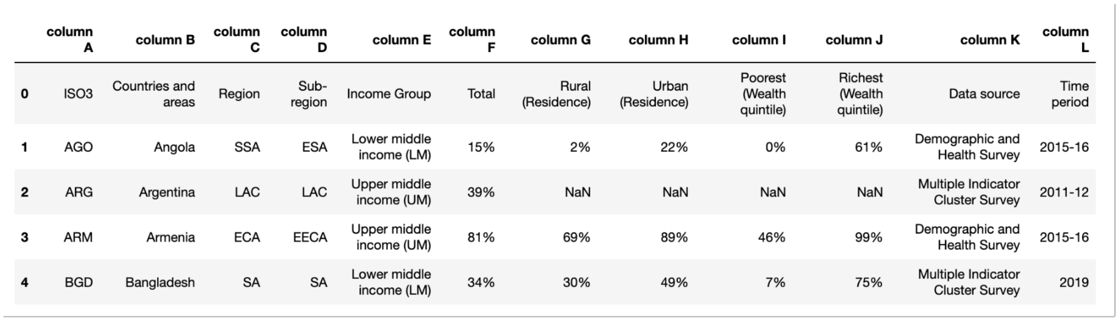
Name: Amay Viswanathan Iyer; Email: s3970066@student.rmit.edu.au; Student ID: s3970066

Data Preparation:

Primary.csv preparation:
In this section, I will detail the steps I took to carefully clean the datasets that we were given so that it could be better analyzed. Let's start with the Primary.csv dataset. In this dataset, I began with understanding what the sheet looked like. For this, I started off with pulling the data from the csv and placing it into a dataframe and displaying the first four rows:



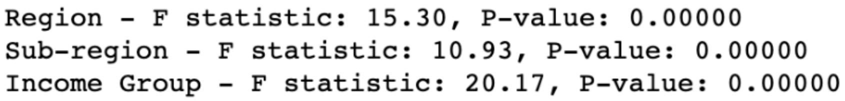
	column A	column B	column C	column D	column E	column F	column G	column H	column I	column J	column K	column L
0	ISO3	Countries and areas	Region	Sub-region	Income Group	Total	Rural (Residence)	Urban (Residence)	Poorest (Wealth quintile)	Richest (Wealth quintile)	Data source	Time period
1	AGO	Angola	SSA	ESA	Lower middle income (LM)	15%	2%	22%	0%	61%	Demographic and Health Survey	2015-16
2	ARG	Argentina	LAC	LAC	Upper middle income (UM)	39%	NaN	NaN	NaN	NaN	Multiple Indicator Cluster Survey	2011-12
3	ARM	Armenia	ECA	EECA	Upper middle income (UM)	81%	69%	89%	46%	99%	Demographic and Health Survey	2015-16
4	BGD	Bangladesh	SA	SA	Lower middle income (LM)	34%	30%	49%	7%	75%	Multiple Indicator Cluster Survey	2019

First quirks that I noticed:

- Column names - The consecutive row consisted of the actual column names. For analysis, I believe it would be helpful to omit this row entirely so that the dataset can be effectively visualized and analyzed.
- String-based values - The values under the Total, Rural, Urban, Poorest, and Richest columns consisted of a '%' sign following the value, effectively making them strings which cannot be evaluated numerically.
- Missing values - Under columns, G, H, I, and J, I can see empty cells with an 'NaN' indicating emptiness. These empty values will need to be substituted with their best applicable analog, which I will derive based on the mean for the best-fitting Categorical values (Region, Sub-region, or Income Group).
- 'Time period' - Another discrepancy I noticed upon viewing the data in one flash (and later opening the csv file and examining it cell-by-cell), is that on the final column, 'Time period', the years are displayed in an uneven fashion. Some values say, '2015-16', while others say '2562' or '2027'. Across the whole column, the data is aggregated only within the 2010's and the reason these values are out of bounds might be as a result of a typing error. Given that I know the 'Time period' cannot be outside of the years 2010-2019, we can manipulate the other years using regular expressions.

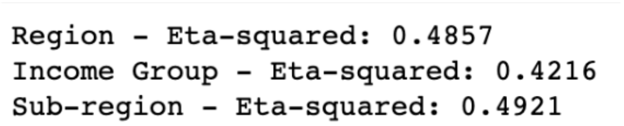
Checking for missing values per column:
When I checked for the missing values per column, I noticed that Columns G, H, I, and J ('Rural (Residence)', 'Urban (Residence)', 'Poorest (Wealth quintile)', and 'Richest (Wealth quintile)' respectively) contain the most number of missing values. This is important because those columns contain numerical values which are integral for the analyses I will be conducting in this Assignment. First, I removed the first row that was simply the lettered Column names ('column A', 'column B', etc.) and stored it into a csv file and dataframe. The data is now structured such that the column names correspond to the data that the column entails.

Substituting the Missing Values: ANOVA, Eta-squared, and Categorical Variable counts to determine which Variable (Region, Sub-region, Income Group) mean to use to substitute mean for missing numerical values:



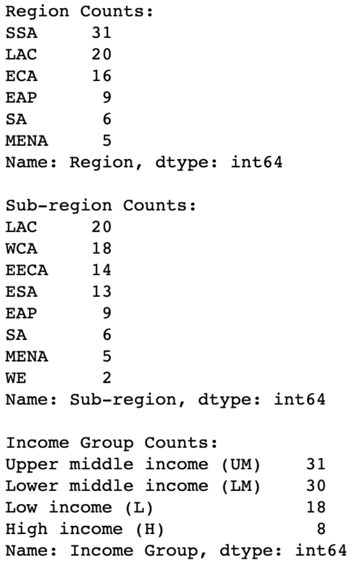
```
Region - F statistic: 15.30, P-value: 0.00000
Sub-region - F statistic: 10.93, P-value: 0.00000
Income Group - F statistic: 20.17, P-value: 0.00000
```

The above is the ANOVA test I conducted to evaluate the best predictor of the 'Total' value. I also conducted an Eta-squared test as displayed below:



```
Region - Eta-squared: 0.4857
Income Group - Eta-squared: 0.4216
Sub-region - Eta-squared: 0.4921
```

In addition to these two, I wanted to see the count for each of the categorical variables I am considering for substituting the mean for the missing values. These are the results for comparison (displayed on the left).

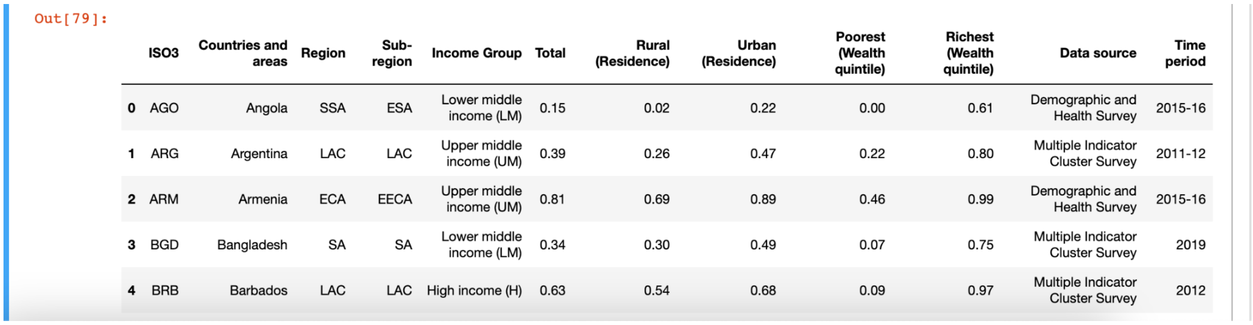


```
Region Counts:
SSA     31
LAC     20
ECA     16
EAP      9
SA       6
MENA     5
Name: Region, dtype: int64

Sub-region Counts:
LAC     20
WCA     18
EECA    14
ESA     13
EAP      9
SA       6
MENA     5
WE       2
Name: Sub-region, dtype: int64

Income Group Counts:
Upper middle income (UM)    31
Lower middle income (LM)    30
Low income (L)              18
High income (H)              8
Name: Income Group, dtype: int64
```

In the ANOVA results, it seems as though that the Income Group is the best predictor of the ‘Total’ percent of primary children with access to Internet, however, the results of the Eta-squared test suggests that the Sub-region might be the best predictor of the ‘Total’ percent of Primary school children with internet access. To pick between ‘Region’ and ‘Sub-region’, I feel it would be best to compare the counts for each of the categorical variables. (Richardson, 2011)

As compared to Sub-region, Region contains more values per category. This gives the variable a higher predictive power as it isn’t limited to a small number of values reserved within its categories. Under the WE Sub-region, there are only two countries, Ukraine and the United Kingdom. Both of these countries have missing values in at least one of their four numerical value columns (‘Rural (Residence)’, ‘Urban (Residence)’, ‘Poorest (Wealth quintile)’, and ‘Richest (Wealth quintile)’). For this reason, I am choosing to substitute a country’s ‘Region’ mean as a replacement for their empty numerical values. This part was tricky and I am sure I might’ve made a mistake in the overall accuracy of the prediction, but I think my reasons were justified. Now that I have given my justification for why I chose ‘Region’ as a substitutive replacement for a country’s missing values, here is how I went about it.

To fill in the missing values, I first turned the strings under ‘Total’, ‘Rural (Residence)’, ‘Urban (Residence)’, ‘Poorest (Wealth quintile)’, and ‘Richest (Wealth quintile)’ into floats so that they can be easily computed. Then, I replaced the empty ‘NaN’ values with the mean of that ‘Countries and areas’ row’s respective ‘Region’. So for example, in the case of Argentina (ARG), which had all four (‘Rural (Residence)’, ‘Urban (Residence)’, ‘Poorest (Wealth quintile)’, and ‘Richest (Wealth quintile)’) values missing, I substituted the ‘NaN’s with the mean of all other non-empty values for that specific column in ‘Region[‘LAC’]’.



Out[79]:

	ISO3	Countries and areas	Region	Sub-region	Income Group	Total	Rural (Residence)	Urban (Residence)	Poorest (Wealth quintile)	Richest (Wealth quintile)	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	0.26	0.47	0.22	0.80	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

Handling the uneven ‘Time period’ values using Regular expressions:
When I looked at the values in the ‘Time period’ column. I inferred that the data is from the 2010’s. I also cross checked this with the original source. (UNICEF, 2021) That means that no value in the ‘Time period’ column can be above 2019 or below 2010. I was able to find 5 different cases of unevenness in the ‘Time period’ column and they are as follows:

- ‘2015-16’

    - Under such circumstances, I want to remove the '-' and ensure that the value is normalized to fit the latest year mentioned. So '2015-16' should be changed to '2016'
- '2562'
    - There is one value in the 'Time period' column that is wrongly typed in. Values like this should be changed back to the closest analog, for 2562, that would be 2012.
- '2027'

    If it is like this, it should be changed to 2017
- '2018-2019'
    - If it is like this, it should be changed to 2019
- '2012-99'
    - Such values should be changed to 2019
- '2076'
    - This is one I had trouble with. I still wasn't able to change the value for 'Nepal (NPL)', which remained 2076 even after trying regular expressions.

Therefore, to summarize, for Data Preparation of Primary.csv, the cleaning process involved handling the following inconsistencies that I felt required fixing:

1. First row: The first row in the original csv was simply labeled as 'column A', 'column B', 'column C', etc. this makes it extremely difficult to pull into a dataframe for analysis and visualization as the actual column names get lumped as data.
2. Missing values in the 'Rural (Residence)', 'Urban (Residence)', 'Poorest (Wealth quintile)', and 'Richest (Wealth quintile)' columns: For the empty cells, I substituted the mean values for the Country's 'Region'. This is after I conducted an ANOVA test, Eta-squared test, and a Variable count to determine which categorical variable would be better for predicting the missing values.
3. Uneven 'Time period' year format: The values in the 'Time period' column weren't in the same standard 2010's format as they should've been. Therefore, I employed regular expressions to ensure they all fall within a 'valid' time period in the 2010's.

Secondary.csv preparation:
Preparing this dataset involved all the steps and measures I took for Primary.csv, but also some additional ones. For example, the 'Total' value for Ukraine was 179%, which likely seems like a typing error. In order to analyze this, 1.79 had to be converted to 0.79. Once this was done, the Secondary School (LM) data seemed like a good mirror to the Primary School (LM) data, but up until the data was cleaned, the data analyses seemed inconclusively challenging.

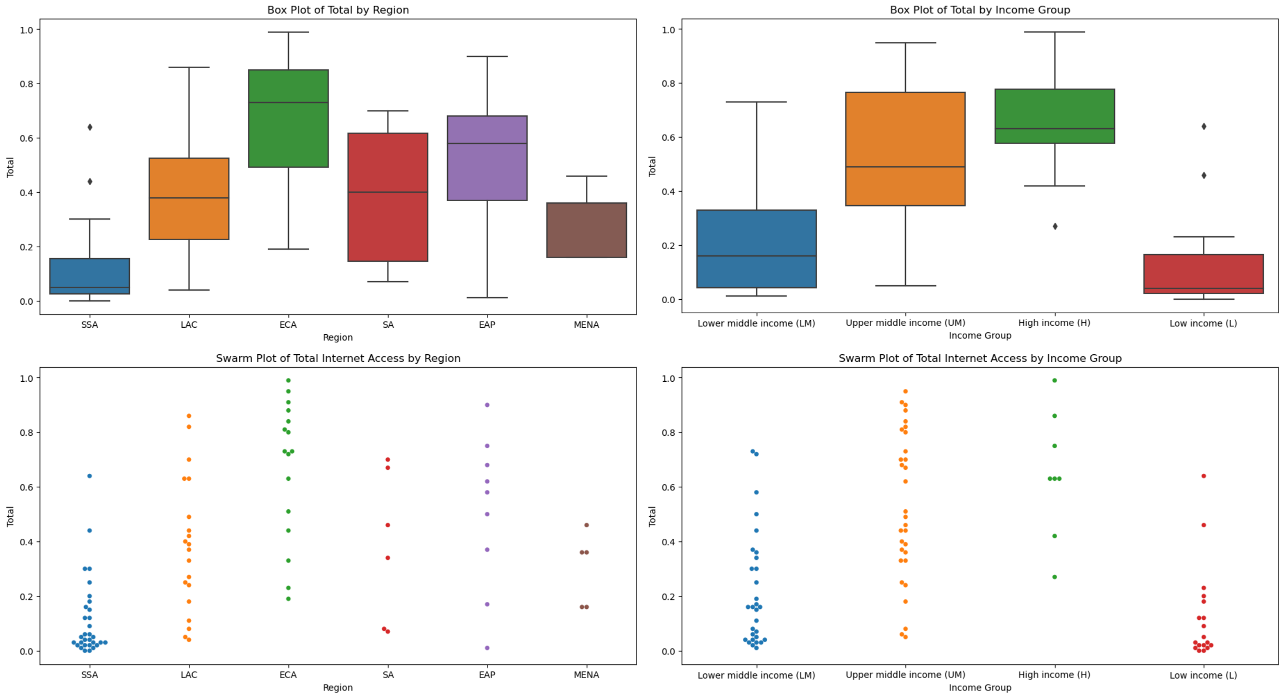
Total School Age.csv preparation:
All the steps involved in the preparation of Primary.csv and Secondary.csv came into play here, however the main difference was that in Task 2.2, I had to use the numerical values in columns which initially consisted of missing values for the analyses. For this reason, it was critical that I employ replicable data science practices to examine the data.

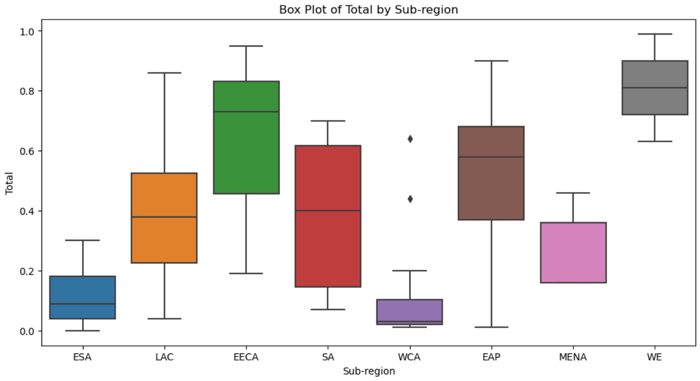
Data Exploration:

Task 2.1: I have chosen 'Region' as my Nominal value column, 'Income Group' as my Ordinal value column, and 'Total' as my Numerical value column, I decided to remove columns G, H, I, J, K, and L. Along with replacing the first row with the accurate column names (which is the subsequent row). These are my reasons for my choice for the variables:

- There are no empty cells in the ‘Total’ column making it the best Numerical variable column for conducting the Data Exploration
- Income group is the best Ordinal variable for predicting the ‘Total’ percentage (as proved by the ANOVA and Eta-squared analyses above)
- Region is a Nominal variable that is specific enough to conduct correlative analyses with, while simultaneously being composed of enough Countries and areas per geographical segment (as compared to Sub-region).

Here are two Boxplots and Swarm plots visualizing the ‘Total’ percentage of primary schoolchildren with Internet access by the Region and Income Group:

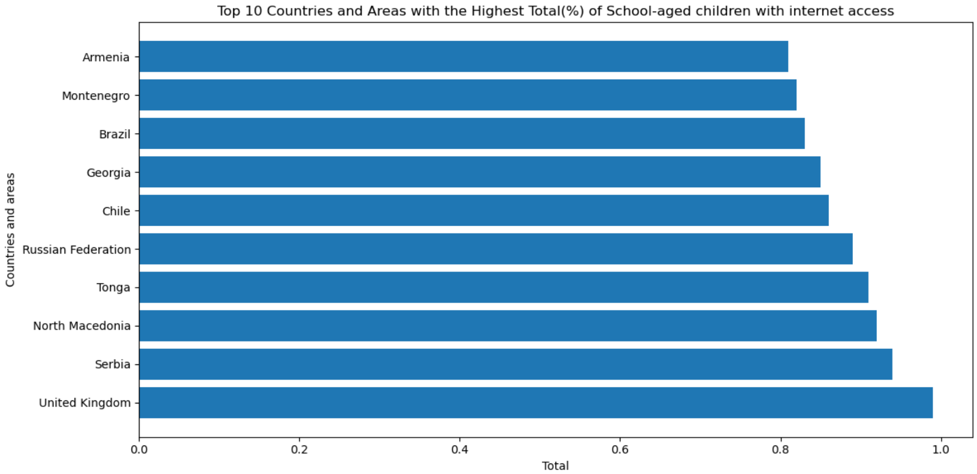


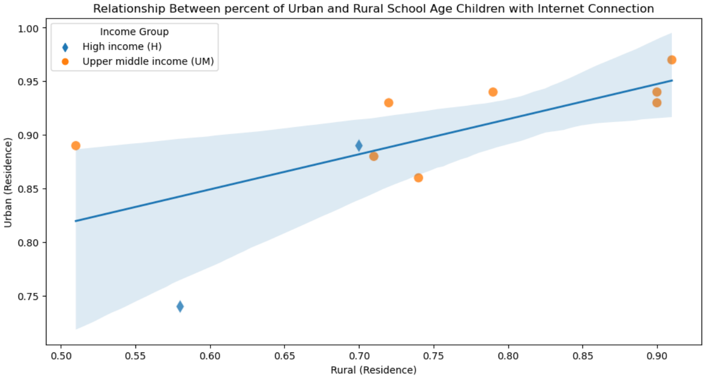


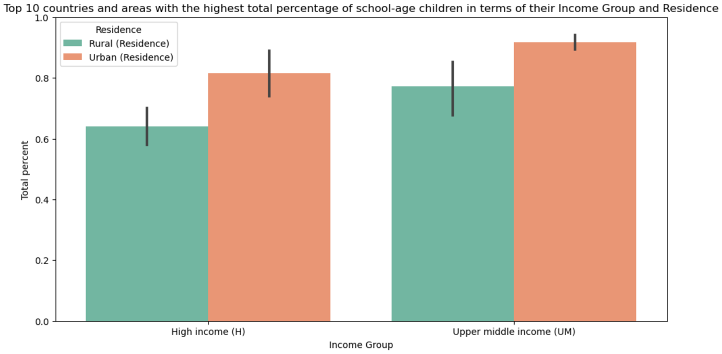
Based on the ANOVA and Eta-squared test, we know that the ‘Region’ is a better predictor of the ‘Total’ percent of primary children with Internet Access value, as compared to ‘Income Group’. (Richardson, 2011) Because of a high count of data points from countries in the SSA Region, LM Income Group, and L Income group, we can say that the data we have is a good representation of those categories. There also seem to be anomalies as visualized in the top left graph in the SSA region. Peeking in, this seems to be the result of the countries Gambia and Sao Tome and Principe, which, after visualizing the ‘Total’ percent boxplots by Sub-region, still seem anomalous (boxplot to the left). I struggled to understand what to do about these two anomalies. (Asli, 2017)

Task 2.2:

Just Below, on the left are the Top 10 countries with the highest 'Total' percent of School-aged children with Internet access. On the right is the data on the Top 10 countries by their Residence and Income Group. We can observe that there is a large correlation between Income Group and Residence, but what is interesting is that it seems as though the disparity between Rural and Urban schoolchildren's access to Internet is larger in High Income (H) countries than it is in Upper middle income (UM) countries.

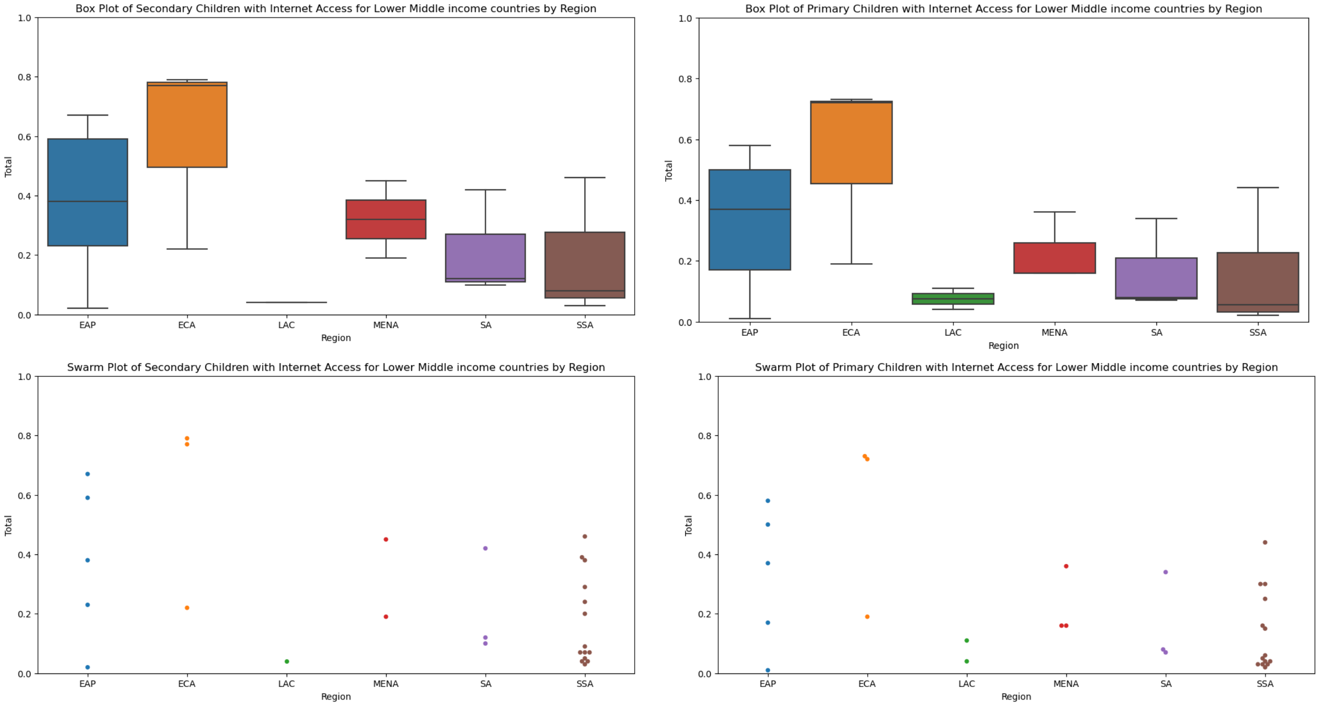






In the top left graph, we can see that that among the Top 10 Countries and areas with the highest percentage of schoolchildren with internet access, while comparing Urban and Rural residences, there seems to be a correlation between a high Total percentage of Urban schoolchildren with Internet access and a high Total percentage of Rural schoolchildren with Internet access. I chose to use a scatter plot as it is a good visualization tool to determine the correlation between two numerical and ordinal variables. (Sainani, 2016) The top right Bar chart with error bars gives us a fascinating insight; that the divide between rural and urban school-aged children's internet connectivity is larger in High income (H) countries as compared to Upper middle income (UM) countries. (Indratmo, 2018)

Task 2.3:
These are the Boxplots and Swarm plots comparing the Total(%) of Primary and Secondary children in Lower middle income (LM) Countries with Internet Access.



Box Plot of Secondary Children with Internet Access for Lower Middle income countries by Region

Box Plot of Primary Children with Internet Access for Lower Middle income countries by Region

Swarm Plot of Secondary Children with Internet Access for Lower Middle income countries by Region

Swarm Plot of Primary Children with Internet Access for Lower Middle income countries by Region

It is self-evident that the spread of data for the Total percent of Primary school children with Internet access mirrors that of Secondary school children with Internet access. Although, because of a larger number of values hailing out of the SSA region means that this visualization is a better representation of SSA than it is for other regions in the Lower middle income (LM) Income Group.

Bibliography:

(Richardson, 2011) John T.E. Richardson, "*Eta squared and partial eta squared as measures of effect size in educational research*", Educational Research Review, Volume 6, Issue 2, 2011, Pages 135-147, ISSN 1747-938X, https://doi.org/10.1016/j.edurev.2010.12.001.

(UNICEF, 2021) United Nations Children's Fund, "*State of the World's Children Statistical Annex 2017"*, UNICEF, New York, July 2021, https://data.unicef.org/resources/dataset/digital-connectivity/

(Sainani, 2016) Sainani, K.L. (2016), The Value of Scatter Plots. PM&R, 8: 1213-1217. https://doi.org/10.1016/j.pmrj.2016.10.018

(Indratmo, 2018) Indratmo, Lee Howorko, Joyce Maria Boedianto, Ben Daniel, The efficacy of stacked bar charts in supporting single-attribute and overall-attribute comparisons, Visual Informatics, Volume 2, Issue 3, 2018, Pages 155-165, ISSN 2468-502X, https://doi.org/10.1016/j.visinf.2018.09.002.

(Asli, 2017) Aslı, Özgün-Koca (2017) Interpretations of Boxplots: Helping Middle School Students to Think Outside the Box, Journal of Statistics Education, 25:1, 21-28, DOI: 10.1080/10691898.2017.1288556