**Econ 860 Midterm Summary Report**

**Clara Cullum**

**PART 1**

Summary of data\_initial (data from charcoal paper minus duplicates)

This data includes all users (minus duplicates) listed on the charcoalpaper webpage along with their repo count, followers, and date created.

Summary Statistics:

Login ID Repo Count Follower Count Member Since

count 702 702.000000 702.000000 702

unique 702 NaN NaN 533

top erinata NaN NaN ( 2012-12-19 )

freq 1 NaN NaN 10

mean NaN 127.029915 570.103989 NaN

std NaN 459.393106 1639.550645 NaN

min NaN -1.000000 -1.000000 NaN

25% NaN 16.000000 24.000000 NaN

50% NaN 39.000000 77.000000 NaN

75% NaN 108.000000 291.750000 NaN

max NaN 8082.000000 12816.000000 NaN

Summary of data\_clean (initial data from charcoal paper minus duplicates and invalid user IDs)

This data includes only the valid user IDs from the charcoalpaper site.

Summary Statistics:

Login ID Repo Count Follower Count Member Since

count 439 439.000000 439.000000 439

unique 439 NaN NaN 379

top erinata NaN NaN ( 2008-04-03 )

freq 1 NaN NaN 7

mean NaN 138.589977 405.797267 NaN

std NaN 472.755085 1187.609022 NaN

min NaN 0.000000 0.000000 NaN

25% NaN 16.000000 27.000000 NaN

50% NaN 42.000000 71.000000 NaN

75% NaN 112.000000 252.500000 NaN

max NaN 8082.000000 12326.000000 NaN

N = 439

# Unique Logins = 702

# Invalid Logins = 263

**PART 2**

Summary of gh\_userid\_data (from GitHub scrape)

This contains all information scraped from Github on the valid users.

Summary Statistics:

Following:

count 439.000000

mean 3723.913440

std 20622.504973

min 0.000000

25% 19.000000

50% 80.000000

75% 348.000000

max 302918.000000

Name: following, dtype: float64

Starred:

count 439.000000

mean 23.886105

std 11.074293

min 0.000000

25% 22.500000

50% 30.000000

75% 30.000000

max 30.000000

Name: starred, dtype: float64

Created Time:

count 439

unique 439

top 20:52:45

freq 1

Name: created\_time, dtype: object

Updated Time

count 439

unique 437

top 14:12:34

freq 2

Name: updated\_time, dtype: object

Comparing datasets:

The data from 2022 has 263 more users than those currently valid on GitHub.

**PART 3**

Scatterplots:

I plotted repositories on followers. I would expect that more repositories is associated with more followers. At this scale, there does not seem to be a very significant linear trend, however, there may still be a correlation with outliers removed.

A graph of blue dots

Description automatically generated

The next plot shows last time updated on number of repositories starred. I would have expected that a more recent update indicates a more active user and therefore they would be more likely to star more repositories. However, the data seem to be distributed broadly with little relationship. Most users either have 0 or >30 starred repositories.

A diagram of blue dots

Description automatically generated

Resources Used

I used ChatGPT in developing this project to fix indentation errors, explain unknown error messages, and correct syntax.

I used Stack Overflow for syntax explanations and to learn how to make a scatterplot.