Student name: Prerak Soni Student ID: S3787596

Data Preparation

Error 1 - Label names being relevant

As the dataset given to us is split into 3 different categories e.g Primary, Secondary and Total school Age. I have decided to load each csv into a dataframe and then work on the curation of the data. I then used the head() function to ensure that the loaded data is correct and matches the csv files. The column labels were Column A,B,C instead of their actual labels so I used the iloc function to convert the 1st row(the labels) into the column header and then removed the 1st row after the labels were changed for all the datasets.

Error 2- NaN values

NaN values. I replaced these values with 0% using the fillna function because this allows us to run operations on the data if needed in the upcoming tasks. This changed the values to string meaning I needed to change the values to int64 so they can be used for operations and visualisations.

Error 3- Data type change to int64

Converting all the percentages to int64 so we can use them to create visualisations and also to compare between the different dataset csvs. While doing this process I found that one of the percentages (‘15%’) was a string meaning I had to convert that into a NaN because the astype function wasn’t working. That being said, we have to remove all the ‘%’ from the columns to convert them into ints. I did that using astype as well after using the replace function to remove ‘%’ from all the columns.

Error 4- Typos/False values

Looking for typos or false values in sub-region columns. I used the isin function to figure if the listed sub regions in the ReadMe file were present in the dataset or not and then printed out the invalid rows so then I can replace them with the correct value. I found this error in the Primary.csv where there were two rows with ‘WE’ as the sub region, one being “Italy” and the other country being “United Kingdom”. As we are not including ‘WE’ as a sub-region according to the ReadMe file so I used np.nan to change the value to NaN.

Error 5- Unrelated Columns

Removing columns that will not be needed in the upcoming tasks to clear the dataset up allows us to visualise the tables better when using head() or shape, etc. I did this using the drop function to remove multiple columns.

Error 6- Whitespace

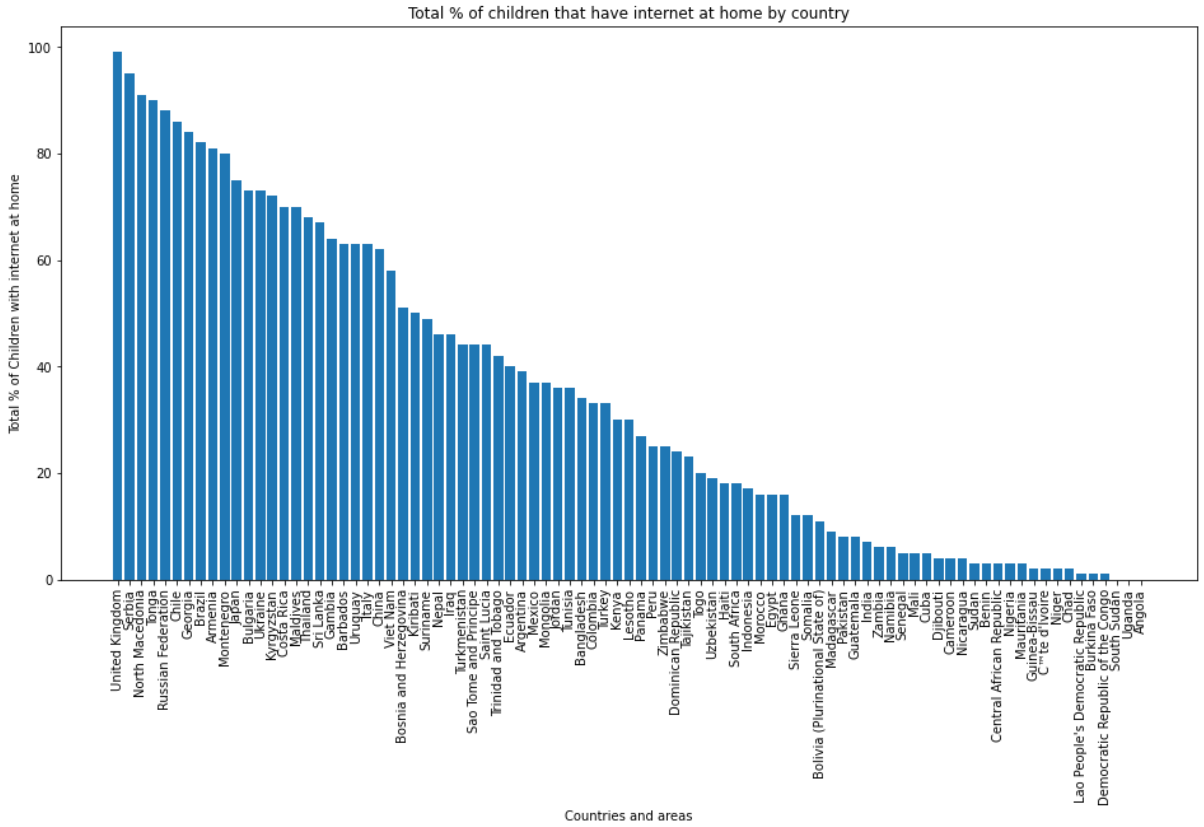
Removing the whitespace in all the dataset csvs by using the strip function which helps clean up the data even more allowing it to be used for analysis.

Data Exploration

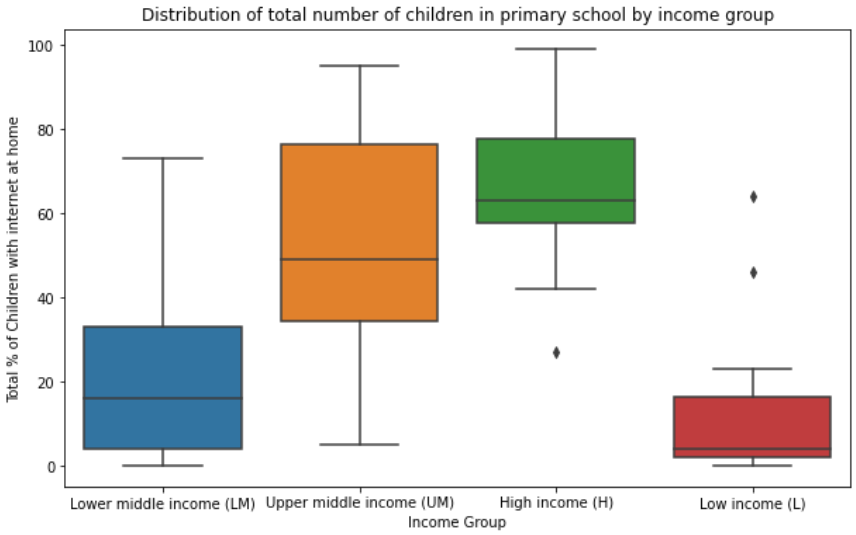
Task 2.1

For this task I picked the ‘Country’ as the nominal column as this allows us to see the different rates of internet that the children have whilst they are in primary school. For the ordinal column I picked the ‘Income Group’ to understand if the amount of income that the child is placed in actually impacts whether they have internet access or not. For both the nominal and ordinal choices I will be visualising them against the ‘Total’. The numerical choice I will compare ‘Rural’ and ‘Urban’ residence and see the difference in percentage.

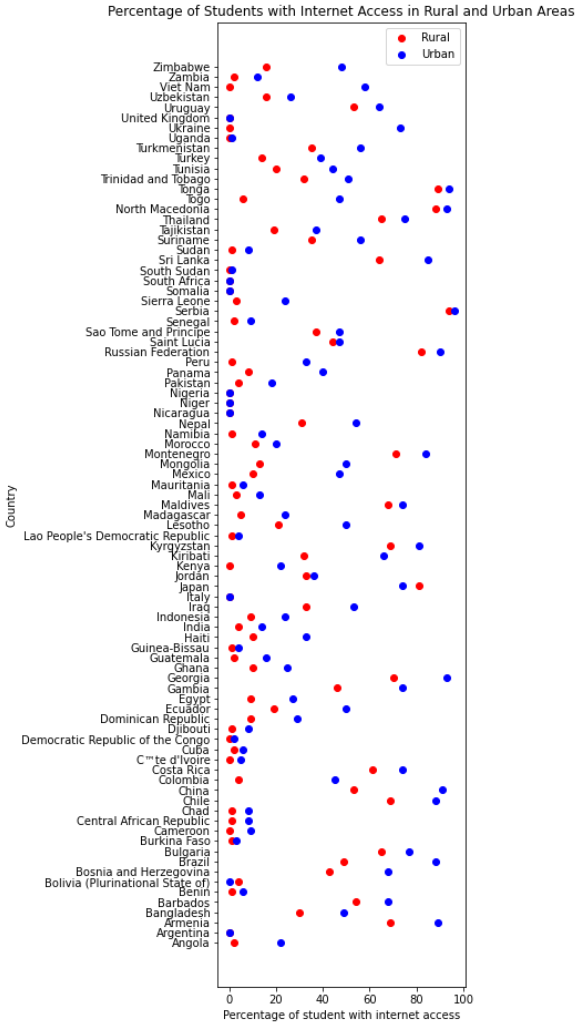
Nominal: I picked Country and decided to plot it against the Total percentage of children that have access to internet at home within the country. By using the bar graph you can see the difference in all the countries that recorded the percentage of children with internet access at home. I sorted the values in descending form showing the highest and the lowest rates.



Ordinal: For the ordinal I picked ‘Income Group’ which displays how the income of a household impacts the access to internet for a child and decided to use a boxplot to display the different income groups. As you can see below the income group heavily impacts the percentage as expected.



Numerical: I used ‘Rural’ & ‘Urban’ graphed with ‘Countries and areas’ using a scatter plot to show how each country places based on the area. Based on the country many of the Rural areas have less internet access compared to Urban areas. We can say this may be due to the wiring and actually connection being built in the areas.

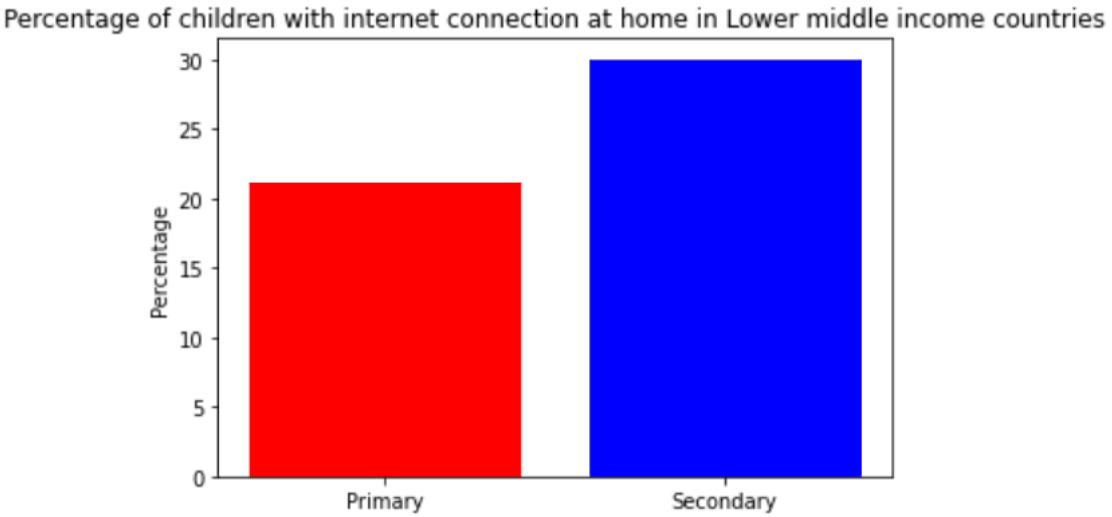


Task 2.2

	ISO3	Countries and areas	Region	Sub-region	Income Group	Total	Rural (Residence)	Urban (Residence)
63	SRB	Serbia	ECA	EECA	Upper middle income (UM)	94	91	97
75	TON	Tonga	EAP	EAP	Upper middle income (UM)	91	90	94
73	MKD	North Macedonia	ECA	EECA	Upper middle income (UM)	92	90	93
36	JPN	Japan	EAP	EAP	High income (H)	78	83	77
59	RUS	Russian Federation	ECA	EECA	Upper middle income (UM)	89	79	94
49	MNE	Montenegro	ECA	EECA	Upper middle income (UM)	82	74	86
28	GEO	Georgia	ECA	EECA	Upper middle income (UM)	85	72	93
4	ARM	Armenia	ECA	EECA	Upper middle income (UM)	81	71	88
16	CHL	Chile	LAC	LAC	High income (H)	86	70	89
40	KGZ	Kyrgyzstan	ECA	EECA	Lower middle income (LM)	74	69	83

These are the top 10 countries with highest internet access based on Rural and Urban areas and the Income group. Majority of the income group here is Upper middle income. We can also see the majority of the region these countries are in is ECA being Europe and Central Asia which shows us that these countries have a more developed internet connection across their country allowing people even in Rural areas to have access to the internet.

Task 2.3



After filtering the total with the Income Group being Lower and middle income. I used the mean function to get the average percentage of children who have access to the internet. As displayed in the figure above Secondary children tend to be at a higher percentage due to the requirement for higher studies and the importance of the internet.