Maji Ndogo Project: Insights Into The Data.

Problem

- Uncleaned raw dataset (missing values, duplications, unformatted cells, lack of data formatting etc.).
- The raw data has insights related to the level of access to water in various countries that need to be uncovered so that feature engineering can be done.
- We need to know the total population of rural and urban individuals and the population count split per region as well as determine the average rate of change of the level of access to water across regions.
- We need quantitative indicators on the change of infrastructural development in each of the regions so that we know which regions to prioritise to increase the level of access to basic water services.
- Prepare the data for the next step which is querying the data using SQL and feature engineering. [deep dive into the data]

Estimates on the use of water (2020)	`name` The country or area name.	`income_group` The country's classification according to income group.
`pop_n` The national population size estimate in thousands.	`pop_u` The urban population share estimate in percentage points (%).	`wat_bas_n` The estimated **national** share of people with at least **basic** service (%)*.
`wat_lim_n` The estimated **national** share of people with **limited** service (%).	`wat_unimp_n` The estimated **national** share of people with **unimproved** service (%).	`wat_sur_n` The estimated **national** share of people with **surface** service (%).

Becoming Familiar with the Dataset

`wat_bas_r`

The estimated **rural** share of people with at least **basic** service (%).

`wat_lim_r`

The estimated **rural** share of people with **limited** service (%).

`wat_unimp_r`

The estimated **rural** share of people with **unimproved** service (%).

`wat_sur_r`

The estimated **rural** share of people with **surface** service (%).

`wat_bas_u`

The estimated **urban** share of people with at least **basic** service (%).

`wat_lim_u`

The estimated **urban** share of people with **limited** service (%).

`wat_unimp_u`

The estimated **urban** share of people with **unimproved** service (%).

`wat_sur_u`

The estimated **urban** share of people with **surface** service (%).

We have a total of 16 features (or columns) in our dataset, 12 of which are service-level percentage shares.

Becoming Familiar with the Dataset

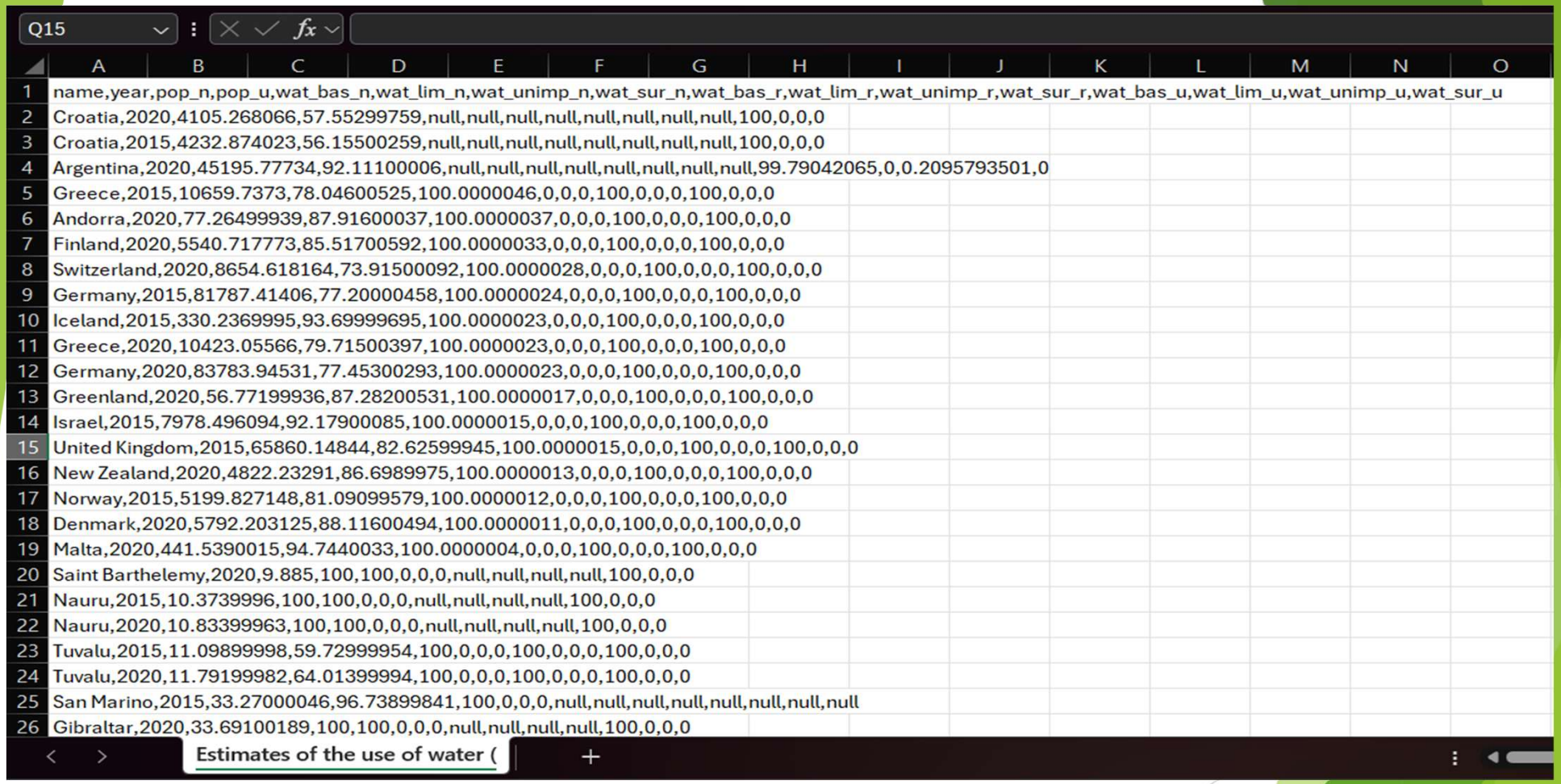
Becoming Familiar with the Dataset

The United Nations (UN) uses Annual Rates of Change (ARC) to see whether the proportion of access to drinking water is declining or increasing. The **Annual Rates of Change (ARC)** is a statistical measure used to express the average yearly change rate of a variable over a certain period of time.

It's calculated by taking the difference between the end and start values of the dataset and dividing the result by the number of years that separate the two values:

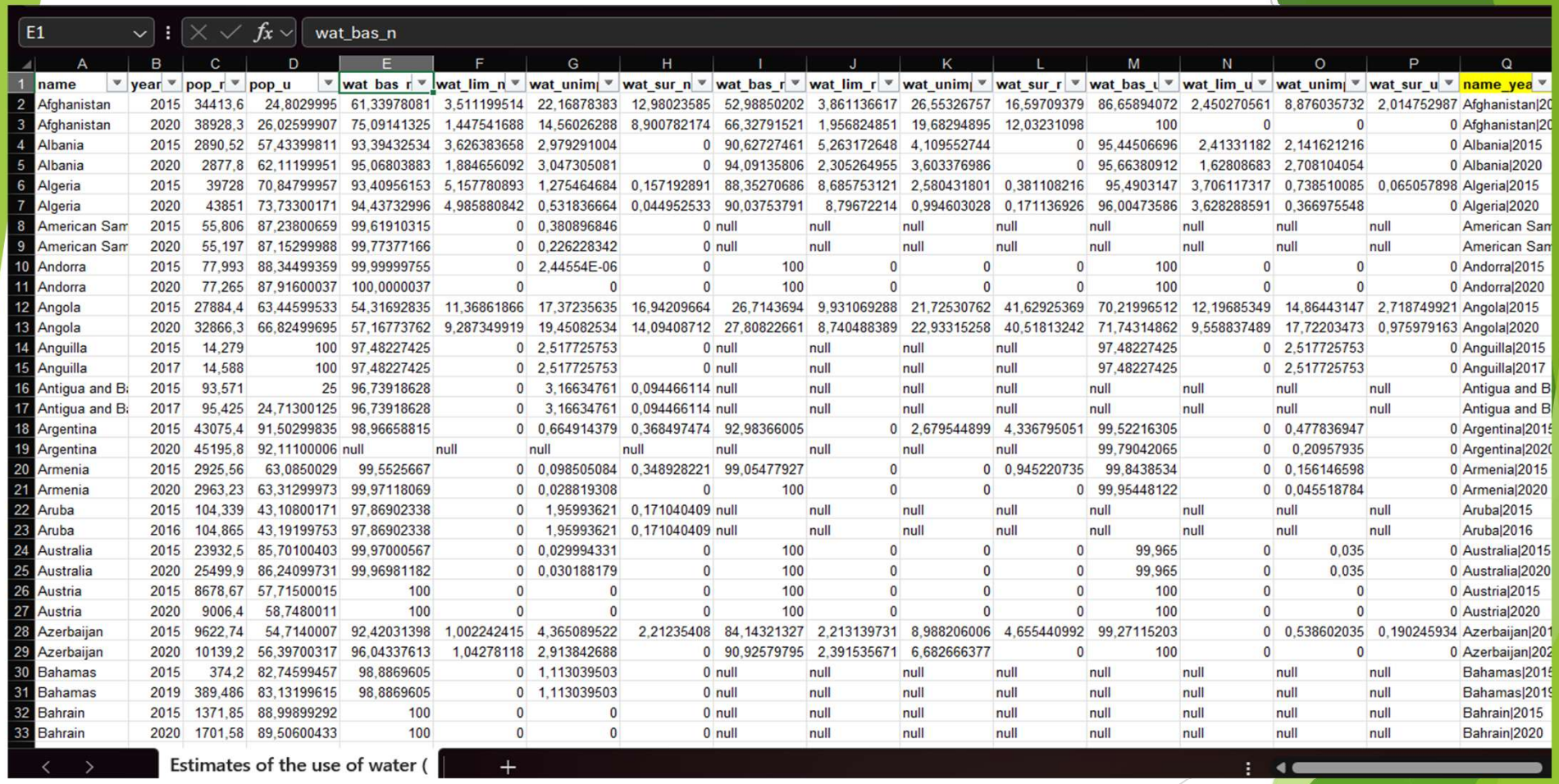
$$ARC_x = \frac{P_{x,y2} - P_{x,y1}}{Y_2 - Y_1}$$

Process:
The ARC of the data set was calculated by taking the national, urban and rural percentage share of the level of access to basic water [wat_bas_()] for each country over 2 separate year periods. The sum of all the countries' individual arcs was then added to get the total sum categorised per region each of the countries fall under.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	name,year,pop_n,pop_u,wat_bas_n,wat_lim_n,wat_unimp_n,wat_sur_n,wat_bas_r,wat_lim_r,wat_unimp_r,wat_sur_r,wat_bas_u,wat_lim_u,wat_unimp_u,wat_sur_u														
2	Croatia,2020,4105.268066,57.55299759,null,null,null,null,null,null,null,null,100,0,0,0														
3	Croatia,2015,4232.874023,56.15500259,null,null,null,null,null,null,null,null,100,0,0,0														
4	Argentina,2020,45195.77734,92.11100006,null,null,null,null,null,null,null,null,99.79042065,0,0.2095793501,0														
5	Greece,2015,10659.7373,78.04600525,100.0000046,0,0,0,100,0,0,0,100,0,0,0														
6	Andorra,2020,77.26499939,87.91600037,100.0000037,0,0,0,100,0,0,0,100,0,0,0														
7	Finland,2020,5540.717773,85.51700592,100.0000033,0,0,0,100,0,0,0,100,0,0,0														
8	Switzerland,2020,8654.618164,73.91500092,100.0000028,0,0,0,100,0,0,0,100,0,0,0														
9	Germany,2015,81787.41406,77.20000458,100.0000024,0,0,0,100,0,0,0,100,0,0,0														
10	Iceland,2015,330.2369995,93.69999695,100.0000023,0,0,0,100,0,0,0,100,0,0,0														
11	Greece,2020,10423.05566,79.71500397,100.0000023,0,0,0,100,0,0,0,100,0,0,0														
12	Germany,2020,83783.94531,77.45300293,100.0000023,0,0,0,100,0,0,0,100,0,0,0														
13	Greenland,2020,56.77199936,87.28200531,100.0000017,0,0,0,100,0,0,0,100,0,0,0														
14	Israel,2015,7978.496094,92.17900085,100.0000015,0,0,0,100,0,0,0,100,0,0,0														
15	United Kingdom,2015,65860.14844,82.62599945,100.0000015,0,0,0,100,0,0,0,100,0,0,0														
16	New Zealand,2020,4822.23291,86.6989975,100.0000013,0,0,0,100,0,0,0,100,0,0,0														
17	Norway,2015,5199.827148,81.09099579,100.0000012,0,0,0,100,0,0,0,100,0,0,0														
18	Denmark,2020,5792.203125,88.11600494,100.0000011,0,0,0,100,0,0,0,100,0,0,0														
19	Malta,2020,441.5390015,94.7440033,100.0000004,0,0,0,100,0,0,0,100,0,0,0														
20	Saint Barthelemy,2020,9.885,100,100,0,0,0,null,null,null,null,100,0,0,0														
21	Nauru,2015,10.3739996,100,100,0,0,0,null,null,null,null,100,0,0,0														
22	Nauru,2020,10.83399963,100,100,0,0,0,null,null,null,null,100,0,0,0														
23	Tuvalu,2015,11.09899998,59.72999954,100,0,0,0,100,0,0,0,100,0,0,0														
24	Tuvalu,2020,11.79199982,64.01399994,100,0,0,0,100,0,0,0,100,0,0,0														
25	San Marino,2015,33.27000046,96.73899841,100,0,0,0,null,null,null,null,null,null,null,null														
26	Gibraltar,2020,33.69100189,100,100,0,0,0,null,null,null,null,100,0,0,0														

Raw Dataset

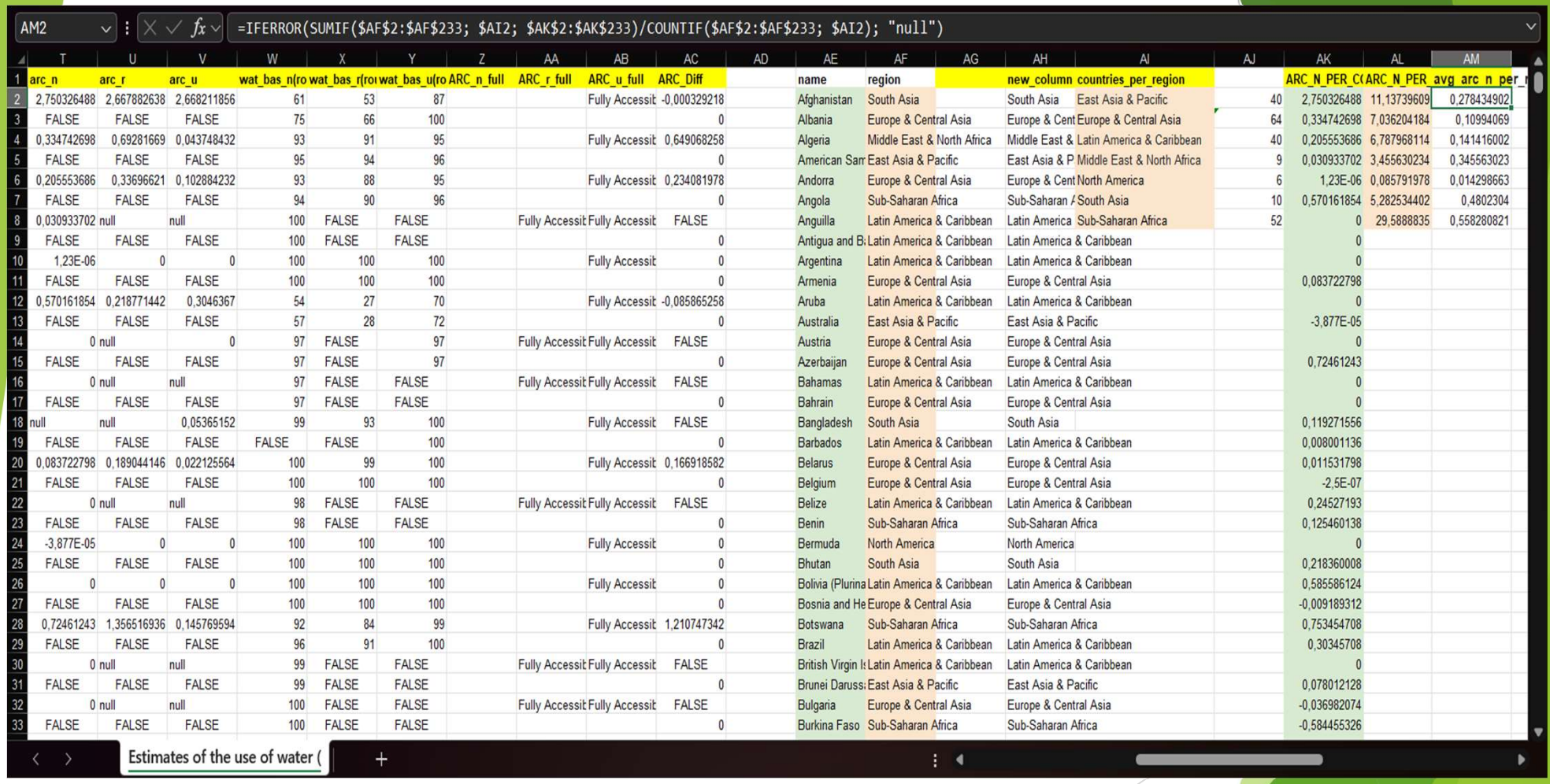


E1 | wat_bas_n

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	name	year	pop_r	pop_u	wat_bas_r	wat_lim_n	wat_unim	wat_sur_n	wat_bas_r	wat_lim_r	wat_unim	wat_sur_r	wat_bas_u	wat_lim_u	wat_unim	wat_sur_u	name_yea
2	Afghanistan	2015	34413,6	24,8029995	61,33978081	3,511199514	22,16878383	12,98023585	52,98850202	3,861136617	26,55326757	16,59709379	86,65894072	2,450270561	8,876035732	2,014752987	Afghanistan|20
3	Afghanistan	2020	38928,3	26,02599907	75,09141325	1,447541688	14,56026288	8,900782174	66,32791521	1,956824851	19,68294895	12,03231098	100	0	0	0	Afghanistan|20
4	Albania	2015	2890,52	57,43399811	93,39432534	3,626383658	2,979291004	0	90,62727461	5,263172648	4,109552744	0	95,44506696	2,41331182	2,141621216	0	Albania|2015
5	Albania	2020	2877,8	62,11199951	95,06803883	1,884656092	3,047305081	0	94,09135806	2,305264955	3,603376986	0	95,66380912	1,62808683	2,708104054	0	Albania|2020
6	Algeria	2015	39728	70,84799957	93,40956153	5,157780893	1,275464684	0,157192891	88,35270686	8,685753121	2,580431801	0,381108216	95,4903147	3,706117317	0,738510085	0,065057898	Algeria|2015
7	Algeria	2020	43851	73,73300171	94,43732996	4,985880842	0,531836664	0,044952533	90,03753791	8,79672214	0,994603028	0,171136926	96,00473586	3,628288591	0,366975548	0	Algeria|2020
8	American San	2015	55,806	87,23800659	99,61910315	0	0,380896846	0	null	null	null	null	null	null	null	null	American San
9	American San	2020	55,197	87,15299988	99,77377166	0	0,226228342	0	null	null	null	null	null	null	null	null	American San
10	Andorra	2015	77,993	88,34499359	99,99999755	0	2,44554E-06	0	100	0	0	0	100	0	0	0	Andorra|2015
11	Andorra	2020	77,265	87,91600037	100,0000037	0	0	0	100	0	0	0	100	0	0	0	Andorra|2020
12	Angola	2015	27884,4	63,44599533	54,31692835	11,36861866	17,37235635	16,94209664	26,7143694	9,931069288	21,72530762	41,62925369	70,21996512	12,19685349	14,86443147	2,718749921	Angola|2015
13	Angola	2020	32866,3	66,82499695	57,16773762	9,287349919	19,45082534	14,09408712	27,80822661	8,740488389	22,93315258	40,51813242	71,74314862	9,558837489	17,72203473	0,975979163	Angola|2020
14	Anguilla	2015	14,279	100	97,48227425	0	2,517725753	0	null	null	null	null	97,48227425	0	2,517725753	0	Anguilla|2015
15	Anguilla	2017	14,588	100	97,48227425	0	2,517725753	0	null	null	null	null	97,48227425	0	2,517725753	0	Anguilla|2017
16	Antigua and B	2015	93,571	25	96,73918628	0	3,16634761	0,094466114	null	null	null	null	null	null	null	null	Antigua and B
17	Antigua and B	2017	95,425	24,71300125	96,73918628	0	3,16634761	0,094466114	null	null	null	null	null	null	null	null	Antigua and B
18	Argentina	2015	43075,4	91,50299835	98,96658815	0	0,664914379	0,368497474	92,98366005	0	2,679544899	4,336795051	99,52216305	0	0,477836947	0	Argentina|2015
19	Argentina	2020	45195,8	92,11100006	null	null	null	null	null	null	null	null	99,79042065	0	0,20957935	0	Argentina|2020
20	Armenia	2015	2925,56	63,0850029	99,5525667	0	0,098505084	0,348928221	99,05477927	0	0	0,945220735	99,8438534	0	0,156146598	0	Armenia|2015
21	Armenia	2020	2963,23	63,31299973	99,97118069	0	0,028819308	0	100	0	0	0	99,95448122	0	0,045518784	0	Armenia|2020
22	Aruba	2015	104,339	43,10800171	97,86902338	0	1,95993621	0,171040409	null	null	null	null	null	null	null	null	Aruba|2015
23	Aruba	2016	104,865	43,19199753	97,86902338	0	1,95993621	0,171040409	null	null	null	null	null	null	null	null	Aruba|2016
24	Australia	2015	23932,5	85,70100403	99,97000567	0	0,029994331	0	100	0	0	0	99,965	0	0,035	0	Australia|2015
25	Australia	2020	25499,9	86,24099731	99,96981182	0	0,030188179	0	100	0	0	0	99,965	0	0,035	0	Australia|2020
26	Austria	2015	8678,67	57,71500015	100	0	0	0	100	0	0	0	100	0	0	0	Austria|2015
27	Austria	2020	9006,4	58,7480011	100	0	0	0	100	0	0	0	100	0	0	0	Austria|2020
28	Azerbaijan	2015	9622,74	54,7140007	92,42031398	1,002242415	4,365089522	2,21235408	84,14321327	2,213139731	8,988206006	4,655440992	99,27115203	0	0,538602035	0,190245934	Azerbaijan|201
29	Azerbaijan	2020	10139,2	56,39700317	96,04337613	1,04278118	2,913842688	0	90,92579795	2,391535671	6,682666377	0	100	0	0	0	Azerbaijan|202
30	Bahamas	2015	374,2	82,74599457	98,8869605	0	1,113039503	0	null	null	null	null	null	null	null	null	Bahamas|2015
31	Bahamas	2019	389,486	83,13199615	98,8869605	0	1,113039503	0	null	null	null	null	null	null	null	null	Bahamas|2019
32	Bahrain	2015	1371,85	88,99899292	100	0	0	0	null	null	null	null	null	null	null	null	Bahrain|2015
33	Bahrain	2020	1701,58	89,50600433	100	0	0	0	null	null	null	null	null	null	null	null	Bahrain|2020

Estimates of the use of water (

Cleaned and Organised Data



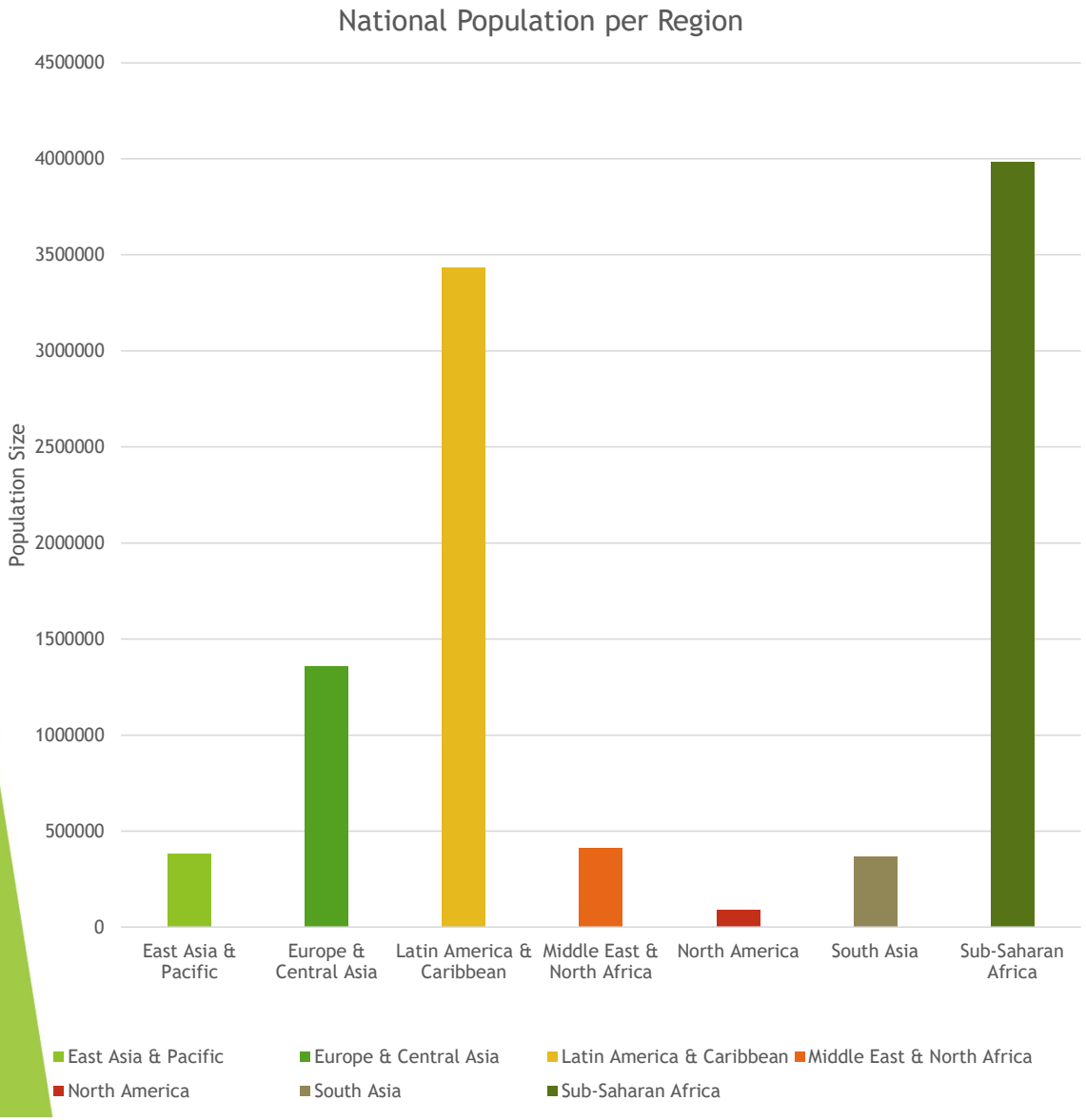
AM2

=IFERROR(SUMIF($AF$2:$AF$233; $AI2; $AK$2:$AK$233)/COUNTIF($AF$2:$AF$233; $AI2); "null")

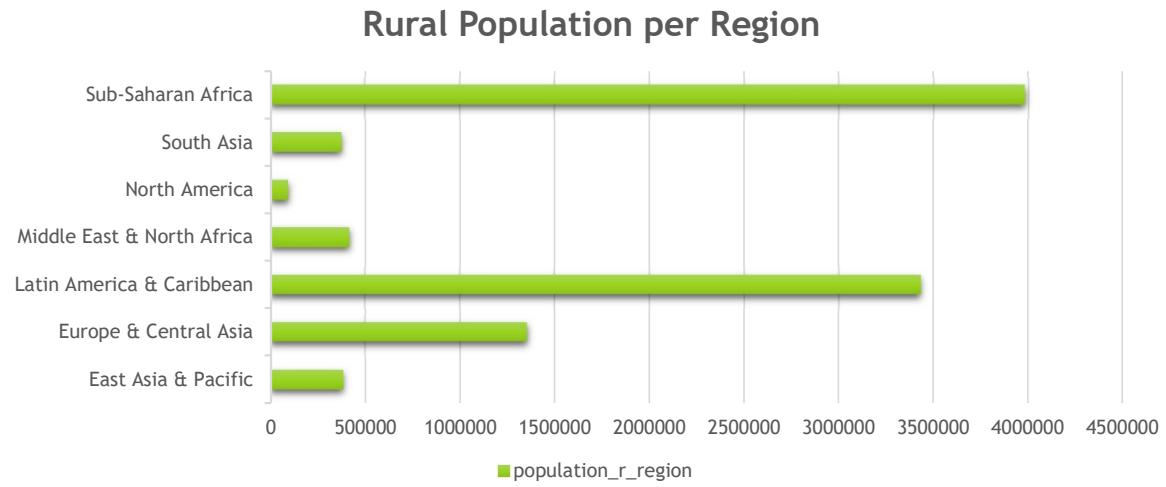
	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM
1	arc_n	arc_r	arc_u	wat_bas_n(ro	wat_bas_r(ro	wat_bas_u(ro	ARC_n_full	ARC_r_full	ARC_u_full	ARC_Diff		name	region		new_column	countries_per_region		ARC_N_PER_C(	ARC_N_PER_	avg_arc_n_per_
2	2,750326488	2,667882638	2,668211856	61	53	87			Fully Accessit	-0,000329218		Afghanistan	South Asia		South Asia	East Asia & Pacific	40	2,750326488	11,13739609	0,278434902
3	FALSE	FALSE	FALSE	75	66	100				0		Albania	Europe & Central Asia		Europe & Cent	Europe & Central Asia	64	0,334742698	7,036204184	0,10994069
4	0,334742698	0,69281669	0,043748432	93	91	95			Fully Accessit	0,649068258		Algeria	Middle East & North Africa		Middle East &	Latin America & Caribbean	40	0,205553686	6,787968114	0,141416002
5	FALSE	FALSE	FALSE	95	94	96				0		American Sam	East Asia & Pacific		East Asia & P	Middle East & North Africa	9	0,030933702	3,455630234	0,345563023
6	0,205553686	0,33696621	0,102884232	93	88	95			Fully Accessit	0,234081978		Andorra	Europe & Central Asia		Europe & Cent	North America	6	1,23E-06	0,085791978	0,014298663
7	FALSE	FALSE	FALSE	94	90	96				0		Angola	Sub-Saharan Africa		Sub-Saharan A	South Asia	10	0,570161854	5,282534402	0,4802304
8	0,030933702	null	null	100	FALSE	FALSE		Fully Accessit	Fully Accessit	FALSE		Anguilla	Latin America & Caribbean		Latin America	Sub-Saharan Africa	52	0	29,5888835	0,558280821
9	FALSE	FALSE	FALSE	100	FALSE	FALSE				0		Antigua and B	Latin America & Caribbean		Latin America & Caribbean			0		
10	1,23E-06	0	0	100	100	100			Fully Accessit	0		Argentina	Latin America & Caribbean		Latin America & Caribbean			0		
11	FALSE	FALSE	FALSE	100	100	100				0		Armenia	Europe & Central Asia		Europe & Central Asia			0,083722798		
12	0,570161854	0,218771442	0,3046367	54	27	70			Fully Accessit	-0,085865258		Aruba	Latin America & Caribbean		Latin America & Caribbean			0		
13	FALSE	FALSE	FALSE	57	28	72				0		Australia	East Asia & Pacific		East Asia & Pacific			-3,877E-05		
14	0	null	0	97	FALSE	97		Fully Accessit	Fully Accessit	FALSE		Austria	Europe & Central Asia		Europe & Central Asia			0		
15	FALSE	FALSE	FALSE	97	FALSE	97				0		Azerbaijan	Europe & Central Asia		Europe & Central Asia			0,72461243		
16	0	null	null	97	FALSE	FALSE		Fully Accessit	Fully Accessit	FALSE		Bahamas	Latin America & Caribbean		Latin America & Caribbean			0		
17	FALSE	FALSE	FALSE	97	FALSE	FALSE				0		Bahrain	Europe & Central Asia		Europe & Central Asia			0		
18	null	null	0,05365152	99	93	100			Fully Accessit	FALSE		Bangladesh	South Asia		South Asia			0,119271556		
19	FALSE	FALSE	FALSE	FALSE	FALSE	100				0		Barbados	Latin America & Caribbean		Latin America & Caribbean			0,008001136		
20	0,083722798	0,189044146	0,022125564	100	99	100			Fully Accessit	0,166918582		Belarus	Europe & Central Asia		Europe & Central Asia			0,011531798		
21	FALSE	FALSE	FALSE	100	100	100				0		Belgium	Europe & Central Asia		Europe & Central Asia			-2,5E-07		
22	0	null	null	98	FALSE	FALSE		Fully Accessit	Fully Accessit	FALSE		Belize	Latin America & Caribbean		Latin America & Caribbean			0,24527193		
23	FALSE	FALSE	FALSE	98	FALSE	FALSE				0		Benin	Sub-Saharan Africa		Sub-Saharan Africa			0,125460138		
24	-3,877E-05	0	0	100	100	100			Fully Accessit	0		Bermuda	North America		North America			0		
25	FALSE	FALSE	FALSE	100	100	100				0		Bhutan	South Asia		South Asia			0,218360008		
26	0	0	0	100	100	100			Fully Accessit	0		Bolivia (Plurina	Latin America & Caribbean		Latin America & Caribbean			0,585586124		
27	FALSE	FALSE	FALSE	100	100	100				0		Bosnia and He	Europe & Central Asia		Europe & Central Asia			-0,009189312		
28	0,72461243	1,356516936	0,145769594	92	84	99			Fully Accessit	1,210747342		Botswana	Sub-Saharan Africa		Sub-Saharan Africa			0,753454708		
29	FALSE	FALSE	FALSE	96	91	100				0		Brazil	Latin America & Caribbean		Latin America & Caribbean			0,30345708		
30	0	null	null	99	FALSE	FALSE		Fully Accessit	Fully Accessit	FALSE		British Virgin I	Latin America & Caribbean		Latin America & Caribbean			0		
31	FALSE	FALSE	FALSE	99	FALSE	FALSE				0		Brunei Daruss	East Asia & Pacific		East Asia & Pacific			0,078012128		
32	0	null	null	100	FALSE	FALSE		Fully Accessit	Fully Accessit	FALSE		Bulgaria	Europe & Central Asia		Europe & Central Asia			-0,036982074		
33	FALSE	FALSE	FALSE	100	FALSE	FALSE				0		Burkina Faso	Sub-Saharan Africa		Sub-Saharan Africa			-0,584455326		

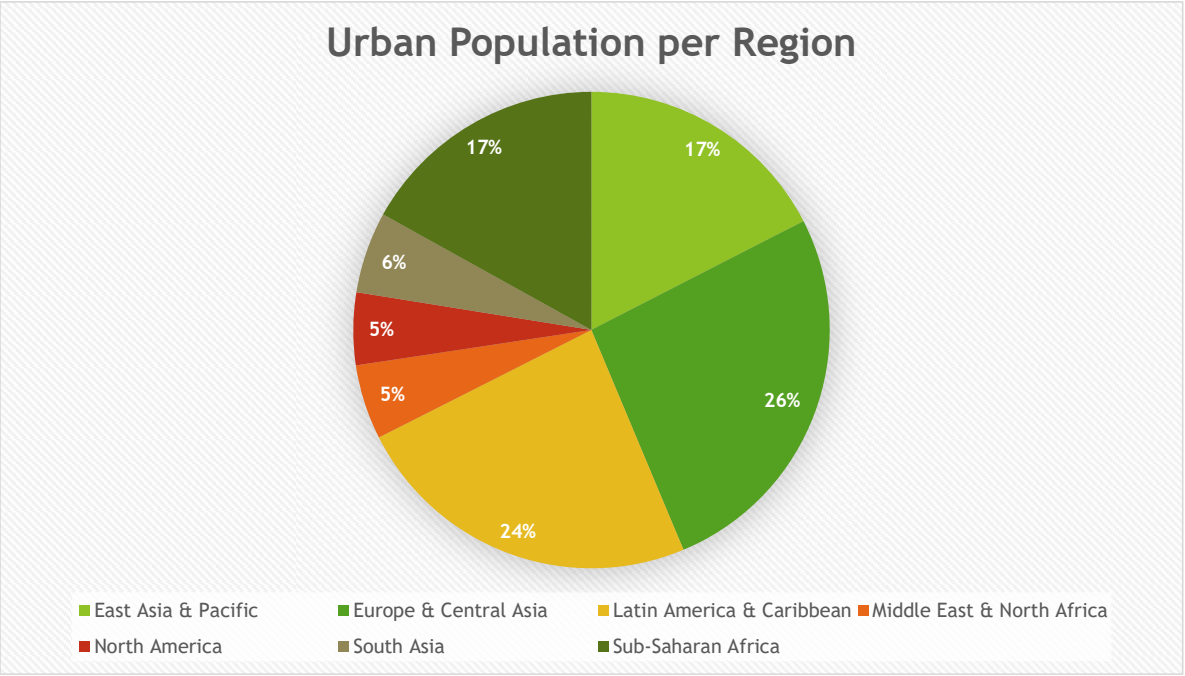
Estimates of the use of water (

Feature Engineering



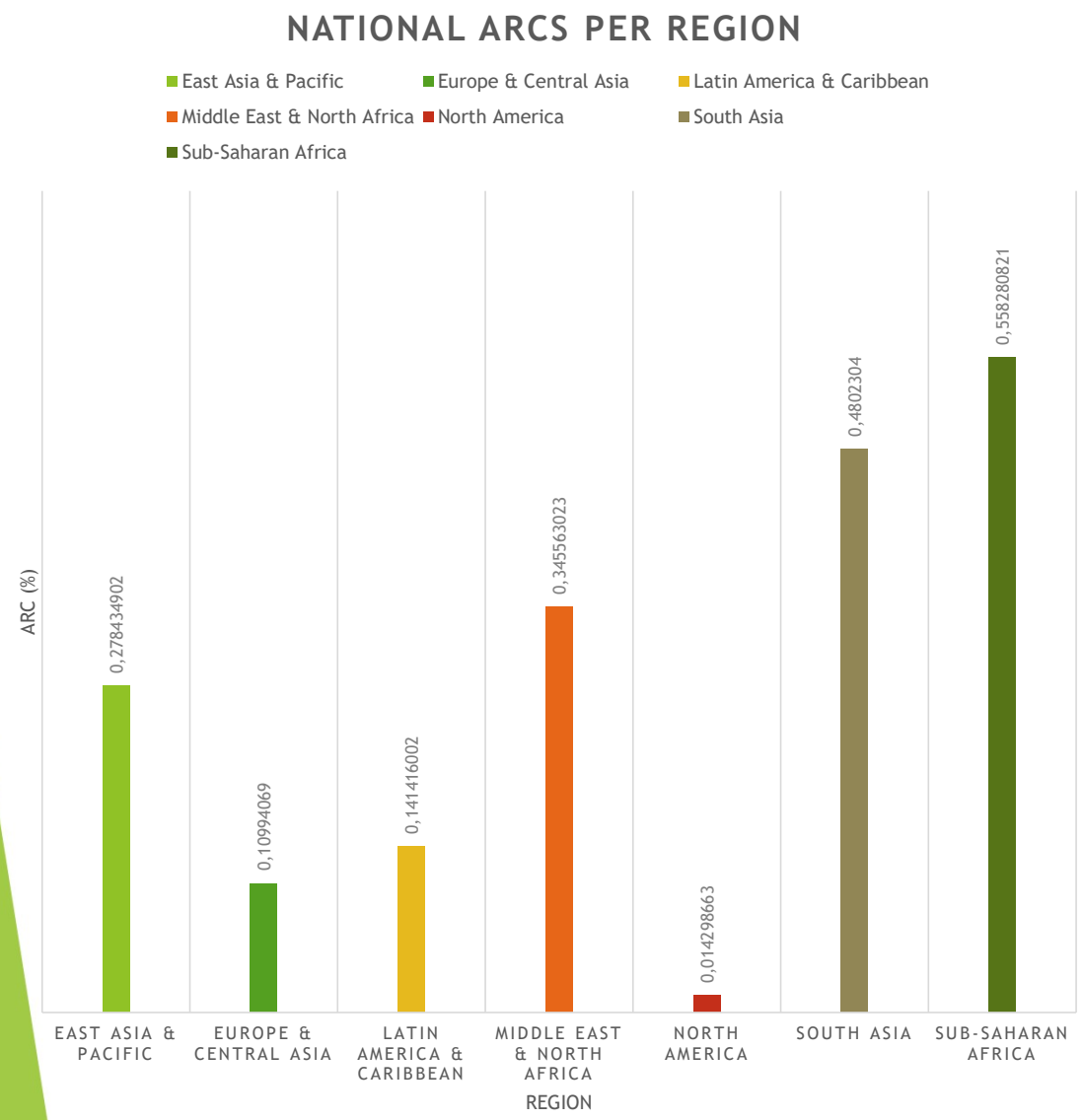
- Sub-Saharan Africa, Latin America and Caribbean and Europe and Central Asia rank among the highest in population count in the Maji Ndogo data set.
- This graph gives us valuable insight into the proportion of individuals affected by basic water access levels in Maji Ndogo and also paint a picture of how much the lack of access to water affects predominantly third world countries. This is a problem that is exacerbated by structural inequalities and inefficiencies experienced by the significantly affected countries.



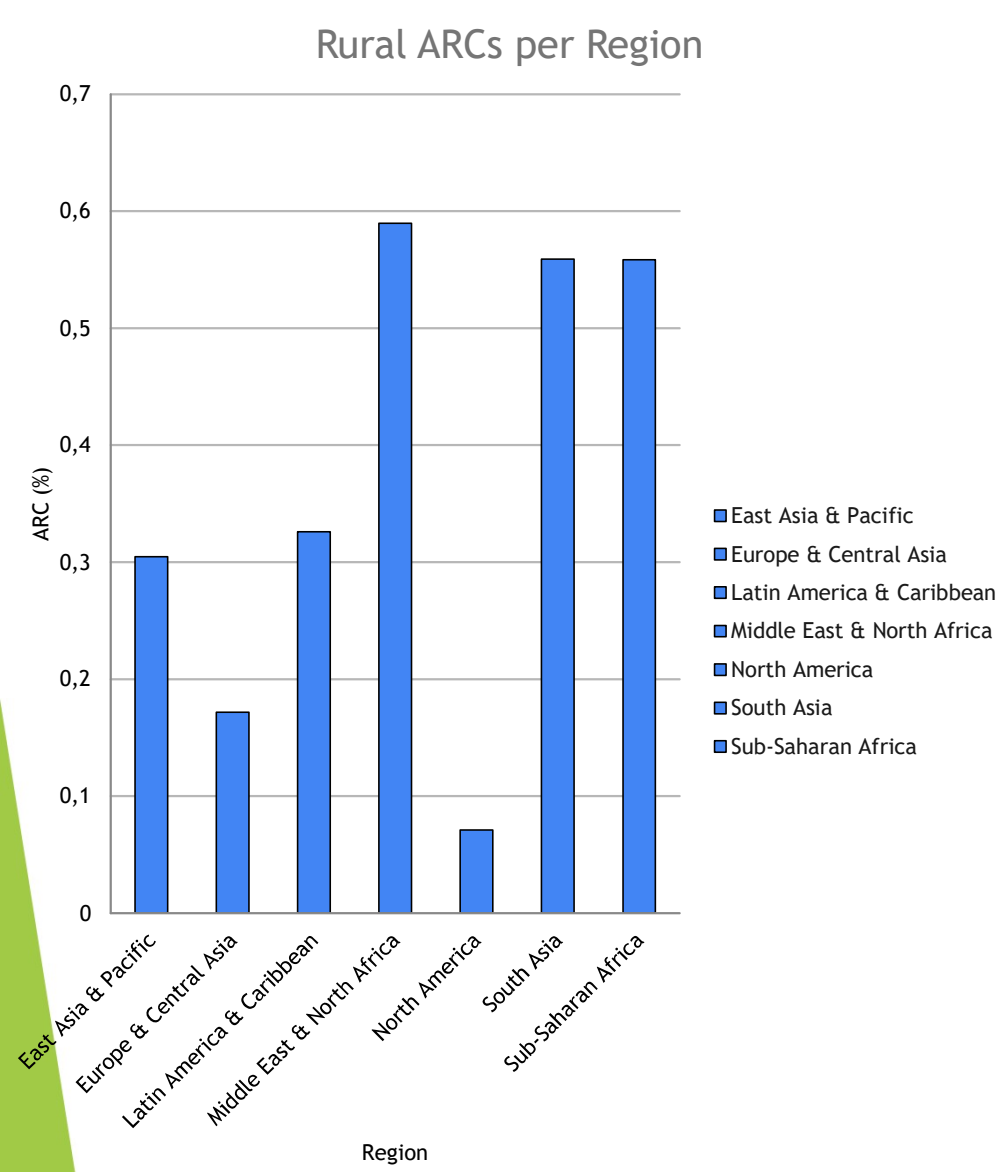


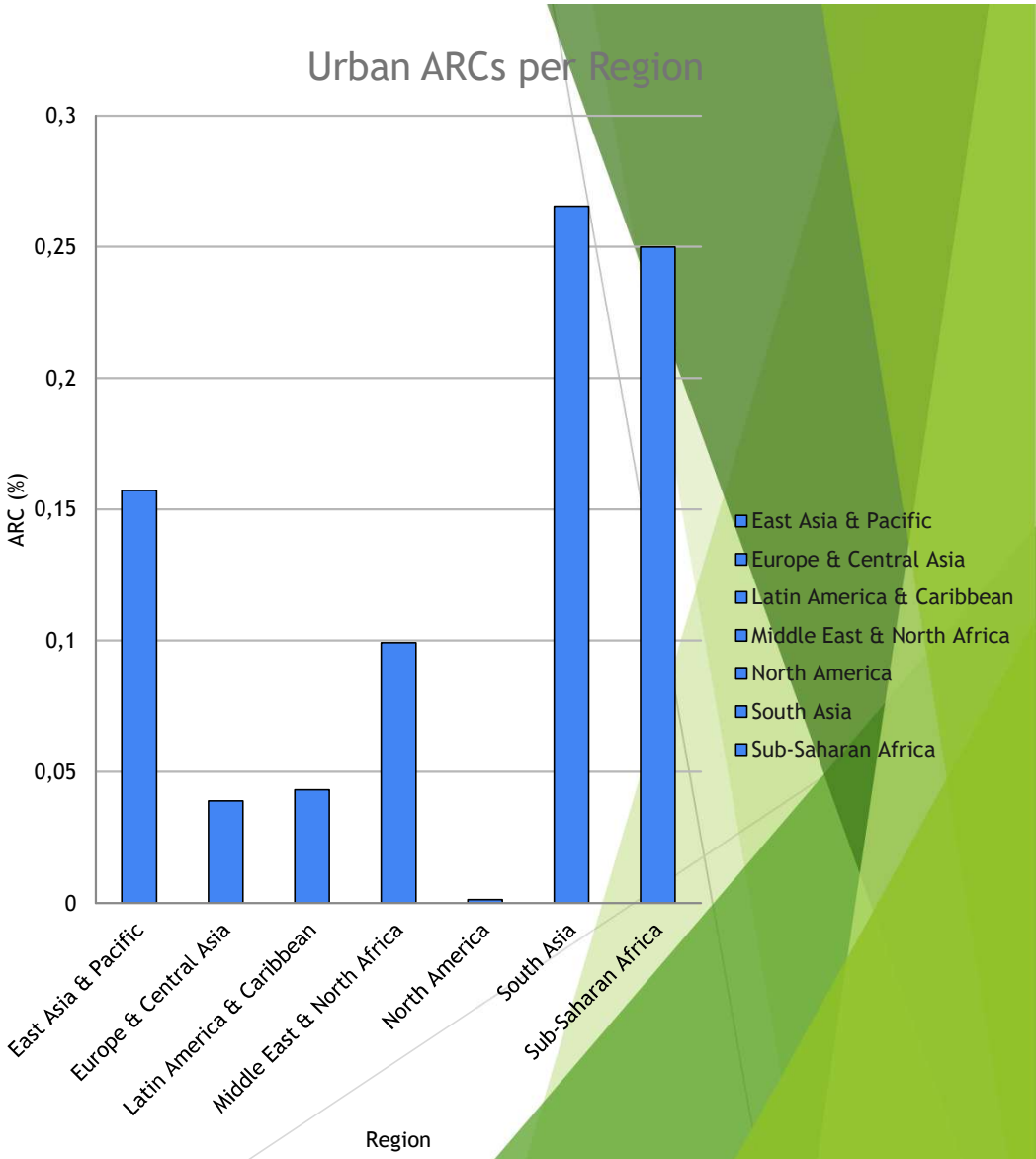
- **Total Rural Population: 10 015 519**
- **Total Urban Population: 14 842**

- Sub-Saharan Africa has the highest proportion of rural population where rural areas are often known for their infrastructural deficiencies especially regarding basic service provision e.g.: water access, healthcare, municipal plumbing and electricity etc. This highly negatively impacts their level of basic water access.
- Efforts should primarily be directed in the rural areas namely; Sub-Saharan Africa and Latin America and Central Asia in order to equal the population's level of access to basic water service with the urban region.



- The ARC values listed here illustrate the rate of change of the share percentage of people's access to basic water services per country per region. A higher ARC value means there have been infrastructural water developments in the region to provide access to basic services to the people.
- Comparing the ARC value against the national population count also give a better picture on the extent and volume of developments undertaken to improve basic water access to different regions.
- Sub-Saharan Africa, Middle East and North Africa and South Asia are among the most underdeveloped regions globally. The are also the most prioritised in this project hence their high ARC values comparatively.


Rural ARCs per Region
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
ARC (%)
East Asia & Pacific
Europe & Central Asia
Latin America & Caribbean
Middle East & North Africa
North America
South Asia
Sub-Saharan Africa
Region


Urban ARCs per Region
0,3
0,25
0,2
0,15
0,1
0,05
0
ARC (%)
East Asia & Pacific
Europe & Central Asia
Latin America & Caribbean
Middle East & North Africa
North America
South Asia
Sub-Saharan Africa
Region

Conclusion/Recommendation

- The data was cleaned and is ready for SQL operations and querying to reveal deeper insights.
- **Rural Infrastructure:** Allocate 60-70% of new funding to rural Sub-Saharan Africa, where the population density vs. access ratio is most critical.
- **Targeted ARC Benchmarks:** Set a minimum target ARC of % for lagging regions to ensure they are attended to with high priority.