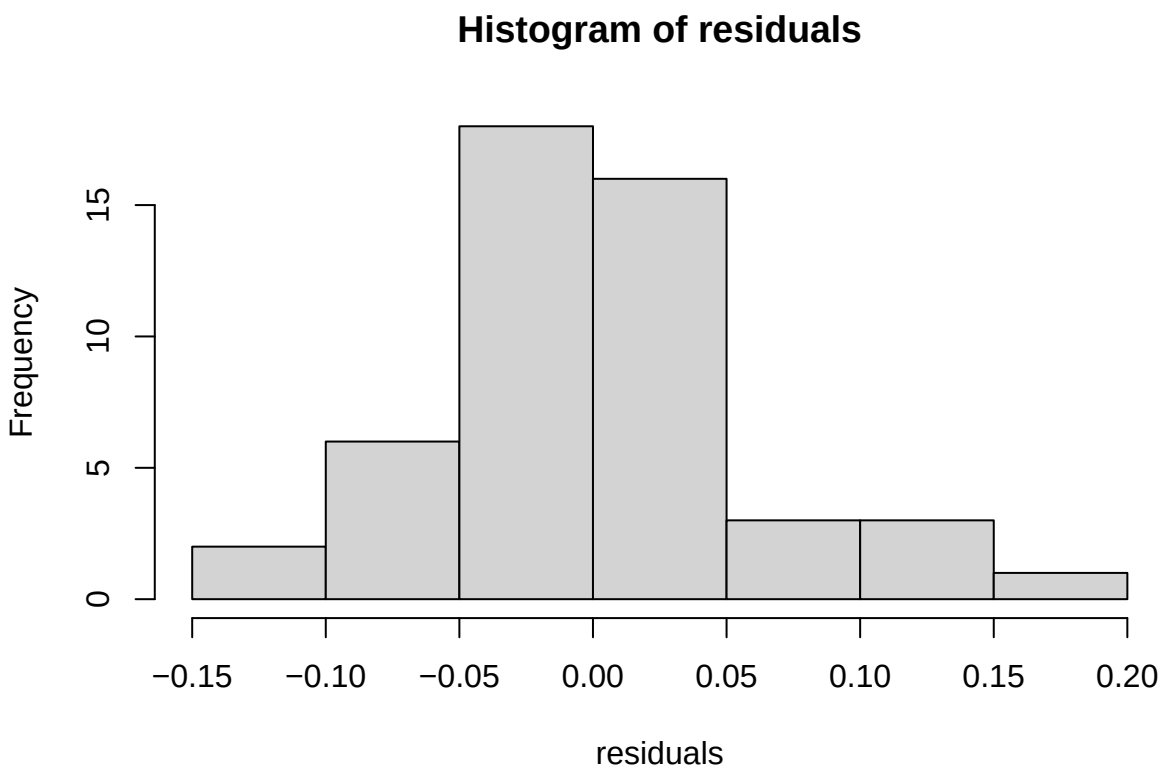
Let's pick four!

Check that residuals are normally distributed

```
pca_resid <- principal(scaled_data_pca, nfactors = 4, rotate = "none")
pca_resid #results. 4 looks good
```

```
## Principal Components Analysis
## Call: principal(r = scaled_data_pca, nfactors = 4, rotate = "none")
## Standardized loadings (pattern matrix) based upon correlation matrix
##                 PC1   PC2   PC3   PC4   h2     u2 com
## Acreage        0.41 -0.49  0.65  0.41 1.00 0.0013 3.4
## Finished Area 0.93  0.05 -0.02  0.00 0.87 0.1345 1.0
## Land Value    0.81 -0.12  0.17 -0.42 0.87 0.1275 1.7
## Total Value   0.92  0.06  0.08 -0.29 0.93 0.0665 1.2
## Bedrooms      0.67  0.04 -0.45  0.49 0.89 0.1075 2.6
## Full Bath     0.82 -0.16 -0.35  0.05 0.83 0.1740 1.4
## Half Bath     0.37  0.83  0.36  0.18 0.99 0.0145 1.9
##
##                        PC1  PC2  PC3  PC4
## SS loadings           3.78 0.98 0.91 0.70
## Proportion Var        0.54 0.14 0.13 0.10
## Cumulative Var        0.54 0.68 0.81 0.91
## Proportion Explained  0.59 0.15 0.14 0.11
## Cumulative Proportion 0.59 0.75 0.89 1.00
##
## Mean item complexity =  1.9
## Test of the hypothesis that 4 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.04
##  with the empirical chi square  1978.72  with prob <  NA
##
## Fit based upon off diagonal values = 0.99
```

```r
#residuals
#require correlation matrix for final data
corMatrix<-cor(scaled_data_pca)
#corMatrix

#next,create an object from the correlation matrix and the pca loading.
residuals<-factor.residuals(corMatrix, pca_resid$loadings)

#call a histogram to check residuals
hist(residuals) #are the residuals normally distributed? Yes!
```



PCA with selected number of components based on interpretation of scree plot and SS loadings

```r
#rotation. Since factors should be related, use oblique technique (promax), if unrelated, use varimax
pca_final <- principal(scaled_data_pca, nfactors = 4, rotate = "promax")
pca_final #results.
```

```
## Principal Components Analysis
## Call: principal(r = scaled_data_pca, nfactors = 4, rotate = "promax")
## Standardized loadings (pattern matrix) based upon correlation matrix
##                  RC1   RC4   RC2   RC3   h2     u2 com
## Acreage         0.01  0.01 -0.01  1.00 1.00 0.0013 1.0
## Finished Area   0.59  0.36  0.13  0.06 0.87 0.1345 1.8
## Land Value      1.07 -0.24 -0.07  0.02 0.87 0.1275 1.1
## Total Value     0.95  0.00  0.10 -0.04 0.93 0.0665 1.0
## Bedrooms       -0.23  1.06  0.06  0.02 0.89 0.1075 1.1
## Full Bath       0.42  0.63 -0.19 -0.05 0.83 0.1740 2.0
## Half Bath       0.04  0.01  0.98 -0.01 0.99 0.0145 1.0
```

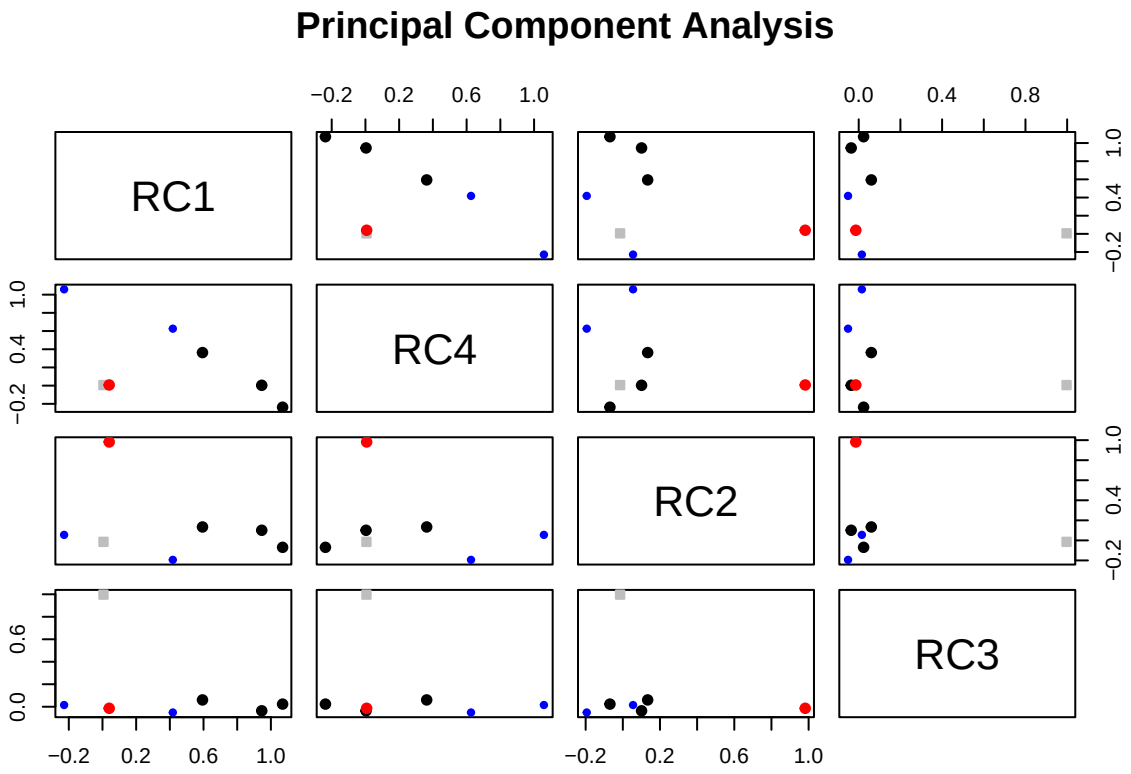
```
##
##                        RC1  RC4  RC2  RC3
## SS loadings           2.62 1.69 1.05 1.01
## Proportion Var        0.37 0.24 0.15 0.14
## Cumulative Var        0.37 0.62 0.77 0.91
## Proportion Explained  0.41 0.27 0.16 0.16
## Cumulative Proportion 0.41 0.68 0.84 1.00
##
##  With component correlations of
##      RC1  RC4  RC2  RC3
## RC1 1.00 0.63 0.24 0.33
## RC4 0.63 1.00 0.18 0.20
## RC2 0.24 0.18 1.00 0.06
## RC3 0.33 0.20 0.06 1.00
##
## Mean item complexity =  1.3
## Test of the hypothesis that 4 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.04
##  with the empirical chi square  1978.72  with prob <  NA
##
## Fit based upon off diagonal values = 0.99
```

```
#let's make the results easier to read. Include loadings over 3 and sort them

print.psych(pca_final, cut = 0.3, sort = TRUE)
```

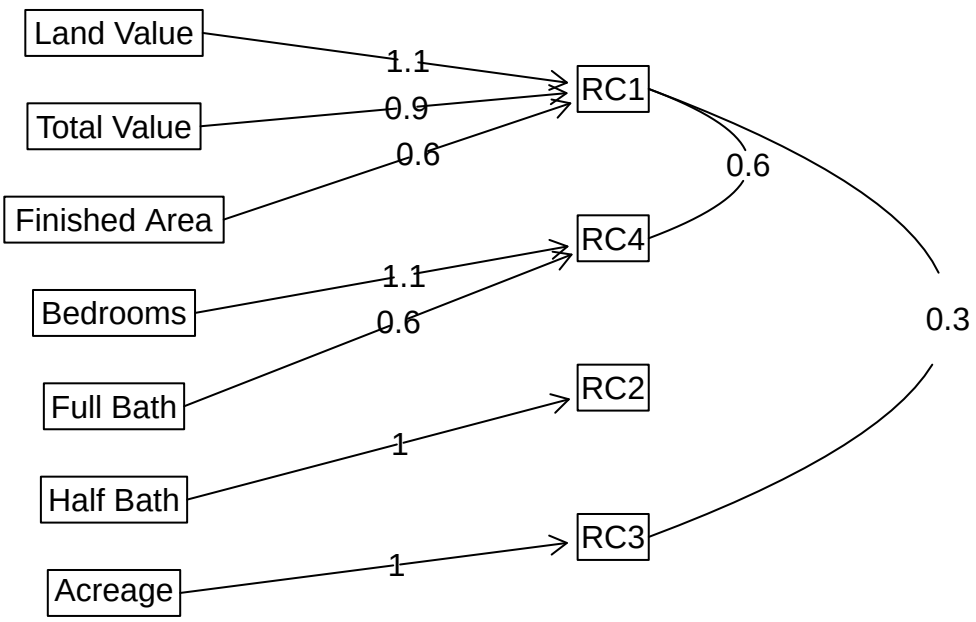
```
## Principal Components Analysis
## Call: principal(r = scaled_data_pca, nfactors = 4, rotate = "promax")
## Standardized loadings (pattern matrix) based upon correlation matrix
##               item   RC1   RC4   RC2   RC3   h2     u2 com
## Land Value       3  1.07                   0.87 0.1275 1.1
## Total Value      4  0.95                   0.93 0.0665 1.0
## Finished Area    2  0.59  0.36             0.87 0.1345 1.8
## Bedrooms         5        1.06             0.89 0.1075 1.1
## Full Bath        6  0.42  0.63             0.83 0.1740 2.0
## Half Bath        7              0.98       0.99 0.0145 1.0
## Acreage          1                    1.00 1.00 0.0013 1.0
##
##                        RC1  RC4  RC2  RC3
## SS loadings           2.62 1.69 1.05 1.01
## Proportion Var        0.37 0.24 0.15 0.14
## Cumulative Var        0.37 0.62 0.77 0.91
## Proportion Explained  0.41 0.27 0.16 0.16
## Cumulative Proportion 0.41 0.68 0.84 1.00
##
##  With component correlations of
##      RC1  RC4  RC2  RC3
## RC1 1.00 0.63 0.24 0.33
## RC4 0.63 1.00 0.18 0.20
## RC2 0.24 0.18 1.00 0.06
## RC3 0.33 0.20 0.06 1.00
##
## Mean item complexity =  1.3
## Test of the hypothesis that 4 components are sufficient.
```

```
##
## The root mean square of the residuals (RMSR) is  0.04
##  with the empirical chi square  1978.72  with prob <  NA
##
## Fit based upon off diagonal values = 0.99
```

```
plot(pca_final)
```




```
fa.diagram(pca_final)
```



Collect Factor Scores for CSV

```
#we need the pca scores
pca_final_scores <- as.data.frame(pca_final$scores)
```

Rename the variables

```
pca_final_scores<- pca_final_scores %>%
  rename(`Home Value` = RC1,
         `Half Bath` = RC2,
         `Acreage` = RC3,
         `Rooms` = RC4)

write.csv(pca_final_scores,"pca_scores_nsah.csv", row.names=FALSE)
```

Discussion

My Principle Component Analysis was already limited because I only started with 8 variables, where it would have been more robust with more input variables. There was not a high dimensional space to begin with but this did help simplify some of the variables that are measuring very similar metrics.

First, we had to check for multicollinearity that was too high to be included in the PCA and I had to remove building value since it was a .9 correlation with total value and less predictive on sales price as a whole. This did reduce the total number of numeric variables we had to work with to seven, which made it an even smaller PCA, but these still hold lots of explanatory power.

Next, I did not need to scale any variables since everything was measured in the same unit of feet or acres so they are comparable.

Next, my visualizations showed that my data set had a very large skew to the fourth quadrant and all of

my variables are pointing in the positive direction. The visualizations were already showing some early explanations of the grouping of the variables, which also made intuitive sense since they were variables measuring similar features of the house. I had to turn off the repel feature which would have made sure my labels did not overlap too much on the graph because there was just so many points, but the bi-plot showed a good visual of why the graph was so skewed, which is because all the variables were in the positive direction for dimension 1 and the variables like acreage seem to hold more weight since they have such big outliers in that direction.

My point of inflection showed that four would be a good number of factors, even if that meant two of my factors were singular variables. I even tested this threshold by rerunning my PCA with three factors and found that it performed much worse and decreased predictability, so I kept it at four factors as my point of inflection based on my scree plot. My residuals were still normally distributed and much more normally distributed with four factors compared to three.

Then I made my four factors and renamed them according to their common theme. The RC1 factor I named Home Value since it was composed of Land Value, Total Value, and Finished Area. It makes sense these would be grouped together since these are all related to the value of the house, but it is surprising that acreage was different enough to Finished Area to be its own factor and not linked. It seems like the property area of the house is its own function and then the finished area of the building is related to the value. This explains about .37 of proportional variance.

The RC2 factor is named Half Bath and the RC3 factor is named Acreage because those are the only variables in the factor somewhat due to the small amount of variables to begin with and also due to the variables predictive value. It is interesting that half bath is separate from full bath but I have seen similar results in all my regression results this semester, so I knew that was a correct assumption and it shows the added value of half baths to a house, which makes sense that full baths are almost a given based on bedroom size but half baths are considered a real addition to the home that would affect price. RC2 explains about .15 of proportional variance and RC3 explains .14 of proportional variance.

The RC4 factor is named Rooms since it contains bedrooms and full bath which shows how related bedrooms and full bathrooms are in a property. RC4 explains .24 of proportional variance.