Capstone Preliminary Results

Author: John R Petz, 03-28-2018

Contents

[Overview 1](#_Toc510001719)

[Q & A 1](#_Toc510001720)

[Charts 4](#_Toc510001721)

[Figure 1 – CO2 emissions by Region 4](#_Toc510001722)

[Figure 2 Highest CO2 Emissions by Country <most recent year available> 5](#_Toc510001723)

[Figure 3 Correlation between GDP and CO2 emissions 5](#_Toc510001724)

[Figure 4 - Full Correlation Matrix 6](#_Toc510001725)

[Figure 5 – Plot of K values per Avg Within Distance 7](#_Toc510001726)

[Figure 6 Cluster Analysis (K=3) for GDP vs Country 8](#_Toc510001727)

[Figures 7A thru 7E - Correlation of GDP to Various Attributes of Significance 9](#_Toc510001728)

[7A - GDP vs Infant Mortality Rate 9](#_Toc510001729)

[7B - GDP vs Internet Usage 10](#_Toc510001730)

[7C - GDP Vs Urban Population 11](#_Toc510001731)

[7D – GDP vs Health Care Expenditure % GDP 11](#_Toc510001732)

[Figure 7E - Miscellaneous Cluster Charts (k=4) 13](#_Toc510001733)

[Figure 8 – Correlation Matrix of GDP vs Other Attributes 13](#_Toc510001734)

# Overview

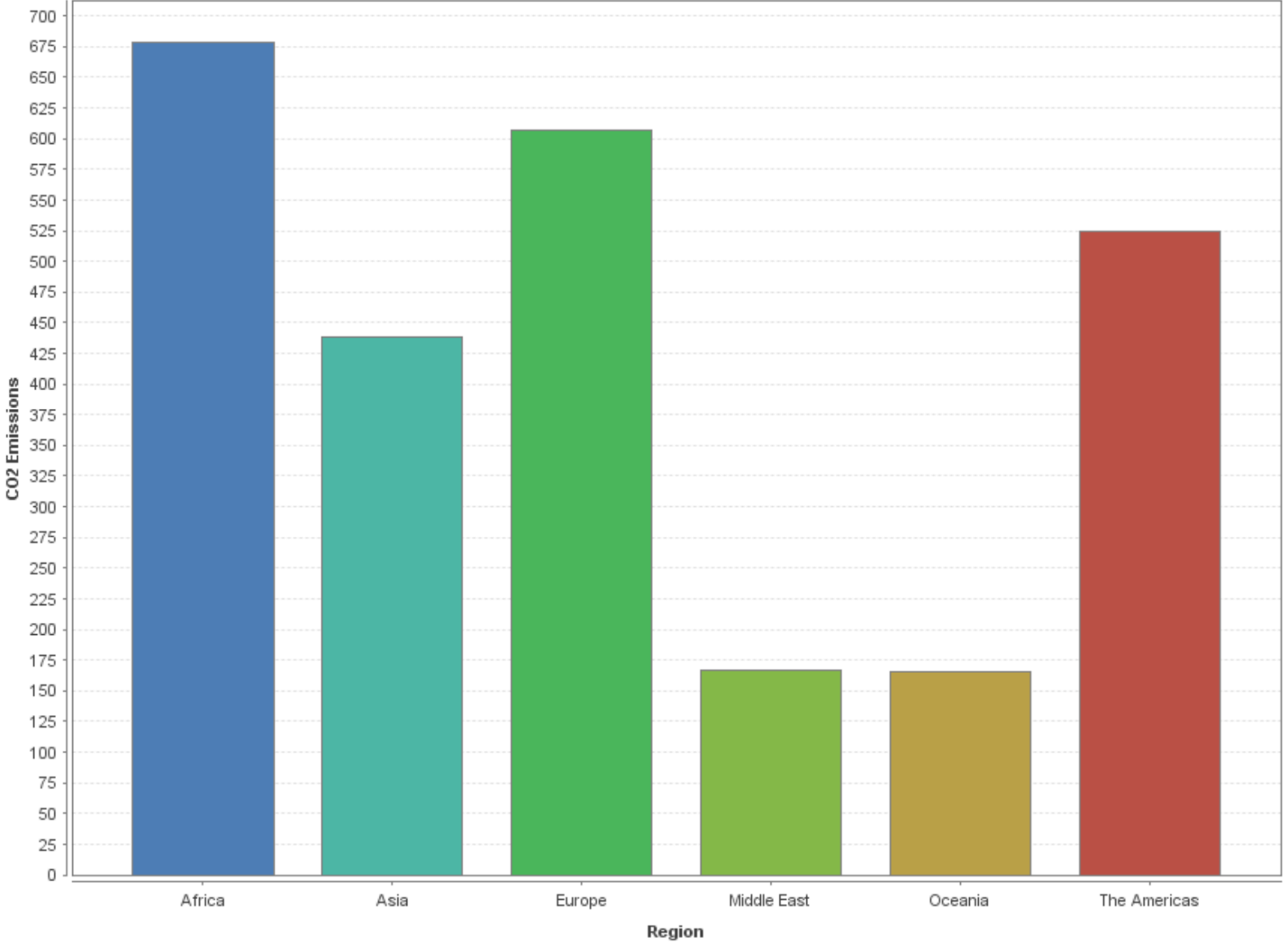
The questions and submitted answers below pertain to an Excel data file “World Indicators” which contains information on world countries pertaining to attributes such as GDP, CO2 emissions, population, birth rate, and a host of other attributes, some of which are highly or moderately correlated.

# Q & A

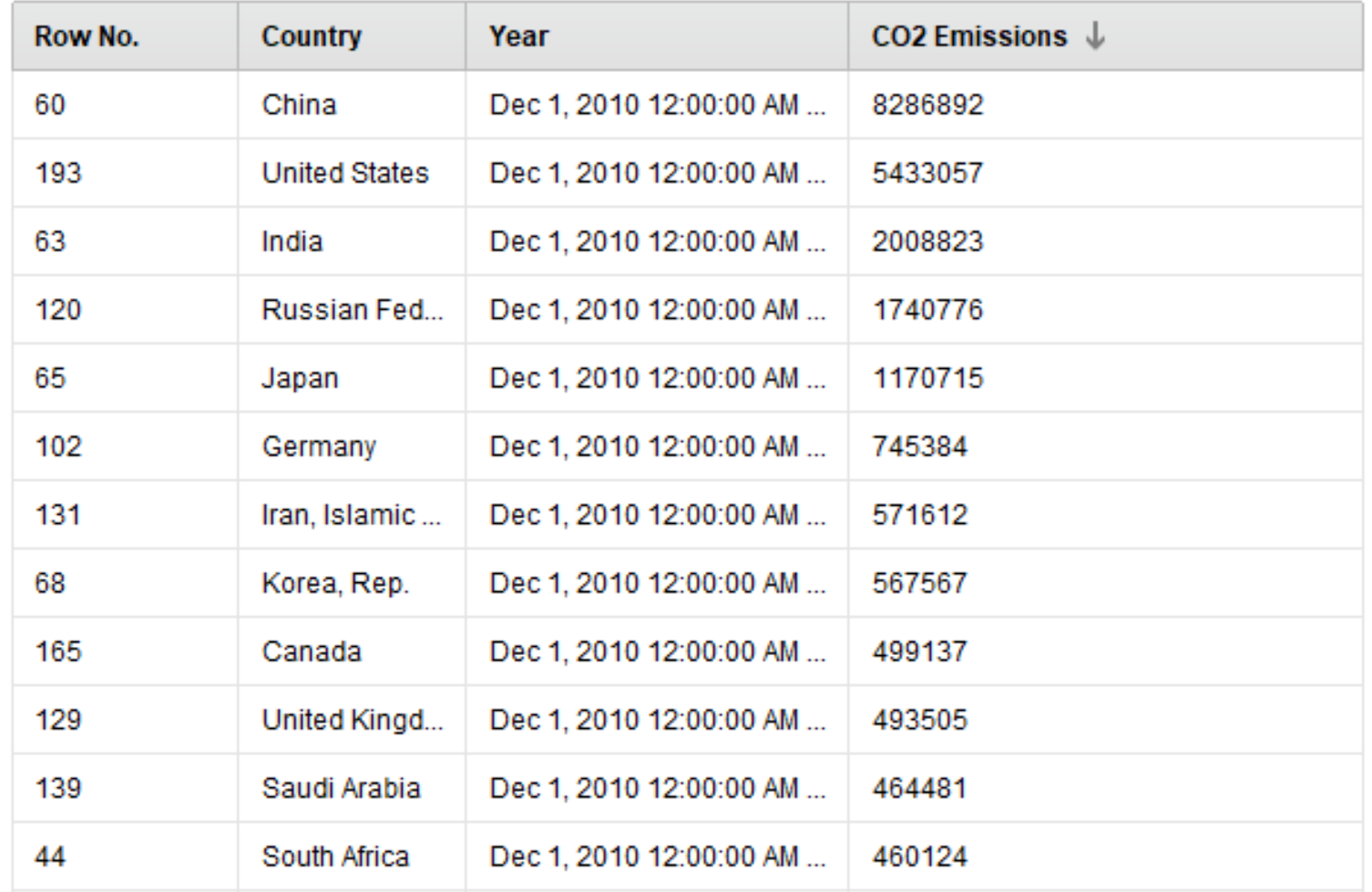
1. Are there any missing data or inconsistencies in this dataset?
   1. Missing data summary (2704 total records in data set)
      1. Birth rate – 119 missing
      2. Biz tax rate – 1281
      3. CO2 emissions – 579
      4. Days to start biz – 986
      5. Ease of biz – 2519
      6. Energy usage – 919
      7. GDP – 210 missing
      8. Health exp % GDP – 309
      9. Health exp/Capita – 309
      10. Hours to do tax – 1288
      11. Infant mortality rate – 260
      12. Internet usage – 173
      13. Lending interest – 824
      14. Life expect female – 136
      15. Life expect male – 136
      16. Mobile phone usage – 167
      17. Population 0-14 – 220
      18. Population 15-64 – 220
      19. Population 65+ - 220
      20. Population urban – 26
      21. Tourism inbound – 368
      22. Tourism outbound – 471
2. Explain your strategies for dealing with missing data for each variable with a missing value. You need to explain why you chose your strategy versus other possible strategies.
   1. Generally use Average replacement with the following exceptions
      1. Since the analysis goal is heavily dependent on GDP and CO2 emissions, remove records with missing values for these 2 attributes
      2. Remove Ease of Business since ~92% missing values.
3. Which region has the most amount of overall CO2 emission?
   1. Plot CO2 emissions by Region – see Figure 1
   2. Per the Plot Africa has the most overall CO2 emission
4. What are the top 3 countries producing CO2?
   1. Some additional context is necessary here since CO2 values change year to year. Assuming the question pertains to the latest year recorded, CO2 emissions by Countries for most recent year available – see Figure 2
   2. Per Figure 2, the top 3 countries are China, US, and India for year 2010
5. Is there any relationship between CO2 emission and the GDP?
   1. Run correlation matric between the 2 attributes – see Figure 3
   2. Per Figure 3 the correlation coeff is 0.643. Per the literature [+/-] 0.70 or greater is considered highly correlated so there is moderate degree of correlation.
6. What are the highly correlated attributes in this dataset?
   1. Run correlation matrix across entire data set – see Figure 4
      1. NOTE: use Nominal to Numerical conversion where necessary
   2. Per Figure 4 a number of attributes are highly correlated with GDP in addition to CO2:
      1. Tourism Inbound/Outbound
      2. Energy Usage [highest at 0.959]
   3. Birth rate is highly correlated with:
      1. Infant Mortality Rate
      2. Life Expected Male/Female
      3. Population ranges 0-14 and 15-64
7. What are the moderately correlated attributes in this dataset?
   1. NOTE: define Moderate correlation as 0.3 < |corr| < 0.7
   2. Per Figure 4 the following attribute correlations appear significant:
      1. Energy Usage vs Tourism
      2. CO2 Emissions vs Tourism
      3. Mobile Phone Usage vs Population
      4. Population vs Life Expectancy
8. We want to group countries into specific categories based on their GDP? How many categories do we need and why?
   1. Based on distribution run K-cluster analysis and look for optimal k
      1. NOTE: need to convert Country (nominal) to numerical – use unique index
   2. Per Figure 5 the optimal value of K appears to be 3. Figure 6 shows the 3 clusters as country categories.
9. Excluding categorical attributes, run a k-means clustering analysis with k=4, can you identify any pattern in the cluster formations?
   1. NOTE: will remove all attributes requiring Nominal to Numerical conversion for k-means
   2. Per Figures 7A-7E there are clearly delineated patterns for numerous attributes relative to GDP. There are similar patterns relative to CO2 emissions such as infant mortality rate. Finally, Figure 7F depicts numerous patterns for various attributes.
10. What are the most important factors impacting GDP?
    1. Per Figure 8 “Tourism Inbound” has the highest correlation with GDP. Other factors are:
       1. Tourism Outbound
       2. Energy Usage
11. What do you think would be the best modeling technique (e.g. regression, decision trees, etc.) in predicting GDP and why? What would be the second best?
    1. Since GDP is impacted by numerous factors, many numeric, Linear Regression is perhaps the best modeling choice. A 2nd choice might be Decision Tree or ANN.
12. Build the models according to your answers to the above questions and calculate and report model performances. Did the models perform the way you theorized they would? Are there any other models that can outperform your suggested models?
    1. TBD <ran out of time>

# Charts

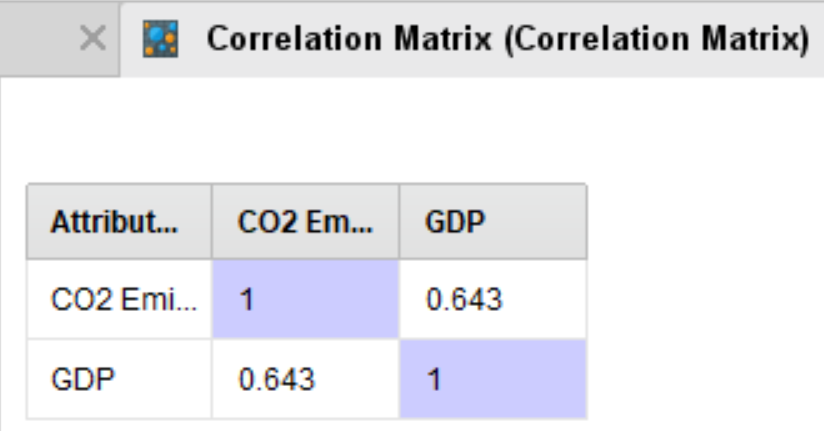
## Figure 1 – CO2 emissions by Region



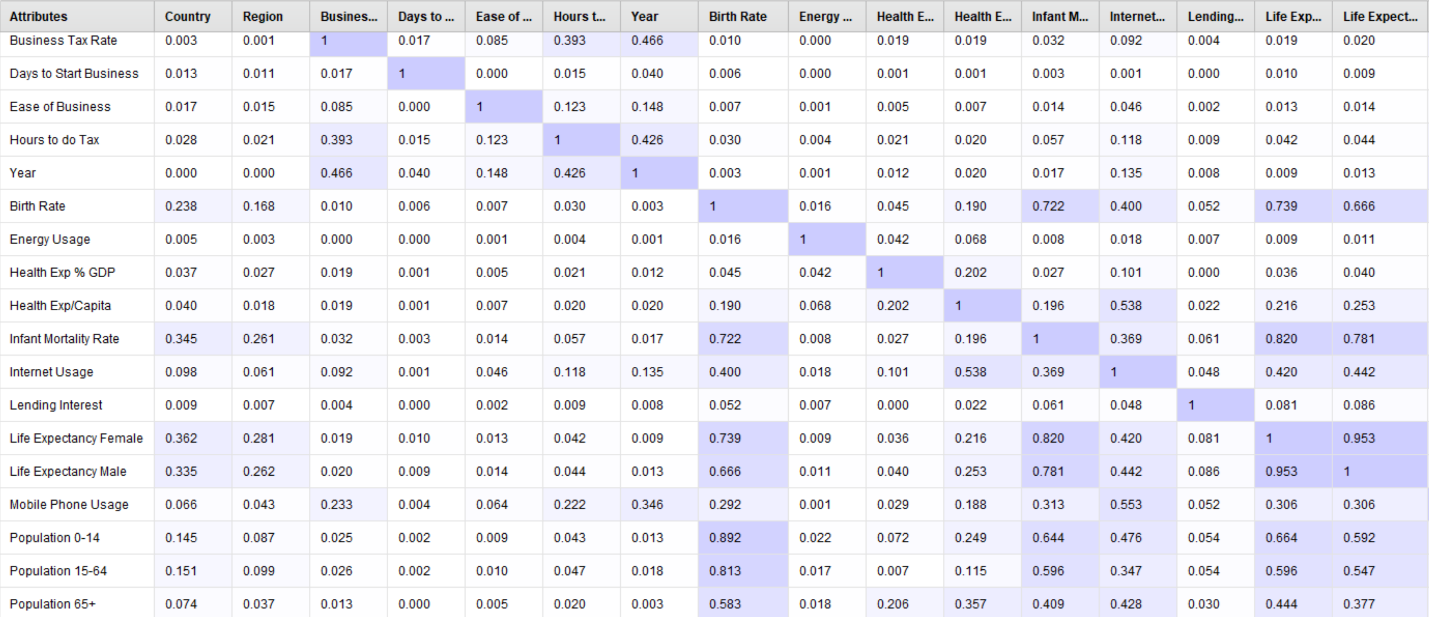
## Figure 2 Highest CO2 Emissions by Country <most recent year available>



## Figure 3 Correlation between GDP and CO2 emissions

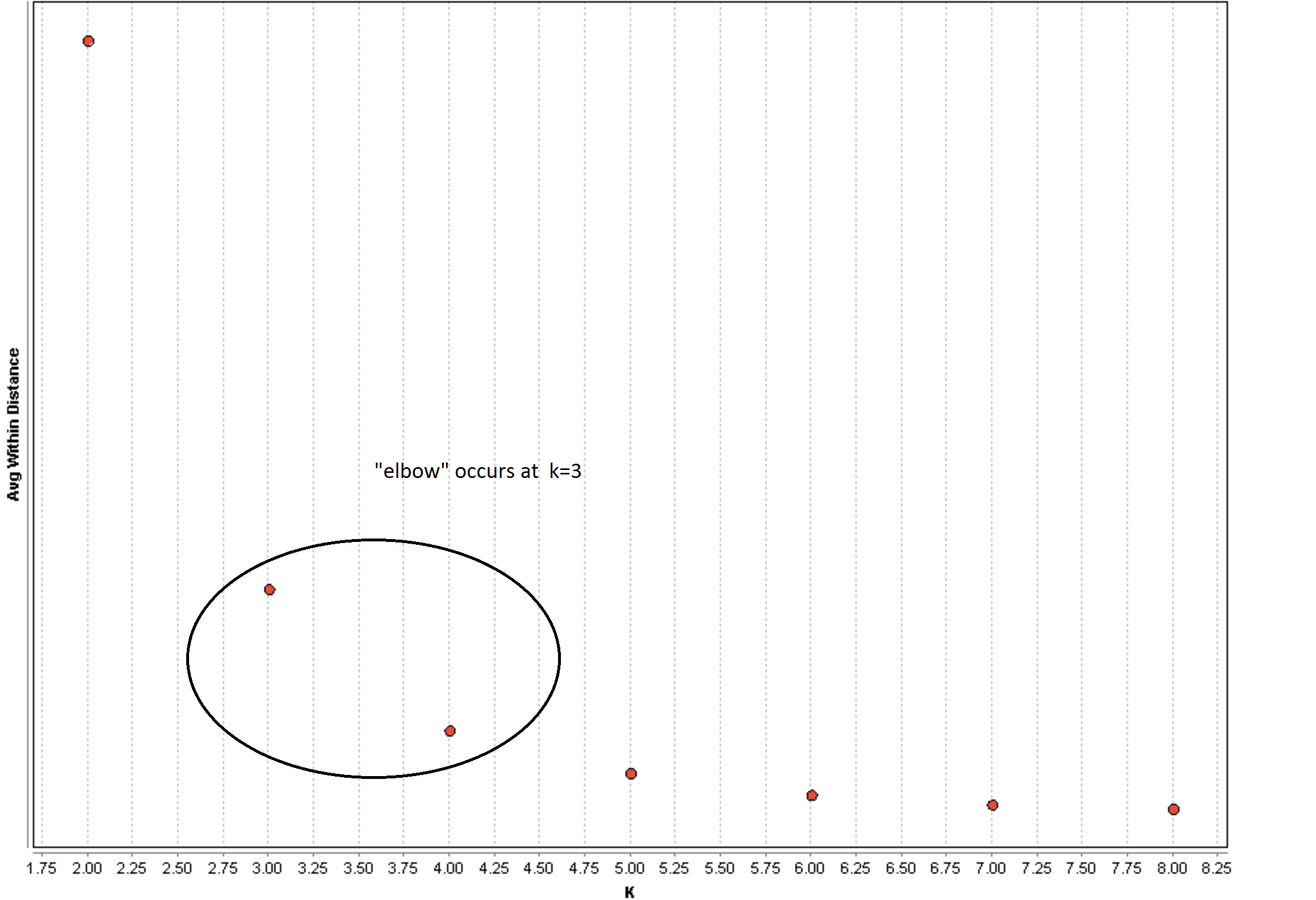


## Figure 4 - Full Correlation Matrix

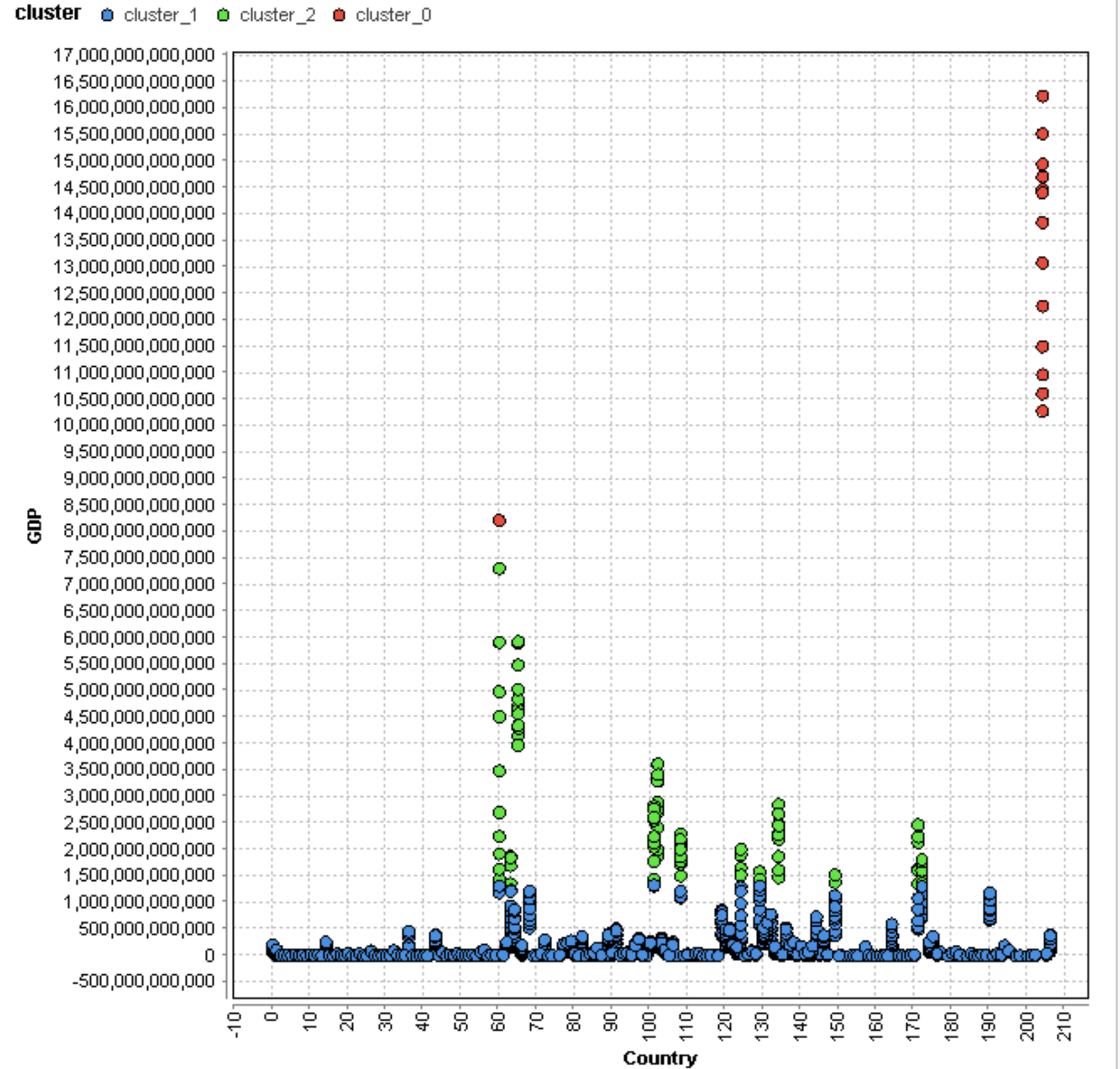




## Figure 5 – Plot of K values per Avg Within Distance

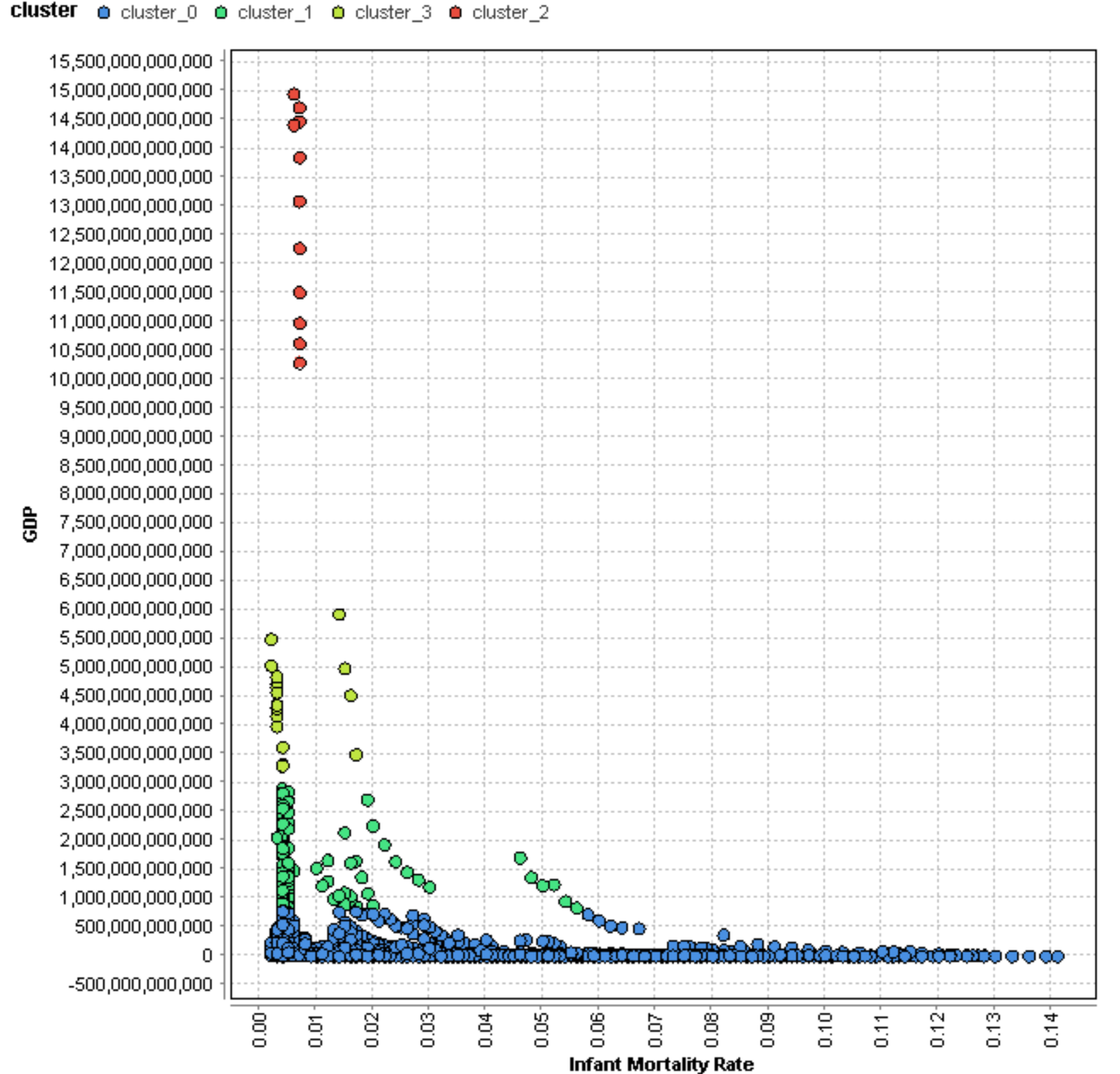


## Figure 6 Cluster Analysis (K=3) for GDP vs Country

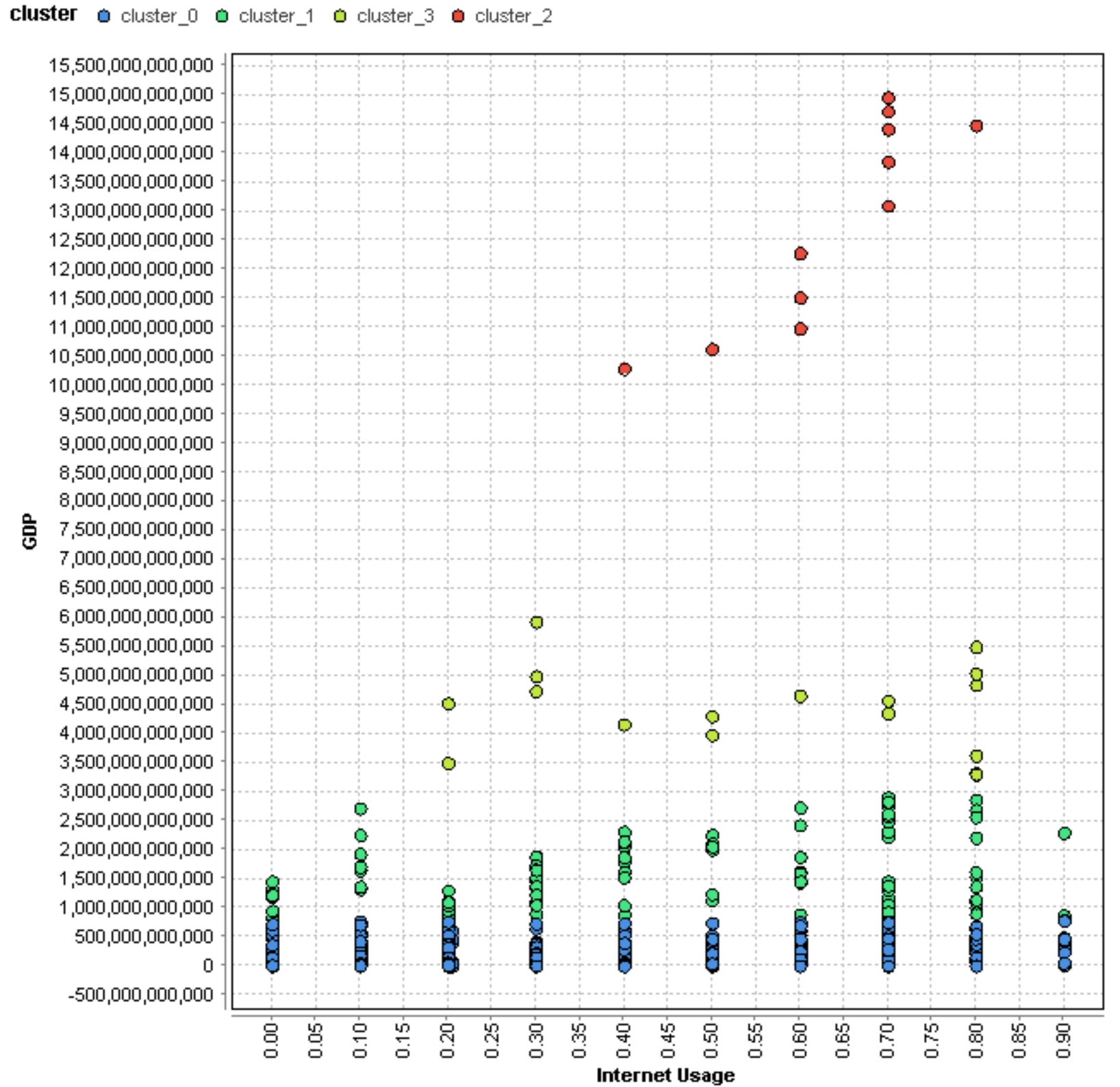


## Figures 7A thru 7E - Correlation of GDP to Various Attributes of Significance

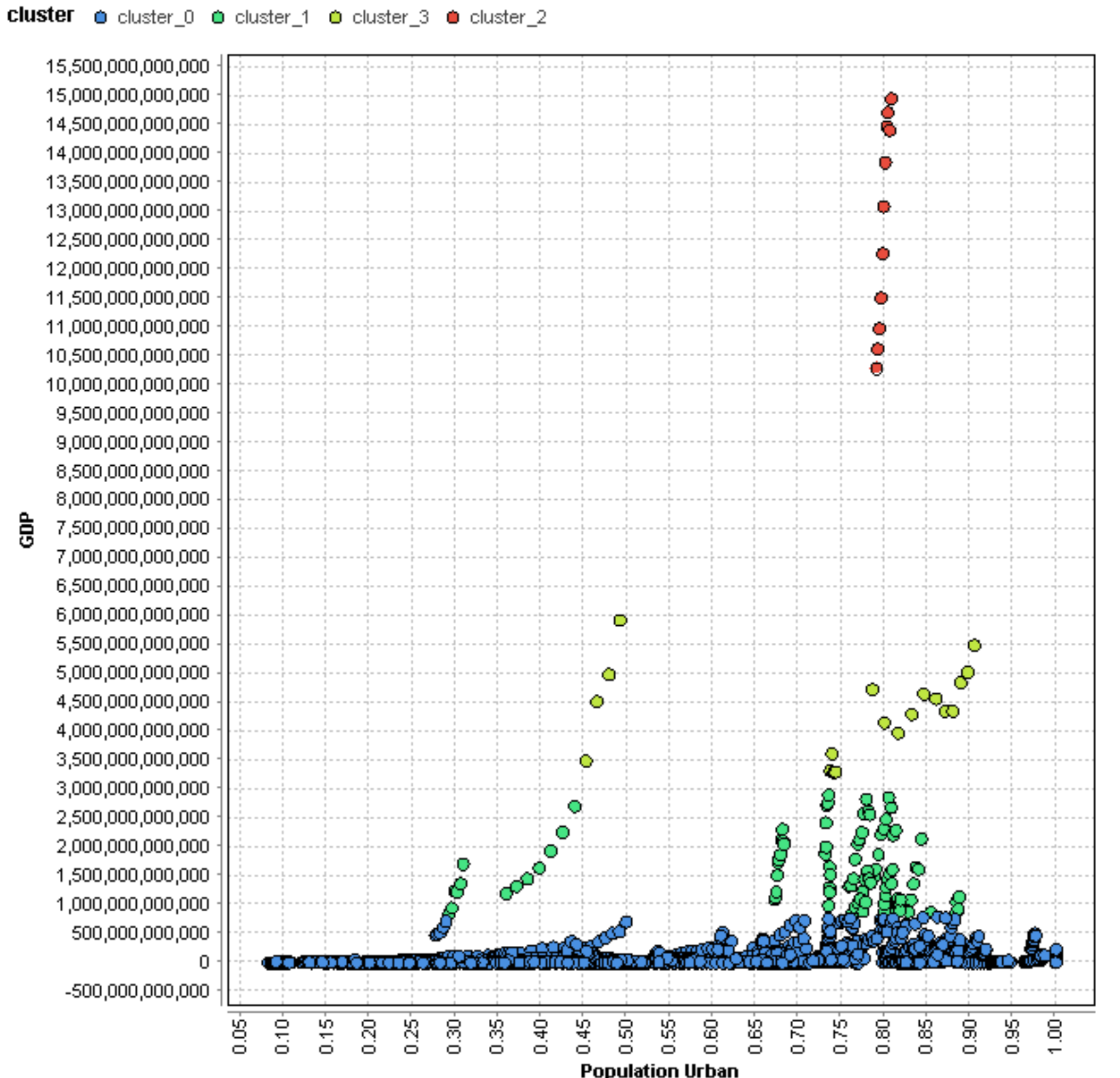
### 7A - GDP vs Infant Mortality Rate



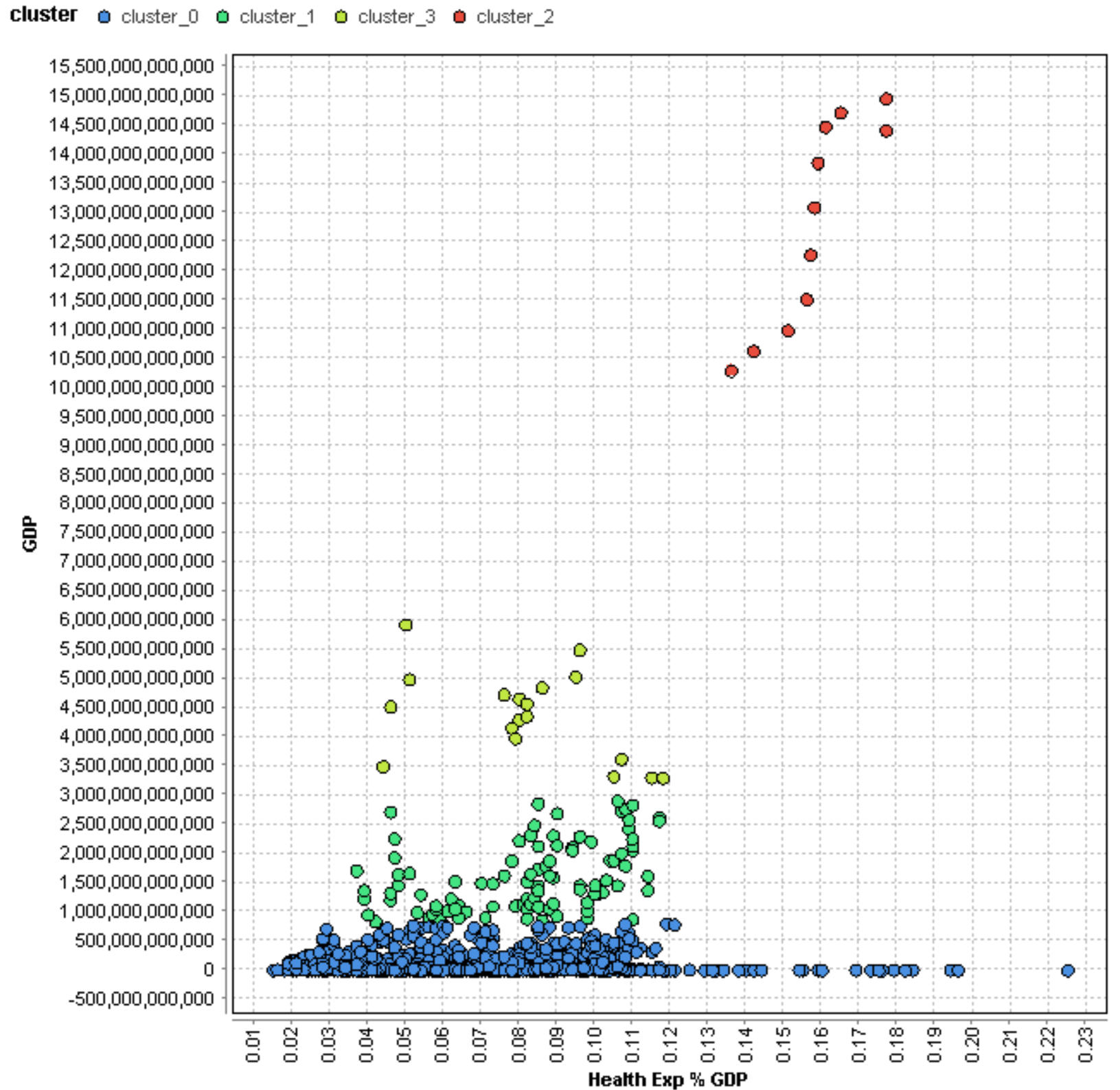
### 7B - GDP vs Internet Usage



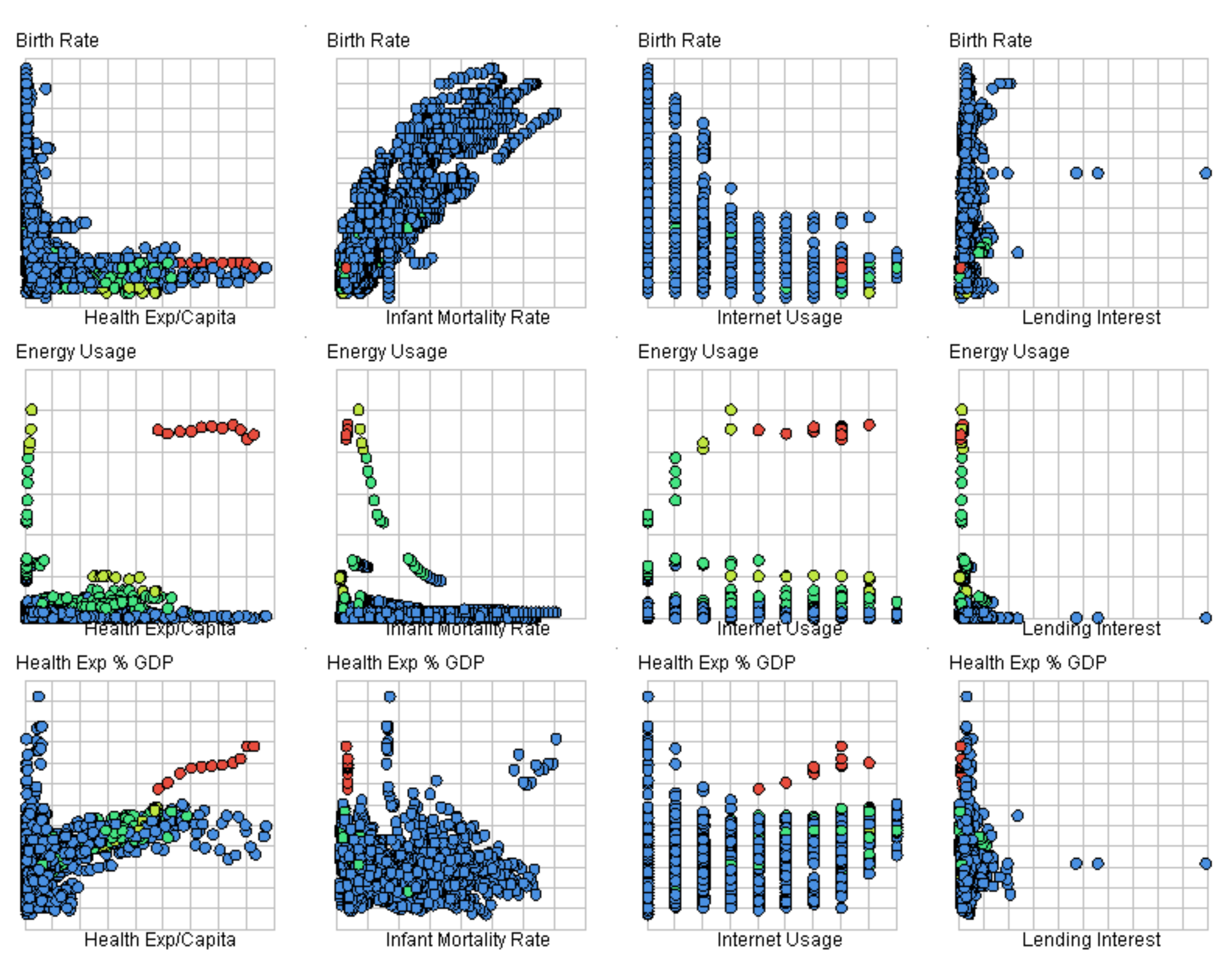
### 7C - GDP Vs Urban Population



### 7D – GDP vs Health Care Expenditure % GDP



### Figure 7E - Miscellaneous Cluster Charts (k=4)



## Figure 8 – Correlation Matrix of GDP vs Other Attributes



