EXCEL PROJECT 1: SHOPEE SALE STRATEGY

THE FIRST STEP IN STRATEGY FORMULATION

Chinh X. Mai

Last update: June 16, 2022, v.01

CONTENTS

1. CASE DESCRIPTION AND PERSONAS ESTABLISHMENT

1.1. CASE DESCRIPTION

Hal is the co-founder of a startup that is aiming to penetrate the Vietnamese market. His company is based in California USA and specializes in providing customized furniture and indoor appliances. The other founder, Huy, is a Vietnamese whose family has a history in assembling and retailing furniture. Huy is the strategy manager of the startup and the one who suggests this expansion. However, Hal is only convinced by statistics and so hires me to carry out the analysis using secondary data provided by a data vendor in Vietnam.

Hal does not know much about the Vietnamese retailing market, so he needs quick information about the potential sales and demand of the market. He possesses a basic understanding of statistics and likes to work with tables and charts. He needs a dashboard that presents important information about revenue, factors affecting revenue, trends in the market and product categories that have the highest revenue. This information is useful for his further discussion with Huy on their tactical decisions and forming their strategies to penetrate the retailing market in Vietnam.

I am given the access to the service of the data vendor in Vietnam. My tasks include extracting data from the vendor, cleaning the data, carrying out the analysis, and designing the dashboard to present my findings. Both Huy and Hal are young businessmen with good vision who focus more on the details and important information rather than the formality. However, it is still necessary to fulfil the aesthetic requirements of the dashboard so that important information is more efficiently delivered to the audience.

1.2. PERSONAS

Persona 1: Hal

"Money does not grow on trees; strategy does not come from thin air. Thoughtful consideration is where a great journey starts"

Being a decision maker in a startup, Hal must make many decisions under high pressure. He is sensitive to changes and a careful thinker. Every decision he made was based on facts and statistics together with thorough discussions with his colleagues.

Role: Executive Officer

Organization: PKT international retailer

Goals:

* Get better understanding about the Vietnamese market

* Evaluate the market potential using sales statistics like revenue

* Know which factors impact the total revenue of a product in the long term

Challenges and Needs:

* Quick facts and figures for thought initiation and discussion

* Statistics is necessary for a good decision, too many statistics can lead to a bad decision

* He likes to focus on the most important points and only discuss the most meaningful findings

2. STORYTELLING DESIGN CHECKLIST

2.1. BACKGROUND OF THE CASE

Main tasks:

- Extract data from the data vendor
- Clean data
- Carry out data analysis to answer certain questions
- Design dashboard

Primary requirements:

- A dashboard for quick information and thought initiation
- Focus on top-level facts with strongly backed-up statistics
- Short enough for a page printout

2.2. AUDIENCE PERSONAS

Executive Officer Hal:

- Needs: quick but accurate information, clear presentation with focus on important points, detection of important factors
- Characteristics: careful thinker, sensitive to changes, prefer data-driven decision making

2.3. CHECKLIST

WHO: Hal is a stakeholder who is thoughtful and pay attention to details, while another audience is Huy who also joins the discussion with Hal with presented statistics. Both have a basic understanding of statistics and can interpret presented charts and tables.

WHAT: The data will be extracted from a data vendor in Vietnam, from the description of how the data is collected, it can be said that the source is quite reliable. Nothing can be said yet about the quality of this data. The data is up to date as the vendor provides constantly updated data on their platform.

WHY: Exploratory analysis on factors that affect sales revenue as well as indicators that can be used to estimate the potential of the market demand.

HOW: Constructing a dashboard to show quick information using charts and tables, which is short enough for a one-page printout
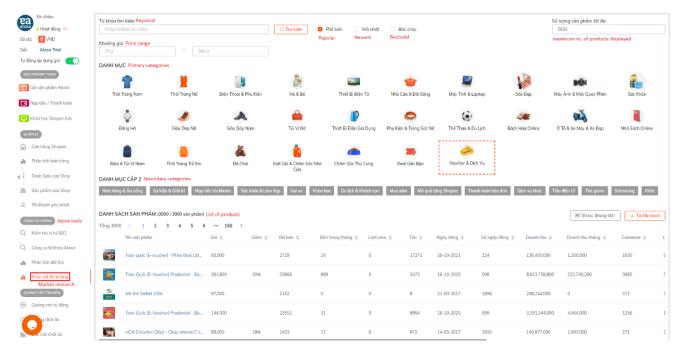
3. DATA EXTRACTION AND CLEANING

3.1. DATA EXTRACTION

Platform description:

Atosa (https://app.atosa.asia/) is a company that provides many marketing automation solutions for Shopee, an ecommerce platform in Vietnam. These solutions include Atosa Shopee Ads, Atosa Keyword Tools, and Atosa Agency. When having subscribed to these services, the user will be provided with many tools, one of which is the Market research tool, whose user interface is presented below

Since Shopee is an e-commerce platform in Vietnam, the user interface (UI) is in Vietnamese, so some texts have been added to the illustration to explain important elements of the UI. The UI is quite intuitive and there are many options for filtering and ordering data.

The basic function of this tool is to provide a list of products following given preferences of its user. The user can filter the products by choosing a set of categories (primary, secondary, and even tertiary categories) and order the results by popularity, novelty, or bestseller. At maximum, up to 3,000 products can be loaded from the database and displayed in the results list. However, this might not be all the available products following a chosen set of conditions for fetching.

Picture 1: The user interface of the platform with key elements translated for a more comprehensive presentation

Data description

The data has many features whose names are displayed in Vietnamese. These features are (after translated if necessary):

- Product name (Tên sản phẩm)
- Price (Giá)
- Discount (Giảm)
- Sold quantity (Đã bán)
- Last month sold quantity (Bán trong tháng)
- Currently in stock (Tồn)
- Posted date (Ngày đăng)
- Posted duration in days (Số ngày đăng)
- Total Revenue (Doanh Thu)
- Last month revenue (Doanh thu tháng)
- Like
- 5 Star
- Rating
- 5 Star
- 4 Star
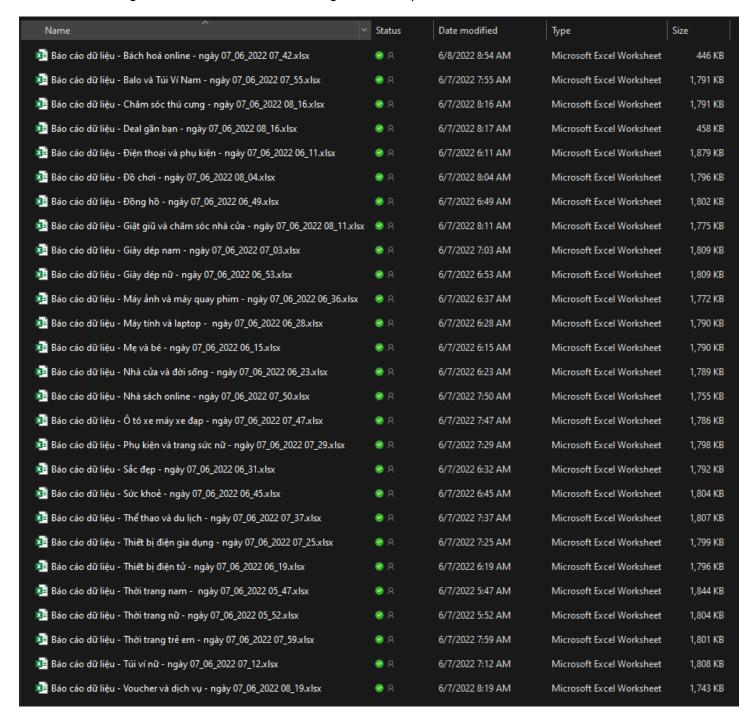- 3 Star
- 2 Star
- 1 Star

Fetched data can be saved as an Excel file. However, the downloaded data does not include the last 5 features (5 Star, 4 Star, 3 Star, 2 Star, and 1 Star), which are necessary for understanding the rating distribution. This data is updated constantly so it might change depending on the collected date.

The given data here is collected on June 7, 2022, following a set of certain fetching and ordering conditions. These conditions are as follows

- Ordering by popularity
- Maximum 3,000 products are displayed
- Data is fetched using only primary categories (so there are 27 categories in total)

In total there are maximally 81,000 observations in the final consolidated data before cleaning.
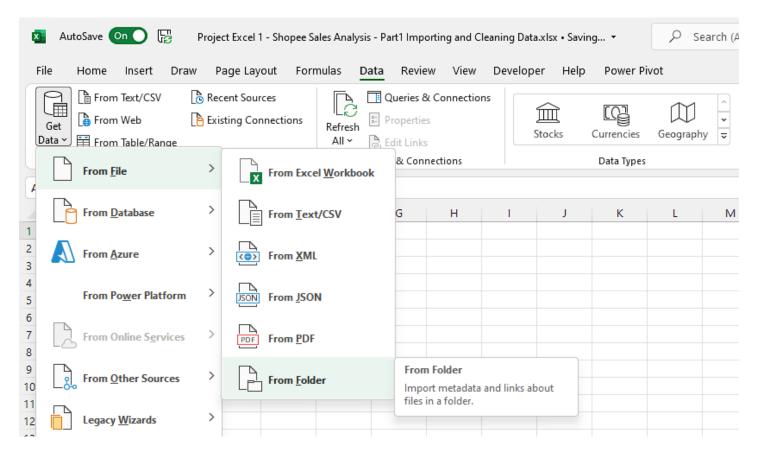
Only 3,000 products of each category are extracted and saved in an Excel file whose name follows the structure "[Data report] – [Category name] – [Date information].xlsx". Since the data does not contain information about its category, this structure makes it more convenient to generate information about data categories later. A picture of the raw data is shown below
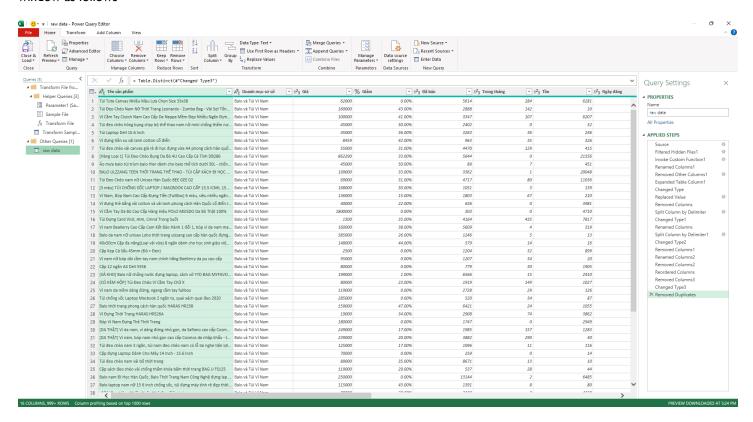

Picture 2: All the raw data files before importing and cleaning

3.2. DATA IMPORTING AND CLEANING

The importing and cleaning process involves in utilizing both the Power Query and Excel functions. Firstly, raw data is imported and put together into a table using Power Query. From the tab Data on the tools ribbon, select Get Data > From File > From Folder to import all the raw data files from their containing folder

This function allows for a mass import of all the data files in one folder, the result is then presented in the Power Query Editor window as follows
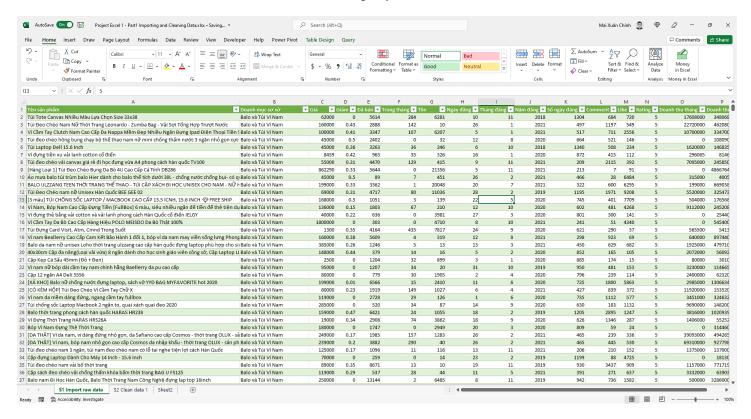

Using this tool, many cleaning steps have been taken. They are summarized on the right side of this windows in the APPLIED STEPS field. These steps include

- Rearranging column order for a more convenient analysis later

- Decomposing the column Posted Date into Posted Day, Posted Month, and Posted Year to formulate different time clusters which can be used to investigate how the product posting activity changes over time
- Deleting unnecessary columns such as the index of each raw data file, No. View (no view data has been recorded), and the column containing the Shopee link to the respective product
- Removing duplicates if existed

The data is then loaded to an Excel sheet where further cleaning steps are carried out:
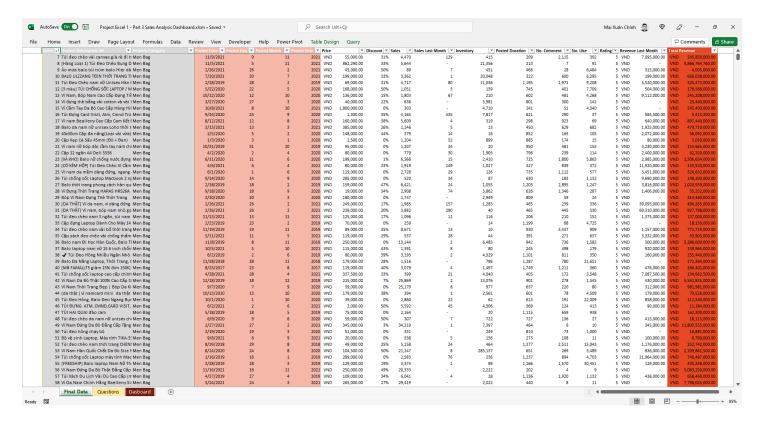

These steps include

- Translating column names to English
- Translating the column `Primary Category` into English using the Sort & Filter > Advanced function in the Data tab to create the list of unique values then translate this list accordingly, after that using the XLOOKUP() function to look up the corresponding English translation of the listed categories in Vietnamese
- Adding the column Index
- Adding the column posted date using the DATE() function

The data is then properly formatted and put into a table for further manipulations. Then we have the final data, which is presented in the picture below. The final data has eighteen columns, including

- Index
- Product Description VIE
- Primary Category
- Posted Date
- Posted Day
- Posted Month
- Posted Year
- Price
- Discount
- Sales
- Sales Last Month
- Inventory

- Posted Duration
- No. Comment
- No. Like
- Rating
- Revenue Last Month
- Total Revenue

These columns are categorized into separate groups featured by their colors for the presentation purpose. For this analysis, I want to focus on the Total Revenue, so it is highlighted by orange.

4. ANALYSIS DETAILS

Before start doing any analysis, it is important to define the questions the analysis needs to answer. For this analysis, the case description has stated the needs of Hal, but this information is still too general and needs to be articulated more clearly. The main points can be summarized shortly as follows:

1. Which factors in Price, Discount, Number of Comments, Number of Likes, and Rating have the significant effect on the Total Revenue of products posted on the E-Commerce platform Shopee until June 7, 2022?
2. How is Total Revenue distributed among Sales Categories of the products posted on the E-Commerce platform Shopee until June 7, 2022?
3. How does the number of posted products on Shopee until June 7, 2022, changes throughout a year? And from year to year?
4. For each Category of the products posted on Shopee until June 7, 2022, what are the Total Revenue, Last Month Revenue, and Distribution of Rating?

In this analysis, Sales statistics will be ignored as its information has already been conveyed by Total Revenue. It should be noted that the data set does not contain data for sales or revenue over time, but it only has the date when a product is posted. This data can be aggregated to illustrate the pattern of posting product on Shopee over time.

The first question can be answered by running a regression model of Total Revenue on Price, Discount, Number of Comments, Number of Likes, and Rating. This is a simple multiple regression model that has the following form
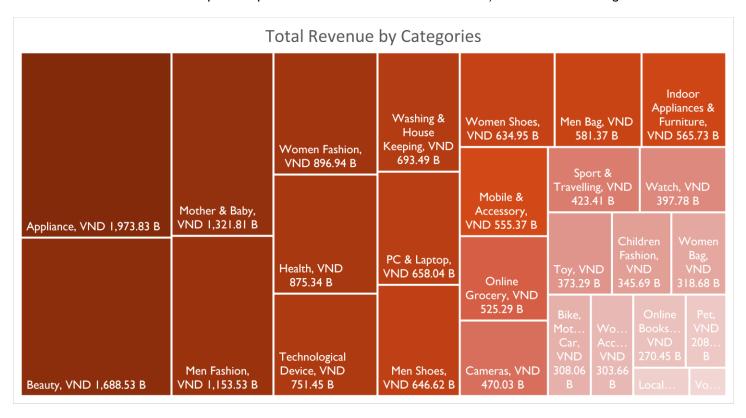
$$\text{Total Revenue} = \beta_0 + \beta_1 \times \text{Price} + \beta_2 \times \text{Discount} + \beta_3 \times \text{No. Comment} + \beta_4 \times \text{No. Like} + \beta_5 \times \text{Rating}$$

Where $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ and β_5 are respectively the intercept, effects of Price, Discount, No. Comment, No. Likes, and Rating. Since this multiple regression is still quite simple, it can be handled well by Excel. Using the Data Analysis function in the Data ribbon, the regression result is summarized in the following table

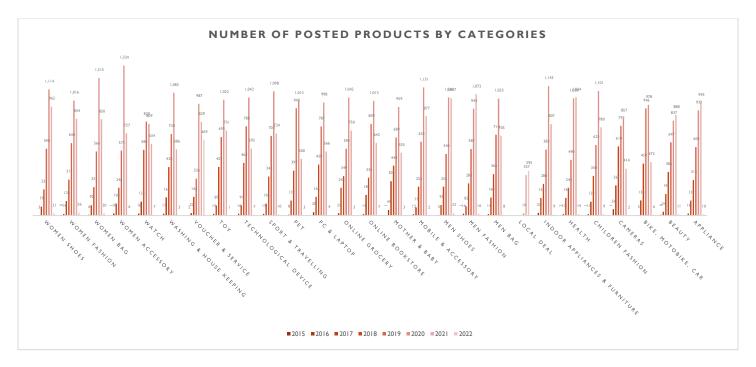| Variable	Coefficients	Standard Error	t-Stat	P-value	Lower 95%	Upper 95%
Intercept	(36,688,510.80)	19,225,143.68	(1.91)	0.06	(74,369,679.06)	992,657.46
Price	46.98	2.26	20.81	0.00	42.56	51.41
Discount	26,464,359.98	17,665,273.79	1.50	0.13	(8,159,472.48)	61,088,192.44
No. Comment	132,607.95	4,482.53	29.58	0.00	123,822.21	141,393.68
No. Like	6,954.91	709.93	9.80	0.00	5,563.45	8,346.36
Rating	35,466,020.25	3,888,292.03	9.12	0.00	27,844,990.80	43,087,049.69

It can be seen from the table that the effect of Price, No. Comment, No. Like and Rating are all statistically significant since their P-value is almost zero. However, the effect of discount is not significant as its P-value is larger than 0.1.

The distribution of Total Revenue among Sales Categories of the posted products can be presented by a tree map. Using a pivot table to aggregate data and construct the map from this data (actually, the version of Excel being used in this analysis still does not allow for the direct construction of a tree map from a pivot table, so a walkaround has been used), we have the following result
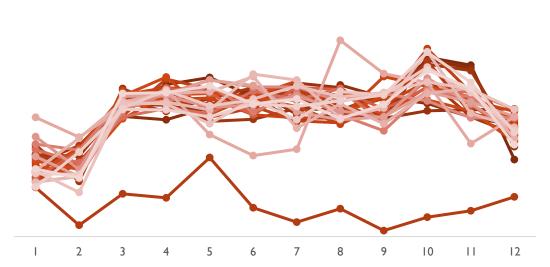

Total Revenue by Categories

Tree map is a proper presentation for this type of information besides bar chart. In this case, I would prefer tree map to bar chart because it shows which category has the highest Total Revenue clearly (as Western People usually read from left to right, and top to bottom) and a sense of relative magnitude between categories, while still saving lots of space due to the more compact size of its illustration. Furthermore, change in color is also used to illustrate this difference in size. From this illustration, Appliance, Beauty, and Mother & Baby are the categories that have the highest Total Revenue.

The posting pattern of products in all categories can be presented by two charts. Due to the substantial number of categories, it is more convenient to separate a supposed-to-be-one illustration into multiple charts so that they are more comfortable to look at and interpret. In this analysis, I choose to visualize the number of posted products aggregated by category, by month, and by year. This offers an insight into the posting pattern throughout the year and from year to year. Moreover, two combinations of illustration are used to present the pros can cons of each means of illustration while allowing me to utilize their strengths as well as minimize their shortcomings. With a slicer, these visualizations can be simplified to compare the patterns between two or more categories or only show the pattern of one category.

NUMBER OF POSTED PRODUCTS BY CATEGORIES

The first chart shows the number of posted products of each category from year to year. Cluster bar chart is chosen to present this information as it shows all categories clearly as well as how each category changes over time. I want to show the magnitude of the number of posted products of each category in each year, so the bar is the best option for presenting this information. It might look a bit messy at first, but the slicer will make the chart look tidier later.

Total Posted Product by Month

The second chart shows how the number of posted products of each category changes throughout the year. Line chart is different from bar chart in such a way that it focuses more on the changes rather than the magnitude within a category. When considering the number of posted products of a category within a year, changes in the number are more important because it allows the user to choose the right time to post their product so that he or she could have tactical advantaged depending on his or her sales strategy. This is different from the number of posted products of a category from year to year as it offers a rough estimate of the market size. These different charts serve different purposes and when combined with the right slicers, they provide very insightful information about the categories of choice.

The final question can be answered easily by a combination of pivot tables. Total Revenue, Last Month Revenue, and Distribution of Rating of each category can be aggregated easily using their corresponding pivot tables. The pivot tables do not need to show the categories because they can be linked to the slicers that are used to control the tables in the previous question. When the slicer is

used, the aggregated statistics will be updated accordingly in the pivot tables of Total Revenue, Last Month Revenue, and Distribution of Rating.
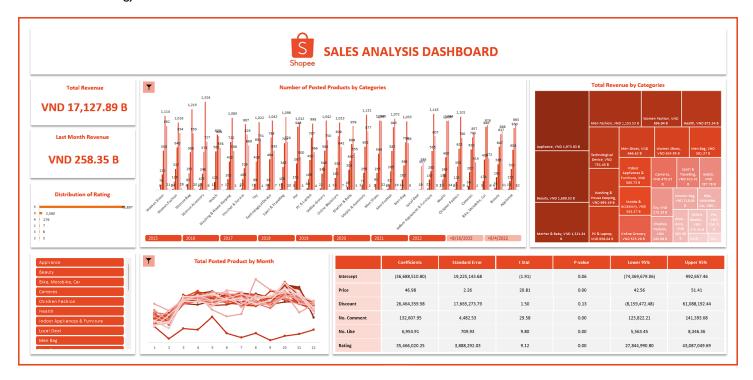
Sum of Total Revenue

VND	17,127.89 B

Sum of Revenue Last Month

VND	258.35 B

Row Labels	Distribution of Rating
2	2
3	6
1	7
4	279
0	2,580
5	75,897
Grand Total	**78,771**

The slicer is an important part of the Excel pivot table because it provides a way to interact with the information being presented, making the visualization become dynamic rather than just static figures.

These charts are then brought together to construct a dashboard to fulfill the needs of Hal, so that he can investigate further to formulate his strategy and tactical choices with the available data.

Overall, this dashboard provides information for the user to estimate the magnitude of market for each category, and important analytical information for forming strategy and making tactical decisions if he or she wants to penetrate the market. With slicers to filter categories and years, the dashboard allows the user to investigate the overall market, the market of each category as well as compare them. Regression table also provides information about the factors affecting Total Revenue, while line chart presents the patterns of how sellers in each category post their products on Shopee throughout the year. Details of the design choices are discussed in the next section of this documentation.

5. DESIGN CHOICES

The dashboard has many elements including information card, table, line chart, vertical bar chart, and tree map. The reasons for these choices are discussed in the previous section. In this dashboard, the alignment of objects as the first pre-attentive attribute. The dashboard has two rows, the upper is for information about the revenue, and the lower is for information about the products posting tactics. The slicers, which are used to interact with the dashboard, are placed on the left side of the dashboard. Since the Western people usually read from top to bottom, left to right, the dashboard user will encounter with important elements and information first following his or her natural eye movement.

The color choice is only to make the dashboard have the same color tone with the icon of Shopee. I did not intend to make it a pre-attentive attribute, so I used the provided color palette of Excel. The palette did a decent job on providing a consistent presentation of the color tone and it seems to give the impression that objects of the darken orange have higher values. However, color is not an effective pre-attentive attribute here because the dashboard does not need to show which values are more important than the others. Moreover, there is too much information to be convey and they seem to be equally important.

Some of the Gestalt principles are employed in the design choices to reduce mental clutter when processing the visualizations, they include:

- Continuity is used to show the changes in the data as the category change from one to another
- Enclosure is used to separate information presented by different charts

As the intended audiences are familiar with the subject and require arguments to be backed up by statistics, they will focus more on the quick, well-articulated overall trends rather than details. As they have no problem with viewing the visualization, these charts are well designed to focus their attention on important points while still providing enough information in case when they need to dive deeper into the statistics and correlation among the presented illustrations.

6. FINAL VISUALIZATION ANALYSIS

7. RELATED INFORMATION AND DATA

7.1. DATASET

The dataset is given together with the dashboard, the raw data is put in a separate folder.

7.2. CHANGELOG

18/06/2022: v0.1 first establishment of the dashboard

APPENDIX AND REFERENCES

REFERENCES