

UCLA Statistics and Data Science Winter Datathon Challenge

Welcome to the 2025 Winter Datathon for the Statistics and Data Science Club! A reminder that this datathon is exclusively for underclassmen and all upperclassmen should participate in Datafest, should they wish to do a data science datathon.

Task:

Your friend Kevin is an average Redditor living in his mom's basement. Kevin has grown tired of posting r/ucla cypypastas, so he has decided to scrape data of the most popular Reddit posts off the front page to learn how to become Reddit famous. Kevin being Kevin, he's asked you to do all the work for him. He wants you to prepare a presentation by Sunday on how best to boost him to the heights of Reddit popularity.

Kevin's collected posts are [here](#). He also wrote a short description of each column [here](#). Make no assumptions about the cleanliness of the data. (What if there are duplicate posts?)

His mother, being an expert software engineer, has provided some extra code [here](#) to help. Feel free to use it and not tell Kevin. Kevin's fabulous friends presented workshops for the UCLA Statistics Club, which you can access for some additional help [here](#).

Some questions to get you started:

- How does engagement differ across subreddits?
- What is the most upvoted sentiment, topic, etc.?
- How can we predict which type of posts receive the highest engagement (upvotes, comments)?
- What is the relationship between comments and upvotes?
- Do positive/negative posts get more engagement?

Presenting:

Your team will be given 36 hours to parse through the dataset, analyze it, and develop and present strategies to **Kevin and Co.** in a 5-minute presentation. Submissions for the presentation are [here](#), due at 11:59 pm on Saturday, 2/22. Late submissions will not be accepted.

If you wish to present in front of judges and be considered for awards, you must sign up for a time slot [here](#), in addition to submitting your presentation in the form above. The sign ups for presentation times will open on Saturday at 12:00 pm on the link above.

UCLA Statistics and Data Science Winter Datathon Challenge

The presentation must be concise (within a 5 minute timeframe) and include the following:

1. Summary
2. Analysis
3. Strategy

More sections may be included at your discretion.

Feel free to message any organizers or anyone on board for any help throughout the datathon!
However, we do recommend you ask more general questions in the #questions Discord channel, to maximize your chances of getting a timely answer.

Criteria:

Your presentation will be judged on the following criteria:

1. Quality of analysis:
 - a. **Comprehensiveness:** How thoroughly does the team analyze the dataset? Has the team derived sufficient information from the data provided?
 - b. **Accuracy:** Are the data findings consistent with the data provided in the dataset?
 - c. **Insightfulness:** Are the insights derived from the data meaningful, actionable, and relevant to the problem at hand?
 - d. **Sophistication:** Were the strategies used by the team sophisticated and advanced?
2. Strategy:
 - a. **Data-driven:** Does the strategy reasonably mirror the analysis on the given dataset?
 - b. **Reasonable:** Is the strategy reasonable, does it consider potential constraints?
3. Quality of presentation:
 - a. **Requirements:** Does the presentation cover all the topics listed above?
 - b. **Clarity:** Does the team present their information in such a way that Kevin could effectively understand the rationale for their strategy and implementation?
 - c. **Data Visualization:** Are graphs, charts, and other visual tools used effectively to communicate key findings?

Good luck! Make Kevin proud!