**EXTRACT**

We had two datasets 2016 & 2021 Olympic Data which we downloaded from Kaggle and saved into a folder called Resources

2016 Olympic Data consist of spreadsheets:

2021 Olympic Data consist of spreadsheets

**TRANSFORMATION**

We loaded the medals\_2021\_file from the Resources folder and printed the dataframe

We loaded the athletes\_2021 csv from the Resources folder and printed the dataframe called ath\_2021\_file and rename column column NOC to country so both datasets have the same columns.

We loaded the countries\_2016\_file from the Resources folder create a dataframe called country\_2016\_df

Then we did the describe function to see the stats of the data and saw the top athlete was Jiana Wang. We then used The clean\_ath\_2016 and sorted by name Kurbanov, to see what other events he participated in and called this dataframe top\_athlete\_2016. This revealed the events of table tennis and athletics.

Take clean\_ath\_2016\_df and choose columns Name, nationality,sex,sport,gold,silver & gold

Then we did the describe function to see the stats of the data and saw the top athlete was Rustan Kurbanov. We then used The clean\_ath\_2021 and sorted by name Kurbanov, to see what other events he participated in and called this dataframe top\_athlete\_2021. This revealed the events of fencing and athletics.

We created a dataframe called medal\_2016 where we took merged\_2016 dataframe and groupby(country and medals: gold, silver and bronze to get the total medals for 2016 and drop all na’s and add a column called total 2016 with the medal totals

Create a dataframe called clean\_country\_df where we rename the column code to nationality in the country\_2016\_df

Create a dataframe called merged\_2016\_df in which you merge clean\_ath\_2016 and clean\_country \_df on nationality with a left join

Take dataframe called cleaned\_merged\_medals and choose columns country total 2016 and total

We took dataframe called cleaned\_merged\_medals and choose columns country total 2016 and total and rename total column to medals2021

Then took dataframe called cleaned\_merged\_medals and dropped all na’s

We created a dataframe called merge\_2021\_df which merges clean\_ath\_2021\_df & medals\_2021\_df and merge on the country column using outer join

Used the merged\_2021\_dataframe get total count of Athletes per sport per country.

Used the merged\_2021\_dataframe and groupby(country and sport) and sum number of athletes per country per sport