CSE 5095: Social Media Mining and Analysis
Fall 2024, Assignment #1, 200 points
Posted – September 10, 2024, Due: September 26, 2024

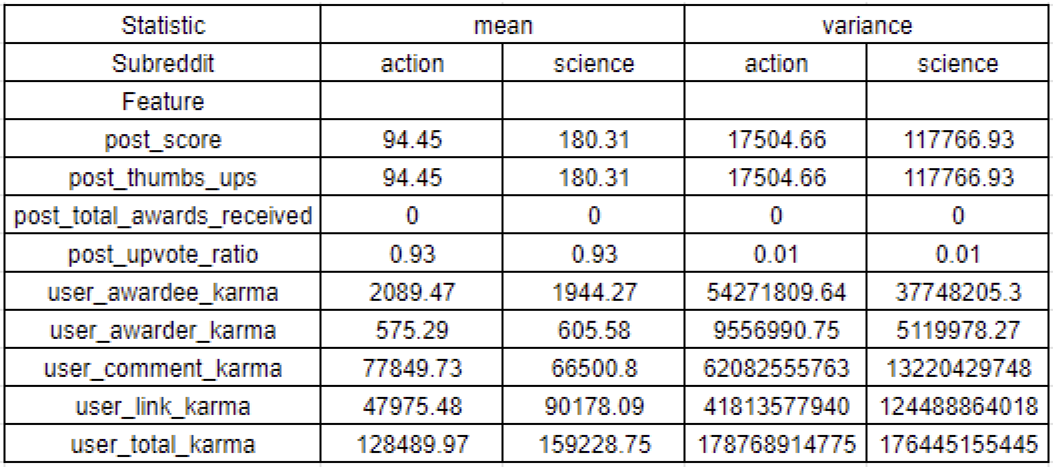
In this assignment, we will explore the statistical properties of the quantitative features associated with each subreddit in your data set. Each data set has observations from two subreddits. In some data sets, each observation is a post, whereas for the other data sets each observation is a compilation of comments for each unique post.

Project #10: (Stance Detection -- Action vs. Science)

Within the collection of subreddits, the climate subreddit is dedicated to the discussion and exchange of truthful information regarding the science of climate change. On the other hand, climate offensive and climate action plan subreddits are dedicated to the discussion of the active measures that are being taken to combat climate change. Build a stance detection framework, to identify whether a post is devoted to either activism or science of climate change. Unique posts from three subreddits are combined into a single data set. The posts from subreddits climate offensive and climate action plan are labeled “action”, and the posts from climate subreddit are labeled “science”. The data contains 1039 posts labeled as “action” and 2235 posts labeled as “science”. Data balancing may be needed in this project.

Task 1: Descriptive Statistics (50 points)

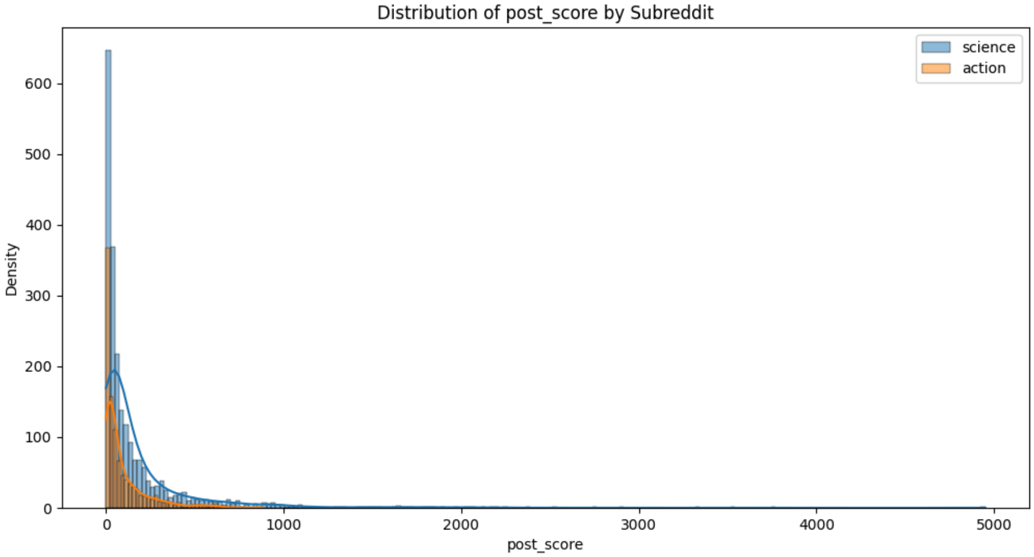
Build a table with the descriptive statistics (mean and variance) of the quantitative features for each subreddit in your data set. For projects based on post-level data, these will include the post-level statistics shown in Table 1. For projects based on comment data, these include the average of average commentlevel statistics as shown in Table 2 (average comment and user statistics are computed for each post). For comment data, include also user-level features listed in Table 2.



Statistic	mean		variance	
Subreddit	action	science	action	science
Feature				
post_score	94.45	180.31	17504.66	117766.93
post_thumbs_ups	94.45	180.31	17504.66	117766.93
post_total_awards_received	0	0	0	0
post_upvote_ratio	0.93	0.93	0.01	0.01
user_awardee_karma	2089.47	1944.27	54271809.64	37748205.3
user_awarder_karma	575.29	605.58	9556990.75	5119978.27
user_comment_karma	77849.73	66500.8	62082555763	13220429748
user_link_karma	47975.48	90178.09	41813577940	124488864018
user_total_karma	128489.97	159228.75	178768914775	176445155445

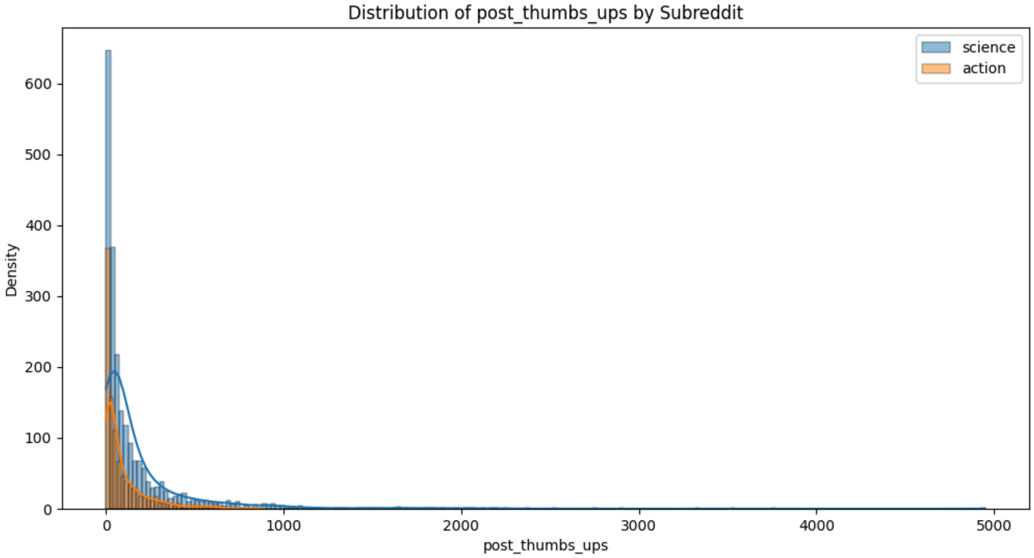
Task 2: Distributions (50 points)

For each quantitative feature, plot the two distributions corresponding to the two subreddits. Comment on the properties of each distribution (symmetrical, left-skewed, right-skewed), and how they compare with each other.



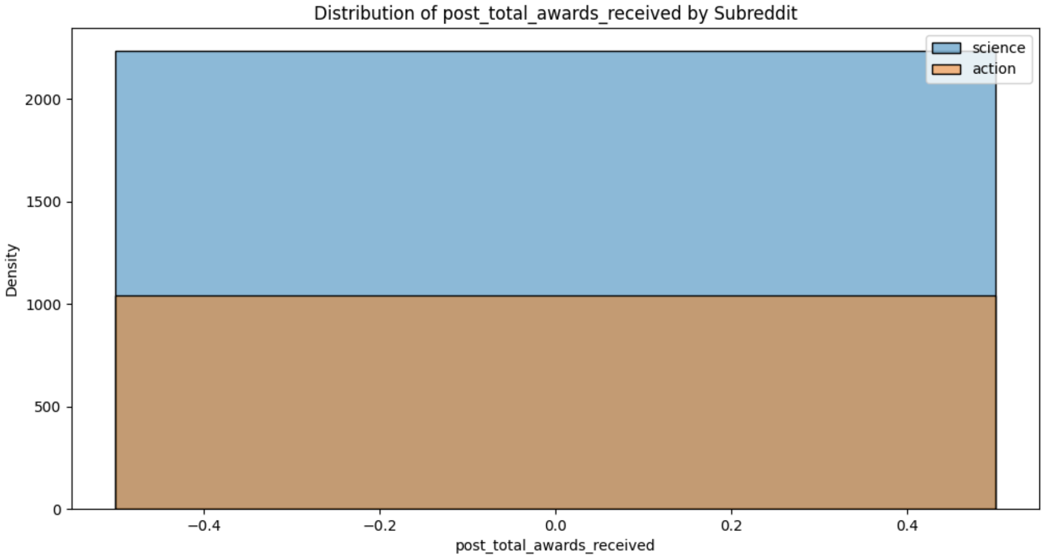
science - post_score:
  Mean: 180.31
  Median: 60.00
  Distribution: right-skewed
  Range: 0.00 to 4951.00

action - post_score:
  Mean: 94.45
  Median: 41.00
  Distribution: right-skewed
  Range: 0.00 to 880.00



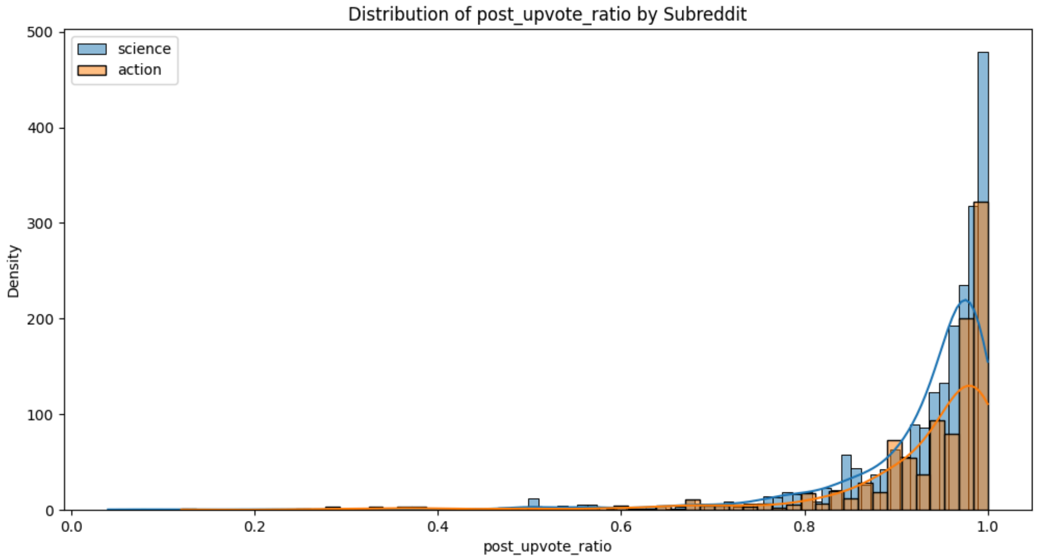
science - post_thumbs_ups:
Mean: 180.31
Median: 60.00
Distribution: right-skewed
Range: 0.00 to 4951.00

action - post_thumbs_ups:
Mean: 94.45
Median: 41.00
Distribution: right-skewed
Range: 0.00 to 880.00



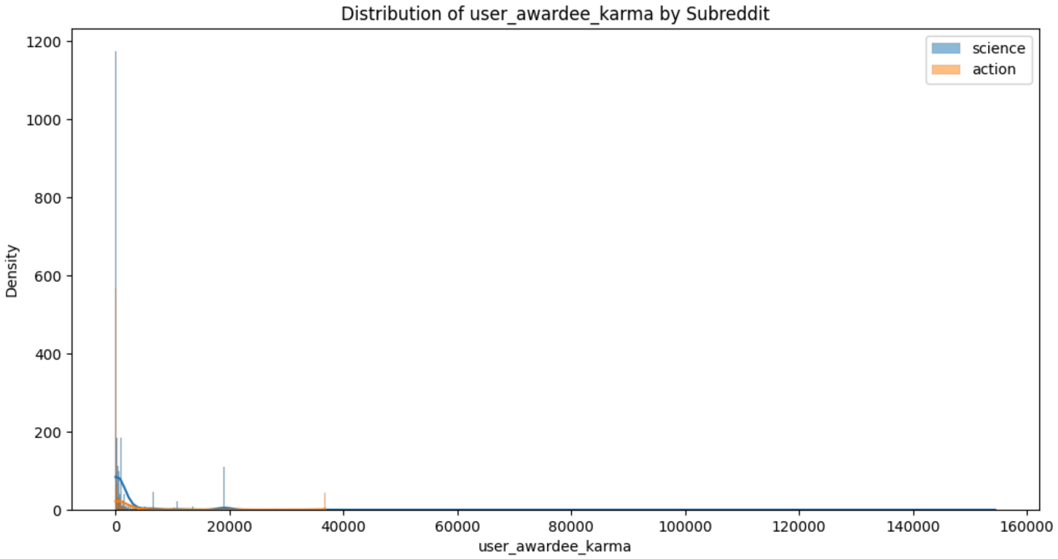
science - post_total_awards_received:
Mean: 0.00
Median: 0.00
Distribution: symmetrical
Range: 0.00 to 0.00

action - post_total_awards_received:
Mean: 0.00
Median: 0.00
Distribution: symmetrical
Range: 0.00 to 0.00



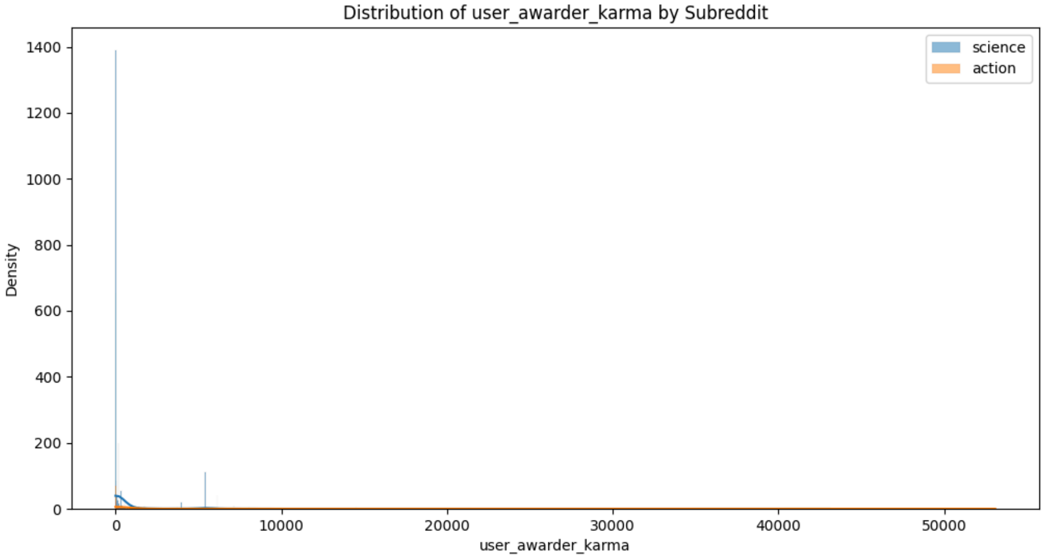
```
science - post_upvote_ratio:
  Mean: 0.93
  Median: 0.96
  Distribution: left-skewed
  Range: 0.04 to 1.00

action - post_upvote_ratio:
  Mean: 0.93
  Median: 0.97
  Distribution: left-skewed
  Range: 0.12 to 1.00
```



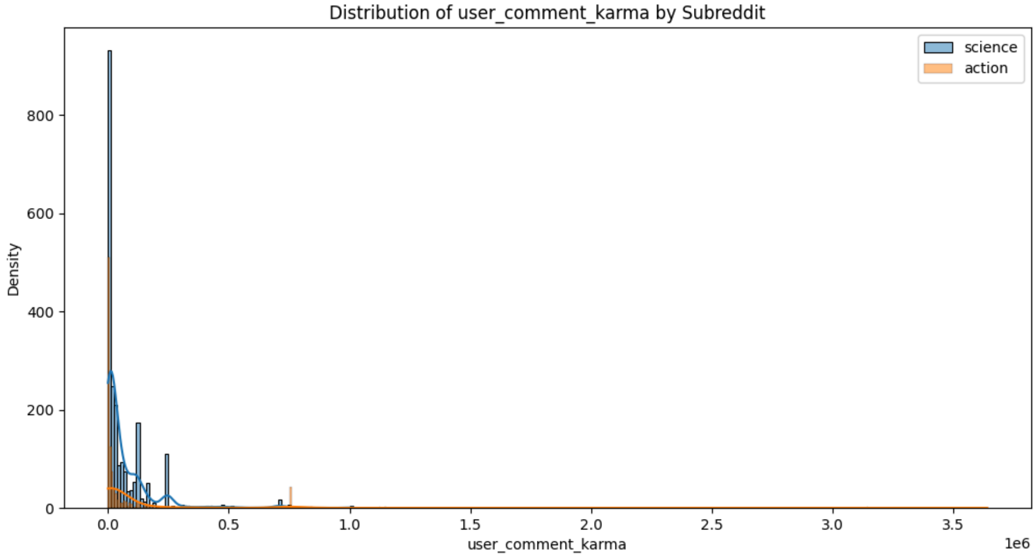
```
science - user_awardee_karma:
  Mean: 1944.27
  Median: 127.00
  Distribution: right-skewed
  Range: 0.00 to 154464.00

action - user_awardee_karma:
  Mean: 2089.47
  Median: 76.00
  Distribution: right-skewed
  Range: 0.00 to 36848.00
```



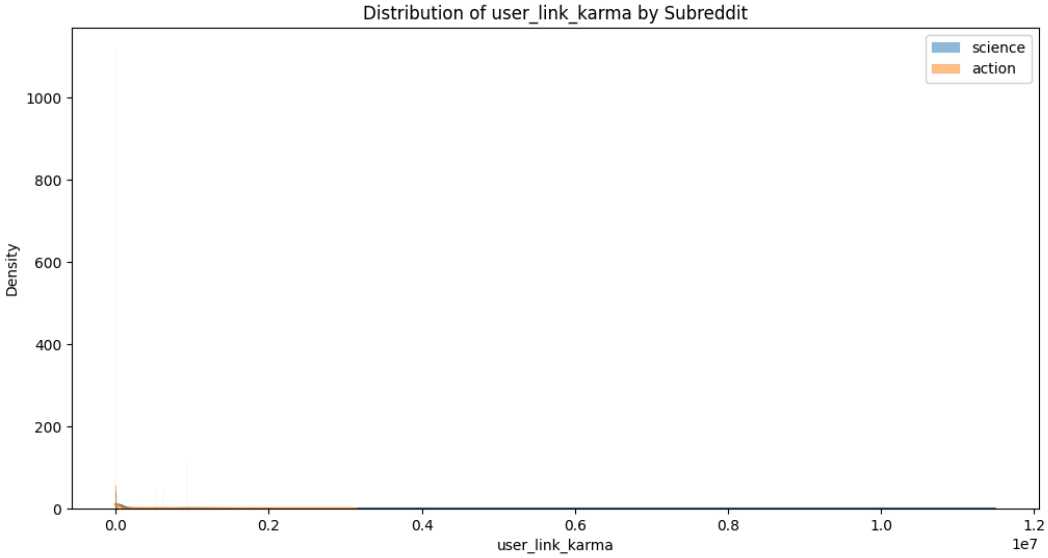
```
science - user_awarder_karma:
  Mean: 605.58
  Median: 0.00
  Distribution: right-skewed
  Range: 0.00 to 43075.00

action - user_awarder_karma:
  Mean: 575.29
  Median: 0.00
  Distribution: right-skewed
  Range: 0.00 to 53073.00
```



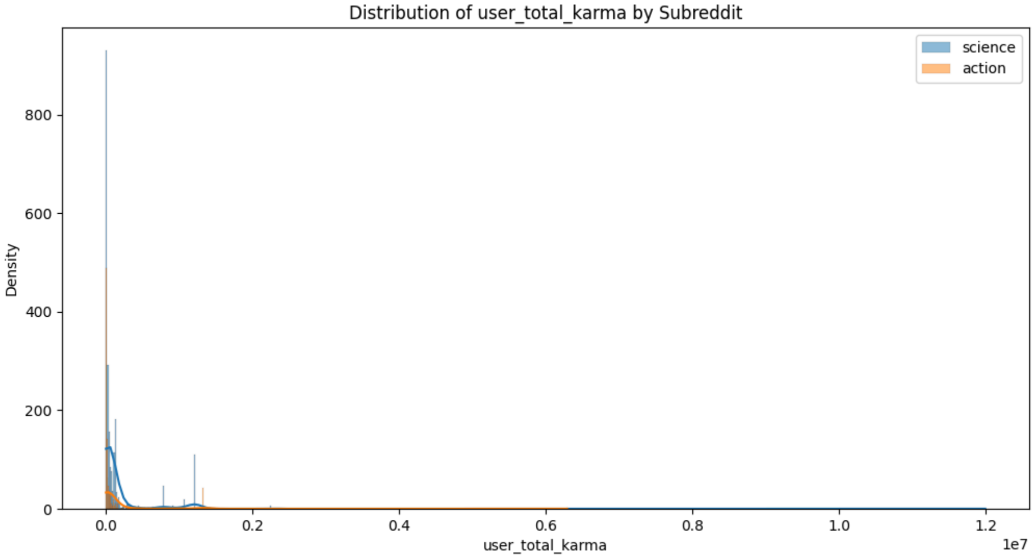
science - user_comment_karma:
  Mean: 66500.80
  Median: 22298.00
  Distribution: right-skewed
  Range: -100.00 to 1015526.00

action - user_comment_karma:
  Mean: 77849.73
  Median: 7851.00
  Distribution: right-skewed
  Range: -100.00 to 3640897.00



```
science - user_link_karma:
  Mean: 90178.09
  Median: 1058.00
  Distribution: right-skewed
  Range: 0.00 to 11490136.00

action - user_link_karma:
  Mean: 47975.48
  Median: 878.00
  Distribution: right-skewed
  Range: 0.00 to 3141592.00
```

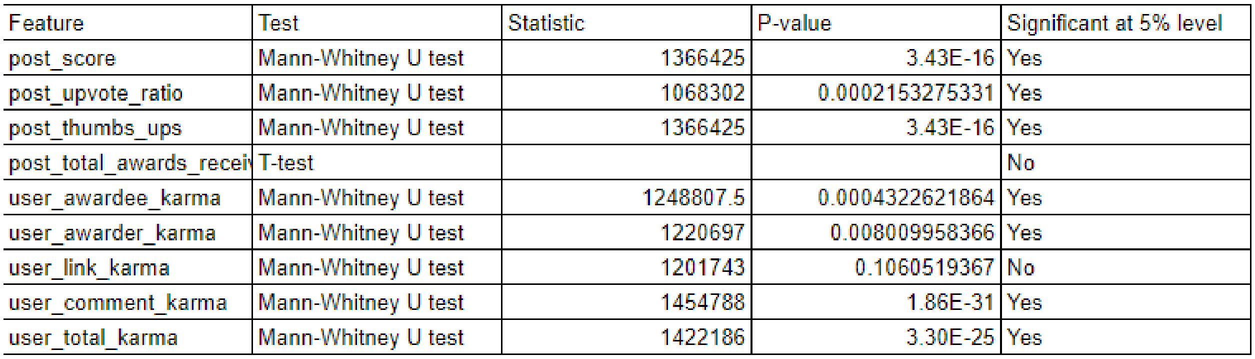


```
science - user_total_karma:
  Mean: 159228.75
  Median: 28091.00
  Distribution: right-skewed
  Range: -99.00 to 11990059.00

action - user_total_karma:
  Mean: 128489.97
  Median: 11608.00
  Distribution: right-skewed
  Range: -80.00 to 6287054.00
```

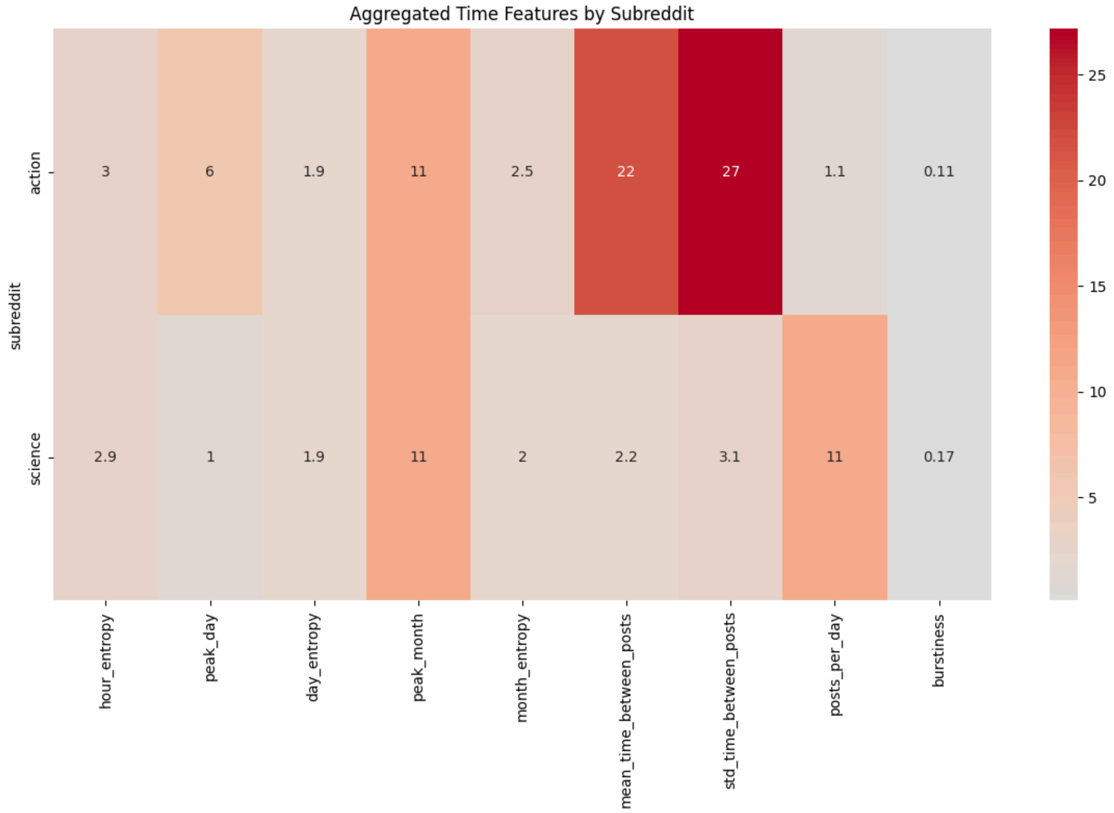
Task 3: Statistical Significance (25 points)

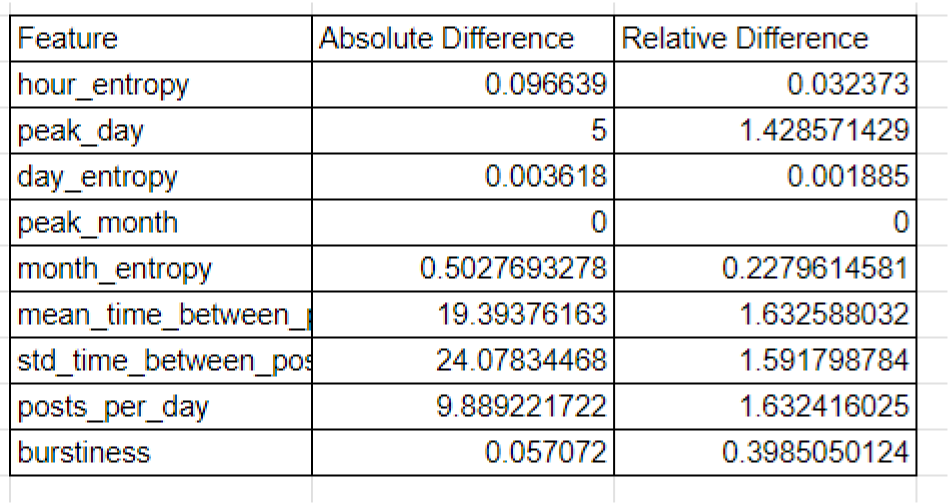
For each quantitative feature from task 1, assess the statistical significance (at 5% level) among the two subreddits. Refer to the distributions of each feature in Task 2 to determine which statistical test would be the most appropriate, for example, if the data follows a near-symmetric distribution them the t-test might be the most appropriate. On the other hand, if the data follows a highly skewed distribution, then a nonparametric test will be appropriate.



Feature	Test	Statistic	P-value	Significant at 5% level
post_score	Mann-Whitney U test	1366425	3.43E-16	Yes
post_upvote_ratio	Mann-Whitney U test	1068302	0.0002153275331	Yes
post_thumbs_ups	Mann-Whitney U test	1366425	3.43E-16	Yes
post_total_awards_recei	T-test			No
user_awardee_karma	Mann-Whitney U test	1248807.5	0.0004322621864	Yes
user_awarder_karma	Mann-Whitney U test	1220697	0.008009958366	Yes
user_link_karma	Mann-Whitney U test	1201743	0.1060519367	No
user_comment_karma	Mann-Whitney U test	1454788	1.86E-31	Yes
user_total_karma	Mann-Whitney U test	1422186	3.30E-25	Yes

<u>Task 4: Feature Computation/Engineering (75 points, 25 points for feature aggregation methods, 25 points for computing the features, and 25 points for testing statistical significance)</u> The time stamps of when the post, comments, and user accounts are created are strings. Each string will be unique, and it would not be informative to include these individual, unique strings in a ML model. All the data sets include the creation times for the posts. Devise a method to aggregate the post creation times into meaningful, compact features suitable to be fed into machine learning models. Compute these features for the post creation times and test their statistical significance for the two subreddits in your data set. This applies to all the projects, regardless of whether the project uses post-level or comment-level data.





Feature	Absolute Difference	Relative Difference
hour_entropy	0.096639	0.032373
peak_day	5	1.428571429
day_entropy	0.003618	0.001885
peak_month	0	0
month_entropy	0.5027693278	0.2279614581
mean_time_between_p	19.39376163	1.632588032
std_time_between_pos	24.07834468	1.591798784
posts_per_day	9.889221722	1.632416025
burstiness	0.057072	0.3985050124