

BABEŞ – BOLYAI UNIVERSITY

Faculty of Economics and Business Administration

Cosmetic Products Data Visualisation

Made by:

Boglárka Póra

Table of Contents

Introduction, topic and data selection	3
Data exploration and normalisation.....	4
The process of creating the data visualisation	6
The advantages and disadvantages of the presentation format	7
Scatter plot.....	7
Column chart.....	7
Packed circles	8
Radial tree	9
Treemap	10
What have I learned?	10

Introduction, topic and data selection

First, I would like to write about the process of choosing a topic, which did not start with a specific idea. I started searching the internet for datasets with a complexity of approximately 1000 data points. I looked in several places, such as Statista, but at last I found a dataset on Kaggle, containing data on cosmetics. I looked into it a little bit more and ended up sticking with the topic because I've been very interested and involved in it for quite a few years.

I downloaded the dataset in .csv format and looked more closely at the data to see what I was dealing with. There are several columns that contain categorical data, such as the type, brand, name of the facial care products, and the names of the ingredients. The market price of the products is shown as a number, as well as the ranking score they received from consumers. In addition, the dataset has columns of the type boolean that indicate whether or not the product is suitable for a given skin type (values 0 and 1).

Out[88]:

	Label	Brand	Name	Price	Rank	Ingredients	Combination	Dry	Normal	Oily	Sensitive
0	Moisturizer	LA MER	Crème de la Mer	175	4.1	Algae (Seaweed) Extract, Mineral Oil, Petrolat...	1	1	1	1	1
1	Moisturizer	SK-II	Facial Treatment Essence	179	4.1	Galactomyces Ferment Filtrate (Pitera), Butyle...	1	1	1	1	1
2	Moisturizer	DRUNK ELEPHANT	Protini™ Polypeptide Cream	68	4.4	Water, Dicaprylyl Carbonate, Glycerin, Ceteary...	1	1	1	1	0
3	Moisturizer	LA MER	The Moisturizing Soft Cream	175	3.8	Algae (Seaweed) Extract, Cyclopentasiloxane, P...	1	1	1	1	1
4	Moisturizer	IT COSMETICS	Your Skin But Better™ CC+™ Cream with SPF 50+	38	4.1	Water, Snail Secretion Filtrate, Phenyl Trimet...	1	1	1	1	1
...
1467	Sun protect	KORRES	Yoghurt Nourishing Fluid Veil Face Sunscreen B...	35	3.9	Water, Alcohol Denat., Potassium Cetyl Phospha...	1	1	1	1	1
1468	Sun protect	KATE SOMERVILLE	Daily Deflector™ Waterlight Broad Spectrum SPF...	48	3.6	Water, Isododecane, Dimethicone, Butyloctyl Sa...	0	0	0	0	0
1469	Sun protect	VITA LIBERATA	Self Tan Dry Oil SPF 50	54	3.5	Water, Dihydroxyacetone, Glycerin, Sclerocarya...	0	0	0	0	0
1470	Sun protect	ST. TROPEZ TANNING ESSENTIALS	Pro Light Self Tan Bronzing Mist	20	1.0	Water, Dihydroxyacetone, Propylene Glycol, PPG...	0	0	0	0	0
1471	Sun protect	DERMAFLASH	DERMAPROTECT Daily Defense Broad Spectrum SPF 50+	45	0.0	Visit the DERMAFLASH boutique	1	1	1	1	1

1472 rows x 11 columns

Data exploration and normalisation

To visualise the data, I decided to use Flourish as the tool, as it offers a lot of possibilities and is easy to use. Next came the step of writing down roughly what type of visualisations I wanted to create and what I wanted to show.

I then proceeded to data normalisation and transforming the data into the correct format for the visualisations I had designed. For this process, I used a Python library called Pandas, I coded in Jupyter Notebook, and I also used Excel in some places.

One of the more interesting moments was when I wanted to calculate an average price and average rating for the cosmetic brands (based on the prices and ratings of the products they are associated with), so that I could visualise it separately, giving an insight into roughly what price and quality category a brand fits into. For this I used Pandas' `groupby()` function and called `mean()` to get the average.

```
In [76]: df2.groupby(['Brand']).mean()
```

```
Out[76]:
```

	Price	Rank
Brand		
ALGENIST	70.777778	3.992593
AMOREPACIFIC	103.523810	4.238095
ANTHONY	32.000000	4.233333
APIVITA	30.000000	4.000000
BAREMINERALS	35.833333	4.116667
...
VITA LIBERATA	49.500000	3.850000
VOLITION BEAUTY	49.400000	4.420000
WANDER BEAUTY	25.000000	4.450000
YOUTH TO THE PEOPLE	46.714286	4.357143
YVES SAINT LAURENT	73.000000	3.400000

116 rows × 2 columns

Perhaps the most difficult task was to dismantle the Ingredients column. First, I saved the Ingredients dimension data in a separate Pandas DataFrame and then split it along the commas.

```
In [19]: ingredients['Ingredients'] = ingredients['Ingredients'].str.split(',')
```

Then I separated the ingredients, which were previously listed in a single field, into separate columns.

```
In [21]: ingredients2 = pd.DataFrame(ingredients['Ingredients'].tolist()).fillna('').add_prefix('Ingredient_')
```

```
In [22]: ingredients2
```

```
Out[22]:
```

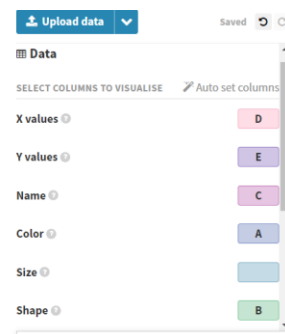
	Ingredient_0	Ingredient_1	Ingredient_2	Ingredient_3	Ingredient_4	Ingredient_5	Ingredient_6	Ingredient_7	Ingredient_8	Ingredient_9
0	Algae (Seaweed) Extract	Mineral Oil	Petrolatum	Glycerin	Isohexadecane	Microcrystalline Wax	Lanolin Alcohol	Citrus Aurantifolia (Lime) Extract	Sesamum Indicum (Sesame) Seed Oil	Eucalyptus Extract
1	Galactomyces Ferment Filtrate (Fitera)	Butylene Glycol	Pentylene Glycol	Water	Sodium Benzoate	Methylparaben	Sorbic Acid			
2	Water	Dicaprylyl Carbonate	Glycerin	Cetearyl Alcohol	Cetearyl Olivat	Sorbitan Olivat	Sclerocarya Birrea Seed Oil	Bacillus/Soybean/Folic Acid Ferment Extract	Nymphaea Alba Root Extract	sh-Oligomeric
3	Algae (Seaweed) Extract	Cyclopentasiloxane	Petrolatum	Glyceryl Distearate	Phenyl Trimethicone	Butylene Glycol	Hydrogenated Vegetable Oil	Cholesterol	Butyrospermum Parkii (Shea Butter)	Stearic Acid
4	Water	Snail Secretion Filtrate	Phenyl Trimethicone	Dimethicone	Butylene Glycol	Butylene Glycol Dicaprylate/Dicaprate	Orbignya Oleifera Seed Oil	Butyloctyl Salicylate	Cetyl Peg/Ppg-10/1 Dimethicone	Cyclopentasiloxane
...

I then converted it into a list, and then into a set data structure to filter out duplicate values. And so I had a one-dimensional Ingredients data set with thousands of rows.

The process of creating the data visualisation

Before I started to create the project in Flourish, I saved the appropriate amount of data in separate .csv files according to the visualizations I had planned. First of all, I created a scatter visualisation on which I placed all the products, showing how they relate to each other in relation to price (x axis) and rank (y axis).

A	B	C	D	E
Label	Brand	Name	Price	Rank
Sun protect	MDSOLARSCIENCES	Quick Dry Body Spray With SolSci-X™ Broad Spectrum SPF 40 UVA-UVB Sunscreen	20	3.9
Sun protect	COOLA	Sport Continuous Spray SPF 30 - Unscented	32	5
Sun protect	MOROCCANOIL	After-Sun Milk Soothing Body Lotion	28	4.7
Sun protect	SUPERGOOP!	Perfect Day 2-in-1 Everywear Lotion Broad Spectrum SPF 50 + Mint Condition Lip Shield SPF 30	19	4.8
Sun protect	PETER THOMAS ROTH	Oily Problem Skin Instant Mineral Powder SPF 30	30	3.7
Sun protect	COOLA	Summer Duo	36	4.8
Sun protect	VITA LIBERATA	Passionflower & Argan Dry Oil Broad Spectrum SPF 50	45	4.2
Sun protect	IT COSMETICS	Anti-Aging Armour™ Super Smart Skin-Perfecting Beauty Fluid SPF 50+	38	4.1
Sun protect	URBAN DECAY	Naked Skin Bronzing Beauty Balm Broad Spectrum SPF 20	34	4.1
Sun protect	KATE SOMERVILLE	Daily Deflector™ Moisturizer Broad Spectrum SPF 50+	48	3.9

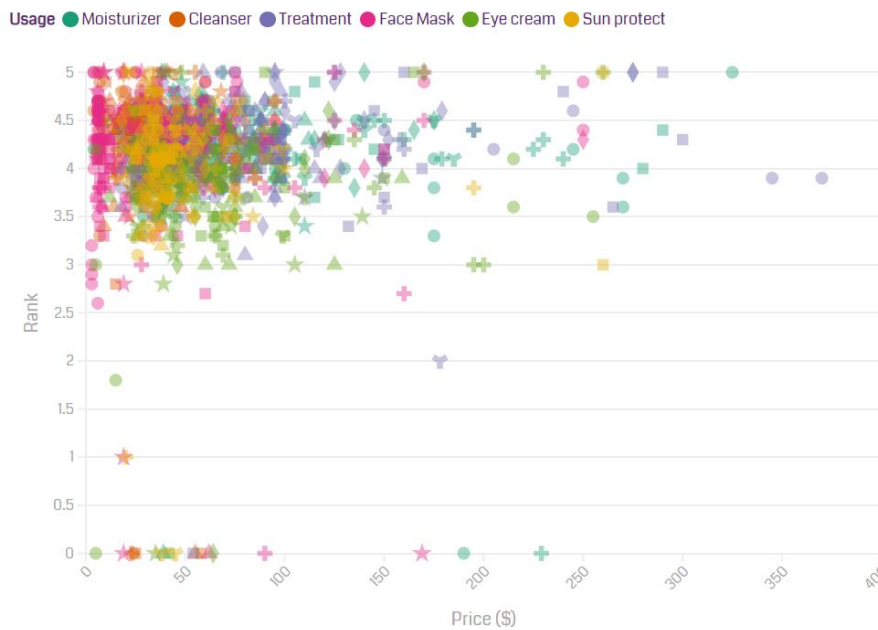


First, I uploaded the file by clicking on the Upload data button, then I started to find the most suitable settings on the right side of the platform. Here we can see that you can differentiate the type of cosmetics by colour, what they are used for, and that the data points are displayed in a different shape for each brand. I then clicked on Preview to set the colours and fonts for the visualisation, I adjusted it a little bit more, I gave it a title and I saved it.

The rest of the visualisations are based on a similar workflow. I have 5 visualisations in total, a Scatter plot, a Column chart, and three Hierarchy charts, including a Packed circles, a Radial tree and a Treemap. The last step was to put together a story, also in Flourish, where I placed the visualisations one after the other, so that they could be viewed together.

The advantages and disadvantages of the presentation format

Scatter plot



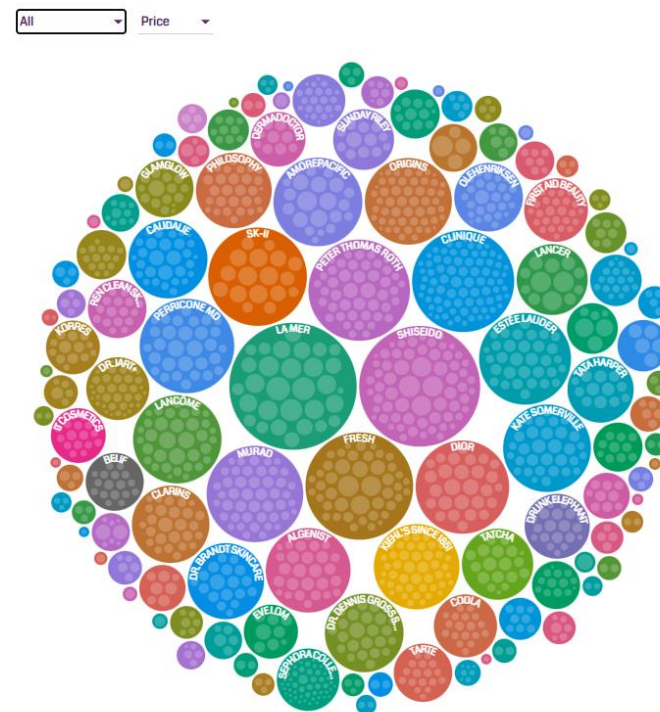
The advantages of scatter plot visualisation include the ability to examine multiple aspects together and to easily see how data points behave in relation to each other. However, it has the disadvantage of being very clustered for similar values.

Column chart



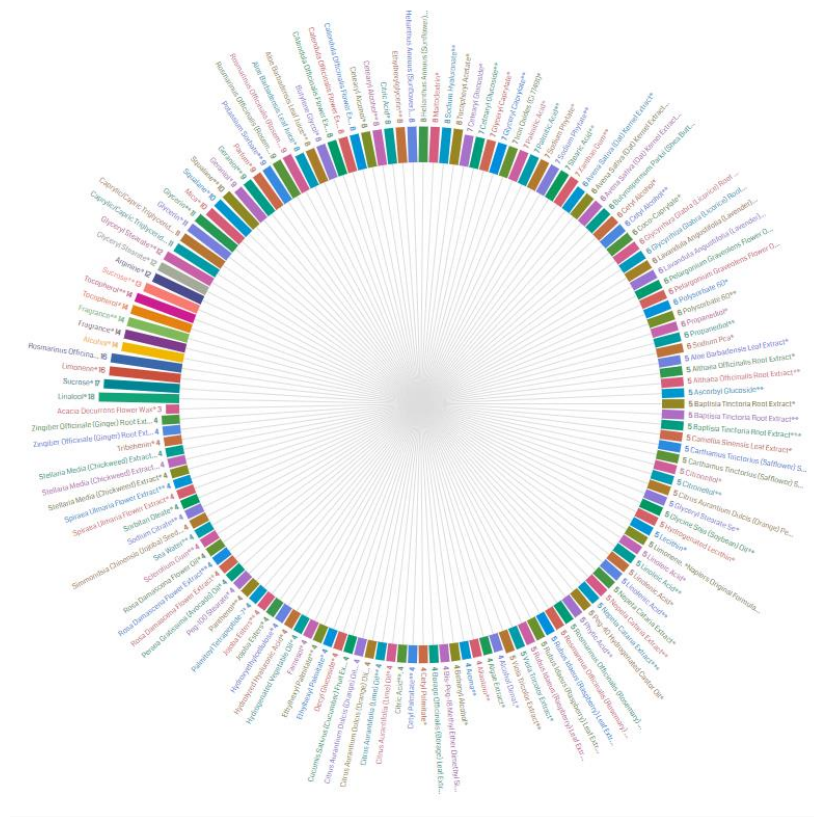
The advantage of the Column chart is that it is very easy to see which coloured column (for which skin type the product is recommended) indicates a positive or negative value. A disadvantage was that I had 0 and 1 values, so I had to rewrite the 0s to -1s in order to visually get the point across more beautifully.

Packed circles



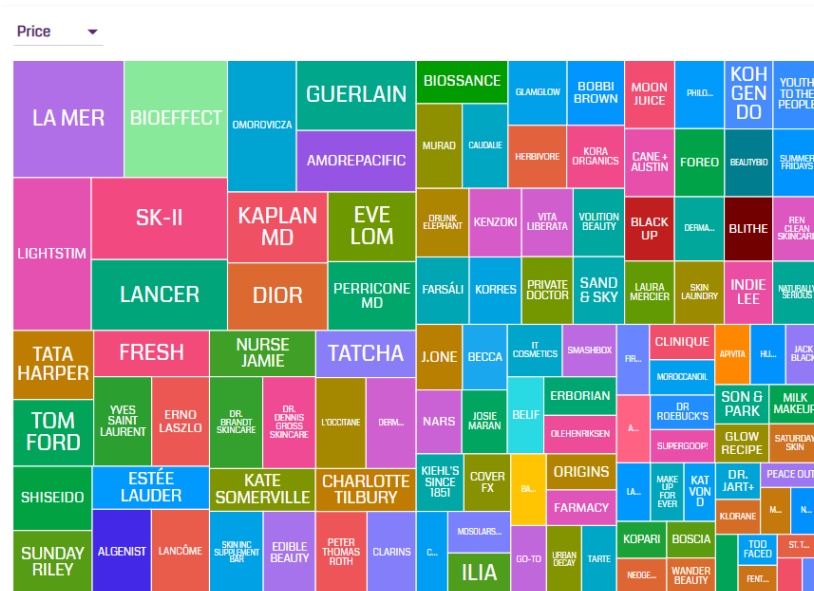
The great advantage of the Packed circles visualisation type is that it is very easy to distinguish categories by colour and to nest them. As you go deeper, you can read more and more detailed information from the diagram. The disadvantage, however, is that when you want to examine all the categories at once, there are so many that the labels are not really visible.

Radial tree



The advantage of using the Radial tree hierarchy diagram was that it could be used to simultaneously read the name of the ingredient around it and also read how many times that ingredient is present in the available products. The disadvantage is that the labels are a bit small.

Treemap



One advantage of the Treemap visualisation is that it is easy to see the average price and ranking ratios and differences between brands, and the colours help to distinguish them. The disadvantage of the chart format is that smaller values are not always visible.

What have I learned?

Throughout the Business Intelligence course and the projects, I learned a lot of new things about the concept of data visualisation, its theoretical and practical benefits, and how to start putting it into practice. Perhaps the most important thing is that I realised how useful visualisation itself is when you want to communicate some information, to make a presentation on a topic. I have learned how to collect data, format it, shape it to get the most transparent and informative result possible. Also, it should not be forgotten that I was able to practice which type of chart works best with which type of data.

Link to the visualisations in Flourish: <https://public.flourish.studio/story/1794448/>