



UNIVERSITAT_{DE}
BARCELONA

PRESENTATION AND VISUALIZATION

MASTER IN FUNDAMENTAL PRINCIPLES OF DATA SCIENCE

TASK 1: DASHBOARD

DESIGN AND IMPLEMENTATION OF A DASHBOARD



Authors

Karol Jurík
Álvaro López Caro
Vladislav Nikolov Vasilev

FACULTY OF MATHEMATICS AND COMPUTER SCIENCE

ACADEMIC YEAR 2021-2022

Introduction

We have chosen the following dataset: SN7216 Living Costs and Food Survey, 2010: Unrestricted Access Teaching Dataset. The Living Costs and Food Survey dataset (LCF) collects information on spending patterns and the cost of living that reflects household budgets across the country. The survey is conducted throughout the year across the whole of the UK and is the most significant consumer survey undertaken in the UK. The LCF is a voluntary sample survey of private households. Each individual aged 16 and over in the household visited is asked to keep diary records of daily expenditure for two weeks. The total sample size for LCF is 5263 observations. It contains 14 variables which include information on total expenditure as well as standard socio-demographic variables (more detailed description of the variables can be seen in the next part of the document). The dataset has no missing values. The data have temporal and also geographical nature since they were collected in the United Kingdom in 2010. The life span of the data is one year and they should be updated every year. However, the life span of the data could be more than one year if we use them for some mathematical modelling and we also include some economical aspects. It depends on the purpose of particular analysis.

Exploratory analysis

As it was mentioned before, the dataset contains 14 variables - 4 scalar variables, 3 ordinal variables, 7 nominal variables. All details about the variables can be seen in the following table.

Variable name	Variable label	Variable label
casenew	Randomly generated case number	Scalar
weighta	Annual weight	Scalar
p550tpr	Total expenditure, by adults & children (top-coded)	Scalar
p344pr	Gross normal weekly household income (top-coded)	Scalar
p425r	Main source of household income	Nominal
a172	Internet connection in household	Nominal
a093r	Economic position of Household Reference Person	Nominal
a094r	NS-SEC 3 class of Household Reference Person	Nominal
a121r	Tenure type	Nominal
sexhrp	Sex of Household Reference Person	Nominal
a049r	Number of persons in household	Ordinal
g018r	Number of adults in household	Ordinal
g019r	Number of children in household	Ordinal
gorx	Government Office Region	Nominal

Table 1: Details of variables

Let's take a look at the most important variables one by one.

- **Randomly generated case number (scalar):** This variable is just a randomly generated case number for every observation/household. Every case number should be unique, but after verifying it we can see that only 4 445 out of 5 262 observation are unique. There is

no additional information about this fact in the documentation, so we will consider all the observation to be unique.

- **Total expenditure, by adults & children (scalar):** The total expenditure is in British Pounds (£). Let's take a look at the summary statistics. We can find them in the following table:

Min.	Q25	Median	Mean	Q75	Max.
7.785	225.319	386.973	446.398	594.071	1175.000

In the figure 1, we can see the histogram of total expenditure and also the kernel estimation of density and the theoretical density (mean and sd were calculated from the dataset). We can observe an unusual behaviour when we look at the households with high total expenditure. When we look at the outliers¹, we can see that we have 231 outliers out of 5262 observations. What is more interesting is that there are 214 households out of 231 with the same total expenditure which is equal to the maximum². There is also no additional information about this fact, so we suppose that some kind of threshold for maximal expenditure was used in this case. We also reject³ the hypothesis of normality at the significance level of 0.95.

NOTE: The original variable was top-coded at 96th percentile, therefore we can observe so many households with the same value of total expenditure.

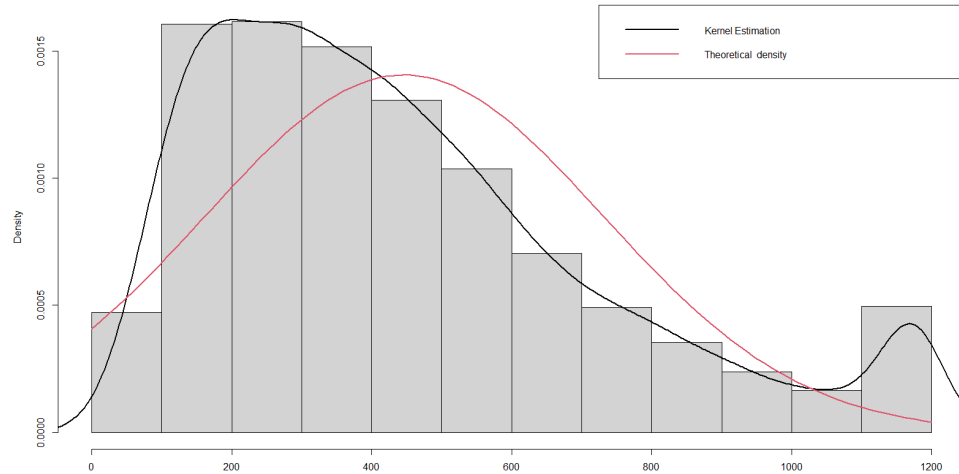


Figure 1: Histogram of total expenditure, by adults & children

- **Gross normal weekly household income (scalar):** This variable is also in British Pounds (£). We can find the summary statistics in the following table.

Min.	Q25	Median	Mean	Q75	Max.
0.0	276.1	511.1	589.6	885.3	1185.0

¹Upper whisker = $\min\{\max(x), Q75 + 1.5IQR\}$

²£ 1175

³Lilliefors (Kolmogorov-Smirnov) test for normality was used.

We can see that we have some households with no normal income. It is exactly 11 households where the persons are either economically inactive or unemployed (and work related Government Training Programmes). It is also people that never worked, are long term unemployed or they are students.

In the figure 2, we can again observe an unusual behaviour. There are 722 households with exactly the same income which is equal to the maximum⁴. We reject the normality at the level of significance of 0.95 again. There is no additional information about why that many households have the same income, so we can assume that some kind of threshold for maximum income was used again. In our further analysis, we should deal with those maximal values, so they will not corrupt the output. We can also look at the correlation between total expenditure and gross income. The correlation⁵ is 0.751, so we can observe strong positive correlation which makes perfect sense.

NOTE: The original variable was top-coded at 96th percentile, therefore we can observe so many households with the same value of gross normal weekly income.

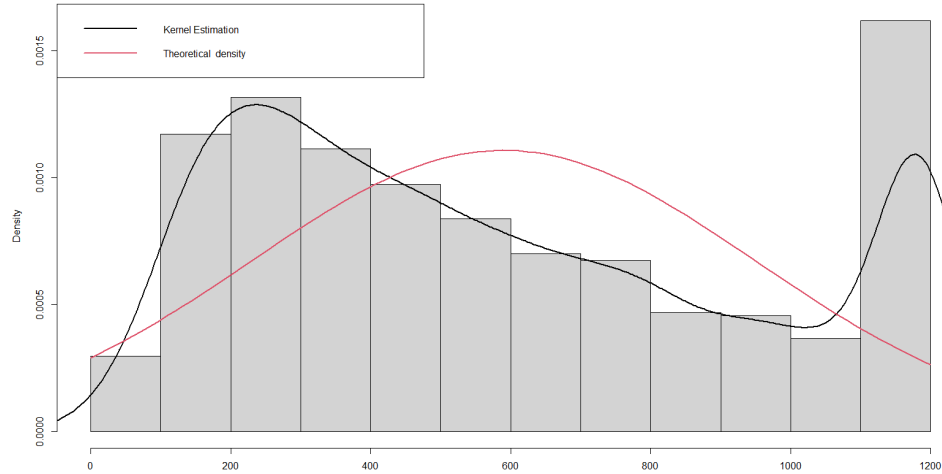


Figure 2: Histogram of gross normal weekly household income

The rest of the variables are either ordinal or nominal. Since there is no space to describe all the variables and present all the summary statistics, let's take look at some interesting associations.

⁴£ 1185

⁵Spearman correlation coefficient was used.

- **Household size**

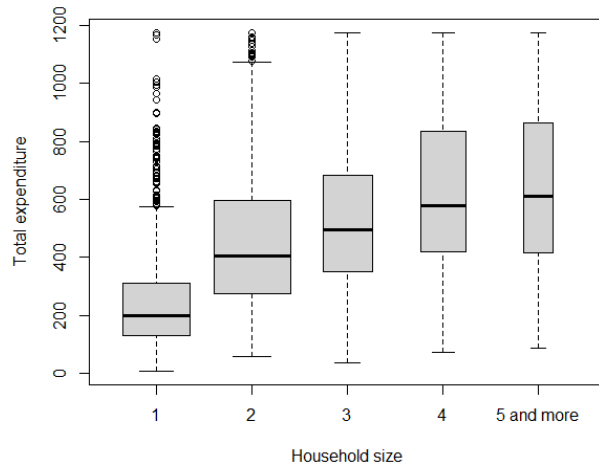


Figure 3: Total expenditure vs. number of persons

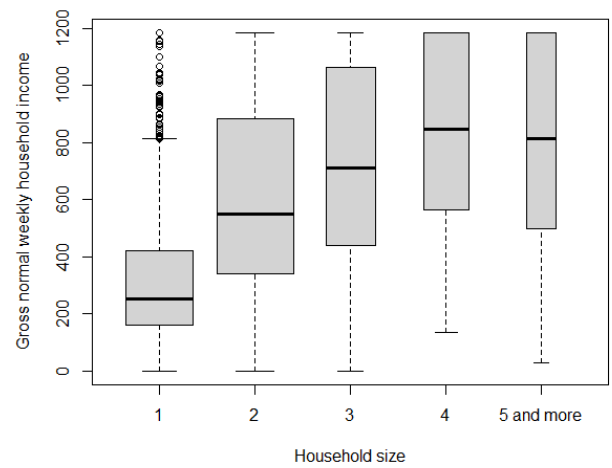


Figure 4: Weekly household income vs. number of persons

In the figure 3 and 4 we can observe that the bigger the household is, the higher expenditure and also income they have. The width of the boxes shows us how big the particular group is. We can see that the majority of households consist of 1 or 2 persons. We can also observe a lot of outliers in the first group. Most of the outliers are owners of the tenure, they are working on full-time and they have higher managerial, administrative and professional occupations. We can also observe this kind of behaviour when we look at the other variables like number of adults and number of children in a household. It is also worth to mention that there are no children in around 70% of households. Majority of households - around 54 % - consist of 2 adults and only 2.5 % of households consist of 5 and more adults.

- **Another interesting facts:**

- around 68 % of the respondents are owners of the tenure (the rest rent the tenure);
- the ratio of men and women in the survey is around 6:4;
- total expenditure and also weekly income is higher in the group of men, although there are much more outliers in the group of women;
- majority of outliers came from the households which consist only of one adult;
- households with no internet connection have significantly lower expenditure and also weekly income;
- the data came from exactly 12 parts of the United Kingdom;
- ...

Conclusion

The dataset consist of many observations and can be used for many purposes. There are a lot of associations and relations that can be found in the dataset but it would require a much thorough analysis (not only 4 pages). Moreover, there are also many outliers when we group the data by some variables. In the following analysis, we should take a closer look at those outliers and try to get some valuable information out of them.

Audience

The first question that needs to be answered is: *Who is our audience?* To make it more clear and easier to understand, let us introduce our internal customer in the first place. As it was already mentioned in the previous section, the data comes from the survey that collects information on spending patterns and the cost of living that reflect household budgets. It is conducted throughout the year, across the whole of the UK, and is the most significant survey on household spending in the UK. It is also an important source of economical and social data for a range of governmental and other research agencies. The results are essential for understanding society and planning to meet their needs. Therefore, our internal customer would be the Government of the United Kingdom, especially The Department for Levelling Up, Housing and Communities in the UK⁶ that supports communities across the UK to thrive, making them great places to live and work.

The audience would be then the ordinary people of the United Kingdom, especially young people who try to decide where to start a family or where to move in order to have better living conditions. As it is well known, the UK is one of the most developed parts of the world. The job market is becoming more and more competitive and the living costs are getting higher every year. Therefore, it is quite difficult for young people to decide where is the best place to start a family. The visualization can help those people to choose the right region for them, they can compare their current living costs to the people from a different region, they also can see if they can afford to have a child and how much does it approximately cost in a particular region. The visualization can also help other people, not just those who want to start a family. People can compare their living standards to other people from different or the same region. If some people consider moving to a different region, they can get information about the population and also living standards in that particular region and so on. Finally, other ministerial departments in the UK can also use the visualization to compare the living standards and the structure of the population within the regions in the UK. Then, they should use the outpost in order to come up with new policies in order to try to diminish the inequalities within the regions (more support for the families with the children, create more working opportunities in some regions, etc.).

The outcome can be also used for other purposes. For example, journalists can use it to create articles, reports, etc. and present the information to the general public. Also, it can be used by research companies for analysis. Nevertheless, the most important audience for the Government should be the ordinary people, the electors. Therefore, the visualization will be mainly used by Government to provide information about the living costs in the United Kingdom. They will publish all of the charts on the official governmental website. And since our internal customer is the Government, they will consult the visualization quite often. If we consider the information that the data are released on yearly basis, then we can assume that the Government would like to have the visualization done as soon as possible, so they can release the outcomes shortly after the survey is done. However, they have to keep in mind that not every single person of the population is an expert in the usage of charts. Thus, they have to make them as accessible and intuitive as possible. The visualizations will probably be consulted at least once a week and will be used by the experts that work for Government or by the politicians themselves. They will

⁶<https://www.gov.uk/government/organisations/departments-for-levelling-up-housing-and-communities/about>

use it for creating a new policy or for proving that their current policy is right. Then, the charts will be shown to the people to support the political statements of the government.

The next important thing is to introduce the persona. In this case, our persona is a young couple from the UK. The young couple consists of people who just finished their studies and want to start a family. So, their goal is to find the right place for living, working, and also raising their children. As it is well known, the UK is one of the most developed countries in the world. The competition in the job market is high, the living costs are getting higher and higher every year. It is becoming more and more difficult to start a family for a young couple. Our young couple wants to start from scratch, they want to plan a future together and they want to find the right place for that. They want to move to the region that suits them best. The population of that region is similar to them, for example in terms of class of occupation. Our young couple also plans to have at least one child in the future. So, they want to move to the region where many households consist of some children too. Finally, our young couple wants to increase their living standards. Obviously, they have some preferences where to live, but they also want to increase their income and lower their expenditures, so they have enough savings to start a family. Taking all the above-mentioned facts into consideration, our young couple wants to make the best decision for themselves. The visualization can help them in reaching their goal. They can get information about all the regions within the UK and they can find the right one that meets their requirements the best. Once they decide on one particular region to move to, their goal is reached. Both, the young couple the government of the UK can benefit from that. The young couple can start the family in the best possible way for them. On the other hand, the government of the UK needs to increase the population of young people and they need to motivate the young families to have some children. If the young couple reaches their goal by finding the right place for living where they can afford to have a child, it will be also beneficial for the whole society, and therefore for the UK government too.

The message we want to convey with our data is simple. We want to find out the patterns that show if the current policies are right. If we can find this information, the government can show the charts to the people in order to show them what they are doing for them, that their policies are right and work for the majority of people. If we find some unusual behavior such as big differences among some groups, we can pass this information to the government, so they can act as soon as possible in order to find a solution for that given problem. In this second scenario, they can also show the charts to the people to support the fact that there is a problem in society and they are trying to address it. In both cases, they can show people that they care about them, that they want to fulfill their needs, and convince them that the current government is the right choice for them.

And what will be the message for the audience? They will be able to see from all of the charts what is the current state of society. They can compare their expenditure/income to similar households, so they can see how they are doing. Also, they can observe how is the situation in different parts of the UK or among different households within their own region. They can observe whether the government treats its population fairly or not and if it does its best to meet the needs of all of the citizens. The people, especially young people, who plan to move in order to increase their living standards or in order to start a family can also find useful information in the visualization. Finally, the visualization can help them to make better decisions for their future.

Napkin design: first approach of the dashboard

In order to present the data and to achieve the previously defined goals, we are going to create a Dashboard. Before implementing, we are going to present an initial design of it:

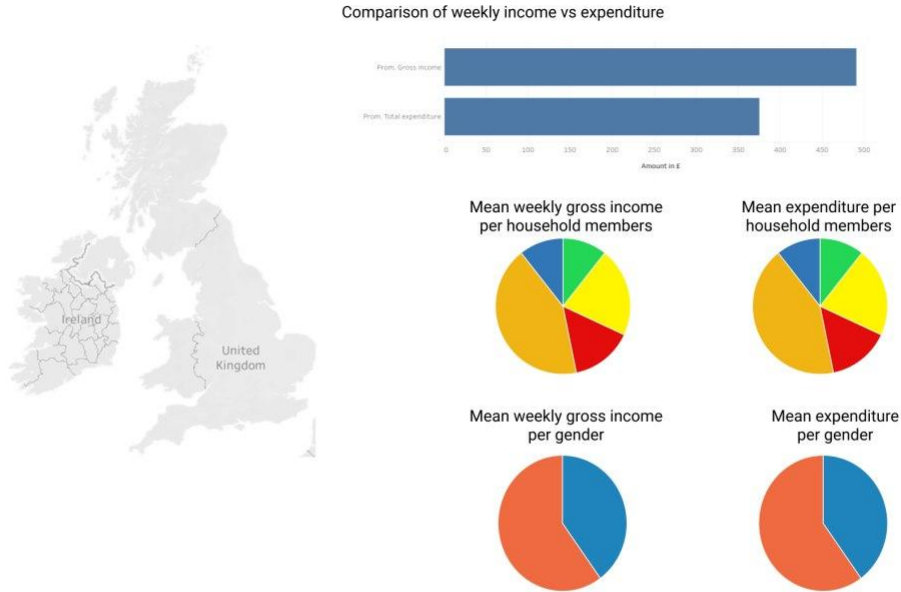


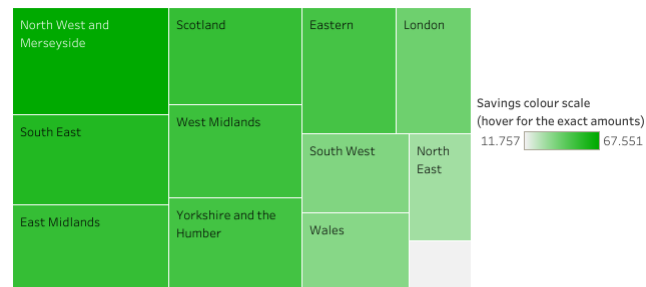
Figure 5: Napkin design of the dashboard.

To convey the previously defined points, we want our dashboard to include a map of the UK along with a bar chart in which the average weekly gross income and expenditure are compared and four pie charts. The pie charts will be divided in two rows:

- The first row will show the mean weekly gross income and expenditure per households members.
- The second row will show the mean weekly gross income and expenditures per gender.

The map will show the different regions in which the survey has been conducted. In each of these regions we can show, for instance, the difference between the income and the expenditure of that particular region. When the user clicks on any of the regions, the previously defined charts will update to show the information corresponding to that region.

Total savings per household (no taxes applied)



Mean income per gender



Gender
1
2

Average Gr..

Mean income per household size



Household size
(ascending order by gross income)
1
2
3
4
5

Average Gr..

Figure 6: Napkin design of the first level of the dashboard.

Comparison of total gross income and expenditure per household

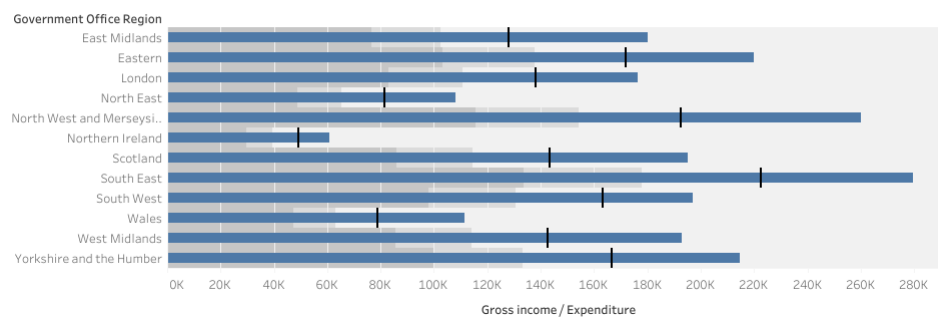


Figure 7: Napkin design of the second level of the dashboard.

Second approach of the dashboard

The dashboard can be seen via the following links:

- https://public.tableau.com/app/profile/karol.jur.k/viz/PV_project/Dashboard1
- https://public.tableau.com/app/profile/karol.jur.k/viz/PV_project/Dashboard2

Since the last delivery, we added/changed the following things:

- we added the map of UK;
- we changed means into medians (because the data are not normally distributed and median is more robust than mean);
- we changed the colour of the pie chart showing median income per gender;
- we changed the pie chart that was showing the mean income per household size into tree map.

The map of the UK shows the median savings per household in the regions of the UK. It works like a heat map, so the regions with the highest median value are shown in the darkest green color. If the user hovers the mouse cursor over a particular region, he/she can see the value of the median in that particular region and also the name of the region. Moreover, the map of the UK is also used as a filter, so it interacts with the other charts. If the user clicks on a particular region, the filter is applied, and all the charts correspond only to the chosen region. In the top right corner, there is also a list of regions where a user can choose any combination of regions. It also works as a filter and interacts with the map, and therefore with all of the other charts. The pie chart and the tree chart can be seen at the bottom of the dashboard. They show median income per gender and the median income per household size respectively. In both cases, if a user hovers the mouse cursor over the chart, he/she can see a value of the median income in a given group.

In the bullet chart, a user can observe the comparison among the regions of the UK. He/she can compare the median income and the median expenditure among the regions. Moreover, if the user is interested in more details (i.e. the value of the income/expenditure), he/she can just hover the mouse cursor over the bar corresponding to the region of his/her interest, and he/she can see the details.

NOTE: In the web version of Public Tableau, the interaction does not work as it should. However, it works correctly in the desktop version. We need to fix this issue for the final delivery.

Things to consider:

- usage of the pie chart for showing the median income (the bar chart or just the table would be more appropriate since income is a continuous variable);
- labeling the charts, so a user can see the values of the income directly;
- usage of the list of regions as a filter;
- a possibility to add more charts;

- colours;
- the bullet chart;
- interaction with the bullet chart;

Final design of the dashboard

Selection of chart and encoding

For the final design and changes of the dashboard, we have taken into account both conventional design principles and "Gestalt" design principles along with preattentive processing.

“**Conventional**” principles are considered the building blocks of geometric and chromatic design, since they determine how the elements of design (e.g. shape, direction and colour) will be used. This will ensure the success of our overall design. For our dashboard, they were included in the following:

- Our charts in the main dashboard are well **aligned**, ensuring a sharper and ordered design by creating a visual connection between elements. The same goes for the differential chart in the secondary dashboard.
- **Size balance** gives more or less importance to some dashboard sections by bigger or smaller size respectively. That helps the reader identifying the main information first, then the details, instinctively. In the main dashboard, the interactive UK map with the household savings per region is shown bigger compared to and followed in size order by the bar charts and the donut charts, due to this principle.
- **Contrast** combines differences in colours, fonts or lines, to emphasize important information in the dashboard. We can leverage contrast to drive the reader’s attention, by indicating where to look first, or which element to interact with. **Colour contrast** is the most powerful type and is the one we considered for our charts.
 - We used **colour saturation to encode numerical quantities and greater saturation to represent greater numerical quantities**. We can see that in the interactive map where different tones of green (the colour of money) represent different median savings per region. The same principle has been used when making the donut charts.
 - We avoided using a saturation sequence to encode a large number of values. Instead, we used other choices.
 - **Light colours will be the best to colour code large areas** because there is more room in colour space in the high-lightness region than in the low-lightness region. This is why we chose green for the UK map.
 - For qualitative colour labelling displayed in the bar charts or the differential, we were **consistent with our colour choices** and didn’t change it between charts.
- **Repetition** strengthens consistency and ease of associations across sections, by repeating the same format multiple times, to describe different information. It also defines the “rhythm” of the dashboard reading. One can see how we applied this in the making of our donut charts.
- We obviously used **different visual channels** to display aspects of our data so that they are visually distinct.

The “**Gestalt**” principles are named after the Gestalt psychology, aiming to understand how humans typically gain meaningful perceptions from the chaotic stimuli around them. All the principles are based on the idea that our mind translates what the eye sees by perceiving a series of individual elements as a whole. We can apply some of those ideas to dashboard design by leveraging the essential laws of the human attempt to find order in disorder. That can enable us when we think about how to get a better dashboard structure, make our charts more straightforward to translate, etc.

In our case, some of its principles such as closure or continuity cannot be applied to our dashboard since we sought **interactive and simple charts** that the viewer can **analytically** interact with over other type of charts that might take those laws into account. Nonetheless, we used the following considerations:

- We didn’t really use **enclosure** but we did take advantage of **proximity** in our main and secondary dashboards in order for closer charts to appear to be more related than things that are spaced farther apart. That’s handy because related elements of the dashboard can be easily related simply leaving some free space, with no need of using too many enclosures, which would cause a heavier appearance.
- Following this last principle, we used **vertical and horizontal symmetry** (differential chart, all the different elements of the main dashboard) to gives us a sense of solidity and order, as well as a sense of symmetrical elements belonging together.
- We also used **colour similarity**, as we just mentioned, to enhance that sensation of elements being related to each other.
- Despite being trivial, UK is very distinct from its map’s white background in order to make it easy and natural to interact with the different regions showing the savings data.

Interaction

The map of the UK shows the median savings per household in the regions of the UK. It works like a heat map, so the regions with the highest median value are shown in the darkest green color. If the user hovers the mouse cursor over a particular region, he/she can see the value of the median in that particular region and also the name of the region. Moreover, the map of the UK is also used as a filter, so it interacts with the other charts in the first level of the dashboard. If the user clicks on a particular region, the filter is applied, and all the charts correspond only to the chosen region.

The second level of the dashboard shows the differential chart of the median gross weekly income and the total expenditure. On the left side, there are single value lists where a user can select a specific combination of variables (e. g. a user wants to see the result only for the households that have 2 or more children, etc.). Based on his/her selection, the differential chart will change, so there is an interaction between the differential chart and the single values lists. In the bottom left corner, there is also a small table with the savings that are calculated from the difference between the income and the expenditure. If it is not obvious from the main chart - differential chart, a user can take a look at the small table in the bottom left corner to see in which region the households have the highest median savings. This table also interacts with the single values lists that are shown above the table.

Worksheet

The worksheets can be found via the following links:

- 1st level dashboard:
https://public.tableau.com/app/profile/karol.jur.k/viz/LCF_dashboard/Dashboard1
- 2nd level dashboard:
https://public.tableau.com/app/profile/karol.jur.k/viz/LCF_dashboard/Dashboard4
- Other worksheets:
https://public.tableau.com/app/profile/karol.jur.k/viz/LCF_dashboard

HTML page

<https://vol0kin.github.io/VIS/dashboard/index.html>