DATA SCIENCE PART I REPORT

COVER PAGE

GROUP NAME: **Zigzagoon**

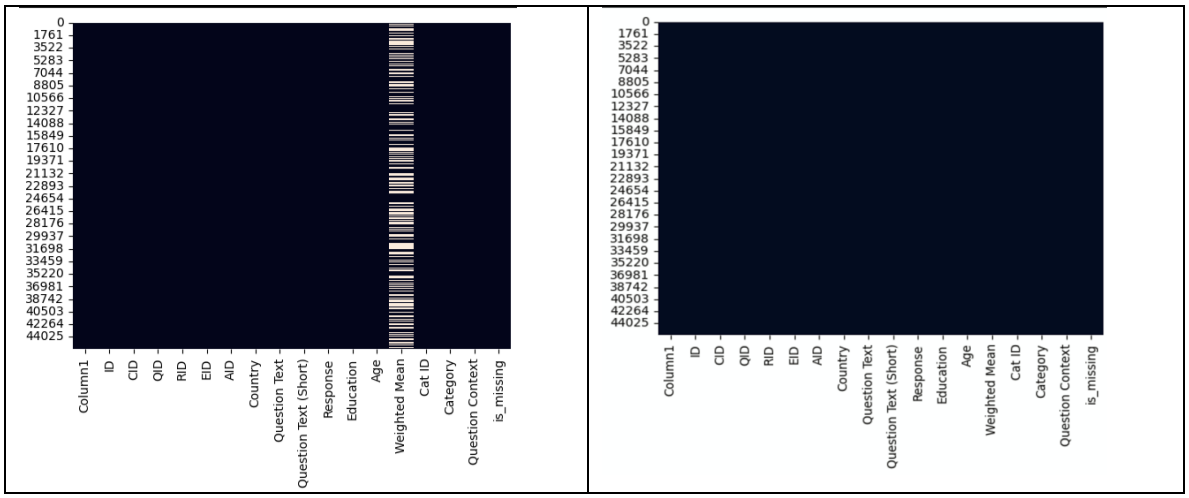
Group Members: **1) Gaurav Dubey(mznq0903)**

2) Tusshar Kashyap(nccm0794)

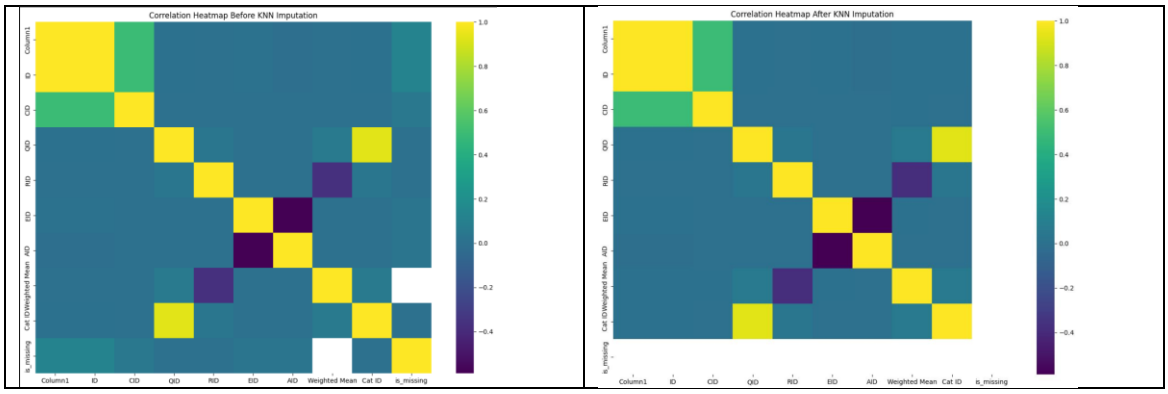
3) Shreya Singh(trbd0672)

4) Yatharth Bali(pmdz0960)

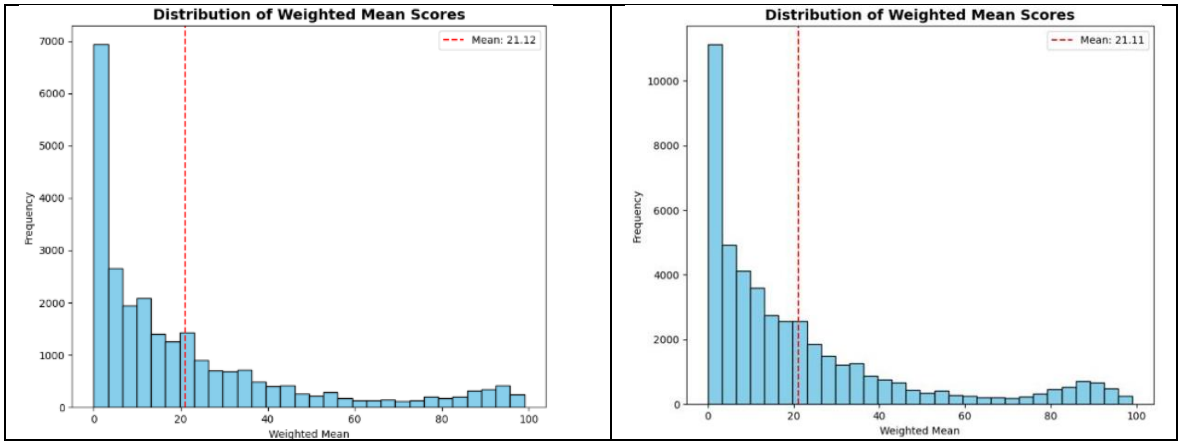
PRE- PROCESSING



In the dataset provided, there are 20261 null values in Weighted Mean column, so we have three approaches to solve it – first one is dropping the values which is not the optimal way as half of the data will get reduced. Second way is filling the data via mean, median or mode method but after doing that all the null values are filled with same numbers, so after consideration we decided to take KNN, but limitation of KNN is that it works slow but it will give the best results out of rest other methods. Apart from KNN, there is one more method that is Imputer, which can be also as good as KNN.

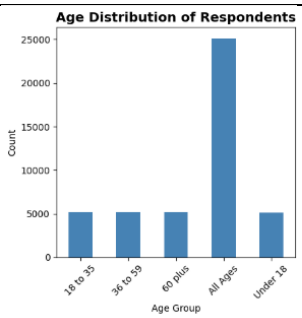


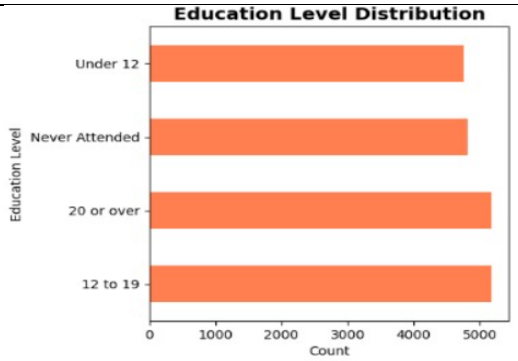
As we can see in the left graph the weighted mean has many missing values because of that its correlation with other columns was weak, but after KNN imputation weighted mean has no missing values and correlation becomes stronger and more realistic. Unlike mean imputation KNN uses multiple nearest rows to estimate missing values.



With reference to above graph of Distribution of Weighted Mean Score we can observe that in first observation there were 20k cases whereas in the second observation there are 45k values.

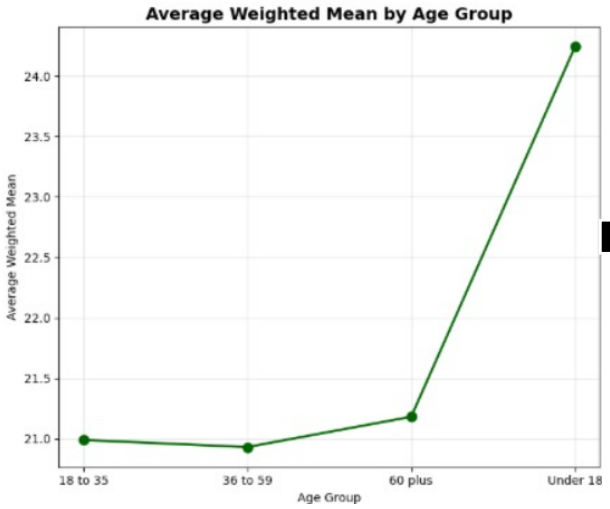
EXPLORATORY DATA ANALYSIS





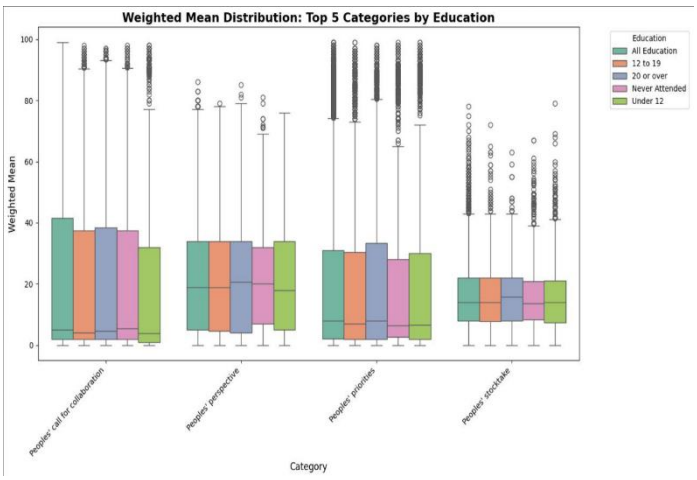
Here we can observe that different age groups have different count of respondents and they have been categorized into 4 distinct age groups and "*all age*" defines the mean of all category.

In this graph there are 4 education levels with nearly equal number of counts and "*all education*" is the average mean of all these levels (24.73+24.96+24.82+25.04 /4) this is equals to 24.88 which is approximate 25.01.

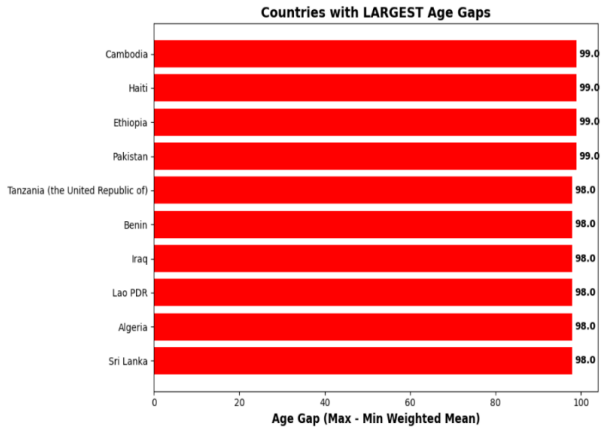


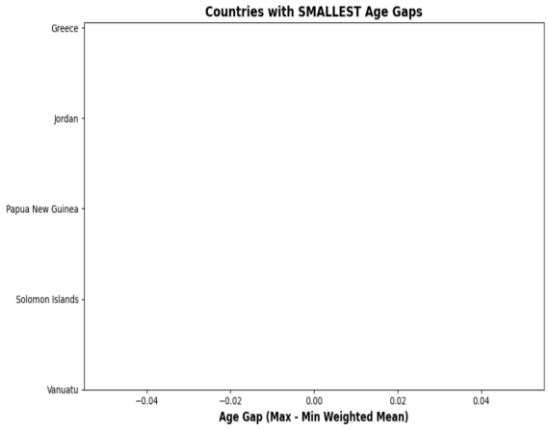
Age from 18-59 represent a consistent value, then age group 60 plus depicts a significant elevated graph which indicates a different response pattern. At last, age group under 18 shows moderately higher values then young adults.

Peoples call for collaboration has highest median, followed by people's perspective with moderate variance, then people's priorities are like collaboration and remain elevated across all education levels then most consistent category is the peoples' social care.



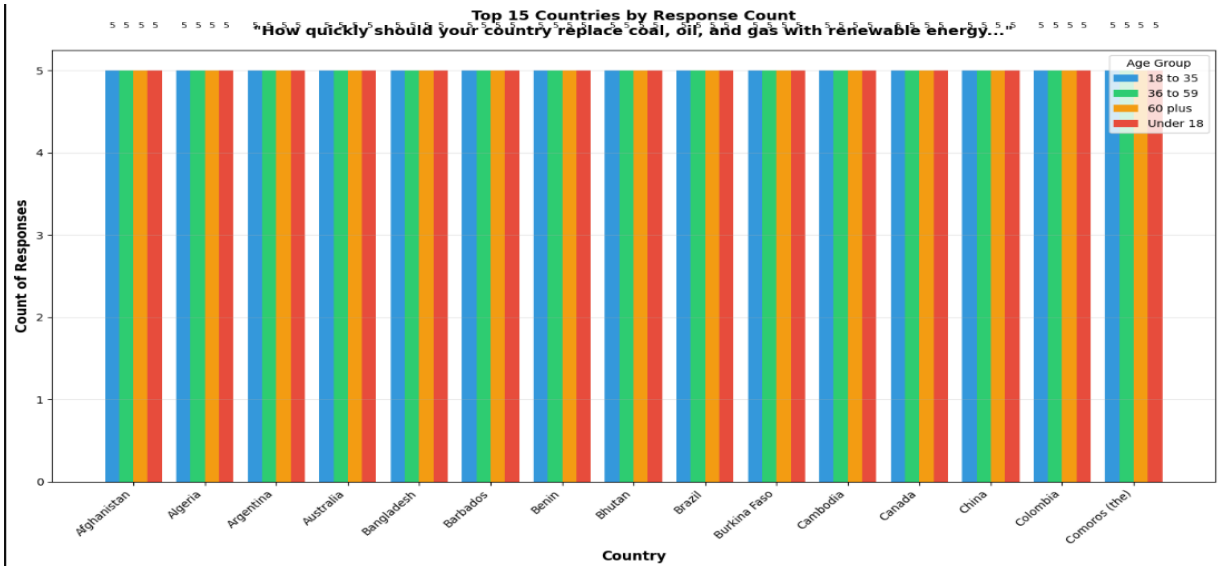
COUNTRIES WITH LARGEST AND SMALLEST AGE GAP





From the above two graphs we find that Cambodia, Haiti, Ethiopia, and Pakistan have the biggest age gaps, while Greece, Jordan, Papua New Guinea, the Solomon Islands, and Vanuatu have the smallest.

Consistent age gap for countries in support for replacing fossil fuels with renewable energy.



From the above chart, there are many countries with consistent age gap and they want replace oil and gas with renewable energy, with columns same or minor difference.

CONCLUSION:

There is overall data quality improvement with implementation of KNN, the distribution of weighted mean score expanded from 20k to 45k. Educational attainment is well balanced and we can find that age group 12-19 has the highest count among all. There is consistent support for replacing oil and gas with renewable energy, many countries including Afghanistan, Algeria etc. from all age group lie on the same plane. Overall, the analysis shows that educational outcomes are balanced and that attitudes across generations are similar in many countries.