Data science Assignment 1: visual analytics

Marts Niewandt s2624893
Daniël Zee s2063131

March 14th 2024

1 Introduction

For this assignment were tasked with creating a dashboard for the fictional software business "Emerald-IT". This company owns the popular app "Complete Reference for Dungeons and Dragons 5", used for playing Dungeons and Dragons by people worldwide. We were provided with data from the second half of 2021 regarding the sales of the app, the ratings and feedback given by the users and the stability of the app. Our job was to create an interactive web-based dashboard which can be used by management to understand and interpret the data of the app. We will explain the functionality of all the visualisations in our dashboard and why various design choices were made to achieve maximum insight into the data. Finally we will answer some strategic question on the app by using our dashboard.

2 Visualization

2.1 Sales Volume

For the sales volume the two sale options are given in a stacked bar plot to show the total sales per week and the difference in *premium* vs *unlockedcharacter-manager* sales. Two contrasting colors were chosen to distinguish the two sales options to help the user to clearly differentiate between them. On hover of a sales option the exact sales volume of that week for the current sales option is shown. We have chosen a weekly time interval because it gives us more insight than grouping by months and is easier to interpret for the user than using daily data.
A line plot was chosen for the weekly generated revenue to easily detect trends in the data. A rising line implies increased revenue and a decreasing line is decreasing revenue, compared to the previous week. Because there are some weeks where the revenue is much higher that the others, we have chosen to set the default zoom to a setting that clearly shows most data points. If the user want to view the values of the outlier, they can use the provided zoom and pan tools on the right of the figure.

2.2 Attribute Segmentation and Filtering

For this segment it is important to make the ratio clear between the two in-app purchase options. As such a pie chart is given. Since there are only two options for sales, a pie chart makes the ratio clear. The same colors are chosen for the *Sku Id* variables as in *Sales Volume* for consistency. On hover the exact sales and percentage of that sales option are shown.
A bar chart is also given to show the amount of sales per sales option.
Amount of sales per sales option are obtained by counting the amount of sales with that sales option in the provided dataset.
Furthermore a bar chart is given to see how sales behave on different days of the week. A pie chart would not have made it clear which days are the busiest, as such a bar chart was chosen. Bar chart makes it possible to see the difference between all variables used, instead of only seeing how big the ratio of sales of the total sales occur on a certain weekday.

2.3 Ratings vs. Stability

A scatterplot shows the *Average Daily Crashes* compared to the *Daily Average Rating*, again grouped by week. This is useful to detect correlations between these two measures. Color is given based on the *Satisfaction Index* and the point of the given week is colored as such.
The individual points can be selected and on being selected the selected week is highlighted in both charts. On hovering on data points in either chart the exact week is given and the exact values used to draw that point on that position. The user can pan the view and zoom in and out to better inspect closely situated data points.
The second chart is a line graph over the *Satisfaction Index* per week with two points also giving the *Crash Index* and *Rating Index* respectively. The line graph depicts the evolution of the satisfaction of the customers. The *Satisfaction Index* is a key performance indicator (KPI) we made up to measure the general satisfaction of the app users by combining the *Average Daily Crashes* and *Daily Average Rating* into a single number using the folllowing formulas:

$$\text{Crash Index} : c = 1 - \frac{\text{Daily Crashes}}{100} \tag{1}$$

$$\text{Rating Index} : r = \frac{\text{Daily Average Rating*}}{5} \tag{2}$$

* Total Average Rating for weeks without ratings.

$$\text{Satisfaction Index} : s = c - \frac{(0.5 - r)(c + r)}{1 + c + r} \tag{3}$$

The *Satisfaction Index* ranges between 0 and 1.33 and has high values when the app users are satisfied. The formula rewards a high combined *Crash Index* and *Rating Index* and punishes a low combined *Crash Index* and *Rating Index*.

2.4 Graphical Development

For the geographical maps there are four different maps, each with a selection option for the month to show the data from. The user can change the shown map using the tabs. We can pan and zoom the map using the tools on the right to get a better view of specific countries or regions. The map view positions for every month in a single category are linked, allowing the user to switch between months without the view resetting. This allows for easy comparisons between months.
The *Geographical Total Average Rating* are shown in a color range that starts at a light yellow, transitions to blue and finally ends in dark purple. This mapping was chosen to clearly differentiate between the three general ranges for the ratings: bad (yellow), neutral (blue), good (purple). Because the colors have many steps we can also differentiate smaller differences well.
Geographical Daily Average tracks the average daily review rating per month. Most countries are colored grey, as not all months may have data for a given country. With this map it is possible to see the trend in review scores for a particular country. It is important to note that we do not know the number of reviews this average is based on.
The *Geographical Sales Volume* makes it clear where most sales are from. This result is in absolute sales numbers that month. Grey countries had no sales and on hover show 'NaN'. The color mapping starts at a light yellow and becomes darker, ending at a dark red. A darker color maps to a higher number of sales, as darker is often associated with more important. Because the sales number of the United States are so much higher than the other countries, we have chosen to end the color range at 45 sales in order to be able to differentiate the other countries.
The last map, *Geographical Sales Difference*, shows the absolute difference in sales per country compared to the previous month. Green is good, as it means more sales compared to last month and red is bad, as that means there were less sales per month. This scales in how much this difference is, with lighter colors meaning less change in sales than the more saturated colors. These colors where chosen because people intuitively associate green with good and red with bad. This map can be used to detect trends in the sales from countries.

3 Decision Making

To detect emerging countries in which marketing could be worthwhile we have to carefully inspect the geographical maps. We could define an emerging country using the data from the last 6 months by the following criteria: 1) The country does not currently provide most of our sales. 2) We have received (mostly) positive reviews from its users. 3) We have above average sales. 4) There is a positive trend visible in the monthly sales.
For criterion 1 we can exclude the United States from being an emerging country as most of our sales are from there according to the *Geographical Sales Volume*

maps. For criterion 2 we look at the *Geographical Daily Average* maps and note which countries have a purple color. For criterion 3 we look at the *Geographical Sales Volume* maps and note which countries have darker colors compared to the rest. For the final criterion we look at the *Geographical Sales Difference* and note which counties have a green color more often than they have a yellow color.
Using this analysis we have identified Canada, the United Kingdom and Germany to be emerging countries in which marketing could be worthwhile.

4 Satisfied and Critical Customers

The find the country or region in which most users rate the app negatively, we look at the *Geographical Total Average Rating* maps. The most negative countries are Botswana and Paraguay, giving an average of 1 out 5. Close second place in negative average reviews are the Dominican Republic and Iraq, that average being 2. There seems no clear indication why these countries have such a low average, as their neighbours are not giving such reviews. A possible explanation could be a low total number of reviews from those countries.
South East Asia is the most positive region, as that region gives on average a 5 as a rating, this is excluding the Philippines, Malaysia and Indonesia. They however rate at least 4.2 out of 5 on average. China and Bangladesh, two countries close to South East Asia, also have an average of 5. However, besides the Philippines and Indonesia, none of these countries left any reviews since June 2021.
Close second in being positive would be Europe, the Baltics, the Swiss, Bosnia, Belarus and Bulgaria all have a perfect average review score. Romania and Czechia have an average review score of at least 4.8, which is also very positive.
To find the date around which the release with a serious bug was released, we take a look at the weekly *Satisfaction Index* and try to find a sudden drop in satisfaction. We see this drop at week 46, where the *Satisfaction Index* drops from 1.091 to 0.657 and continues to drop from there in the follwing weeks. So we suspect that round week 46 of 2021, the version was published that introduced the crucial bug. From week 46 onward the *Average Daily Crashes* passed the 30 mark and eventually even doubled and almost tripled compared to weeks before the 46th. For simplicity sake we assume that the version was introduced on November 15th, the first day of week 46.