The Harvard Crimson - Grant Data Project

Tech Board - Data Journalism Initiative

Contributors: Annika Huprikar, Samuel Lin, Leonard Tang

NSF

- We determined how NSF organizes and codes their grants
    - NSF Directorates
        - NSF Organizations

  Example:
    - NSF Directorate: MPS (Mathematical and Physical Sciences)
        - NSF Organizations: AST (Division of Astronomical Sciences),
          CHE (Division of Chemistry)
          DMR (Division of Materials Research)
          DMS (Division of Mathematical Sciences)
          PHY (Division of Physics)

- We searched the awards/grants database with the keyword *Harvard University* and downloaded the csv with all the results that were generated from this search

- Data Cleaning - ex: needed to recode award funding column as a numeric variable, remove . and , from denoting award integer, remove $ symbol, needed to recode some NSF directorate/org keys that were incorrectly coded, etc.

- **Data Analysis conducted in R and Python**

In R:

- Created data frames and according models of average funding by directorate (converted to thousands of dollars) and proportions of funding allocated to each directorate, both overall figures for 1960-2021 and within the past 6 years (2015-2021)

- Wrote abstract functions that perform a similar process within each directorate: for a given directorate D, modeled average funding by organization and proportions of funding allocated to each organization. This modeling was done for 1960-2021 as well as just a snapshot of the year 2020

- Hypothesis Testing: NSF only released the median annual grant size by directorate. Specifically, the 2019 and 2020 figures were analyzed for these tests. One-sample Wilcoxon signed rank tests were conducted to compare sample medians for Harvard-specific data with the NSF “population” medians. 95% confidence (i.e. alpha level of 0.05) is assumed for level of significance. Results of these tests are commented in the *NSF_Data_RNotebook* file.

In Python:

- Used Python's Plotly library to create data frames that group data by year of allocation and found the average amount of funding per year for each of the 8 directorates.

- Generated a grouped vertical bar chart where the horizontal axis represents the years between 2000 and 2020 and the vertical axis represents the average award amount per year in dollars. Toggled dynamically with the directorates displayed to emphasize individual directorates.

- Created/cleaned dataframes of funding by directorate and computed average funding (mostly deprecated work LOL, Sam and Annika rewrote this for more general situations)

See next page for NIH Methodology

NIH

- Very similar structure to NSF methodology

- We determined how NIH organizes and codes their grants (known as Administering ICs)
- We searched the awards/grants database with the keyword *Harvard University* and downloaded the csv with all the results that were generated from this search

- Data Cleaning - ex: needed to recode award funding column as a numeric variable, remove . and , from denoting award integer, remove $ symbol, needed to recode some NSF directorate/org keys that were incorrectly coded, handle missing values etc.

- **Data Analysis conducted in R and Python**

In R:

- Conducted one-sample t tests for means
- Hypothesis Testing: Located and collected data capturing 2019 and 2020 average funding amounts by IC (serving as population means from NIH) and used these in the t tests against the Harvard sample means to see if Harvard gets significantly more money for a given IC. 95% confidence (i.e. alpha level of 0.05) is assumed for level of significance. Results of these tests are commented in the *NIH_Data_RNotebook* file.

In Python:

- Used Python's Plotly library to create data frames that group data by year of allocation and found the average amount of funding per year for each of the 41 "Administering ICs" in the NIH dataset

- Generated a grouped vertical bar chart with all 41 administering ICs where the horizontal axis represents the years between 2011 and 2021 and the vertical axis represents the the average award amount per year in dollars

- Generated a second grouped vertical bar chart with only the 8 administering ICs with more than 800 listed projects and subprojects (note: top categories by number of projects, not by total funding), with the horizontal and vertical axes representing the same information.

- Created/cleaned dataframes of funding by each of 41 Administering ICs and computed average funding to date, plotting the results in a bar chart sorted by average funding