## CA682 Data management and visualisation

|  |  |
| --- | --- |
| Name | Lavleen Bhat |
| Student Number | 17210637 |
| Programme | M.Sc in Computing (Data Analytics) |
| Module Code | CA682 |
| Assignment Title | Data Visualisation |
| Submission date | 10/12/2017 |
| Module coordinator | Suzanne Little |

I declare that this material, which I now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work. I understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. I have read and understood the Assignment Regulations set out in the module documentation. I have identified and included the source of all facts, ideas, opinions, and viewpoints of others in the assignment references. Direct quotations from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged, and the source cited are identified in the assignment references. This assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other course of study.

I have read and understood the referencing guidelines found recommended in the assignment guidelines.

Name: Lavleen Bhat Date: 10/12/2017

**Introduction**

The data is about the food from all over the world having multiple countries. The data contains all the food purchased from stores and have different dimensions like the country food product was purchased from, all the ingredients and nutritional information, brands and others. In this data visualization, we’re going to look at the data mainly from European countries and explore different aspects from this data about the food consumed in those countries.

**Dataset**

The dataset is available at Kaggle under the title **Open Food Facts**. It has 356028 rows and 161 columns. The dataset is huge and messy. It has a lot of null values in each column and the format of the data in those columns is not even in the expected format.

**Process**

Data is available in TSV format on Kaggle. File was saved in CSV format and Pandas’ CSV reader function is used to read the data into Python notebook. The data processing and cleansing is done in Python language using different tools and libraries.

**Libraries used:** Numpy, Matplotlib, DateTime, Pandas, Plotly(for interactive visualization)

The data seems to be straight forward but is extremely messy which gives a hard time in cleaning the data and bringing it into a structured format for proper visualization.

**Challenges:**

1. Dataset is not very consistent. The number of data varies from year to year. Dataset has huge number of NaN cells in all the columns.
2. For each visualization, a new subset had to be created to take only required columns and exclude all the messy string and null values.
3. Most of the data is from France and hence, did not include the country in most of the visualizations (except last one) to avoid skewness.
4. Next challenge was to sort out the countries because the column “COUNTRIES\_EN” has multiple countries separated by comma for the same product. Though it looked quite simple and easy once I learned how to split the country and arrange the data but because the dataset is huge, it posed challenge when processing this much huge data.

Similarly, column “BRANDS\_TAGS” has brands separated by comma in the same column. So, had to split them first to make the data understandable.

Same is the case with column “FIRST\_PACKAGING\_CODE\_GEO” having coordinates.

**Results**

We’re going to look at the below three aspects of the dataset.

1. Nutritional Information
2. Environmental
   1. Carbon Footprint
   2. Plastic Consumption
3. Different brands prominence in the region/Sales of the brands
4. **Nutritional Information:**

Here, we’ll look at the nutritional information for European regions. We have categorized the countries into three parts – North Europe, Central Europe and South Europe. The dimensions explored for nutritional information are Fat, carbohydrates, sugar and protein consumption per 100 grams of the quantity.

We can see that the fat percentage and sugar consumption slightly decrease from north Europe to south Europe. The carbohydrates consumption increases from north to south. Though the variations are not much but it gives a little insight into the **food habits** of the European countries.

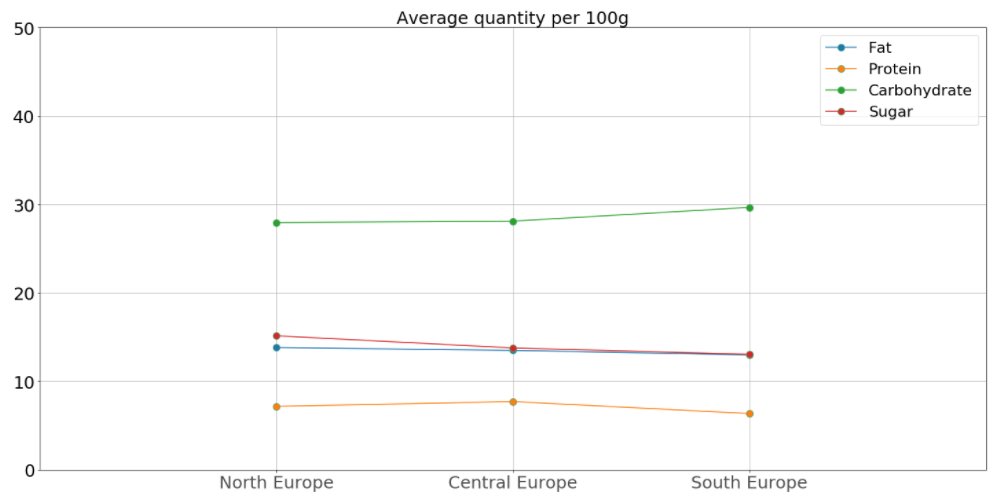


Fig. 1 – Nutritional Information

1. **Environmental**

Under environmental aspect, we’ll be looking at two aspects of the same.

1. Carbon footprint from year 2012 to 2017
2. Plastic consumption from year 2015 to 2017
   1. **Carbon Footprint from year 2012 to 2017**

The first that comes first in mind when some says “Food” is the delicious dishes we eat at the festivals. Some have a sweet tooth and some look at the calories before consuming the food. But not many people look at the aspect that how much impact our food makes at the environment. Dataset has column ‘carbon-footprint\_100g’ which gives information about the carbon footprint each product has on the environment.

Let’s look at the results.

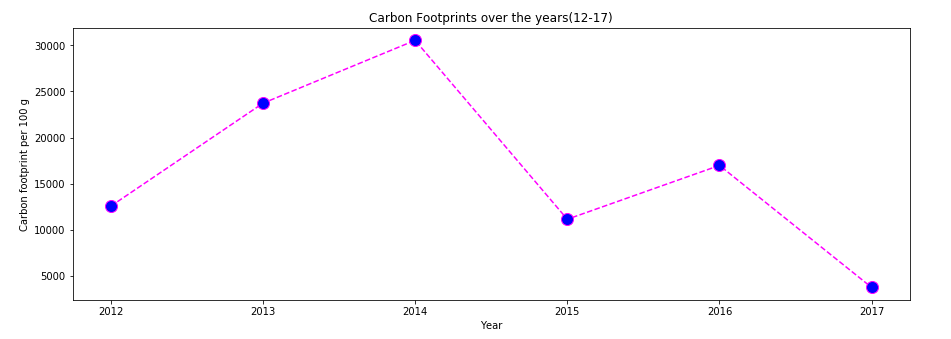


Fig. 2 – Carbon footprint (12-17)

The graph above has the total carbon footprint varying over the years. The main point to note here is that years 2012, 2015 and 2017 have the least carbon footprint among all. This is because of the less data present for these three years. I did not exclude these years just to have a look at the total variation. Years, 2013, 2014 and 2016 have comparable number of data rows and hence, these can be considered to be the legit graphical representation of the carbon footprint.

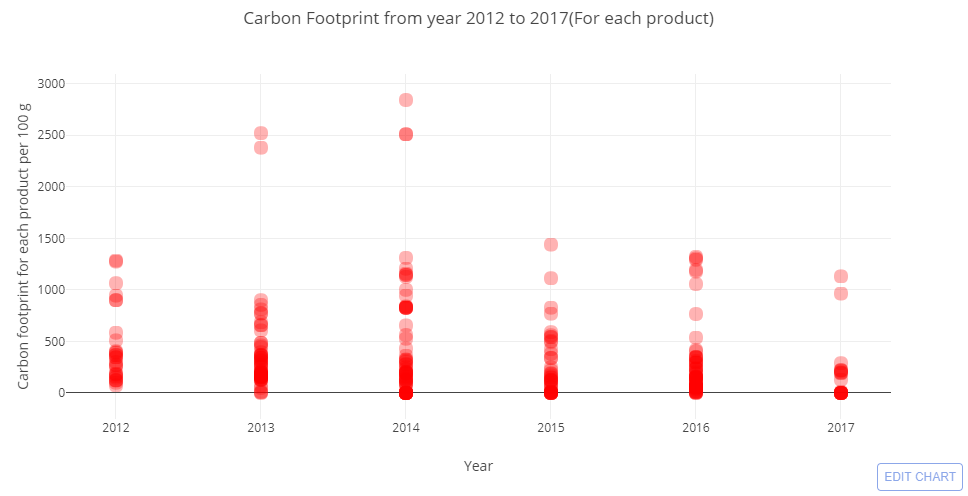


Fig. 3 – Carbon footprint by each product

The graph above shows the carbon footprint made by each product. As most of the products have carbon footprint in the range of 0 to 1000 but there are a few products that have carbon footprints between 2300 and 3000 per 100 grams of the quantity. These number can be significantly high when the quantity of the product consumed is higher.

* 1. **Plastic consumption from year 2015 to 2017**

The plastic consumption is calculated from the packaging of the product. The products packed in plastic are considered for this visualization and not to mention that these products indeed have a bad impact on the environment, from which we can conclude that the use of plastic should be limited.

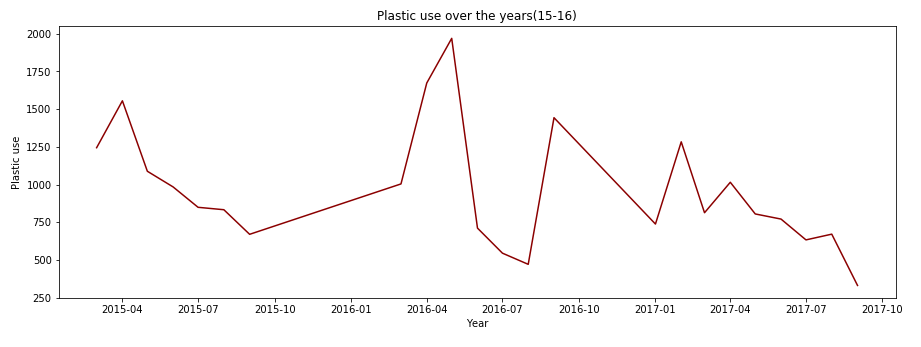


Fig 4. Plastic consumption from 2015 to 2017

The graph shows that the plastic consumption peaked between April 2016 and June 2016. The graph shows the declining use of plastic in year 2017 which can be considered a good gesture towards the environment.

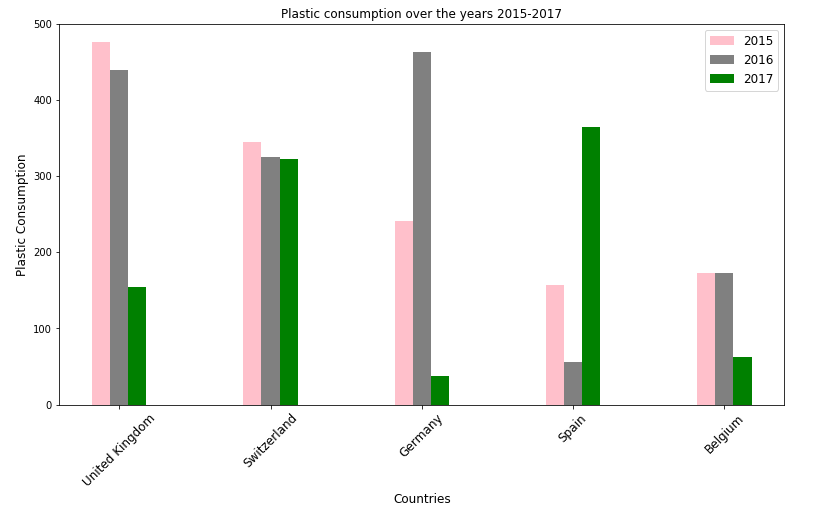


Fig. 5 – Plastic consumption country wise (2015-2017)

Looking at the graph above, we can see the plastic consumption from year 2015 to 2017 for five different countries. Belgium has the lowest consumption of plastic overall. Germany seems to be at peak in 2016 and dipped drastically in 2017.

1. **Presence of different brands in France**

Next dimension explored in the dataset is the presence of difference brands in each country. We have taken France here for reference.

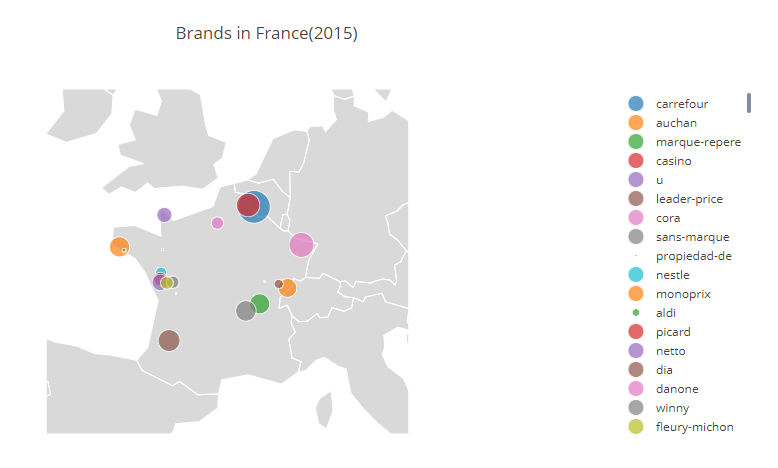


Fig. 6 – Brands in France (year 2015)

The size of the bubble on the map shows its prominence in the country. The bigger the bubble, the more its sales/consumption is. Similarly, we can visualize for other years data as well.

**Conclusion**

We have looked at three different aspects from the dataset. We have seen the food habits of European countries, then looked at the environmental factor of the products and the food brands presence in France. Such visualization can be used in different kinds of situations. For example, nutritional information visualization can be used to find out about the food habits of people in different countries. Then, visualizations done for environmental factor can be used to have an idea about the impact of the food product on the environment. If negative, then take the proper action against the manufacturer and to curb the environmental damage. Brand prominence visualization can be used to understand the market of different companies in different regions. Looking at all of them, it can be concluded that the dataset is huge and there are still many different aspects that can be explored.