Assignment 1: visual analytics

Data Science course

Introduction

Congratulations! For the next 3 weeks you are employed as all-round Data Science expert of the Leiderdorp-based app and software business “Emerald-IT”. One of their apps is the *Complete Reference for Dungeons and Dragons 5*, a popular app used for playing Dungeons and Dragons by people of all ages and genders worldwide, primarily making money from selling in-game modules that allow the user to unlock additional functionality. Your job is to provide the company's management with information on the status and trends with respect to the monetization of the app, the ratings and feedback given by users and the stability (crashes) of the app. To do this, you will use data from the company's Developer Console logs, containing all purchases, ratings, feedback and crashes reported by the customers. Ultimately, your data analysis, visualization, interpretation and tools should allow company management to make decisions on future steps to expand the app.

Goals

The goals of the assignment are:

1. create a web-based **dashboard** which can be used by management to understand and interpret the data of the app.
2. Answer a few **strategic questions** on the app using the data and/or the dashboard in a short assignment report.

The report that you hand in for this assignment should contain a short introduction to the data, the company and the dashboard, as well as the answers to the questions.

Submission information

Deadline: **March 14 (Monday).**

Submission in Brightspace (Assignments, Assignment 1 visible once you enrolled in a group)

Submit:

- Your final assignment report (as separate PDF, not in the zip file), addressing the dashboard requirements and the strategic questions listed in the attached PDF.
- Your Dashboard URL (in the submission comments)
- All relevant source code in one zip-file or tarball

Preliminaries

- Follow this tutorial “Interactive Data Visualization in Python with Bokeh”: https://realpython.com/python-data-visualization-bokeh/
- You can work locally on your computer or in Google Colab, https://colab.research.google.com/
- If you work locally, I advise you to work in a Python IDE such as Anaconda or Pycharm, or use Jupyter notebooks (checkout Visual Studio Code if you like).
- Make sure you have installed Python 3 and the following packages: `pandas`, `numpy`, `geopandas`, `bokeh`.
- *You can use Overleaf (pro) via the University:* https://www.overleaf.com/edu/leidenuniv

Data

The data for this assignment comes in five different file types: `sales`, `reviews`, `crashes`, `ratings_overview and ratings_country`. The main data tables to be studied is sales, crashes and ratings_country, which contain a few hundred records spanning a time period of 7 months in 2021.

The sales files contain more apps than the one we would like to visualize; you need to clean the data set first. (See Product id below). The files are stored per month (using YYYYMM in the file name) . The sales file has the following important attributes:

- *Transaction Date*: The date of the transaction.
- *Transaction Type*: either Charge or Google fee (you only need to process Charge type).
- *Product id:* Which product is sold (app), only use charges for **com.vansteinengroentjes.apps.ddfive**. The sales data of the other apps is optional to use in the Dashboard.
- *Sku Id:* The in app purchase, for the ddfive app there are 2 options: **unlockcharactermanager** and **premium**
- *Buyer Country:* The country code where the customer bought the item.
- *Buyer Postal Code:* Zipcode of the purchase.
- *Amount (Merchant Currency):* The amount in euros.

The crashes file has the following important attributes:

- *Date:* the date.
- *Package Name:* constant since we look at one app.
- *Daily Crashes:* number of crashes on date.
- *Daily ANRs:* number of not responding apps on date.

The ratings_country file has the following important attributes:

- *Date:* the date.
- *Package Name:* constant since we look at one app.
- *Country:* country code.
- *Daily Average Rating:* average day rating for that country

- *Total Average Rating:* total average rating per country

In addition, the reviews file contains text reviews from users and optional developer replies.

Tasks

1. Dashboard

The web-based dashboard consists of various (at least four) widgets that should each visualize (in a different way) the aspects of the data given below.

Use `Bokeh` for the interactive visualization. You first have to process the data from the csv format into `pandas` dataframes. Note that data pre-processing typically takes up a substantial part of the work. Hint: the date column has to be converted to pandas format using the function `pd.to_datetime`.

1. [10p] **Sales Volume**: Visualize the sales over time (for example, per month or per day) in terms of at least two measures. For example: real money (Amount) and transaction count (row count).
2. [15p] **Attribute Segmentation and Filtering**: Present sales volume (as above) segmented per attribute: at least the SKU id (in-app purchase option) attribute should be included, but you can also think of the day of the week, time of the day or the country of the customer.
3. [15p] **Ratings vs Stability**: Can you come up with some Key Performance Indicators (metrics and scores) that help management understand how the app is doing in terms of stability and user satisfaction? Visualize them in a nice way. For example, the number of crashes in correlation with the daily average rating.
4. [10p] **Geographical Development**: visualize the sales volume (as above) and the average rating per country in a geographical setting (using the `geopandas` package, see more information below) , for example the number of customers per country over time. The goal is again to give management as much geographic insight as possible.

- [10p] You are very much encouraged to make the dashboard **visually appealing**, and to use non-traditional visualization techniques to give management an astonishing insight in the data of their app. In particular, the **interaction** and degrees of freedom in visualizing and exploring the data should be as large as possible.

  Hint: For implementing the interactive features you can use JavaScript callbacks: https://docs.bokeh.org/en/latest/docs/user_guide/interaction/js_callbacks.html

Export the final dashboard as HTML and make it online accessible in your own webspace at LIACS. Each user should have SSH-access to the LIACS webserver:

- Login with `ssh ulcnname@ssh.liacs.nl` (enter your ULCN password)

You can create your own website by placing files, e.g., an `index.html` file, in the `/webhome/ulcnname/public_html/` folder on this server.

- Transfer files with
  `scp file.html ulcnname@ssh.liacs.nl:/webhome/ulcnname/public_html/`
- Your website is visible at `http://liacs.leidenuniv.nl/~ulcnname`

2. Report writing

- [10p] Explain the dashboard functionalities (each of the components) in the report. No need to list actual code in the report.

The following questions should be answered by querying the data or using the dashboard. For each question, always elaborately motivate your answers based on the data, for example by giving queries or instructions to use the dashboard.

1. [10p] **Visualization** For each component of your dashboard, explain how you mapped data attributes to visual attributes, and why you made various visualization or interaction design choices. How do these choices help management understand the data without having to understand your code?
2. [10p] **Decision Making**: For next year, company management has budget to start a marketing campaign in three countries. Based on the data, what are the emerging countries in which marketing could be worthwhile? State how you define 'emerging'. Try to do more than counting transactions. For example, look for a trend, or also incorporate customer satisfaction.
3. [10p] **Satisfied and Critical Customers**: Which country or region has most users that rate the app negatively, and which country or region has the highest average rating?
   In the 5-month duration, several app versions have been released. In one of the releases was a serious bug that affected all android 7 users. Around which date do you think this happened? Look at the crashes and the average daily rating.

To help you write for the company's management, use the tips & tricks from the take into account your audience and purpose chapter. Avoid commonly made mistakes when writing in English and improve your English writing with the language module.

Additional information on the use of geopandas

For an in-depth example of how to create a geopandas map with bokeh, see this link: https://dmnfarrell.github.io/bioinformatics/bokeh-maps (Open in a private browser if the site tells you that you need to upgrade to premium). Shape file for countries: https://www.naturalearthdata.com/downloads/110m-cultural-vectors/110m-admin-0-countries/