

**TASK**

**Exploratory Data Analysis on the Fuel Imports of South Africa Data Set**

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

**Introduction**

The data set that I have chosen to analyze, consists of all the fuel imports to South Africa between the period 2010 to 2011. Within the dataset, there contained descriptive values showing: From which continent South Africa received its imports; From which countries South Africa received its imports; The year and month that South Africa received imports; The quantity in KG that South Africa received; The Price South Africa paid for its imports; The mode of transport by which these imports were delivered; as well as the region within South Africa that Imports were delivered.

The dataset can be found [here](https://data.openup.org.za/dataset/fuel-imports-5mk8-wndg)

**DATA CLEANING**

Before we can analyze the data set; we must first establish all descriptive values within the set. We must also determine if there are any values that need to be formatted to easily work with Python. We need to ensure the data does not contain any missing values. We need to determine the volume of unique values in case we have multiple descriptive variables contributing to the same value. We also need to determine if there are any columns within the set that do not contribute in a meaningful way to our analysis.

#### **Column Descriptions**

Based on the initial reading of the data set, we can immediately see there are columns that are either redundant, or that add no meaningful value to our analysis.

We know that we are analyzing Imports of Fuel; therefore, the columns 'TradeType', 'Country\_of\_Destination', 'Country\_of\_Destination\_Name' are meaningless to us. We can however, use the 'District\_Office\_Name' column to show where in South Africa the imports are being delivered to.

We have also been provided with dates for each import. There are 2 columns; One to give us a value on the year and month; the other gives us a value on only the year. This will be useful to us as we can analyze how imports fluctuate on a monthly basis as well as how our total imports compare for the 2 years.

#### **Redundant Columns**

'districtofficecode' and 'districtofficename' provide the same descriptive data. We know CTN represent the city Cape Town and DBN represents the city Durban. To make it easier for users of our analysis, I will be dropping the 'districtofficecode' column.

'Country\_Of\_Origin' and 'Country\_Of\_Origin\_Name' provide the same details as above. From an ease of access perspective, we are not interested in the code of the country, but more so the country's full name. I will be dropping the 'Country\_Of\_Origin' column.

We have multiple columns that give us an ID number (transportcode, chapter), followed by the description of that ID (transportcodedescription, chapterdescription). When we look at our data set we find that the ID description contains the ID as well. This allows us to to drop any columns containing only an ID number

### **Formatting Data**

.

We have also been given the quantity of fuel imports. It is currently measured in KG; however, due to the large volume recorded, I believe that it would be better to convert this value into Tonnes.

### **Unique Value Identifiers**

I will be going through each column to check whether each variable listed is unique and identifiable. If values are extremely similar, I may replace them to match one another.

We need to check the values listed under the 'districtofficename' column. We need to be able to determine that each entry is distinct from one another and identifiable.

Based on our assessment, we can see that there are descriptive values that are identifiable, yet they are not unique. Cape Town and Cape Town Int Airport are found in the same region.

We do know that the methods of transportation are either by rail, road, or an unknown. We therefore do not need to have an airport as a separate unique variable and can rename the variable to match with the province in which its situated. This still allows us to easily ascertain which provinces received what quantity of imports.

The objective of our analysis is to visually represent Imports of Fuel to South Africa. Currently in our data set we have 2 values of imports from unknown regions as well as 11 values of Imports from South Africa itself. I do not feel as though these values accurately represent true Imports to South Africa, nor do I believe that they contribute in any significant way for our analysis. I am of the opinion that I would drop these specific rows as they only factor 1.3% of the total dataset.

We see that Each Fuel Type is preceded by a ID number. For our analysis, The ID number bears no meaning to a general users and the ID number + description would be far too long to neatly place on a graph. My goal is to remove the ID number so that we are just left with the description of the Type of Fuel Imported to South Africa.

I am going to go through each description of Fuel Type and replace it by either 1 or 2 words that will still accurately describe the type of fuel we have imported, but also make it easier to be plotted on an axis, because a value, such as: 'Containing Polychlorinated Biphenyls (Pcbs), Polychlorinated Terphenyls (Pcts) Or Polybrominated Biphenyls (Pbbs)', is far too long and unnecessary to be placed on an axis.

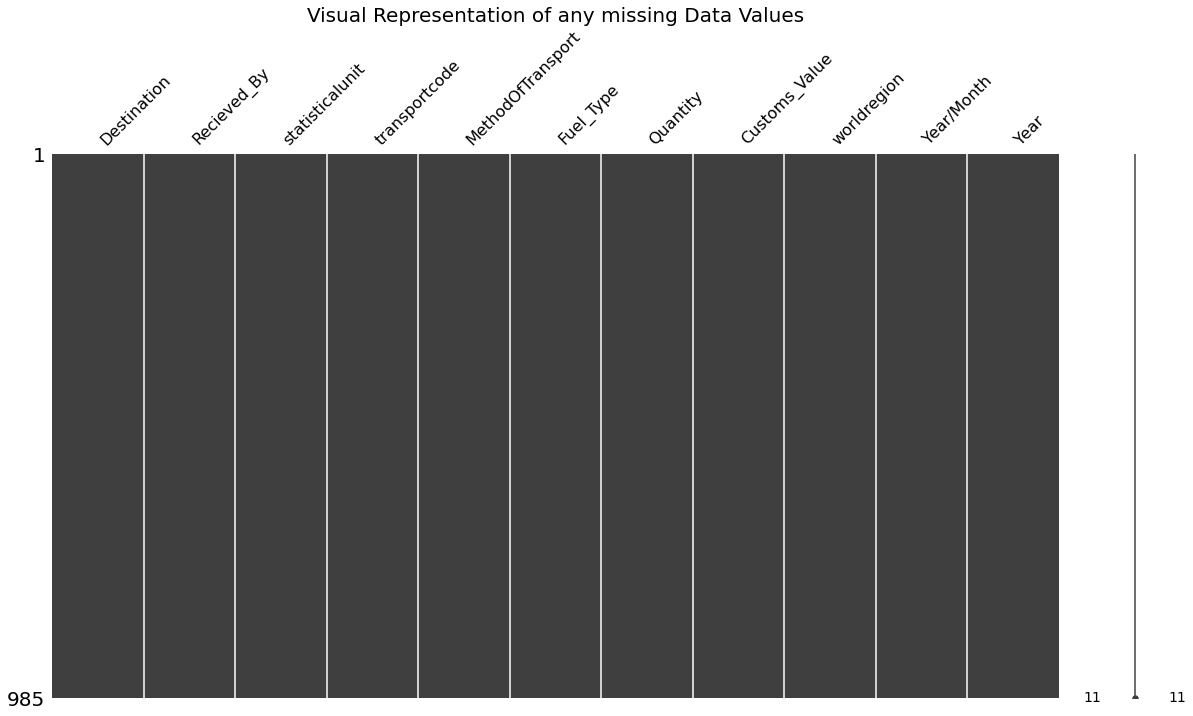
MISSING DATA

Our initial dataset consisted of 21 columns with 1000 entries.

After cleaning and preparing the data to be worked on, we were left with a new dataset that consisted of 11 columns and 985 entries.

We then need to test for any missing values in our new data set. In order to do this, we make use of the missingno library. What this Library does, is that it creates a visual to depict any missing values.

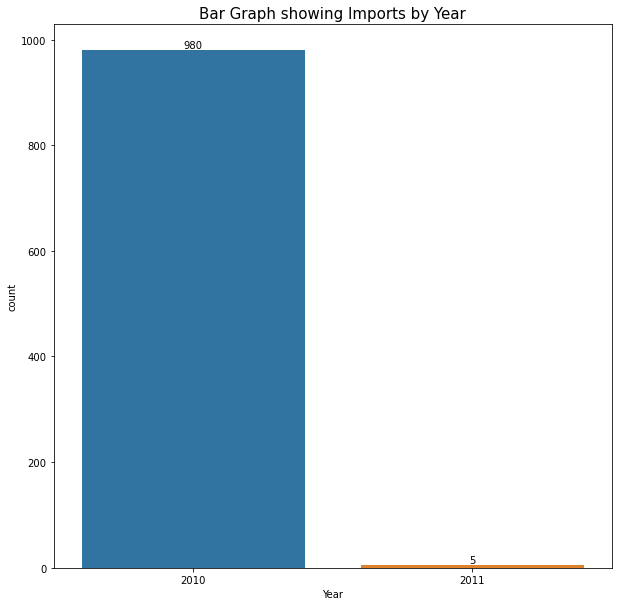
Below is an import of that image I created to show whether or not our new dataset contains any missing values.



What this image shows us, is that currently the cleaned data set I will be using to perform my analysis, contains no missing values.

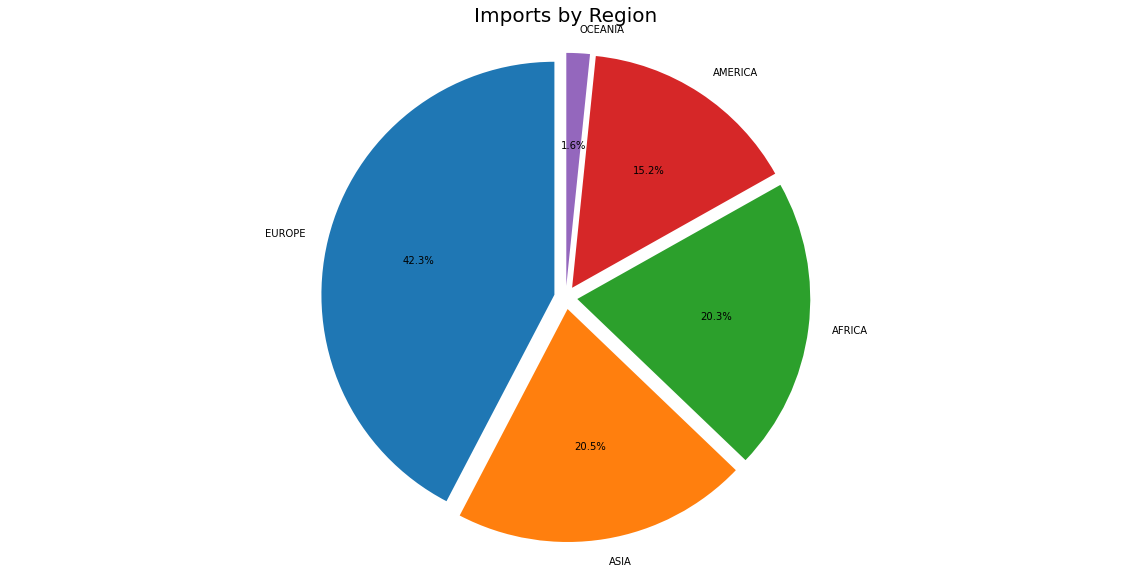
DATA STORIES AND VISUALIZATIONS

The objective for my analysis will be to visually represent:

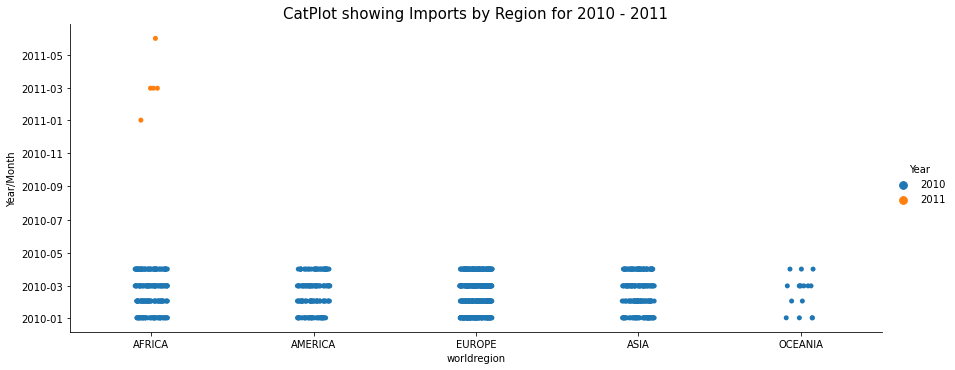
1. which countries supply South Africa with what kind of fuel and in what quantities;
2. The average price South Africa pays for its imports;
3. The main type of fuel South Africa imports;
4. To determine a correlation between total count of imports and total quantity per import;
5. which regions of South Africa receive the greatest amount of deliveries.

Based on the Graph as shown above; Majority of Imports that we have available to us in the current Dataset, occur in the year 2010. What this means, is that, we can not do a comparison of imports by year in terms of quantity or region procured. The graph also indicates to us that we may not pay heed to the year 2011, as we have only 5 values to account for in that period.

### **Graphs Displaying Information Regarding South Africa**

****

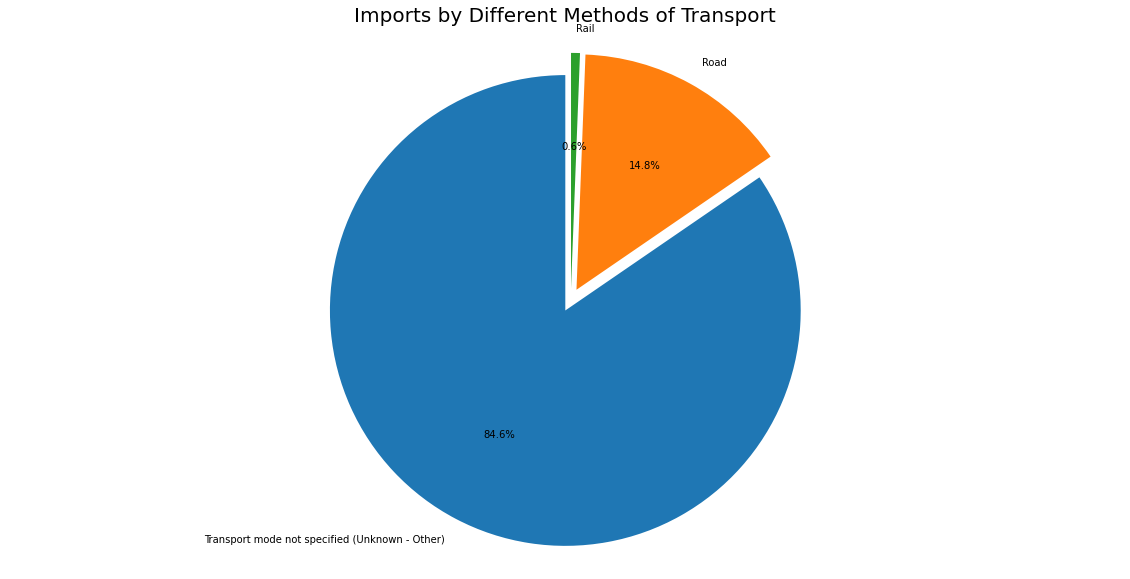
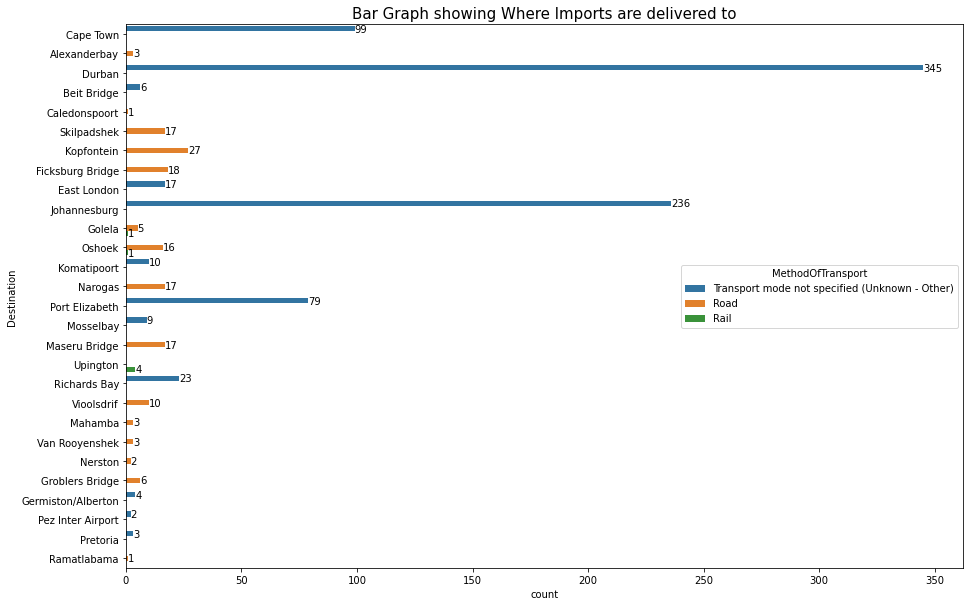
Here we can show the number of Imports made to South Africa by continent. This gives us an indication by means of who South Africa is most reliant on for their Imports.Of the 985 Imports made, we see that 42.3% of those came from Europe. Asia and Africa both have very similar import counts with them making up 20.5% and 20.3% respectively. We do not Import a large amount from the Americas or the Oceania Islands.



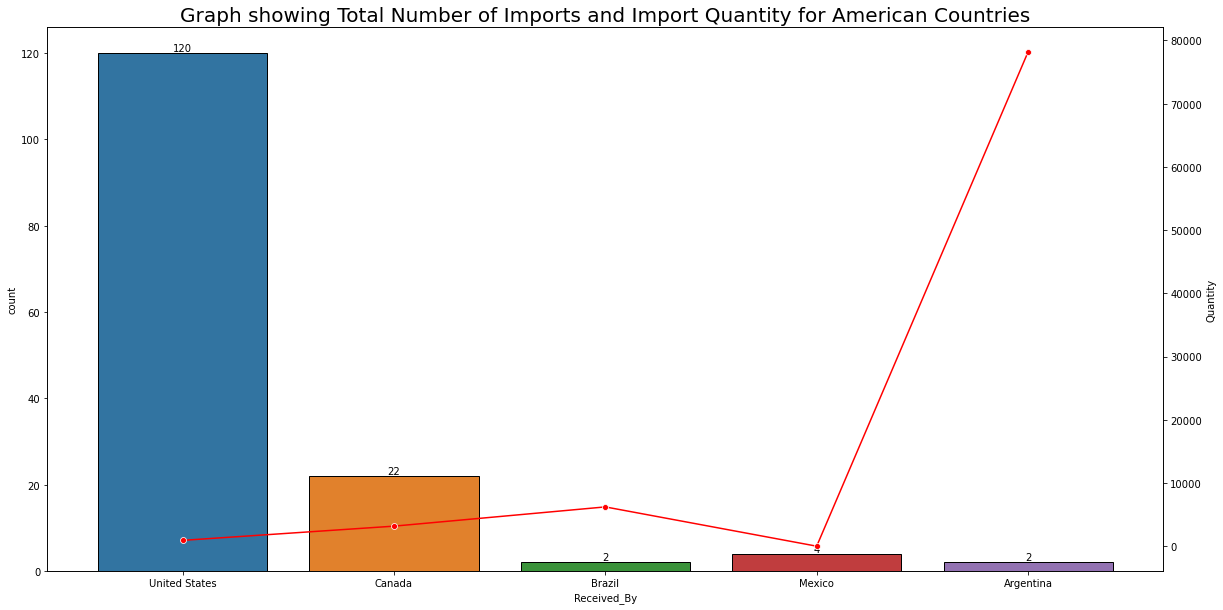
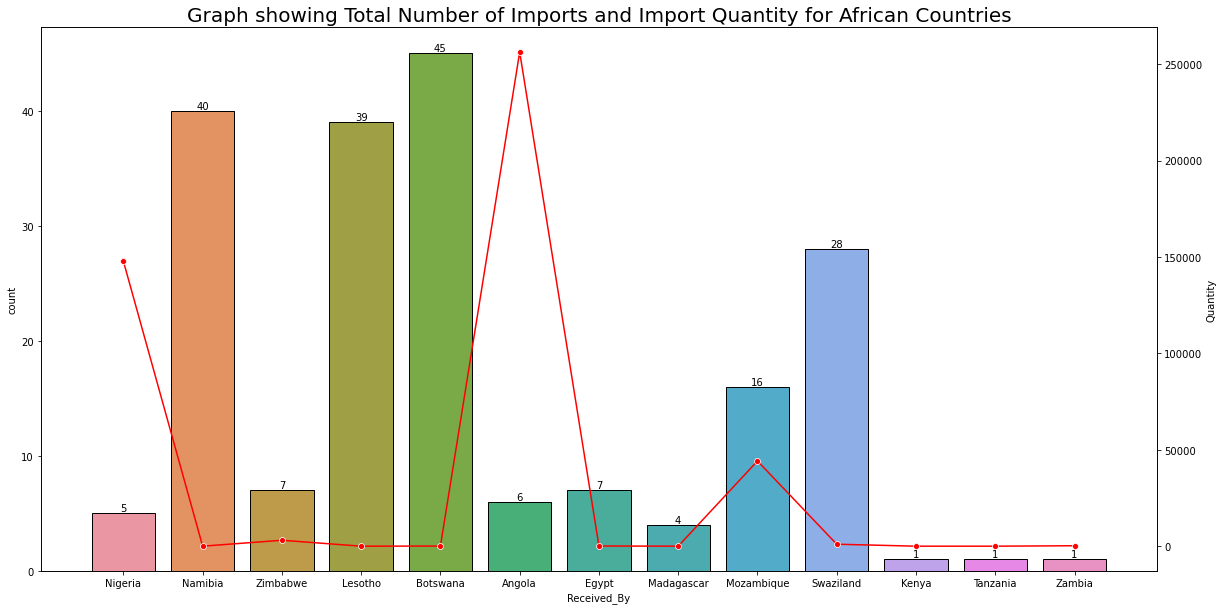
The CatPlot tells us from which region we have attained imports from and during which month of the year. Majority of our imports of fuel have been recorded in the period 1/January - 1/April; occurring each month.

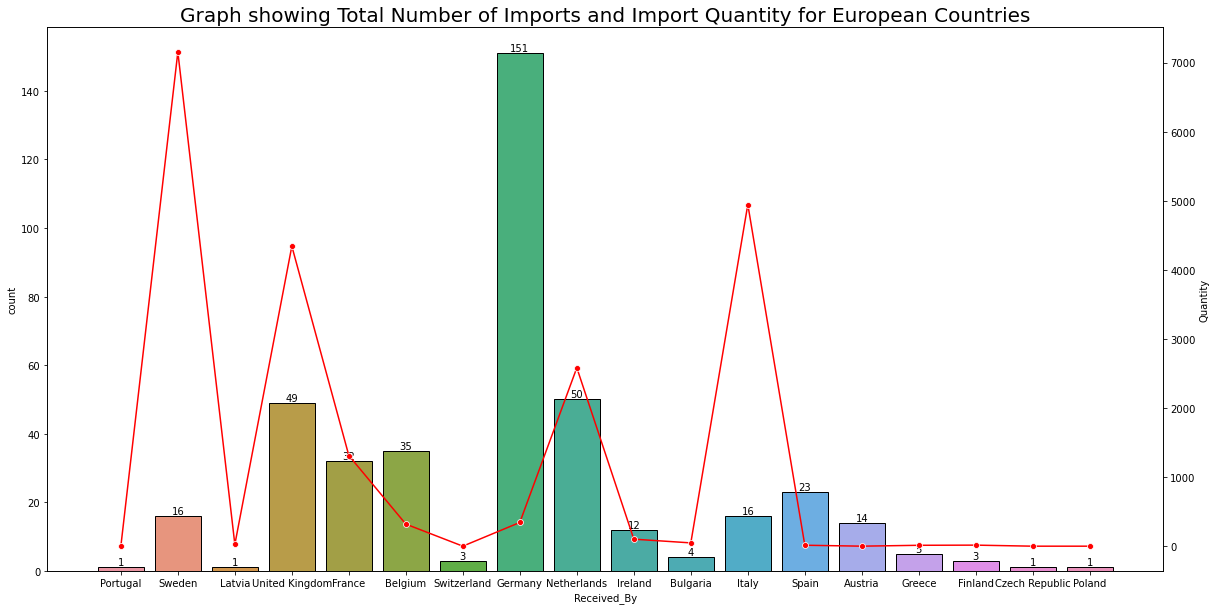
What is interesting; although we only have 5 values for the year 2011; we can see that those imports only occurred from African countries and they do not match the trend found in the year 2010.

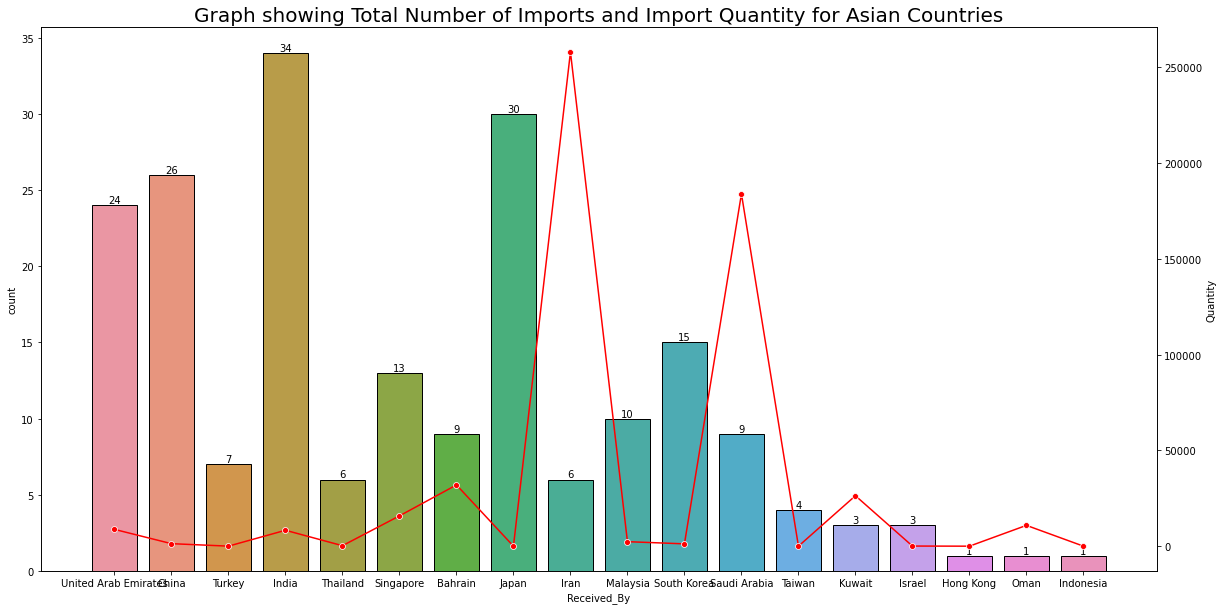
Instead of following the Imports at the start of each month for the first 4 months, we see we have imports on the 1st month, the 3rd month and then on the 6th month. Now we do not have sufficient data to produce a trend, but it is something interesting to look at.

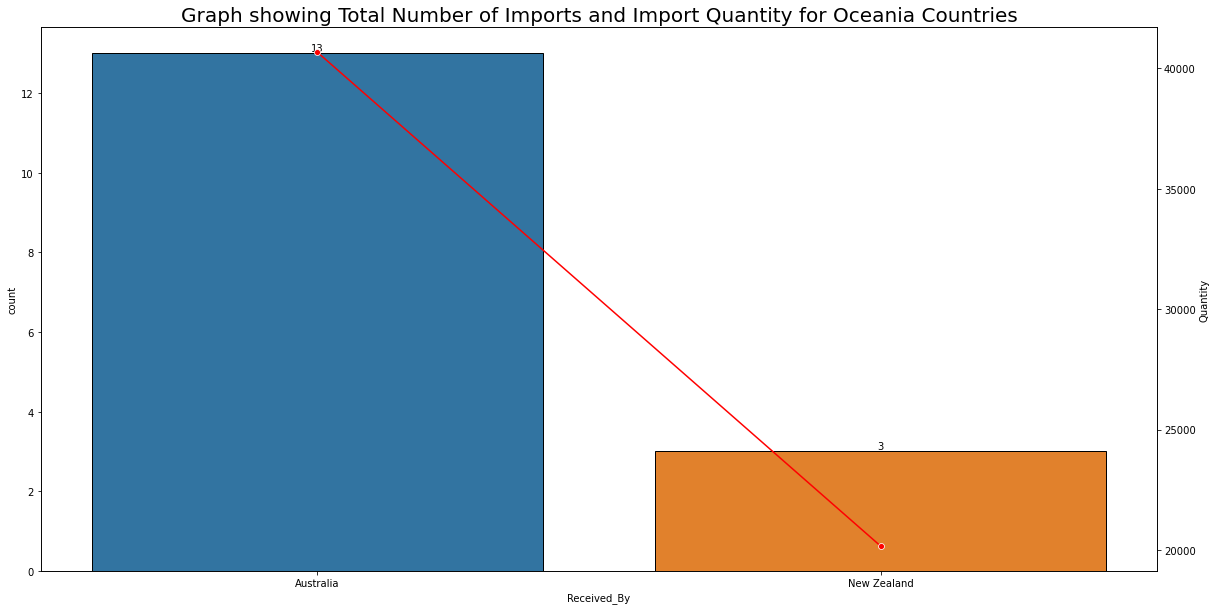
The graphs above gives us an indication as to where Imports are being delivered to within South Africa and the mode of transportation. We see that Majority of Imports are received by Durban area with most not having a specified mode of transport. We can however, assume that unclassified imports were made by sea or air as Durban is a coastal city with a port. The same can be said for Cape Town area. However. for Johannesburg we can only assume that the unclassified mode of transport was done by air.

### **Graphs Displaying Volume and Quantity of Imports to South Africa**





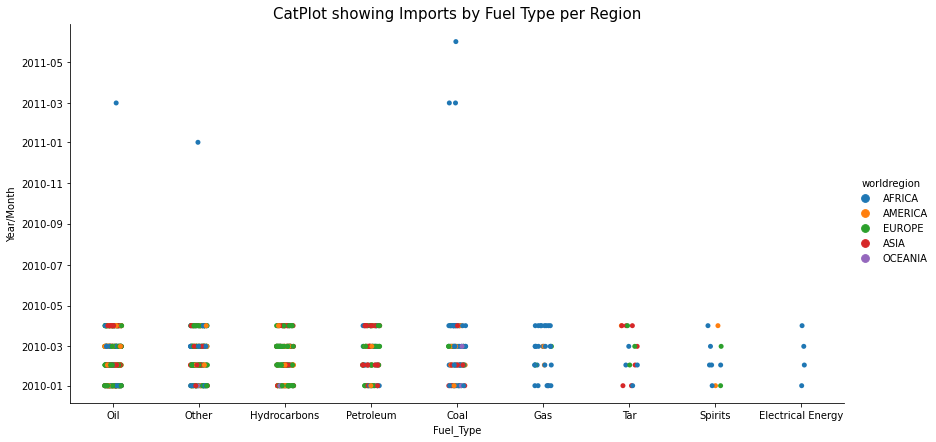




The underlying Objective for these 5 graphs is to visually represent the countries that South Africa Imports from; the total number of imports that South Africa has made; the total quantity of imports made by South Africa from each country and finally to determine is there is any correlation between number of imports and total quantity by each country.

The bar graphs display the total number of imports that South Africa has made from each respective country and each graph is accompanied by a line plot displaying the total quantity.

What we find from these graphs, is that there is no correlation between number of imports and quantity of imports. Looking at the Americas; we see that we imported a total of 120 units from the United States, however the quantity only amounted to 919 tonnes. Whereas we only imported a total of 2 times from Argentina, yet the total quantity amounted to 78 000 tonnes. The same trend is found when looking at Europe. South Africa only imported 15 times from Sweden, yet the total quantity of imports amounts to 7 000 tonnes, while Germany made 151 imports to South Africa yet Quantity only amounts to 300 tonnes.



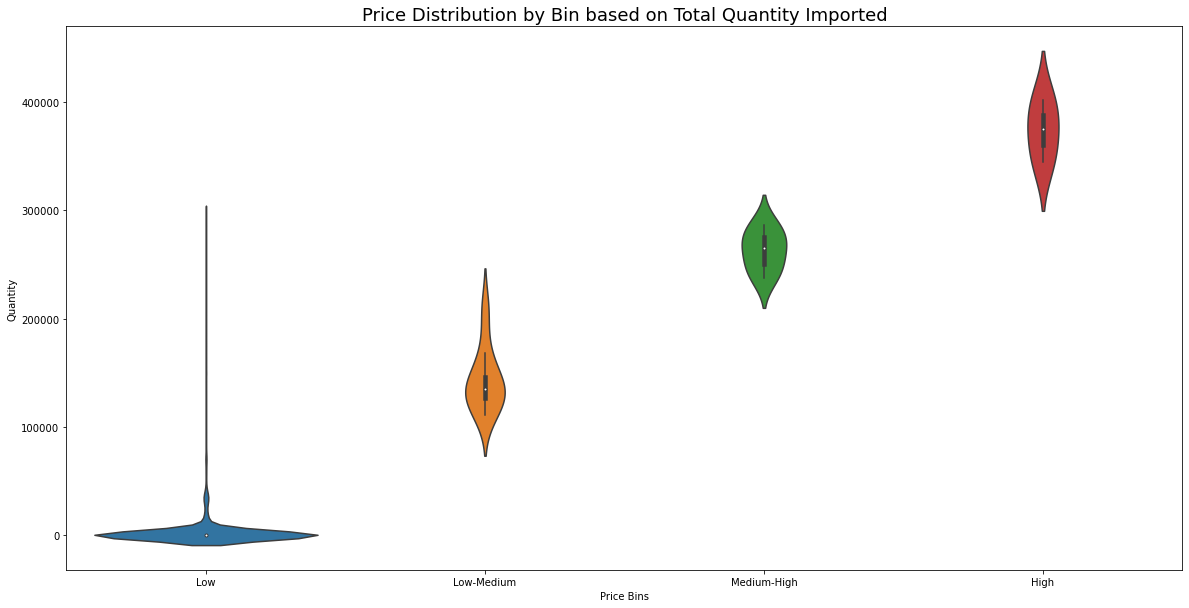
From the graph above; we can get a representation of the period when South Africa imported certain Fuel Types and from which region those imports occurred.

The reason why I have included the year 2011 is that, even though we only have 5 values; what is interesting to note is that, these values do not occur in the same time period, nor do they only form 1 fuel type; yet, we do see that these 5 values all came from the African continent.

Not really important as the data is in no way indicative of baring any meaning, yet one may always ask: ‘of those values we do have, why are they only originating from Africa?’

**Price Analysis**

When analyzing the price paid for imports, we see that our Price range is quite large. We may want to use binning to categorise our prices into low, medium and high. Even though the data may include a value that attributes a low cost to a Fuel Type in one entry and a high cost to the same fuel type in another entry; this just ensures a greater accuracy for our predictive model.

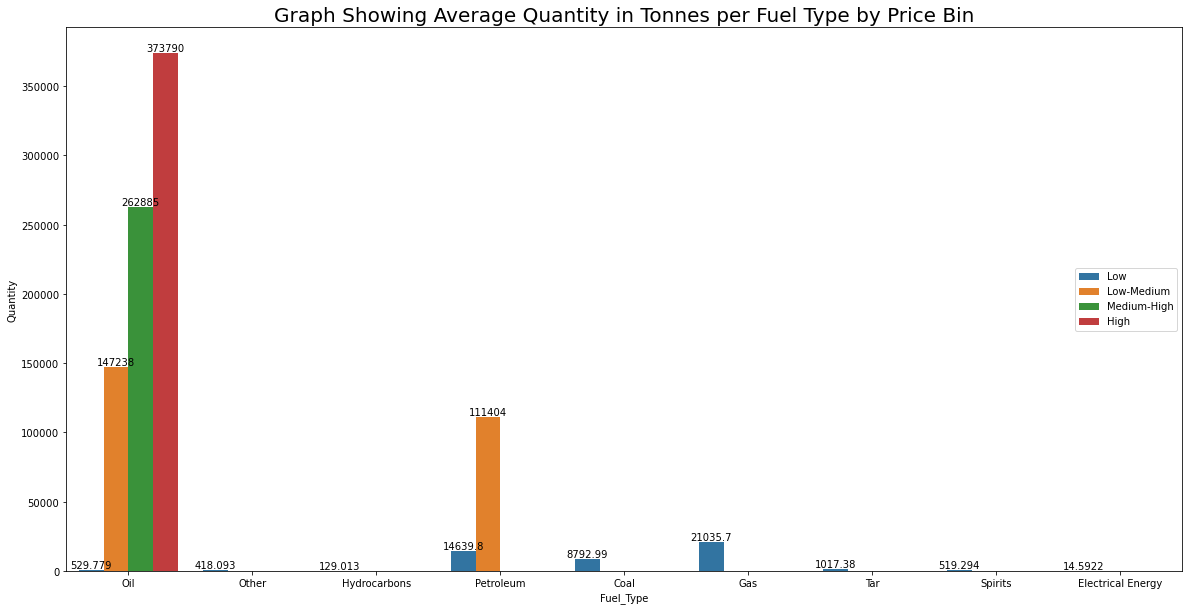


This is a Violin Plot graph to represent the distribution of Price per quantity bought in the respective price bins.

What is interesting from this observation, is that we have many overlapping quantities bought in different price ranges.

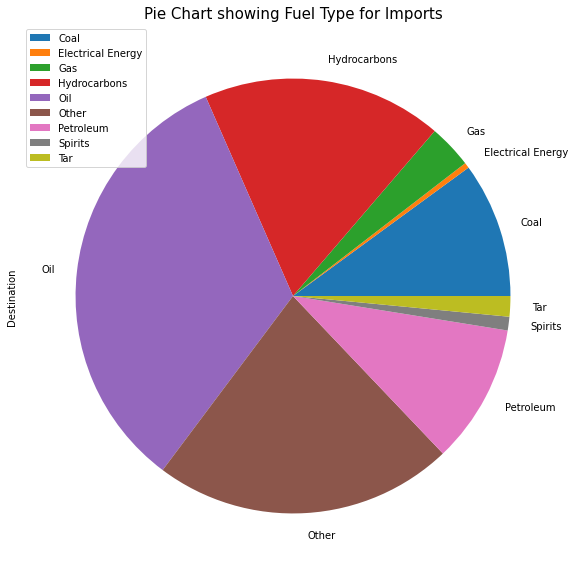
We can infer from this data that there were instances where South Africa purchased a large quantity, yet paid a low amount for it.

However, one thing that is clear from this graph, is that there is a trend that correlates total quantity purchased to its relative price bins.

****

What the graph above represents, is an average Quantity per price bin for each Fuel Type that South Africa imports. We see that we have 4 price bins. Each is allocated to the 9 different types of Fuel we import. Currently we observe that only the Upper 2 price bins (Medium-High and High) are allocated to Oil. Based on this, we have an understanding that oil is in fact a costly import for South Africa. Petroleum is second, yet only contains price bins for Low and Low-Medium. Every other Fuel type contains its quantities in the Low price bin.

Now this graph does not display the price or the average price that South Africa paid, it merely conveys, within each price bin, what the average quantity purchased is.

****

Lastly, I wanted to show the distribution of Fuel Types that South Africa imports. Majority of our imports consist of Oil, Petroleum and Hydrocarbons.

Other is defined as consisting of Polychlorinated Biphenyls and Terphenyls.

What is interesting from this Pie Chart is that South Africa imports Coal products, even though we are a Country that Exports Coal.

**Summary of Analysis**

The objective was to show:

* The Regions and countries that supply South Africa with its imports
* the Quantity of South African imports
* The Cost to South Africa for its imports
* The regions within South Africa that receive imports
* Whether there was any correlation between number of imports and total quantity

Based on my observations, we can conclude that South Africa is an active participant on a global scale. We have received imports from every continent for a range of different fuels.

We have observed that there is no concrete evidence to suggest that number of imports equates to a higher quantity.

We see that our main regions of delivery are Durban, Cape Town and Johannesburg. Although majority of the data contained unclassified modes of transport for delivery, it is easy to infer that if it is not listed as road or rail, then for coastal areas that contain a port and an airport; the mode of transport falls under either by air or by sea. For regions inland that we can assume the unclassified delivery was made by air.

We see that South Africa imports mainly Oil and thus incurs a large cost for this import.

A key point to note is that Hydrocarbons make up the Third largest contributor to imports to South Africa, however, based on the average quantity to price bin, we see that it is a very low contributor. Which means that we do not pay a high fee for the importation of Hydrocarbons.

**THIS REPORT WAS WRITTEN BY:**

**Christopher Myburgh**

