Homework #4. Exploratory Data Analysis

Context

It's assumed that your data is already downloaded on your machine (in h/w #3). Now you're going to explore it. This means that you're a data analytic who got a dataset and the task to explore the data from different points of view and find as many insights as data consists.

Data

You can use either "personal data" data (**minimal number of messages 100k per chat**) or russian propaganda data (downloaded from telegram).

"russian propaganda" dataset was collected by Kate Burovova, and this data should be used only for course purposes. It can not be shared or uploaded online (please don't commit in the git).
https://drive.google.com/drive/folders/1nhcgHqo_mjQtl7eVw58Jvb_qajdS2pzn?usp=drive_link.

Once you made a data decision - you should stick to it till the end of the course.

Task #1

Calculate the time you spent on this task.

1. Create a file `2_eda.ipynb`. Copy the first cell from the file 1_data_review.ipynb, update h/w name (Homework #4. Exploratory Data Analysis), and time spent on the h/w.
2. Make explorative data analysis on top of the telegram data you've decided to use **personal data** or russian propaganda data downloaded previously (dialogues and messages).
   [**For personal data**] During this investigation, you're going to take a look at each dataset separately, but also **you should merge them and analyze the new (merged) dataset**.
3. Generate a pdf report on top of your .ipynb file.

I know that people tend to choose the shortest path (do as minimum work as they can), and tasks without strictly defined deliverables lead to misunderstanding, so I'll give you a clue how I'm going to evaluate your work:

- I'll find ~5 the best works; they will be estimated at 100%, all other works will be estimated proportionally to the amount of work the person did relative to our "best works".
- I do expect you'll get answers to 20+ questions (it's your work to create questions) during the research. The deeper your questions are - the better.

- Exploration should be mostly **visual**, so I do expect **a lot of data visualizations** in the report. **Each visualization should consist of**: title (+caption if you want/need) with information that is shown on the picture, x-, and y-axis have meaningful names, legend (if needed) has all entities named meaningfully.
- [**For personal data**] You **should merge two datasets** (dialogues and messages) and analyze this merged dataset. Also, do not hesitate to add any additional data which you think can work together.
- This work is intended to initiate your creativity and analytics skills, skills to find and ask the right questions. It's okay if it takes time; the more you work, the better you get.
- **Please, don't do word cloud!**

Recommendations

It can be helpful to revisit Lecture #3 (Dataviz), and read this page - https://en.wikipedia.org/wiki/Exploratory_data_analysis to remind EDA purpose and flow.

I **strongly recommend** you to go through all links from the materials section to get some inspiration and ideas.

Intermediate results presentation

02/11/2024, 23:59 (4pt)

Expected Outcome

1. PDF report (file named "4_0_<your_id>.pdf") exported from the file `2_eda.ipynb`. This report should consist of your (some) exploration for datasets, a list of questions (assumptions, hypothesis, ideas to test) you have to the data, and some drawings of basic information (distributions, etc).
   It's just a preliminary report, I'm going to score the fact that you've submitted a file with the information above but with remaining work (not finalized) to be done during the second week of work.

Final Deadline

09/11/2024, 23:59 (max: 16pt + extra)

Expected Outcome

1. PDF report (file named "4_<your_id>.pdf") exported from the file `2_eda.ipynb` with the well-structured report, with storytelling, drawings, questions, and answers (in the visual form mostly, but not only).
2. 1 csv file named "4_<your_id>.csv" with all fields filled.
   ATTENTION: naming is critical as I'm using an automated approach to merge your answers for the review.

File with fields to fill can be downloaded here (I'll open after the first week)- https://docs.google.com/spreadsheets/d/192Swa-OdpdQWKr2Qfdf8eoyuoNIHGLraz2wW8ryvkvQ/edit?usp=sharing

Your ID can be found here - https://docs.google.com/spreadsheets/d/1U9hFib_R1OxEo8G6ZNNNJEFeBzjIdL27WdAVLY4qct8/edit?gid=0#gid=0

Materials

Lecture #3

Some basic tools (you can apply to the data)

- Text Mining: Word Relationships
  https://uc-r.github.io/word_relationships
- Count basic statistics (Basics of Text Mining in R - Bag of Words)
  http://rstudio-pubs-static.s3.amazonaws.com/256588_57b585da6c054349825cba46685d8464.html

Visual ideas

1. https://www.data-to-viz.com/
2. https://datavizproject.com/#
3. https://datavizcatalogue.com/
4. https://github.com/Financial-Times/chart-doctor/tree/main/visual-vocabulary

Colours

1. https://projects.susielu.com/viz-palette
2. https://colorbrewer2.org/#type=sequential&scheme=BuGn&n=3
3. https://learnui.design/tools/data-color-picker.html#palette

ggplot2

1. GG plot usage
   http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html
2. Themes
   https://github.com/jrnold/ggthemes
   https://github.com/Mikata-Project/ggthemr

GEO

1. https://kepler.gl/#/
   https://github.com/keplergl/kepler.gl
2. https://datashader.org/
   https://github.com/holoviz/datashader/

3. https://geoffboeing.com/2016/11/osmnx-python-street-networks/ https://github.com/gboeing/osmnx
4. https://github.com/python-visualization/folium
5. https://github.com/karimbahgat/PyGeoj
6. https://github.com/riatelab/linemap

List of tools

1. https://docs.google.com/spreadsheets/d/1miN1dUvPMmnLdhunhZCil80LwlJ3xj8gRXrUNA-KrIE/edit#gid=0

Dataviz ideas

1. https://pudding.cool/
2. https://exhibits.stanford.edu/dataviz
3. https://setosa.io/#/
4. http://czrt.by/projects/
5. https://www.makeovermonday.co.uk/makeovers/makeovers-2018/
6. https://www.bloomberg.com/graphics/2015-whats-warming-the-world/
7. https://www.visualisingdata.com/resources/