542\_Final

Brittany Dymond, Hailey Fagerness, Cecilia Liang, Kathy Wang

3/16/2020

# Fossil Fuels & Economic Development

## Rationale of our two questions

How do fossil fuels relate to social and economic development in different countries? \* Can we identify groups of countries with similar oil production and GDP and if/how oil production impacts a country’s GDP? \* Can we identify groups of countries with similar population and fossil fuel usage and if/how population size affects fossil fuel usage?

Data: \* Population \* GDP per Capita \* Oil Production (92 countries) \* Fossil fuel use (as % of total electricity generating capacity)

## *QUESTION 1 CLUSTERING CODE START*

RESEARCH QUESTION: ‘Can we identify groups of countries with similar oil production and GDP and if/how oil production impacts a country’s GDP?’

Data used: \* Oil Production: [from U.S. Energy Information Administration] For calendar year 2019, on a comparable best-estimate basis \* GDP per Capita: [from Wikipedia] Converted at market exchange rates to current U.S. dollars, divided by the population for the same year

Prep to cluster OilProduction and GDP\_pc Getting data from github and initializing：

## Warning: package 'dplyr' was built under R version 3.6.3

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

Removing rows where OilProduction == 0：

### Clustering Part

Preparing to cluster oil production & GDP：

## OilProduction GDP\_pc  
## Albania 22915 5372  
## Algeria 1348361 3980  
## Angola 1769615 3037  
## Argentina 510560 9887  
## Australia 289749 53825  
## Austria 15161 50022

This is for replicability of results.

### Partitioning Technique: PAM

1. Apply function and indicate the amount of clusters required
2. Clustering results

TABLE OF CLUSTERS：

##   
## 1 2 3 4   
## 58 14 17 3

1. Evaluate Results

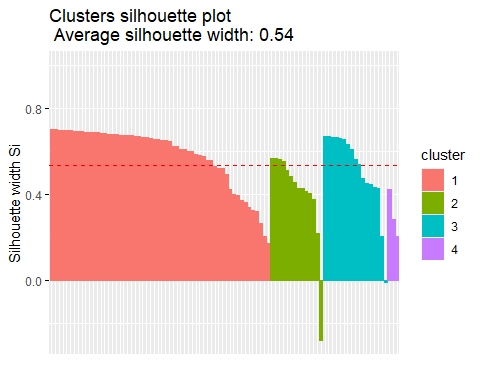
AVG SILHOUETTES：

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 3.6.3

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

## cluster size ave.sil.width  
## 1 1 58 0.58  
## 2 2 14 0.41  
## 3 3 17 0.52  
## 4 4 3 0.30



DECTECTING ANOMALIES：

## cluster neighbor sil\_width  
## Vietnam 1 3 0.7014283  
## Congo, Republic of the 1 3 0.7006665  
## Papua New Guinea 1 3 0.6977134  
## Ghana 1 3 0.6972409  
## Timor-Leste 1 3 0.6957549  
## Tunisia 1 3 0.6955229

Requesting negative silhouettes：

## cluster neighbor sil\_width  
## Italy 2 3 -0.277837863  
## Romania 3 1 -0.008107932

### Hierarchizing/Agglomerative Technique: AGNES

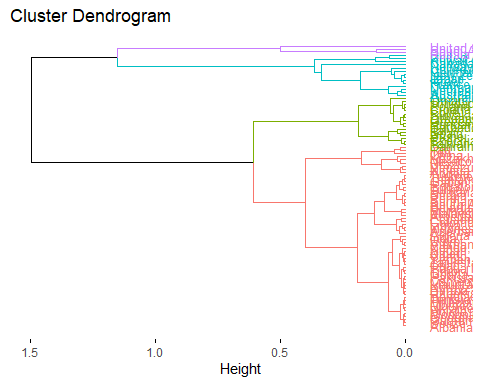
1. Apply function and indicate the amount of clusters required
2. Clustering results

TABLE OF CLUSTERS:

##   
## 1 2 3 4   
## 59 14 16 3

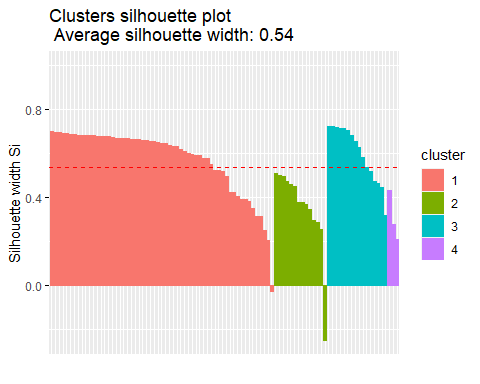
1. Evaluate results

DENDOGRAM:



AVG SILHOUETTES:

## cluster size ave.sil.width  
## 1 1 59 0.57  
## 2 2 14 0.35  
## 3 3 16 0.60  
## 4 4 3 0.30



DECTECTING ANOMALIES:

## cluster neighbor sil\_width  
## Vietnam 1 3 0.6977023  
## Congo, Republic of the 1 3 0.6970225  
## Egypt 1 3 0.6946234  
## Ghana 1 3 0.6884544  
## Papua New Guinea 1 3 0.6880476  
## Timor-Leste 1 3 0.6860947

Requesting negative silhouettes:

## cluster neighbor sil\_width  
## Romania 1 3 -0.02769694  
## Kuwait 2 3 -0.25164833

### Hierarchizing/Divisive Technique: DIANA

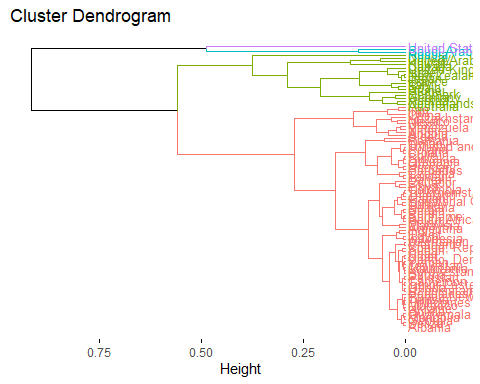
1. Apply function and indicate the amount of clusters required
2. Clustering results

TABLE OF CLUSTERS:

##   
## 1 2 3 4   
## 72 17 2 1

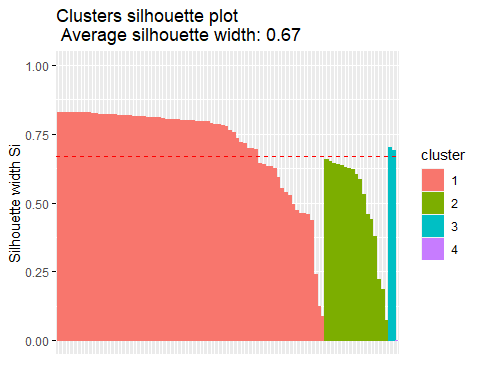
1. Evaluate results

DENDOGRAM:



AVG SILHOUETTES:

## cluster size ave.sil.width  
## 1 1 72 0.72  
## 2 2 17 0.50  
## 3 3 2 0.70  
## 4 4 1 0.00



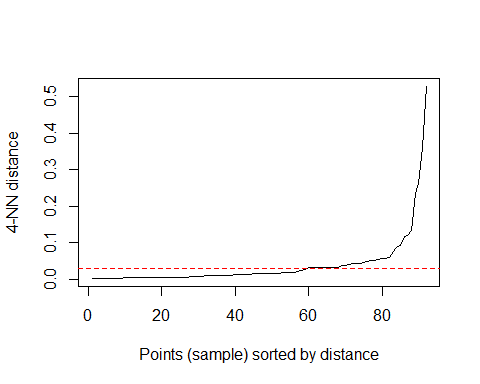
DECTECTING ANOMALIES:

## cluster neighbor sil\_width  
## Mongolia 1 2 0.8295996  
## Bolivia 1 2 0.8294036  
## Ukraine 1 2 0.8291026  
## Guatemala 1 2 0.8285038  
## Tunisia 1 2 0.8284611  
## Georgia 1 2 0.8284214

Requesting negative silhouettes:

## [1] cluster neighbor sil\_width  
## <0 rows> (or 0-length row.names)

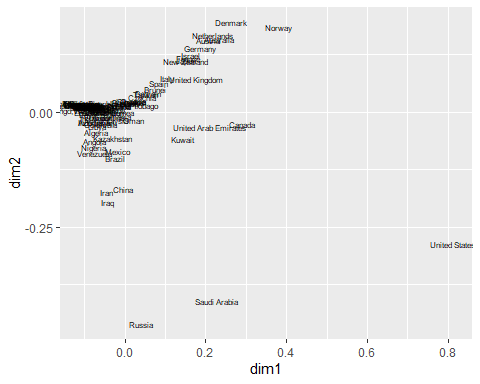
### Density Based Clustering: DBSCAN



HOW MANY OUTLIERS? (0 identified outliers)

## DBSCAN clustering for 92 objects.  
## Parameters: eps = 0.03, minPts = 4  
## The clustering contains 3 cluster(s) and 20 noise points.  
##   
## 0 1 2 3   
## 20 53 14 5   
##   
## Available fields: cluster, eps, minPts

Save coordinates to original data frame:

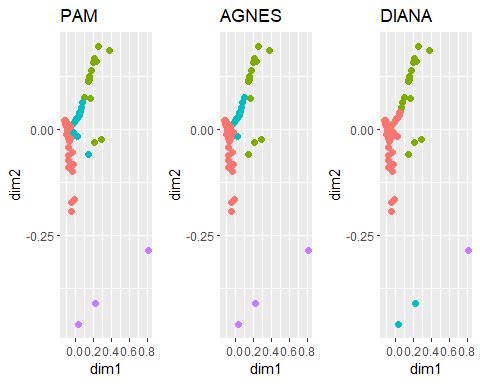


* Plot PAM：
* Plot AGNES：
* Plot DIANA：

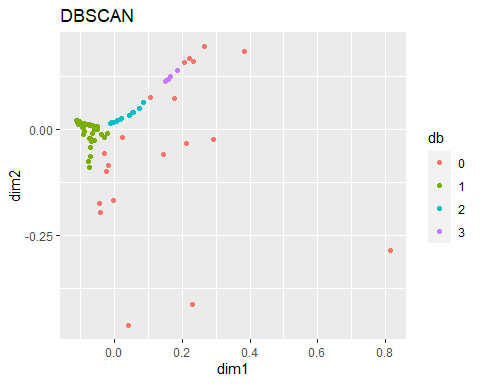
Compare results visually：

## Warning: package 'ggpubr' was built under R version 3.6.3

## Loading required package: magrittr

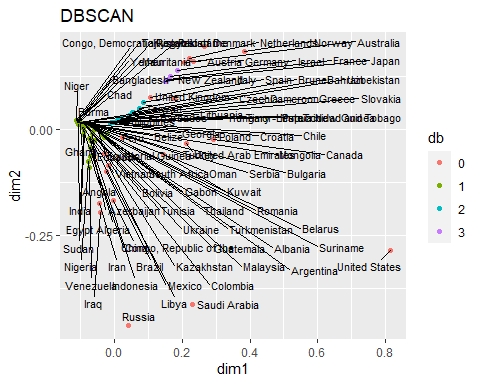


* Plot DBSCAN：

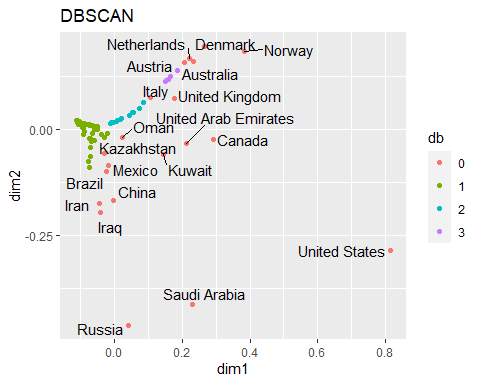


Annotating:

## Warning: package 'ggrepel' was built under R version 3.6.3



Annotating Outliers:



BASED ON CLUSTERING, WE WILL USE DBSCAN. This cluster had high production &/OR high GDP (outliers).

## [1] "Australia" "Austria" "Brazil"   
## [4] "Canada" "China" "Denmark"   
## [7] "Iran" "Iraq" "Italy"   
## [10] "Kazakhstan" "Kuwait" "Mexico"   
## [13] "Netherlands" "Norway" "Oman"   
## [16] "Russia" "Saudi Arabia" "United Arab Emirates"  
## [19] "United Kingdom" "United States"

This cluster had higher production & lower GDP.

## [1] "Albania" "Algeria"   
## [3] "Angola" "Argentina"   
## [5] "Azerbaijan" "Bangladesh"   
## [7] "Belarus" "Belize"   
## [9] "Bolivia" "Bulgaria"   
## [11] "Burma" "Cameroon"   
## [13] "Chad" "Colombia"   
## [15] "Congo, Democratic Republic of the" "Congo, Republic of the"   
## [17] "Ecuador" "Egypt"   
## [19] "Equatorial Guinea" "Gabon"   
## [21] "Georgia" "Ghana"   
## [23] "Guatemala" "India"   
## [25] "Indonesia" "Kyrgyzstan"   
## [27] "Libya" "Malaysia"   
## [29] "Mauritania" "Mongolia"   
## [31] "Morocco" "Niger"   
## [33] "Nigeria" "Pakistan"   
## [35] "Papua New Guinea" "Peru"   
## [37] "Philippines" "Romania"   
## [39] "Serbia" "South Africa"   
## [41] "Sudan" "Suriname"   
## [43] "Tajikistan" "Thailand"   
## [45] "Timor-Leste" "Tunisia"   
## [47] "Turkey" "Turkmenistan"   
## [49] "Ukraine" "Uzbekistan"   
## [51] "Venezuela" "Vietnam"   
## [53] "Yemen"

This cluster had lower production & lower GDP.

## [1] "Bahrain" "Barbados" "Brunei"   
## [4] "Chile" "Croatia" "Czechia"   
## [7] "Greece" "Hungary" "Lithuania"   
## [10] "Poland" "Slovakia" "Spain"   
## [13] "Taiwan" "Trinidad and Tobago"

This cluster had lower production & higher GDP.

## [1] "France" "Germany" "Israel" "Japan" "New Zealand"

## *QUESTION 1 REGRESSION START*

* Hypothesis:
  + Model 1: GDP Per Capita ~ Oil Production
  + Model 2: GDP Per Capita ~ Oil Production + Continent
* Continuous Outcome – – GDP Per Capita
* Independent variable – – Oil Production
* Control variable – – Continent
* Rationale for hypothesis
  + Oil infrastructure supports GDP
  + OPEC // many economies heavily rely on oil income
  + Oil price wars (like now with Saudi Arabia and Russia) impact oil prices and thus GDP

Preparing to regress Oil Production & GDP

## 'data.frame': 92 obs. of 12 variables:  
## $ Country : chr "Albania" "Algeria" "Angola" "Arge"..  
## $ fossilFuel\_PctTotalElec: num 0.05 0.96 0.34 0.69 0.72 0.25 0.84 ..  
## $ OilProduction : num 22915 1348361 1769615 510560 289749..  
## $ Population : int 2880917 43053054 31825295 44780677 ..  
## $ GDP\_pc : int 5372 3980 3037 9887 53825 50022 468..  
## $ Continent : Factor w/ 7 levels "Africa","Asia",..: 3..  
## $ pam : Factor w/ 4 levels "1","2","3","4": 1 1 ..  
## $ agn : Factor w/ 4 levels "1","2","3","4": 1 1 ..  
## $ dia : Factor w/ 4 levels "1","2","3","4": 1 1 ..  
## $ db : Factor w/ 4 levels "0","1","2","3": 2 2 ..  
## $ dim1 : num -0.0795 -0.0701 -0.0722 -0.0399 0.2..  
## $ dim2 : num 0.0111 -0.0435 -0.063 -0.0116 0.159..

### EXPLANATORY APPROACH

1.State the hypotheses

hypo1=formula(GDP\_pc ~ OilProduction)  
hypo2=formula(GDP\_pc ~ OilProduction + Continent)

2.Save colums needed and varify data types

## 'data.frame': 92 obs. of 3 variables:  
## $ OilProduction: num 22915 1348361 1769615 510560 289749 ...  
## $ GDP\_pc : int 5372 3980 3037 9887 53825 50022 4689 25273 1905 18069 ...  
## $ Continent : Factor w/ 7 levels "Africa","Asia",..: 3 1 1 7 6 3 4 2 2 5 ...

3.Compute regression models

4.Hypothesis results

* First Hypothesis:

##   
## Call:  
## glm(formula = hypo1, family = "gaussian", data = DataRegGauss)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -20272 -11232 -7177 5968 61852   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.337e+04 1.882e+03 7.103 2.76e-10 \*\*\*  
## OilProduction 1.673e-03 7.318e-04 2.286 0.0246 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 280755365)  
##   
## Null deviance: 2.6736e+10 on 91 degrees of freedom  
## Residual deviance: 2.5268e+10 on 90 degrees of freedom  
## AIC: 2054.7  
##   
## Number of Fisher Scoring iterations: 2

* Second Hypothesis:

summary(gauss2)

##   
## Call:  
## glm(formula = hypo2, family = "gaussian", data = DataRegGauss)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -29507 -8171 -1380 5466 47217   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.227e+03 3.288e+03 0.677 0.500143   
## OilProduction 2.140e-03 6.498e-04 3.294 0.001448 \*\*   
## ContinentAsia 7.450e+03 4.231e+03 1.761 0.081897 .   
## ContinentEurope 2.500e+04 4.421e+03 5.656 2.09e-07 \*\*\*  
## ContinentEurope/Asia -2.640e+02 7.171e+03 -0.037 0.970725   
## ContinentNorth America 1.499e+04 6.409e+03 2.340 0.021675 \*   
## ContinentOceania 2.990e+04 8.669e+03 3.449 0.000882 \*\*\*  
## ContinentSouth America 3.519e+03 5.678e+03 0.620 0.537102   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 193147596)  
##   
## Null deviance: 2.6736e+10 on 91 degrees of freedom  
## Residual deviance: 1.6224e+10 on 84 degrees of freedom  
## AIC: 2026  
##   
## Number of Fisher Scoring iterations: 2

5.Searching for a better model

## Analysis of Deviance Table  
##   
## Model 1: GDP\_pc ~ OilProduction  
## Model 2: GDP\_pc ~ OilProduction + Continent  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 90 2.5268e+10   
## 2 84 1.6224e+10 6 9043584826 2.03e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

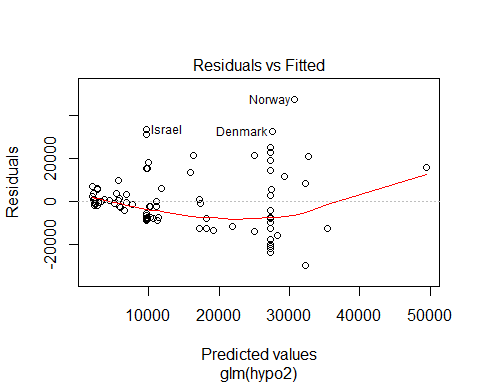
Model for the Second hypothesis is chosen. This is the RSquared:

## Warning: package 'rsq' was built under R version 3.6.3

## [1] 0.3425815

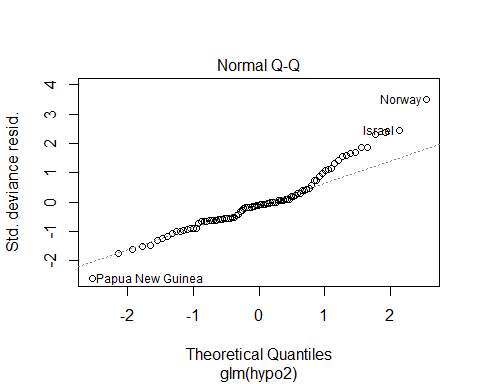
6.Verify the situation of chosen model:

6.1. Linearity between dependent variable and predictors is assumed, then these dots should follow a linear and horizontal trend:

 The linear trend is not obvious, and the distribution range goes wider when the predicted values increase. I’d like to say it represents the linearity between our variables in a certain level. Further research upon outliers are necessary.

6.2. Normality of residuals is assumed:

Visual exploration

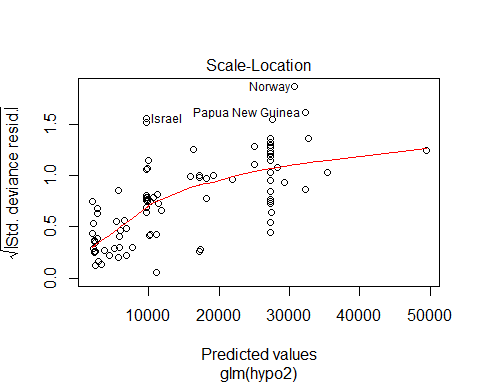


Mathematical exploration:

##   
## Shapiro-Wilk normality test  
##   
## data: gauss2$residuals  
## W = 0.94464, p-value = 0.000681

6.3. Homoscedasticity is assumed, so check if residuals are spread equally along the ranges of predictors

Visual exploration:



Mathematical exploration:

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

##   
## studentized Breusch-Pagan test  
##   
## data: gauss2  
## BP = 19.735, df = 7, p-value = 0.006171

6.4. We assume that there is no colinearity, that is, that the predictors are not correlated.

## Warning: package 'car' was built under R version 3.6.3

## Loading required package: carData

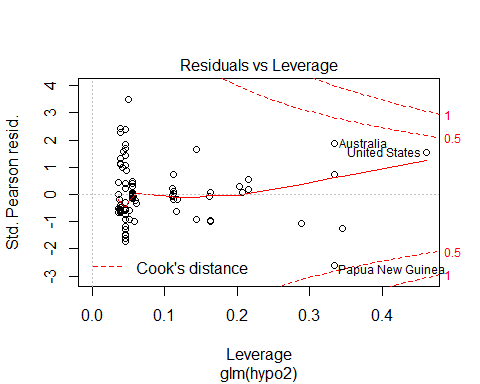
##   
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':  
##   
## recode

## GVIF Df GVIF^(1/(2\*Df))  
## OilProduction 1.146212 1 1.070613  
## Continent 1.146212 6 1.011437

6.5. Analize the effect of atypical values. Determine if outliers (points that are far from the rest, but still in the trend) or high-leverage points (far from the trend but close to the rest) are influential

Visual exploration:



Querying:

gaussInf=as.data.frame(influence.measures(gauss2)$is.inf)  
gaussInf[gaussInf$cook.d,]

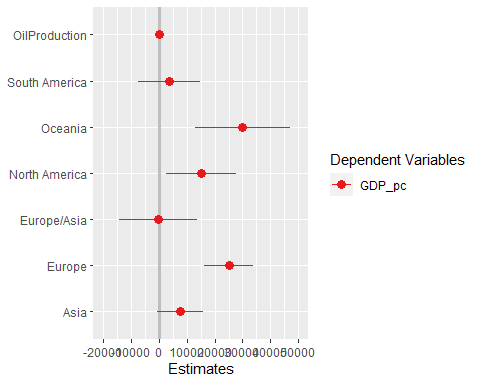
## [1] dfb.1\_ dfb.OlPr dfb.CntA dfb.CntE dfb.CE/A dfb.CnNA dfb.CntO dfb.CnSA  
## [9] dffit cov.r cook.d hat   
## <0 rows> (or 0-length row.names)

1. Finally, a nice summary plot of our work

## Warning: package 'sjPlot' was built under R version 3.6.3

## Registered S3 methods overwritten by 'lme4':  
## method from  
## cooks.distance.influence.merMod car   
## influence.merMod car   
## dfbeta.influence.merMod car   
## dfbetas.influence.merMod car

## Learn more about sjPlot with 'browseVignettes("sjPlot")'.



### PREDICTIVE APPROACH

1. Splitting the data set

## Warning: package 'caret' was built under R version 3.6.3

## Loading required package: lattice

1. Regress with train data

##   
## Call:  
## NULL  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -29470 -9388 -1480 7635 44909   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.204e+03 4.180e+03 0.527 0.59982   
## OilProduction 2.670e-03 8.371e-04 3.189 0.00221 \*\*   
## ContinentAsia 8.393e+03 5.127e+03 1.637 0.10657   
## ContinentEurope 2.646e+04 5.685e+03 4.655 1.68e-05 \*\*\*  
## `ContinentEurope/Asia` -1.646e+03 8.083e+03 -0.204 0.83925   
## `ContinentNorth America` 1.532e+04 7.708e+03 1.988 0.05109 .   
## ContinentOceania 2.986e+04 9.604e+03 3.109 0.00280 \*\*   
## `ContinentSouth America` 3.651e+03 7.890e+03 0.463 0.64509   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 224537488)  
##   
## Null deviance: 2.3585e+10 on 71 degrees of freedom  
## Residual deviance: 1.4370e+10 on 64 degrees of freedom  
## AIC: 1598.4  
##   
## Number of Fisher Scoring iterations: 2

1. Evaluate performance

## RMSE Rsquared MAE   
## 10322.575720 0.536489 7547.290591

## *QUESTION 2 CLUSTERING START*

RESEARCH QUESTION: ‘Can we identify groups of countries with similar population and fossil fuel usage and if/how population size affects fossil fuel usage?’

Data used: \* fossilFuel\_PctTotalElec: [from CIA World Factbook] percentage of total electricity generating capacity that comes from fossil fuels \* Population: [UN Dept of Economic and Social Affairs] World population estimates

Prep to cluster fossilFuel\_PctTotalElec and Population

## fossilFuel\_PctTotalElec Population  
## Albania 0.05 2880917  
## Algeria 0.96 43053054  
## Angola 0.34 31825295  
## Argentina 0.69 44780677  
## Australia 0.72 25203198  
## Austria 0.25 8955102

Set random seed for replicability of results:

Setting distance matrix:

Defining number of clusters for each method (NumCluster = 5) Clustering via pam method:

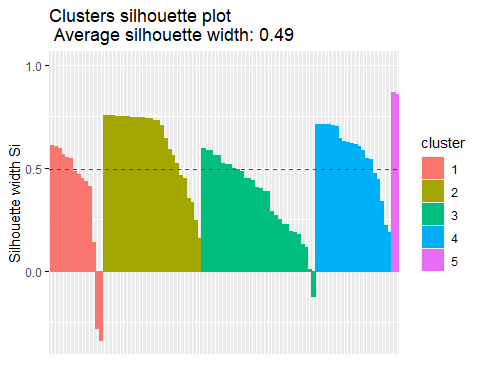
Adding pam results to original DF (DFnew1)

REPORT: Table of Cluster:

##   
## 1 2 3 4 5   
## 14 26 30 20 2

REPORT: Evaluate Results:

## cluster size ave.sil.width  
## 1 1 14 0.38  
## 2 2 26 0.63  
## 3 3 30 0.36  
## 4 4 20 0.57  
## 5 5 2 0.86



REPORT: Detecting Anomalies

Saving individual silhouettes

## cluster neighbor sil\_width  
## Tajikistan 1 4 0.6090716  
## Albania 1 4 0.6060691  
## Norway 1 4 0.5978116  
## Timor-Leste 1 4 0.5667970  
## Congo, Democratic Republic of the 1 4 0.5525901  
## France 1 4 0.5468275

Requesting negative silhouettes:

## cluster neighbor sil\_width  
## Angola 1 4 -0.2800818  
## Georgia 1 4 -0.3377824  
## Ghana 3 4 -0.1223473

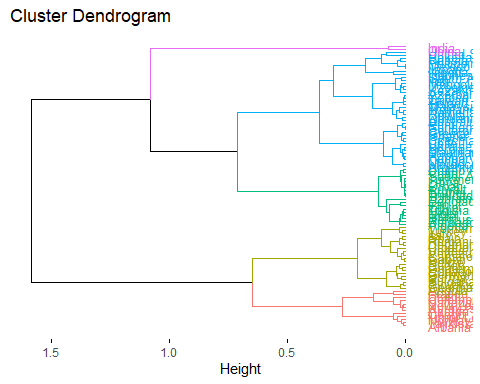
Cluster via agnes method; indicate number of clusters (NumCluster):

Adding agn results to original DF (DFnew1)

REPORT: Table of clusters:

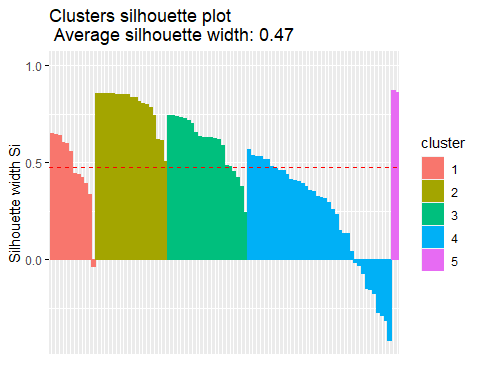
##   
## 1 2 3 4 5   
## 12 19 21 38 2

Evaluating results:



REPORT: Average silhouettes

## cluster size ave.sil.width  
## 1 1 12 0.47  
## 2 2 19 0.79  
## 3 3 21 0.60  
## 4 4 38 0.23  
## 5 5 2 0.87



REPORT: Detecting anomalies

## cluster neighbor sil\_width  
## Tajikistan 1 3 0.6486823  
## Albania 1 3 0.6461141  
## Norway 1 3 0.6392016  
## Timor-Leste 1 3 0.6056001  
## Congo, Democratic Republic of the 1 3 0.5963339  
## France 1 3 0.5571241

Requesting negative silhouettes:

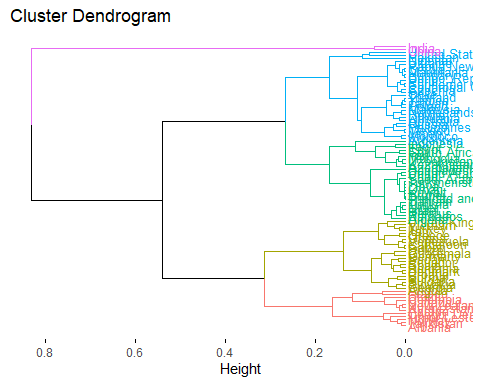
## cluster neighbor sil\_width  
## Colombia 1 3 -0.03279474  
## Indonesia 4 2 -0.01212078  
## Iran 4 2 -0.03057378  
## Chile 4 3 -0.06866262  
## Azerbaijan 4 2 -0.14697731  
## South Africa 4 2 -0.15226216  
## Ghana 4 3 -0.17299110  
## Greece 4 3 -0.27237486  
## Uzbekistan 4 2 -0.28525827  
## Kazakhstan 4 2 -0.31159271  
## Mongolia 4 2 -0.41514761

Cluster via diana method; indicate number of clusters (NumCluster):

Adding diana results to original DF (DFnew1):

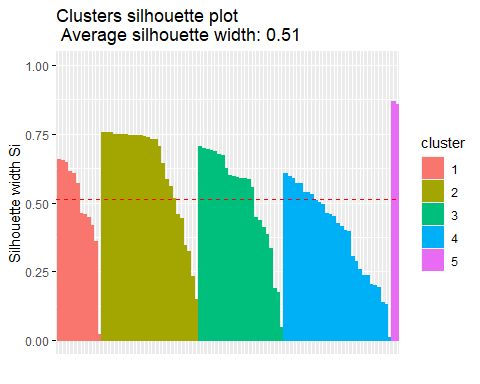
REPORT: Table of clusters

##   
## 1 2 3 4 5   
## 12 26 23 29 2

REPORT: Dendrogram 

REPORT: Average silhouettes

## cluster size ave.sil.width  
## 1 1 12 0.49  
## 2 2 26 0.62  
## 3 3 23 0.52  
## 4 4 29 0.39  
## 5 5 2 0.86



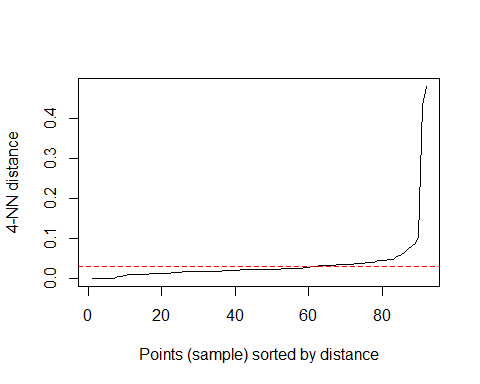
REPORT: Detecting anomalies:

## cluster neighbor sil\_width  
## Tajikistan 1 3 0.6564827  
## Albania 1 3 0.6539019  
## Norway 1 3 0.6467797  
## Timor-Leste 1 3 0.6133818  
## Congo, Democratic Republic of the 1 3 0.6057145  
## France 1 3 0.5720875

Requesting negative silhouettes:

## [1] cluster neighbor sil\_width  
## <0 rows> (or 0-length row.names)

Cluster via DBSCAN method; indicate minimum neighbors (4):



Setting distance (epsilon):

REPORT: Number of clusters and outliers produced

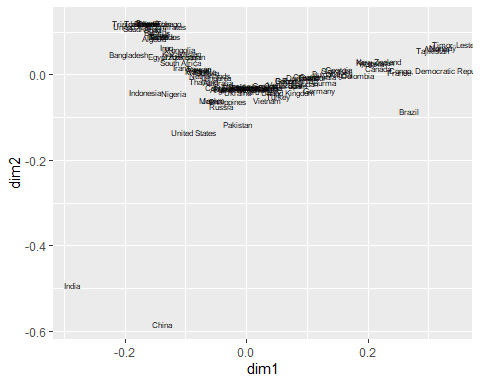
## DBSCAN clustering for 92 objects.  
## Parameters: eps = 0.03, minPts = 4  
## The clustering contains 4 cluster(s) and 11 noise points.  
##   
## 0 1 2 3 4   
## 11 4 52 4 21   
##   
## Available fields: cluster, eps, minPts

Saving results:

Comparing clusters

Prepare a bidimensional map:

View bidimensional map:



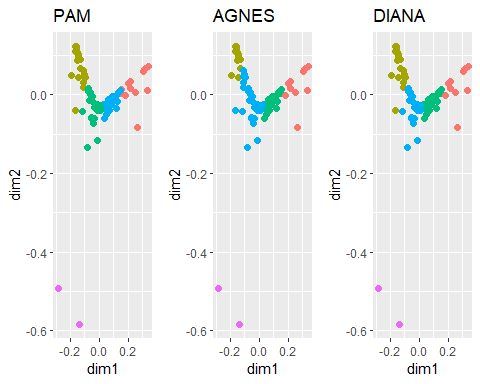
Results from pam:

Results from agnes:

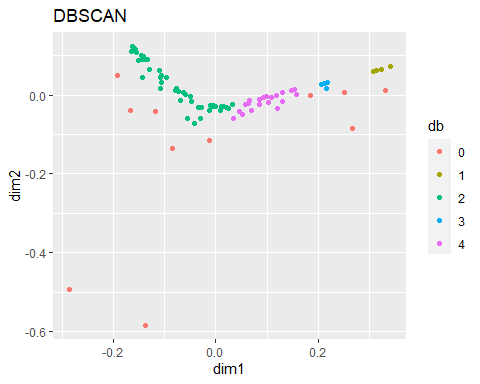
Results from diana：

Compare visually:

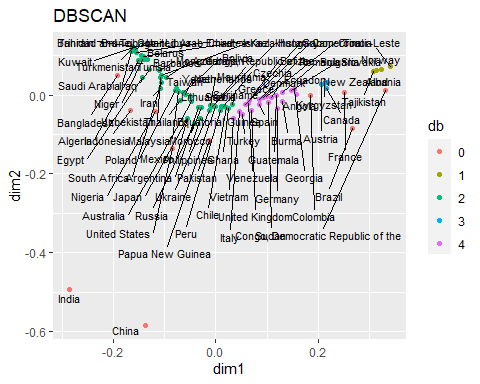
Viewing pam, agnes, and diana plots side by side



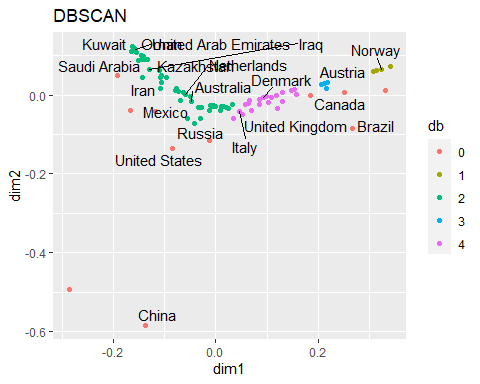
Plot results from DBSCAN：



Annotating graph with country names：



Annotating just the outlier countries：



CHOOSING DIANA METHOD DUE TO HAVING ZERO NEGATIVE SILHOUETTES。

## *QUESTION 2 REGRESSION CODE START*

* Hypothesis:
  + Model 1: FF ~ Population
  + Model 2: FF ~ Population + Developed
* Method:
  + Binary Outcome – – FF usage (Median percentage FF use of total electricity capacity)
* Control variable – – Developed (Median GDP per capita)
* Independent variable – – Population
* Rationale for hypothesis:
  + Larger populations would exhibit higher fossil fuel usage as a percent of total electricity capacity

Changing dtype for population：

## Country fossilFuel\_PctTotalElec OilProduction   
## "character" "numeric" "numeric"   
## Population GDP\_pc   
## "numeric" "integer"

Changing dtype for GDP\_pc(gdp):

## Country fossilFuel\_PctTotalElec OilProduction   
## "character" "numeric" "numeric"   
## Population GDP\_pc   
## "numeric" "numeric"

Filtering out non-oil producing countries & creating new DF (teamnew):

## Country fossilFuel\_PctTotalElec OilProduction  
## 2 Albania 0.05 22915  
## 3 Algeria 0.96 1348361  
## 4 Angola 0.34 1769615  
## 6 Argentina 0.69 510560  
## 9 Australia 0.72 289749  
## 10 Austria 0.25 15161  
## 11 Azerbaijan 0.84 833538  
## 13 Bahrain 1.00 40000  
## 14 Bangladesh 0.97 4189  
## 15 Barbados 0.93 1000  
## 16 Belarus 0.96 25000  
## 18 Belize 0.51 2000  
## 21 Bolivia 0.76 58077  
## 24 Brazil 0.17 2515459  
## 25 Brunei 1.00 109117  
## 26 Bulgaria 0.39 1000  
## 28 Burma 0.39 15000  
## 32 Cameroon 0.52 93205  
## 33 Canada 0.23 3662694  
## 35 Chad 0.98 110156  
## 36 Chile 0.59 4423  
## 37 China 0.62 3980650  
## 38 Colombia 0.29 897784  
## 40 Congo, Democratic Republic of the 0.02 20000  
## 41 Congo, Republic of the 0.64 308363  
## 43 Croatia 0.45 13582  
## 45 Czechia 0.60 2333  
## 46 Denmark 0.46 140637  
## 50 Ecuador 0.43 548421  
## 51 Egypt 0.91 490000  
## 53 Equatorial Guinea 0.61 227000  
## 60 France 0.17 16418  
## 61 Gabon 0.51 210820  
## 63 Georgia 0.35 400  
## 64 Germany 0.41 46839  
## 65 Ghana 0.58 100549  
## 66 Greece 0.57 3172  
## 68 Guatemala 0.41 8977  
## 75 Hungary 0.64 13833  
## 77 India 0.71 715459  
## 78 Indonesia 0.85 833667  
## 79 Iran 0.84 3990956  
## 80 Iraq 0.91 4451516  
## 82 Israel 0.95 390  
## 83 Italy 0.54 70675  
## 85 Japan 0.71 3918  
## 87 Kazakhstan 0.86 1595199  
## 91 Kuwait 1.00 2923825  
## 92 Kyrgyzstan 0.24 1000  
## 98 Libya 1.00 1003000  
## 99 Lithuania 0.73 2000  
## 104 Malaysia 0.78 661240  
## 109 Mauritania 0.65 5000  
## 111 Mexico 0.71 2186877  
## 114 Mongolia 0.87 23426  
## 116 Morocco 0.68 160  
## 121 Netherlands 0.75 18087  
## 122 New Zealand 0.23 35574  
## 124 Niger 0.95 13000  
## 125 Nigeria 0.80 1999885  
## 127 Norway 0.03 1647975  
## 128 Oman 1.00 1006841  
## 129 Pakistan 0.62 80000  
## 131 Papua New Guinea 0.63 56667  
## 133 Peru 0.61 40266  
## 134 Philippines 0.67 20000  
## 135 Poland 0.79 20104  
## 139 Romania 0.47 504000  
## 140 Russia 0.68 10800000  
## 147 Saudi Arabia 1.00 12000000  
## 149 Serbia 0.65 20000  
## 153 Slovakia 0.36 200  
## 156 South Africa 0.85 2000  
## 158 Spain 0.47 2667  
## 160 Sudan 0.44 255000  
## 161 Suriname 0.61 17000  
## 164 Taiwan 0.79 196  
## 165 Tajikistan 0.06 180  
## 167 Thailand 0.76 257525  
## 168 Timor-Leste 0.00 60661  
## 171 Trinidad and Tobago 1.00 60090  
## 172 Tunisia 0.94 48757  
## 173 Turkey 0.53 49497  
## 174 Turkmenistan 1.00 230779  
## 177 Ukraine 0.65 31989  
## 178 United Arab Emirates 0.99 3106077  
## 179 United Kingdom 0.50 939760  
## 180 United States 0.70 15043000  
## 182 Uzbekistan 0.86 52913  
## 184 Venezuela 0.51 2276967  
## 185 Vietnam 0.56 301850  
## 186 Yemen 0.79 22000  
## Population GDP\_pc  
## 2 2880917 5372  
## 3 43053054 3980  
## 4 31825295 3037  
## 6 44780677 9887  
## 9 25203198 53825  
## 10 8955102 50022  
## 11 10047718 4689  
## 13 1641172 25273  
## 14 163046161 1905  
## 15 287025 18069  
## 16 9452411 6603  
## 18 390353 4925  
## 21 11513100 3670  
## 24 211049527 8796  
## 25 433285 27871  
## 26 7000119 9518  
## 28 54045420 1244  
## 32 25876380 1514  
## 33 37411047 46212  
## 35 15946876 861  
## 36 18952038 15399  
## 37 1433783686 10098  
## 38 50339443 6508  
## 40 86790567 500  
## 41 5380508 2534  
## 43 4130304 14949  
## 45 10689209 23213  
## 46 5771876 59795  
## 50 17373662 6249  
## 51 100388073 3046  
## 53 1355986 8927  
## 60 65129728 41760  
## 61 2172579 8112  
## 63 3996765 4289  
## 64 83517045 46563  
## 65 28833629 2223  
## 66 10473455 19974  
## 68 17581472 4616  
## 75 9684679 17463  
## 77 1366417754 2171  
## 78 270625568 4163  
## 79 82913906 5506  
## 80 39309783 5738  
## 82 8519377 42823  
## 83 60550075 32946  
## 85 126860301 40846  
## 87 18551427 9139  
## 91 4207083 29266  
## 92 6415850 1292  
## 98 6777452 5019  
## 99 2759627 19266  
## 104 31949777 11136  
## 109 4525696 1392  
## 111 127575529 10118  
## 114 3225167 4132  
## 116 36471769 3345  
## 121 17097130 52367  
## 122 4783063 40634  
## 124 23310715 405  
## 125 200963599 2222  
## 127 5378857 77975  
## 128 4974986 17791  
## 129 216565318 1388  
## 131 8776109 2742  
## 133 32510453 7046  
## 134 108116615 3294  
## 135 37887768 14901  
## 139 19364557 12482  
## 140 145872256 11162  
## 147 34268528 22865  
## 149 8772235 7397  
## 153 5457013 19547  
## 156 58558270 6100  
## 158 46736776 29961  
## 160 42813238 714  
## 161 581372 6310  
## 164 23773876 24827  
## 165 9321018 877  
## 167 69037513 7791  
## 168 1293119 2262  
## 171 1394973 16365  
## 172 11694719 3287  
## 173 83429615 8957  
## 174 5942089 7816  
## 177 43993638 3592  
## 178 9770529 37749  
## 179 67530172 41030  
## 180 329064917 65111  
## 182 32981716 1831  
## 184 28515829 2547  
## 185 96462106 2740  
## 186 29161922 943

Converting USDollar to a factor variable

Calling new variable ‘Developed’

Converting fossilFuel\_PctTotalElec to a factor variable

Calling new variable ‘FF’

Checking dtypes:

## 'data.frame': 92 obs. of 7 variables:  
## $ Country : chr "Albania" "Algeria" "Angola" "Arge"..  
## $ fossilFuel\_PctTotalElec: num 0.05 0.96 0.34 0.69 0.72 0.25 0.84 ..  
## $ OilProduction : num 22915 1348361 1769615 510560 289749..  
## $ Population : num 2880917 43053054 31825295 44780677 ..  
## $ GDP\_pc : num 5372 3980 3037 9887 53825 ...  
## $ Developed : Factor w/ 2 levels "0","1": 1 1 1 2 2 2 ..  
## $ FF : Factor w/ 2 levels "0","1": 1 2 1 2 2 1 ..

Defining ‘Population’ as independent variable:

Defining columns needed：

Verify dtypes for colsNeededDico:

## 'data.frame': 92 obs. of 3 variables:  
## $ FF : Factor w/ 2 levels "0","1": 1 2 1 2 2 1 2 2 2 2 ...  
## $ Population: num 2880917 43053054 31825295 44780677 25203198 ...  
## $ Developed : Factor w/ 2 levels "0","1": 1 1 1 2 2 2 1 2 1 2 ...

Create subset

Rename indexes by country

Define & compute regression models

Results of hypo3:

At p-value of 0.634, this model is not statistically significant

##   
## Call:  
## glm(formula = hypo3, family = "binomial", data = DataRegLogis)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.436 -1.129 -1.126 1.223 1.230   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -1.231e-01 2.217e-01 -0.555 0.579  
## Population 4.972e-10 1.045e-09 0.476 0.634  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 127.37 on 91 degrees of freedom  
## Residual deviance: 127.13 on 90 degrees of freedom  
## AIC: 131.13  
##   
## Number of Fisher Scoring iterations: 4

Results of hypo4:

At p-values of 0.631 and 0.673, this model also is not statistically significant:

##   
## Call:  
## glm(formula = hypo4, family = "binomial", data = DataRegLogis)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.479 -1.141 -1.089 1.192 1.268   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -2.120e-01 3.063e-01 -0.692 0.489  
## Population 5.028e-10 1.048e-09 0.480 0.631  
## Developed1 1.768e-01 4.184e-01 0.422 0.673  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 127.37 on 91 degrees of freedom  
## Residual deviance: 126.95 on 89 degrees of freedom  
## AIC: 132.95  
##   
## Number of Fisher Scoring iterations: 4

Analysis of variance between models:

## Analysis of Deviance Table  
##   
## Model 1: FF ~ Population  
## Model 2: FF ~ Population + Developed  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
## 1 90 127.13   
## 2 89 126.95 1 0.17869 0.6725

## Recommendations

### First Question

Oil production could be an important component of GDP, but higher oil production rate does not lead to higher GDP. If we want to evaluate the relationship between GDP and oil production, we also need to know what is the percentage of the GDP generated by oil production.

* *Same level variables are more easy to be compared*
* Too many countries that their oil production is close to zero
* *Try other control variables like export/import*
* Higher oil production does not lead to higher GDP necessarily

### Second Question

Neither model is statistically significant; no further analysis required.

Recommendations for future analysis of question #2 include:

* *Incorporate country-specific income levels as an additional variable*
* Remove major outliers from sample population
* Use actual fossil fuel usage data in lieu of ratios