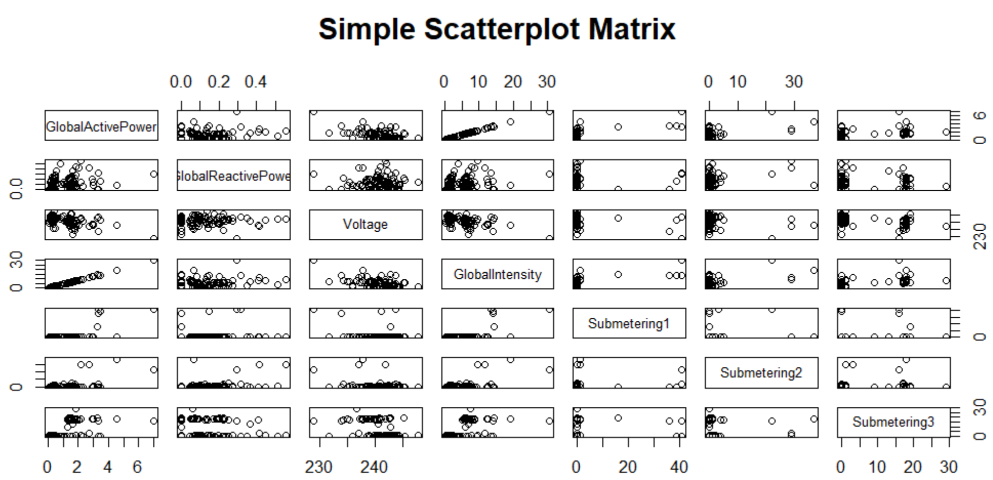
Statistics Assignment 7

- Joshua Boryer 41497475

1a)




Observations:

Visually Submetering1, Submetering2, and Submetering3 have no obvious trend or correlation within each other.

Reading the numbers from the correlation method matrix using:

```
cor(power[, c("Submetering1", "Submetering2", "Submetering3")], use = "complete.obs")
```

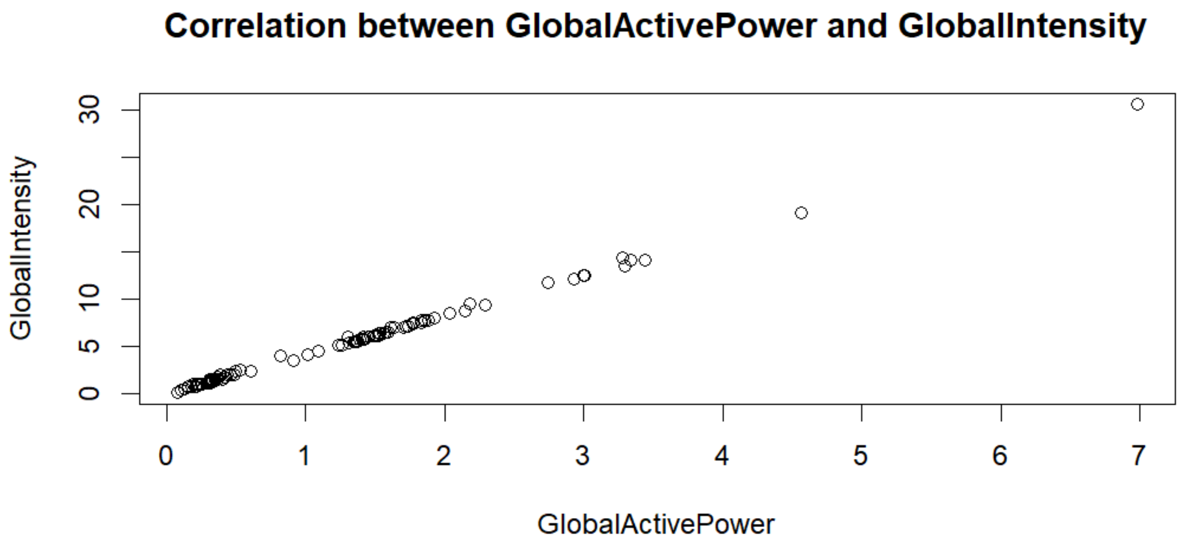
Submetering1 has a moderate correlation with Submetering3, Submetering2 has a moderate correlation with Submetering3.

```
             Submetering1 Submetering2 Submetering3
Submetering1    1.0000000   0.17906976   0.04284170
Submetering2    0.1790698   1.00000000   0.04801754
Submetering3    0.0428417   0.04801754   1.00000000
```

Visually GlobalActivePower and GlobalIntensity are strongly correlated in a positive linear trend

```
> cor(power[, c("GlobalActivePower", "GlobalIntensity")])
                  GlobalActivePower GlobalIntensity
GlobalActivePower         1.0000000       0.9990185
GlobalIntensity           0.9990185       1.0000000
```

The correlation numbers of GlobalActivePower and GlobalIntensity show a near perfect positive correlation.




The correlation scatterplot between GlobalActivePower and GlobalIntensity shows a positive linear trend.

Visually unknown correlation between Voltage and GlobalReactivePower and further digging provides evidence for this

```
> cor(power[, c("Voltage", "GlobalReactivePower")])
                        Voltage GlobalReactivePower
Voltage              1.00000000         -0.03223281
GlobalReactivePower -0.03223281          1.00000000
```

1b)

After putting in the given assignment code it only showed the variance of the first four variables, another command was needed to gather the full result.

```
# A tibble: 1 × 7
  GlobalActivePower GlobalReactivePower Voltage GlobalIntensity
              <dbl>               <dbl>   <dbl>           <dbl>
1              1.23              0.0147    8.57            22.3
# ℹ 3 more variables: Submetering1 <dbl>, Submetering2 <dbl>,
#   Submetering3 <dbl>
```

Using:

```
as.data.frame(power %>% summarise(across(everything(), var)))
```

Gave the results below.

```
  GlobalActivePower GlobalReactivePower  Voltage
1          1.229233          0.01465218 8.565778
  GlobalIntensity Submetering1 Submetering2 Submetering3
1         22.2656     61.87586     34.18545     75.98424
```

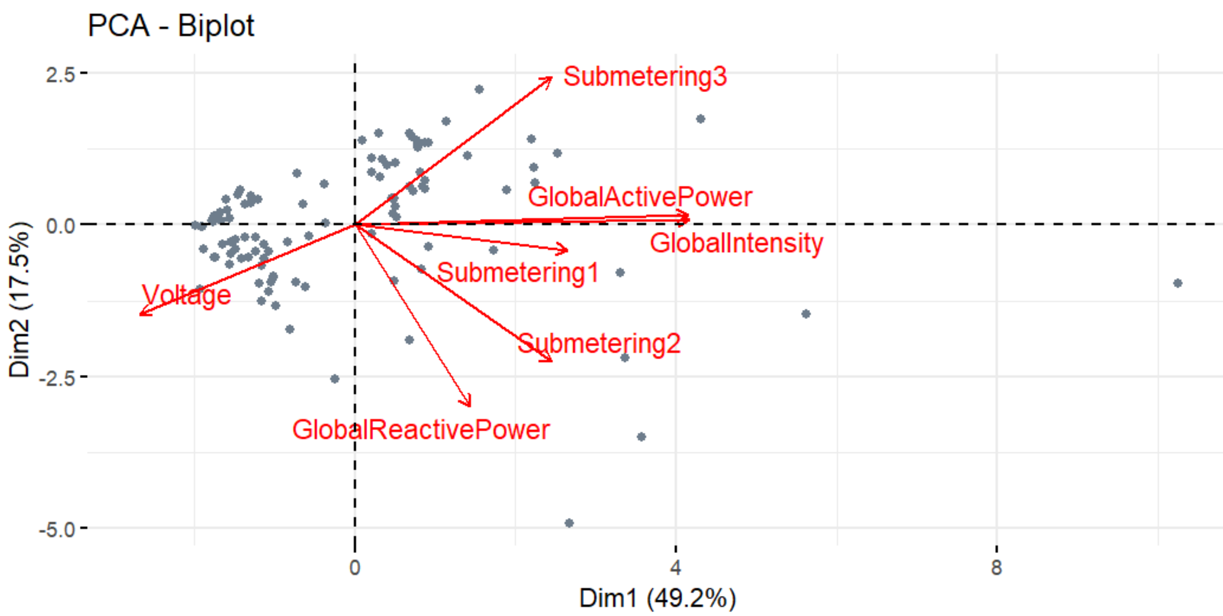
PCA is sensitive to the variance of each variable. If variables are measured on different scales, the larger variances will dominate the principal components not because they carry more information but because of their units. GlobalActivePower is measured in kilowatts while Voltage is in volts. In the data Voltage has a variance of 8.56 and GlobalActivePower has a variance of only 1.3. Without scaling the variables, the PCA will treat Voltage as more influential simply because of its larger variance.

To scale these variables the data frame scaled_power is created using:

```
scaled_power <- scale(power)
```

This dataframe is ready to be used for the PCA test which will help reduce the dimensionality of the dataset reducing the risk of overfitting.

Looking at the code given in the assignment it was unnecessary to create a new data frame as the PCA test function in R has a built-in argument for Scaling which is used for this PCA test. The PCA plot is below.




Using the rotation matrix code provided in lectures to help read this the results gave me this:

```
power_pca$rotation
```

```
                           PC1         PC2         PC3
GlobalActivePower    0.5237772  0.03297193  0.06539276
GlobalReactivePower  0.1801171 -0.63364151 -0.37444377
Voltage             -0.3386465 -0.31095910  0.14892904
GlobalIntensity      0.5254762  0.01752332  0.06567994
Submetering1         0.3334862 -0.08840849  0.76293694
Submetering2         0.3095976 -0.47610234 -0.21756703
Submetering3         0.3087798  0.51567350 -0.44678154
                            PC4        PC5         PC6
GlobalActivePower   -0.10121910 -0.1651036  0.43420748
GlobalReactivePower -0.46290084  0.4554943 -0.06325285
Voltage             -0.52225199 -0.7015634  0.03746501
GlobalIntensity     -0.09072406 -0.1468357  0.42899990
Submetering1        -0.16419306  0.1354486 -0.50356449
Submetering2         0.57270983 -0.4293406 -0.34319140
Submetering3        -0.37355534 -0.2209845 -0.50066043
                             PC7
GlobalActivePower    0.703031649
GlobalReactivePower  0.009745337
Voltage             -0.010057531
GlobalIntensity     -0.710929946
Submetering1         0.003907453
Submetering2         0.005497533
Submetering3        -0.009137664
```

The variable with the lowest positive loading on the first principal component (PC1) is GlobalReactivePower with a value of 0.18.

The two variables with the highest quality representation by the first two principal components

| | PC1 | PC2 |
|---|---|---|
| GlobalActivePower | 0.5237772 | 0.03297193 |
| GlobalReactivePower | 0.1801171 | -0.63364151 |
| Voltage | -0.3386465 | -0.31095910 |
| GlobalIntensity | 0.5254762 | 0.01752332 |
| Submetering1 | 0.3334862 | -0.08840849 |
| Submetering2 | 0.3095976 | -0.47610234 |
| Submetering3 | 0.3087798 | 0.51567350 |

Squaring all the values of PC1 and PC2 gives the amount of variance explained per variable and adding each square by its appropriate row counterpart shows how much of the variable is captured in 2 dimensions

| | | |
|---|---|---|
| GlobalActivePower | GlobalReactivePower | Voltage |
| 0.2754297 | 0.4339437 | 0.2113770 |
| GlobalIntensity | Submetering1 | Submetering2 |
| 0.2764323 | 0.1190291 | 0.3225241 |
| Submetering3 | | |
| 0.3612641 | | |

Looking at these findings, the two variables with the highest quality representation by PC1 and PC2 are the variables with the highest value which happen to be GlobalReactivePower which has a sum square of 0.43 and Submetering3 with a value of 0.36

The measurements at Submetering 2 and 3 have a weak correlation seen visually by the angle in the PCA test plot. The vectors difference is around 75 - 80 degrees

1c)

**PCA summary**

```
Importance of components:
                          PC1    PC2    PC3     PC4     PC5
Standard deviation     1.8563 1.1067 0.9653 0.83382 0.74874
Proportion of Variance 0.4922 0.1750 0.1331 0.09932 0.08009
Cumulative Proportion  0.4922 0.6672 0.8003 0.89963 0.97972
                           PC6    PC7
Standard deviation     0.37584 0.0264
Proportion of Variance 0.02018 0.0001
Cumulative Proportion  0.99990 1.0000
```

This summary data of the PCA shows that:

The standard deviation for PC1, the first principle component, is 1.86. Variance is standard deviation squared = (1.8563)^2 = 3.445 = 3.45

The proportion of variance that the second principle component explains is the value 0.175 so 17.5% of the variance is explained by PC2.

6 components are needed to explain 99% of the variance in this dataset as seen by the cumulative proportion under PC6 (0.9999) meaning it's safe to remove PC7 which accounts for 0.00001% of the variance in the dataset.

2a)

The summary for the population dataset outputs the following image

```
               Country       Density           MedianAge    
 Afghanistan        :  1   Min.   :   2.11   Min.   :15.20  
 Albania            :  1   1st Qu.:  31.40   1st Qu.:21.65  
 Algeria            :  1   Median :  82.70   Median :29.00  
 Angola             :  1   Mean   : 151.07   Mean   :29.92  
 Antigua and Barbuda:  1   3rd Qu.: 158.50   3rd Qu.:38.10  
 Argentina          :  1   Max.   :1800.00   Max.   :48.40  
 (Other)            :173                                    
     Over60          Urban            Growth      
 Min.   : 3.23   Min.   : 13.30   Min.   :-1.720  
 1st Qu.: 5.65   1st Qu.: 40.75   1st Qu.: 0.458  
 Median :10.10   Median : 58.40   Median : 1.200  
 Mean   :13.00   Mean   : 57.97   Mean   : 1.269  
 3rd Qu.:20.45   3rd Qu.: 76.40   3rd Qu.: 2.085  
 Max.   :34.30   Max.   :100.00   Max.   : 4.120  

     Under5            Cluster  
 Min.   : 3.700   Cluster 1: 7  
 1st Qu.: 5.950   Cluster 2:62  
 Median : 8.890   Cluster 3:57  
 Mean   : 9.701   Cluster 4:53  
 3rd Qu.:13.445                 
 Max.   :19.790                 
```

This data shows all of the variables, giving the range of their units, mean average, and median. From this data it looks like everything has loaded correctly.

Before performing a PCA test the variables Country and Cluster must be removed as cluster which is a constant and country which is Identifier variable both don't contribute to any of the variance in the data.

After removing these variables and saving that to a data frame, the new data frame summary now looks like this:

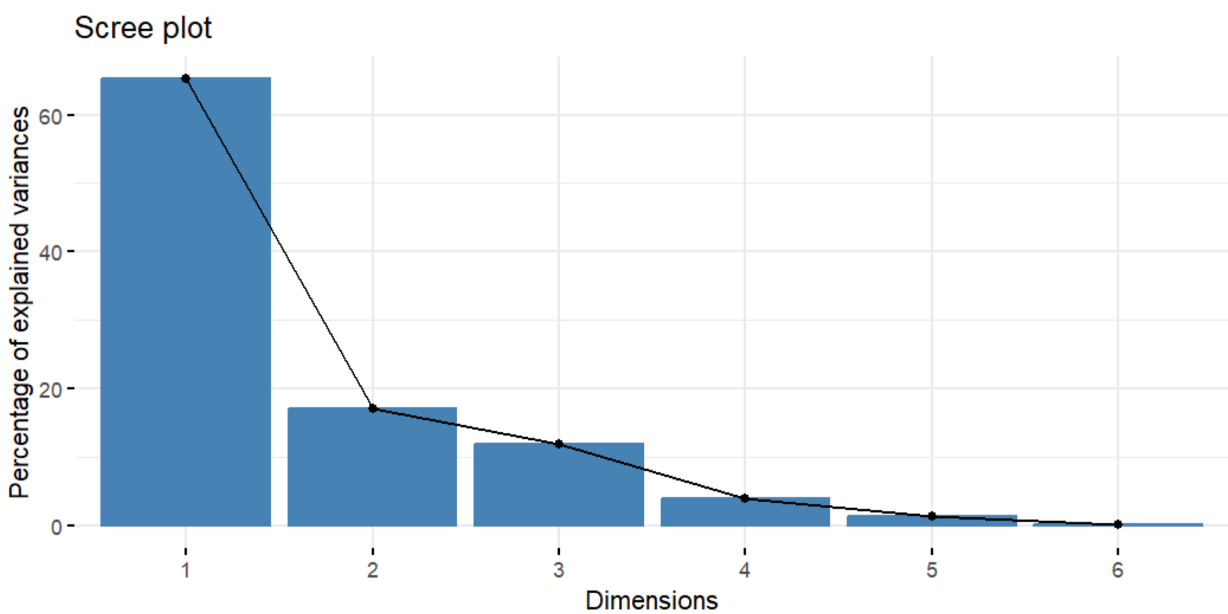
```
    Density           MedianAge          Over60     
 Min.   :   2.11   Min.   :15.20   Min.   : 3.23  
 1st Qu.:  31.40   1st Qu.:21.65   1st Qu.: 5.65  
 Median :  82.70   Median :29.00   Median :10.10  
 Mean   : 151.07   Mean   :29.92   Mean   :13.00  
 3rd Qu.: 158.50   3rd Qu.:38.10   3rd Qu.:20.45  
 Max.   :1800.00   Max.   :48.40   Max.   :34.30  
     Urban           Growth           Under5      
 Min.   : 13.30   Min.   :-1.720   Min.   : 3.700  
 1st Qu.: 40.75   1st Qu.: 0.458   1st Qu.: 5.950  
 Median : 58.40   Median : 1.200   Median : 8.890  
 Mean   : 57.97   Mean   : 1.269   Mean   : 9.701  
 3rd Qu.: 76.40   3rd Qu.: 2.085   3rd Qu.:13.445  
 Max.   :100.00   Max.   : 4.120   Max.   :19.790  
```

2b)

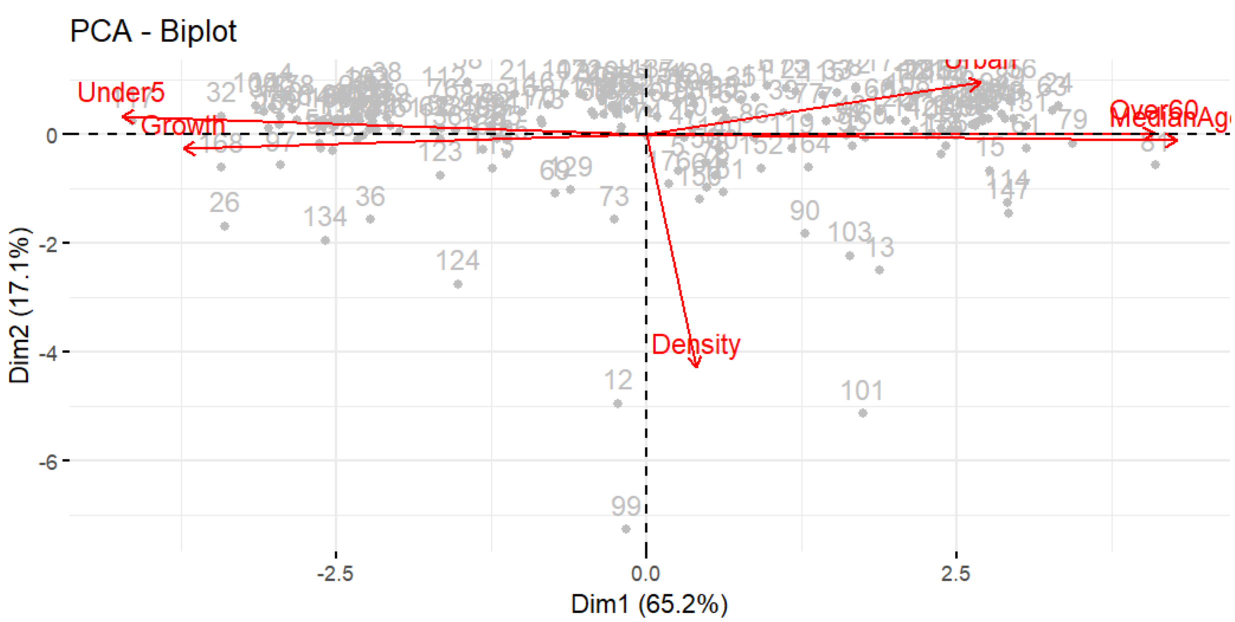
**PCA test summary with scaling**

```
Importance of components:
                          PC1    PC2    PC3     PC4     PC5
Standard deviation     1.9781 1.0121 0.8479 0.49163 0.29826
Proportion of Variance 0.6521 0.1707 0.1198 0.04028 0.01483
Cumulative Proportion  0.6521 0.8229 0.9427 0.98300 0.99782
                           PC6
Standard deviation     0.11427
Proportion of Variance 0.00218
Cumulative Proportion  1.00000
```

Together PC1 and PC2 explain 82% of the variance therefore choosing 2 components with the elbow rule is the most optimal approach as seen by the pca test summary cumulative proportion under PC2 and the “elbow” line in the scree plot which shows where the variance starts to level off.




2c)



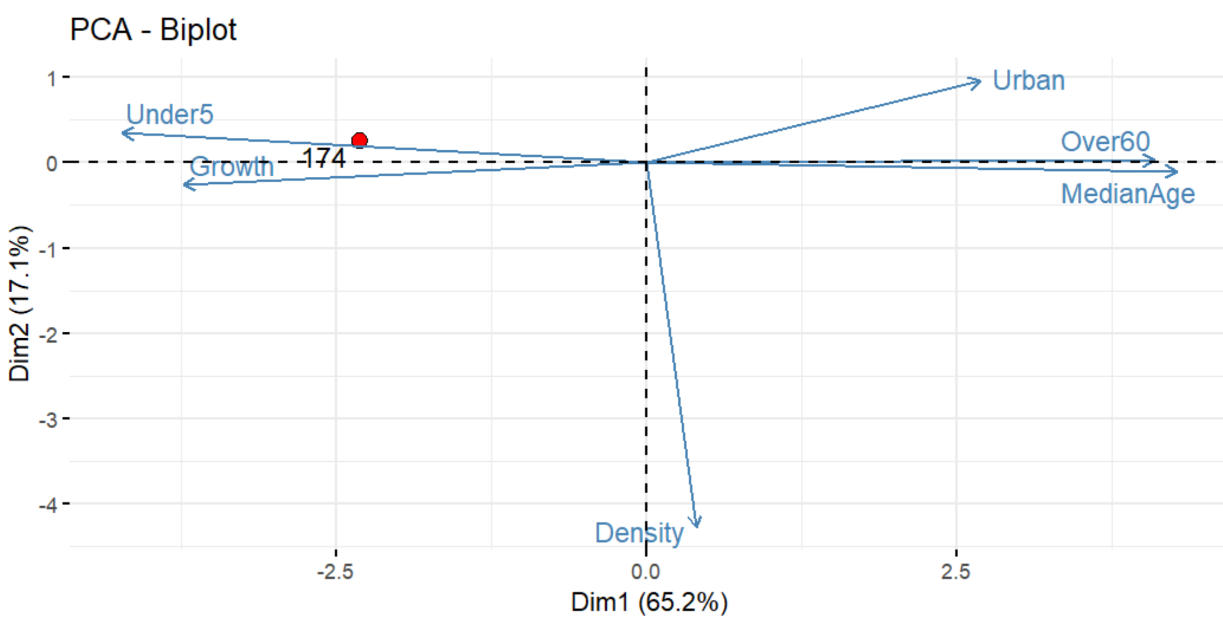
**Findings**

Because the growth vector and Over60 vector point in obvious opposite directions they're negatively correlated. This suggests that countries with higher population growth tend to have lower proportions of population over 60.

The vector that has the longest magnitude on the vertical axis is density so the variable with the highest absolute loading on the second principal component is Density. This means that PC2 primarily reflects differences in population density between countries separating countries based on how densely populated they are

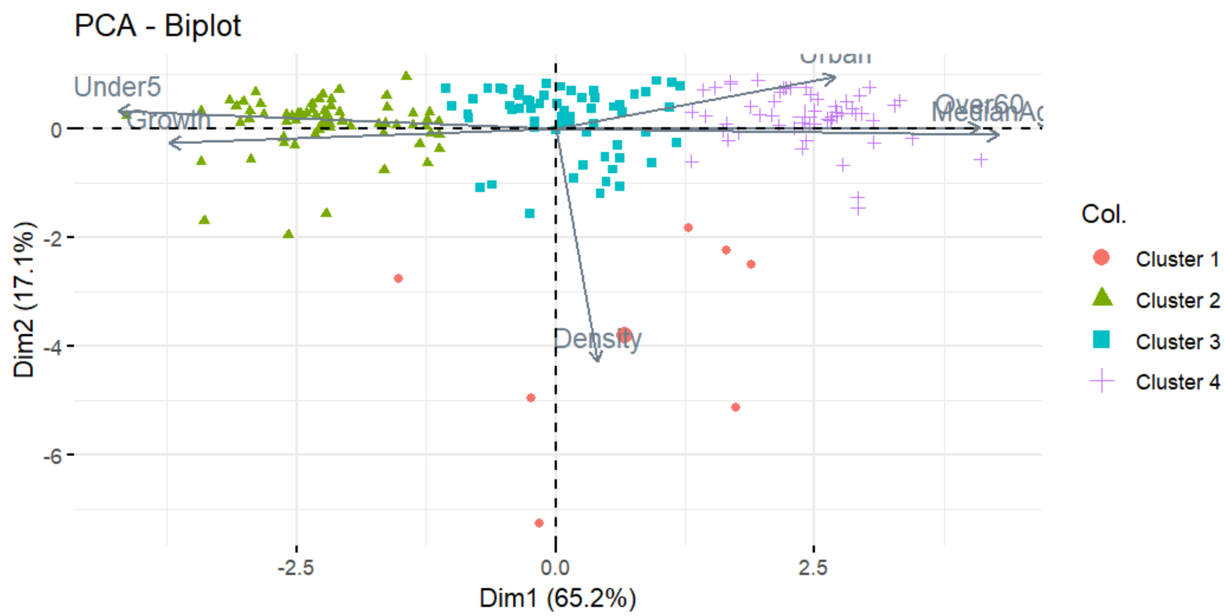
PC1 on the horizontal axis has high values for the variables Over60, Median age, and Urban in the rightward direction as well as high values for age structure and Under 5 and growth in the leftward direction. This means PC1 explains the ages and growth of a country. PC2 on the vertical axis explains density.

Because it's impossible to find the point Rwanda on the PCA biplot within all the other points, Isolating the point with country “Rwanda” is identified by the number 174 which has been isolated in this biplot.




This shows that Rwanda has a low proportion of Urban population and is relatively rural.

2d)




A new biplot above is made using the cluster variables. Cluster 2 is top leftward cluster 4 is top rightward. Cluster 2 countries have higher growth and a larger population ratio of under 5's in comparison to cluster 4 countries which have much higher urban population, less growth, and a higher age population seen by the over60 and median age variable vectors.