Assignment 1, CSE 474/574

The number of points per question are in parentheses here (but not in the jupyter

notebook).

Notes on grading:

• For 474, the points here add up to 75. The remaining 25 will be based on

code spot checks.

• For 575, the points here add up to 90. The spot checks will be worth 30

points, and we will then normalize your score out of 120 to get a final

grade out of 100.

**Part 1.1 - Understanding APIs (5 points)**

• **1.1.1 (2)** How many API calls were required to collect the submissions?

The process of a client application submitting a request to an API and the API getting the desired data from an external server is known as an API call.

Only, 60 request can be made in a minute and per request we can fetch 100 items.

In the given question, we had to fetch 2991 items so total requests that we required to make was 30.

• **1.1.2 (1)** Why did we set the submission limit at 1000?

To understand the rules which we must adhere to while making the API calls

For machine learning project we need more data to train the algorithm so we tried to fetch more than 1000 post at a time but we couldn’t as it’s the limit set by reddit API. The limit is set to such a number because a normal user wouldn’t look more than 1000 top posts in a subreddit. Therefore, post over 1000 wouldn’t impact and will over fit our model

so obtain more than 1000 we have to use push API, which is third party API.

• **1.1.3 (2)** How long, in minutes, would it take you to collect 1000 posts

from 25 different subreddits? What about from 500 different subreddits?

*Hint: You’ll have to consider how many API requests you are allowed to*

*Make*

*As par rules for API, we can make at max 60 request and each request can fetch up to 100 post so in a minute we can fetch at max 6000 items so in 4 minute we can fetch 24000 post and rest thousand will require 0.003 sec. So it takes almost 4 min to fetch 25000 items from praw.*

*And similarly for 500 subreddit it will take around 83 minutes to fetch.*

**Part 1.2 Thinking about your sample (3 points)**

• **1.2.1 (1)** Do you think these posts are representative of **all** the posts on

that subreddit?

Yes, as most normal users wouldn’t look at post over 1000 so this data represents all the top viewed post on a particular subreddit.

• **1.2.2 (2)** Why or why not? That is, if you think so, why do you think

there’s not much sampling bias here? If not, what do you think might be

different about these top posts than other posts?

As most users will be looking at top 1000 post most of the time so post over 1000 is not much related and wont be affecting more model in any ways and in 1000 top post are post were most people reacted to it which it was most relevant to them or were useful for them.

**Part 2.1 - Univariate descriptive analyses (13**

**points)**

• **2.1.1 (1)** What are the names (subreddit\_name\_prefixed) of the 25

different subreddits that are in part2\_data.csv?

The 25 different subreddits chosen were :

'Jokes', 'news', 'science', 'WritingPrompts',

'Showerthoughts', 'worldnews', 'todayilearned',

'learnprogramming', 'announcements', 'funny', 'food',

'sports', 'gadgets', 'aww', 'mildlyinteresting', 'memes',

'technology', 'travel', 'books', 'gaming', 'cats',

'conspiracy', 'PoliticalHumor', 'hockey'

• **2.1.2 (3)** How many reddit authors (author\_name) have a post in more

than one unique subreddit in part2\_data.csv (e.g. they have a top post

in both r/news and r/hockey)?

569 Authors were found to have a post in more than one unique subreddit.

• **2.1.3 (1)** What is the mean number of upvotes (ups) for posts in r/Jokes?

41057.78 is the mean number of upvotes for posts in r/Jokes

• **2.1.4 (1)** What is the variance of the number of upvotes in r/news?

variance of the number of upvotes in r/news is 600,707,867.6254

• **2.1.5 (2)** What is the standard deviation of the number of upvotes received

across the entire dataset?

Standard Deviation of the number of upvotes received is 43102.484474

• **2.1.6 (1)** (No code for this) Mathematically, what is the relationship

between the standard deviation of the number of upvotes and the variance

of upvotes?

The relation between the standard Deviation and variance of upvotes is :

600,707,867.6254 = (24509.342456)\*\*2

S tan d a r d space D e v i a t i o n open parentheses u p v o t e s close parentheses equals space square root of V a r i a n c e open parentheses u p v o t e s close parentheses end root

• **2.1.7 (1)** Which subreddit had the third highest median number of upvotes?

'r/aww' is the sub reddit with third highest median number of upvotes.

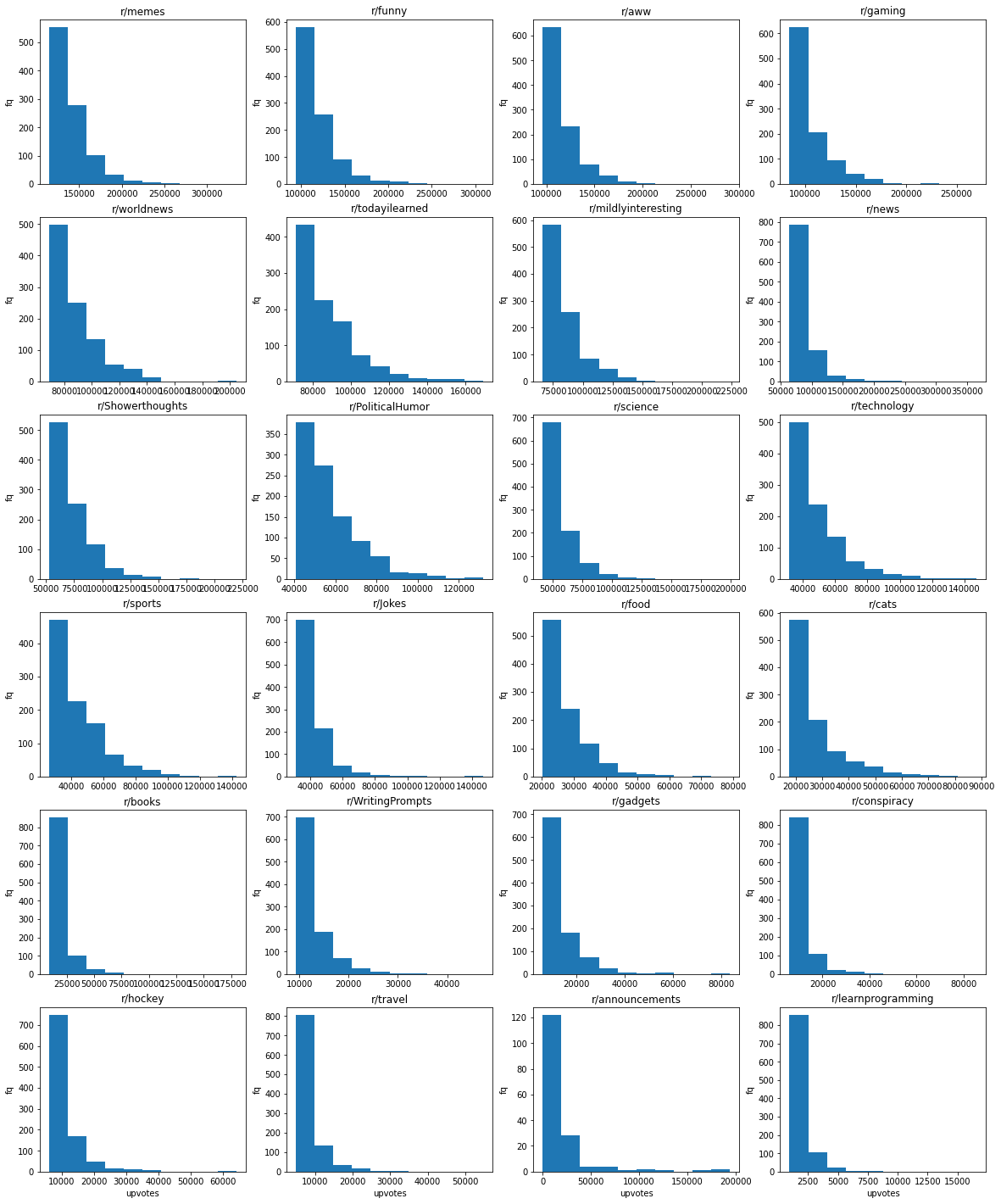
• **2.1.8 (3)** What is the conditional probability of an author having a top

post in r/news, given that they have a top post in r/worldnews?

1. P(r/news|r/worldnews) = 0.100763

**Part 2.2 - Plotting (12 points)**

• **2.2.1 (3)** - Submit your histogram image in your assignment



• **2.2.2 (2)** - Based on your histogram, which subreddit would you say is the

*least* popular? (Note, there is more than one reasonable answer here. We

are looking mostly for how you justify your response using the histogram)

learn Programing explanation

As “learnprogramming” has 800 posts with only 2500 upvotes which is far less when compared to other subreddits.

• **2.2.3 (2)** - **Approximately (within 1-2 percentage points)** what

percent of top posts for each of the three subreddits plotted below have

less than 100,000 upvotes? (Give answers for each subreddit)s

News with 84.40644%

Science with 98.5873%

Worldnews with 79.27565%

• **2.2.4 (2)** - **Approximately (within 1-2 percentage points)** what is

the probability that a post on each of the three subreddits plotted below

has more than 70,000 upvotes? (Give answers for each subreddit)

P(post=science > 70,000 upvotes) = 1 - 0.8758829 = 0.1241171

P(post= world news > 70,000 upvotes) = 1 - 0.0331992 = 0.9668008

P(post= news > 70,000 upvotes) = 1 - 0.2706237 = 0.7293763

• **2.2.5 (1)** - How many posts in the dataset were sent in 2010?

In 2010, there were only 35 posts sent as per the dataset given.

• **2.2.6 (2)** - In your report, provide a table (a screenshot of a pandas

dataframe is fine) that shows the average number of upvotes for r/memes

each year from 2015 to 2020. The table should be sorted by year (i.e. 2015,

then 2016, etc.). Note again, if a year does not have data, there should be

zeros in this table!

Table

Description automatically generated

• **2.2.7 (3)** - Plot a line graph of the temporal trend of mean upvotes from

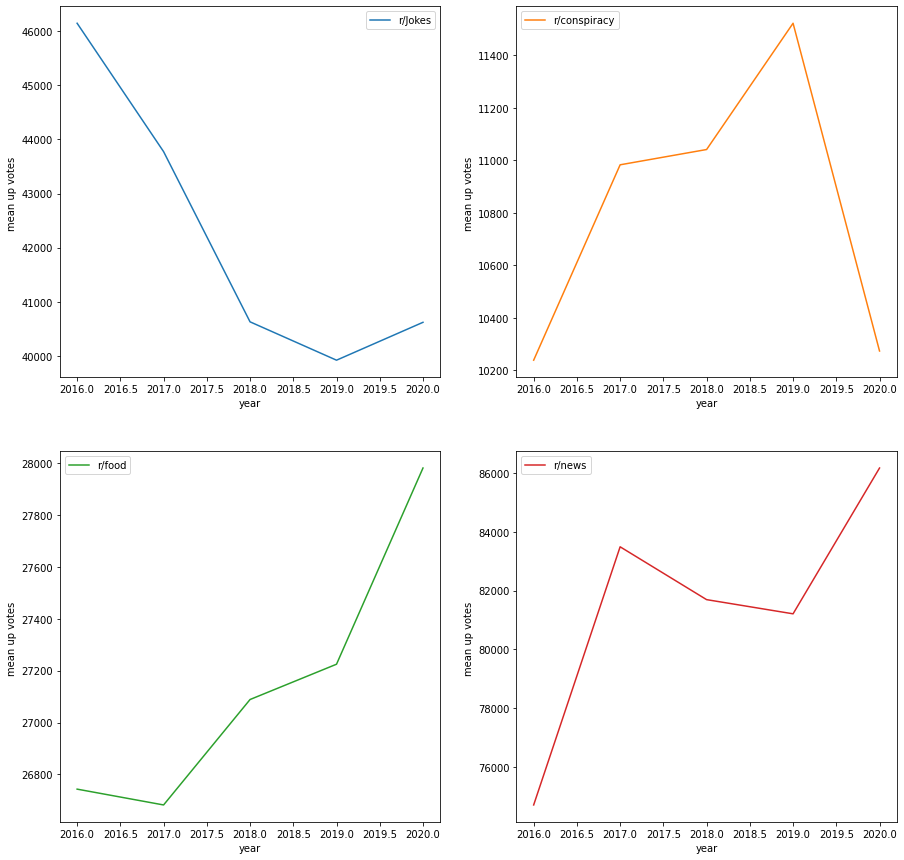
2016-2020 for the following subreddits: r/Jokes, r/food,r/conspiracy, and

r/news . You can plot them individually, or use the faceting approach from

above. Write your code for this in the cell below; copy the resulting plot

to your PDF report. **Hint: Doing part 2.2.8 will be easiest if you**

**make sure that the plot for each subreddit has its own y-axis!**.



• **2.2.8 (2)** - Using what you have plotted, make an argument for which of

the four subreddits is the most “up and coming” - i.e. the one that seems

to be getting more popular over time. NOTE: There is more than one

reasonable answer here. We are looking for how you justify your answer

using the (plotted) data.

r/food and r/news can be considered as the most “up and coming” subreddits.

r/food is the most “up and coming” as the trend of upvotes had been consistently increasing from 2017 to 2020.

And when coming to r/news even though it had a decline in the number of upvotes between the years 2017-2019, it also has a highly significant increase in the number of upvotes in the year 2016-2017 and recently in 2019-2020 to be considered as “up and coming”.

**Part 2.3 - Data Cleaning & Regression-related**

**Analyses (14 points)**

• **2.3.1 (2)**- There are two continuous variables that are very clearly not

going to be useful for our analysis. Identify them, and explain why they are

not useful (**note: you do NOT need to know why these variables**

**take on the values they do in our data. You just need to know**

**why we don’t want to use them!**)

• **2.3.2 (2)**- There are two (supposedly) binary variables that are very clearly

not going to be useful for our analysis. Identify them, and explain why

they are not useful.

• **2.3.3 (2)** - Explain why we it is not useful to use *both* subreddit\_id and

subreddit\_name\_prefixed in any predictive analysis of per-post upvotes.

• **2.3.4 (2)** - Explain why it is not useful to use permalink in any predictive

analysis of per-post upvotes.

• **2.3.5** - Plot the relationship between num\_comments and upvotes as a

scatterplot with log-scaled axes, with the posts from different subreddits

as different color points. Paste this plot into your PDF writeup

• **2.3.6 (2)** - Describe, briefly (a sentence) the relationship between

num\_comments and upvotes.

• **2.3.7 (2)** - Which of these has the strongest positive correlation with ups?

• **2.3.8 (2)** - Which of these has the weakest positive correlation with ups?

**Part 3.1 - Regression Basics (23 points)**

• **3.1.1 (5)** - Report your error on the test data, in RMSE. State what this

metric means for the expected error in terms of the number of upvotes

(not log upvotes!) you should expect to be off on any given prediction

• **3.1.2 (2)** - What did the whole one-hot encoding thing on

subreddit\_name\_prefixed actually do?

• **3.1.3 (1)** - What does the argument drop = "first" do for us when we

are doing that to subreddit\_name\_prefixed?

• **3.1.3 (1)** - Why did we need to add one to the outcome variable before

using log?

• **3.1.4 (3)** - What does the StandardScaler do? Why do we want to do

that?

• **3.1.5 (4)** - Provide a scatterplot that compares the true values in y\_test

to the absolute value of the difference between y\_test and your predictions.

**The axes should be on the original scale** (i.e. not the log scale you’re

predicting on.

• **3.1.6 (2)** - What does this plot suggest about how well your model fits

the data as the true number of upvotes changes?

• **3.1.7 (3)** - What is the new RMSE with the logged independent variables?

• **3.1.8 (2)** - How did this compare to the old RMSE? Why do you think

that is? Hint: It may help to re-plot the same figure as you did in 3.1.5,

3

but with the new model, in order to answer this question.

**Part 3.2 - Interpreting Regression Coefficients (5**

**points)**

• **3.2.1 (3)** - What is the strongest positive predictor of upvotes? How many

more log(upvotes+1) does a one standard deviation increase in the feature

correspond to?

• **3.2.2 (2)** - What is the strongest negative predictor of upvotes? How

many fewer log(upvotes+1) does a one standard deviation increase in the

feature correspond to?

**Part 3.3 - 574 Only - Attempting to Improve Your**

**Predictions**

• **3.3.1 (10)** - Describe at least two changes you made – at least one to

the feature set, and at least one different model – to try to improve

prediction. Explain *why* you think that these changes make sense, given

the Exploratory analyeses above, or any other exploratory analysis you

choose to do.

• **3.3.2 (5)** - By how much did your RMSE improve? Which change that

you made improved it the most? How do you know?

4