Data for Good Hackathon SDG Forecast

Team 8: Asmita Ghoshal, Cristhian Gutierrez

3 June 2021

```
##
## 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

## [1] 254

## [1] 144

## [1] 138

## [1] 6

## [1] "Bolivia"                    "Congo, Democratic Republic"
## [3] "C<f4>te d'Ivoire"           "Moldova"
## [5] "Venezuela"                  "Vietnam"

## 'data.frame':    149932 obs. of  22 variables:
##  $ Ã¯..Goal          : int  4 4 4 4 4 4 4 4 4 4 ...
##  $ Target           : Factor w/ 9 levels "4.1","4.2","4.3",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Indicator        : Factor w/ 11 levels "4.1.1","4.1.2",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ SeriesCode       : Factor w/ 29 levels "DC_TOF_SCHIPSL",..: 26 26 26 26 26 26 26 26 26 26 ...
##  $ SeriesDescription: Factor w/ 29 levels "Adjusted gender parity index for completion rate, by sex,
##  $ GeoAreaCode      : int  4 4 4 4 4 4 4 4 8 8 ...
##  $ GeoAreaName      : Factor w/ 254 levels "Afghanistan",..: 1 1 1 1 1 1 1 1 3 3 ...
##  $ TimePeriod       : int  2013 2013 2016 2016 2016 2016 2016 2016 2000 2000 ...
##  $ Value            : num  11 13 21.5 22.5 22 ...
##  $ Time_Detail      : Factor w/ 30 levels "2000","2001",..: 16 16 21 21 21 21 21 21 1 1 ...
##  $ BasePeriod       : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ Source           : Factor w/ 306 levels "Adult Education Survey (AES).",..: 170 169 168 168 168 1
##  $ FootNote         : Factor w/ 14 levels "","age 15-49",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Age              : Factor w/ 4 levels "","16-65","M36T47",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Education.level  : Factor w/ 8 levels "","_T","GRAD23",..: 6 6 3 3 3 3 3 3 4 4 ...
##  $ Location         : Factor w/ 4 levels "","ALLAREA","RURAL",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Nature           : Factor w/ 3 levels "C","CA","E": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Quantile         : Factor w/ 7 levels "","_T","Q1","Q2",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Reporting.Type   : Factor w/ 1 level "G": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Sex              : Factor w/ 4 levels "","BOTHSEX","FEMALE",..: 2 2 4 3 2 4 3 2 4 3 ...
##  $ Type.of.skill    : Factor w/ 14 levels "","ARSP","CMFL",..: 11 12 12 12 12 11 11 11 12 12 ...
##  $ Units            : Factor w/ 3 levels "CON_USD","PERCENT",..: 2 2 2 2 2 2 2 2 2 2 ...

## [1] 21

## [1] 19

## [1] 1
```

Numder of NA's acoss dimensions.

```
## $Ã¯..Goal
## [1] 0
##
## $Target
## [1] 0
##
## $Indicator
## [1] 0
##
## $SeriesCode
## [1] 0
##
## $SeriesDescription
## [1] 0
##
## $GeoAreaCode
## [1] 0
##
## $GeoAreaName
## [1] 0
##
## $TimePeriod
## [1] 0
##
## $Value
## [1] 10675
##
## $Time_Detail
## [1] 0
##
## $BasePeriod
## [1] 147592
##
## $Source
## [1] 0
##
## $FootNote
## [1] 0
##
## $Age
## [1] 0
##
## $Education.level
## [1] 0
##
## $Location
## [1] 0
##
## $Nature
## [1] 0
##
## $Quantile
## [1] 0
##
## $Reporting.Type
## [1] 0
##
## $Sex
## [1] 0
##
## $Type.of.skill
## [1] 0
##
## $Units
## [1] 0
```

As we are building a forecasting model using data for 2000-2018 and the variable “Baseperiod”" only takes the value 2018 for some countries we deem this variable not significant in terms of modelling SDG.

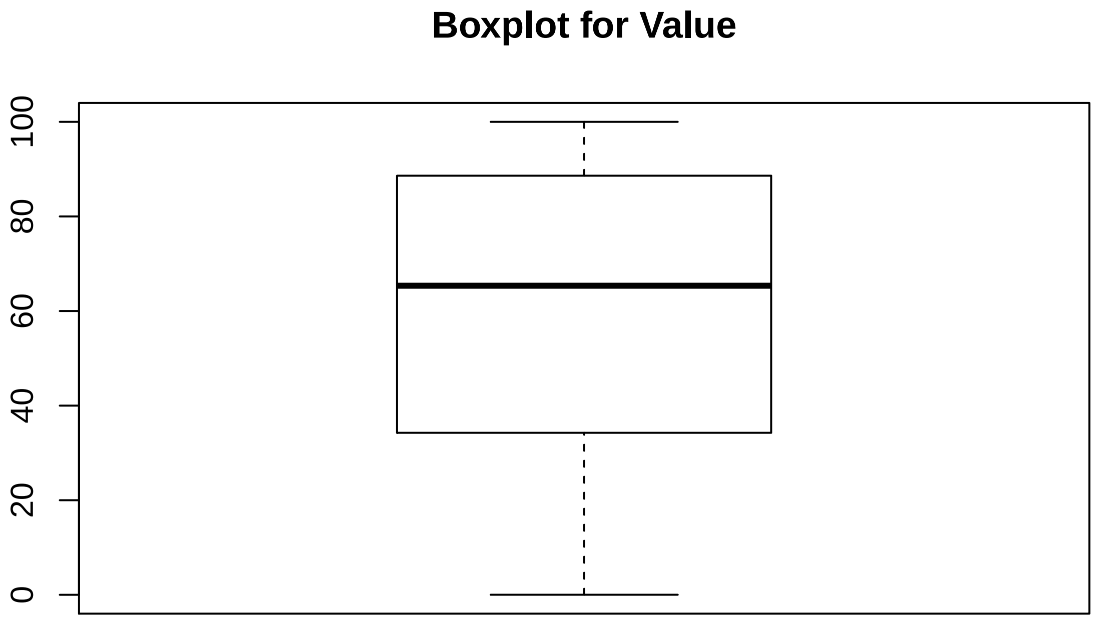
```
## [1] G
## Levels: G
```

Based on the information available in ‘https://en.unesco.org/gem-report/sdg-goal-4’ we deem that it makes sense to buid 9 different models corresponding to each of the 9 targets. The indicators are used to scrutinize subdivisions of the 9 major aspects aka targets in this context.Due to the time constraint the best choice is building different models across different targets.

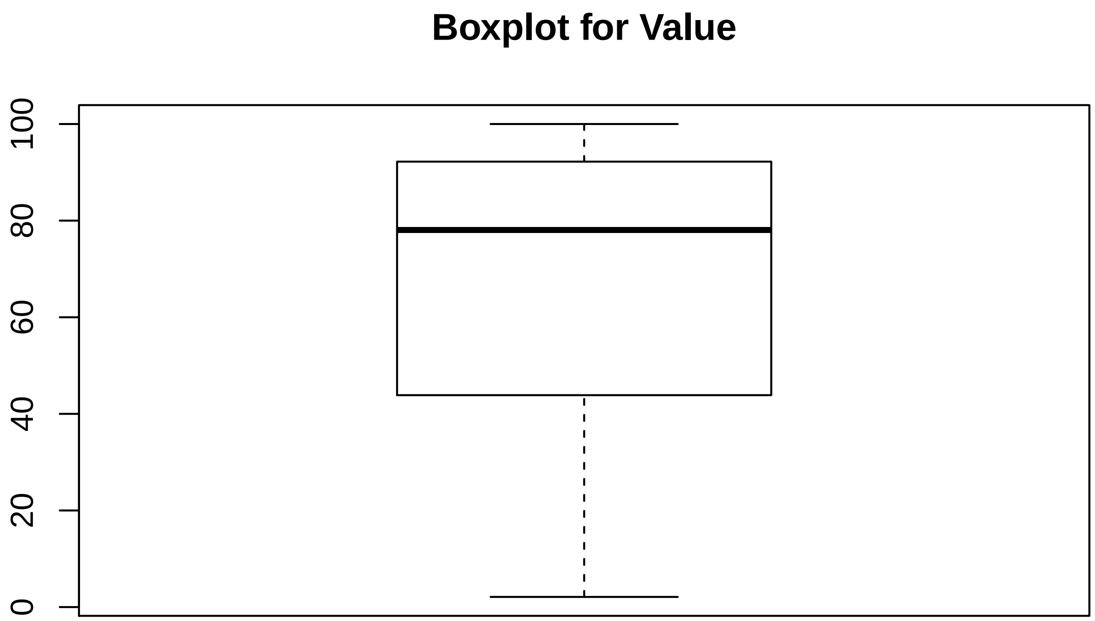
Exploratory Data Analysis by summarising the key information across different dimensions

We will be conducting an exploratory data analysis for each target.

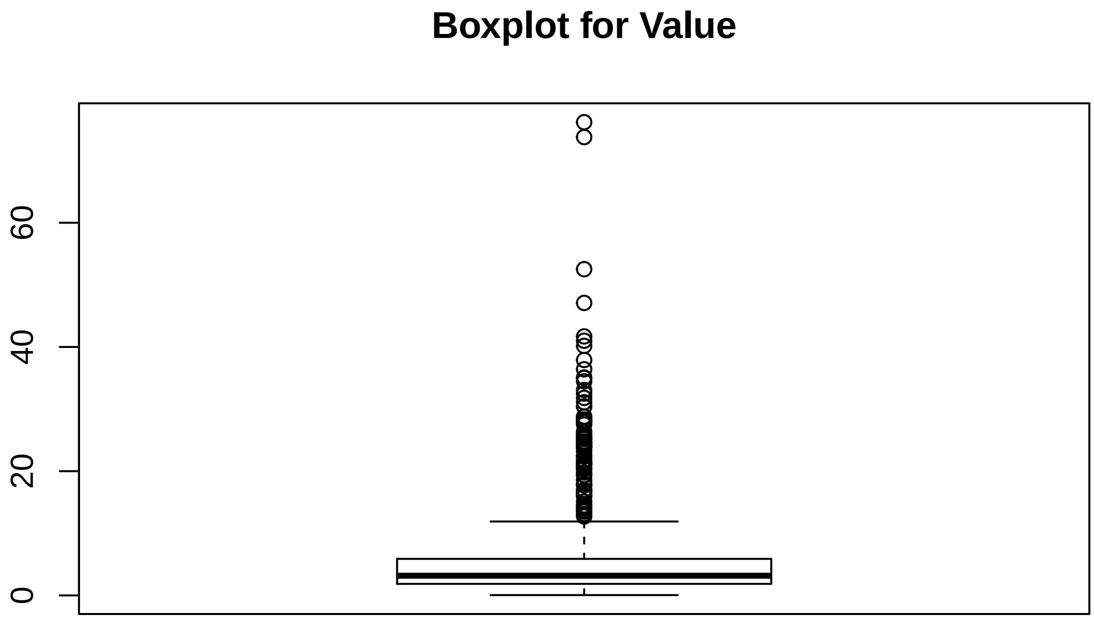
```
## [1] "EDA corresponding to the sectors highlighted under the target 4.1"
```



```
##    Age Education.level Location     Sex Type.of.skill Target    n
## 1               GRAD23          BOTHSEX    SKILL_MATH    4.1   74
## 2               GRAD23          BOTHSEX    SKILL_READ    4.1   86
## 3               GRAD23           FEMALE    SKILL_MATH    4.1   71
## 4               GRAD23           FEMALE    SKILL_READ    4.1   78
## 5               GRAD23             MALE    SKILL_MATH    4.1   71
## 6               GRAD23             MALE    SKILL_READ    4.1   78
## 7               LOWSEC          BOTHSEX    SKILL_MATH    4.1  180
## 8               LOWSEC          BOTHSEX    SKILL_READ    4.1  163
## 9               LOWSEC           FEMALE    SKILL_MATH    4.1  171
## 10              LOWSEC           FEMALE    SKILL_READ    4.1  154
## 11              LOWSEC             MALE    SKILL_MATH    4.1  171
## 12              LOWSEC             MALE    SKILL_READ    4.1  154
## 13              LOWSEC  ALLAREA BOTHSEX                  4.1 2931
## 14              LOWSEC  ALLAREA  FEMALE                  4.1 2481
## 15              LOWSEC  ALLAREA    MALE                  4.1 2481
## 16              LOWSEC    RURAL BOTHSEX                  4.1 2475
## 17              LOWSEC    RURAL  FEMALE                  4.1 2368
## 18              LOWSEC    RURAL    MALE                  4.1 2368
## 19              LOWSEC    URBAN BOTHSEX                  4.1 2614
## 20              LOWSEC    URBAN  FEMALE                  4.1 2506
## 21              LOWSEC    URBAN    MALE                  4.1 2506
## 22              PRIMAR          BOTHSEX    SKILL_MATH    4.1   86
## 23              PRIMAR          BOTHSEX    SKILL_READ    4.1   76
## 24              PRIMAR           FEMALE    SKILL_MATH    4.1   77
## 25              PRIMAR           FEMALE    SKILL_READ    4.1   63
## 26              PRIMAR             MALE    SKILL_MATH    4.1   77
## 27              PRIMAR             MALE    SKILL_READ    4.1   63
## 28              PRIMAR  ALLAREA BOTHSEX                  4.1 2847
## 29              PRIMAR  ALLAREA  FEMALE                  4.1 2467
## 30              PRIMAR  ALLAREA    MALE                  4.1 2467
## 31              PRIMAR    RURAL BOTHSEX                  4.1 2461
## 32              PRIMAR    RURAL  FEMALE                  4.1 2368
## 33              PRIMAR    RURAL    MALE                  4.1 2368
## 34              PRIMAR    URBAN BOTHSEX                  4.1 2600
## 35              PRIMAR    URBAN  FEMALE                  4.1 2506
## 36              PRIMAR    URBAN    MALE                  4.1 2506
## 37              UPPSEC  ALLAREA BOTHSEX                  4.1 2913
## 38              UPPSEC  ALLAREA  FEMALE                  4.1 2463
## 39              UPPSEC  ALLAREA    MALE                  4.1 2463
## 40              UPPSEC    RURAL BOTHSEX                  4.1 2457
## 41              UPPSEC    RURAL  FEMALE                  4.1 2350
## 42              UPPSEC    RURAL    MALE                  4.1 2350
## 43              UPPSEC    URBAN BOTHSEX                  4.1 2596
## 44              UPPSEC    URBAN  FEMALE                  4.1 2488
## 45              UPPSEC    URBAN    MALE                  4.1 2488
## [1] "EDA corresponding to the sectors highlighted under the target 4.2"
```

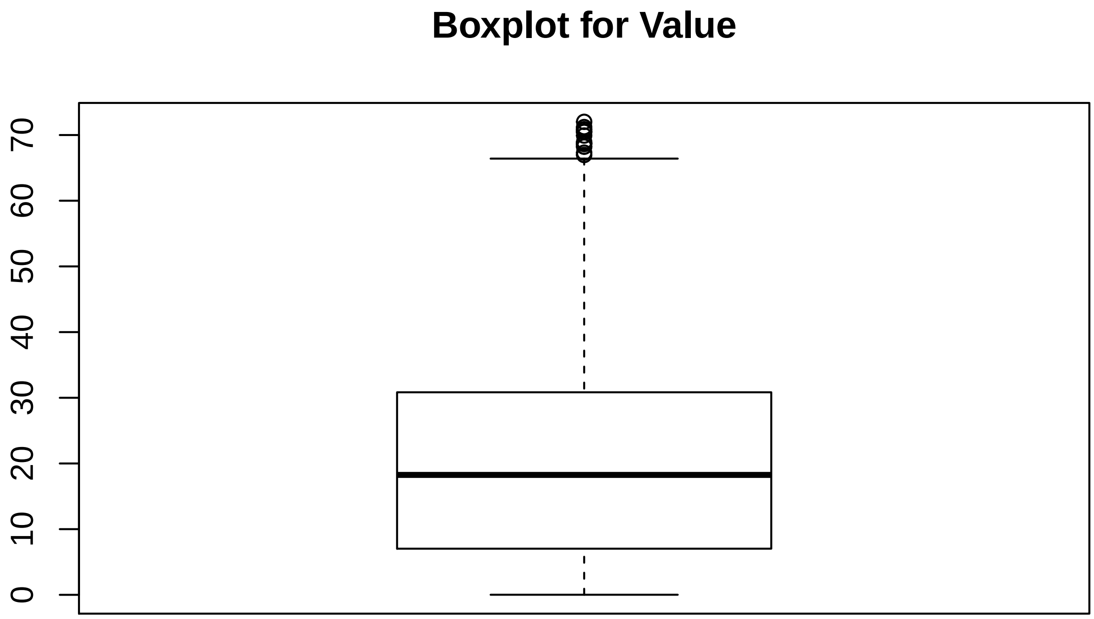


```
##      Age Education.level Location     Sex Type.of.skill Target   n
## 1                                 BOTHSEX                  4.2 939
## 2                                  FEMALE                  4.2 919
## 3                                    MALE                  4.2 919
## 4 M36T47                          BOTHSEX                  4.2   1
## 5 M36T47                           FEMALE                  4.2   1
## 6 M36T47                             MALE                  4.2   1
## 7 M36T59                          BOTHSEX                  4.2  50
## 8 M36T59                           FEMALE                  4.2  49
## 9 M36T59                             MALE                  4.2  49
## [1] "EDA corresponding to the sectors highlighted under the target 4.3"
```

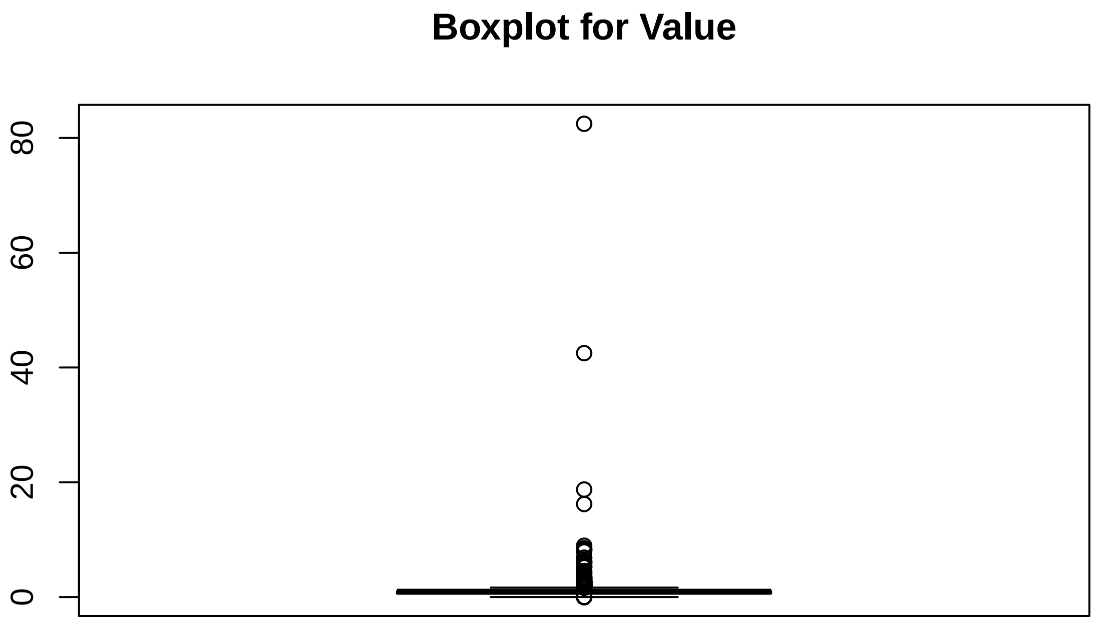


```
##   Age Education.level Location     Sex Type.of.skill Target   n
## 1                              BOTHSEX                  4.3 241
## 2                               FEMALE                  4.3 206
## 3                                 MALE                  4.3 209
## [1] "EDA corresponding to the sectors highlighted under the target 4.4"
```

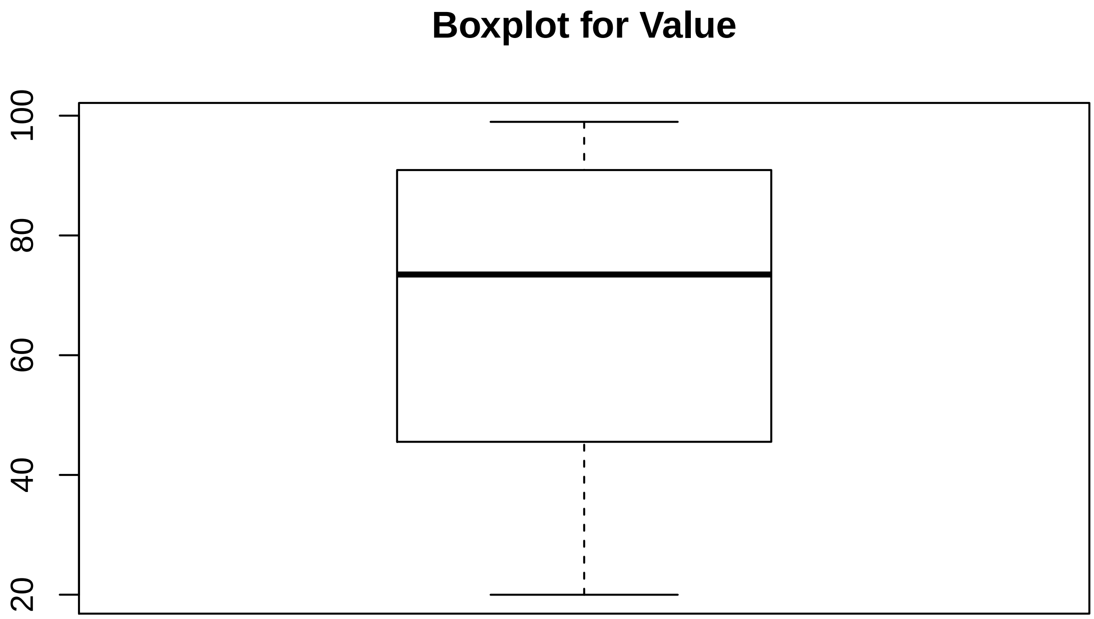
Boxplot for Value



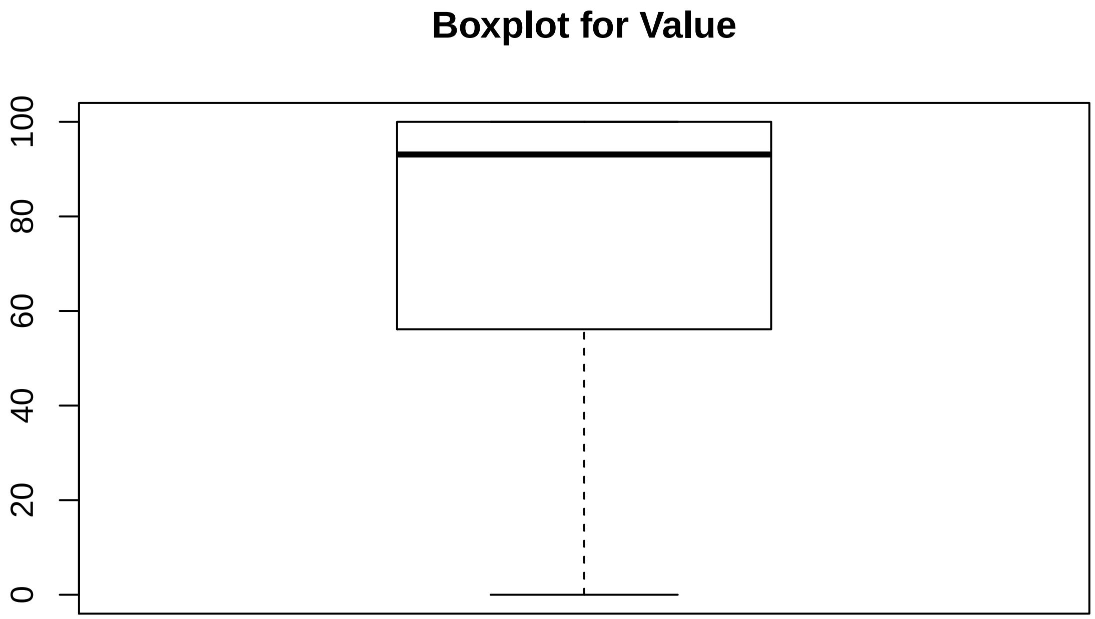
```
##     Age Education.level Location     Sex Type.of.skill Target   n
## 1                               BOTHSEX          ARSP    4.4  97
## 2                               BOTHSEX          CMFL    4.4 103
## 3                               BOTHSEX          COPA    4.4  86
## 4                               BOTHSEX         EMAIL    4.4  79
## 5                               BOTHSEX          EPRS    4.4 108
## 6                               BOTHSEX          INST    4.4  78
## 7                               BOTHSEX          PCPR    4.4  99
## 8                               BOTHSEX          SOFT    4.4  89
## 9                               BOTHSEX          TRAF    4.4 105
## 10                               FEMALE          ARSP    4.4  82
## 11                               FEMALE          CMFL    4.4  86
## 12                               FEMALE          COPA    4.4  70
## 13                               FEMALE         EMAIL    4.4  57
## 14                               FEMALE          EPRS    4.4  89
## 15                               FEMALE          INST    4.4  65
## 16                               FEMALE          PCPR    4.4  83
## 17                               FEMALE          SOFT    4.4  74
## 18                               FEMALE          TRAF    4.4  86
## 19                                 MALE          ARSP    4.4  79
## 20                                 MALE          CMFL    4.4  83
## 21                                 MALE          COPA    4.4  67
## 22                                 MALE         EMAIL    4.4  54
## 23                                 MALE          EPRS    4.4  86
## 24                                 MALE          INST    4.4  62
## 25                                 MALE          PCPR    4.4  80
## 26                                  MALE          SOFT    4.4  71
## 27                                  MALE          TRAF    4.4  83
## [1] "EDA corresponding to the sectors highlighted under the target 4.5"
```



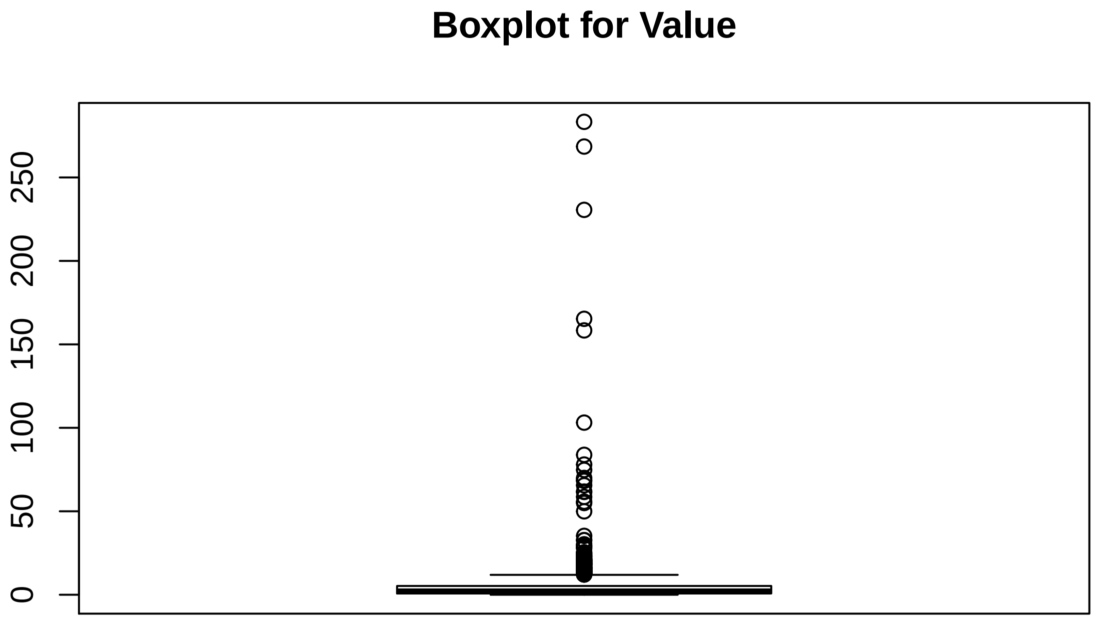
```
##    Age Education.level Location      Sex Type.of.skill Target    n
## 1                                                         4.5 1127
## 2                                                 ARSP    4.5   10
## 3                                                 CMFL    4.5   11
## 4                                                 COPA    4.5    9
## 5                                                EMAIL    4.5    4
## 6                                                 EPRS    4.5   11
## 7                                                 INST    4.5    5
## 8                                                 LITE    4.5    3
## 9                                                 NUME    4.5    3
## 10                                                PCPR    4.5   10
## 11                                                SOFT    4.5    9
## 12                                                TRAF    4.5   10
## 13                  _T                            LITE    4.5   28
## 14                  _T                            NUME    4.5   16
## 15              GRAD23                      SKILL_MATH    4.5  191
## 16              GRAD23                      SKILL_READ    4.5  188
## 17              LOWSEC                                    4.5  430
## 18              LOWSEC                      SKILL_MATH    4.5  754
## 19              LOWSEC                      SKILL_READ    4.5  689
## 20              LOWSEC           BOTHSEX                  4.5 2475
## 21              LOWSEC            FEMALE                  4.5 2368
## 22              LOWSEC             MALE                   4.5 2368
## 23              LOWSEC  ALLAREA                           4.5 2481
## 24              LOWSEC  ALLAREA BOTHSEX                   4.5  481
## 25              LOWSEC  ALLAREA  FEMALE                   4.5  391
## 26              LOWSEC  ALLAREA    MALE                   4.5  391
## 27              LOWSEC    RURAL                           4.5 2368
## 28              LOWSEC    RURAL BOTHSEX                   4.5  391
## 29              LOWSEC    RURAL  FEMALE                   4.5  391
## 30              LOWSEC    RURAL    MALE                   4.5  391
## 31              LOWSEC    URBAN                           4.5 2506
## 32              LOWSEC    URBAN BOTHSEX                   4.5  414
## 33              LOWSEC    URBAN  FEMALE                   4.5  414
## 34              LOWSEC    URBAN    MALE                   4.5  414
## 35              PREPRI                                    4.5  449
## 36              PRIMAR                                    4.5  854
## 37              PRIMAR                      SKILL_MATH    4.5  233
## 38              PRIMAR                      SKILL_READ    4.5  172
## 39              PRIMAR          BOTHSEX                   4.5 2461
## 40              PRIMAR           FEMALE                   4.5 2368
## 41              PRIMAR             MALE                   4.5 2368
## 42              PRIMAR  ALLAREA                           4.5 2467
## 43              PRIMAR  ALLAREA BOTHSEX                   4.5  467
## 44              PRIMAR  ALLAREA  FEMALE                   4.5  391
## 45              PRIMAR  ALLAREA    MALE                   4.5  391
## 46              PRIMAR    RURAL                           4.5 2368
## 47              PRIMAR    RURAL BOTHSEX                   4.5  391
## 48              PRIMAR    RURAL  FEMALE                   4.5  391
## 49              PRIMAR    RURAL    MALE                   4.5  391
## 50              PRIMAR    URBAN                           4.5 2506
## 51              PRIMAR    URBAN BOTHSEX                   4.5  414
## 52              PRIMAR    URBAN  FEMALE                   4.5  414
## 53              PRIMAR    URBAN    MALE                   4.5  414
## 54              SECOND                                    4.5  488
## 55              UPPSEC                                    4.5  360
## 56              UPPSEC          BOTHSEX                   4.5 2457
## 57              UPPSEC           FEMALE                   4.5 2350
## 58              UPPSEC             MALE                   4.5 2350
## 59              UPPSEC  ALLAREA                           4.5 2463
## 60              UPPSEC  ALLAREA BOTHSEX                   4.5  478
## 61              UPPSEC  ALLAREA  FEMALE                   4.5  388
## 62              UPPSEC  ALLAREA    MALE                   4.5  388
## 63              UPPSEC    RURAL                           4.5 2350
## 64              UPPSEC    RURAL BOTHSEX                   4.5  388
## 65              UPPSEC    RURAL  FEMALE                   4.5  388
## 66              UPPSEC    RURAL    MALE                   4.5  388
## 67              UPPSEC    URBAN                           4.5 2488
## 68              UPPSEC    URBAN BOTHSEX                   4.5  411
## 69              UPPSEC    URBAN  FEMALE                   4.5  411
## 70              UPPSEC    URBAN    MALE                   4.5  411
## [1] "EDA corresponding to the sectors highlighted under the target 4.6"
```



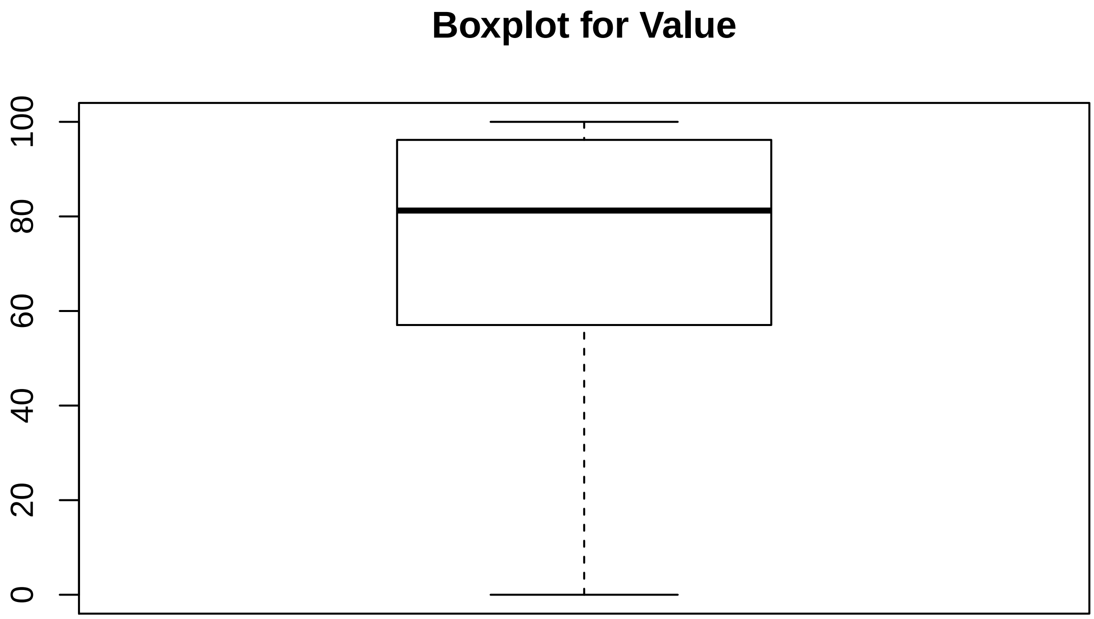
```
##     Age Education.level Location     Sex Type.of.skill Target  n
## 1 16-65                          BOTHSEX          LITE    4.6 29
## 2 16-65                          BOTHSEX          NUME    4.6 10
## 3 16-65                           FEMALE          LITE    4.6 11
## 4 16-65                           FEMALE          NUME    4.6  9
## 5 16-65                             MALE          LITE    4.6 11
## 6 16-65                             MALE          NUME    4.6  9
## [1] "EDA corresponding to the sectors highlighted under the target 4.a"
```



```
##   Age Education.level Location Sex Type.of.skill Target    n
## 1             LOWSEC                                 4.a  931
## 2             PRIMAR                                 4.a 1010
## 3             UPPSEC                                 4.a  964
## [1] "EDA corresponding to the sectors highlighted under the target 4.b"
```



```
##   Age Education.level Location Sex Type.of.skill Target    n
## 1                                                   4.b 1672
## [1] "EDA corresponding to the sectors highlighted under the target 4.c"
```



```
##    Age Education.level Location     Sex Type.of.skill Target   n
## 1              LOWSEC         BOTHSEX                  4.c 481
## 2              LOWSEC          FEMALE                  4.c 431
## 3              LOWSEC            MALE                  4.c 430
## 4              PREPRI         BOTHSEX                  4.c 614
## 5              PREPRI          FEMALE                  4.c 552
## 6              PREPRI            MALE                  4.c 550
## 7              PRIMAR         BOTHSEX                  4.c 974
## 8              PRIMAR          FEMALE                  4.c 854
## 9              PRIMAR            MALE                  4.c 854
## 10             SECOND         BOTHSEX                  4.c 572
## 11             SECOND          FEMALE                  4.c 489
## 12             SECOND            MALE                  4.c 488
## 13             UPPSEC         BOTHSEX                  4.c 413
## 14             UPPSEC          FEMALE                  4.c 361
## 15             UPPSEC            MALE                  4.c 360
```

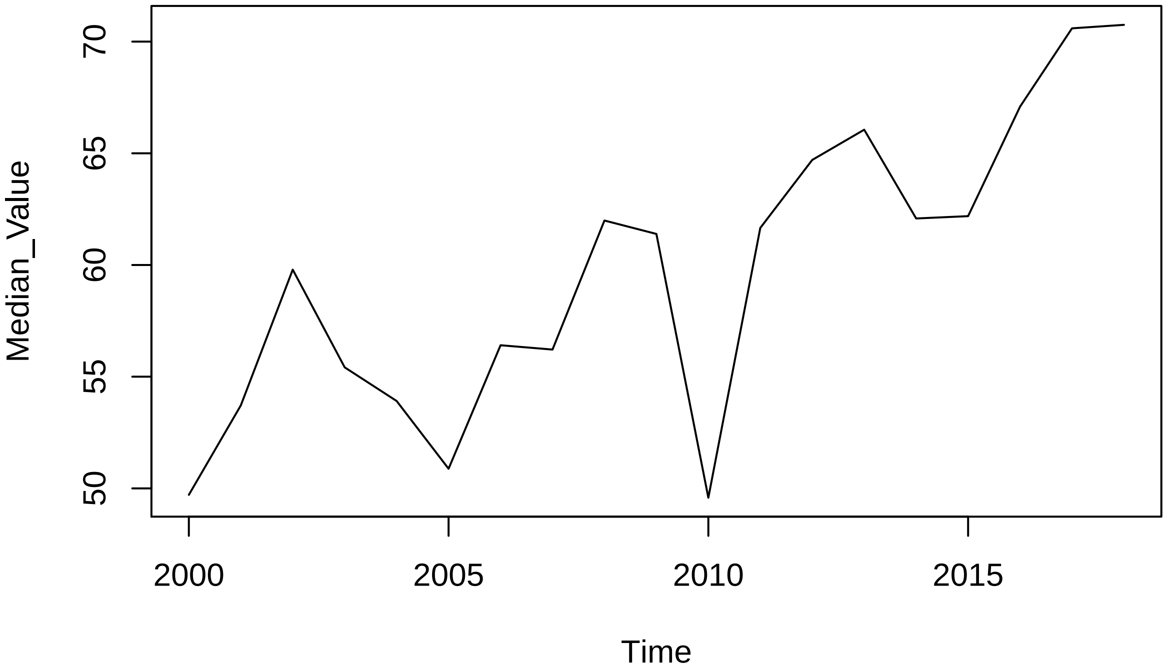
Replacing missing values with median for numerical variables and with the majority class for categorical variables across different years (Time period).As we are interested in the quality education and lifelong learning for all there is no point in creating a segregstion across different countries.

Identifying missing values and replacing them with the median for the variable named "Value" aggregated by time period and target.Count of missing values across dimensions.

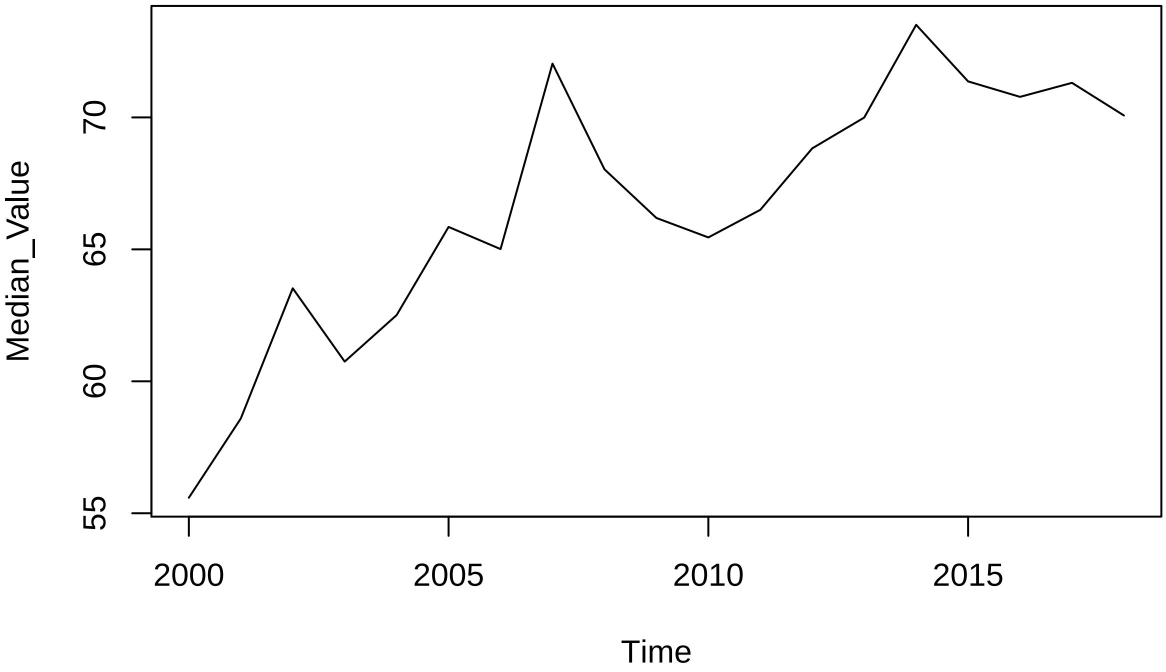
```
## $Target
## [1] 0
##
## $GeoAreaName
## [1] 0
##
## $TimePeriod
## [1] 0
##
## $Value
## [1] 10675
##
## $Age
## [1] 149034
##
## $Education.level
## [1] 8748
##
## $Location
## [1] 48386
##
## $Nature
## [1] 0
##
## $Quantile
## [1] 26821
##
## $Sex
## [1] 32638
##
## $Type.of.skill
## [1] 142735
##
## $Units
## [1] 0
```

Here I will be using the variable “Values” for the purpose of forcasting. Essentially I will fit an ARMA model on the median corresponding to each group specified by target.Here, we are conducting a one step forecast for the year 2019 across different target groups using the median values from 200-2018.

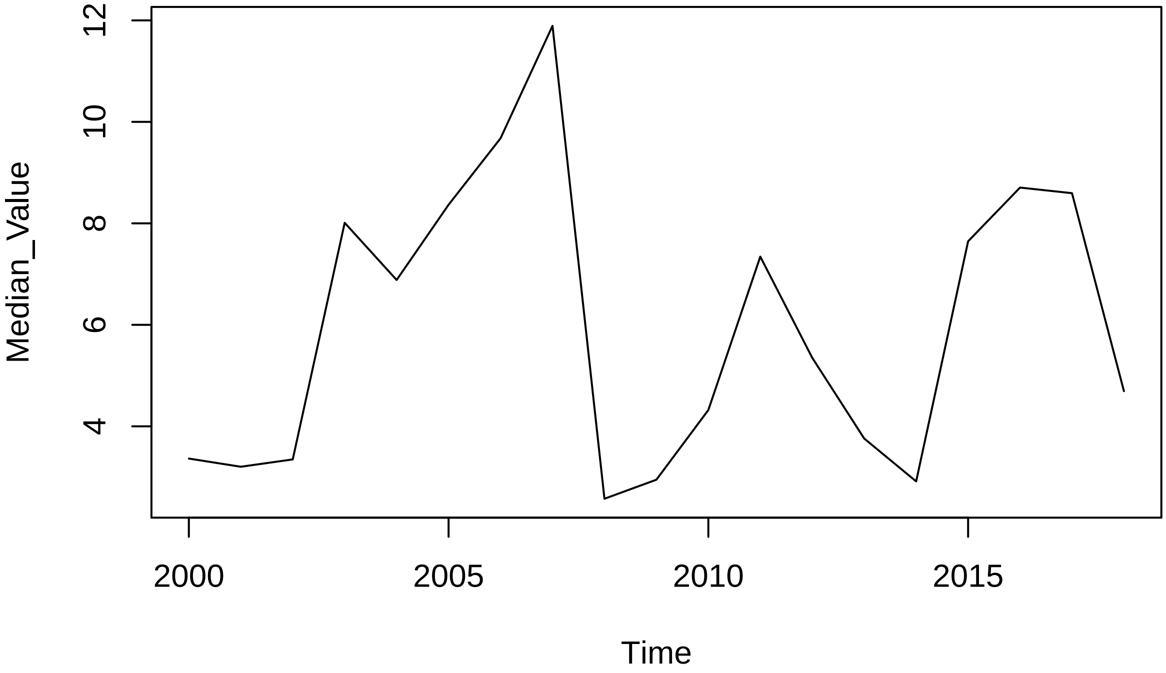
```
## [1] "Median Value Forecast for the year 2019 corresponding to the sectors highlighted under the targe
```



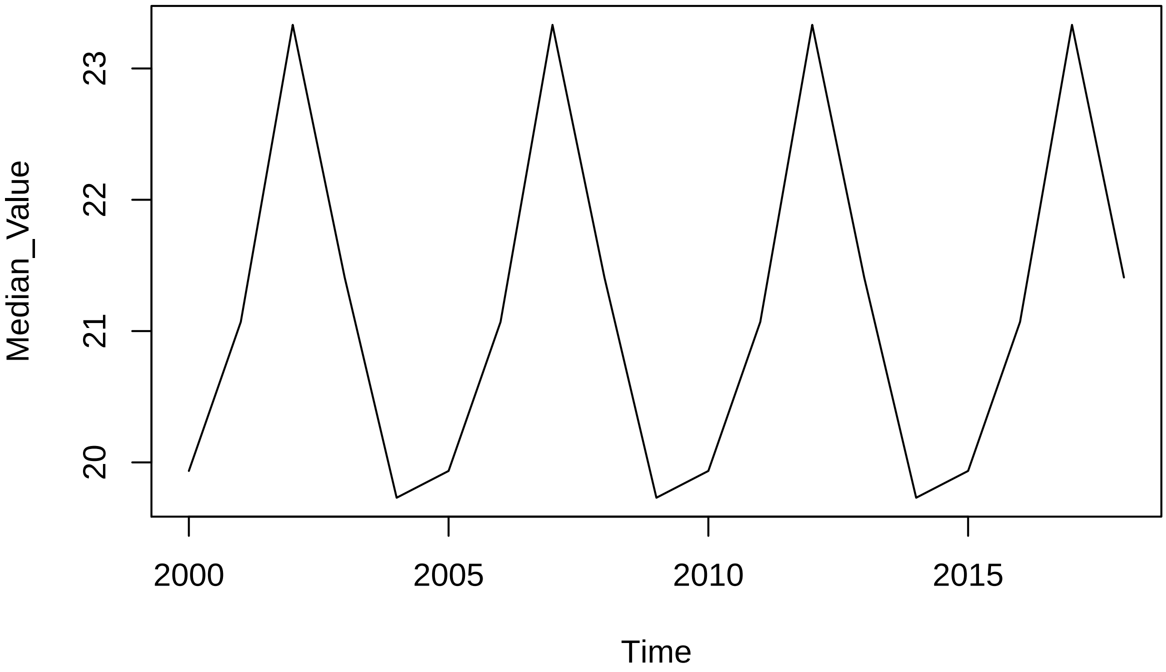
```
## [1] "The predicted median value for the year 2019 is by 69.3694100964184"
## [1] "Median Value Forecast for the year 2019 corresponding to the sectors highlighted under the targ
```



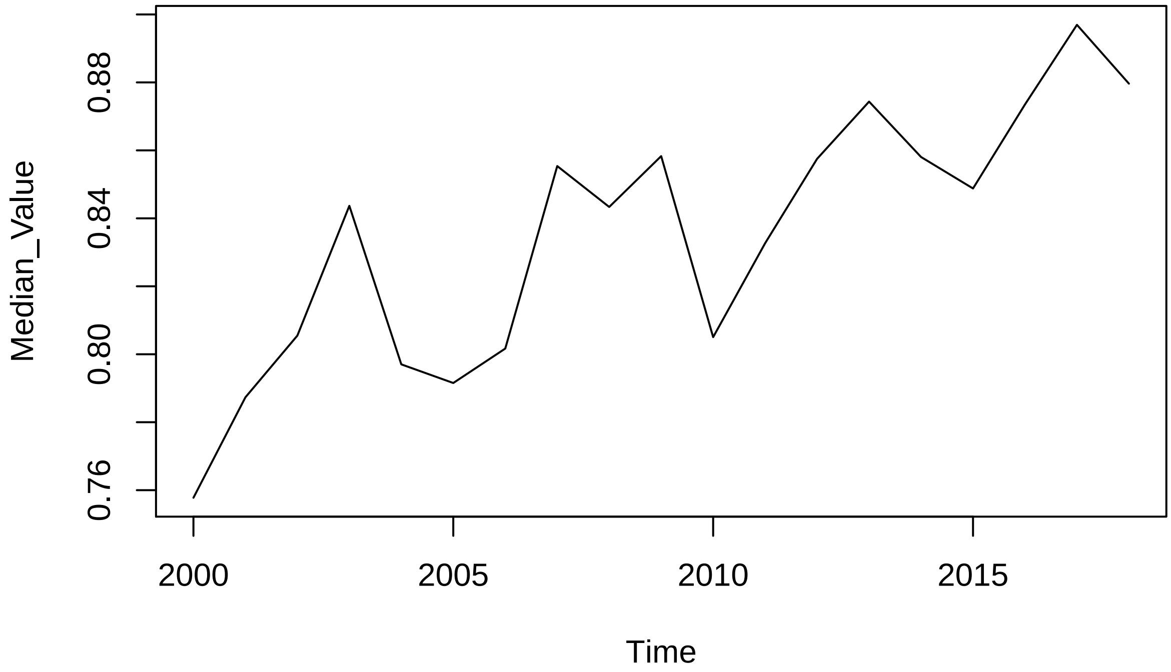
```
## [1] "The predicted median value for the year 2019 is by 70.2593394258754"
## [1] "Median Value Forecast for the year 2019 corresponding to the sectors highlighted under the targ
```



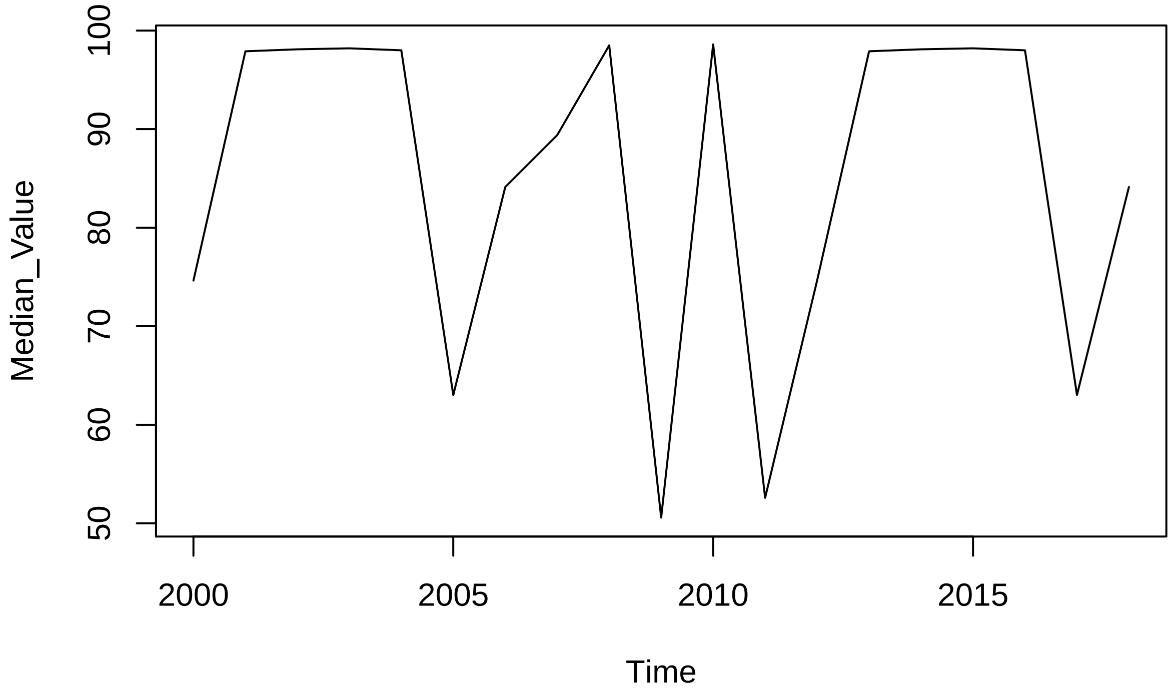
```
## [1] "The predicted median value for the year 2019 is by 6.21047353178157"
## [1] "Median Value Forecast for the year 2019 corresponding to the sectors highlighted under the targ
```



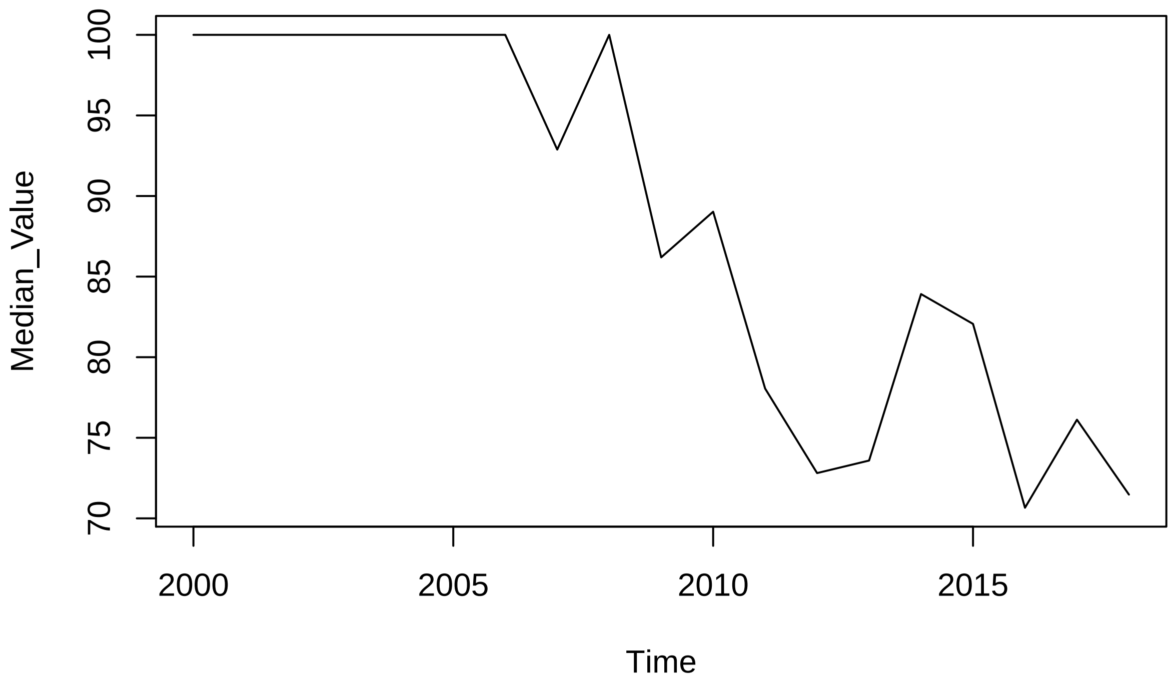
```
## [1] "The predicted median value for the year 2019 is by 20.788332644812"
## [1] "Median Value Forecast for the year 2019 corresponding to the sectors highlighted under the targ
```



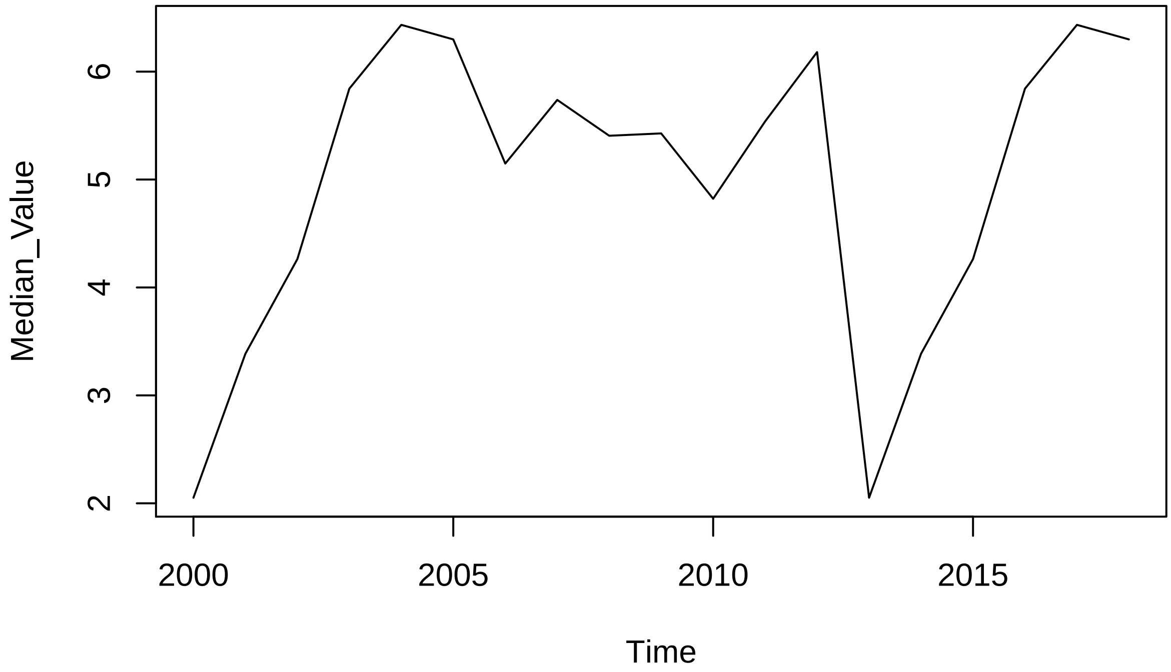
```
## [1] "The predicted median value for the year 2019 is by 0.881813047684052"
## [1] "Median Value Forecast for the year 2019 corresponding to the sectors highlighted under the targ
```



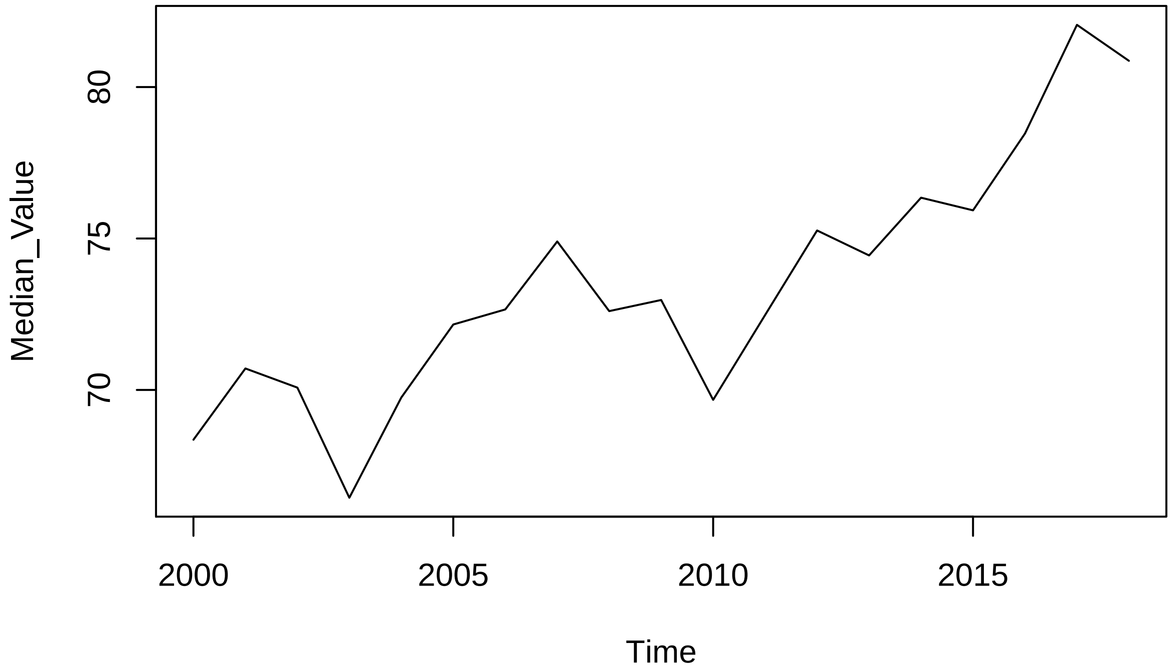
```
## [1] "The predicted median value for the year 2019 is by 85.1392411026445"
## [1] "Median Value Forecast for the year 2019 corresponding to the sectors highlighted under the targ
```



```
## [1] "The predicted median value for the year 2019 is by 72.55965801076"
## [1] "Median Value Forecast for the year 2019 corresponding to the sectors highlighted under the targe
```



```
## [1] "The predicted median value for the year 2019 is by 6.30479373798973"
## [1] "Median Value Forecast for the year 2019 corresponding to the sectors highlighted under the targ
```



```
## [1] "The predicted median value for the year 2019 is by 80.9550933142788"
```