

**TASK**

**Exploratory Data Analysis on the fuel imports Data Set**

[](http://www.hyperiondev.com/portal/)

**Introduction**

The fuel imports dataset contains data of the imports of fuels into South Africa and Botswana. The dataset provides information on where the fuel is from, the tariffs associated with the fuel, where the fuel is imported to, the statistical unit used to measure the fuel (which can be used to determine the type of fuel), detailed descriptions of the fuel, how the fuel has been transported to its destination and the customs value of the fuel. In this report I provide insight into the trade regarding fuel imports to South Africa.

**DATA CLEANING**

In order to get an overview of all the columns provided in the data set, the datatypes of each column were viewed. This allowed to us to both see the columns provided, as well as ascertain which datatypes the values in the columns were saved as.

I found that the columns labelled “yearmonth” and “calendaryear” were saved as int64 datatypes. Since both these columns represent time, they were converted to datatime datatypes. I created new columns for these new values and saved them under the newly created “yearmonth\_parsed” and “calendaryear\_parsed” columns.

The dataset contained imports to not only South Africa, but imports to Botswana as well. Since I was only concerned with analysing imports to South Africa, I dropped any rows that provide information regarding the importation of fuels to Botswana.

Variables which were irrelevant and redundant (features which were not used in the analysis) were also dropped. This made the dataset easier to read and more compact. Several columns provided postal abbreviations for countries, which were available in other columns. These were dropped since they were represented elsewhere in the data set. These columns were: districtofficecode, countryoforigin and countryofdestination. Columns which provided a code were also represented elsewhere in the data set and, therefore, could also be dropped. These columns included: transportcode, section and chapter. Columns which were dropped due irrelevance were: statisticalunit and statisticalquantity. The columns ‘chapteranddescription’ and ‘sectionanddescription’ had no variation and could therefore be dropped as well.

The column tariffanddescription contains numbers, separated by a dash (-), followed by a description of the fuel. These values needed to be separated so that only the relevant information (the description) was represented in the column. This was done by using the str.split( ) function, with the dash (-) used as the delimiter. I then created a new column for these values named ‘tariffdescription’.

MISSING DATA

Once the dataset was cleaned, I then checked to see if any missing data occurred, and also the percentage of missing data. I found that only ‘countryoforigin’ had missing data. However, I dropped this column during the cleansing process and was thereby dealt with.

DATA STORIES AND VISUALIZATIONS

Figure 1: Number of imports per country bar graph

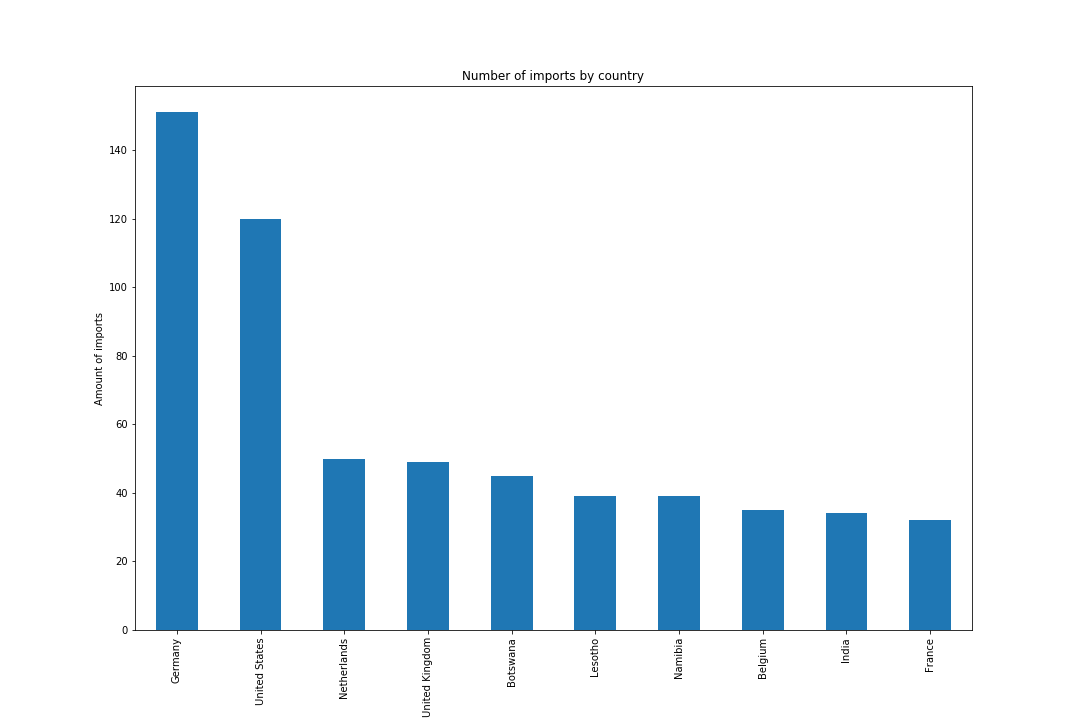


Figure 2: Number of imports per country word cloud

By checking the frequency of each country in the dataset, I was able to determine which countries South Africa imports the most fuel from. This showed that the majority of the imports originated from Germany, and the United States. These two countries provided South Africa with more than 50% of its fuel. We see that other countries appear to provide South Africa with a substantial amount of fuel as well. Countries like the United Kingdom, Netherlands, Namibia, Lesotho and Botswana all play a significant role in the importation of fuel to South Africa. (Figure 1, figure 2)

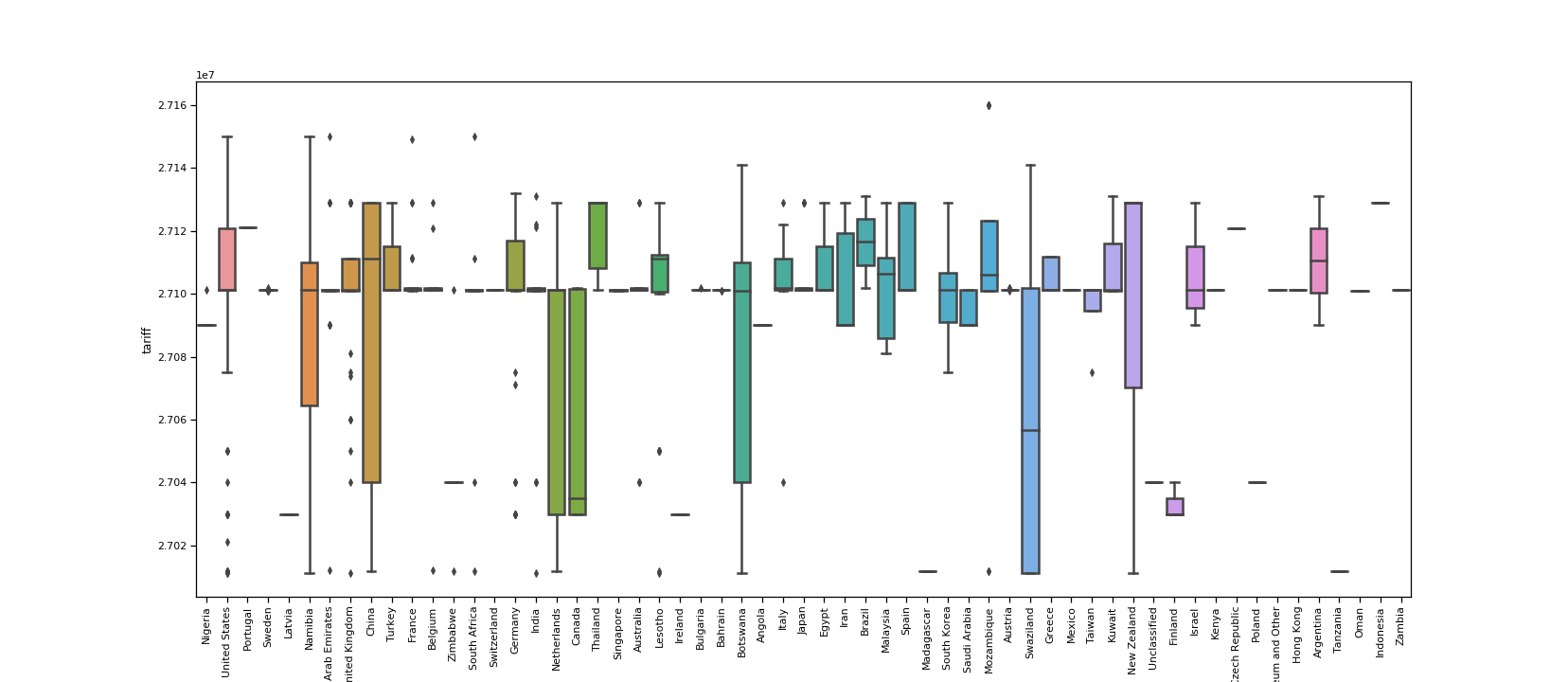
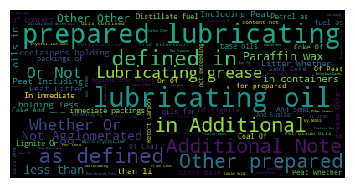


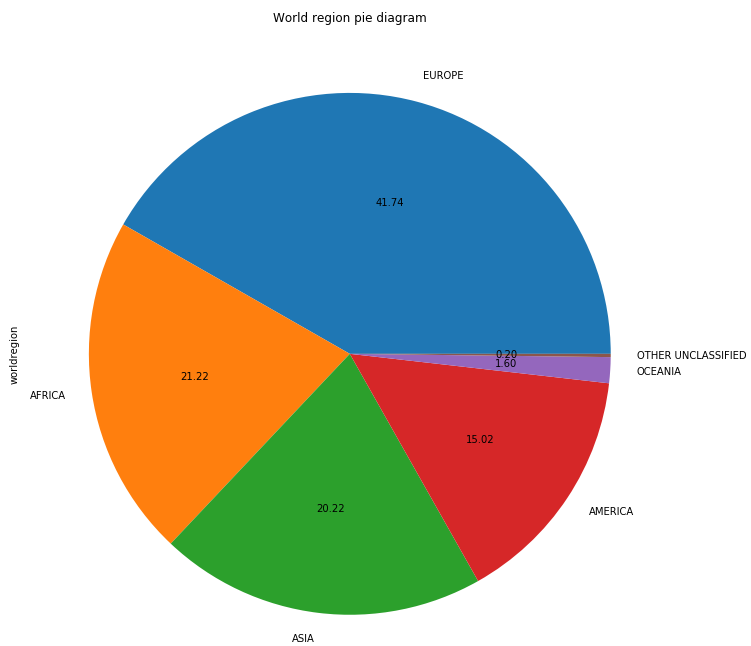
Figure 3: Tariff rates per country box-and-whiskers plot

By grouping the dataset by country, and then sorting the data by tariffs, I found that Indonesia, Portugal, the Czech Republic, Thailand and Brazil had the highest tariffs on fuels. There was a significant difference in tariff rates when compared to those countries which have the lowest tariffs i.e. Tanzania, Madagascar, Latvia, Ireland and Finland. These differences can be attributed to factors like interest rates, political stability and terms of trade of the country of origin. (Figure 3)

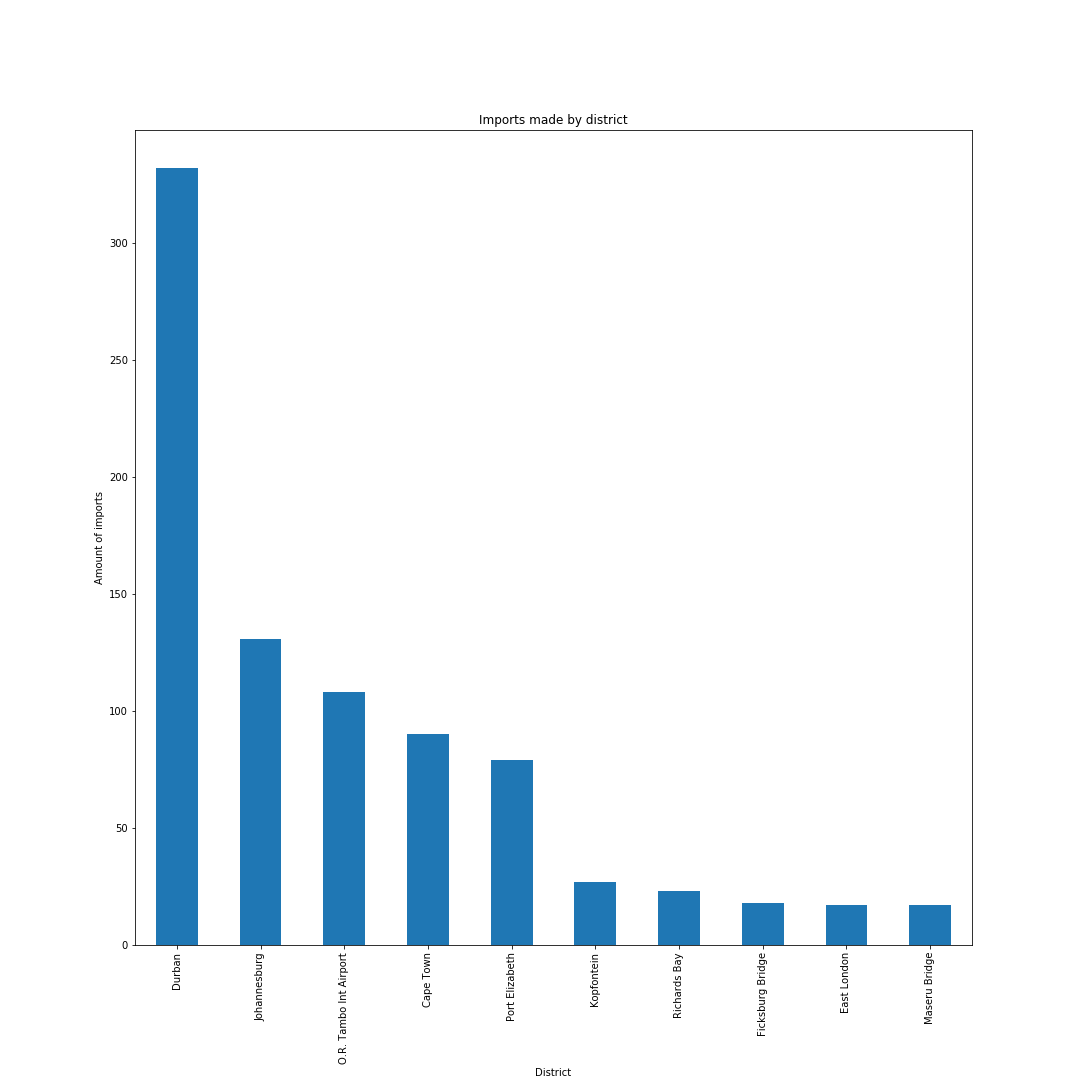
Figure 4: Frequency of fuels word cloud

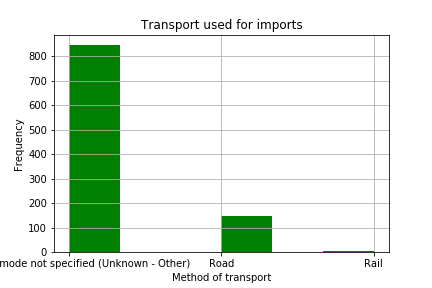
The most commonly imported fuels are lubricating fuels, and fuels that are used for lubrication in some form. Other commonly imported fuels include paraffin wax and peat. Of these fuels, the ones with the highest tariffs are those used in the production of electrical energy and emulsions. The fuels with the lowest tariffs include anthracite, bituminous coal, other coals, briquettes and lignite. (Figure 4)

In order to determine the total cost of an import, I added the tariffs to customs value. Once this was done, I was able to find the most expensive fuels in our data set. I found that petroleum oils are the most expensive fuels, regardless of the origin.

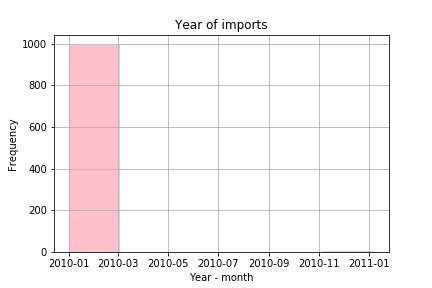
Figure 5: Number of imports per region pie chart

I found that 41.7% of the fuel is imported from Europe. This is almost double the amount of imports compared to the second and third largest importers, which are from Africa (21.22%) and Asia (20.22%). (Figure 5)

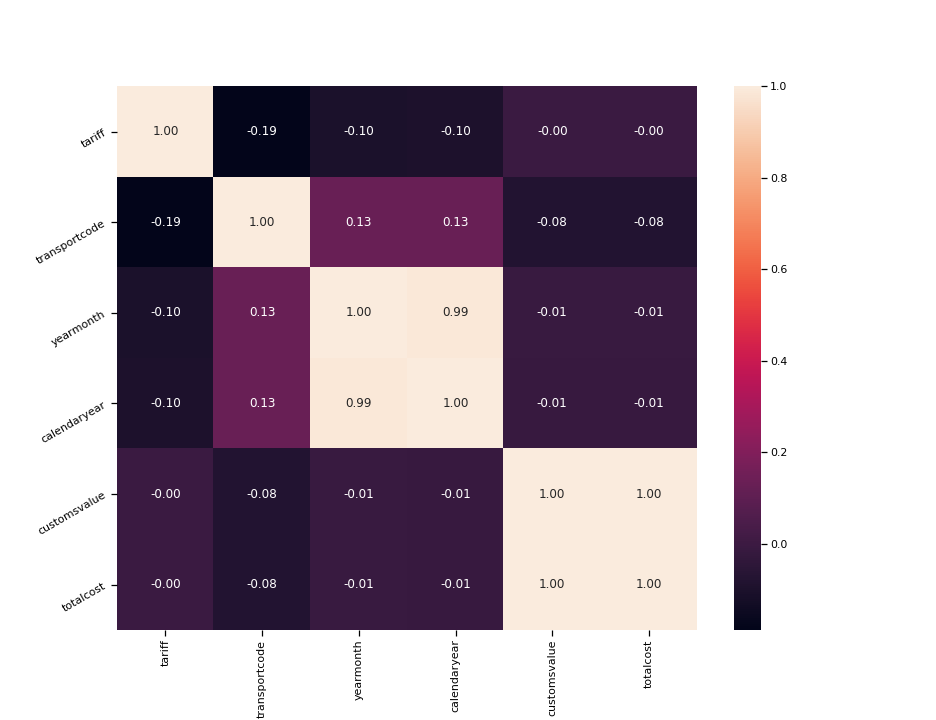
Figure 6: District of imports bar graph

Figure 7: Transport used for imports histogram

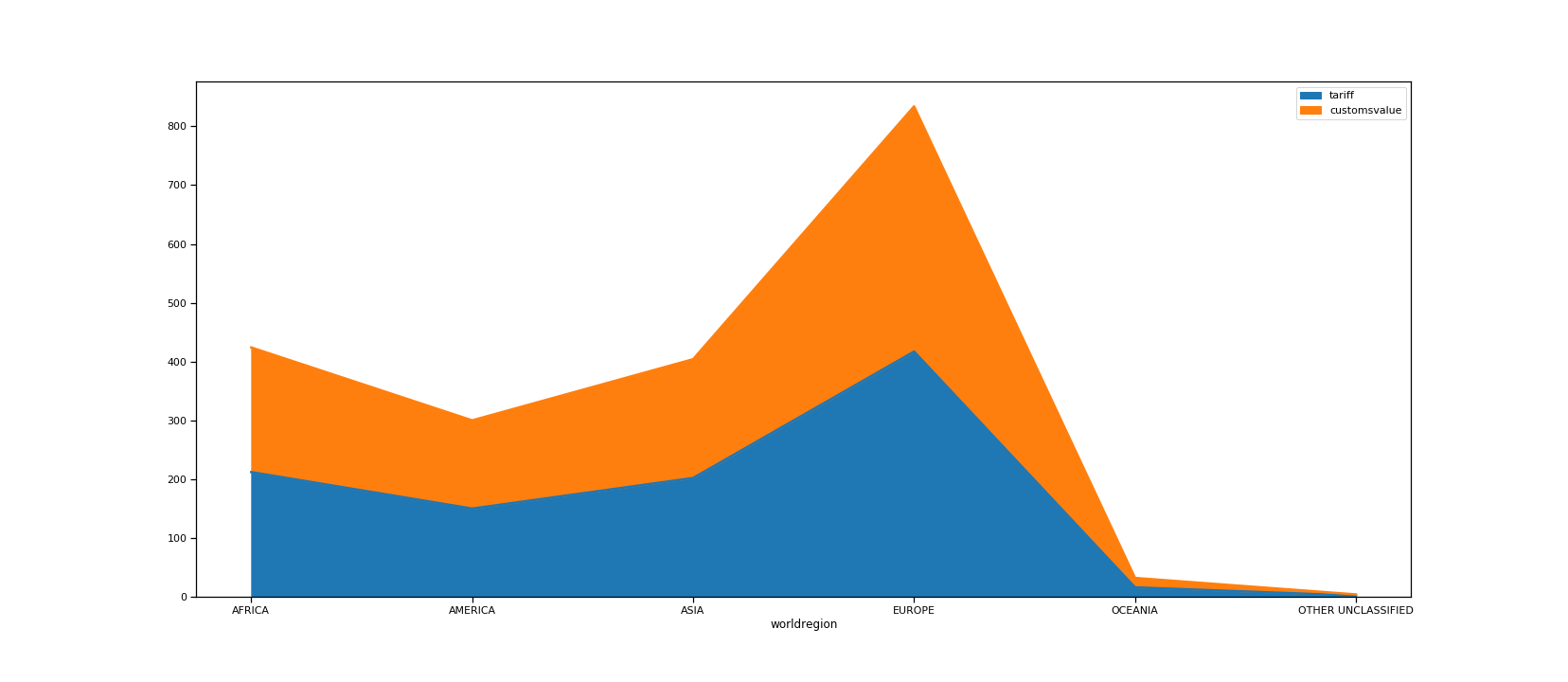
Most of imports were made to the Durban district. When comparing this to the methods used to transport fuels, we find that majority of the fuels were transported by ‘unknown’ means. Since imports into Durban are done so to various harbours around the Durban district, we can ascertain that large a portion of transport used in the ‘unknown’ variable is via freight. (Figure 6, figure 7)

Figure 8: Year-month of imports histogram

When plotting the month of the year that the imports were made, I found that almost all imports were made during the first three months of the calendar year (January 2010 – March 2010). No other imports were then made until November 2010 to January of the following year. It is worth noting that the imports made during this period (November – January) were exceptionally low. (Figure 8)

Figure 9: Numerical data heatmap

The heatmap shows how the different numerical features in our dataset relate to each other. The colour gradient easily shows the strongest and weakest relations within our dataset. We find that tariff rates will be influenced by the transport code. However, the transport code will not have as much as an influence on customs value. The time of the year (yearmonth) will also have a slight influence on tariff rates. (Figure 9)

Figure 10: Tariff/ customs value area graph

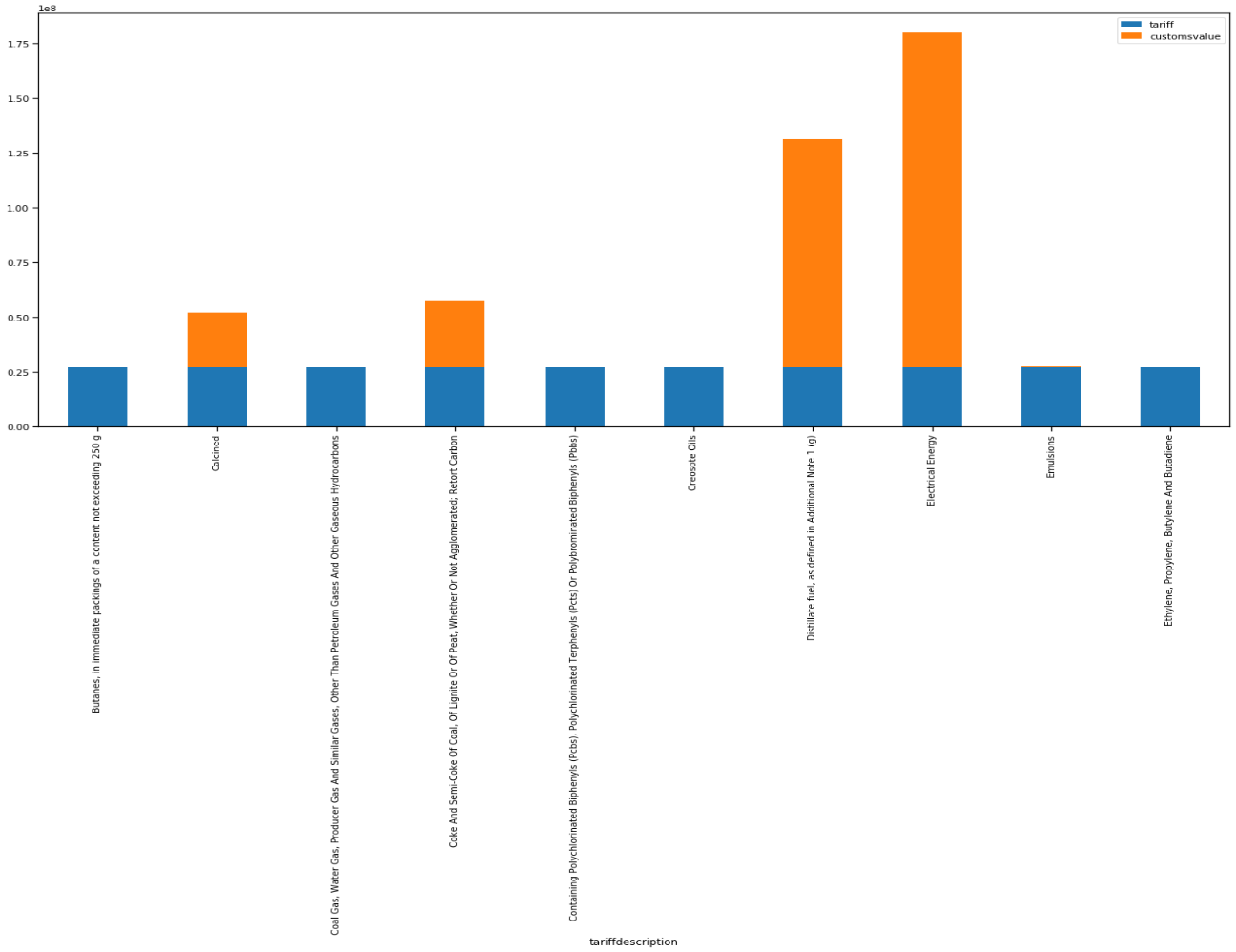


Figure 11: Tariff/ customs value per fuel bar graph

Tariff and customs values tend to be directly proportional, regardless of where the fuel has originated from. However, when observing how tariff and customs value relate to the types of fuel, I found that tariff rates stay constant throughout, whereas customs values vary notably between fuel types. This once again shows that fuel used in the production of electrical energy is much higher than other fuels. (Figure 10, figure 11)

**THIS REPORT WAS WRITTEN BY : Curtis van Wyk**

