Project 3

GROUP 8- Primary Education

INITIAL DATA

- 87 Rows - 12 columns.

	ISO3	Countries and areas	Region	Sub-region	Income Group	Total	Residence Rural	Residence Urban	Wealth quintile Poorest	Wealth quintile Richest	Data source	Time period
0	AGO	Angola	SSA	ESA	Lower middle income (LM)	0.15	0.02	0.22	0.00	0.61	Demographic and Health Survey	2015-16
1	ARG	Argentina	LAC	LAC	Upper middle income (UM)	0.39	NaN	NaN	NaN	NaN	Multiple Indicator Cluster Survey	2011-12
2	ARM	Armenia	ECA	EECA	Upper middle income (UM)	0.81	0.69	0.89	0.46	0.99	Demographic and Health Survey	2015-16
3	BGD	Bangladesh	SA	SA	Lower middle income (LM)	0.34	0.30	0.49	0.07	0.75	Multiple Indicator Cluster Survey	2019
4	BRB	Barbados	LAC	LAC	High income (H)	0.63	0.54	0.68	0.09	0.97	Multiple Indicator Cluster Survey	2012
...	...	...	...	...	...	...	...	...	...	...	...	...
82	URY	Uruguay	LAC	LAC	High income (H)	0.63	0.53	0.64	0.37	NaN	Multiple Indicator Cluster Survey	2012-99
83	UZB	Uzbekistan	ECA	EECA	Lower middle income (LM)	0.19	0.16	0.26	0.00	0.69	UNICEF Nutrition Survey 2017	2017
84	VNM	Viet Nam	EAP	EAP	NaN	0.58	NaN	0.58	NaN	NaN	STEP Skills Measurement Household Survey 2012 ...	2012
85	ZMB	Zambia	SSA	ESA	Lower middle income (LM)	0.06	0.02	0.12	0.00	0.28	Demographic and Health Survey	2018-19
86	ZWE	Zimbabwe	SSA	ESA	Lower middle income (LM)	0.25	0.16	0.48	0.04	0.60	Multiple Indicator Cluster Survey	2018-19

87 rows × 12 columns

PROCESS

- Import file
- Analyzing data (Data type, value_counts, generating graphs)
- Drop NA per Row
- Identify duplicates
- Identify outliers (Gaussian distribution method)
- Dropping outliers (entire row)
- Imputing zero values with median (for each column)
- Encoding for (Region,Sub-region_Encoder, Income Group,Data source)
- Dataframe to MySQL

HIGHLIGHTS

- Imputing zero values with median. Why?

```sql
SELECT Region_Encoder,SUM(Total) AS Total FROM Exercise
GROUP BY Region_Encoder;

SELECT `Data source_Enncoder`,count(`Data source_Enncoder`) AS Total FROM Exercise
GROUP BY `Data source_Enncoder`;

SELECT `Time period`,count(`Time period`) AS Total FROM Exercise
GROUP BY `Time period`;
```

- SQL

FINAL DATA

- 60 Rows , 12 Columns

	ISO3	Countries and areas	Total	Residence Rural	Residence Urban	Wealth quintile Poorest	Wealth quintile Richest	Time period	Region_Encoder	Sub-region_Encoder	Income Group_Encoder	Data source_Enncoder
0	AGO	Angola	0.15	0.02	0.22	0.02	0.61	2015-16	5	2	2	3
2	ARM	Armenia	0.81	0.69	0.89	0.46	0.99	2015-16	1	1	3	3
3	BGD	Bangladesh	0.34	0.30	0.49	0.07	0.75	2019	4	5	2	5
4	BRB	Barbados	0.63	0.54	0.68	0.09	0.97	2012	2	3	0	5
5	BEN	Benin	0.03	0.01	0.06	0.02	0.14	2017-18	5	6	1	3
7	BIH	Bosnia and Herzegovina	0.51	0.43	0.68	0.03	0.95	2011-12	1	1	3	5
8	BRA	Brazil	0.82	0.49	0.88	0.86	1.00	2018	2	3	3	0
9	BGR	Bulgaria	0.73	0.65	0.77	0.24	0.98	2013	1	1	3	1

COMPARISON

- Initial Data
- Mean: 0.911
- Standard Deviation: 5.12

Final Data

Mean: 0.34

Standard Deviation: 0.29