Estonians on the r/place canvas

Jaanus Raudsik, Johan Erik Pukk, Tanel Marran

TASK 2

- **Identifying your business goals**
  - **Background** - On an exceptionally rainy Tuesday evening, our heroes - Jaanus, Johan and Tanel - were, after discussing a few more critical issues, brainstorming ideas for their data science project. They all agreed that they didn't want anything that was too difficult, yet nothing boring either. After spitballing a few ideas, they landed and stuck with r/place because they were all familiar with it, there were loads of different datasets about it, it would enable them to conduct an Estonian-centric project and, perhaps most importantly, r/place had a visualization aspect to it that really interested them.
  - **Business goals** - Our goal with this project is to identify the presence and activities of Estonian contributors. We'd like to gather statistics on the average r/place Estonian. For instance, besides the obvious "average pixels placed" or "number of Estonian structures found" it would also be interesting to see the so called feeding grounds of these users. Where were they most active? Was there any noticeable pattern to their movements? Were all of these users collaborating with each other or were they acting alone? These are "noble" goals of this project, but beyond actual tangible results, we also hope that this project can sell us on the idea that "data science can be fun".
  - **Business success criteria** - We shall consider the project a success if we can collect at the very least 3 interesting bits of statistics, info, patterns or other trivia about the Estonian contributors of r/place that we personally feel are noteworthy. Additionally, if we're able to create a few (at least 2) fun visualizers of our efforts, such as the travel logs of Estonian users, that could also be considered a success.
- **Assessing your situation**
  - **Inventory of resources** - Besides the entirety of our Introduction to data science course and everything it has taught us about data analysis, we also have at least one considerably large Estonian flag on the r/place canvas from where we can get the user hashes that interest us. It's important to note here that while the default dataset about r/place is absolutely huge, a lot of that data doesn't have any context before we find our Estonian users.
  - **Requirements, assumptions, and constraints** - The most important requirement for this project to even have a chance of succeeding is the existence and identifiability of Estonian contributors. Our main assumption is that mostly Estonians would contribute to the creation of Estonian flags, hearts or other such structures. This assumption is the baseline upon which we will start building our project. However, such an assumption is coupled with an equally large restriction - there is no real way to prove that the users that we are following and analysing in our project are actually Estonians.

    - **Risks and contingencies** - As mentioned before, the largest risk we are taking is that we are analysing users that very well might not all be Estonians.
    - **Terminology** - Beyond some statistical terms such as mean, distribution, deviation and the like, we do not believe we will use any complicated terminology. We may however use terms relating to data visualisation such as a heat and line map.
    - **Costs and benefits** - This project will not cost anything besides the valuable time of three students. On the other hand, it will ensure that upon a successful project, these three students will get some well deserved points, after which they will be well on their way to completing the associated course. :)
- **Defining your data-mining goals**
    - **Data-mining goals** - Our goal is to gather and report data about the average pixels placed by Estonian users (henceforth referred to as simply users), the color placed most often by users and whether or not that coincides with the most commonly used color of Estonians as accounted by Reddit in another dataset, the timeframes where users were most active and the average deviation of pixels placed by users. We'd also like to report the heat and line maps of users on the canvas (preferably in an interactive form). Furthermore if we have the time and knowledge, it would be interesting to develop a rudimentary model for detecting Estonian flags on the canvas (this is yet to be decided on however and may very well be omitted).
    - **Data-mining success criteria** - There is no quantifiable success criteria when it comes to data-mining in this project, since we are not creating a prediction model.

TASK 3

- **Gathering data**
    - **Outline data requirements** - Not applicable since we looked up the dataset before even considering the project. We came up with this project after seeing that the dataset was publicly available and this project does not require additional data sources.
    - **Verify data availability** - Data about the event was made public by the site's admins, see here. The dataset we will be using can be downloaded here 120MB download, unzips to 1GB.
    - **Define selection criteria** - We will be using the full dataset given in the link above, we will be figuring out which users, as denoted by their user_hash in the dataset, contributed to making the estonian flag and heart. We will figure this out by looking at the coordinate ranges where these objects were built and looking at users who placed several pixels of the correct colours there.
- **Describing data** - Dataset contains contributions of 1.2 million redditors colouring 1000x1000 pixel canvas to build the largest collaborative art project in history. Over the course of 72h they painted over 16.5 million tiles in 16 colours and info about those actions are the dataset. The data contains the color of the pixel the user placed, a unique number that identifies that user, x and y coordinates of the placed pixel and a timestamp of when the pixel was placed. The dataset is too large to open all at once in Excel.
- **Exploring data** - The x and y coordinates range from 0 to 999.. The colours are represented by integers which range from 0 to 15 and each integer has colour code associated with it. User_hash is number that is uniquely connected to user who contributed the datapoint. The timestamp is a standard Unix timestamp. The data has several visualizations and timelapses available online and can be seen here.
- **Verifying data quality** - The data has already been cleaned up for use by the site's data science team. We did not find any errors or abnormalities when looking over the data so we trust that the team did a good job. After analyzing previous points we decided that this dataset is good enough to support our project’s goals.

TASK 4

Project plan

	Tools & methods	Contribution time	Additional notes
Try to identify the users who built the estonian flag and heart	Pandas	Tanel: 3.5 hours Jaanus: 3 hours Johan: 3.5 hours Total: 10 hour	Outputs a list of users who contributed to building the estonian flag and heart
Map the contributions of the users identified in the previous task on the rest of the canvas	Pandas, Matplotlib	Tanel: 5 hours Jaanus: 5 hours Johan: 5 hours Total: 15 hour	Outputs a dataset containing all the contributions of the aforementioned users
Create a heatmap of the contributions of the identified users	Pandas, Seaborn	Tanel: 6 hours Jaanus: 7 hours Johan: 7 hours Total: 20 hour	In this point we can see if there are any clusters.
Identify regions and objects where the identified users contributed to	Pandas, Visual analysis, Matplotlib	Tanel: 7 hours Jaanus: 6 hours Johan: 7 hours Total: 20 hours	
Making an informative and attractive poster about our findings from data	Photoshop	Tanel: 5 hours Jaanus: 5 hours Johan: 5 hours Total: 15 hours	