Project Diary

ABSTRACT

This project is focused on the clear and light visualization of data gotten from VK analytics. Despite of the fact that I had selected information about outer space groups, the given template could be easily extended for any sort of groups (with minor changes in design). Project is based on the d3 library for visualization of data and Bootstrap framework for page layout, what gives the opportunity to concentrate on the beautify process actually.

KEYWORDS

LINE CHART, CHORD DIAGRAM, D3, VISUALISATION, DATA MUNGING, SOCIAL NETWORKS

“The beginning is always today”

INTRODUCTION

This project had started in November, when the data scope had been chosen – information from Russian social network about 4 groups of Space field. As it was noticed in Proposal, the team was represented by 4 people, *Table 1.*

Table 1 - Project team at the very beginning

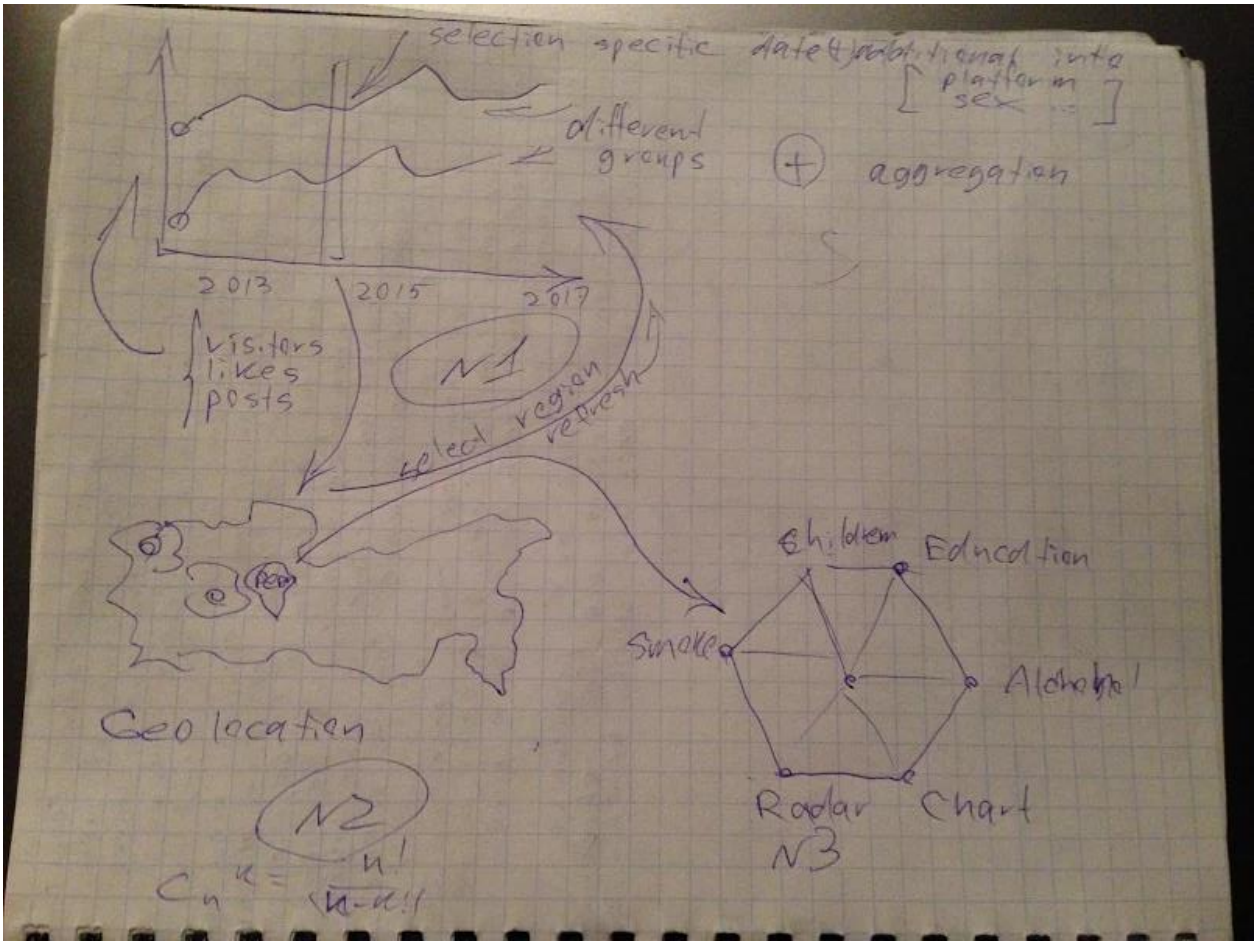
	Group	Role
Alexey	Big Data	Dev 1
Oleg	Big Data	Dev 2
Anna	Science Communication	
Anastasia	Science Communication	

21 November 2017

The meeting with team, discussing the concept of final visualisation, sketching.

22 November 2017

Handmade sketch.



ABOUT RAW DATA

The raw data has been downloaded from group admin's panel and represents the time series data, *Table 2.*

Date	Name	Par1	Par2	Value
03.12.2017	views			222
03.12.2017	visitors			165
03.12.2017	gender	Ж		54
03.12.2017	gender	M		104

Day-by-day statistics for next parameters:

- Visitors
- Gender
- Ages (8 ranges)
- Countries
- Cities
- Feedback
  - Likes
  - Shares
  - Comments
- Members
  - New
  - Left

4 files with approximately 60k rows, what made them awful for pure integration in project. For better performance has been implemented the data munging process, what gives as a result the json file with hierarchical structure (for more information have a look at /py folder).

Another part of data is public users information has been crawled via VK API, total number of unique users is **1 262 272** and approximately **1GB of size**, what makes this raw data totally unfitted for front-end usage. However, we can aggregate it before moving to visualization actually. For this purpose, I've selected typical Data Science package - Pandas. The main focus was on following fields:

- Relations
  - **1** – single;
  - **2** – in a relationship;
  - **3** – engaged;
  - **4** – married:
  - **5** – it's complicated;
  - **6** – actively searching;
  - **7** – in love.
- Personal views
  - Political
    - **1** – Communist;

        - **2** – Socialist;
        - **3** – Moderate;
        - **4** – Liberal;
        - **5** – Conservative;
        - **6** – Monarchist;
        - **7** – Ultraconservative;
        - **8** – Apathetic;
        - **9** – Libertian.
    - Languages
        - Russian
        - English
        - Deutch
    - Religion
    - People main
        - **1** – intellect and creativity;
        - **2** – kindness and honesty;
        - **3** – health and beauty;
        - **4** – wealth and power;
        - **5** – courage and persistance;
        - **6** – humour and love for life.
    - Life main
        - **1** – family and children;
        - **2** – career and money;
        - **3** – entertainment and leisure;
        - **4** – science and research;
        - **5** – improving the world;
        - **6** – personal development;
        - **7** – beauty and art ;
        - **8** – fame and influence;
    - Smoking
        - **1** – very negative;
        - **2** – negative;
        - **3** – neutral;
        - **4** – comprisable;
        - **5** – positive.
    - Alcohol
        - **1** – very negative;
        - **2** – negative;
        - **3** – neutral;
        - **4** – comprisable;
        - **5** – positive.
- Education
    - High edu: Yes or No
    - University name

- Education status?

ATTENTION! the final implementation on defence could be without some data due to force majeure situations.

IDEA of PROJECT

Based on the raw data or even on the graphs from group admin panel, we cannot show the dynamics of users' action as well as their regional distribution and personal portrait. Current work could give the opportunity for analyzing the set of groups about Space: such as typical views/likes, and more rare gender/regions, what as a result could help to find out some patterns of interest in outer space based on the users' behavior.

The core of visualization is interactive line chart with days as xAxis, and selected statistics of groups as value. The second part is a map of Russia, where color of regions is filled based on the max value of views from this region from groups, what in other words encodes the regional distribution of groups members. The last part is chord diagram for showing the common members of groups, what is critical for advertisements and cooperation for selected groups as well as sponsors.

“A deadline is negative inspiration.
Still, it's better than no inspiration at all.”

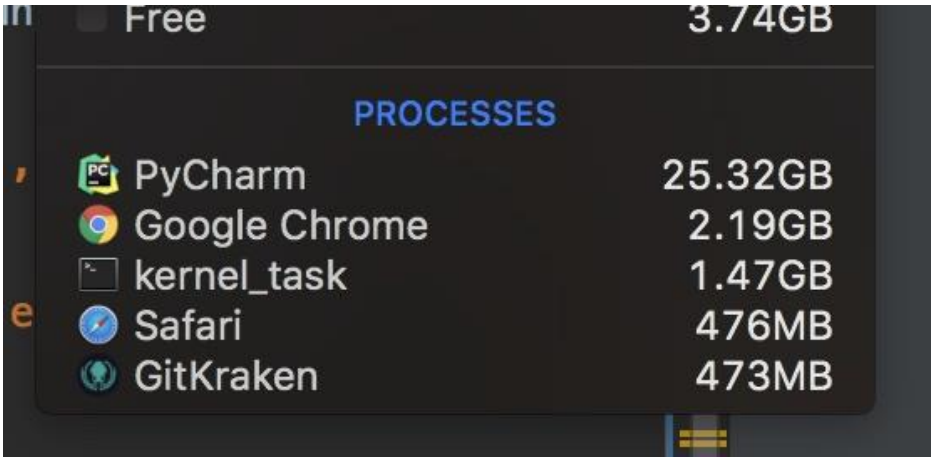
IMPLEMENTATION

As it was mentioned previously, the data has required a vast amount of work for further usage in project, so let’s start from data preparation.

8 January 2018

NOTE: for this purpose, I used Python 3 and pandas.

The crawled data had been aggregated – in words of size 1.2GB to 16KB, or in words of *data as is* from users’ info to information about groups.



9 January 2018

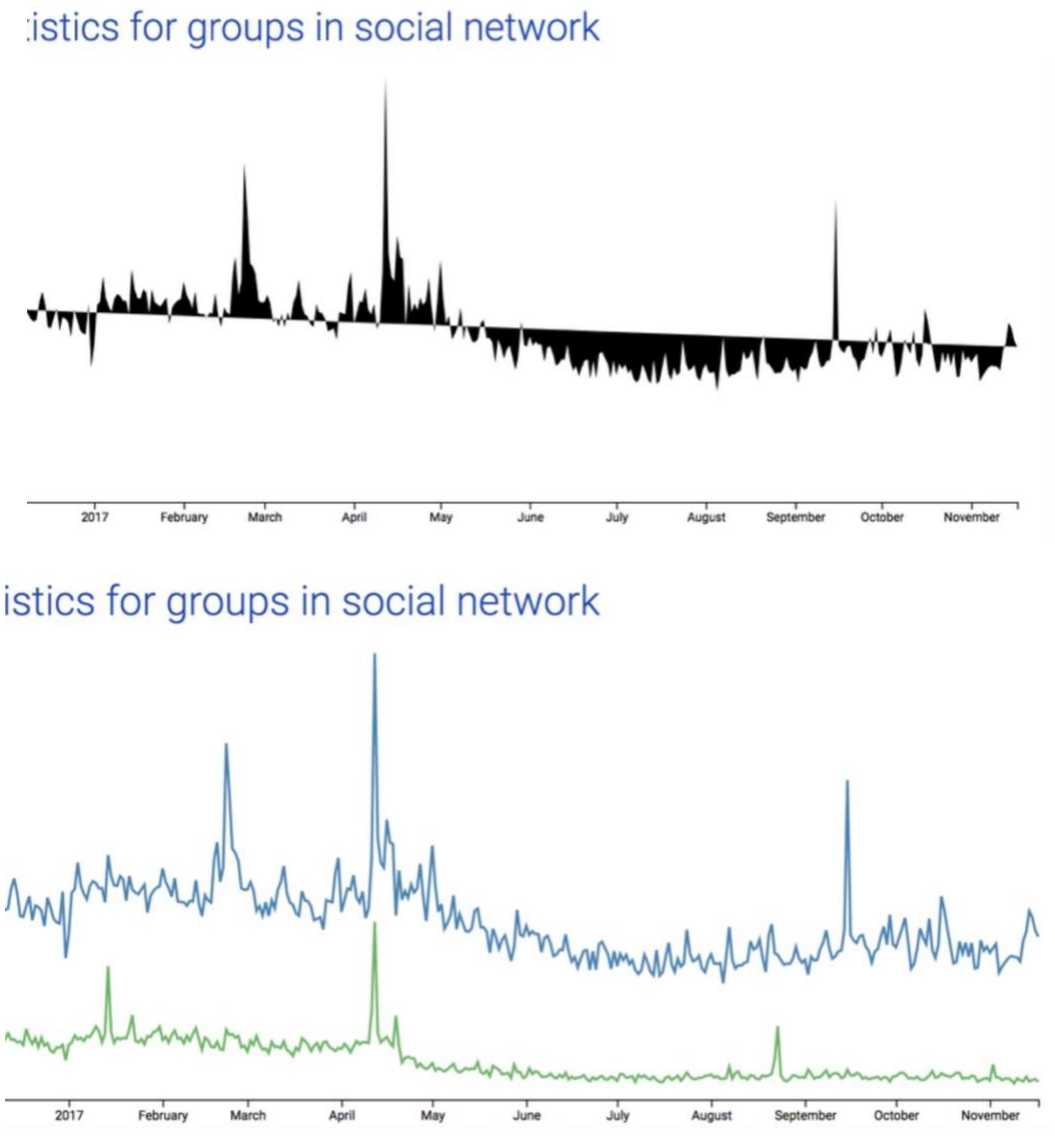
Still worked with data, was trying to choose the best structure for js: array of dictionaries, dictionary of dictionaries and etc. Managed the project structure: folders, humans.txt and robot.txt.

10 January 2018

Built the layout for HTML page (thx Bootstrap), implement the simple line chart (non-interactive) and selection option. Realised, that the data in json format which had been given me from another teammate is wrong, what was pushing me to the decision of leaving team. Had implemented py-script for refactoring data structure: array of dictionaries relative to group with stat info inside as array.

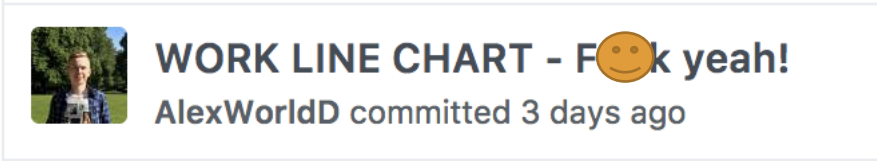
11 January 2018

Left team – now I could rely just on myself. First of all, done pretty Axis according the selected encoding options. But the line was looks awful – no CSS was a bad idea, so, the style for lines were added.



12 January 2018

Animation was fixed. Line char had been completely finished. Honestly, the best short description of the results of this day is summary for commit:



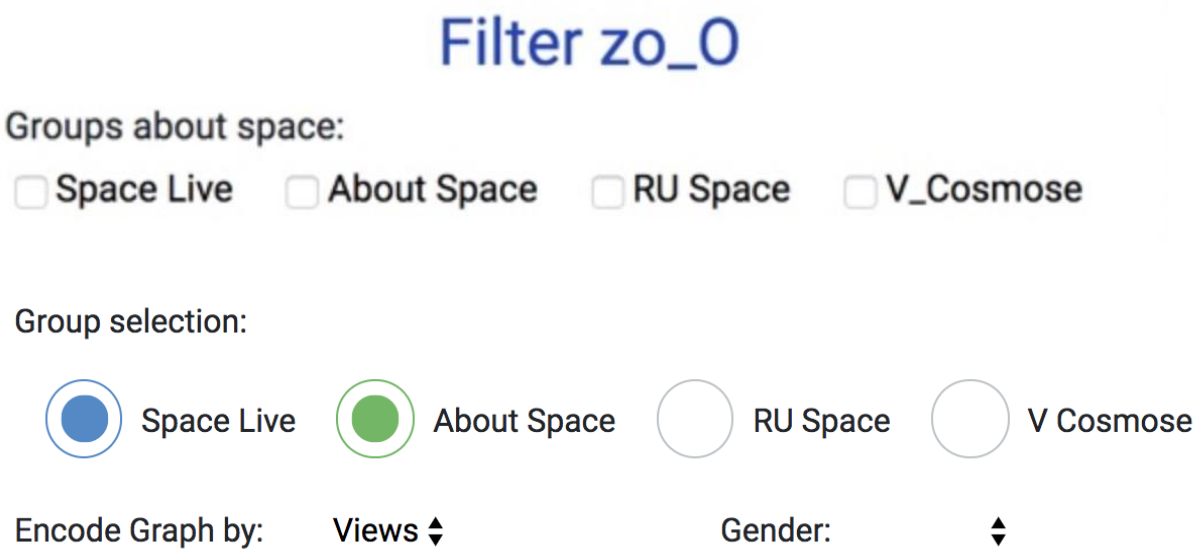
13 January 2018

The final day before *code-ready* deadline and the most productive one. During this endless day were implemented:

- Encoding selection – adaptive, for instance, choosing Vies follows the gender selection;
- Data had been modified one more time: the cities names turned into region code (additional info in /py folder);
- Map had been implemented: topoJson as a start point, basic colors filling done. Added animation and transition as well as dependencies from line chart to map.

14 January 2018

Header had been updated: add customs checkboxes as well as images changed to quotes about Space.



“Success is walking from failure to failure with no loss of enthusiasm.”

RESULTS

LIVE DEMO - https://alexworldd.github.io/SpaceVK/

CONCLUSION

During moving from idea and concept to final realisation were studied or improved some critical competences for Data Scientist such as traditional finding and munging data and more specific skills to visualise data in good way.

Justification of final implementation:

- ✓ For core part was selected line chart due to some pros against bar chart, like more intuitive for time series data and clearer with wide range of values, low number of intersections between data items;
- ✓ The legend on graphs has been missed, cause the selection checkboxes are based on the same color encoding as data in graphics;
- ✓ The white background has been selected for better experience in showing with projector; plus, it allows to concentrate on the data.
- ✓ The back-forward dependences from map to line chart has been missed due to the lack of data (the main part of line is 0-value for specific cities/region);
- ✓ Map has been selected because it is the most compact representation for GEO information (for instance, table for regions could require the whole page and more actions from user as consequence);