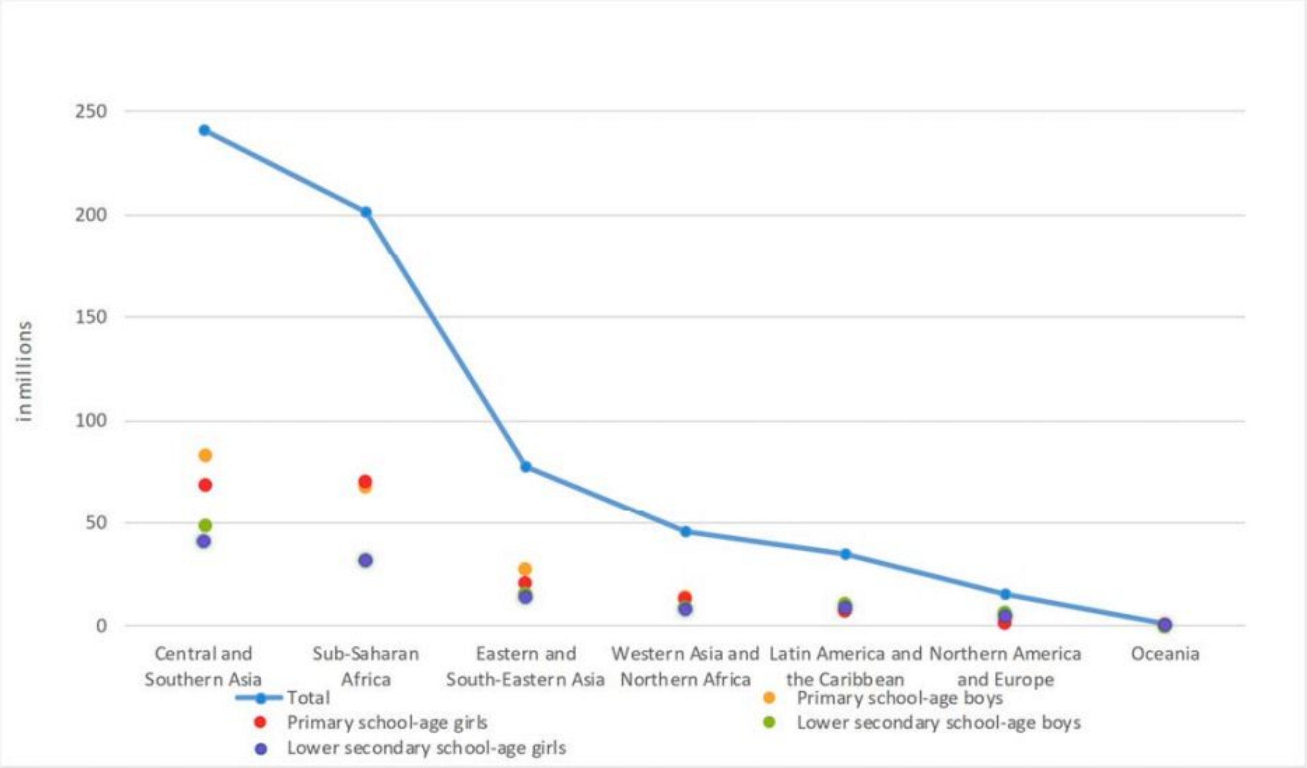
JPMC Data for Good Hackathon Team 19

Ting Xu, Reema Yadav, Yutong Wu, Michael Wieck-Sosa, Lavanya Velagala

Inclusive and equitable quality education is still a main issue

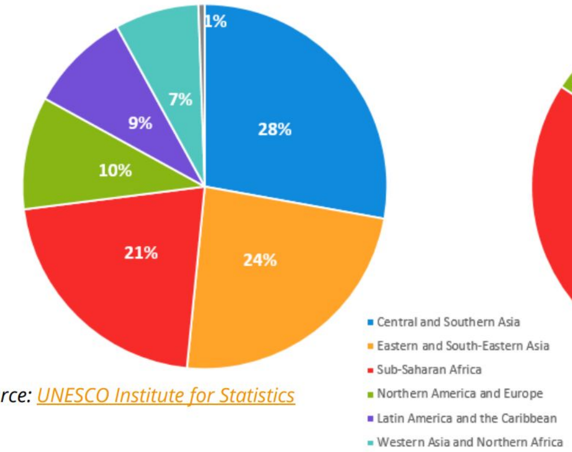


UNESCO. More Than One-Half of Children and Adolescents Are Not Learning Worldwide. UNESCO Inst. Stat. 2017, 67 (46), 25.

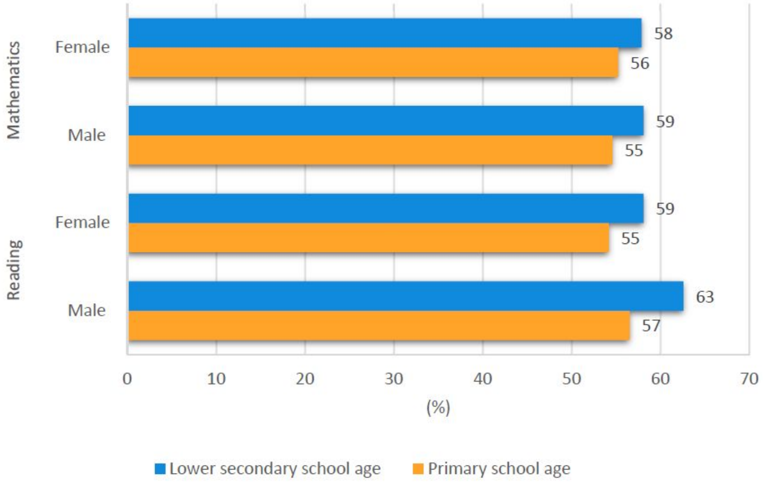
Gender and country can influence education level

Figure 3a. Distribution of the primary and lower secondary school-age population, by region

Figure 3b. Distribution of children and adolescents not learning, by region



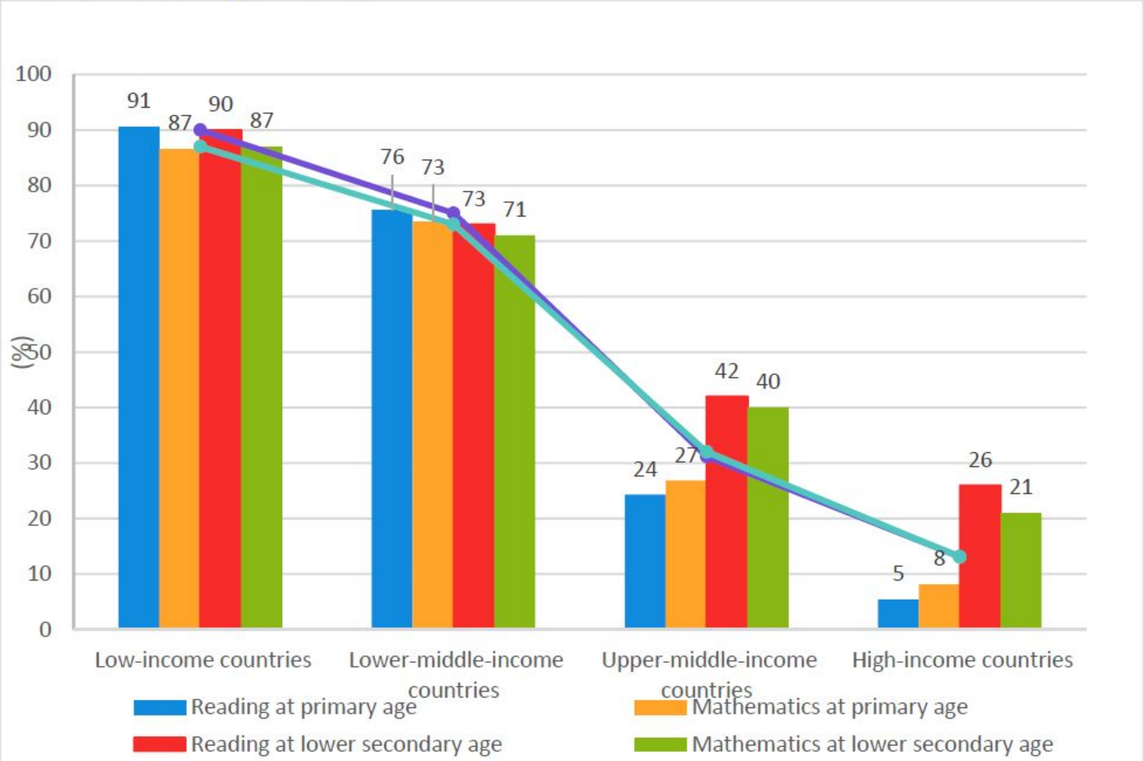
Source: UNESCO Institute for Statistics



UNESCO. More Than One-Half of Children and Adolescents Are Not Learning Worldwide. UNESCO Inst. Stat. 2017, 67 (46), 25.

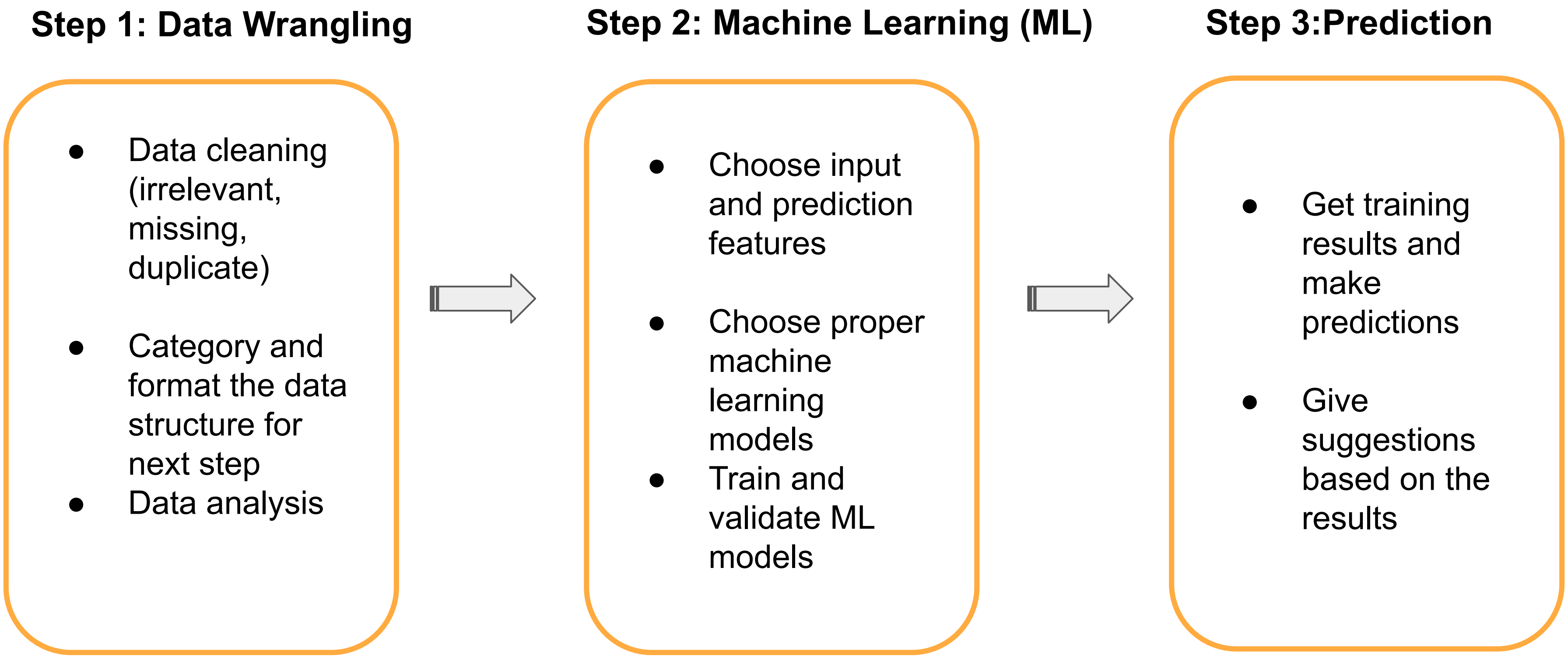
Wealth can all influence the education level

Figure 12. Proportion of children and adolescents not achieving MPLs, by domain and country income grouping



UNESCO. More Than One-Half of Children and Adolescents Are Not Learning Worldwide. UNESCO Inst. Stat. 2017, 67 (46), 25.

Goal: To better understand the current status and predict the inclusive and equitable quality education



Step 1: Data Wrangling

- Data cleaning (irrelevant, missing, duplicate)
- Category and format the data structure for next step
- Data analysis

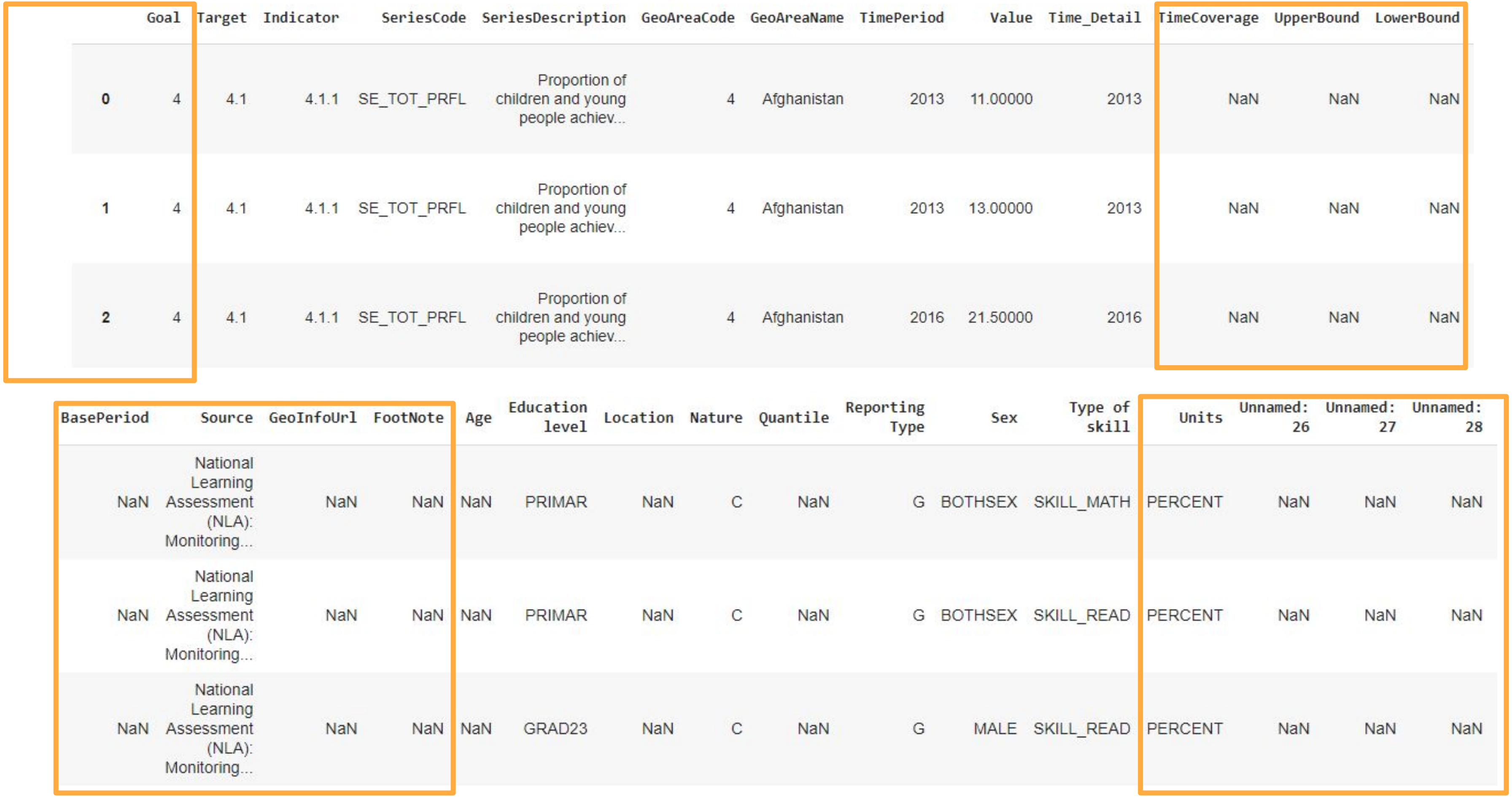
Step 2: Machine Learning (ML)

- Choose input and prediction features
- Choose proper machine learning models
- Train and validate ML models

Step 3:Prediction

- Get training results and make predictions
- Give suggestions based on the results

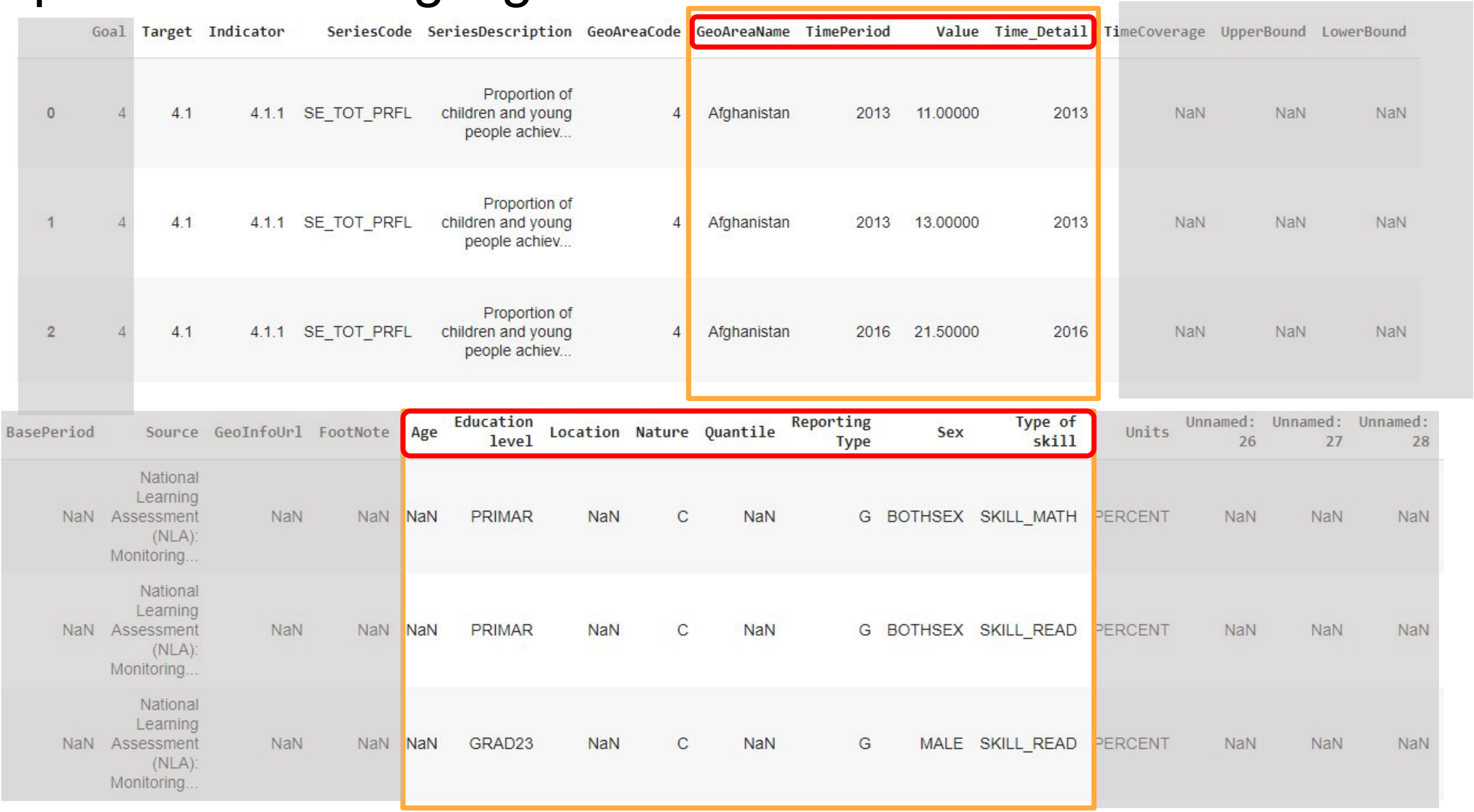
Step 1: Data Wrangling: deleting useless data



	Goal	Target	Indicator	SeriesCode	SeriesDescription	GeoAreaCode	GeoAreaName	TimePeriod	Value	Time_Detail	TimeCoverage	UpperBound	LowerBound
0	4	4.1	4.1.1	SE_TOT_PRFL	Proportion of children and young people achiev...	4	Afghanistan	2013	11.00000	2013	NaN	NaN	NaN
1	4	4.1	4.1.1	SE_TOT_PRFL	Proportion of children and young people achiev...	4	Afghanistan	2013	13.00000	2013	NaN	NaN	NaN
2	4	4.1	4.1.1	SE_TOT_PRFL	Proportion of children and young people achiev...	4	Afghanistan	2016	21.50000	2016	NaN	NaN	NaN

BasePeriod	Source	GeoInfoUrl	FootNote	Age	Education level	Location	Nature	Quantile	Reporting Type	Sex	Type of skill	Units	Unnamed: 26	Unnamed: 27	Unnamed: 28
NaN	National Learning Assessment (NLA): Monitoring...	NaN	NaN	NaN	PRIMAR	NaN	C	NaN	G	BOTHSEX	SKILL_MATH	PERCENT	NaN	NaN	NaN
NaN	National Learning Assessment (NLA): Monitoring...	NaN	NaN	NaN	PRIMAR	NaN	C	NaN	G	BOTHSEX	SKILL_READ	PERCENT	NaN	NaN	NaN
NaN	National Learning Assessment (NLA): Monitoring...	NaN	NaN	NaN	GRAD23	NaN	C	NaN	G	MALE	SKILL_READ	PERCENT	NaN	NaN	NaN

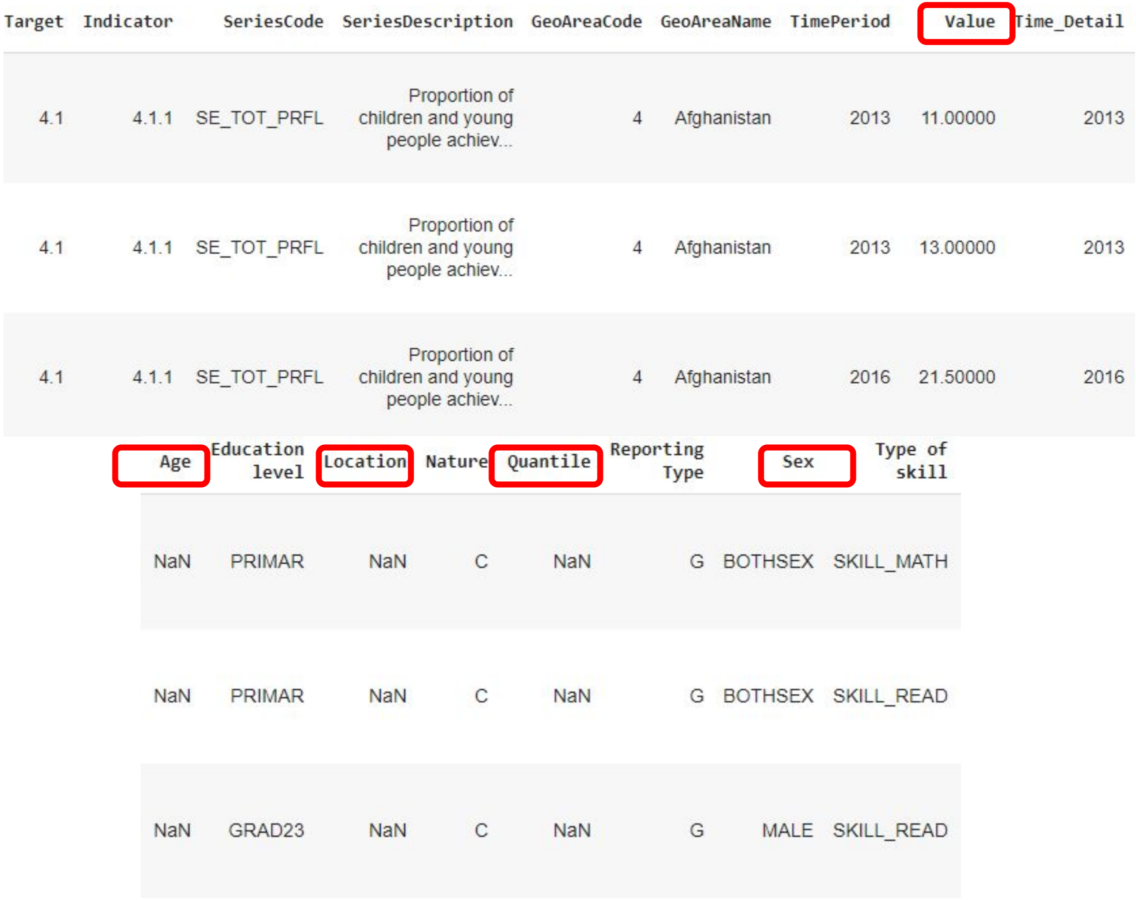
Step 1: Data Wrangling: select useful data



	Goal	Target	Indicator	SeriesCode	SeriesDescription	GeoAreaCode	GeoAreaName	TimePeriod	Value	Time_Detail	TimeCoverage	UpperBound	LowerBound
0	4	4.1	4.1.1	SE_TOT_PRFL	Proportion of children and young people achiev...	4	Afghanistan	2013	11.00000	2013	NaN	NaN	NaN
1	4	4.1	4.1.1	SE_TOT_PRFL	Proportion of children and young people achiev...	4	Afghanistan	2013	13.00000	2013	NaN	NaN	NaN
2	4	4.1	4.1.1	SE_TOT_PRFL	Proportion of children and young people achiev...	4	Afghanistan	2016	21.50000	2016	NaN	NaN	NaN

BasePeriod	Source	GeoInfoUrl	FootNote	Age	Education level	Location	Nature	Quantile	Reporting Type	Sex	Type of skill	Units	Unnamed: 26	Unnamed: 27	Unnamed: 28
NaN	National Learning Assessment (NLA): Monitoring...	NaN	NaN	NaN	PRIMAR	NaN	C	NaN	G	BOTHSEX	SKILL_MATH	PERCENT	NaN	NaN	NaN
NaN	National Learning Assessment (NLA): Monitoring...	NaN	NaN	NaN	PRIMAR	NaN	C	NaN	G	BOTHSEX	SKILL_READ	PERCENT	NaN	NaN	NaN
NaN	National Learning Assessment (NLA): Monitoring...	NaN	NaN	NaN	GRAD23	NaN	C	NaN	G	MALE	SKILL_READ	PERCENT	NaN	NaN	NaN

Step 1: Data Wrangling: filling missing data



Target	Indicator	SeriesCode	SeriesDescription	GeoAreaCode	GeoAreaName	TimePeriod	Value	Time_Detail	Age	Education level	Location	Nature	Quantile	Reporting Type	Sex	Type of skill
4.1	4.1.1	SE_TOT_PRFL	Proportion of children and young people achiev...	4	Afghanistan	2013	11.00000	2013	NaN	PRIMAR	NaN	C	NaN	G	BOTHSEX	SKILL_MATH
4.1	4.1.1	SE_TOT_PRFL	Proportion of children and young people achiev...	4	Afghanistan	2013	13.00000	2013	NaN	PRIMAR	NaN	C	NaN	G	BOTHSEX	SKILL_READ
4.1	4.1.1	SE_TOT_PRFL	Proportion of children and young people achiev...	4	Afghanistan	2016	21.50000	2016	NaN	GRAD23	NaN	C	NaN	G	MALE	SKILL_READ

- Fill [value] based on median of the pair of country and indicator
- Fill [Sex] based on the mode
- Fill [Location] based on the mode of the pair country and indicator
- Fill [Age] based on the education level
- Fill [education level] based on mode of the pair of country and indicator

Step 1: Data Wrangling: Structuring data

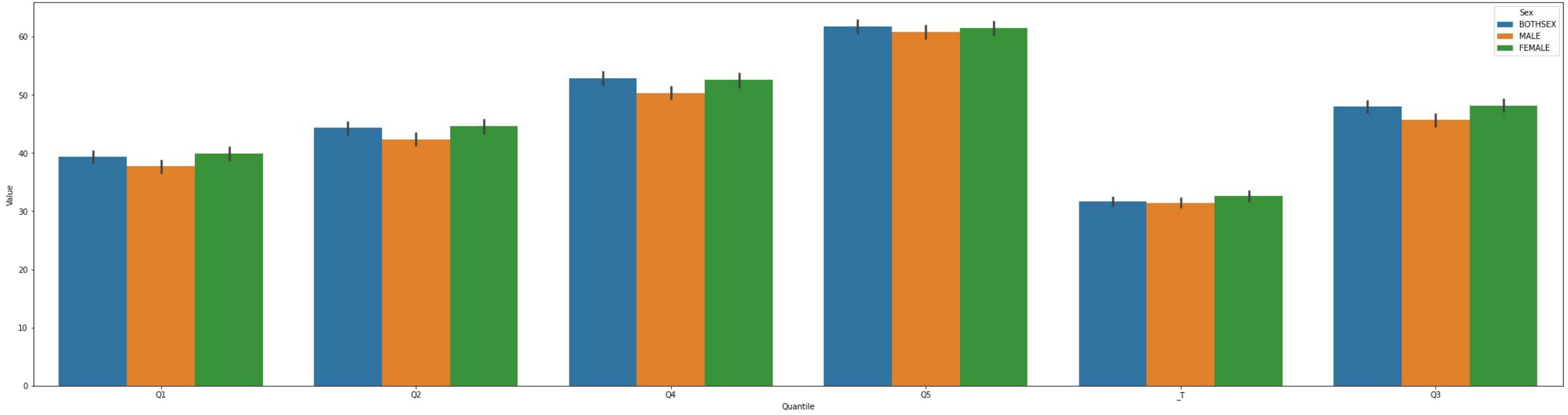
Structure data

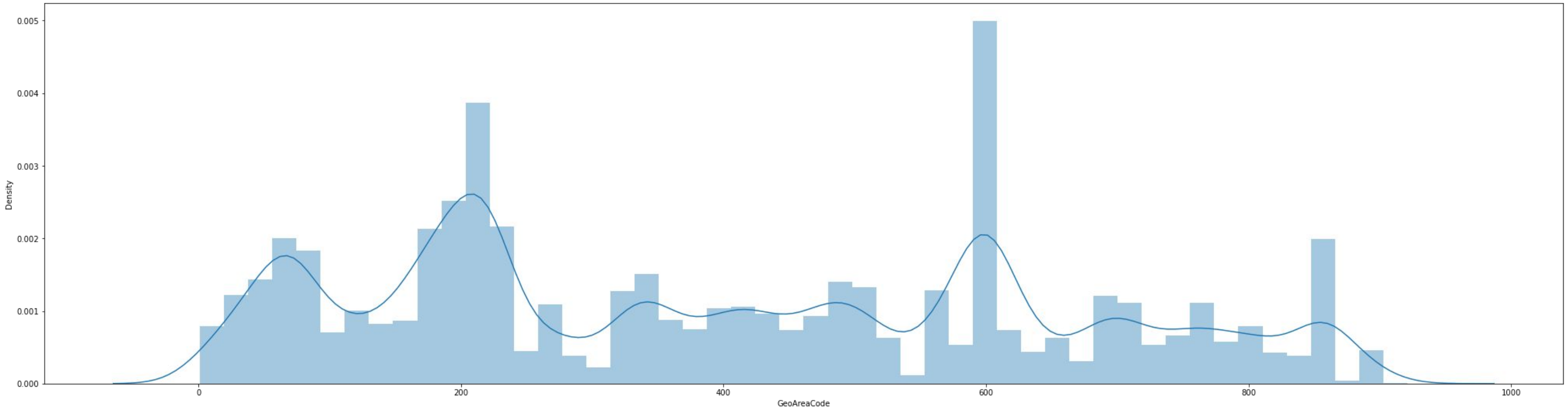
- Constructed 20 features with combinations of sex, type of skill, and education level

Linear Interpolation

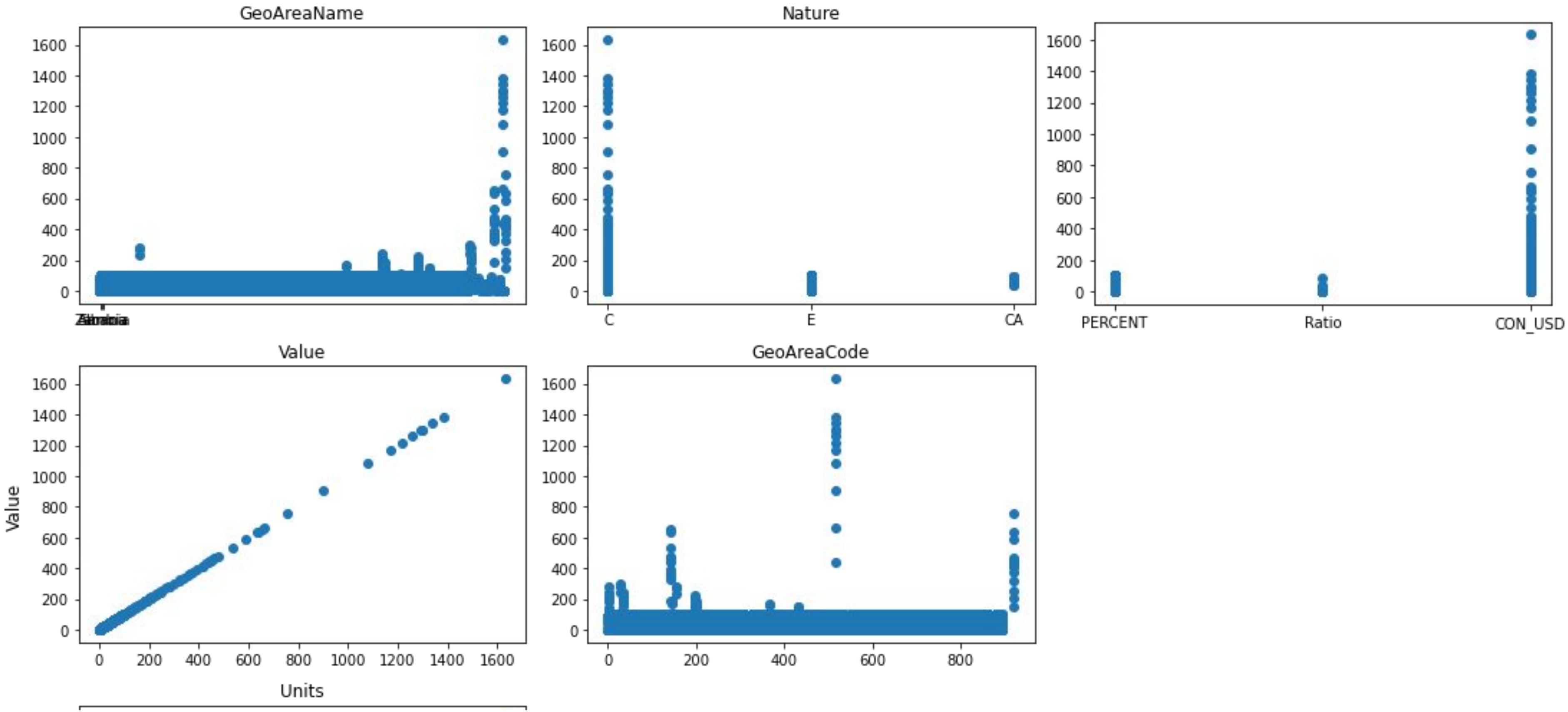
- Some countries only had 2 observations (e.g. 2006, 2013)
- Used linear interpolation to get the average change to impute the values for each year 2000-2019

Step 1: Data Wrangling: Data Analysis





Step 1: Data Wrangling: Data Analysis



Step 2: Statistical Modeling

Input features

- Reduce data dimensions from 10,000+
- Sex, Type of skill Education level

Model selection

- Vector autoregression (VAR)
- Reason: An interpretable model, multiple target features
- Time series analysis and time series cross validation and metrics

VAR Model for Columbia

```
  Summary of Regression Results
==================================
Model:                         VAR
Method:                        OLS
Date:           Thu, 03, Jun, 2021
Time:                     22:07:43
--------------------------------------------------------------------
No. of Equations:         18.0000    BIC:                    12.4234
Nobs:                     2779.00    HQIC:                   11.9571
Log likelihood:          -86884.4    FPE:                    119803.
AIC:                      11.6936    Det(Omega mle):         105975.
--------------------------------------------------------------------
```

VAR Model Prediction for Columbia 2019

```
['BOTHSEX.SKILL_READ.GRAD23', 'BOTHSEX.SKILL_READ.LOWSEC', 'FEMALE.SKILL_MATH.GRAD23',

 'FEMALE.SKILL_MATH.LOWSEC', 'MALE.SKILL_MATH.GRAD23', 'BOTHSEX.SKILL_MATH.PRIMAR',

 'MALE.SKILL_READ.GRAD23', 'MALE.SKILL_MATH.LOWSEC', 'FEMALE.SKILL_READ.LOWSEC',
'BOTHSEX.SKILL_READ.PRIMAR', 'FEMALE.SKILL_READ.GRAD23', 'BOTHSEX.SKILL_MATH.LOWSEC',
'MALE.SKILL_READ.PRIMAR', 'FEMALE.SKILL_READ.PRIMAR', 'BOTHSEX.SKILL_MATH.GRAD23',
'MALE.SKILL_MATH.PRIMAR', 'MALE.SKILL_READ.LOWSEC', 'FEMALE.SKILL_MATH.PRIMAR']
```

```
[[82.12420274 63.68178493 81.22458488 50.81520381 76.69042084 62.53068801

  80.36472478 53.95159255 62.64138252 56.80552639 82.30100092 52.37763895

  53.3741399  62.27456359 78.87735422 64.73495496 44.92262381 60.82812349]]
```

VAR Model for Costa Rica

```
Summary of Regression Results
==================================
Model:                         VAR
Method:                        OLS
Date:           Thu, 03, Jun, 2021
Time:                     22:09:12
--------------------------------------------------------------------
No. of Equations:         18.0000    BIC:                    12.4234
Nobs:                     2779.00    HQIC:                   11.9571
Log likelihood:          -86884.4    FPE:                    119803.
AIC:                      11.6936    Det(Omega_mle):         105975.
```

VAR Model Prediction for Costa Rica 2019

```
['BOTHSEX.SKILL_READ.GRAD23', 'BOTHSEX.SKILL_READ.LOWSEC', 'FEMALE.SKILL_MATH.GRAD23',

 'FEMALE.SKILL_MATH.LOWSEC', 'MALE.SKILL_MATH.GRAD23', 'BOTHSEX.SKILL_MATH.PRIMAR',

 'MALE.SKILL_READ.GRAD23', 'MALE.SKILL_MATH.LOWSEC', 'FEMALE.SKILL_READ.LOWSEC',
'BOTHSEX.SKILL_READ.PRIMAR', 'FEMALE.SKILL_READ.GRAD23', 'BOTHSEX.SKILL_MATH.LOWSEC',
'MALE.SKILL_READ.PRIMAR', 'FEMALE.SKILL_READ.PRIMAR', 'BOTHSEX.SKILL_MATH.GRAD23',
'MALE.SKILL_MATH.PRIMAR', 'MALE.SKILL_READ.LOWSEC', 'FEMALE.SKILL_MATH.PRIMAR']
```

```
[[82.12420274 63.68178493 81.22458488 50.81520381 76.69042084 62.53068801

  80.36472478 53.95159255 62.64138252 56.80552639 82.30100092 52.37763895

  53.3741399  62.27456359 78.87735422 64.73495496 44.92262381 60.82812349]]
```

VAR Model for Guatemala

```
Summary of Regression Results
==================================
Model:                         VAR
Method:                        OLS
Date:           Thu, 03, Jun, 2021
Time:                     22:09:12
--------------------------------------------------------------------
No. of Equations:         18.0000    BIC:                    12.4234
Nobs:                     2779.00    HQIC:                   11.9571
Log likelihood:          -86884.4    FPE:                    119803.
AIC:                      11.6936    Det(Omega_mle):         105975.
```

VAR Model Prediction for Guatemala

```
['BOTHSEX.SKILL_READ.GRAD23', 'BOTHSEX.SKILL_READ.LOWSEC', 'FEMALE.SKILL_MATH.GRAD23',

 'FEMALE.SKILL_MATH.LOWSEC', 'MALE.SKILL_MATH.GRAD23', 'BOTHSEX.SKILL_MATH.PRIMAR',

 'MALE.SKILL_READ.GRAD23', 'MALE.SKILL_MATH.LOWSEC', 'FEMALE.SKILL_READ.LOWSEC',
'BOTHSEX.SKILL_READ.PRIMAR', 'FEMALE.SKILL_READ.GRAD23', 'BOTHSEX.SKILL_MATH.LOWSEC',
'MALE.SKILL_READ.PRIMAR', 'FEMALE.SKILL_READ.PRIMAR', 'BOTHSEX.SKILL_MATH.GRAD23',
'MALE.SKILL_MATH.PRIMAR', 'MALE.SKILL_READ.LOWSEC', 'FEMALE.SKILL_MATH.PRIMAR']
```

```
[[82.12420274 63.68178493 81.22458488 50.81520381 76.69042084 62.53068801

  80.36472478 53.95159255 62.64138252 56.80552639 82.30100092 52.37763895

  53.3741399  62.27456359 78.87735422 64.73495496 44.92262381 60.82812349]]
```

VAR Model for Mexico

```
  Summary of Regression Results
==================================
Model:                         VAR
Method:                        OLS
Date:           Thu, 03, Jun, 2021
Time:                     22:11:24
--------------------------------------------------------------------
No. of Equations:         18.0000    BIC:                    12.4234
Nobs:                     2779.00    HQIC:                   11.9571
Log likelihood:          -86884.4    FPE:                    119803.
AIC:                      11.6936    Det(Omega_mle):         105975.
```

VAR Model Prediction for Mexico

```
['BOTHSEX.SKILL_READ.GRAD23', 'BOTHSEX.SKILL_READ.LOWSEC', 'FEMALE.SKILL_MATH.GRAD23',

 'FEMALE.SKILL_MATH.LOWSEC', 'MALE.SKILL_MATH.GRAD23', 'BOTHSEX.SKILL_MATH.PRIMAR',

 'MALE.SKILL_READ.GRAD23', 'MALE.SKILL_MATH.LOWSEC', 'FEMALE.SKILL_READ.LOWSEC',
'BOTHSEX.SKILL_READ.PRIMAR', 'FEMALE.SKILL_READ.GRAD23', 'BOTHSEX.SKILL_MATH.LOWSEC',
'MALE.SKILL_READ.PRIMAR', 'FEMALE.SKILL_READ.PRIMAR', 'BOTHSEX.SKILL_MATH.GRAD23',
'MALE.SKILL_MATH.PRIMAR', 'MALE.SKILL_READ.LOWSEC', 'FEMALE.SKILL_MATH.PRIMAR']
```

```
[[82.12420274 63.68178493 81.22458488 50.81520381 76.69042084 62.53068801

  80.36472478 53.95159255 62.64138252 56.80552639 82.30100092 52.37763895

  53.3741399  62.27456359 78.87735422 64.73495496 44.92262381 60.82812349]]
```

Summary:

- We did the data cleaning based on the feature importance, and fill in the missing data based on specific information. Then we structure the data and select the training features.

- Vector regression model (VAR) was selected to fit the training data and from the time series analysis we find that all of the sex, type of skill, and education level features will increase in 2019 (see previous slides for forecasts)