Data Visualization Exam

Magnus Olesen

Study number: 20194512

28. January 2021

Table of Contents

1 Introduction

Data visualization is a common tool in the field of data analysis. It enables very efficient communication, by exploiting the human visual system and known visual phenomenons such as popout. Popout is when a specific colored item is immediately noticeable compared to the other items with another color. Besides communicating interesting findings and results, visualizations can also aid in getting an overview of datasets. For large datasets, this overview is not a simple as only showing one view, as different idioms or views may be necessary to properly encode all the relevant data. Interactive graphs can help with this by allowing the user to change the view thereby reducing visual cluster and allowing multiple different encoding styles and idioms [1, p. 244].

Moreover, there may not exist one "correct" way of showing the data. Depending on the tasks the designer wants the user to be able to fulfill, the chart type may vary [1, Chapter 1].

1.1 The Dataset

The data used in this report pertains the *The World Happiness Report*, which is a survey concerning the worlds happiness, where 158 countries are assigned a happiness score. This score is based upon a series of questions asked in a *Gallup World Poll*. Furthermore, the dataset also contains some explanatory attributes such as GDP, life expectancy or government trust etc. These attributes are estimates that try to explain to which extend these factors contribute to the overall happiness score. That also means that they have no direct impact on the happiness score.

The dataset is a static set of the table type, with it containing 12 unique attributes and 158 unique items, one row for each measured country. Because of the fact that the data is concerning geographical countries, the argument could be made that the dataset is a combination of the two types; table and spatial data.

The dataset can be found on this link to Kaggle, where an even more in-depth explanation of the dataset also exists. On Kaggle five datasets containing the World Happiness Report 2015-2019 can be found, but in this report, the focus will be on the 2015 one.

1.2 Tasks

Now the tasks for the users will be defined. With the support of the constructed visualizations, these tasks should be able to be answered effectively.

- What are the happiest countries in the world?
- Can a correlation between the region of the countries and their happiness score be found?
- What factors seem to contribute toward or explain the countries happiness score?

2 Idioms & Design Choices

In an effort to answer the tasks as best as possible, three different interactive idioms have been designed.

In this chapter, the design choices made for each visualization idiom, will be explained. This will be done in reference to the "What, Why, How" framework, as seen in Figure 1. Furthermore, the strength and weaknesses of each idiom as well as their signifiers and affordances will also be investigated.

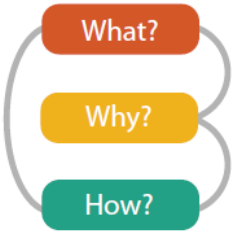

Figure 1: The "What, Why, How" framework [1, p. 16-17].

The visualizations can be found on https://moleseaau.github.io/, and the source code can be found this Github repository.

2.1 Bar Idiom

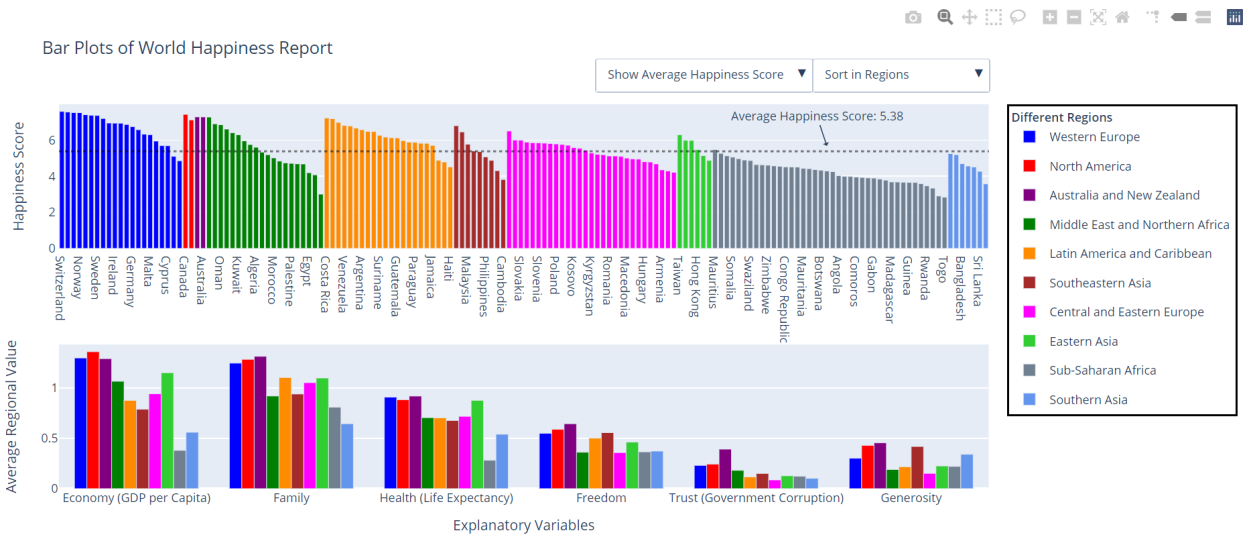

Figure 2: Linked bar charts of World Happiness Report.

The main visualization in this report is the bar idiom, see Figure 2. This visualization consists of two bar plots. One at the top for the happiness score and one at the bottom for the explanatory attributes. The two plots are linked together by the legend, which affords clicking to filter out data.

What

The top bar plot encodes three attributes. The first being the categorical attribute country on the x-axis, the sequential quantitative attribute happiness score on the y-axis, and the color encoding the categorical attribute region. Furthermore, the derived attributes rank and average happiness score can also be seen.

The second bar plot also contains the categorical attribute region, which is shared between the two plots. Then it encodes information containing the six sequential ordered quantitative explanatory attributes. The attributes shown however, are actually the derived average regional values.

Why

The goal for this idiom is for the user to discover new information which maybe was not previously known. Furthermore, a choice was made to produce new information in the idiom, such as an annotation for the derived attribute average happiness score.

The actions available for the user to discover this new information is to search in the idiom. Depending if the target is known or unknown, the user can locate, browse or explore. The user can most likely not lookup the target as the location of the target is probably unknown for the user. When the search is successful, the user can either identify one target, compare two targets or look at a summarization of multiple targets.

The idioms target is to look at both the dependency between the explanatory attributes and the happiness score, but also to identify any (dis)similarities for the countries.

How

Now it will be explained how the visualization idiom has been designed to support the intended targets and actions.

A bar plot like this one, encodes two attributes by itself. Using line markers with vertical spatial channel for the happiness score in combination with the horizontal spatial position for the country attribute [1, p. 97, Figure 5.4]. Furthermore, the color channel hue has been added to the idiom, thereby encoding a second categorical attribute. This is for the top bar plot, but the same asserts itself in the bottom bar plot.

Initially the countries have been separated in categories, i.e. the region of the country. The option exists for the user to change the encoding in the idiom by ordering the countries by rank. The user can also manipulate the view in different ways. They can filter out one or multiple regions in the legend, or change the view by using the zoom-in function.

Signifiers & Affordance

When designing interactive visualizations, it is important for the designer to consider how the visualization signify where it affords the user clicking.

In this vis idiom, three signifiers have been added. Two signifiers have been added in the form of drop downs. In these drop downs, the sign of a downward facing arrow has been added, further signifying that they afford clicking. When clicked, they reveal the option for the user to either show the average happiness score or to sort the top plot by rank. The

text in these drop downs signal feed-forward, i.e. what will happen if these buttons are pressed.

Furthermore, a black box have been added to the legend. The legend is what allows the user to filter out regions, and also what links the two plots together. As a legend in itself does not necessarily signify affordance, the black box have been added to somewhat resemblance a button, and to draw attention to it.

Finally, the package used to construct these visualizations, called Plotly [2], automatically adds an interactive panel signifying different options, see Figure 3. This panel utilizes iconic signs, such as the camera to save a picture or the loop to zoom. Each sign indicates feed forward when hovering it with the mouse, with a textbook describing what will happen if clicking the sign.

Figure 3: The built in interactive Plotly panel.

Indicators exists in these interactive elements, signalling what state the system is currently in. The drop down option and the Plotly panel signs will be highlighted if selected, and the filtered out regions will be greyed out in the legend.

With all this taken into account, the vis idiom contains the interaction techniques hovering, encoding, zoom and sorting.

Strengths & Weaknesses

Now some of the strengths and weaknesses of this visualization idiom, will be investigated. Starting with the strengths, the combination of these two bar plots effectively show the majority of the data. Even though the explanatory attributes are the regional averages, and not each single value for each country, it still functions as a good summarization of the data. Furthermore, the top plot supports an effective way to compare regions by filtering out uninteresting ones using the legend. It is also easy to sort the bars after rank, making targeting features and outliers more obvious.

Moving on to the weaknesses, this plot does not support the action of lookup very well. If the user knows the target, he/she will have to zoom in on the appropriate region and find it manually, as the location is unknown. Another weakness for this plot, is the option to target the explanatory attributes and look for any correlation between these and the happiness score. In this plot, these attributes better affords comparison between the regional average values, and not to the happiness score.

In the following sections, two new idioms will be shown and explained. The idea behind these idioms is that they each eliminate one of these weaknesses by changing the way the data is shown. The design choices for these idioms will also be explained, however it may not be as in-depth as for the bar plot.

2.2 Bubble Idiom

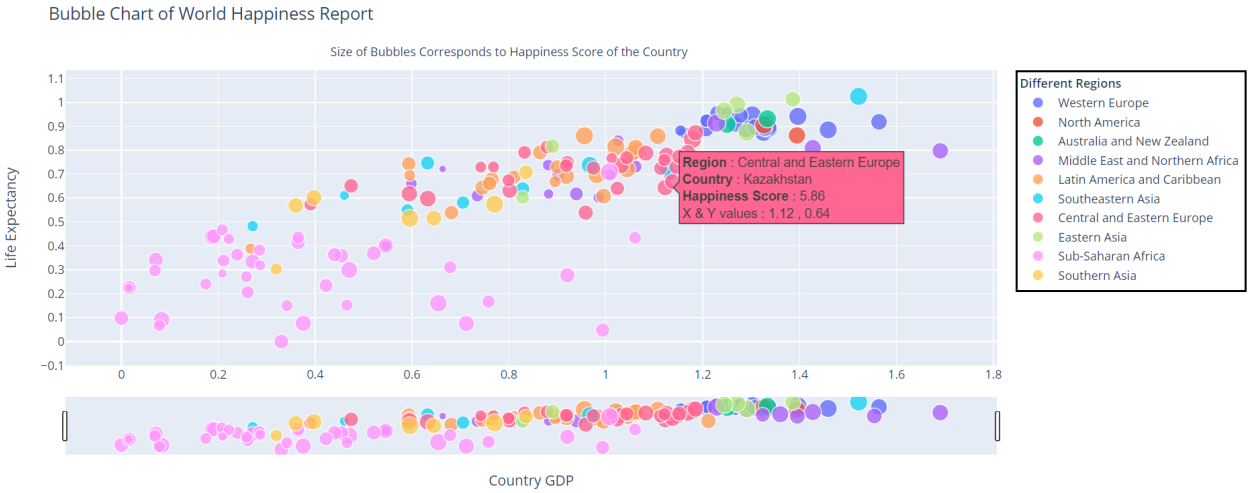

Figure 4: Bubble chart of World Happiness Report.

What, Why, How

This bubble idiom encodes almost the same data as the bar idiom. The only difference is the rank of country is not included in this bubble chart.

The bubble chart utilizes points markers to represent each country, with the vertical and horizontal spatial positions of the point markers to encode the quantitative attributes GDP and life expectancy. The color hue channel is used to encode the categorical attribute region and the size channel is used to encode the third quantitative attribute happiness score. This is a quite effective way of expressing channels, as both position on a common scale, spatial region, color hue and area of markers rank generally high in effectiveness [1, p. 94].

As seen by this breakdown, four different attributes have been encoded in just a single plot. Compare this to the bar plot where two linked bar plots was needed to show the same information. This makes one of the bubble charts strengths very evident.

The main action in this bubble idiom is also for the user to discover new information. The idiom supports all the search actions, however like the bar idiom, the lookup function is not distinctly clear. There is no logical way to know the location of the target. Here the user either has to filter out data by using the legend, or have an idea where the country would be placed in the plot, by having a general idea of the countries GDP or life expectancy.

Strengths & Weaknesses

As for this visualization idioms strengths, the bubble chart better encodes the explanatory attributes into the main visualization. This in turn, better affords the user to target the correlation between happiness score and GDP/life expectancy. If one wanted to further expand upon this bubble chart, an interactive option could be added to change what explanatory variables are shown on the axes, thereby also altering the view. However, it also has some

weaknesses. There is no way to sort the countries by rank, making comparison of happiness scores between countries or regions difficult. It can also be difficult to directly read the link between a bubbles size and its corresponding happiness score.

Signifiers & Affordance
To combat some of these weaknesses some interactive elements have been added to the idiom. For comparing, the user can again use the legend to filter out uninteresting regions, thereby increasing the visual clarity of the remaining markers. In this idiom a black box, meant to represent a button, has again been added to the legend, signifying that the legends items afford clicking. A range slider has also been added to the bottom of the plot. This is so the user can easily keep an overview of the data, while zoomed in. As for the difficulty of reading the happiness score, this information has been added as a functional affordance, as it pops up in the hover box, when placing the mouse over the bubbles.

2.3 Spatial Idiom

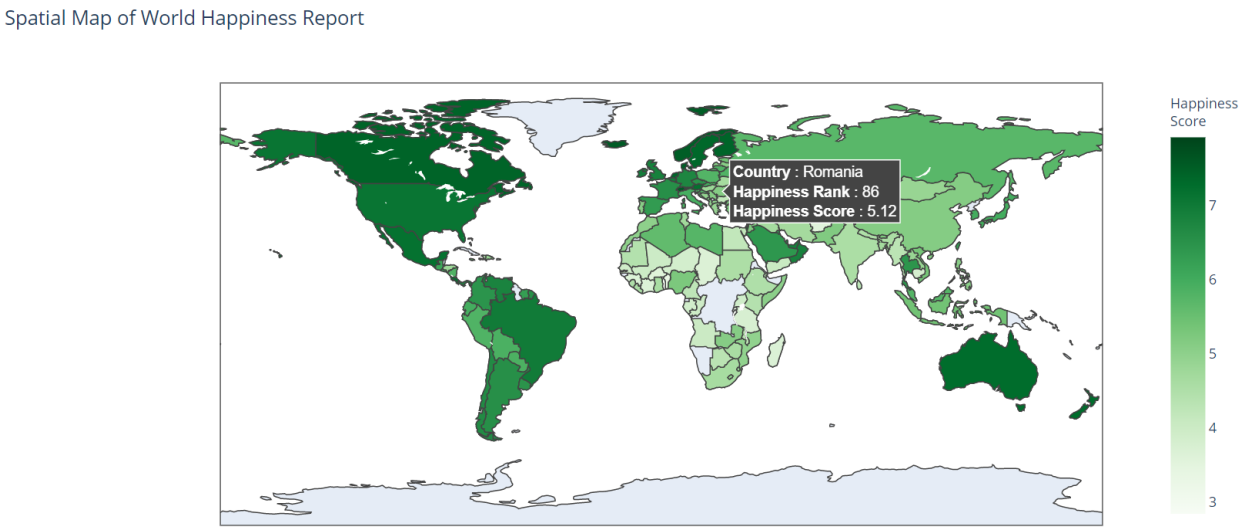

Figure 5: Choropleth map of World Happiness Report.

Since the dataset contains the spatial attribute of country, the option to use a spatial idiom such as a choropleth map exists. If choosing to use the spatial data to guide the layout, a lot of encoding options is removed, since the position element of an idiom for marks and channels is completely taken up by the map. Therefore, it is important to consider if the spatial relation of the data are the primary focus of the defined tasks.

In this choropleth map the country attribute defines the geometry of the map, and the happiness score is used as the quantitative attribute. This basically fills the possible options for encoding data in a choropleth map. However, the option to add extra attributes to the hover information can be utilized, as seen in Figure 5 where the happiness rank can be seen in the hover box. These limitations is perhaps also the greatest weakness of a geo-visualizations.

The spatial idiom supports the search action very well, even the lookup search target (Which neither the bar idiom nor the bubble idiom supported very well), as most users would know the location of their target on a world map. This is the biggest strength of using geo-visualizations.

The color scale used to encode the continuous quantitative attribute happiness score utilizes the magnitude channel of luminance. The darker the green color gets, the higher the countries happiness score is. This supports the action of comparing the different scores for countries.

3 User Evaluation

As briefly talked about in the introduction many different ways of constructing a valid design exists for the vis designer. Therefore, it can be useful to validate ones choices made during the design phase, i.e. if ones "What, Why, How" decisions actual make sense in the context of the given tasks and visualizations. To do this Tanzara Munzer refers to a validation framework, see Figure 6. .

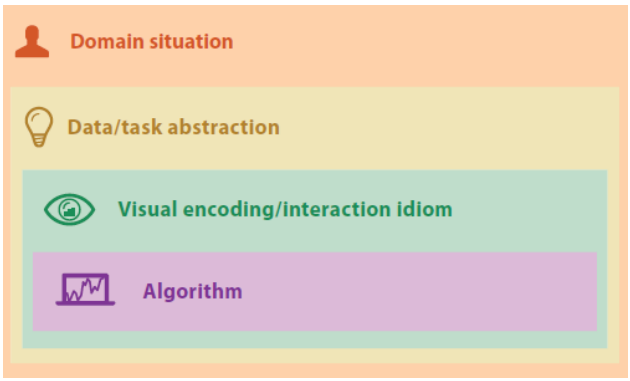

Figure 6: The validation framework as described by T. Munzer [1, Chapter 4]

.

In this chapter, a user test will be conducted, with the focus being on step 2 and 3 of the framework. Step 2 focuses on understanding if the data shown supports the given tasks. Step 3 focuses on the way this data is shown, i.e. can the user effectively and with reasonable effort extract the desired information.

In this test, the user will be given the three tasks from Chapter 1.2. Furthermore, supplementing tasks will also be given. The purpose of these is to test some of the interactive elements, as well as give the user a feel for how to use interactivity in the visualizations. The user will get the chance to answer these task by first using the bar idiom, and then the option to to use the two alternative idioms for support will be given.

During the evaluation the user will also be asked to explain what information they are shown in the visualization, and explain their thought process when clicking interactive elements or changing the view. To capture all relevant events and information, the test will be recorded.

This makes it possible to go through the user test multiple test, in case something important was missed during the real time test.

3.1 Feedback

The users feedback and general observations done by the designer can be seen below. The points have been divided by the given tasks.

Can you find the average happiness score?
The intended action for this task was for the user to identify the derived attribute. Furthermore, this task required the user to look around in the visualization, leading the person to discover some of the interactive functions such as drop downs and zoom. The task was quite easily solved.

Who would you say are the top five happiest countries in the world?
Here the intended action is for the user to identify and locate these top countries by changing the encoding of the happiness score by ordering by rank. With the knowledge of drop downs from the previous task the user quickly sees the second drop down which function is to sort by rank. However, when zooming in on the top countries, the user accidentally zooms in to far, and does not know how to zoom out again. The Plotly panel actually has a reset axes option, but this was not known for the user.

At what rank does France lie on?
Here the user needed to identify one country, but first she needed to locate it. The idea was to use the reduce option, by filtering out unnecessary regions from the view. However, the user had no knowledge of the functionality of the legend, so the person ended up zooming in on the top countries where they thought France would be located. Also in the bar idiom, the rank attribute was not actually encoded anywhere, which was a mistake in the design. Using the spatial idiom instead, this task was much easier. As the user knew the target and location they could easily navigate and look up the country and rank which was actually encoded in the hover box.

Does the region of the countries seem to have any relevance on their happiness?
For this task, the user had to compare different regions and look for any dependency between happiness score and region. Using the bar idiom, this task was a bit difficult. The general answer was no, however the user did acknowledge that the Sub-Saharan African countries was somewhat lower than the other regions. The user was then directed to use the bubble idiom to answer this task instead. Here, the user had to identify clusters of regions and try to look for trends in the size of the bubbles. However, this did not prove easy, as it was very difficult for the user to distinguish between the sizes of the bubbles.

Can you see any factors besides region, that could contribute toward or explain the countries happiness score?
The purpose of this task, was to get feedback on how the explanatory attributes were shown, and if their meaning was clear. Again the user had to compare different attributes and see if any dependency existed. From the bar idiom, it was clear that the GDP, family and health attributes was the three highest attributes. Using the bubble idiom, where only two of the

explanatory attributes were encoded, the user had to identify if the spatial position of the bubbles had any correlation with the happiness score. Again, the score was hard to identify for the user. However, she did notice that generally the western countries was located in the top right corner.

General Observations
Two important general observations were done by the designer during testing. The first being that the legend did not signify well enough that it afforded clicking. This is not good, as it was meant to be a important tool for some of the tasks. Secondly, the explanatory attributes actual meaning was confusing, and the units on the axes was not specified (because there is no unit).

3.2 Second Iteration

Based on the most important feedback from the user test, a second iteration of the idioms could have been constructed. For the second iteration the focus should definitely be on better explaining and encoding the explanatory attributes, making the legend and Plotly panel popout more, properly encode the rank in the bar idiom, and perhaps find a way to combine all the three idioms into one visualization. This way, each idioms strengths could compliment the others weaknesses.

However, due to time constraints this will not be done in this report.

4 Ethical Concerns

In this chapter some common ethical concerns regarding data visualization, and the context of which they appear for the constructed visualizations in this report, will be investigated. The risen concerns are based on Michael Corrells paper "Ethical Dimensions of Visualization Research" [3].

4.1 Clarity of Visualizations

As a visualization designer it is important to think about the ethical implications of ones visualization idioms. Well designed idioms can be a powerful communication and persuasion tool, but less well designed ones can just as easily confuse or mislead the users. Visualizations are not necessarily intuitive, neutral or objective (biases in the visualization designer exists) nor does the use of graphics reduce the need for further explanations. This is why properly contemplating and explaining the design choices for each idiom, as well as considering the possible misunderstandings in ones idiom, is very important.

In the context of this reports visualizations, the biggest offender of confusing attributes is perhaps the explanatory variables. Without reading the report or the Kaggle/World Happiness Report description of the dataset, the context of these attributes can be hard to decipher from the visualizations. This makes these attribute obvious for misunderstandings. For example, the GDP could be interpreted as quantitative objective numbers that describe

the countries actual GDP, and not the possible explanation of the countries GDP's influence on the happiness score.

Another considerations for these happiness visualizations, and the World Happiness Report in general, is the risk of exploitation of the findings. The intent of the report is generally to explore how the world evolves and how the our happiness evolves with it, with the focus being moving towards a more sustainable future where well-being is in a primary element (at least for the 2015 report). However, just looking at the visualizations could lead to poor conclusions on the worlds more "unhappy" regions. Perhaps the western readers would conclude that the Sub-Saharan African or Middle Eastern countries are generally a non-safe unhappy poor place, further alienating and secluding them from the rest of the world.

This can be linked to a general problem with data. Typically data is somewhat abstract, and the human aspect of the quantitative numbers are often lost. It can be hard to look behind the visualizations and see the actual humans they impact or report. Some suggest, such as Correll, that anthropomorphizing the data is the way forward when dealing with sensitive visualizations, as to not loose the human aspect of human data.

4.2 Neutral/Objective Data

Initial collection of data also has ethical considerations. Data is not a natural occurring phenomenon in the sense that the world does not quantify, collect and store data for us. In most cases data collection is a political act with some agenda, which is usually hidden for the visualization designer.

In the case of the World Happiness Report, the data sources and variable definitions are defined in their statistical appendix [4]. The happiness scores that they report comes from the Gallup World Poll and is based on this question:

> *"Please imagine a ladder, with steps numbered from 0 at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?"* - [4, p. 1]

Defining a nations happiness from this question, surely makes the happiness score a subjective measurement. Moreover, a couple of the explanatory attributes also comes from questions such as these. Based on this, it could be important to emphasise for the vis user that these scores are not objective truths and measurements. With Gallup being an American analytics company [5], the question can be raised if they are not themselves biased in their own western thinking, perhaps further making the poll questions and the associated data non objective.

Bibliography

[1] Tamara Munzner. *Visualization Analysis & Design*. 7th ed. CRC Press, 2014. ISBN: 978-1-4665-0893-4.

[2] Plotly Technologies Inc. *Collaborative Data Science*. 2021. URL: https://plotly.com/ (visited on 01/25/2021).

[3] Michael Correll. *Ethical Dimensions of Visualization Research*. 2018. URL: https://arxiv.org/pdf/1811.07271.pdf (visited on 01/25/2021).

[4] Sustainable Development Solutions Network. *Statistical Appendix for Chapter 2*. 2015. URL: https://s3.amazonaws.com/happiness-report/2015/StatisticalAppendixWHR3-April-16-2015.pdf (visited on 01/25/2021).

[5] Gallup Inc. *WHO WE ARE*. 2021. URL: https://www.gallup.com/corporate/212381/who-we-are.aspx (visited on 01/25/2021).