Hyperiondev

TASK

Exploratory Data Analysis on the Chocolate Imports to South Africa from 2010 to 2015

Visit our website

Introduction

The chocolate imports to South Africa dataset contains the custom values and statistical quantity of the chocolate imported from different countries in the years 2010 to 2015. The data set is available from the OpenUp data portal. OpenUp has shortened the dataset to 1000 rows, the complete dataset can be obtained from the South African Revenue Services (SARS) website, therefore this report is conducted purely to analyse the data and not to make any predictions.

The dataset consists of the name and code of the exporting country, the South African receiving port, the statistical unit and quantity of the chocolate, the customs value of the chocolate and the year and month in which the import occurred. Wikipedia defines customs valuation as a process whereby the goods or service being imported are assigned a monetary value.

I have used the Python programming language and number of python libraries such as pandas, NumPy to handle arrays and matrices and Matplotlib and Seaborn to create plots. The code file is created in Jupyter notebook.

The purpose of this report is to explore the change in chocolate imports to South Africa over the period of 2010 to 2015. To explore the proportion of countries and regions that import chocolate to South Africa and how that changed over the same period. What was the most used mode of transport for chocolate imports to South Africa? What are the top importing countries of chocolate to SA?

DATA CLEANING

The dataset is in csv format, to read in the data use the *read_csv()* function. To view a snapshot of the DataFrame I used the head() method, the method returns the first 5 rows of the DataFrame.

The original data contains the following features

- Tradetype – All the records are stored as Imports.
- Districtofficecode - The abbreviated name for the district office receiving the imported chocolate.
- Districtofficename - The name of the district office receiving the imported goods.
- Countryoforigin- The exporting country code
- Countryoforiginname - The exporting country geographical name
- Countryofdestination - The South African country code.
- Countryofdestinationname - South Africa is the final destination for the imported goods.
- Tariff - The tariff code for chocolate goods.
- Statisticalunit - Chocolate imports are measured in kilograms
- Transportcode - The numerical code for the mode of transport
- Transportcodedescription - Description for the mode of transport; 0 – unknown, 1 – Road, 3 – Sea/Maritime, 4 - Air
- Yearmonth - The year and month in which the import occurs
- Calendaryear - The year in which the import occurs
- Tariffanddescription - The tariff and description of the chocolate imports
- Statisticalquantity - The chocolate quantities imported
- Customsvalue - The value of the chocolate imports
- Worldregion - The geographical continent of each exporting country.

After reading the dataset into a DataFrame, I noted that several columns represent the same information. The countryoforigin and countryoforiginname columns both contain the label of the exporting country. For the purpose of readability I will use drop() to drop the countryoforigin column. The following columns were also dropped;

- Tradetype, each record within the dataset represents an import
- Countryofdestination and countrydestinationname, we are solely looking at chocolate imports to South Africa.
- Districtofficecode, Tariff, countryoforigin, tariffanddescription are also dropped.

View the dtypes of each column using the info() function, which returns:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 17 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   tradetype                 1000 non-null   object
 1   districtofficecode        1000 non-null   object
 2   districtofficename        1000 non-null   object
 3   countryoforigin           989 non-null    object
 4   countryoforiginname       1000 non-null   object
 5   countryofdestination      1000 non-null   object
 6   countryofdestinationname  1000 non-null   object
 7   tariff                    1000 non-null   int64
 8   statisticalunit           1000 non-null   object
 9   transportcode             1000 non-null   int64
 10  transportcodedescription  1000 non-null   object
 11  yearmonth                 1000 non-null   int64
 12  calendaryear              1000 non-null   int64
 13  tariffanddescription      1000 non-null   object
 14  statisticalquantity       1000 non-null   float64
 15  customsvalue              1000 non-null   int64
 16  worldregion               1000 non-null   object
dtypes: float64(1), int64(5), object(11)
memory usage: 132.9+ KB
```

It should be noted that the data in calendaryear and yearmonth columns is stored as int64 instead of datetime data types. To convert the dtypes to datetime we compute the following line of code:

```
chocolate_df['yearmonth'] = pd.to_datetime(chocolate_df['yearmonth'],
format = '%Y%m')
```

The argument ‘format’ will return the data values in yearmonth in the format: YYYY-MM-DD, because I did not specify the format for the days in the date it will do to the default ’01’ for each record. I have extracted the year and month from yearmonth and created new columns year and month in the DataFrame.

To confirm that the only unit of measure used for the quantity of chocolate imported we pass the ‘statisticalunit’ column through the unique() function. I note that all the import quantities are measured in kilograms.

To calculate the correlation between the columns I used the corr() function. The correlation between customs value and statistical quantity is +0.95 which is a strong positive correlation. This correlation value is expected because if a high/low quantity of chocolate is imported the customs value of that chocolate will be high/low. Put in another way if a region exports a large quantity of chocolate to South Africa, the customs value of the imported chocolate will be high.

To determine the skewness of data in the DataFrame I used the skew() function. The results of which are shown below:

```
transportcode          0.528882
calendaryear           0.137343
statisticalquantity    4.080484
customsvalue           4.846044
year                   0.137343
month                 -0.015922
dtype: float64
```

I can see that customsvalue and statisticalquantity are highly skewed, each feature has a skew value of 4.

MISSING DATA

Explore the presence of any missing values within the dataset using the isna() function, the following line of code will return the total sum of missing numbers per column.

```
chocolate_df.isna().sum()
```

There are no missing values identified.

```
districtofficename          0
countryoforiginname         0
statisticalunit             0
transportcode               0
transportcodedescription    0
yearmonth                   0
calendaryear                0
statisticalquantity         0
customsvalue                0
worldregion                 0
year                        0
month                       0
dtype: int64
```

There is no need to use the missingno library to visualise the distribution of the missing values because the dataset contains no missing values. It should be noted that there are (3) records that have ‘unclassified’ as a ‘countryoforiginname’. Three out of 1000 records is not a cause for worry because each record has a valid ‘worldregion’ value assigned to it.

Let's have a closer look at the time period that the data set covers by using the function unique().

```
chocolate_df['year'].unique()
```

We obtain the following:

```
array([2010, 2011, 2013, 2012, 2014, 2015], dtype=int64)
```

DATA STORIES AND VISUALIZATIONS

Time series data is a collection of observations over time. The chocolate imports dataset contains the customs value of chocolate imports to South Africa from the year 2010 through to the year 2015.

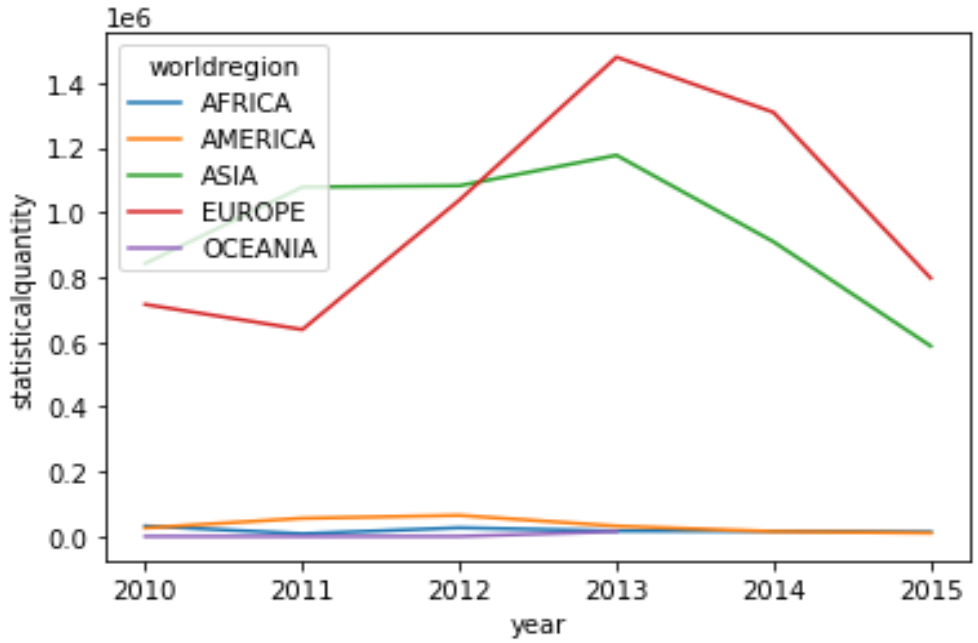
This EDA seeks to identify trends such as changes in the quantity of chocolate imports over the years per region, changes in the quantity of chocolate imports per month in a year, changes in the customs value over time, changes in customs value over regions or country of origin.

Year to Year Trends

Using the groupby() method, I group the data by year and then by worldregion. To obtain the total imports per year I use the following lines of code:

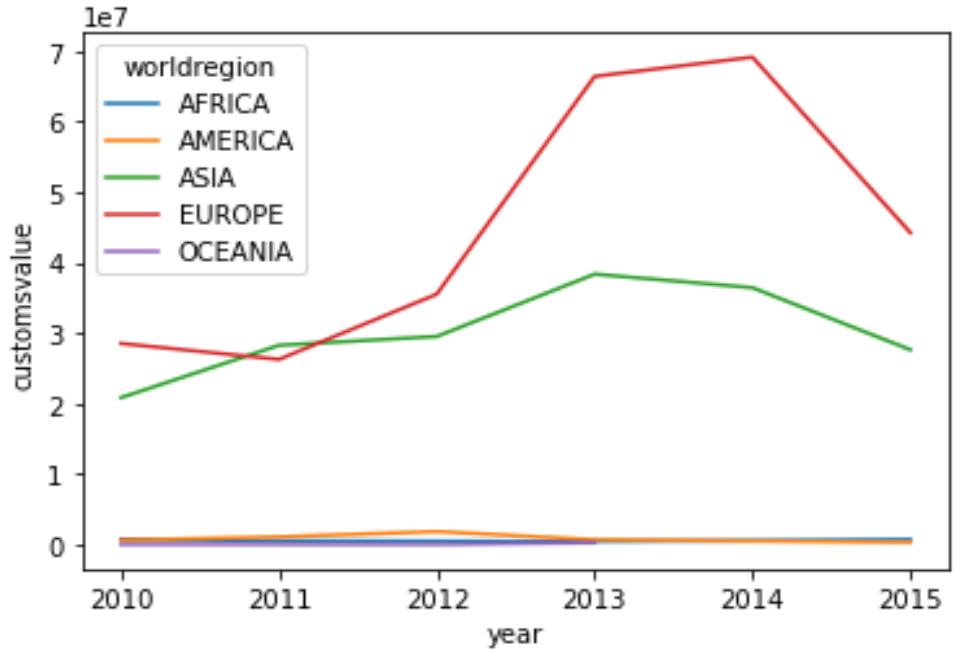
```
grouped_imports = chocolate_df.groupby(['year', 'worldregion'])
total_imports = grouped_imports.sum()
total_imports
```

The line graph below plots the statistical quantity of all the chocolate imports per region from 2010 to 2015.



From the line graph above, Asia and Europe are importing a much larger quantity throughout the time period of 2010 to 2015 than the other regions. There is a steady increase in imports from Europe from 2011 to 2013, but after the year 2013 it gradually decreases. Imports from Asia decreased are at steadier rate than Europe from 2013, this is evident from the slope of the green line above.

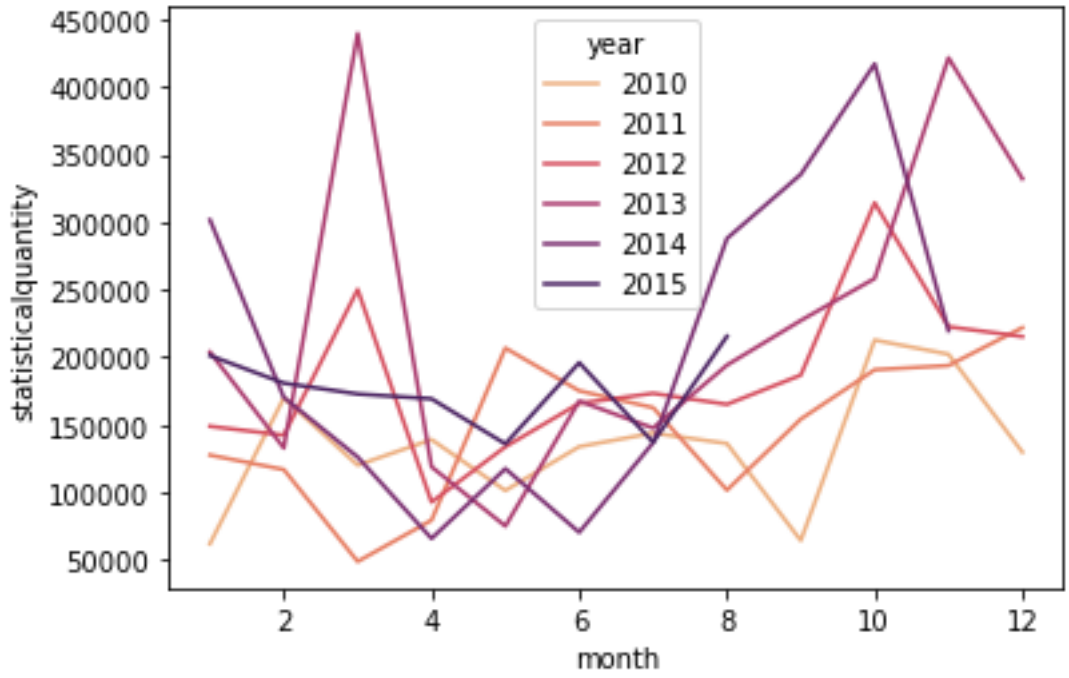
The line graph below plots the customs value of all the chocolate imports per region from 2010 to 2015.



The line graph above indicates a much higher customs value for the chocolate imported from Europe than any other region. This is expected because Europe was exporting larger quantities of chocolate to South Africa. The high customs value can be attributed to the fact that it costs more Rands in exchange for one Euro compare to one Yen.

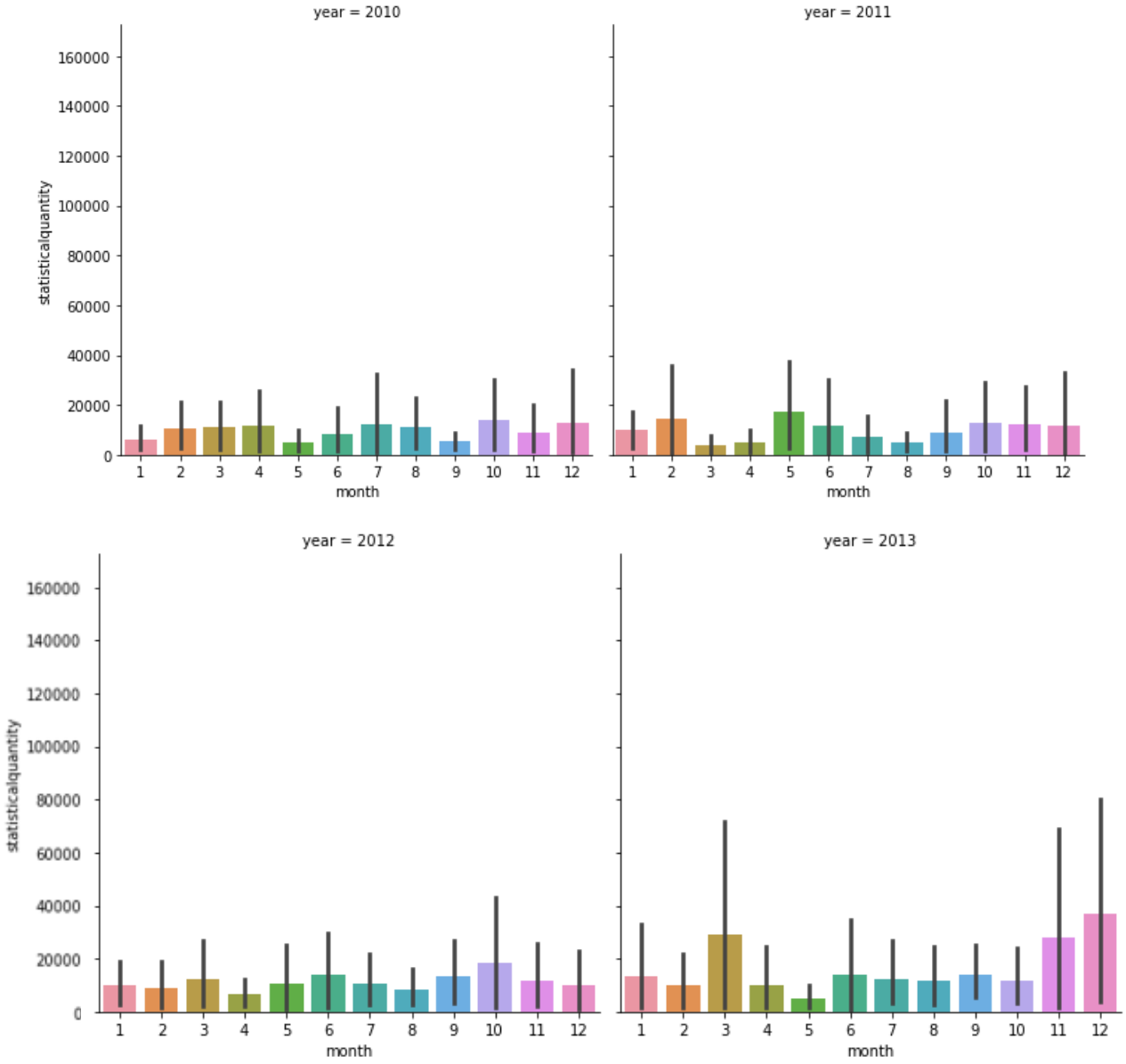
Month to Month Trends

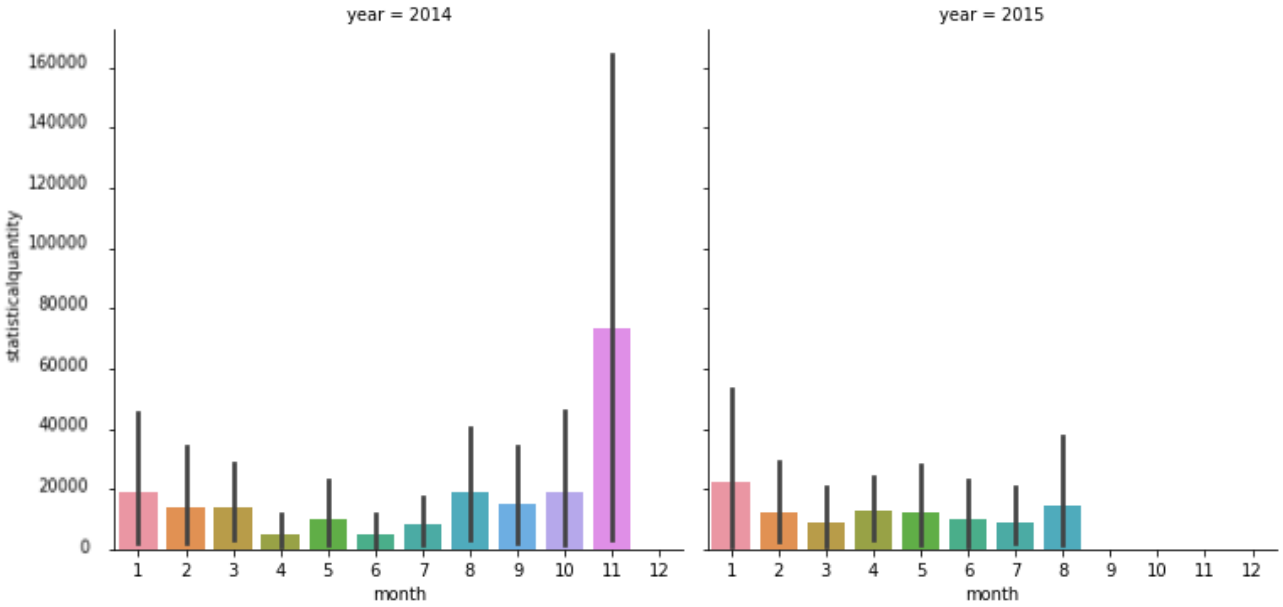
Use groupby() method to group the DataFrame by the year then the month and analyse any trends regarding the months that had the most imports. The line graph below indicates the quantity of chocolate imports per month for each year.



There are high quantities of chocolate imports between February and March after which a great dip in imports occurs. The quantity of imports steadily increases again after August and peaks in October. Once again, we can see that after 2011 there was an overall increase in chocolate imports to South Africa.

Let's have a closer look at each month of 2010 to 2015 using bar graphs.



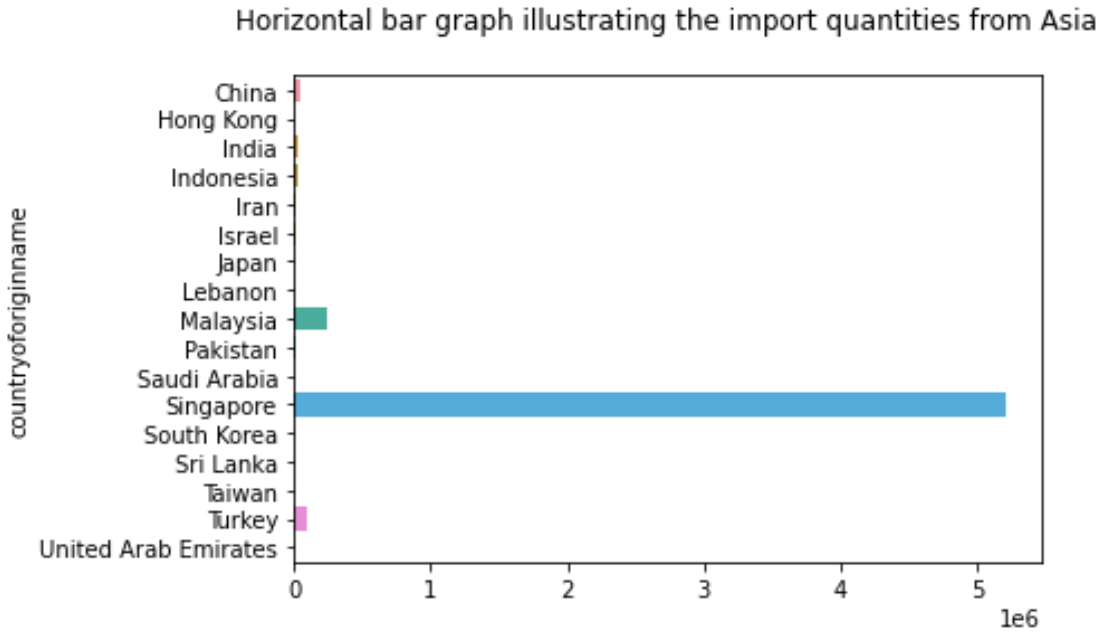


From 2010 to 2012, there seems to be a pattern where imports increase in the months of March, June and October. The quantity of imports fluctuates throughout the year but remains consistent from year to year.

I can see that the reason for high imports in the year 2013 is high imports during in March, November and December. It is odd that in 2014 there is a large quantity of chocolate imported in November and suddenly no imports in December. December is the most festive month in South Africa, this decline may be due to an incomplete dataset. For the year 2015, there is no data for the months after August. I cannot tell if imports were high towards the end of the year or not.

Highest exporting countries in Asia and Europe

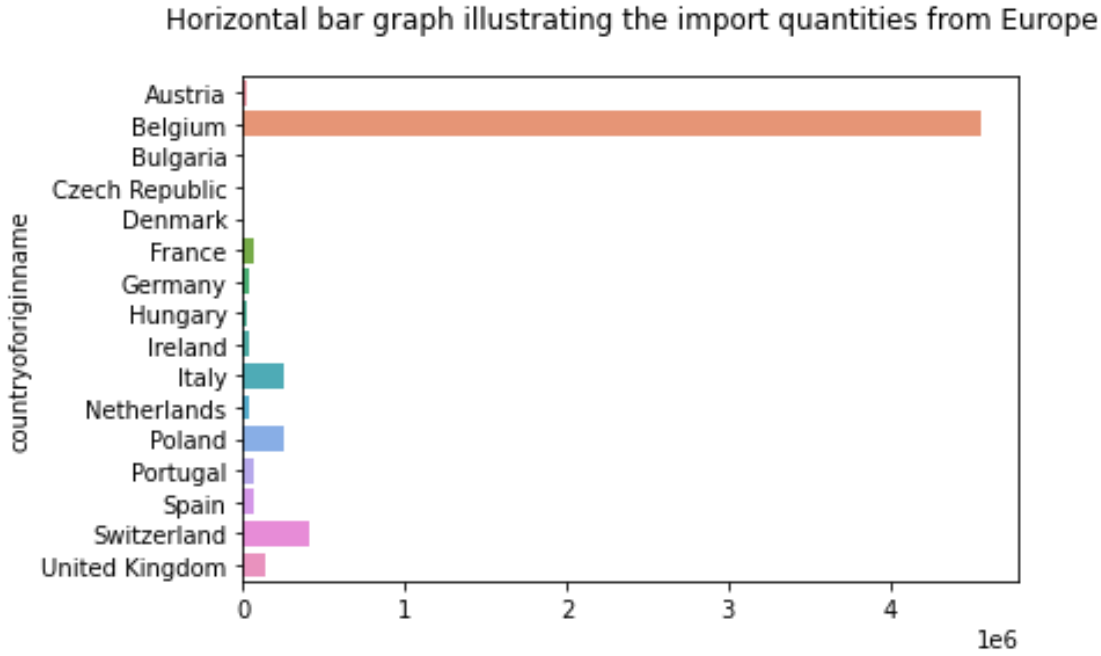
Using the groupby() method to group the data into regions and countries to obtain the quantity that is exported per country. The horizontal bar graph below illustrates the quantity of chocolate exported from Asia.



Singapore was by far exporting the largest quantity of chocolate to South Africa. The other Asian countries combined do not export even half of what Singapore was exporting.

Let's look at the European countries.

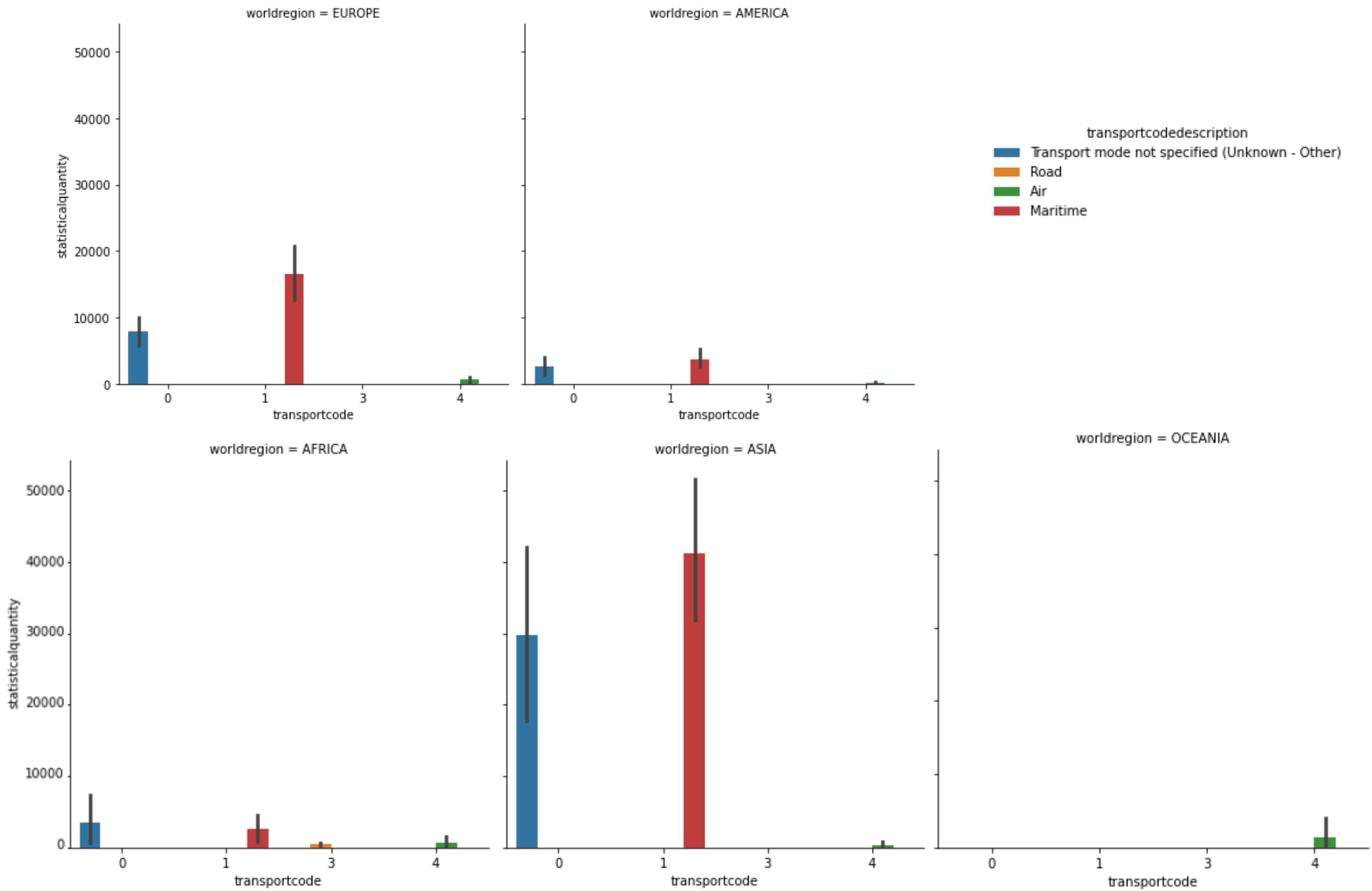
The horizontal bar graph below illustrates the quantity of chocolate exported from Europe.



Belgium is the largest exporting country of chocolate in Europe, followed by Switzerland, Italy and Poland. There are more countries in Europe than Asia that export chocolate to South Africa.

Mode of Transportation

Let's visualise the trend in the mode of transportation that each region used to export to South Africa over the years. The bar graphs below indicate the mode of transport used to transport the chocolate per region.



The graphs above indicate that the most used mode of transport is maritime. There is a large number of imports that do not have a known mode of transport. Further investigation needs to be undertaken as to why the information is missing because every item entering a country needs to be accounted for.

Conclusion

The chocolate imports dataset only contained 1000 rows which represented imports from 2010 to August 2015. Within this time period, Europe was the largest exporting region followed by Asia. The year 2013 saw an increase in imports and the most popular mode of transport was Maritime.

References

1. Chocolate Imports into South Africa - 2010 - 2017 - Datasets - OpenUp Data Portal
2. https://eriikcasstro.medium.com/eda-for-time-series-b2ea7b36c65a#:~:text=EDA%20for%20time%20series%20is%20realtively%20short%2C%20however,%28as%20it%20can%20be%20done%20in%20this%20example%29.
3. https://medium.com/analytics-vidhya/how-to-guide-on-exploratory-data-analysis-for-time-series-data-34250ff1d04f
4. https://seaborn.pydata.org/generated/seaborn.lineplot.html
5. https://python-graph-gallery.com/242-area-chart-and-faceting
6. https://seaborn.pydata.org/generated/seaborn.scatterplot.html
7. https://pandas.pydata.org/pandas-docs/stable/user_guide/groupby.html

THIS REPORT WAS WRITTEN BY : Natacha Nsengiyumva