Data 621 Final Project: Predicting Fertility Rates

December 20, 2018

Authors

Vikas Sinha

Luisa Velasco

Dan Wigodsky

Sarah Wigodsky

Zhenni Xie

Abstract

The aim is to create a model that will predict the fertility rate of a country. Models were built using data taken from the United Nations and chose to consider the effect of carbon dioxide emissions, cellular subscriptions, employment to population ratio, the poorest quintile's share in income, maternal mortality per 100,000 births, the percent of children sleeping under insecticide treated bed nets, the percent unmet family planning need, the percent of urban population living in slums, the ratio of literacy rates between women and men, and the net migration rate. There was a great deal of missing data. To address this, multiple strategies were used, including downloading data from other sources and the imputing the median value of the variable by region in the world. Multiple linear regression, principal component analysis, Poisson, negative binomial and zero inflated models were built. The most successful model was built using principal component analysis and showed a positive correlation between fertility rate and the percent of children sleeping under insecticide treated bed nets, the percent unmet family planning need and the percent of urban population living in slums. There is a negative correlation between fertility rate and ratio of literacy rates between women and men and cellular subscriptions.

Key Words

Fertility Rate - Average number of children a woman has over her life

Slums - Area in a city that is inhabited by impoverished people. It is overcrowded, lacking in safety and the homes are of poor quality.

Literacy Rate - Ratio of people over 15 years old who can read and write as compared with the total population over 15

Introduction

Accurate knowledge of population trends is needed for formulation of effective policies for addressing the changing needs and requirements of populations across the globe. Forecasts based on projections of fertility, mortality and migration rates are used for many purposes, such as predicting the demands for food, water, medical services and education, and the impact on labor markets, pension systems and the environment [2].

Although current data on population is available from sources such as the United Nations and the World Bank, there is a need for accurate population projections for longer time horizons. Prediction of fertility rates is a key component of population projections.

Together with migration and mortality rates, fertility rates is a key indicator of demographic change, including the age structure of future populations. The total fertility rate (TFR) is the average number of children a woman would bear if she survived through the end of the reproductive age span, experiencing at each age the age-specific fertility rates of that period [2].

Prior Work

Probabilistic methods are commonly used to project future fertility rates [1]. Since the assumptions underlying the probabilistic models affect the sensitivity of the projections, a number of different projections are produced, corresponding with each underlying assumption. In the current literature, those underlying assumptions are: (1) medium-fertility assumption; (2) high-fertility assumption; (3) low-fertility assumption; (4) constant-fertility assumption; (5) instant-replacement assumption. The median trajectory of the projections constitutes the medium-fertility assumption.

In *Modelling Fertility: A Semi-Parametric Approach* (Oberhofer and Reichsthaler, 2004) the authors present a categorical model of fertility based on the Generalized Linear Model. For predictor variables, they used only one factor – the age of the mother. This variable was modeled as a Bernoulli random variable, and the technique of Local Likelihood Estimation was used. Other factors such as marital status and ethnic origin were available but not used in the model.

Different Models for Different Phases of Growth

The latest models are based on a demographic transition theory to model demographic changes: the theory postulates three distinct phases of demographic growth to account for different patterns observed across countries of the world. Phase I is distinguished by high fertility rates and is prior to transitions. Phase II corresponds to low fertility rates. In Phase III, a time series model is used to project further fertility change, based on the assumption that in the long run the levels fluctuate around country-specific levels based on a Bayesian hierarchical model.

The revised model of United Nations Population Division uses three separate models for Fertility rate projections based on the different phases of the Fertility Transition. The model is described in detail in Alkema et al (2011). Further, it is believed that all countries have begun or already completed their Fertility Transition. The pace of the fertility decline is based on a double-logistic (A double-logistic function is a sum of two logistic functions) decline function which depends on the current fertility level. The model is hierarchical because in addition to the country-level information, a second level is used to determine the parameters of the double-logistic function; this second level takes into account the information for all countries) [3]. Thus the hierarchical model includes two levels for country and world.

DATA EXPLORATION

install.packages("tidyr")

```
##   country_code fertility_rate_children_per_woman carbon_dioxide_emissions
## 1          ABW                             1.800                       NA
## 2          AFG                             5.255                       NA
## 3          AGO                             5.950                       NA
## 4          AIA                             1.740                       NA
## 5          ALB                             6.230                       NA
## 6          AND                             1.220                       NA
##   cell_subs_per_100 employment_to_pop_ratio female_employment_to_pop_ratio
## 1         135.06589                      NA                             NA
## 2          74.88284                      NA                             NA
## 3          63.47921                      NA                             NA
## 4         179.80636                      NA                             NA
## 5         105.46997                    44.5                           43.7
## 6          82.64319                      NA                             NA
##   lowest_quint_income_share.csv maternal_mortality
## 1                            NA                 NA
## 2                            NA                 NA
## 3                            NA                 NA
## 4                            NA                 NA
## 5                          8.85                  0
## 6                            NA                 NA
##   percent_children_malaria_nets unmet_family_planning_need
## 1                            NA                         NA
## 2                            NA                         NA
## 3                          25.9                         NA
## 4                            NA                         NA
## 5                            NA                       12.9
## 6                            NA                         NA
##   urban_pop_in_slums women_men_literacy_ratio net_migration_per_1000
## 1                 NA                       NA                 -0.875
## 2               62.7                       NA                 -5.773
## 3               55.5                  0.83211                  0.795
## 4                 NA                       NA                     NA
## 5                 NA                       NA                -14.443
## 6                 NA                       NA                     NA
##          name region_num        region
## 1       Aruba          2     caribbean
## 2 Afghanistan          1          asia
## 3      Angola          7 middle_africa
## 4    Anguilla          2     caribbean
## 5     Albania          5        europe
## 6     Andorra          5        europe
```

The data set is collected from http://data.un.org/Explorer.aspx?d=WHO. It consists of 214 rows, where each row represents the data from a different country. A model to predict the fertility rate will be built using the following variables:

- carbon_dioxide_emissions - carbon dioxide emissions in kilotonnes
- cell_subs_per_100 - Cell subscriptions per 100 population(2014)
- employment_to_pop_ratio - employment to total population ratio
- female_employment_to_pop_ratio - female employee to population ratio
- lowest_quint_income_share - poorest quintile's share in income
- maternal_mortality - maternal mortality per 100,000 live births
- percent_children_malaria_nets - percent of children sleeping under insecticide-treated bed nets
- unmet_family_planning_need - percent unmet family planning need
- urban_pop_in_slums - percent urban population living in slums
- women_men_literacy_ratio - women to men parity index, as ratio of literacy rates
- net_migration_per_1000 - net migration rate per 1000
- region_num - number that signifies the region the country resides in
  - 0 - Antarctica
  - 1 - Asia
  - 2 - Caribbean

  - 3 - Central America
  - 4 - Eastern Africa
  - 5 - Europe
  - 6 - European Union
  - 7 - Middle Africa
  - 8 - Middle East
  - 9 - North America
  - 10 - Northern Africa
  - 11 - Oceania
  - 12 - South America
  - 13 - South Africa
  - 14 - Western Africa

The characterizations of the regions was taken from https://internetworldstats.com/list1.htm.

A summary of the data can be seen below:

```
##  country_code       fertility_rate_children_per_woman
##  Length:214         Min.   :1.192
##  Class :character   1st Qu.:1.746
##  Mode  :character   Median :2.309
##                     Mean   :2.835
##                     3rd Qu.:3.690
##                     Max.   :7.599
##
##  carbon_dioxide_emissions cell_subs_per_100 employment_to_pop_ratio
##  Min.   :     83          Min.   :  0.00    Min.   :29.40
##  1st Qu.:  22989          1st Qu.: 74.58    1st Qu.:49.45
##  Median :  50573          Median :105.76    Median :56.90
##  Mean   : 329279          Mean   :106.21    Mean   :55.25
##  3rd Qu.: 316745          3rd Qu.:132.12    3rd Qu.:61.25
##  Max.   :5375003          Max.   :322.59    Max.   :86.90
##  NA's   :172              NA's   :12        NA's   :107
##  female_employment_to_pop_ratio lowest_quint_income_share.csv
##  Min.   :11.20                  Min.   :3.200
##  1st Qu.:38.15                  1st Qu.:4.218
##  Median :47.00                  Median :5.755
##  Mean   :46.27                  Mean   :5.910
##  3rd Qu.:54.70                  3rd Qu.:7.430
##  Max.   :79.50                  Max.   :8.910
##  NA's   :111                    NA's   :190
##  maternal_mortality percent_children_malaria_nets
##  Min.   :  0.00     Min.   : 0.00
##  1st Qu.: 21.05     1st Qu.:15.20
##  Median : 65.00     Median :35.70
##  Mean   :137.57     Mean   :32.68
##  3rd Qu.:176.00     3rd Qu.:46.10
##  Max.   :711.00     Max.   :80.60
##  NA's   :139        NA's   :155
##  unmet_family_planning_need urban_pop_in_slums women_men_literacy_ratio
##  Min.   : 1.70              Min.   : 5.50      Min.   :0.4360
##  1st Qu.:10.62              1st Qu.:25.20      1st Qu.:0.9968
##  Median :16.85              Median :43.50      Median :1.0002
##  Mean   :17.89              Mean   :44.55      Mean   :0.9741
##  3rd Qu.:24.48              3rd Qu.:61.50      3rd Qu.:1.0028
##  Max.   :55.90              Max.   :93.30      Max.   :1.1345
##  NA's   :78                 NA's   :133        NA's   :146
##  net_migration_per_1000          name       region_num           region
##  Min.   :-23.129       Afghanistan:  1   1      :37   asia          :37
##  1st Qu.: -3.137       Albania    :  1   6      :27   european_union:27
##  Median : -0.606       Algeria    :  1   2      :22   caribbean     :22
##  Mean   :  1.138       Andorra    :  1   4      :19   eastern_africa:19
##  3rd Qu.:  2.271       Angola     :  1   5      :18   europe        :18
##  Max.   :127.251       Anguilla   :  1   14     :17   western_africa:17
##  NA's   :20            (Other)    :208   (Other):74   (Other)       :74
```

The number of countries in each region is found.

A challenge of the data set is the large number of missing values. Carbon dioxide emissions is missing 172 values, cellular subscriptions per 100 population is missing 12 values, employment to population ratio is missing 107 values, female employment to population ratio is missing 111 values, lowest quintile share in income is missing 190 values, maternal mortality per 100,000 live births is missing 139 values, percent of children sleeping under insecticide treated bed nets is missing 155 values, unmet family planning need is missing 78 values, percentage of urban population living in slums in missing 133 values, the literacy ratio between women to men is missing 146 values, and the net migration per 1000 is missing 20 values.

Percent of Children Sleeping Under Insecticide Treated Bed Nets

There are 155 missing values. The likelihood of these values being zero We will be investigated by exploring the relationship between this variable and region.

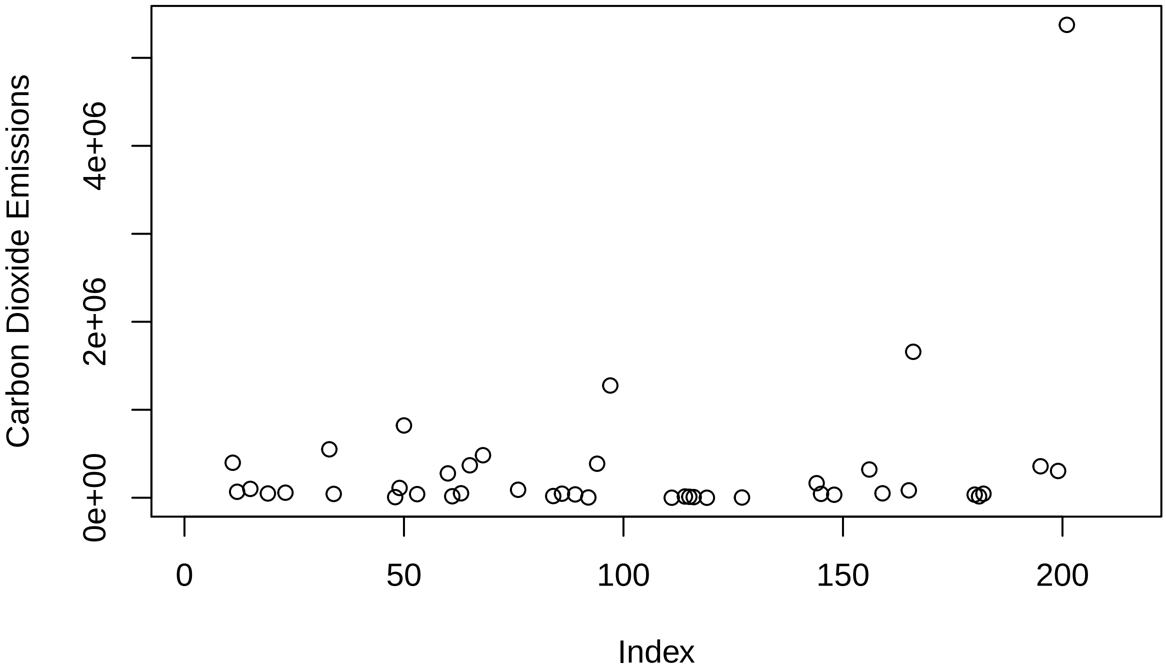
```
##    region_num          region NumCountriesPerRegion NumCountriesNA
## 1           0      antarctica                     0              0
## 2           1            asia                    37             27
## 3           2       caribbean                    22             21
## 4           3 central_america                     8              7
## 5           4  eastern_africa                    19              4
## 6           5          europe                    18             18
## 7           6  european_union                    27             27
## 8           7   middle_africa                     9              0
## 9           8     middle_east                    14             13
## 10          9   north_america                     3              3
## 11         10 northern_africa                     6              5
## 12         11         oceania                    16             14
## 13         12   south_america                    13             11
## 14         13 southern_africa                     5              3
## 15         14  western_africa                    17              2
##    PerentageCountriesInRegion
## 1                         NaN
## 2                    72.97297
## 3                    95.45455
## 4                    87.50000
## 5                    21.05263
## 6                   100.00000
## 7                   100.00000
## 8                     0.00000
## 9                    92.85714
## 10                  100.00000
## 11                   83.33333
## 12                   87.50000
## 13                    84.61538
## 14                    60.00000
## 15                    11.76471
```

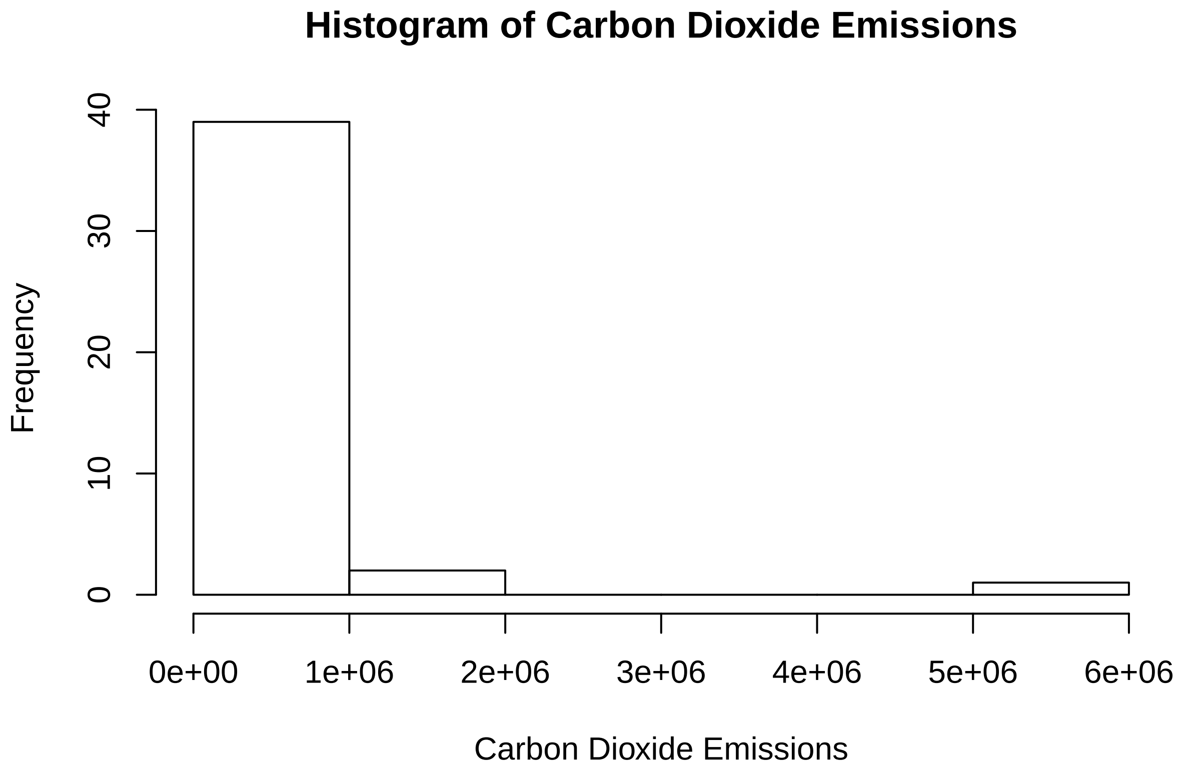
The data table above displays the percentage of countries in each region that are missing values for the percentage of children sleeping under insecticide treated bed nets. The average value for the percentage of children sleeping under insecticide treated bed nets in each region is calculated.

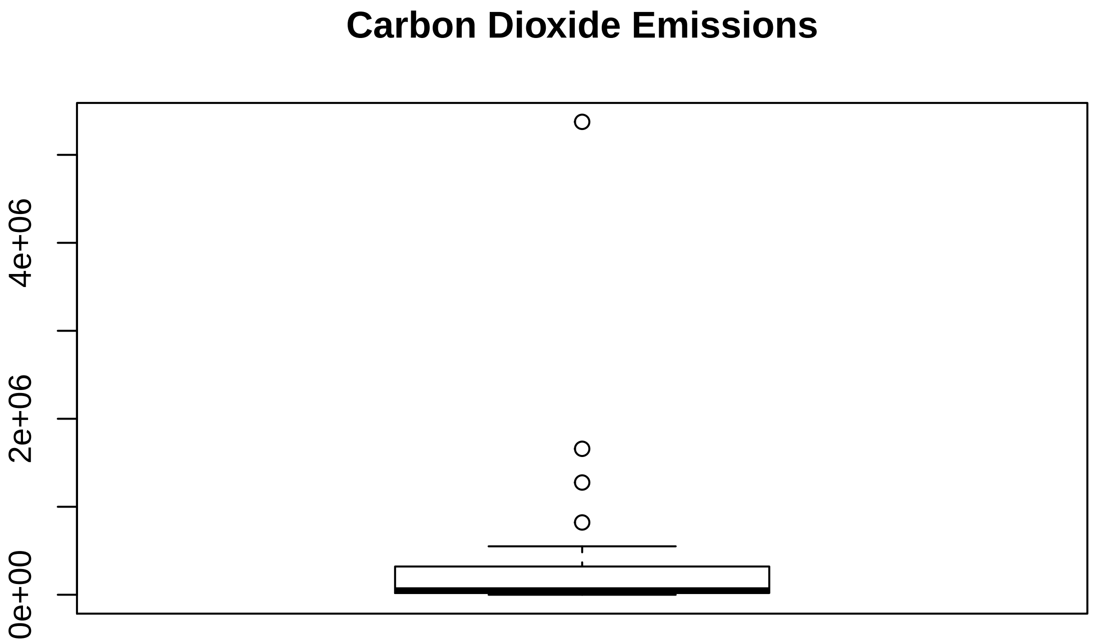
```
## # A tibble: 11 x 3
## # Groups:   region_num [11]
##    region_num region          AvgBedNetsInRegion
##    <fct>      <fct>                        <dbl>
##  1 7          middle_africa                32.6
##  2 1          asia                         11.7
##  3 4          eastern_africa               42.9
##  4 14         western_africa               44.5
##  5 3          central_america               1
##  6 12         south_america                33.9
##  7 2          caribbean                    12
##  8 8          middle_east                   0
##  9 13         southern_africa               3.55
## 10 10         northern_africa              28
## 11 11         oceania                      45.5
```

We will assume that countries that have missing data for the percentage of children sleeping under insecticide treated nets, have a percentage of zero and we will impute zero for the missing values.

Exploring Carbon Dioxide Emissions

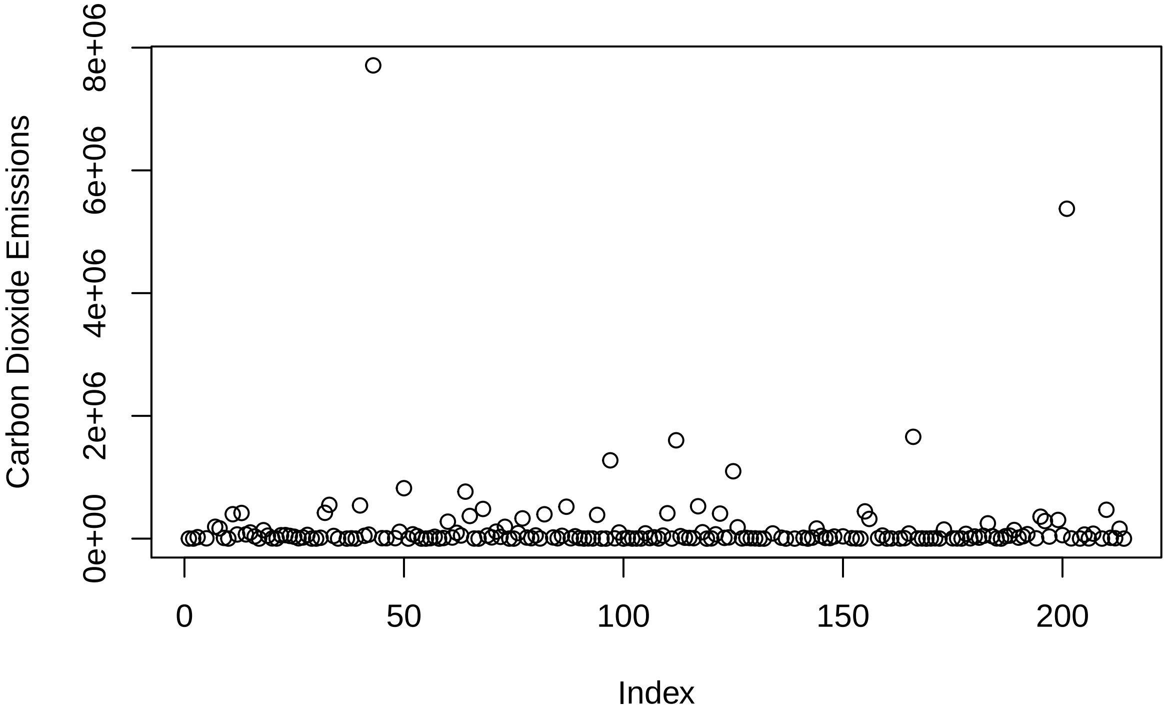

Carbon Dioxide Emissions
0e+00
2e+06
4e+06
0
50
100
150
200
Index

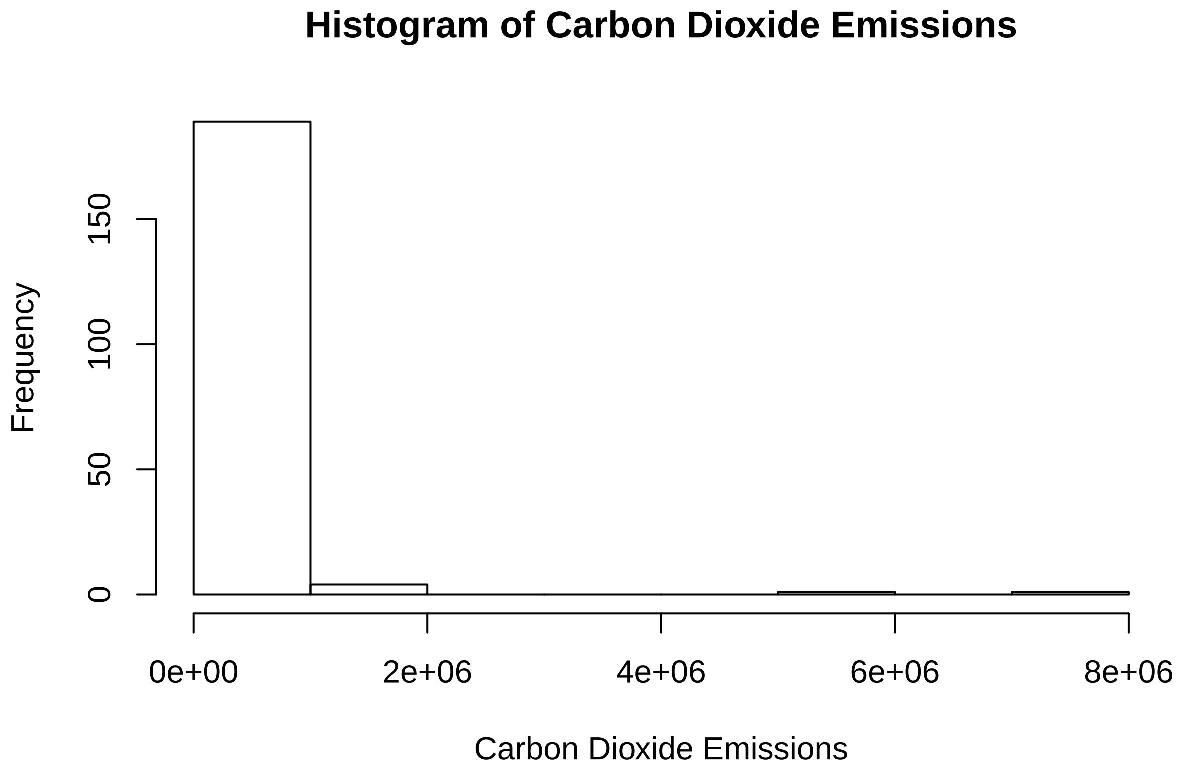

Histogram of Carbon Dioxide Emissions
Frequency
0
10
20
30
40
0e+00
1e+06
2e+06
3e+06
4e+06
5e+06
6e+06
Carbon Dioxide Emissions

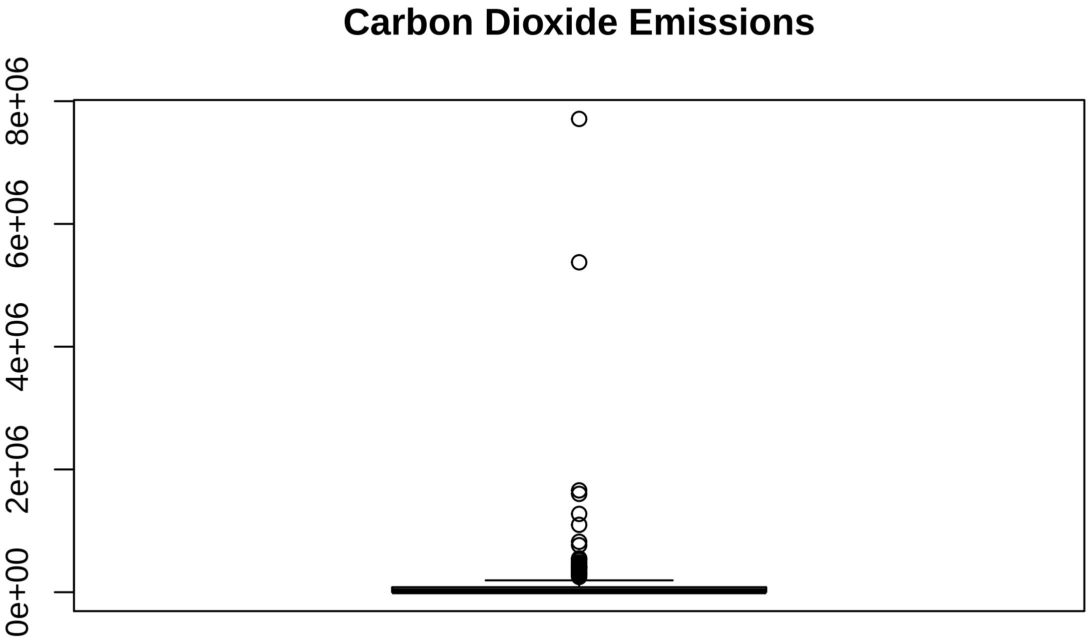


```
##   country_code fertility_rate_children_per_woman carbon_dioxide_emissions
## 1          USA                             1.877                  5375003
##   cell_subs_per_100 employment_to_pop_ratio female_employment_to_pop_ratio
## 1          98.40686                    58.6                           53.2
##   lowest_quint_income_share.csv maternal_mortality
## 1                            NA                 NA
##   percent_children_malaria_nets unmet_family_planning_need
## 1                             0                          8
##   urban_pop_in_slums women_men_literacy_ratio net_migration_per_1000
## 1                 NA                       NA                  3.335
##            name region_num        region
## 1 United States          9 north_america
```

There are 172 missing values for carbon dioxide emissions. Of the data that is present, most is between 0 and 1,000,000 kilotonnes. The United States is an outlier with a very high emission value. This value does not take into account the size of the country or population. Imputing values presents a significant challenge for this reason as well as the variation between countries' emissions that depends on their development and commitment to environmental action. To address this, values reported by the Guardian in 2011 wil be used to supplement the data. The data can be found here: https://www.theguardian.com/news/datablog/2011/jan/31/world-carbon-dioxide-emissions-country-data-co2

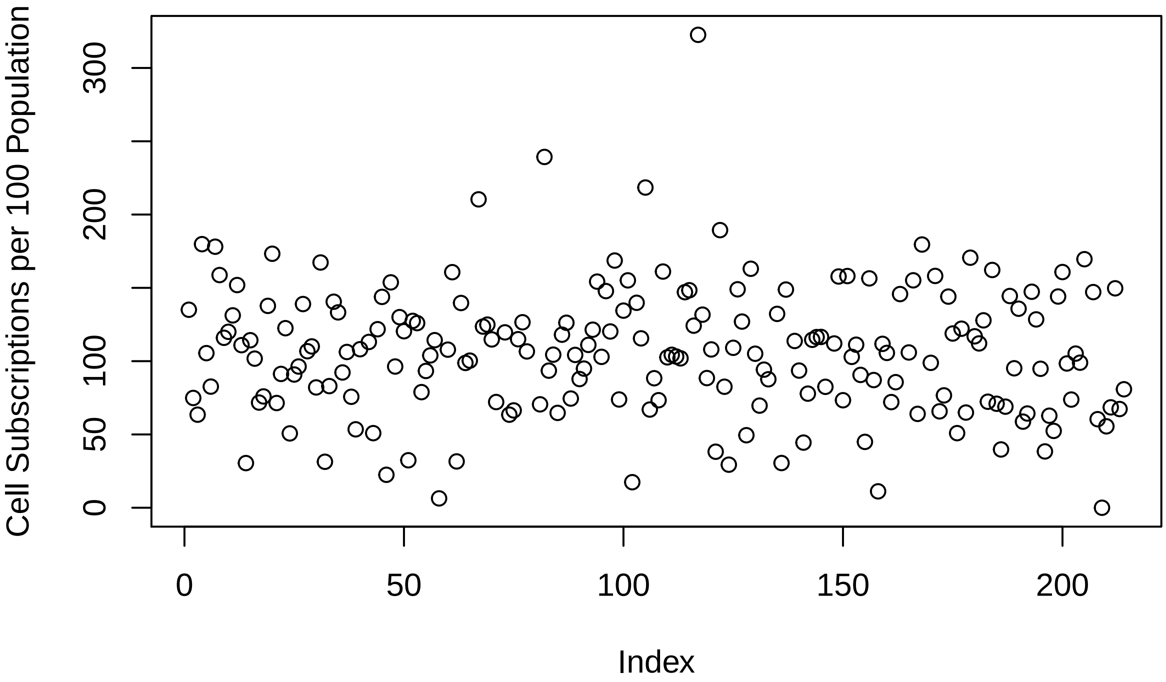

Carbon Dioxide Emissions
0e+00
2e+06
4e+06
6e+06
8e+06
0
50
100
150
200
Index

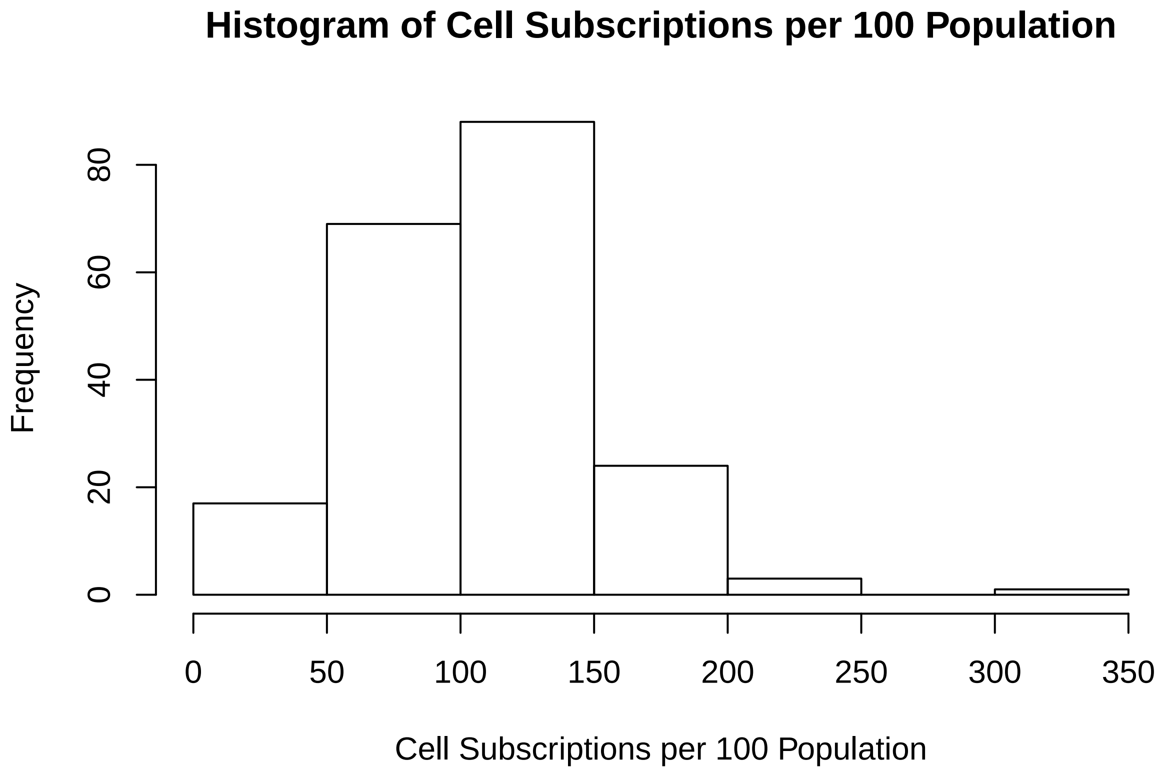

Histogram of Carbon Dioxide Emissions
Frequency
0
50
100
150
0e+00
2e+06
4e+06
6e+06
8e+06
Carbon Dioxide Emissions



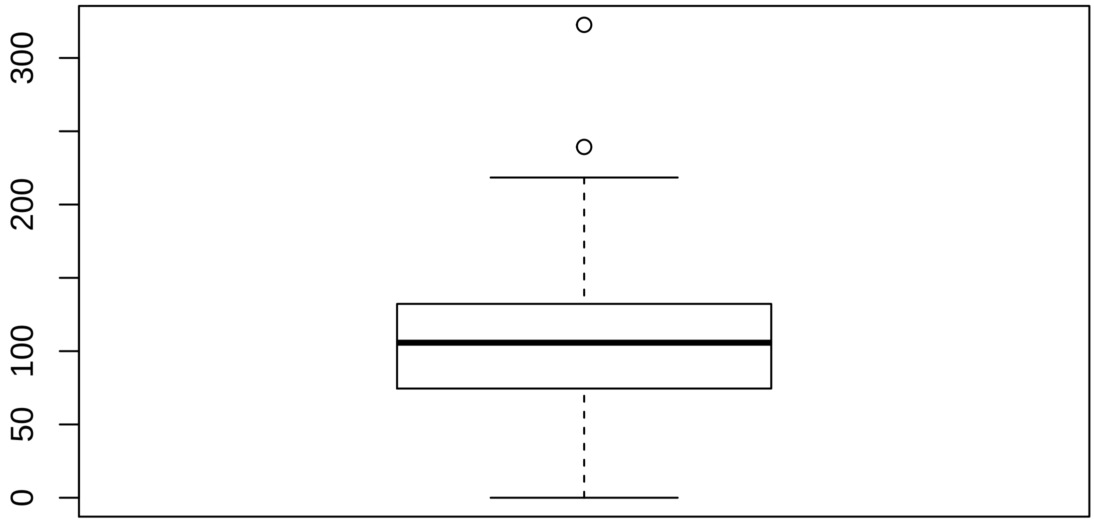
Now there are only 18 missing values for CO2 emissions. Because of the extent to which outliers will increase the mean, the median will be imputed for missing values of CO2 emissions.

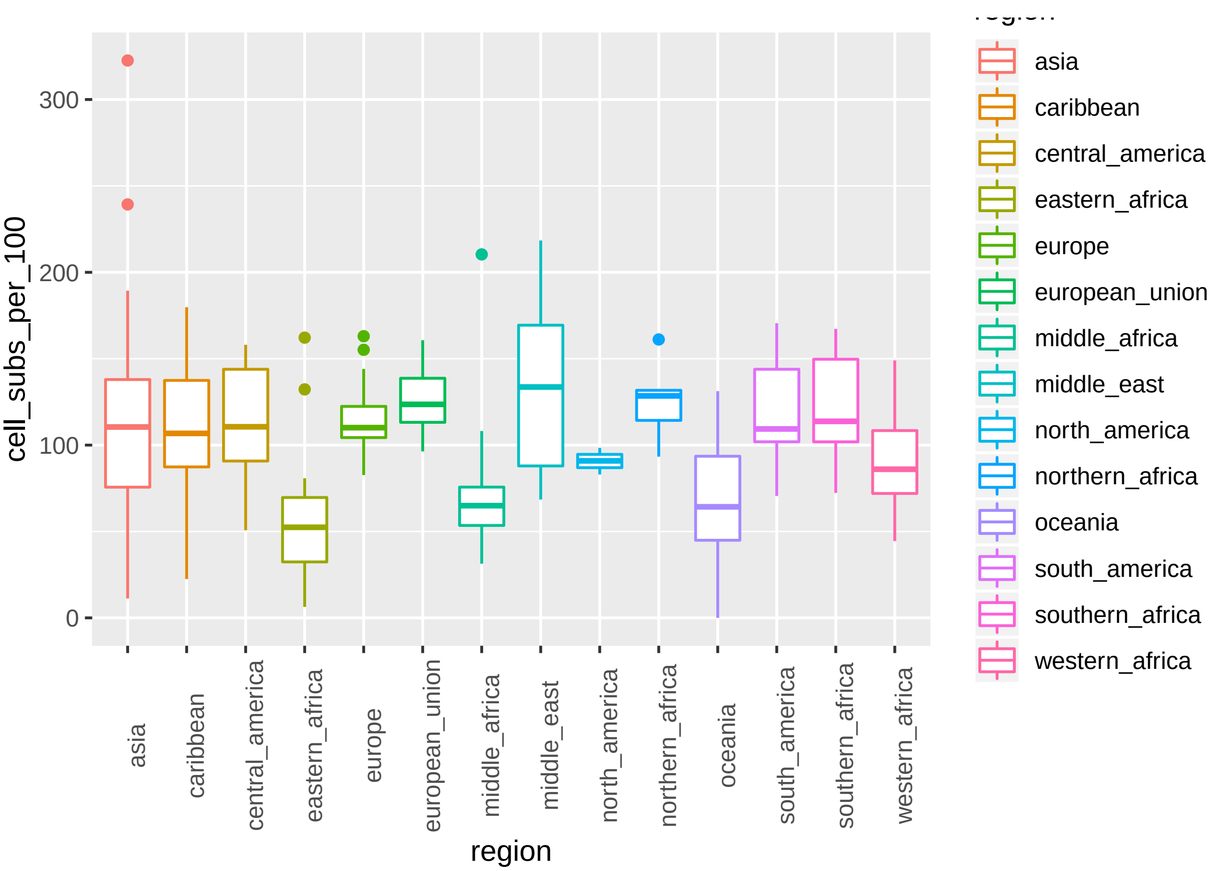
Exploring Cell Subscriptions per 100 Population


Cell Subscriptions per 100 Population
0
50
100
200
300
0
50
100
150
200
Index


Histogram of Cell Subscriptions per 100 Population
Frequency
0
20
40
60
80
0
50
100
150
200
250
300
350
Cell Subscriptions per 100 Population

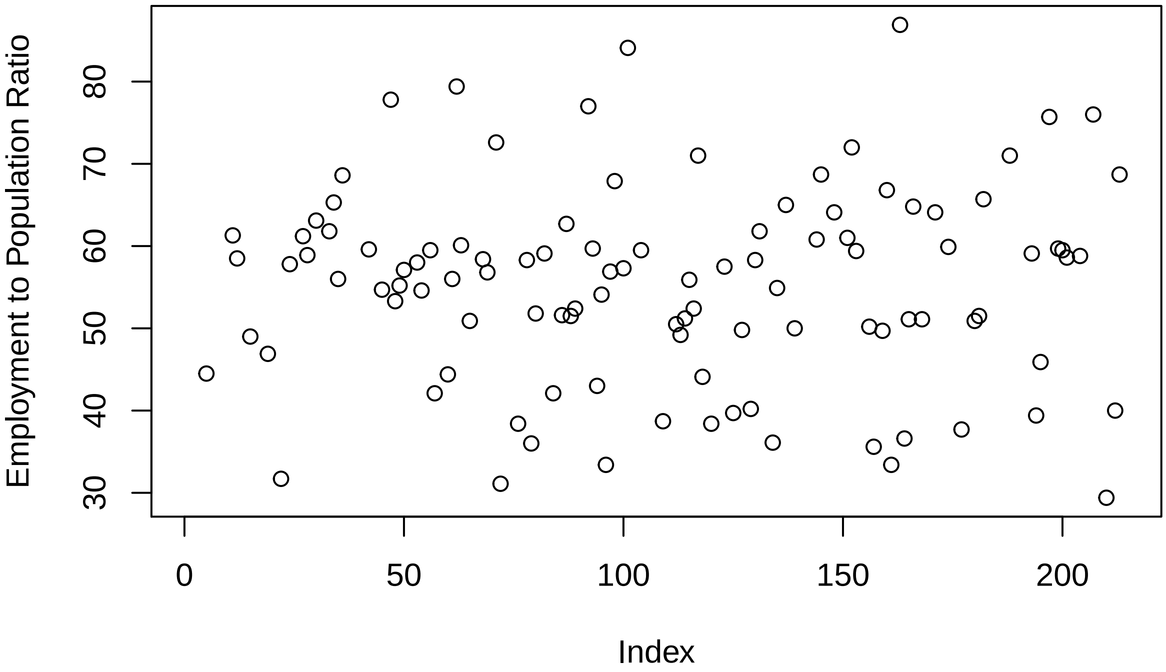
Cell Subscriptions per 100 Population

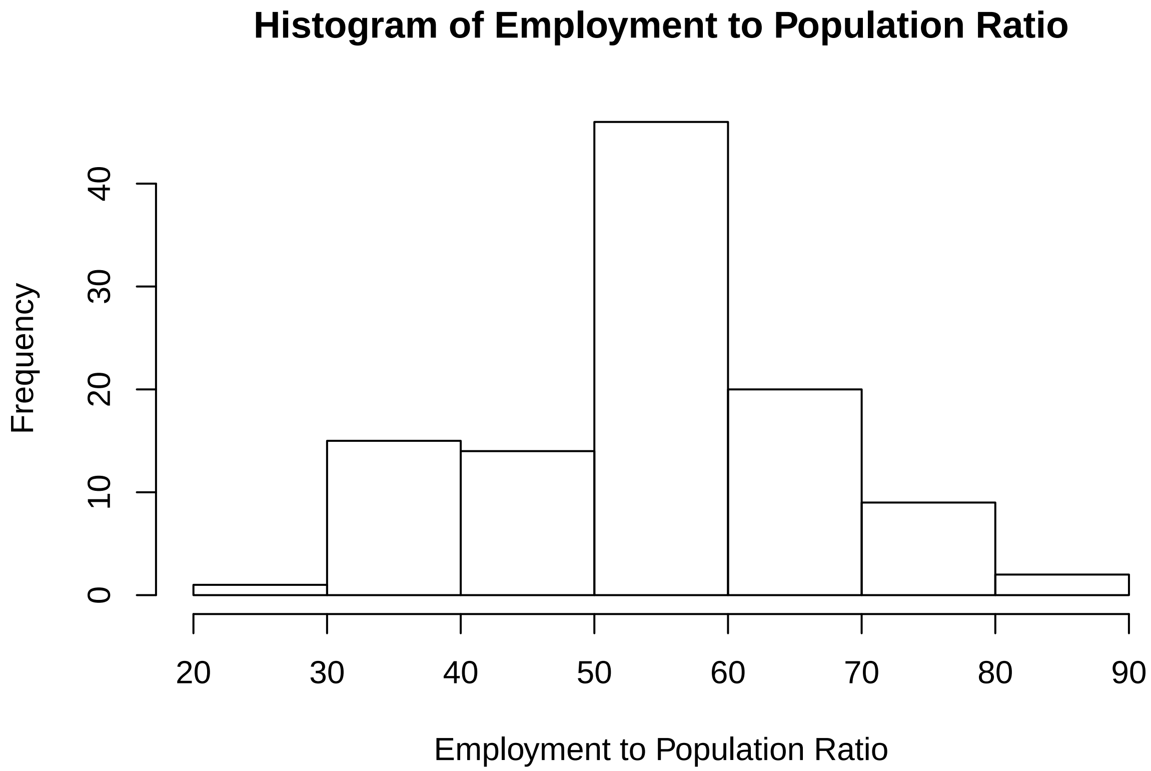


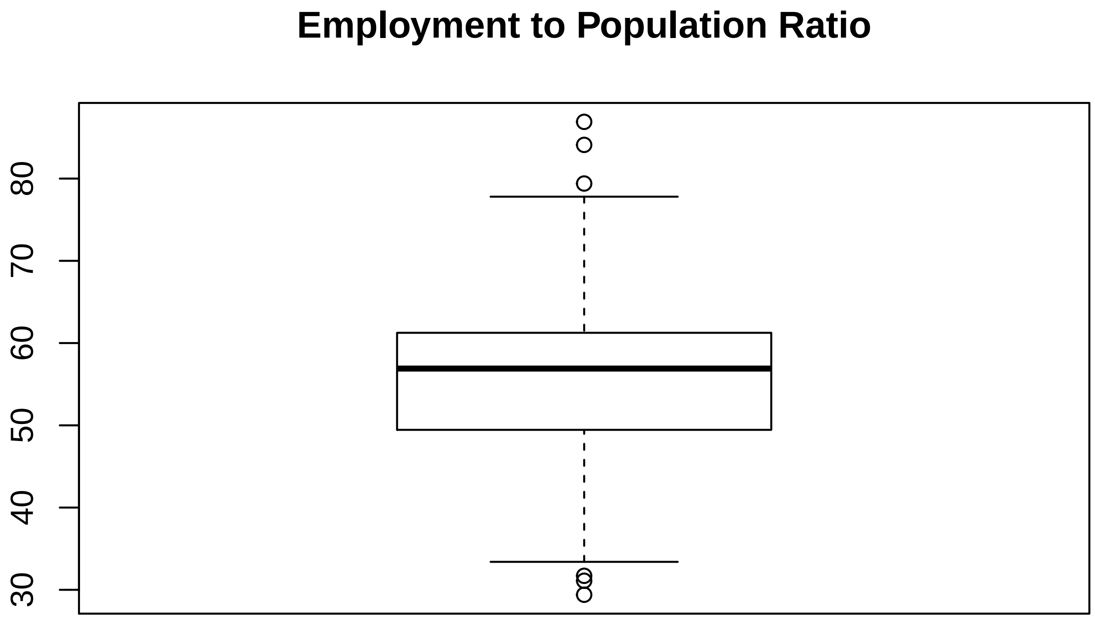


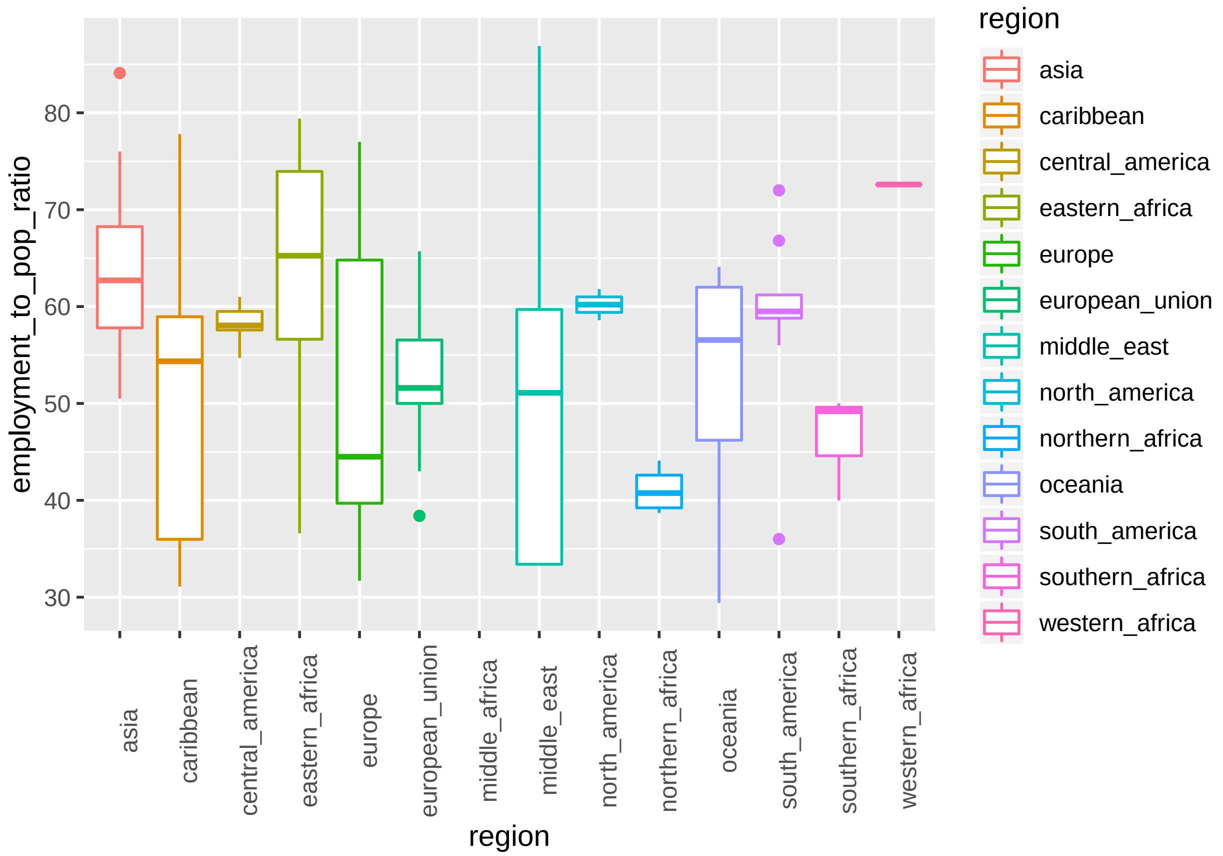
The median and mean of cell subscriptions per 100 population are about the same. However when exploring the variation by region, there are more noticable differences. There are few outliers. The region's median will be imputed for the 18 missing values.

Exploring Employment to Population Ratio


Employment to Population Ratio
30
40
50
60
70
80
0
50
100
150
200
Index

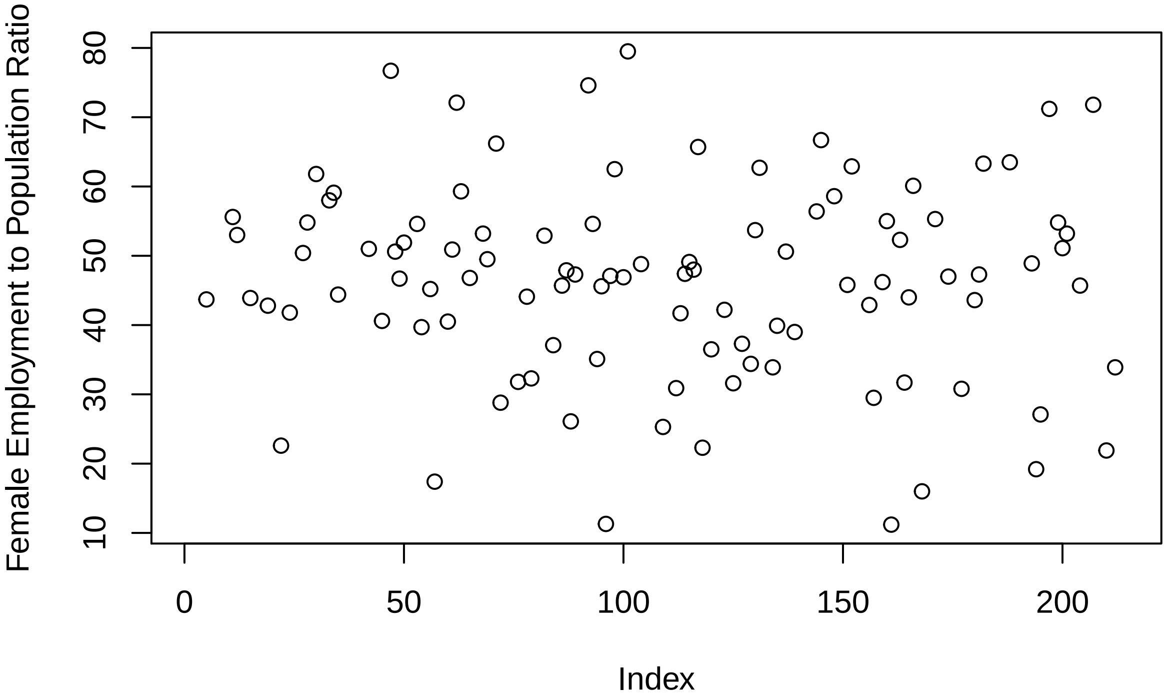

Histogram of Employment to Population Ratio
Frequency
0
10
20
30
40
20
30
40
50
60
70
80
90
Employment to Population Ratio

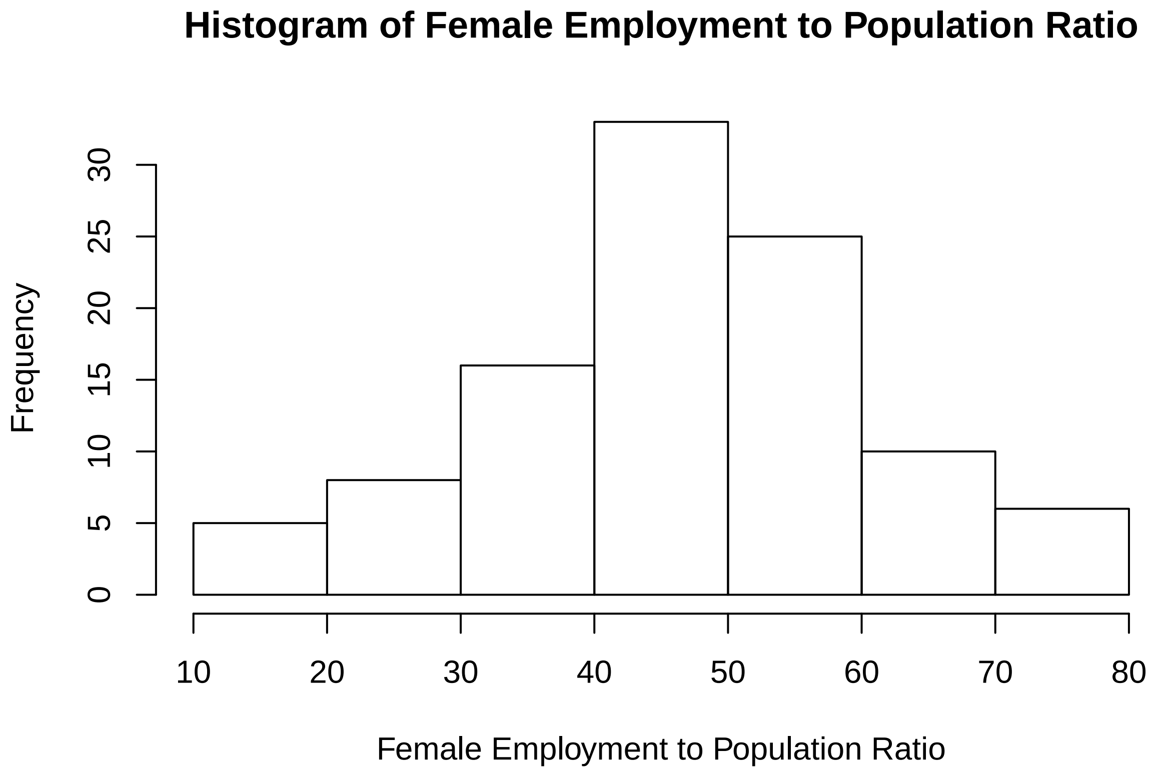

Employment to Population Ratio
30
40
50
60
70
80

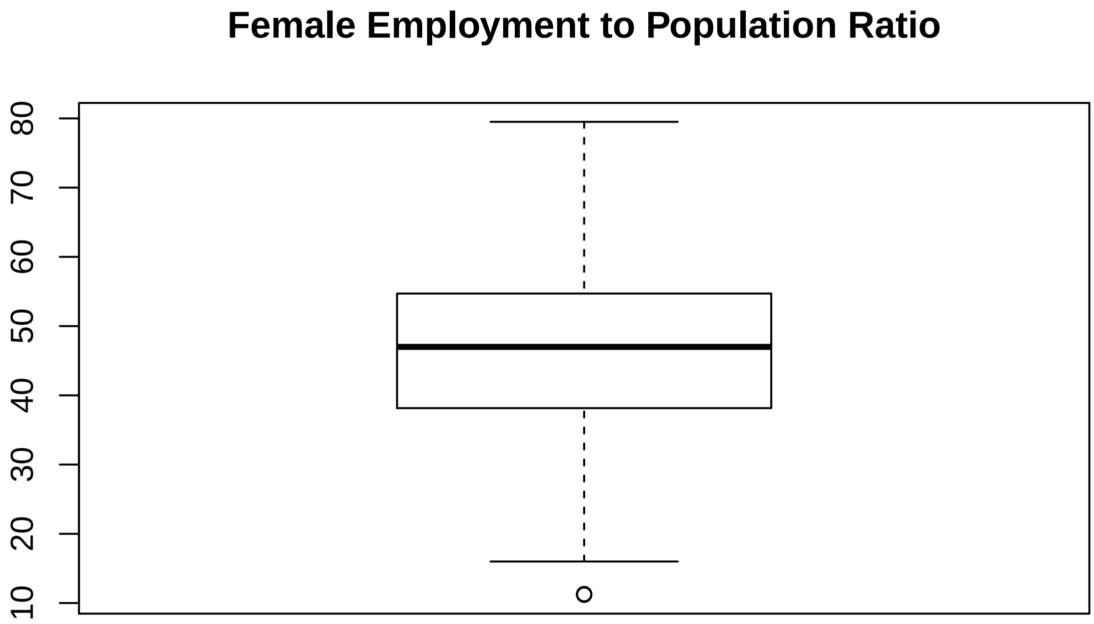


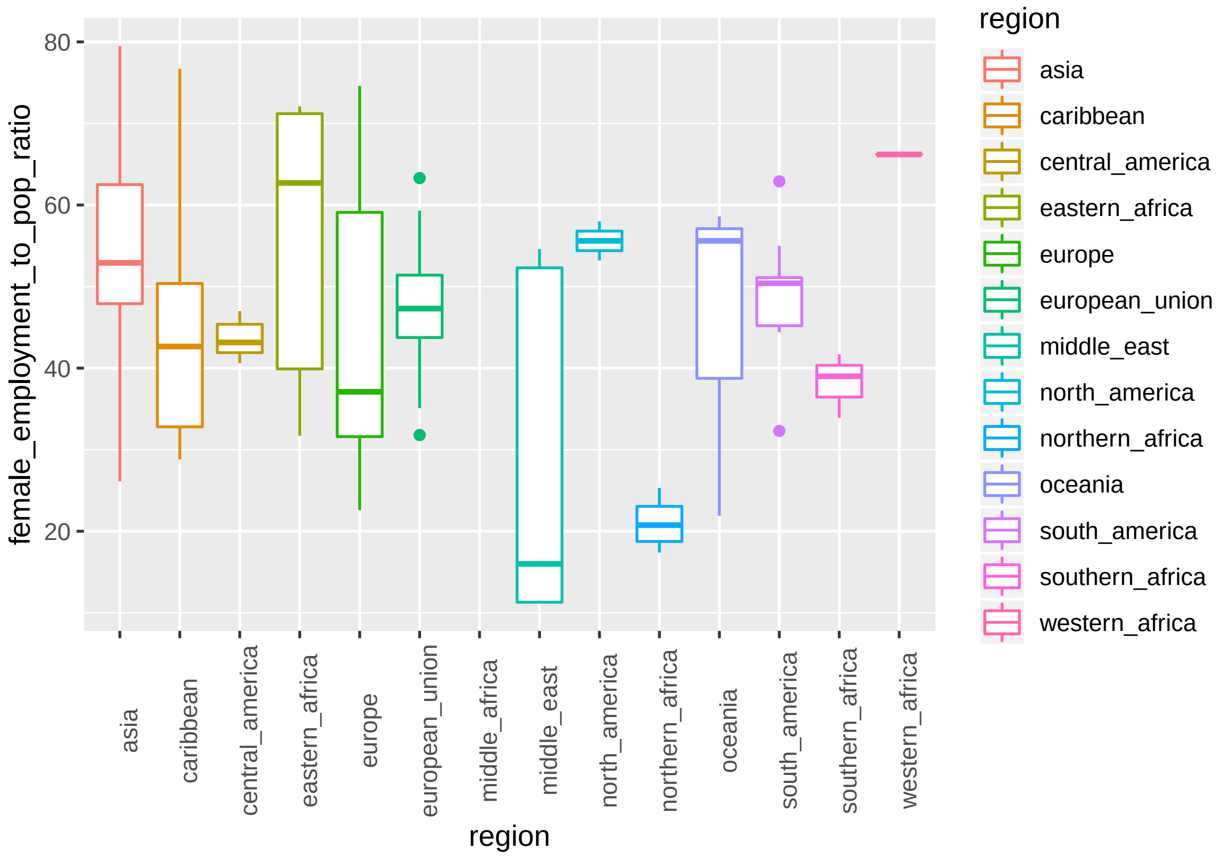
The employment to population value varies dramatically by region. The median for the employment to population ratio per region will be imputed for the missing values. There are no values for middle Africa so the median from eastern Africa will be imputed for countries in middle Africa.

Exploring Female Employment to Population Ratio


Female Employment to Population Ratio
10
20
30
40
50
60
70
80
0
50
100
150
200
Index

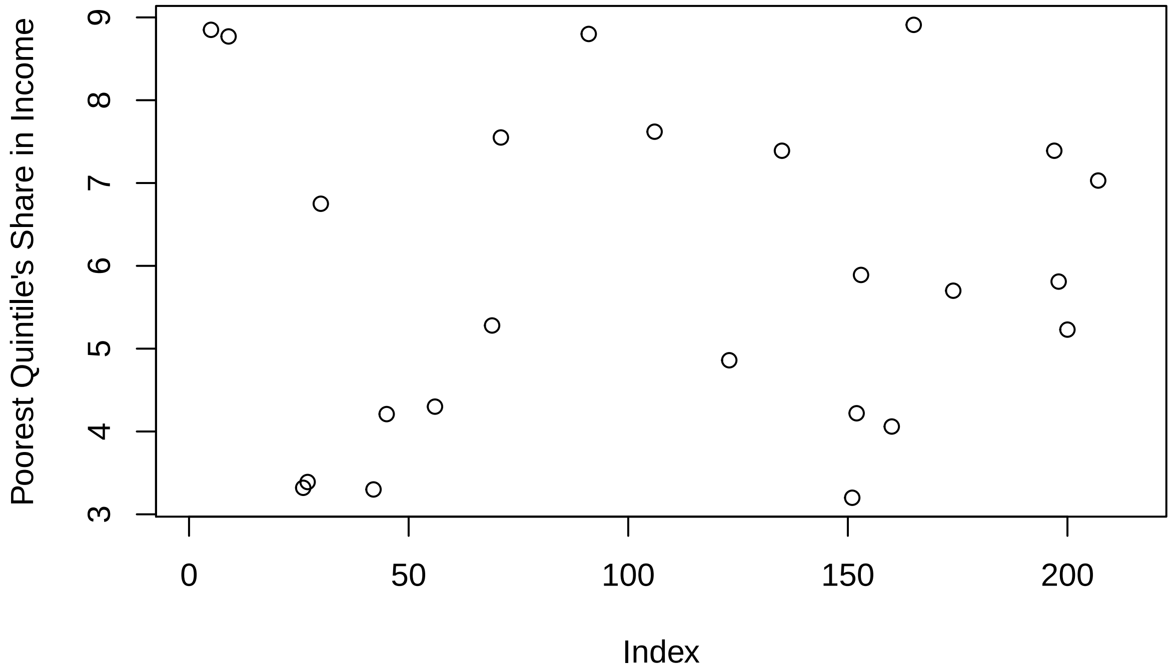

Histogram of Female Employment to Population Ratio
Frequency
0
5
10
15
20
25
30
10
20
30
40
50
60
70
80
Female Employment to Population Ratio

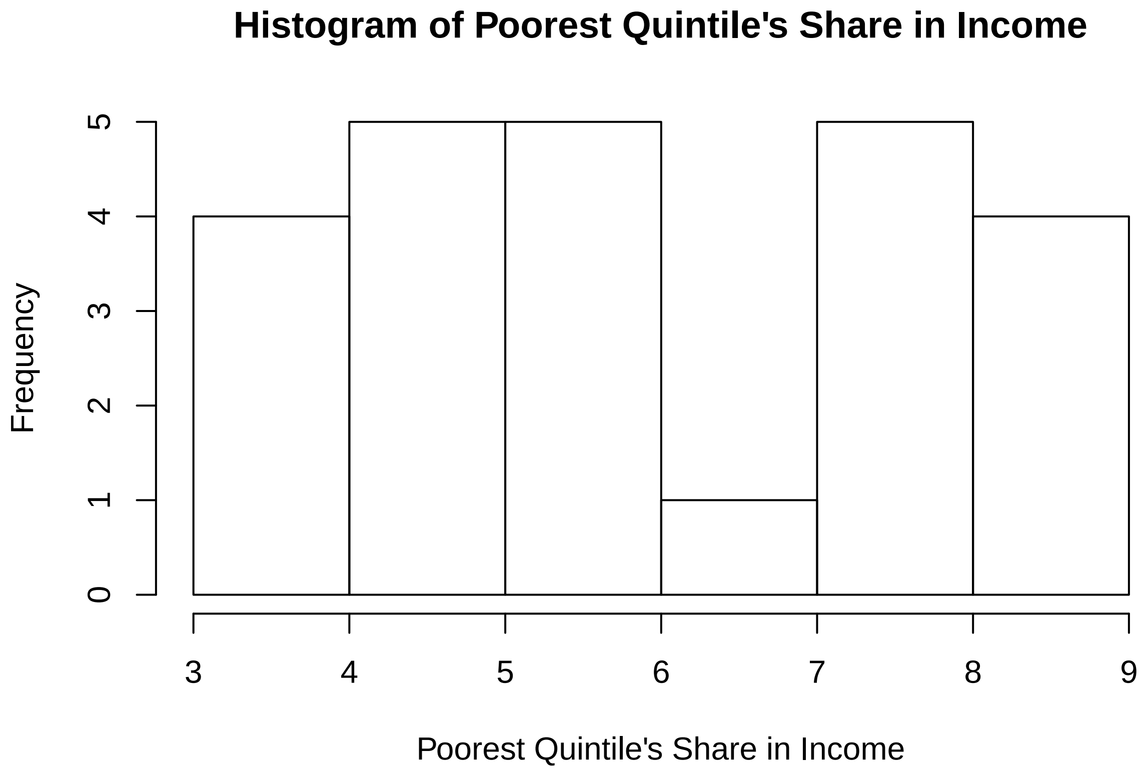

Female Employment to Population Ratio
80
70
60
50
40
30
20
10

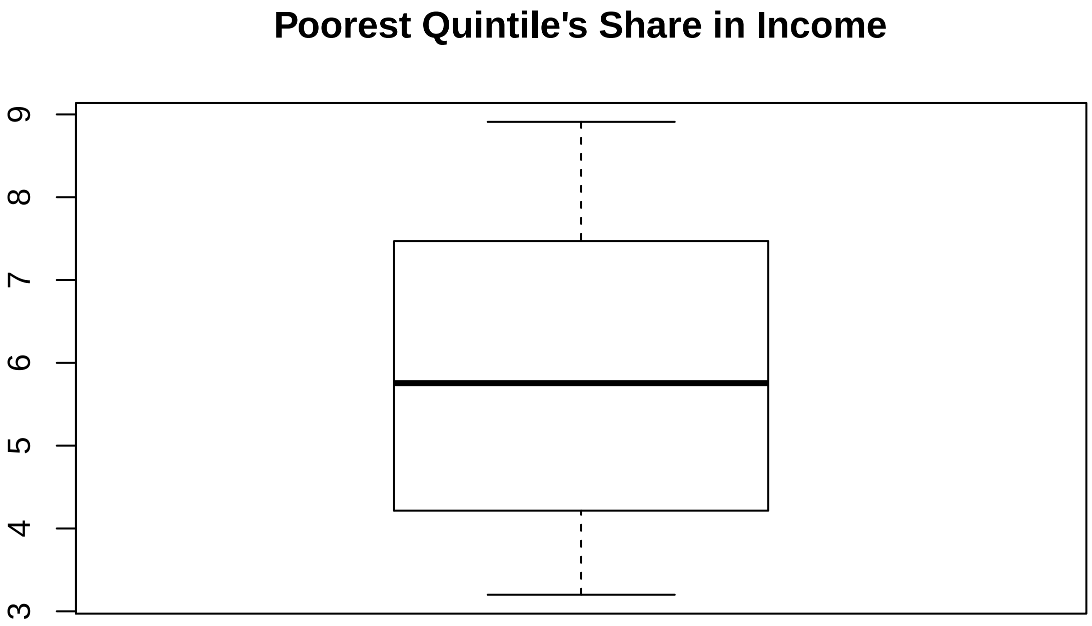


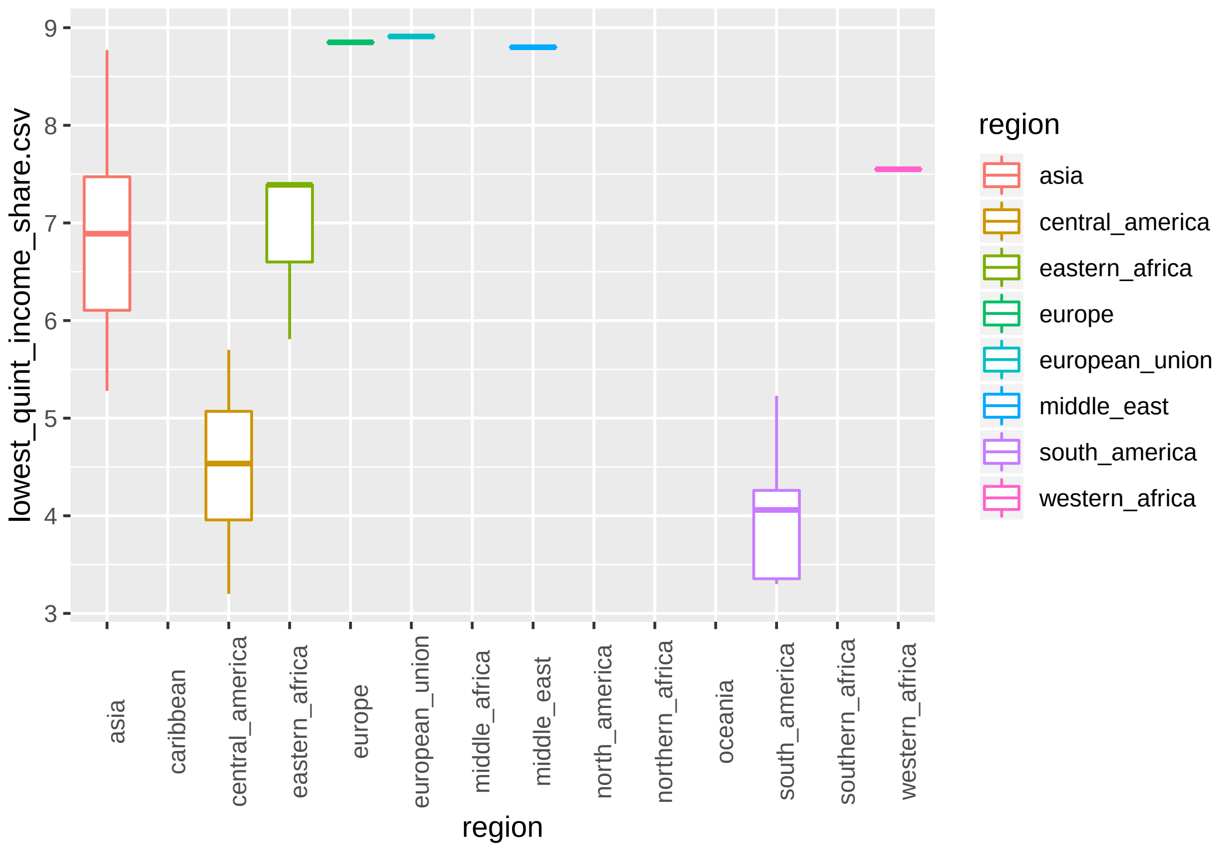
There are 111 missing values. The female employment to population value varies dramatically by region. The median value by region will be imputed for the missing values of female employment to population ratio. There are no values for middle Africa so the median from eastern Africa will be imputed for middle Africa.

Exploring Lowest Quintile Income Share


Poorest Quintile's Share in Income
3
4
5
6
7
8
9
0
50
100
150
200
Index


Histogram of Poorest Quintile's Share in Income
Frequency
0
1
2
3
4
5
3
4
5
6
7
8
9
Poorest Quintile's Share in Income


Poorest Quintile's Share in Income
3
4
5
6
7
8
9

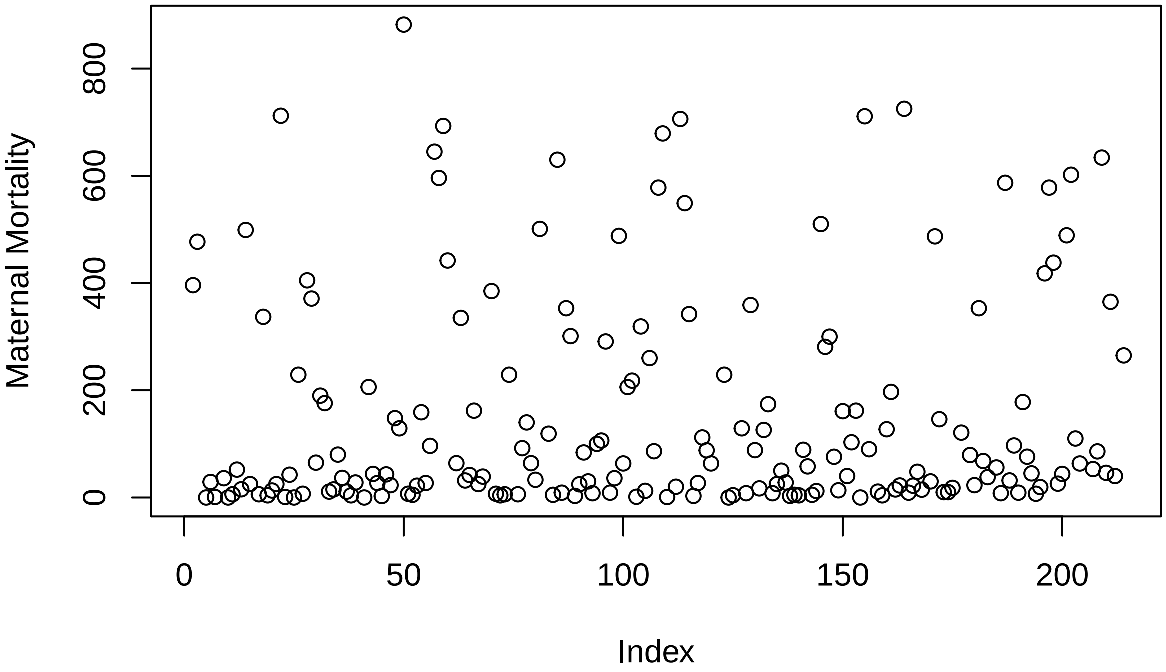


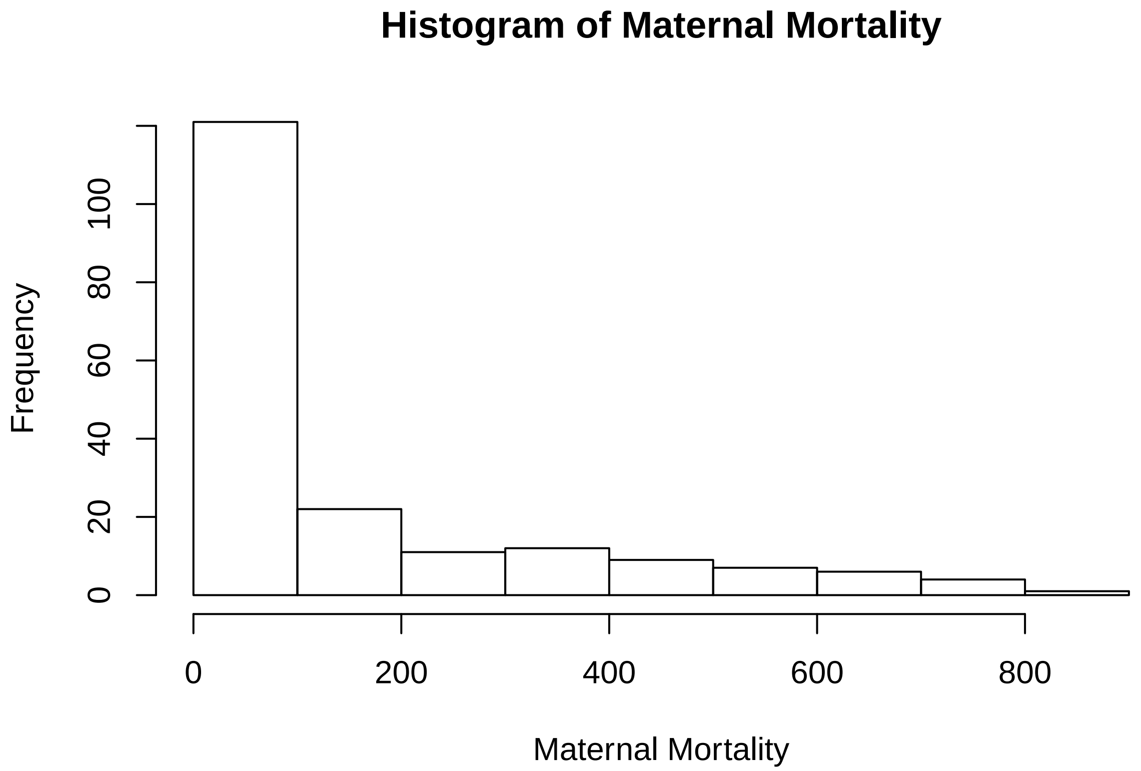
There are 190 missing values. There are so many missing values that there is not a meaningful way to impute missing values. This variable will therefore be removed.

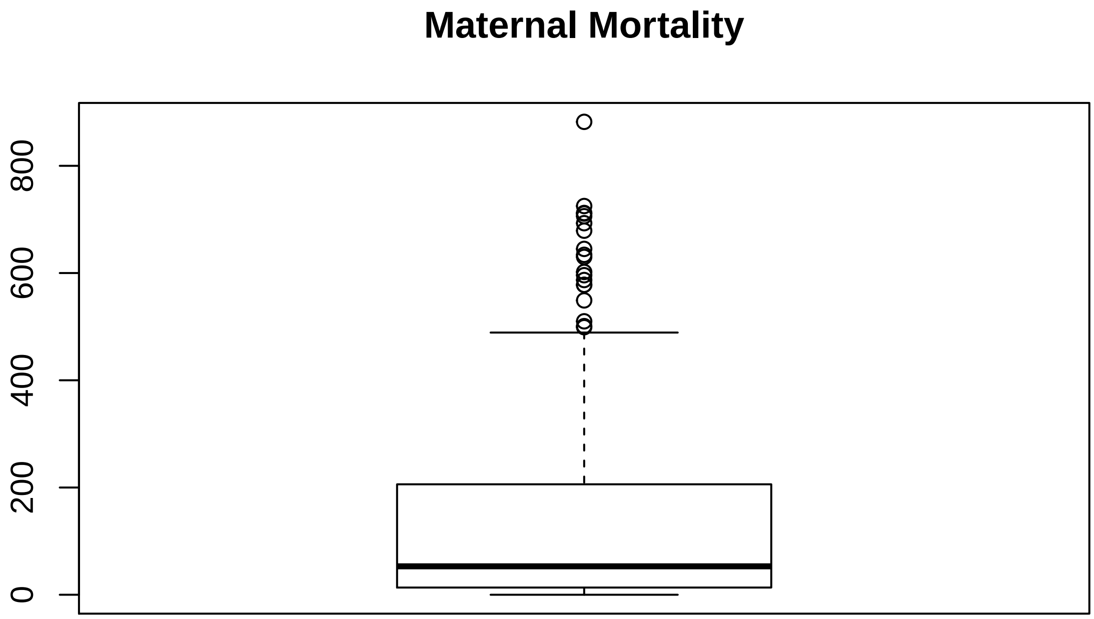
Exploring Maternal Mortality

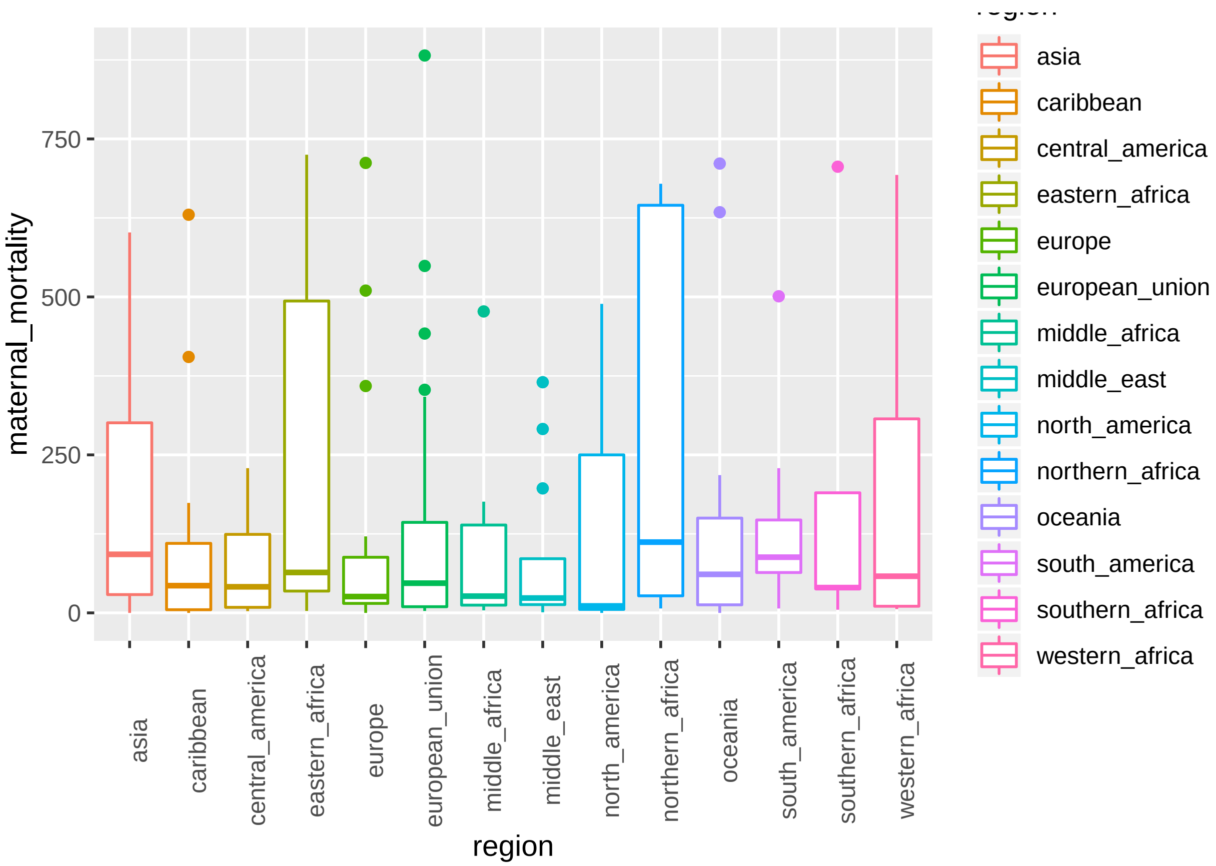
There are 139 missing values. To address this large number of missing values,data will be taken from UNICEF at the following website https://data.unicef.org/topic/maternal-health/maternal-mortality/ and the values posted there for 2015 will be used for the missing values.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##     0.0    13.4    53.0   152.4   206.0   882.0      21 
```


Maternal Mortality
0
200
400
600
800
Index
0
50
100
150
200


Histogram of Maternal Mortality
Frequency
0
20
40
60
80
100
0
200
400
600
800
Maternal Mortality

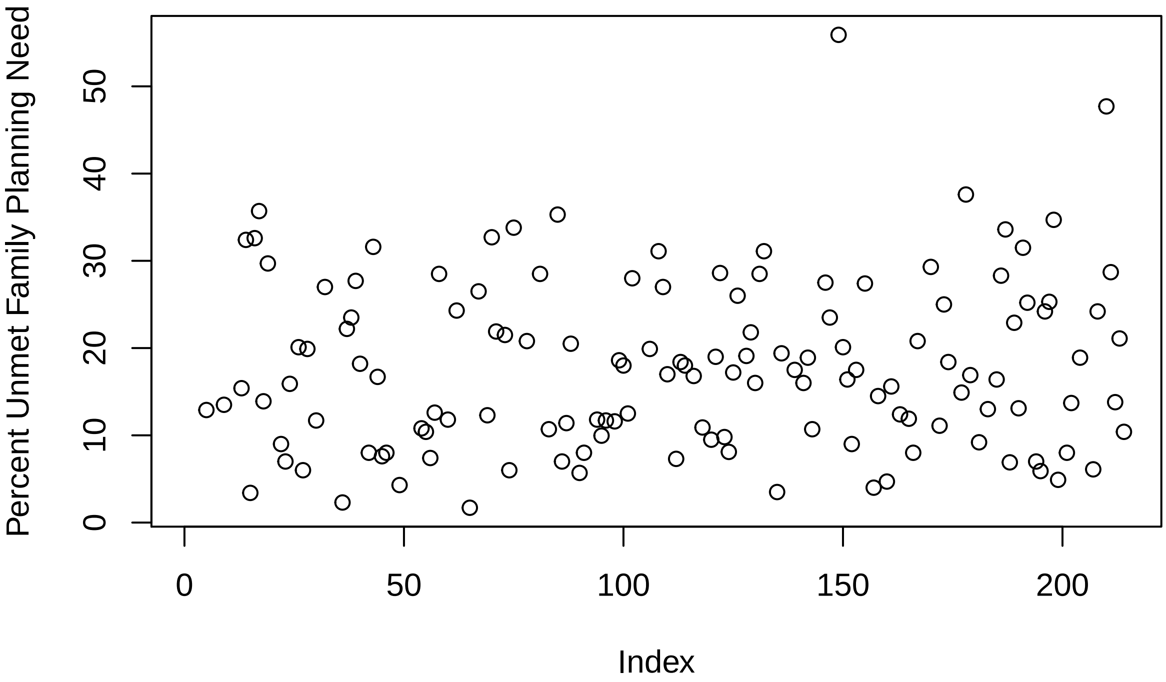

Maternal Mortality
0
200
400
600
800

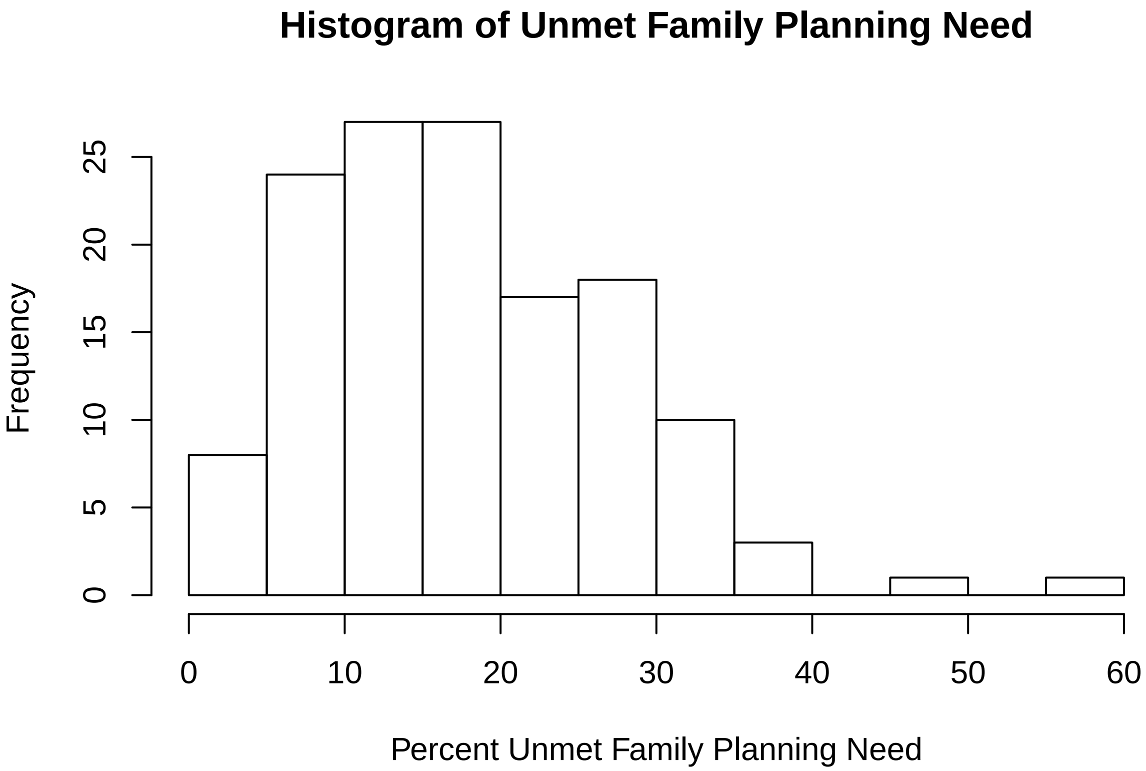


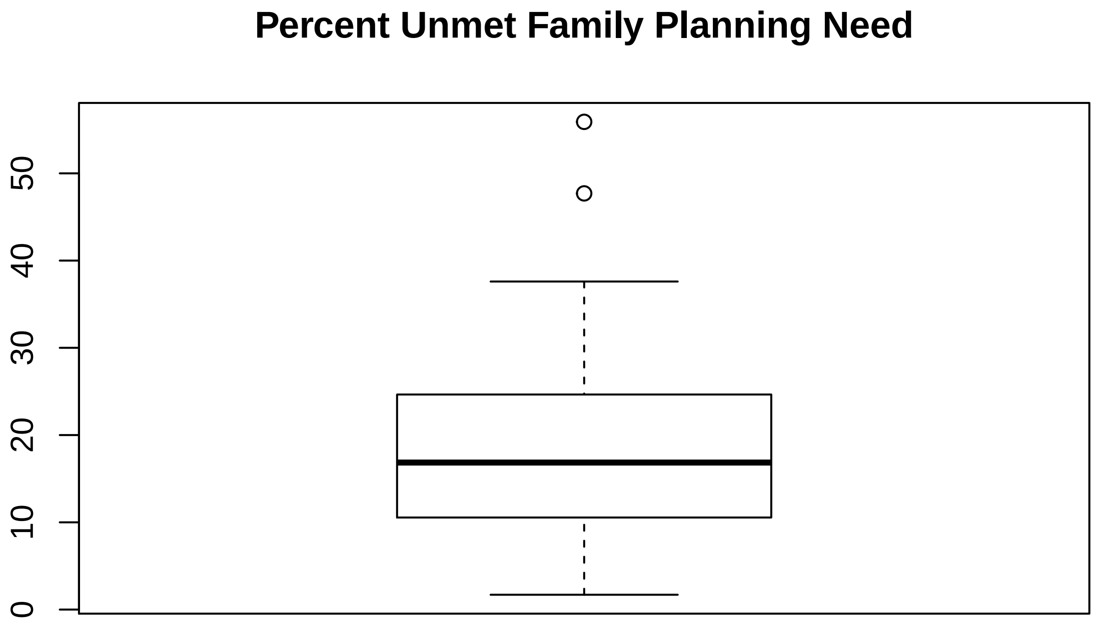
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##     0.0    13.4    53.0   152.4   206.0   882.0      21
```

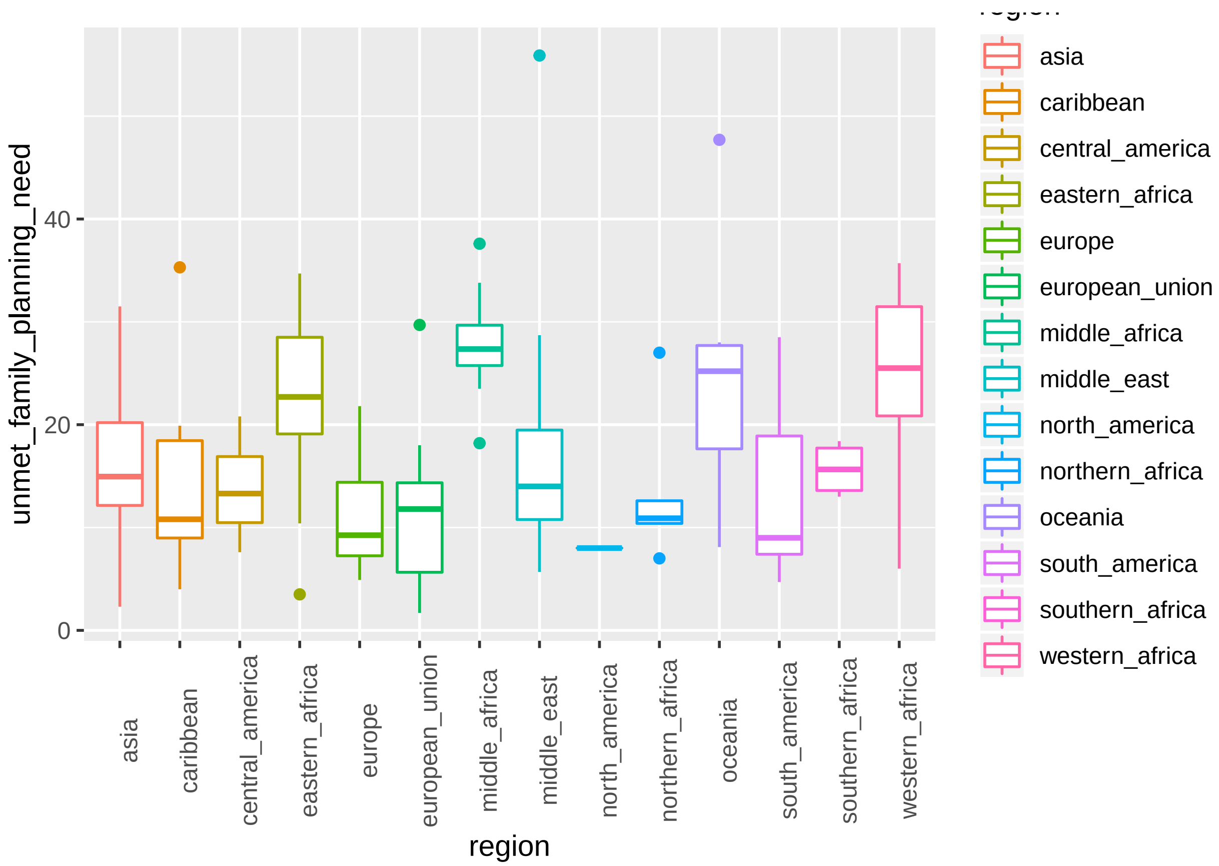
With the added data, now there are only 21 missing values for maternal mortality. The data is skewed to the right, with the presence of outliers above a value of 500. There is a relationship between region and maternal mortality so the missing values will be imputed with the median maternal mortality for the region the country is in.

Exploring Unmet Family Planning Need


Percent Unmet Family Planning Need
0
10
20
30
40
50
0
50
100
150
200
Index

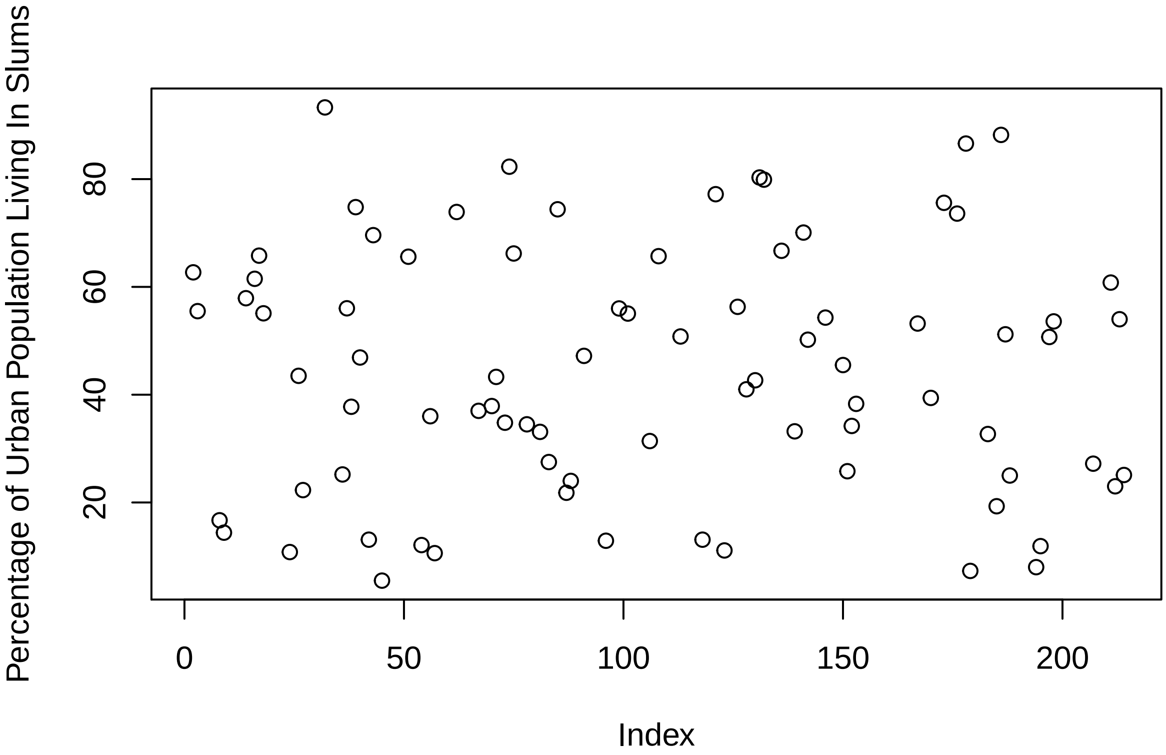

Histogram of Unmet Family Planning Need
Frequency
0
5
10
15
20
25
0
10
20
30
40
50
60
Percent Unmet Family Planning Need

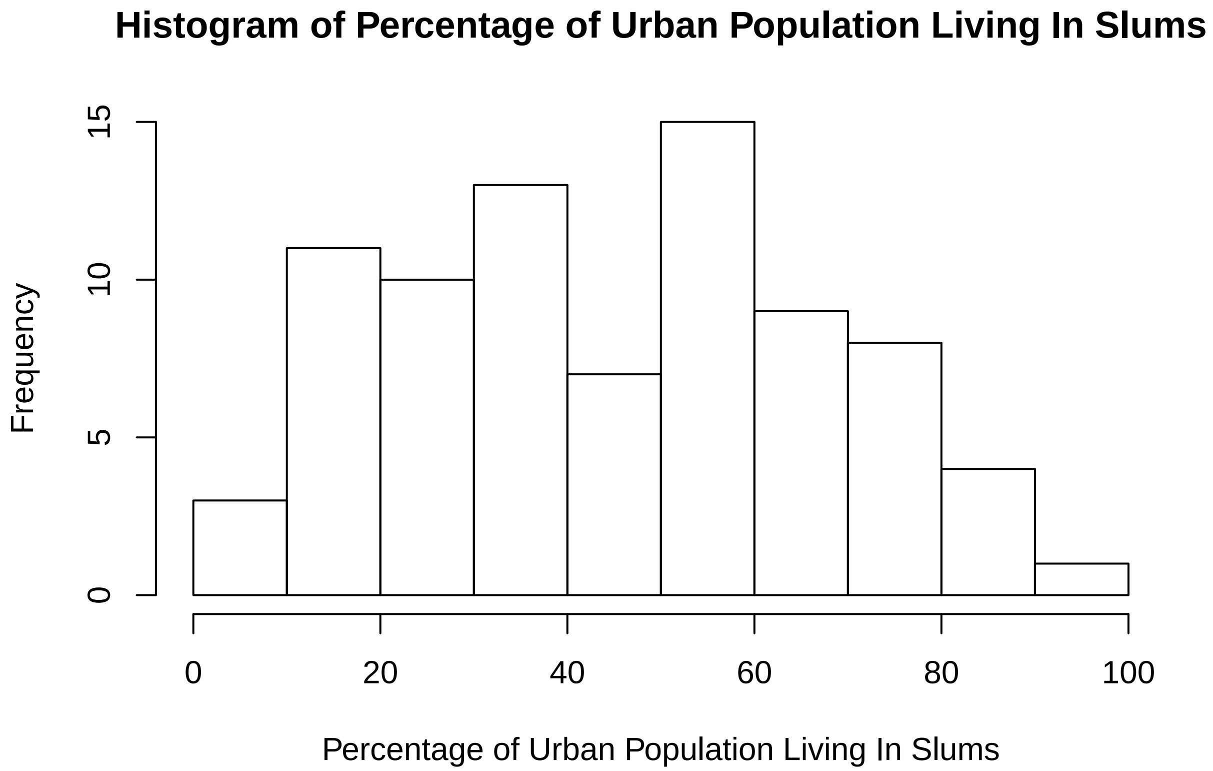

Percent Unmet Family Planning Need
0
10
20
30
40
50

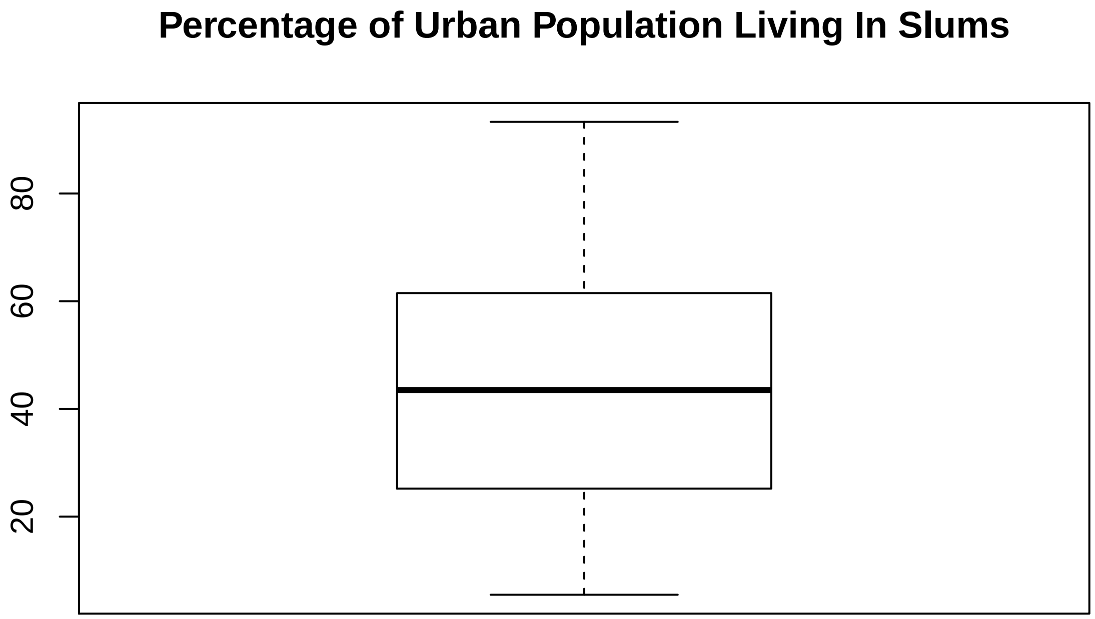


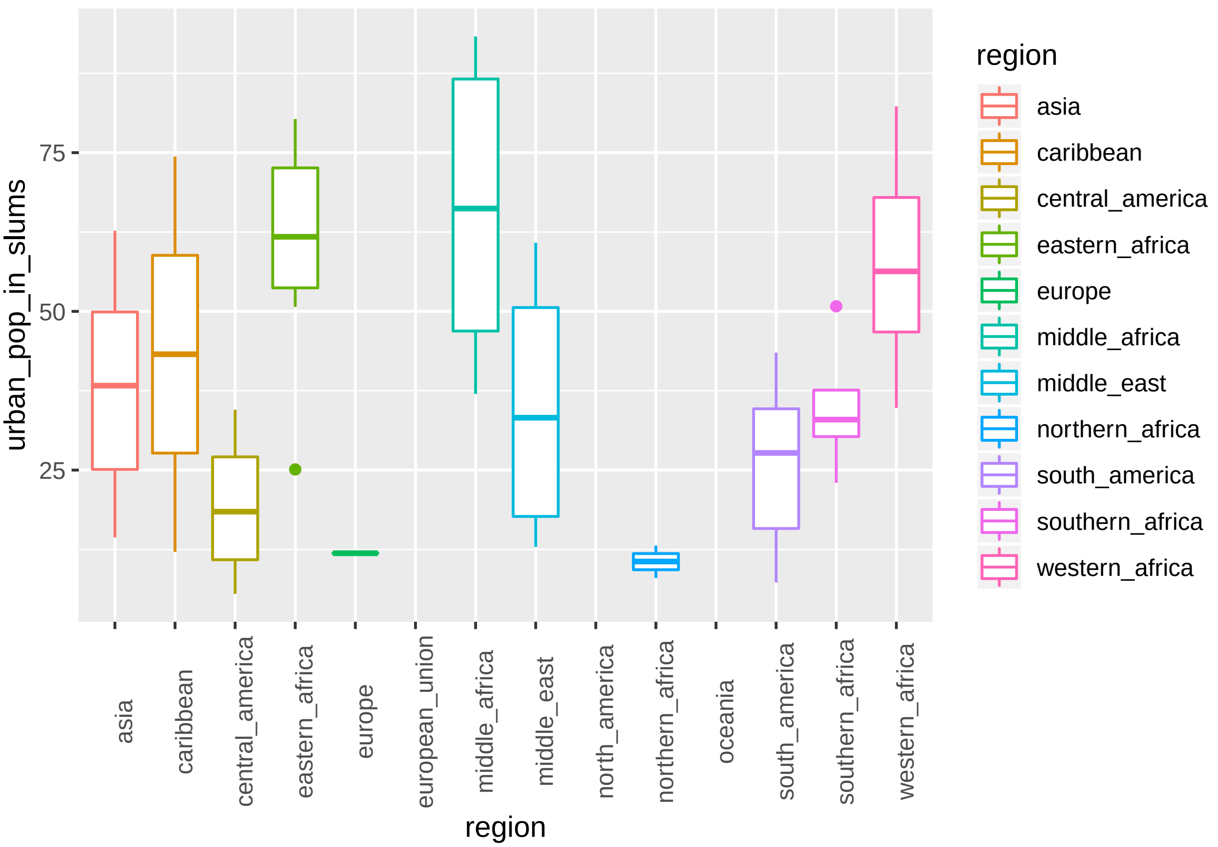
There are 78 missing values. There are differences in median according to region. The median of the region will beimputed for the missing values.

Exploring Percentage of Urban Population Living In Slums


Percentage of Urban Population Living In Slums
20
40
60
80
0
50
100
150
200
Index


Histogram of Percentage of Urban Population Living In Slums
Frequency
0
5
10
15
0
20
40
60
80
100
Percentage of Urban Population Living In Slums


Percentage of Urban Population Living In Slums
80
60
40
20



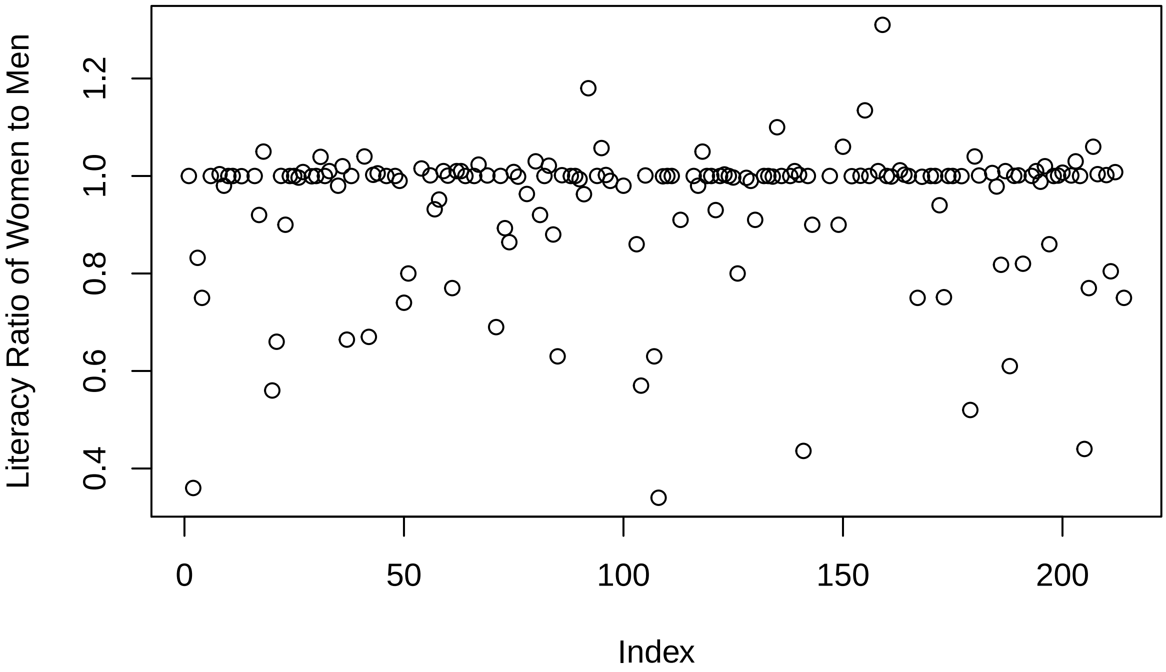
There are 133 missing values. There are differences in median according to region but some regions are missing values entirely. The median of the region wil be imputed for the missing values and if there are no values for a region, zero will be imputed for the missing values.

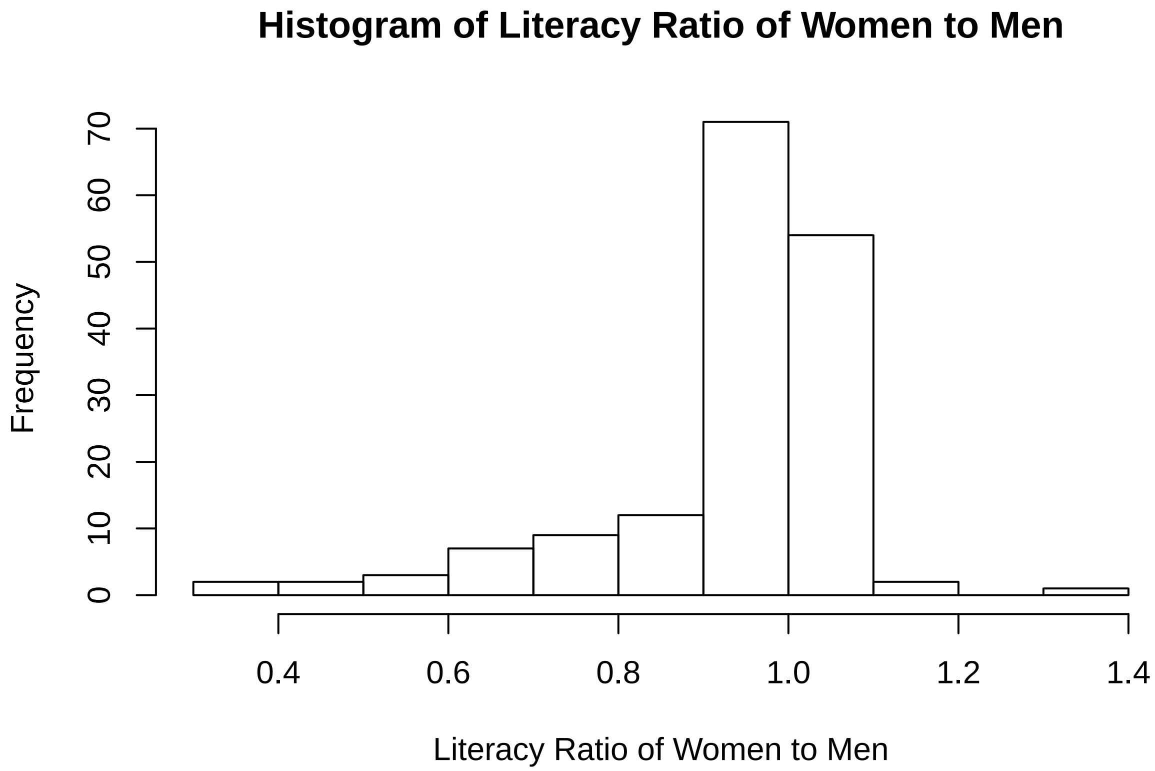
Exploring The Literacy Ratio of Women to Men

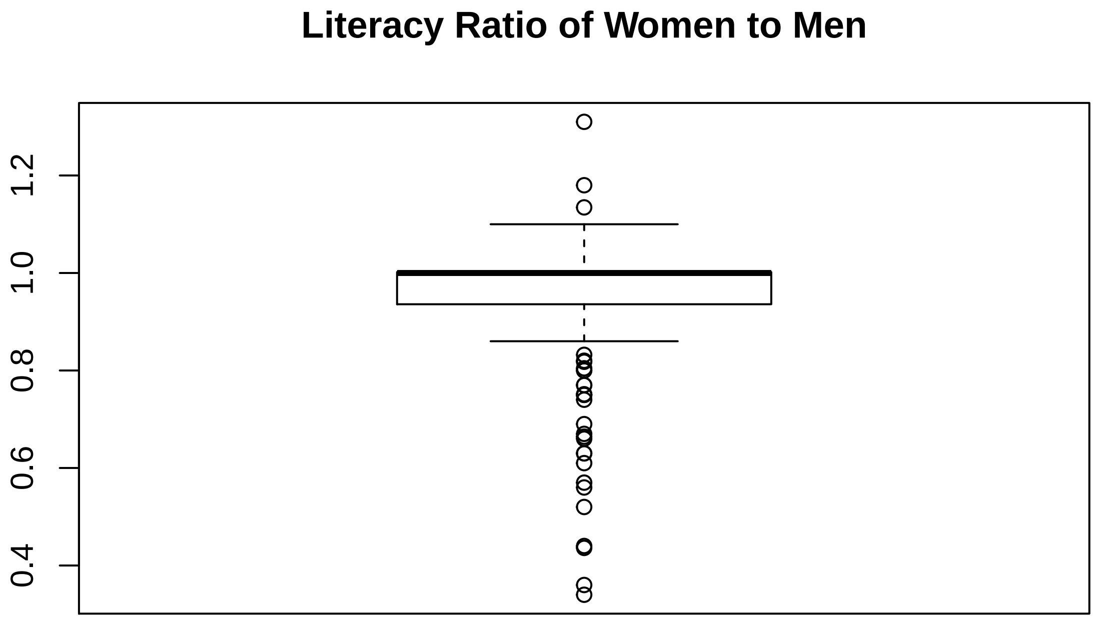
There are 146 missing values. To address this large number of missing values, data will be taken from the following website https://www.nationmaster.com/country-info/stats/Education/Women-to-men-parity-index/As-ratio-of-literacy-rates/Aged-15--24#amount and used for the missing values.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##  0.3400  0.9359  1.0000  0.9424  1.0019  1.3100      51
```

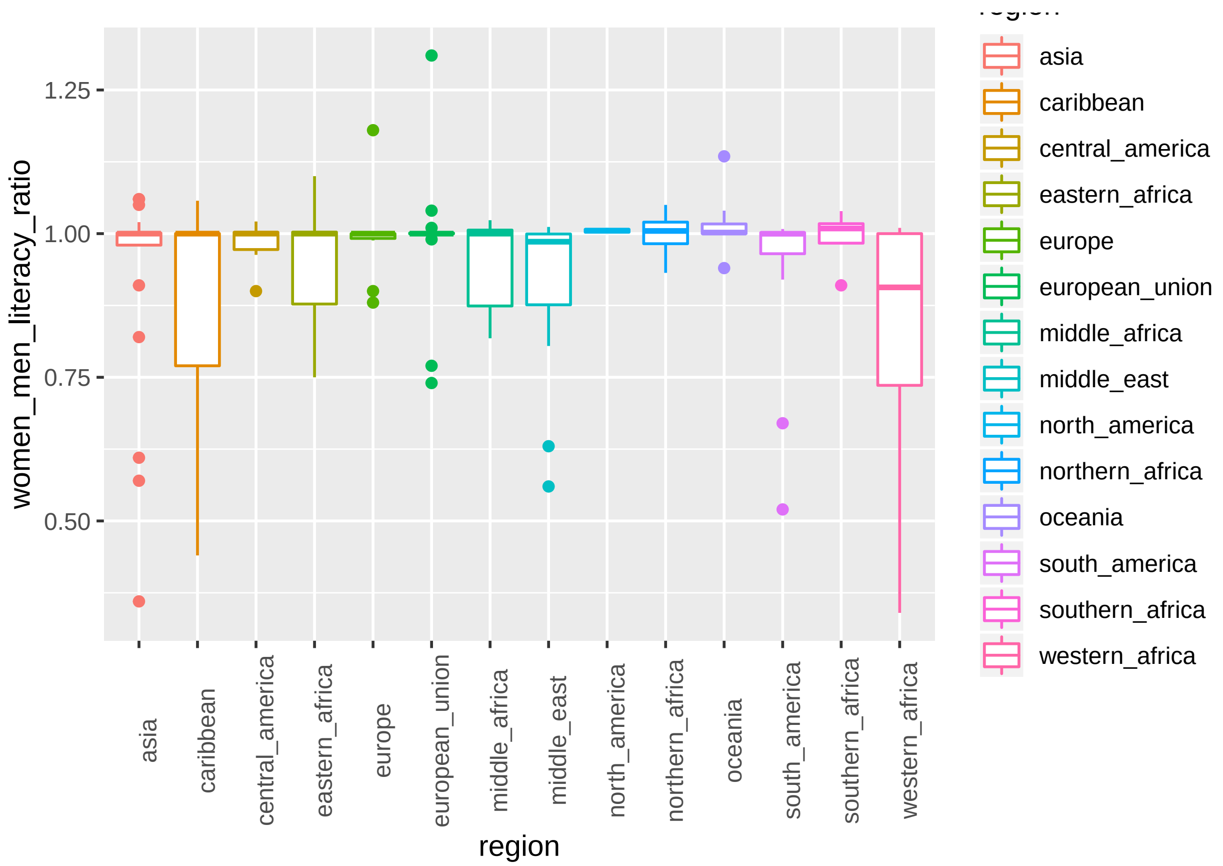
Now there are only 51 missing values.


Literacy Ratio of Women to Men
0.4
0.6
0.8
1.0
1.2
0
50
100
150
200
Index


Histogram of Literacy Ratio of Women to Men
Frequency
0
10
20
30
40
50
60
70
0.4
0.6
0.8
1.0
1.2
1.4
Literacy Ratio of Women to Men



Literacy Ratio of Women to Men



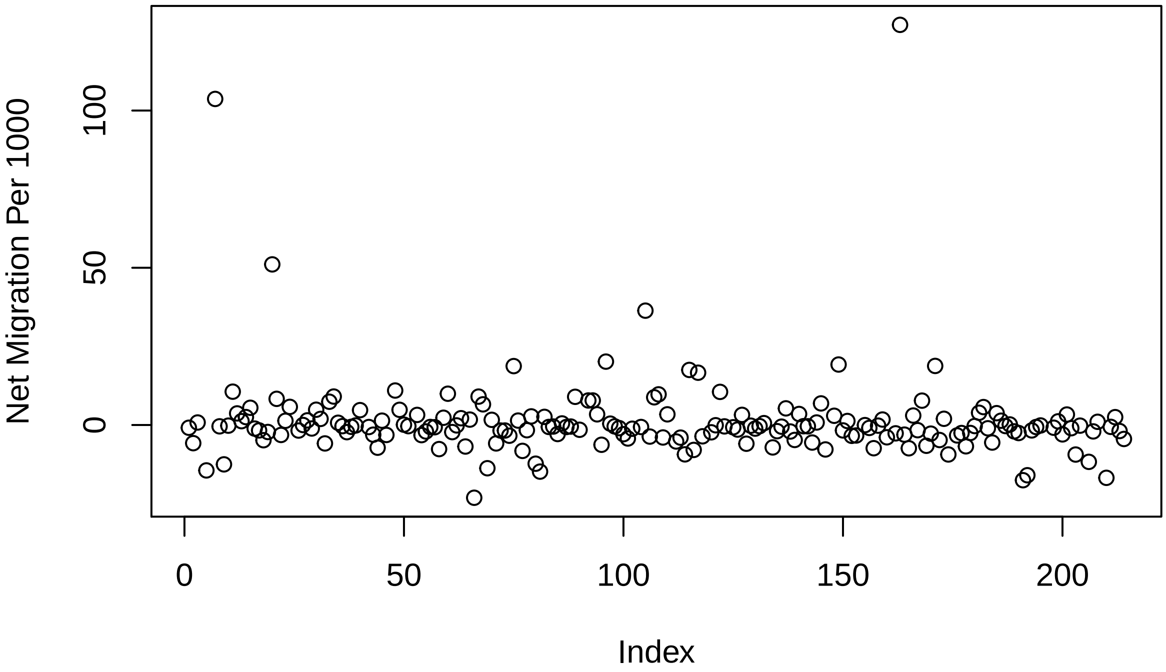
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##  0.3400  0.9359  1.0000  0.9424  1.0019  1.3100      51
```

The median for most regions is close to 1, which means that literacy rate for women is equal to the literacy rate for men. The third quartile for most regions is close to the median of 1. There are some variations be region. The median of the region will be imputed for the missing values.

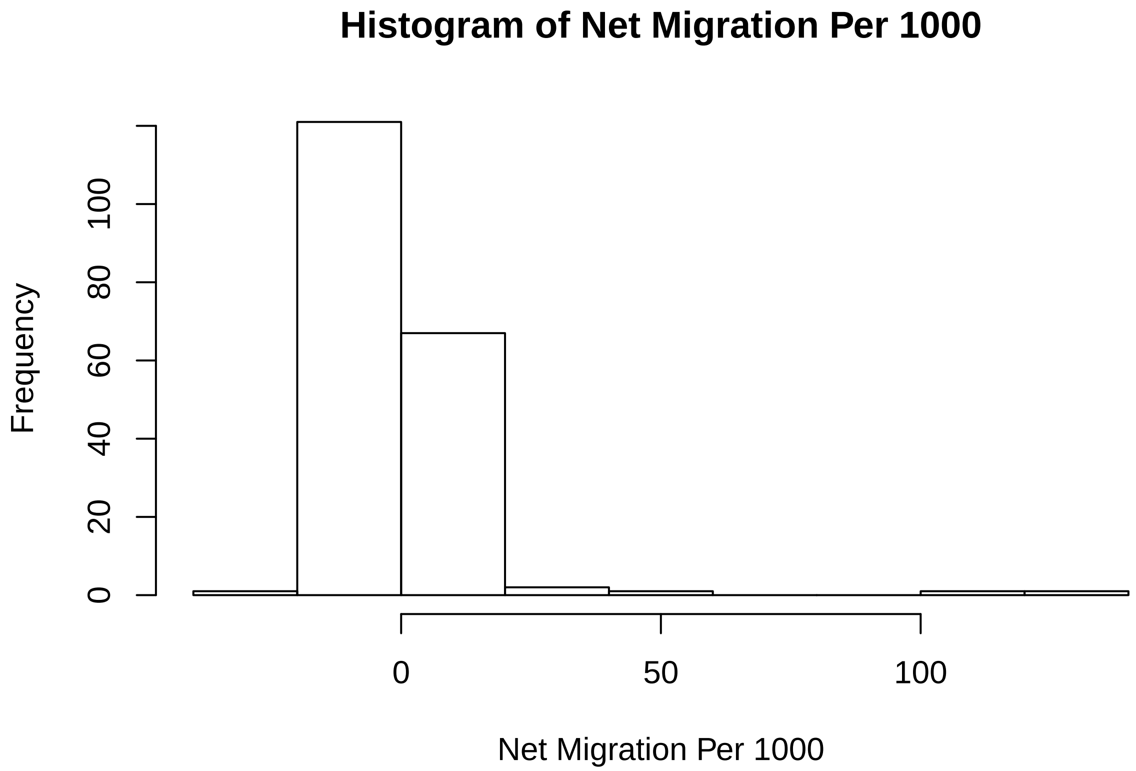
```
##  country_code       fertility_rate_children_per_woman
##  Length:214         Min.   :1.192                    
##  Class :character   1st Qu.:1.746                    
##  Mode  :character   Median :2.309                    
##                     Mean   :2.835                    
##                     3rd Qu.:3.690                    
##                     Max.   :7.599                    
##                                                      
##  carbon_dioxide_emissions cell_subs_per_100 employment_to_pop_ratio
##  Min.   :     40          Min.   :  0.00    Min.   :29.40          
##  1st Qu.:   3345          1st Qu.: 75.08    1st Qu.:51.12          
##  Median :  15750          Median :106.78    Median :56.85          
##  Mean   : 161234          Mean   :106.13    Mean   :56.47          
##  3rd Qu.:  69922          3rd Qu.:130.93    3rd Qu.:62.70          
##  Max.   :7710500          Max.   :322.59    Max.   :86.90          
##                                                                    
##  female_employment_to_pop_ratio maternal_mortality
##  Min.   :11.20                  Min.   :  0.00    
##  1st Qu.:42.31                  1st Qu.: 17.25    
##  Median :47.30                  Median : 51.00    
##  Mean   :47.13                 Mean   :142.79
##  3rd Qu.:55.23                 3rd Qu.:175.50
##  Max.   :79.50                 Max.   :882.00
##
##  percent_children_malaria_nets unmet_family_planning_need
##  Min.   : 0.000               Min.   : 1.70
##  1st Qu.: 0.000               1st Qu.:10.80
##  Median : 0.000               Median :14.95
##  Mean   : 9.009               Mean   :17.20
##  3rd Qu.: 1.375               3rd Qu.:24.82
##  Max.   :80.600               Max.   :55.90
##
##  urban_pop_in_slums women_men_literacy_ratio net_migration_per_1000
##  Min.   : 0.00      Min.   :0.3400           Min.   :-23.129
##  1st Qu.:12.95      1st Qu.:0.9815           1st Qu.: -3.137
##  Median :38.30      Median :1.0000           Median : -0.606
##  Mean   :35.55      Mean   :0.9543           Mean   :  1.138
##  3rd Qu.:55.40      3rd Qu.:1.0009           3rd Qu.:  2.271
##  Max.   :93.30      Max.   :1.3100           Max.   :127.251
##                                              NA's   :20
##      name             region_num           region
##  Length:214         1      :37   asia          :37
##  Class :character   6      :27   european_union:27
##  Mode  :character   2      :22   caribbean     :22
##                     4      :19   eastern_africa:19
##                     5      :18   europe        :18
##                     14     :17   western_africa:17
##                     (Other):74   (Other)       :74
```

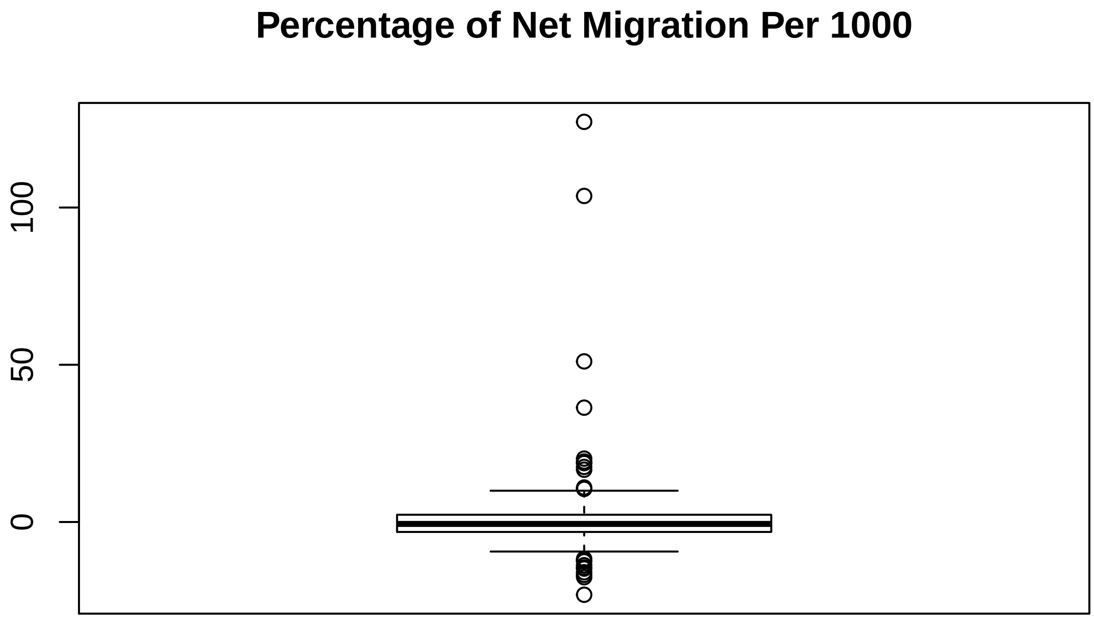
```
## # A tibble: 6 x 15
## # Groups:   region [4]
##   country_code fertility_rate_~ carbon_dioxide_~ cell_subs_per_1~
##   <chr>                   <dbl>            <dbl>            <dbl>
## 1 ABW                      1.8              1090            135.
## 2 AFG                      5.26              830             74.9
## 3 AGO                      5.95            24000             63.5
## 4 AIA                      1.74            15750            180.
## 5 ALB                      6.23             4620            105.
## 6 AND                      1.22            15750             82.6
## # ... with 11 more variables: employment_to_pop_ratio <dbl>,
## #   female_employment_to_pop_ratio <dbl>, maternal_mortality <dbl>,
## #   percent_children_malaria_nets <dbl>, unmet_family_planning_need <dbl>,
## #   urban_pop_in_slums <dbl>, women_men_literacy_ratio <dbl>,
## #   net_migration_per_1000 <dbl>, name <chr>, region_num <fct>,
## #   region <fct>
```

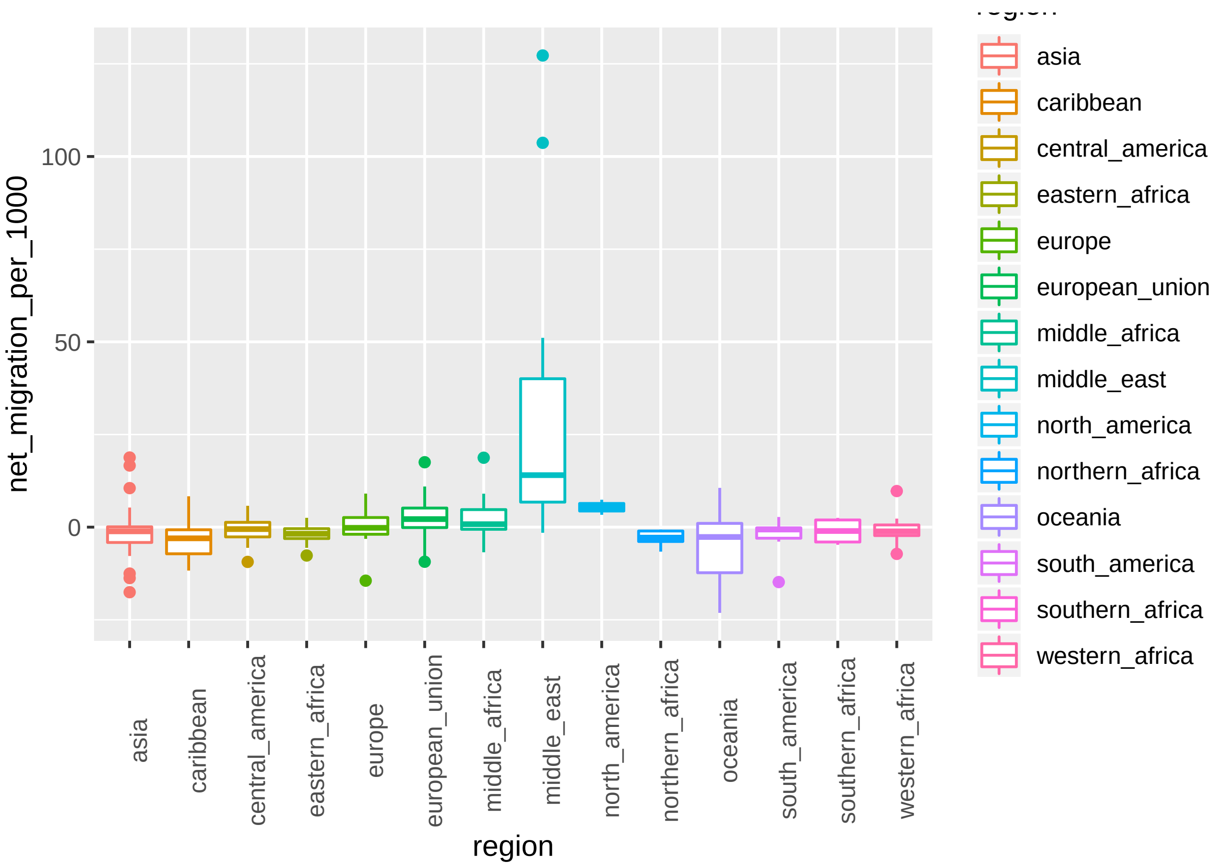
Exploring Net Migration Per 1000



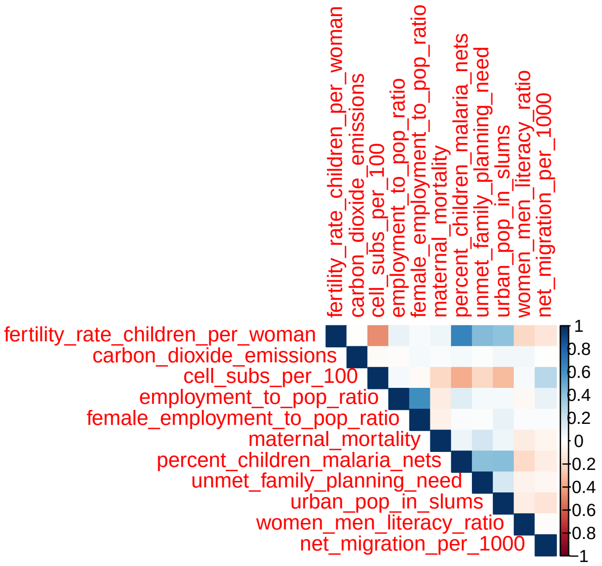
Net Migration Per 1000
0
50
100
0
50
100
150
200
Index


Histogram of Net Migration Per 1000
Frequency
0
20
40
60
80
100
0
50
100
Net Migration Per 1000


Percentage of Net Migration Per 1000
100
50
0



There are 20 missing values. The war in Syria is part of what accounts for striking difference in migration between the middle east and other parts of the world. The median of the region will be imputed for missing values.



Fertility rate is negatively correlated with cell subscriptions per 100 population and high ratios of literacy rates of women to men.

Fertility rate is positively correlated with the percent of children sleeping under insecticide treated nets, unmet family planning need and the percent of urban population living in slums.

Having a high fertility rate may lead to a high unmet family planning need.

The female employment to population ratio is positively correlated with the employment to population ratio.

Cellular subscriptions per 100 in the population is negatively correlated with the percent of children sleeping under insecticide treated bed nets and the percentage of the urban population living in slums.

Build Models

Creating a Test Set and Training Set

Backward Elimination - Linear Regression Model - Model 1

A linear regression model will be built using the backward elimination model. Initially all of the variables will be present, and then they will be removed one at a time. The variable with the highest p value, which has the least effect on fertility rate, will be eliminated first. Variables will be removed until every predictor has a p value below 0.05.

Female employment to population ratio has the (highest p value) lowest effect on fertility rate and will be removed first.

Employment to population ratio has the (highest p value) lowest effect on fertility rate and will be removed next.

Maternal mortality has the (highest p value) lowest effect on fertility rate and will be removed next.

Carbon dioxide emissions has the (highest p value) lowest effect on fertility rate and will be removed next.

Net migration per 100 has the (highest p value) lowest effect on fertility rate and will be removed next.

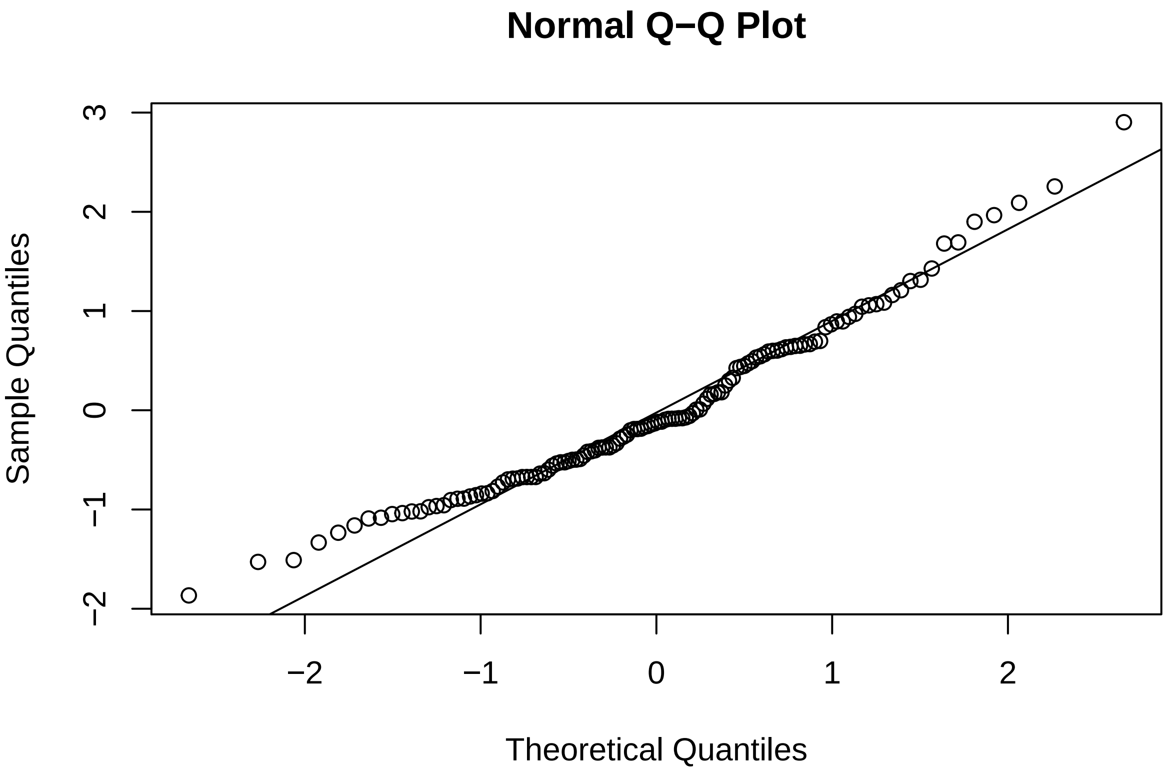
Urban population in slums has the (highest p value) lowest effect on fertility rate and will be removed next.

```
##
## Call:
## lm(formula = fertility_rate_children_per_woman ~ cell_subs_per_100 +
##     percent_children_malaria_nets + unmet_family_planning_need +
##     women_men_literacy_ratio, data = train1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8656 -0.6466 -0.1245  0.5997  2.9027
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    4.310632   0.618005   6.975 1.66e-10 ***
## cell_subs_per_100             -0.007759   0.002017  -3.847 0.000191 ***
## percent_children_malaria_nets  0.039203   0.005690   6.889 2.57e-10 ***
## unmet_family_planning_need     0.026109   0.011569   2.257 0.025791 *
## women_men_literacy_ratio      -1.546420   0.557318  -2.775 0.006387 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8735 on 123 degrees of freedom
## Multiple R-squared:  0.5824, Adjusted R-squared:  0.5688
## F-statistic: 42.89 on 4 and 123 DF,  p-value: < 2.2e-16
```

Fertility rate = 0.039percent_children_malaria_nets + 0.026unmet_family_planning_need - 0.008cell_subs_per_100 - 1.55women_men_literacy_ratio + 4.3

Countries with a higher percentage of children sleeping under insecticide treated nets have higher fertility rates. Countries with higher levels of unmet family planning need have higher fertility rates. Countries with higher levels of cellular subscriptions have lower fertility rates. Countries with a higher ratio of women's literacy to men's literacy have lower fertility rates.

The R squared value is 0.5688. 56.88% of the variation in fertility rate is accounted for by this model.



The residuals at the lower end are not nearly normal. However much of the residuals do follow a normal distribution.

Prediction from Model 1

The root mean square error from model 1 is

```
## [1] 1.149661
```

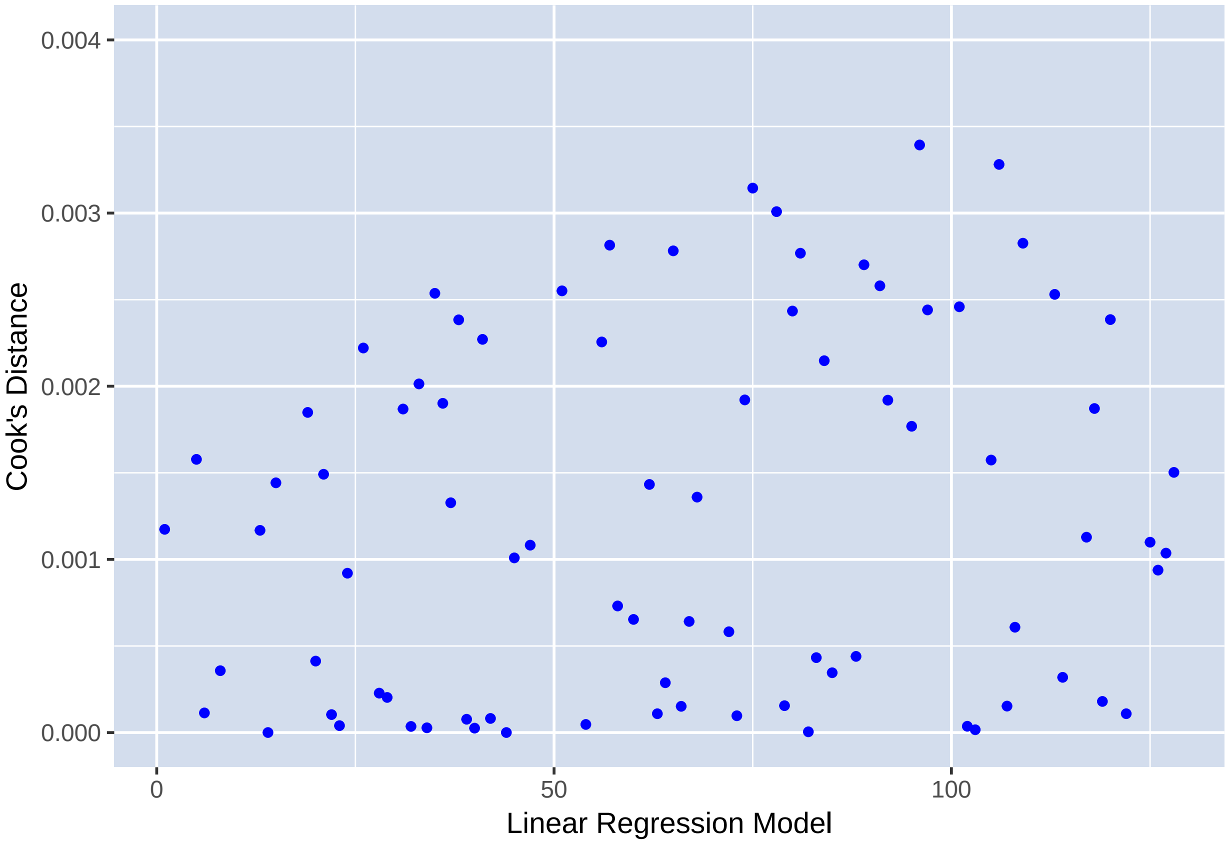
On average, the predicion for the fertility rate, is off by 1.15.

Mulitcolinearity Test

```
##            cell_subs_per_100 percent_children_malaria_nets
##                     1.121764                      1.590792
##    unmet_family_planning_need      women_men_literacy_ratio
##                     1.404217                      1.066811
```

Since each of the variance inflation factor values are below 5, there is not an issue with multicollinearity.

Cook's Distance



The values of Cook's distance are very low. This indicates that there is not an issue with outliers.

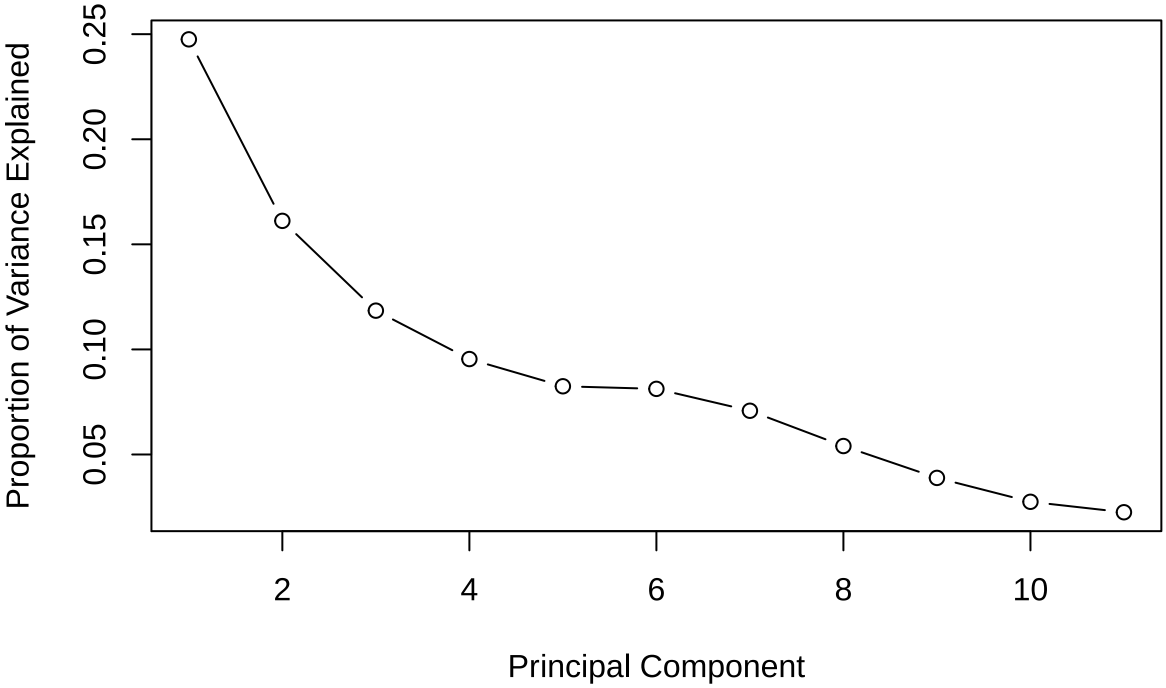
Model 2 Principal Component Analysis

The code to create the PCA models was taken from: http://www.gastonsanchez.com/visually-enforced/how-to/2012/06/17/PCA-in-R/ and https://www.analyticsvidhya.com/blog/2016/03/practical-guide-principal-component-ana

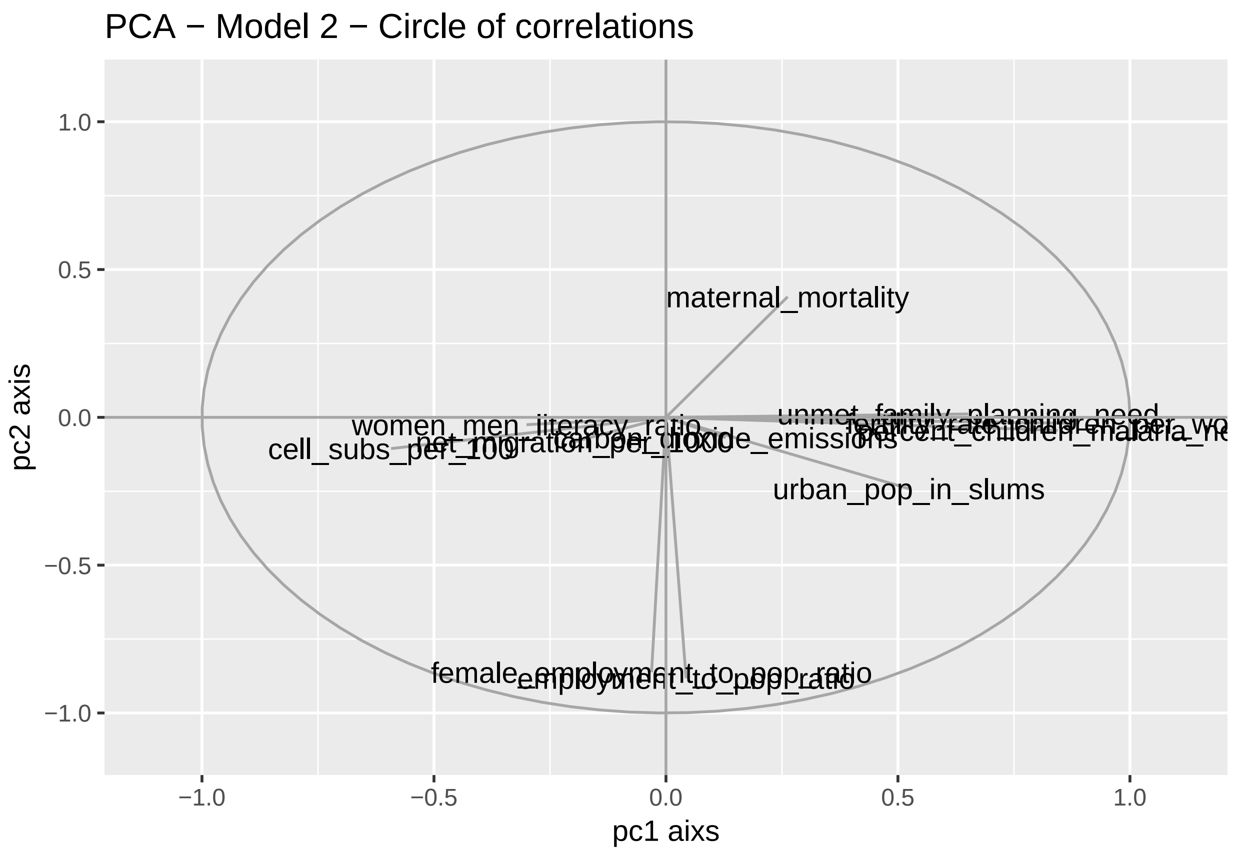
```
## [1] 0.8292908
```

The root mean square error is 0.83. On average, a prediction of the fertility rate is off by 0.83.

Scree Plot



About 99% of the variance is accounted for using 11 principal components.



Model 3 Principal Component Analysis

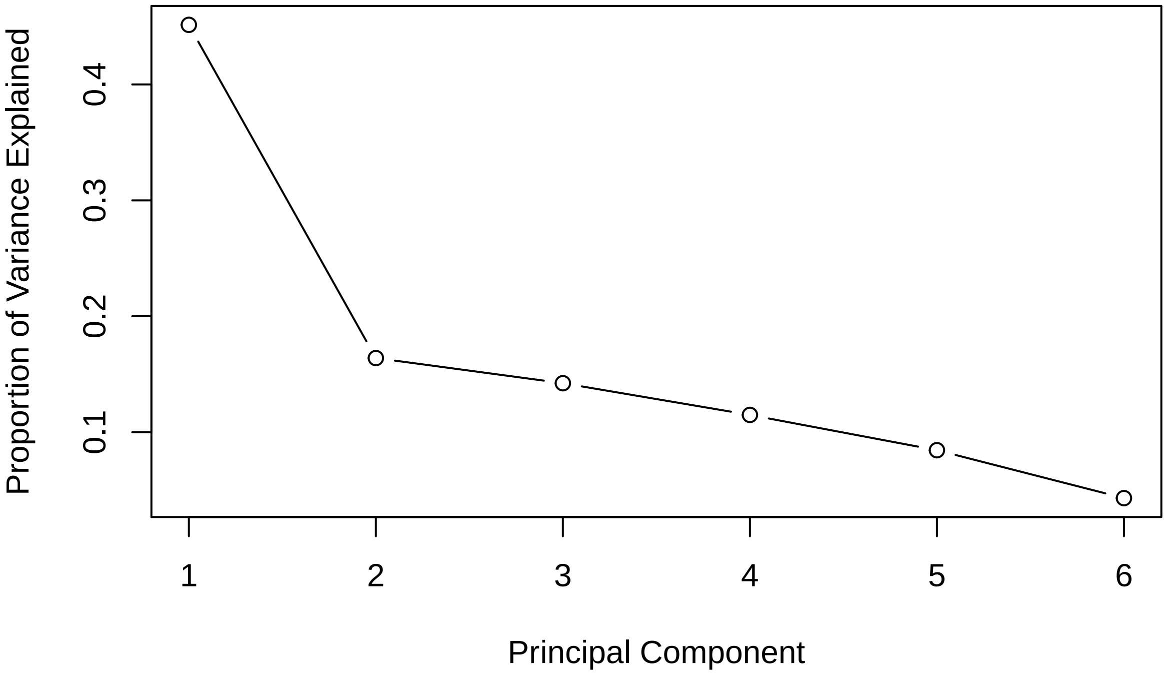
It is hard to distinguish the variables in the graph above. The next model will be built only based on the variables the correlation plot indicated as being correlated to fertility rate.

Creating a Test Set and Training Set

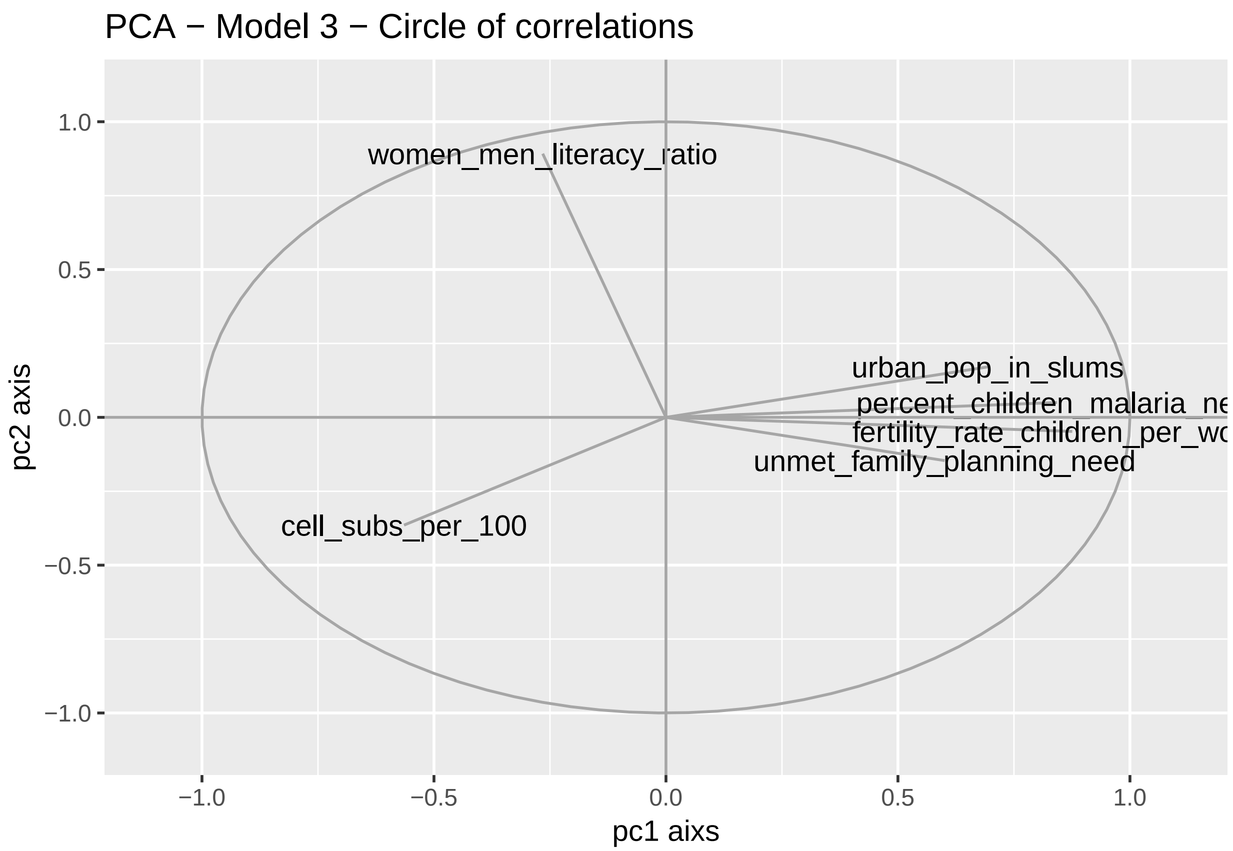
```
## [1] 0.6996821
```

The root mean square error is 0.70. On average, a prediction of the fertility rate is off by 0.70.

Scree Plot



About 98% of the variance in the data set is accounted for using 6 principal components. The scree plot becomes close to horizontal after the second principal component. About 86% of the variance is accounted for by the first 2 principal components.



The correlation plot above displays the variables urban population in slums, the percent of children sleeping under insecticide treated nets, unmet family planning need and fertility rate having positive effects on principal component 1. The ratio of the rate of women's literacy to men's literacy and the number of cellular subscriptions have a negative effect on principal component 1. The ratio of the rate of women's literacy to men's literacy has a strong positive effect on principal component 2 and cellular subscriptions has a negative effect on principal component 2.

```
##                                          PC1         PC2         PC3
## fertility_rate_children_per_woman  0.5321283 -0.04801175 -0.12502973
## urban_pop_in_slums                 0.4213866  0.17207043  0.31838562
## unmet_family_planning_need         0.3649607 -0.14743722 -0.71156312
## percent_children_malaria_nets      0.5121694  0.05048778 -0.06530044
## women_men_literacy_ratio          -0.1613968  0.89925179 -0.37682960
## cell_subs_per_100                 -0.3430890 -0.36762223 -0.48001186
##                                           PC4         PC5         PC6
## fertility_rate_children_per_woman -0.01738691 -0.38406863 -0.74235496
## urban_pop_in_slums                -0.54587207  0.62310075 -0.07228316
## unmet_family_planning_need         0.32582479  0.45980541  0.14546849
## percent_children_malaria_nets     -0.27789758 -0.49636511  0.63817233
## women_men_literacy_ratio          -0.11780957 -0.05879721 -0.07720440
## cell_subs_per_100                 -0.71024885 -0.05459213 -0.09642996
```

The first principal component has a large positive correlation between fertility rate, the percent of children sleeping under insecticide treated nets, urban population in slums and unmet family planning need. The first principal component has a negative correlation with the number of cellular subscriptions. The first principal component therefore can b e considred as a measurement of the extent of a country's poverty level. The greater the poverty level, the higher the fertility rate. The second principal component has a large positive association with the ratio of the literacy rate between women and men, and a negative association with cellular subscriptions. This component is more difficult to categorize. Since it is so highly correlated to literacy ratio, perhaps this component is a measure of women's education. There is a negative correlation between fertility rate and the literacy ratio between women and men.

Model 4 Regression Subset Selection

```
regsubsets.out <-
    regsubsets(fertility_rate_children_per_woman ~ .,
               data = train1,
               nbest = 1,       # 1 best model for each number of predictors
               nvmax = NULL,    # NULL for no limit on number of variables
               force.in = NULL, force.out = NULL,
               method = "exhaustive")
summary.out <- summary(regsubsets.out)
```

Number of variables with highest adjusted (R^2). Variables marked with TRUE are the ones that will be chosen.

```
which.max(summary.out$adjr2)
```

```
## [1] 5
```

```
summary.out$which[5,]
```

```
##                    (Intercept)       carbon_dioxide_emissions
##                           TRUE                          FALSE
##              cell_subs_per_100        employment_to_pop_ratio
##                           TRUE                          FALSE
## female_employment_to_pop_ratio             maternal_mortality
##                          FALSE                          FALSE
##  percent_children_malaria_nets     unmet_family_planning_need
##                           TRUE                           TRUE
##             urban_pop_in_slums       women_men_literacy_ratio
##                           TRUE                           TRUE
##         net_migration_per_1000
##                          FALSE
```

```
summary(best.model <- lm(fertility_rate_children_per_woman ~
                   cell_subs_per_100 +
                   percent_children_malaria_nets +
                   urban_pop_in_slums +
                   unmet_family_planning_need +
                   women_men_literacy_ratio, data = train1))
```

```
##
## Call:
## lm(formula = fertility_rate_children_per_woman ~ cell_subs_per_100 +
##     percent_children_malaria_nets + urban_pop_in_slums + unmet_family_planning_need +
##     women_men_literacy_ratio, data = train1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7201 -0.6103 -0.1320  0.5978  2.8103
##
## Coefficients:
```

```
##                                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                    4.142134   0.639026   6.482 2.01e-09 ***
## cell_subs_per_100             -0.007439   0.002040  -3.647 0.000392 ***
## percent_children_malaria_nets  0.037642   0.005886   6.395 3.09e-09 ***
## urban_pop_in_slums             0.003866   0.003744   1.033 0.303852    
## unmet_family_planning_need     0.026379   0.011569   2.280 0.024339 *  
## women_men_literacy_ratio      -1.540002   0.557203  -2.764 0.006599 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.8733 on 122 degrees of freedom
## Multiple R-squared:  0.586,  Adjusted R-squared:  0.5691 
## F-statistic: 34.54 on 5 and 122 DF,  p-value: < 2.2e-16
```

Prediction from Model 4

```
pred.model4 <- predict(best.model, newdata=test1, type="response")
error.model4 <- pred.model4-test1$fertility_rate_children_per_woman
rmse.model4 <- sqrt(mean(error.model4^2))
rmse.model4
```

```
## [1] 1.136879
```

On average, the prediction for the fertility rate is off by 1.14.

Model 5 Count Models

Interpreting the fertility rate as the number of children a woman would have (a whole number), the dependent variable would be transformed to be able to try out the 3 count models (Poisson, Negative Binomial, Zero Inflated)

Create Count Models

```
## 
## Call:
## glm(formula = fertility_rate_children_per_woman ~ ., family = "poisson", 
##     data = count.train1)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -44.194  -21.225   -3.254   13.892   62.403  
## 
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)    
## (Intercept)                     8.005e+00  2.144e-02  373.37   <2e-16 ***
## carbon_dioxide_emissions       -3.881e-08  2.300e-09  -16.88   <2e-16 ***
## cell_subs_per_100              -5.350e-03  6.919e-05  -77.32   <2e-16 ***
## employment_to_pop_ratio         5.300e-03  2.941e-04   18.02   <2e-16 ***
## female_employment_to_pop_ratio -2.839e-03  1.949e-04  -14.56   <2e-16 ***
## maternal_mortality             -2.657e-04  1.176e-05  -22.59   <2e-16 ***
## percent_children_malaria_nets   1.246e-02  1.441e-04   86.50   <2e-16 ***
## unmet_family_planning_need      2.430e-02  3.754e-04   64.73   <2e-16 ***
## urban_pop_in_slums              2.097e-03  1.144e-04   18.33   <2e-16 ***
## women_men_literacy_ratio       -9.832e-01  1.356e-02  -72.53   <2e-16 ***
## net_migration_per_1000         -1.086e-02  3.666e-04  -29.62   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for poisson family taken to be 1)
## 
##     Null deviance: 136492  on 127  degrees of freedom
## Residual deviance:  64540  on 117  degrees of freedom
## AIC: 65674
## 
## Number of Fisher Scoring iterations: 5
```

```
## 
## Call:
## glm.nb(formula = fertility_rate_children_per_woman ~ cell_subs_per_100 + 
##     employment_to_pop_ratio + female_employment_to_pop_ratio + 
##     percent_children_malaria_nets + unmet_family_planning_need + 
##     women_men_literacy_ratio, data = count.train1, init.theta = 1.864811808, 
##     link = log)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -4.1992  -0.9390  -0.1246   0.5274   2.1100  
## 
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)    
## (Intercept)                     8.032977   0.636982  12.611  < 2e-16 ***
## cell_subs_per_100              -0.007210   0.001719  -4.194 2.75e-05 ***
## employment_to_pop_ratio         0.014971   0.009020   1.660  0.09697 .  
## female_employment_to_pop_ratio -0.014110   0.006294  -2.242  0.02497 *  
## percent_children_malaria_nets   0.016008   0.004972   3.219  0.00128 ** 
## unmet_family_planning_need      0.020704   0.009869   2.098  0.03592 *  
## women_men_literacy_ratio       -0.760415   0.468137  -1.624  0.10430    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for Negative Binomial(1.8648) family taken to be 1)
## 
##     Null deviance: 226.95  on 127  degrees of freedom
## Residual deviance: 141.75  on 121  degrees of freedom
## AIC: 2078.9
## 
## Number of Fisher Scoring iterations: 1
## 
## 
##               Theta:  1.865 
##           Std. Err.:  0.220 
## 
##  2 x log-likelihood:  -2062.949
```

```
## 
## Call:
## zeroinfl(formula = fertility_rate_children_per_woman ~ . | percent_children_malaria_nets, 
##     data = count.train1)
## 
```

```
## Pearson residuals:
##      Min       1Q   Median       3Q      Max 
## -28.0672  -6.0188  -0.9492   5.2420  49.3347 
## 
## Count model coefficients (poisson with log link):
##                                  Estimate Std. Error z value Pr(>|z|)    
## (Intercept)                     7.955e+00         NA      NA       NA    
## carbon_dioxide_emissions       -3.777e-08  0.000e+00     Inf   <2e-16 ***
## cell_subs_per_100              -5.048e-03         NA      NA       NA    
## employment_to_pop_ratio         5.279e-03         NA      NA       NA    
## female_employment_to_pop_ratio -2.649e-03         NA      NA       NA    
## maternal_mortality             -2.571e-04         NA      NA       NA    
## percent_children_malaria_nets   1.255e-02         NA      NA       NA    
## unmet_family_planning_need      2.417e-02         NA      NA       NA    
## urban_pop_in_slums              2.302e-03         NA      NA       NA    
## women_men_literacy_ratio       -9.758e-01         NA      NA       NA    
## net_migration_per_1000         -1.065e-02         NA      NA       NA    
## 
## Zero-inflation model coefficients (binomial with logit link):
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -4.554         NA      NA       NA
## percent_children_malaria_nets  -55.149         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Number of iterations in BFGS optimization: 28 
## Log-likelihood: -3.252e+04 on 13 Df
```

Prediction from Count Models

```
##  [1] -1.220 -3.100 -1.885 -5.200 -1.591 -1.770 -1.414 -6.230 -2.221 -1.497
## [11] -1.878 -1.380 -3.454 -4.000 -5.950 -1.600 -1.820 -2.200 -4.668 -1.428
## [21] -1.820 -3.838 -2.036 -4.994 -2.170 -1.489 -3.130 -4.010 -2.096 -1.750
## [31] -1.269 -2.106 -1.390 -4.400 -7.599 -4.630 -1.900 -3.325 -4.058 -1.204
## [41] -2.008 -6.400 -6.000 -2.190 -2.599 -1.591 -4.100 -2.000 -2.252 -4.200
## [51] -2.650 -4.875 -3.154 -2.800 -1.996 -2.071 -1.329 -1.740 -2.605 -1.706
## [61] -1.488 -2.450 -3.788 -2.290 -2.900 -2.111 -6.610 -6.310 -1.982 -5.135
## [71] -1.800 -2.050 -1.233 -1.343 -1.717 -2.438 -2.640 -4.000 -1.780 -4.750
## [81] -3.000 -4.164 -5.450 -4.950 -4.900 -5.646
```

```
##  [1] -1.220 -3.100 -1.885 -5.200 -1.591 -1.770 -1.414 -6.230 -2.221 -1.497
## [11] -1.878 -1.380 -3.454 -4.000 -5.950 -1.600 -1.820 -2.200 -4.668 -1.428
## [21] -1.820 -3.838 -2.036 -4.994 -2.170 -1.489 -3.130 -4.010 -2.096 -1.750
## [31] -1.269 -2.106 -1.390 -4.400 -7.599 -4.630 -1.900 -3.325 -4.058 -1.204
## [41] -2.008 -6.400 -6.000 -2.190 -2.599 -1.591 -4.100 -2.000 -2.252 -4.200
## [51] -2.650 -4.875 -3.154 -2.800 -1.996 -2.071 -1.329 -1.740 -2.605 -1.706
## [61] -1.488 -2.450 -3.788 -2.290 -2.900 -2.111 -6.610 -6.310 -1.982 -5.135
## [71] -1.800 -2.050 -1.233 -1.343 -1.717 -2.438 -2.640 -4.000 -1.780 -4.750
## [81] -3.000 -4.164 -5.450 -4.950 -4.900 -5.646
```

```
##  [1] -1.220 -3.100 -1.885 -5.200 -1.591 -1.770 -1.414 -6.230 -2.221 -1.497
## [11] -1.878 -1.380 -3.454 -4.000 -5.950 -1.600 -1.820 -2.200 -4.668 -1.428
## [21] -1.820 -3.838 -2.036 -4.994 -2.170 -1.489 -3.130 -4.010 -2.096 -1.750
## [31] -1.269 -2.106 -1.390 -4.400 -7.599 -4.630 -1.900 -3.325 -4.058 -1.204
## [41] -2.008 -6.400 -6.000 -2.190 -2.599 -1.591 -4.100 -2.000 -2.252 -4.200
## [51] -2.650 -4.875 -3.154 -2.800 -1.996 -2.071 -1.329 -1.740 -2.605 -1.706
## [61] -1.488 -2.450 -3.788 -2.290 -2.900 -2.111 -6.610 -6.310 -1.982 -5.135
## [71] -1.800 -2.050 -1.233 -1.343 -1.717 -2.438 -2.640 -4.000 -1.780 -4.750
## [81] -3.000 -4.164 -5.450 -4.950 -4.900 -5.646
```

```
##        count.models       count.rmse
## 1           Poisson 1.20702379052436
## 2 Negative Binomial 1.20716261676315
## 3     Zero Inflated 1.20701084749699
```

On average, the prediction for the fertility rate is off by 1.21.

Discussion and Conclusion

It is possible to predict the fertility rate in a country based on the percent of children sleeping under insecticide treated bed nets, the percentage of the urban population in slums, unmet family planning need, the number of cellular subscriptions and the ratio of the literacy rate between women and men. We created 7 different models and all gave relatively similar results. The model with the lowest root mean square error was model 3, which employed principal component analysis. In conclusion, higher fertility rates are associated with having more children sleeping under insecticide treated bed nets, higher percentages of the urban population living in slums and higher unmet family planning need. From our analysis, it is not possible to ascertain whether a high fertility rate is the effect of these variables, the cause of these variables or simply correlated with them. For example, does having a large number of children sleeping under insecticide treated nets correlate with a higher fertility rate or a consequence of a high fertility rate? A high literacy rate among women compared with men as well as a large number of cellular subscriptions is associated with countries that have lower fertility rates. A source of further research would involve uncovering the nature of these connections to identify causation for fertility rates, not just correlations. In closing, high fertility rates are associated with higher poverty levels and lower levels of education between women and men.

References

[1] United Nations, Department of Economic and Social Affairs. World Population Prospects, The 2017 Revision. Retrieved from https://esa.un.org/unpd/wpp/Publications/Files/WPP2017_Methodology.pdf.

[2] Alkema et al (2011). Probabilistic Projections of the Total Fertility Rate for All Countries. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3367999/.

[3] National Institute of Health (2011). NIH-funded study proposes new method to predict fertility rates. Retrieved from https://www.nih.gov/news-events/news-releases/nih-funded-study-proposes-new-method-predict-fertility-rates.

[4] United Nations, Department of Economic and Social Affairs. Population Trends. Retrieved from http://www.un.org/en/development/desa/population/theme/trends/index.shtml

[5] Walter Oberhofer and Thomas Reichsthaler (2004). Modelling Fertility: A Semi-Parametric Approach. Retrived from https://epub.uni-regensburg.de/4511/1/rdisb396.pdf

Appendix

library(tidyr) library(dplyr) library(corrplot) library(ggplot2) library(car) library(rpart)

fertility <- read.csv("https://raw.githubusercontent.com/swigodsky/Data621/master/fertility_rate_predictionUN.csv", stringsAsFactors = FALSE) country_code_df <- read.csv("https://raw.githubusercontent.com/swigodsky/Data621/master/country_codes.csv", stringsAsFactors = TRUE) country_code_df$region_num < -as.factor(country_code_df$region_num) fertility <- left_join(fertility,country_code_df,by="country_code") fertility <- subset(fertility, select=-c(order,country_number)) head(fertility) nrow(fertility)

summary(fertility)

country_per_region <- fertility country_per_region <- country_per_region %>% select(region_num,region) country_per_region <- as.data.frame(table(country_per_region\$region_num)) colnames(country_per_region) <- c(“region_num”,“NumCountriesPerRegion”) country_per_region

bed_nets_region <- fertility bed_nets_region <- bed_nets_region %>% filter(is.na(percent_children_malaria_nets)) %>% select(region_num)

count_na_by_region <- as.data.frame(table(bed_nets_region\$region_num)) colnames(count_na_by_region) <- c(“region_num”,“Freq”) count_na_by_region <- count_na_by_region %>% left_join(country_code_df,by=“region_num
%>% left_join(country_per_region,by=“region_num”) %>% select(region_num,region,Freq,NumCountriesPerRegion) %>% mutate(NumCountriesNA=Freq) %>% mutate(NumCountriesPerRegion=NumCountriesPerRegion) %>% mutate(PerentageCountriesInRegion=Freq*100/NumCountriesPerRegion) %>% distinct(region_num, region, NumCountriesNA,NumCountriesPerRegion, PerentageCountriesInRegion) count_na_by_region

bed_nets_region_avg <- fertility bed_nets_region_avg <- bed_nets_region_avg %>% filter((percent_children_malaria_net
%>% select(region_num,region,percent_children_malaria_nets) %>% group_by(region_num) %>% mutate(AvgBedNetsInRegion=mean(percent_children_malaria_nets)) %>% distinct(region_num, region, AvgBedNetsInRegion) bed_nets_region_avg

fertility$percent_{c}hildren_{m}alaria_{n}ets[is.na(fertility$percent_children_malaria_nets)] <- 0

plot(fertility$carbon_{d}ioxide_{e}missions, ylab = ”CarbonDioxideEmissions”)hist(fertility$carbon_dioxide_emissions, xlab=“Carbon Dioxide Emissions”,main=“Histogram of Carbon Dioxide Emissions”) boxplot(fertility\$carbon_dioxide_emission
main=“Carbon Dioxide Emissions”) filter(fertility, carbon_dioxide_emissions>4000000)

guardianCO2 <- read.csv(“https://raw.githubusercontent.com/swigodsky/Data621/master/co2emissions.csv”, stringsAsFactors = FALSE) fertility <- fertility %>% left_join(guardianCO2, by =“name”) fertility$carbon_{d}ioxide_{e}missions[is.na(fertility$carbon_dioxide_emissions)] <- fertility$co2Guardian fertility <$ $-subset(fertility, select = -c(emissions, co2Guardian))summary(fertility$carbon_dioxide_emissions)

plot(fertility$carbon_{d}ioxide_{e}missions, ylab = ”CarbonDioxideEmissions”)hist(fertility$carbon_dioxide_emissions, xlab=“Carbon Dioxide Emissions”,main=“Histogram of Carbon Dioxide Emissions”) boxplot(fertility\$carbon_dioxide_emission
main=“Carbon Dioxide Emissions”)

medianCO2 <- median(fertility$carbon_{d}ioxide_{e}missions, na.rm = TRUE)fertility$carbon_dioxide_emissions[is.na(fertility\$ca
<- medianCO2

plot(fertility$cell_{s}ubs_{p}er_{1}00, ylab = ”CellSubscriptionsper100Population”)hist(fertility$cell_subs_per_100, xlab=“Cell Subscriptions per 100 Population”,main=“Histogram of Cell Subscriptions per 100 Population”) boxplot(fertility\$cell_subs_per_100, main=“Cell Subscriptions per 100 Population”) ggplot(fertility, aes(x=region,y=cell_subs_per_100)) + geom_boxplot(aes(color=region)) + theme(axis.text.x = element_text(angle=90))

fertility <- fertility %>% group_by(region) %>% mutate(med_cell=median(cell_subs_per_100,na.rm=TRUE)) fertility$cell_{s}ubs_{p}er_{1}00[is.na(fertility$cell_subs_per_100)] <- fertility\$med_cell fertility = subset(fertility, select=-c(med_cell))

plot(fertility$employment_{t}o_{p}op_{r}atio, ylab = ”EmploymenttoPopulationRatio”)hist(fertility$employment_to_pop_ratio, xlab=“Employment to Population Ratio”,main=“Histogram of Employment to Population Ratio”) boxplot(fertility\$employment_to_pop_ratio, main=“Employment to Population Ratio”) ggplot(fertility, aes(x=region,y=employment_to_pop_ratio)) + geom_boxplot(aes(color=region)) + theme(axis.text.x = element_text(angle=90))

fertility <- fertility %>% group_by(region) %>% mutate(med_employ=median(employment_to_pop_ratio,na.rm=TRUE)) fertility$employment_{t}o_{p}op_{r}atio[is.na(fertility$employment_to_pop_ratio)] <- fertility$med_{e}mployeast_{a}fr <$ $-filter(fertility, region == ”eastern_{a}frica”)fertility$employment_to_pop_ratio[is.na(fertility\$employment_to_pop_ratic
<- east_afr\$med_employ fertility = subset(fertility, select=-c(med_employ))

plot(fertility$female_{e}mployment_{t}o_{p}op_{r}atio, ylab = "FemaleEmploymenttoPopulationRatio")hist(fertility$female_employme xlab=“Female Employment to Population Ratio”,main=“Histogram of Female Employment to Population Ratio”) boxplot(fertility$female_employment_to_pop_ratio, main=“Female Employment to Population Ratio”) ggplot(fertility, aes(x=region,y=female_employment_to_pop_ratio)) + geom_boxplot(aes(color=region)) + theme(axis.text.x = element_text(angle=90))

fertility <- fertility %>% group_by(region) %>% mutate(med_fem_employ=median(female_employment_to_pop_ratio,na.r fertility$female_{e}mployment_{t}o_{p}op_{r}atio[is.na(fertility$female_employment_to_pop_ratio)] <- fertility$med_{f}em_{e}mployeast_{a}fr$ $-filter(fertility, region == "eastern_{a}frica")fertility$female_employment_to_pop_ratio[is.na(fertility$female_employmen <- east_afr$med_fem_employ fertility = subset(fertility, select=-c(med_fem_employ))

plot(fertility$lowest_{q}uint_{i}ncome_{s}hare.csv, ylab = "PoorestQuintile'sShareinIncome")hist(fertility$lowest_quint_income_s xlab=“Poorest Quintile’s Share in Income”,main=“Histogram of Poorest Quintile’s Share in Income”) boxplot(fertility$lowest_quint_income_share.csv, main=“Poorest Quintile’s Share in Income”) ggplot(fertility, aes(x=region,y=lowest_quint_income_share.csv)) + geom_boxplot(aes(color=region)) + theme(axis.text.x = element_text(angle=90))

fertility = subset(fertility, select=-c(lowest_quint_income_share.csv))

mat_mortUNICEF <- read.csv(“https://raw.githubusercontent.com/swigodsky/Data621/master/maternal_mortality2015.csv”, stringsAsFactors = FALSE) fertility <- fertility %>% left_join(mat_mortUNICEF, by =“country_code”) fertility$maternal_{m}ortality[is.na(fertility$maternal_mortality)] <- fertility$maternal_{m}ortalityUNICEFfe$ $-subset(fertility, select = -c(maternal_{m}ortalityUNICEF))summary(fertility$maternal_mortality)

plot(fertility$maternal_{m}ortality, ylab = "MaternalMortality")hist(fertility$maternal_mortality, xlab=“Maternal Mortality”,main=“Histogram of Maternal Mortality”) boxplot(fertility$maternal_{m}ortality, main = "MaternalMortality")ggplot(fertility, aes(x = region, y = maternal_{m}ortality))+geom_{b}oxplot(aes(color = region)) + theme(axis.text.x = element_{t}ext(angle = 90))summary(fertility$maternal_mortality)

fertility <- fertility %>% group_by(region) %>% mutate(med_mother_mort=median(maternal_mortality,na.rm=TRUE)) fertility$maternal_{m}ortality[is.na(fertility$maternal_mortality)] <- fertility$med_mother_mort fertility = subset(fertility, select=-c(med_mother_mort))

plot(fertility$unmet_{f}amily_{p}lanning_{n}eed, ylab = "PercentUnmetFamilyPlanningNeed")hist(fertility$unmet_family_planni xlab=“Percent Unmet Family Planning Need”,main=“Histogram of Unmet Family Planning Need”) boxplot(fertility$unmet_family_planning_need, main=“Percent Unmet Family Planning Need”) ggplot(fertility, aes(x=region,y=unmet_family_planning_need)) + geom_boxplot(aes(color=region)) + theme(axis.text.x = element_text(angle=90))

fertility <- fertility %>% group_by(region) %>% mutate(med_unmet_fam=median(unmet_family_planning_need,na.rm=TI fertility$unmet_{f}amily_{p}lanning_{n}eed[is.na(fertility$unmet_family_planning_need)] <- fertility$med_unmet_fam fertility = subset(fertility, select=-c(med_unmet_fam))

plot(fertility$urban_{p}op_{i}n_{s}lums, ylab = "PercentageofUrbanPopulationLivingInSlums")hist(fertility$urban_pop_in_slums xlab=“Percentage of Urban Population Living In Slums”,main=“Histogram of Percentage of Urban Population Living In Slums”) boxplot(fertility$urban_pop_in_slums, main=“Percentage of Urban Population Living In Slums”) ggplot(fertility, aes(x=region,y=urban_pop_in_slums)) + geom_boxplot(aes(color=region)) + theme(axis.text.x = element_text(angle=90))

fertility <- fertility %>% group_by(region) %>% mutate(med_slums=median(urban_pop_in_slums,na.rm=TRUE)) fertility$urban_{p}op_{i}n_{s}lums[is.na(fertility$urban_pop_in_slums)] <- fertility$med_{s}lumsfertility = subset(fertility, select = -c(med_{s}lums))fertility$urban_pop_in_slums[is.na(fertility$urban_pop_in_slums)] <- 0

literacy <- read.csv(“https://raw.githubusercontent.com/swigodsky/Data621/master/women_men_literacy.csv”, stringsAsFactors = FALSE) fertility <- fertility %>% left_join(literacy, by =“name”) fertility$women_{m}en_{l}iteracy_{r}atio[is.na(fertility$women_men_literacy_ratio)] <- fertility$women_{m}en_{l}iteracy_{r}atioNEWfertili$ $-subset(fertility, select = -c(women_{m}en_{l}iteracy_{r}atioNEW))summary(fertility$women_men_literacy_ratio) plot(fertility$women_men_literacy_ratio, ylab = "LiteracyRatioofWomentoMen")hist(fertility$women_men_literacy_ratio, xlab=“Literacy Ratio of Women to Men”,main=“Histogram of Literacy Ratio of Women to Men”) boxplot(fertility$women_men_literacy_ratio, main = "LiteracyRatioofWomentoMen")ggplot(fertility, aes(x = region, y = women_men_literacy_ratio)) + geom_boxplot(aes(color = region)) + theme(axis.text.x = element_text(angle = 90))summary(fertility$women_men_literacy_ratio)

fertility <- fertility %>% group_by(region) %>% mutate(med_lit=median(women_men_literacy_ratio,na.rm=TRUE)) fertility$women_men_literacy_ratio[is.na(fertility$women_men_literacy_ratio)] <- fertility$med_lit fertility = subset(fertility, select=-c(med_lit)) summary(fertility) head(fertility)

plot(fertility$net_migration_per_1000, ylab = "NetMigrationPer1000")hist(fertility$net_migration_per_1000, xlab=“Net Migration Per 1000”,main=“Histogram of Net Migration Per 1000”) boxplot(fertility$net_migration_per_1000, main=“Percentage of Net Migration Per 1000”) ggplot(fertility, aes(x=region,y=net_migration_per_1000)) + geom_boxplot(aes(color=region)) + theme(axis.text.x = element_text(angle=90))

fertility <- fertility %>% group_by(region) %>% mutate(med_mig=median(net_migration_per_1000,na.rm=TRUE)) fertility$net_migration_per_1000[is.na(fertility$net_migration_per_1000)] <- fertility$med_mig fertility = subset(fertility, select=-c(med_mig))

fertility_variables <- fertility fertility_variables <- subset(fertility_variables, select=-c(region,name,region_num,country_code

correlation <- cor(fertility_variables, method = “pearson”,use=“complete.obs”) corrplot(correlation, type=“upper”, method=“color”)

set.seed(15) n <- nrow(fertility_variables) shuffle_df1 <- fertility_variables[sample(n),] train_indeces <- 1:round(0.6*n) train1 <- shuffle_df1[train_indeces,] test_indeces <- (round(.6*n)+1):n test1 <- shuffle_df1[test_indeces,]

fert_lm <- lm(fertility_rate_children_per_woman ~ ., data=train1) summary(fert_lm)

fert_lm <- update(fert_lm, .~. -female_employment_to_pop_ratio, data = train1) summary(fert_lm)

fert_lm <- update(fert_lm, .~. -employment_to_pop_ratio, data = train1) summary(fert_lm)

fert_lm <- update(fert_lm, .~. -maternal_mortality, data = train1) summary(fert_lm)

fert_lm <- update(fert_lm, .~. -carbon_dioxide_emissions, data = train1) summary(fert_lm)

fert_lm <- update(fert_lm, .~. -net_migration_per_1000, data = train1) summary(fert_lm)

fert_lm <- update(fert_lm, .~. -urban_pop_in_slums, data = train1) summary(fert_lm)

qqnorm(resid(fert_lm)) qqline(resid(fert_lm))

pred1 <- predict(fert_lm, newdata=test1, type=“response”) error <- pred1-test1$fertility_rate_children_per_woman rmse1 <- sqrt(mean(error^2)) rmse1

vif(fert_lm)

ggplot()+geom_point(aes(x=seq_along(cooks.distance(fert_lm)),y=cooks.distance(fert_lm)),color=‘blue’,shape=20,size=2)+ theme(panel.background = element_rect(fill = ‘#d3dded’))+labs(x=‘Linear Regression Model’,y=“Cook’s Distance”)+ylim(0,.004)

prin_comp <- prcomp(train1, scale. = T)

train.data <- data.frame(fert = train1$fertility_rate_children_per_woman, prin_comp$x)

fert_rpart <- rpart(fert ~ .,data = train.data, method = “anova”)

test.data <- predict(prin_comp, newdata = test1) test.data <- as.data.frame(test.data)

pred2 <- predict(fert_rpart, test.data)

error <- pred2-test1$fertility_rate_children_per_woman rmse2 <- sqrt(mean(error^2)) rmse2 std_dev <- prin_comp$sdev pr_var <- std_dev^2 prop_varex <- pr_var/sum(pr_var) plot(prop_varex, xlab = “Principal Component”, ylab = “Proportion of Variance Explained”, type = “b”)

circle <- function(center = c(0, 0), npoints = 100) { r = 1 tt = seq(0, 2 * pi, length = npoints) xx = center[1] + r * cos(tt) yy = center[1] + r * sin(tt) return(data.frame(x = xx, y = yy)) } corcir = circle(c(0, 0), npoints = 100)

correlations = as.data.frame(cor(train1, prin_comp$x))

arrows = data.frame(x1 = c(0,0,0,0,0,0,0,0,0,0,0), y1 = c(0,0,0,0,0,0,0,0,0,0,0), x2 = correlations$PC1, y2 = correlations$PC2)

ggplot() + geom_path(data = corcir, aes(x = x, y = y), colour = “gray65”) + geom_segment(data = arrows, aes(x = x1, y = y1, xend = x2, yend = y2), colour = “gray65”) + geom_text(data = correlations, aes(x = PC1, y = PC2, label = rownames(correlations))) + geom_hline(yintercept = 0, colour = “gray65”) + geom_vline(xintercept = 0, colour = “gray65”) + xlim(-1.1, 1.1) + ylim(-1.1, 1.1) + labs(x = “pc1 aixs”, y = “pc2 axis”) + ggtitle(“PCA - Model 2 - Circle of correlations”)

fertility3 <- fertility fertility3 <- subset(fertility3, select=c(fertility_rate_children_per_woman,urban_pop_in_slums, unmet_family_planning_need, percent_children_malaria_nets, women_men_literacy_ratio, cell_subs_per_100))

set.seed(25) n <- nrow(fertility_variables) shuffle_df3 <- fertility3[sample(n),] train_indeces <- 1:round(0.6*n) train3 <- shuffle_df3[train_indeces,] test_indeces <- (round(.6*n)+1):n test3 <- shuffle_df3[test_indeces,]

prin_comp3 <- prcomp(train3, scale. = T)

train.data3 <- data.frame(fert3 = train3$fertility_rate_children_per_woman, prin_comp$3x)

fert_rpart3 <- rpart(fert3 ~ .,data = train.data3, method = “anova”)

test.data3 <- predict(prin_comp3, newdata = test3) test.data3 <- as.data.frame(test.data3)

pred3 <- predict(fert_rpart3, test.data3)

error <- pred3-test3$fertility_rate_children_per_woman rmse3 <- sqrt(mean(error^2)) rmse3

std_dev3 <- prin_comp3$sdev pr_var3 <- std_dev3^2 prop_varex3 <- pr_var3/sum(pr_var3) plot(prop_varex3, xlab = “Principal Component”, ylab = “Proportion of Variance Explained”, type = “b”)

circle <- function(center = c(0, 0), npoints = 100) { r = 1 tt = seq(0, 2 * pi, length = npoints) xx = center[1] + r * cos(tt) yy = center[1] + r * sin(tt) return(data.frame(x = xx, y = yy)) } corcir = circle(c(0, 0), npoints = 100)

correlations3 = as.data.frame(cor(train3, prin_comp3$x))

arrows = data.frame(x1 = c(0,0,0,0,0,0), y1 = c(0,0,0,0,0,0), x2 = correlations3$PC1, y2 = correlations$3PC2)

ggplot() + geom_path(data = corcir, aes(x = x, y = y), colour = “gray65”) + geom_segment(data = arrows, aes(x = x1, y = y1, xend = x2, yend = y2), colour = “gray65”) + geom_text(data = correlations3, aes(x = PC1, y = PC2, label = rownames(correlations3))) + geom_hline(yintercept = 0, colour = “gray65”) + geom_vline(xintercept = 0, colour = “gray65”) + xlim(-1.1, 1.1) + ylim(-1.1, 1.1) + labs(x = “pc1 aixs”, y = “pc2 axis”) + ggtitle(“PCA - Model 3 - Circle of correlations”)

prin_comp3$rotation

code reference https://rstudio-pubs-static.s3.amazonaws.com/2897_9220b21cfc0c43a396ff9abf122bb351.html

if (!require(‘leaps’)) (install.packages(‘leaps’))

library(leaps)

regsubsets.out <- regsubsets(fertility_rate_children_per_woman ~ ., data = train1, nbest = 1, # 1 best model for each number of predictors nvmax = NULL, # NULL for no limit on number of variables force.in = NULL, force.out = NULL, method = “exhaustive”) summary.out <- summary(regsubsets.out)

which.max(summary.out$adjr2)

summary.out$which[5,]

summary(best.model <- lm(fertility_rate_children_per_woman ~ cell_subs_per_100 + percent_children_malaria_nets + urban_pop_in_slums + unmet_family_planning_need + women_men_literacy_ratio, data = train1))

pred.model4 <- predict(best.model, newdata=test1, type="response") error.model4 <- pred.model4-test1$fertility_rate_children_per_woman rmse.model4 <- sqrt(mean(error.model4^2)) rmse.model4

if (!require('pscl')) (install.packages('pscl')) if (!require('MASS')) (install.packages('MASS')) library(MASS) library(pscl)

count.train1 <- train1 count.test1 <- test1

int.min.train1 <- (min(count.train1$fertility_rate_children_per_woman)*1000)int.min.test1 < -(min(count.test1$fertility_rate_c
* 1000)

count.train1$fertility_rate_children_per_woman < -(count.train1$fertility_rate_children_per_woman * 1000) - int.min.train1 count.test1$fertility_rate_children_per_woman < -(count.test1$fertility_rate_children_per_woman * 1000) - int.min.test1

summary(poisson.model <- glm(fertility_rate_children_per_woman ~ ., data=count.train1, family="poisson"), trace = FALSE)

summary(nb.model <- step(glm.nb(fertility_rate_children_per_woman ~ . , data = count.train1), trace = FALSE))

summary(zeroinf.model <- zeroinfl(fertility_rate_children_per_woman ~ . |percent_children_malaria_nets, data = count.train1))

pred.model5 <- ((predict(poisson.model, newdata=count.test1, type="response") + + int.min.test1) / 1000) error.model5 <- ((pred.model5 + int.min.test1) / 1000) - test1$fertility_rate_children_per_woman rmse.model5 <- sqrt(mean(error.model5^2))

pred.model6 <- ((predict(nb.model, newdata=count.test1, type="response") + + int.min.test1) / 1000) error.model6 <- ((pred.model6 + int.min.test1) / 1000) - test1$fertility_rate_children_per_woman rmse.model6 <- sqrt(mean(error.model6^2))

pred.model7 <- ((predict(zeroinf.model, newdata=count.test1, type="response") + + int.min.test1) / 1000) error.model7 <- ((pred.model7 + int.min.test1) / 1000) - test1$fertility_rate_children_per_woman rmse.model7 <- sqrt(mean(error.model7^2))

count.models <- c("Poisson", "Negative Binomial", "Zero Inflated") count.rmse <- c(rmse.model5, rmse.model6, rmse.model7) count.prediction.results <- as.data.frame(cbind(count.models,count.rmse)) count.prediction.results