Project Proposal

Liam Albright, Jennifer Tao, Christine Zhang

Info 201

https://github.com/jentao/eu-project

Project: EU Horizon 2020 analysis

Our project is going to be an analysis of the EU research projects under Horizon 2020. The data is compiled by The EU Open Data Portal which is a project set up by the European Union to add more transparency to their actions. The data itself is aggregated by the Community Research and Development Information Service.

The data sets we are going to work with are the H2020 Organizations and H2020 Projects data sets. The Organizations dataset contains data on the organizations that are part of the H2020 mission and includes data such as role, name, and address. The Projects dataset contains data on the projects involved with the H2020 mission. The data in this set contains attributes such as project name, cost, objective, and participating countries. These datasets can be found on the [EU Open Data Portal](https://data.europa.eu/euodp/data/dataset/cordisH2020projects?fbclid=IwAR0QM-YgHTwobcWksrDtLBj7kKwGD_WN2NrUincxO7DUKVdEvAqLnEQm6zo).

There are several target audiences for our project. Mainly we want to offer analysis to anyone interest in the economics or political economy in the European Union. This could be students, professors, or the general public. Our audience will be to learn several issues from our projects. First, we want to illuminate which projects the EU is prioritizing. We are doing that by asking the question “Which projects is the EU fully funding and how much?”. Another thing we want to illustrate is the EU’s focus on Environmental issues. To do this we are going to ask the question “How many projects funded have environmentally focused objectives and how many countries are participating?”. Lastly, we are going to ask “ does the location of an organization has any effect on the funding of the project?”. In addition to answering those questions, we would like to make this data more visual and digestible for lay users.

Technical Description

The final product of this project will be a responsive Shiny App.

The datasets from EU Open Data Portal can be downloaded as .csv files.

The data we are using are separated in different files: H2020 Projects (includes project status, name, topic, cost, participants, etc.), H2020 Organisations (includes organization information), and H2020 Report summaries (contains the work performed and final results). These tables all have project Record Control Number (RCN) which we can perform joins to get the desired information.

We are planning to visualize these datasets in form of maps of EU to answer questions about the relationship between location and funding. Thus we need to include “maps” and “mapproj” packages to plot boundary maps.

Project Set-up

Github Issues:

1. Read the "Projects" dataset and filter the fully funding projects, then sort data in descending order. And make a histogram of this data using ggplot.
2. Read data and filter projects with environmentally focused objectives. Create a histogram base on countries to see the environmental awareness of each country (by looking at how many projects they participated in)
3. Create a boundary map by countries in EU. And have country name labeled with boundaries.
4. Write a function that takes country names as variable and generates scatter plots of individual project fundings of that country.
5. Create the basic layout of our Shiny App. Include three pages: About (about the project, about the dataset), Plots (histogram, line graph, pie graph, etc.), EU Map (location of organization).  
   All the plots come from previous issues.

We expect a couple of problems with our project. One of the major challenges we expect to find is so noise in our data. Particularly the objectives attribute in the projects data set has a lot of words. What this means is it is going to be hard to parse out exactly what each objective is about for each object of data. Another challenge in the project dataset is figuring out what some of the attributes mean, particularly we are unsure exactly what the data in the funding scheme column means. Figuring out this column will require further research.