# KAPSARC Task

By Sultan Alharbi

## Overview:

This task is part of KAPSARC GDP evaluation for my skills, I was given a dataset related to oil production, and the goal was to clean it, create some visualizations, and build a web interface to add the visualizations.

## Data:

The data contains 123 attributes and 16 features, each attribute indicates a country and its related numbers, and in terms of features the first feature indicates the country name, and the other 15 indicates each months from March 2022 up until May 2023.

## Data Acquisition:

The data was downloaded from the tool bar in the provided website, I have downloaded the data in Microsoft Excel Format A.K.A (.xls) .

## Pre-Processing:

| | Joint Organisations Data Initiative - Primary (15 months) | Unnamed: 1 | Unnamed: 2 | Unnamed: 3 | Unnamed: 4 | Unnamed: 5 | Unnamed: 6 | Unnamed: 7 | Unnamed: 8 | Unnamed: 9 | Unnamed: 10 | Unnamed: 11 | Unnamed: 12 | Unnamed: 13 | Unnamed: 14 | Unnamed: 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| 1 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| 2 | Unit | Thousand Barrels per day (kb/d) | Product | Crude oil | BALANCE | Production | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| 3 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| 4 | Time | Mar2022 | Apr2022 | May2022 | Jun2022 | Jul2022 | Aug2022 | Sep2022 | Oct2022 | Nov2022 | Dec2022 | Jan2023 | Feb2023 | Mar2023 | Apr2023 | May2023 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 118 | United States of America | 11700.8065 | 11668.4 | 11629.129 | 11797.2667 | 11844 | 12002.4839 | 12337.3333 | 12416.871 | 12379.2667 | 12148.5806 | 12568.4516 | 12525.0714 | 12695.5161 | 12614.7 | 12274.1935 |
| 119 | Uruguay | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 120 | Venezuela | 728 | 775 | 735 | 727 | 629 | 723 | 666 | 717 | 693 | 669 | 732 | 704 | 754 | 810 | 819 |
| 121 | Vietnam | 187.8828 | 183.0435 | 180.1768 | 192.0438 | 174.9067 | 181.3884 | 176.544 | 166.5778 | 174.8514 | 179.1542 | 172.0589 | 176.1143 | 0 | 0 | 0 |
| 122 | Yemen | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

123 rows × 16 columns

As we can notice from the figure above, there are several issues with the data:

1- The unnamed columns.

2- The null values from the first row to the fourth row.

So, what I have done, was the following:

1- used (skiprows = 5 ), alongside with pd.read_xls, to skip the rows with nulls.

| | Time | Mar2022 | Apr2022 | May2022 | Jun2022 | Jul2022 | Aug2022 | Sep2022 | Oct2022 | Nov2022 | Dec2022 | Jan2023 | Feb2023 | Mar2023 | Apr2023 | May2023 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Country | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| 1 | Albania | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0 | 0.0000 |
| 2 | Algeria | 996.0000 | 1006.0000 | 1015.0000 | 1027.0000 | 1040.0000 | 1053.0000 | 1058.0000 | 1060.0000 | 1021.0000 | 1009.0000 | 1012.0000 | 1014.0000 | 1008.0000 | 999.0 | 962.0000 |
| 3 | Angola | 1133.0000 | 1183.0000 | 1162.0000 | 1175.0000 | 1180.0000 | 1179.0000 | 1091.0000 | 1051.0000 | 1088.0000 | 1088.0000 | 1105.0000 | 1064.0000 | 972.0000 | 1063.0 | 1111.0000 |
| 4 | Argentina | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0 | 0.0000 |

Now this is what the dataframe look like after skipping the rows, we can notice that we have Time, Mar2022, … May 2023 as columns, but the first row contains ‘Country’ in Time column and null values in the others, so we will drop the row since it will not be informative for us.

2- After dropping the first row, we can notice that the Time columns doesn’t indicates any time related values, instead of that it indicates country name, So we need to rename the Time column to Country.

| | Country | Mar2022 | Apr2022 | May2022 | Jun2022 | Jul2022 | Aug2022 | Sep2022 | Oct2022 | Nov2022 | Dec2022 | Jan2023 | Feb2023 | Mar2023 | Apr2023 | May2023 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Algeria | 996.0000 | 1006.0000 | 1015.000 | 1027.0000 | 1040.00 | 1053.0000 | 1058.0000 | 1060.0000 | 1021.0000 | 1009.000 | 1012.0000 | 1014.0000 | 1008.0000 | 999.0000 | 962.0000 |
| 1 | Angola | 1133.0000 | 1183.0000 | 1162.000 | 1175.0000 | 1180.00 | 1179.0000 | 1091.0000 | 1051.0000 | 1088.0000 | 1088.000 | 1105.0000 | 1064.0000 | 972.0000 | 1063.0000 | 1111.0000 |
| 2 | Argentina | 0.0000 | 0.0000 | 0.000 | 0.0000 | 0.00 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 3 | Armenia | 0.0000 | 0.0000 | 0.000 | 0.0000 | 0.00 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 4 | Australia | 287.6753 | 306.7265 | 323.285 | 315.1372 | 236.55 | 255.6266 | 288.3282 | 295.5603 | 283.0715 | 288.184 | 271.1423 | 277.6646 | 258.6789 | 241.0182 | 256.1353 |

This is how the dataframe looks after applying step number 2, but we still have a little problem, the numbers in the dataframe are showing 4 numbers after the comma , so we need to fix that since we are dealing with large numbers ( Thousands per day ) and the values after the comma will not be that useful to us, so we will round it to one number after the comma.

| | Country | Mar2022 | Apr2022 | May2022 | Jun2022 | Jul2022 | Aug2022 | Sep2022 | Oct2022 | Nov2022 | Dec2022 | Jan2023 | Feb2023 | Mar2023 | Apr2023 | May2023 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Algeria | 996.0 | 1006.0 | 1015.0 | 1027.0 | 1040.0 | 1053.0 | 1058.0 | 1060.0 | 1021.0 | 1009.0 | 1012.0 | 1014.0 | 1008.0 | 999.0 | 962.0 |

This is how it looks after rounding the numbers to one number after the comma.

3- We need to check for any null values after what we did.

```
Country     0
Mar2022     0
Apr2022     0
May2022     0
Jun2022     0
Jul2022     0
Aug2022     0
Sep2022     0
Oct2022     0
Nov2022     0
Dec2022     0
Jan2023     0
Feb2023     0
Mar2023     0
Apr2023     0
May2023     0
dtype: int64
```

**Note:** there are some zeros in other countries data, but we don't have to drop it because it could be true since not all of the countries in the datasets are oil producing countries.

4- Checking the type of our features, and it's all good.

```
Country      object
Mar2022     float64
Apr2022     float64
May2022     float64
Jun2022     float64
Jul2022     float64
Aug2022     float64
Sep2022     float64
Oct2022     float64
Nov2022     float64
Dec2022     float64
Jan2023     float64
Feb2023     float64
Mar2023     float64
Apr2023     float64
May2023     float64
dtype: object
```

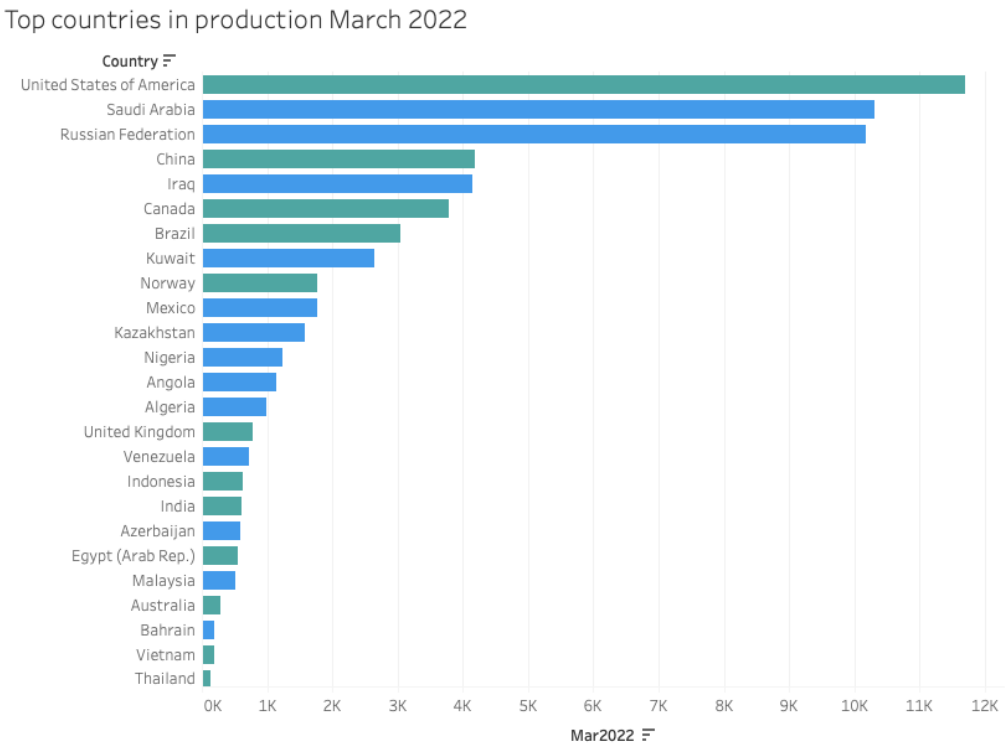
5- Finally, saving our cleaned dataframe into .xls file for the data analysis phase.

**Note:** the preprocessing was done using Python, for more details please check the notebook provided.

## Data Visualization:

In this section I will go through the visualization I made from the data provided.

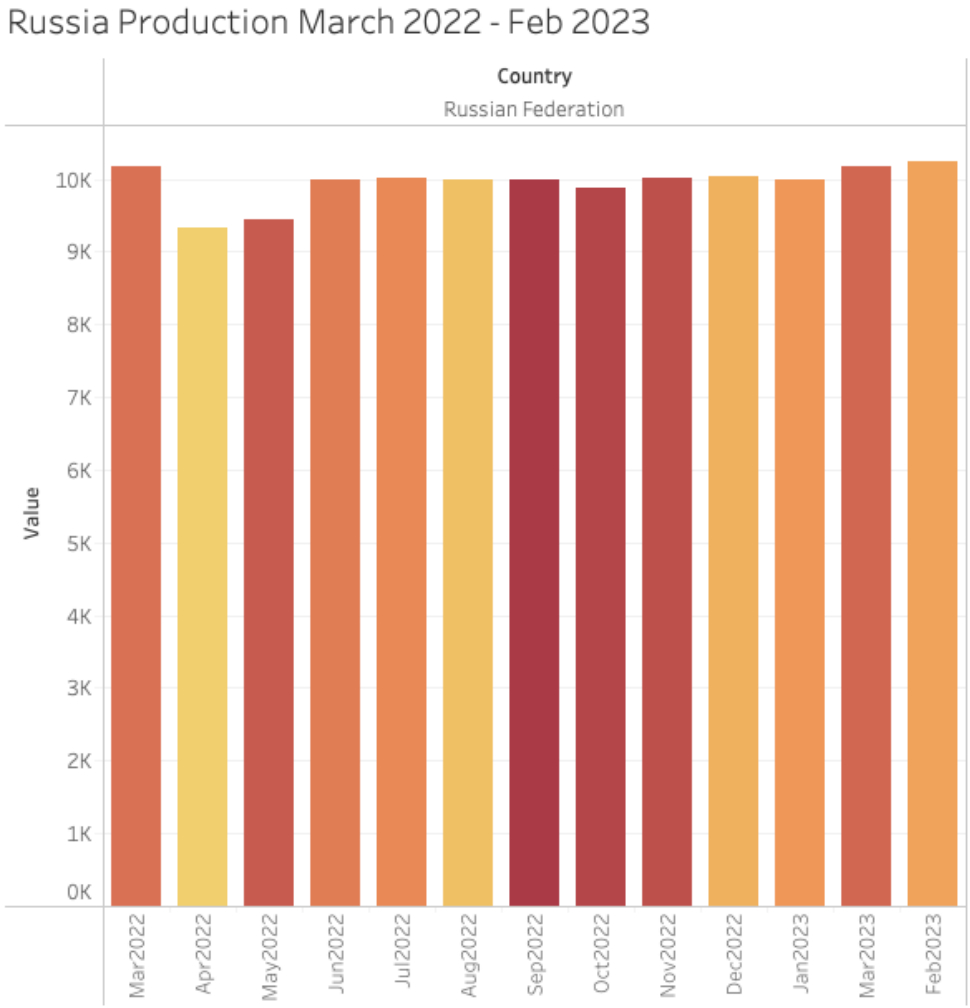
## Top Countries In Terms Of Production On March 2022:




We can notice that there are two colors, the blue is for OPEC+ countries and the green is for other countries.

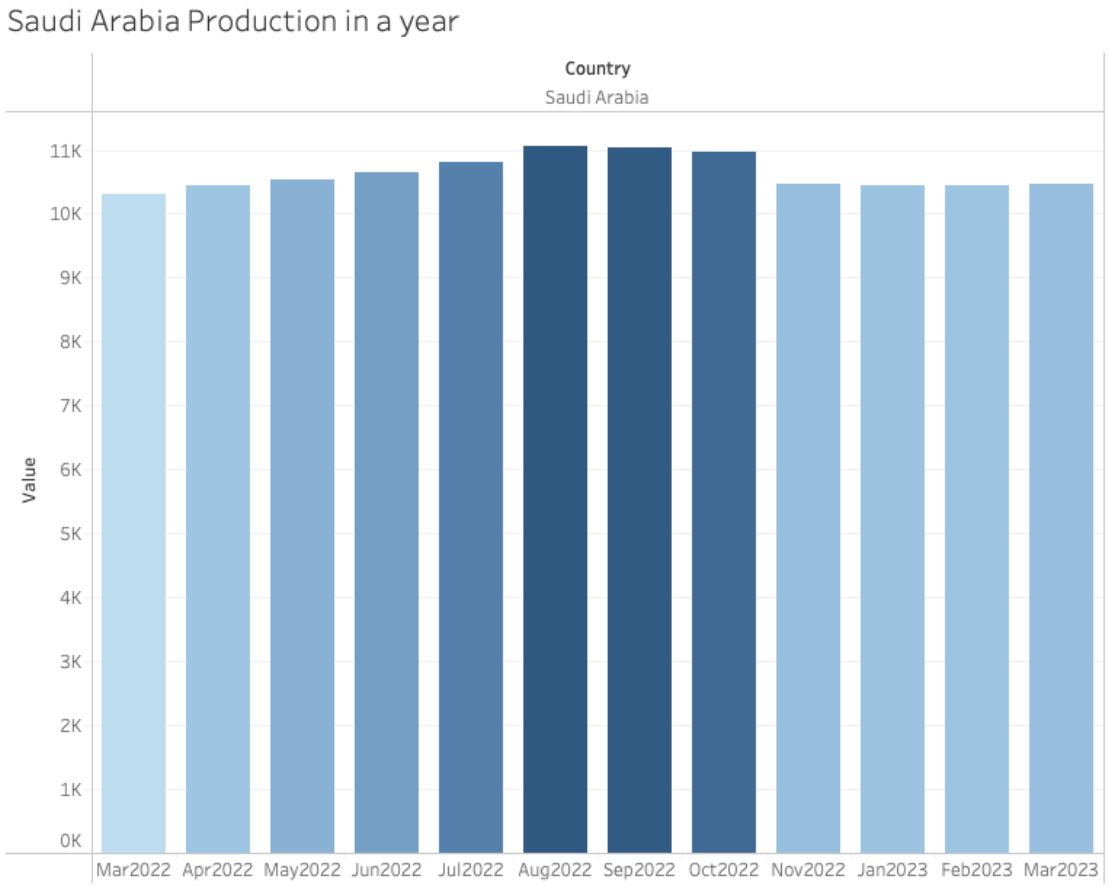
Also we can notice that United States is the highest in terms of oil production, with over than 12 Million barrel per day, even though it's not from OPEC+ organization.

**Russia's Production for the period from March 2022 until Feb 2023:**




We can notice that Russia's production amount was always around 10 Million barrel per day, but it has dropped to approximately 9.3 Million in April 2022 and 9.4 Million in May 2022, just after two to three months from the beginning of the war on Ukraine, and the production reach it's peak in February 2023 exceeding 10.2 Million Barrel Per Day.

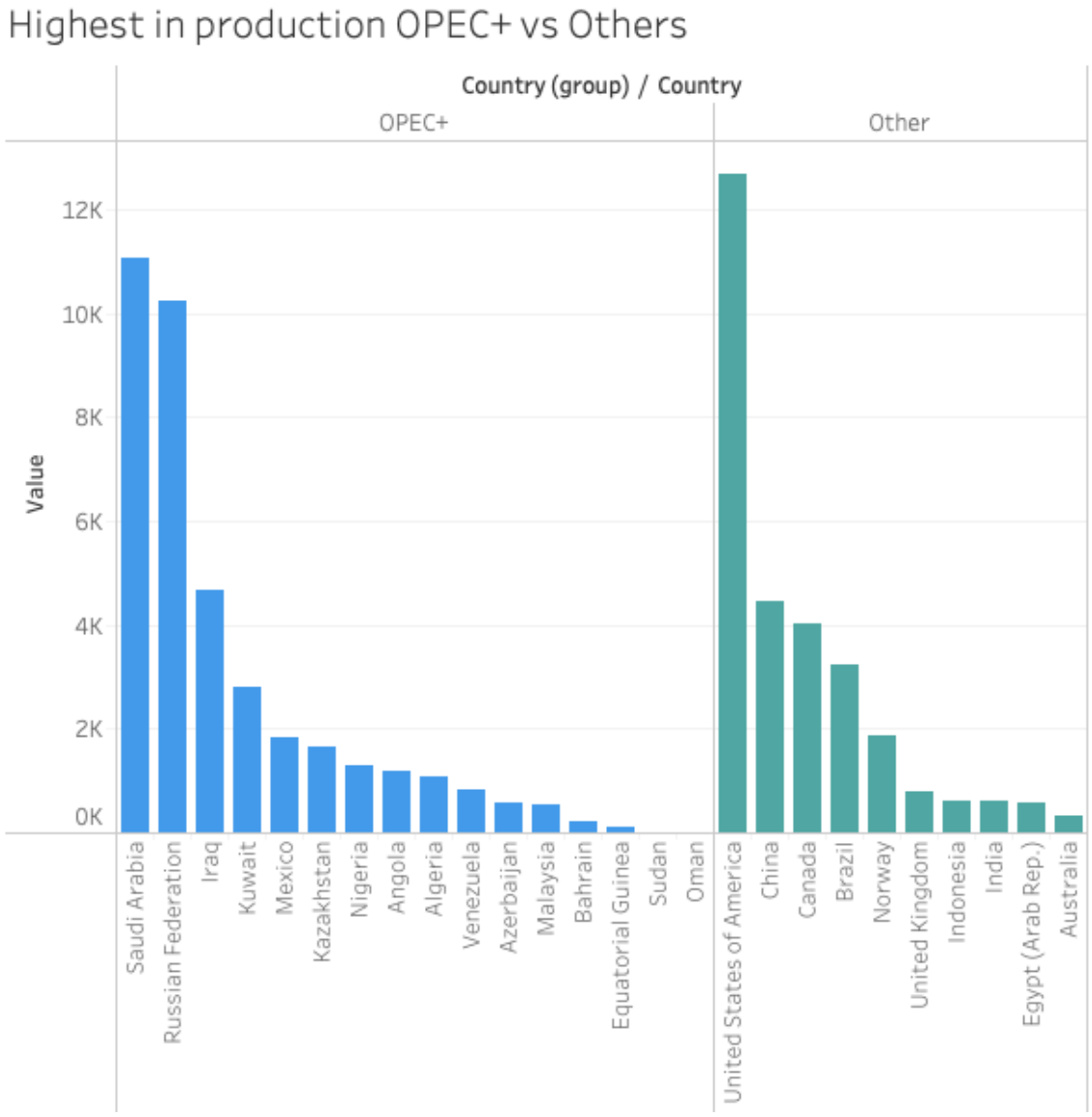
## Saudi Arabia's Production in a year:




This chart shows the oil production of Saudi Arabia for the time period from March 2022 until March 2023.

We can notice that the amount of oil produced nearly stable at 10.4 Million per day, but from Jun 2022 it started to increase until it reach its peak in August 2022, with 11.05 Million per day , and it goes back to the normal amount in November 2022.

**Top Production OPEC+ vs Other:**




We can notice that the Saudi Arabia and Russia are the highest in terms of production in OPEC+ with a huge difference from Iraq which comes third, and on the Other side, USA is the highest country outside OPEC+ umbrella, with a huge difference from China and Canada.

Best Regards,
Sultan Alharbi.