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?

1. 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 requests can be made in a minute and per request, we can fetch 25 items.

In the given question, we had to fetch 2991 items so the total requests that we were required to make were 120.

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

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

For training a Machine learning model, we need more data for the model to train on. So we tried to fetch more than 1000 posts 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 at more than 1000 top posts in a subreddit. Therefore, post over 1000 wouldn’t impact and will overfit our model.

To 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*

1. As per rules for API, we can make a max of 60 requests per minute and each request can fetch up to 25 posts so in a minute we can fetch a max of 1500 items. So it takes almost 4 min to fetch 25000 items from praw.

And similarly, for 500 subreddit, it will take around 333 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?

1. 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?

1. As most users will be looking at the top 1000 posts most of the time, so post over 1000 is not much related and won’t be affecting our model in anyways and these 1000 top posts are the posts where most people reacted to it and were 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?

1. 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)?

**A.** 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?

1. 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?

1. 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?

1. 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?

1. 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?

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

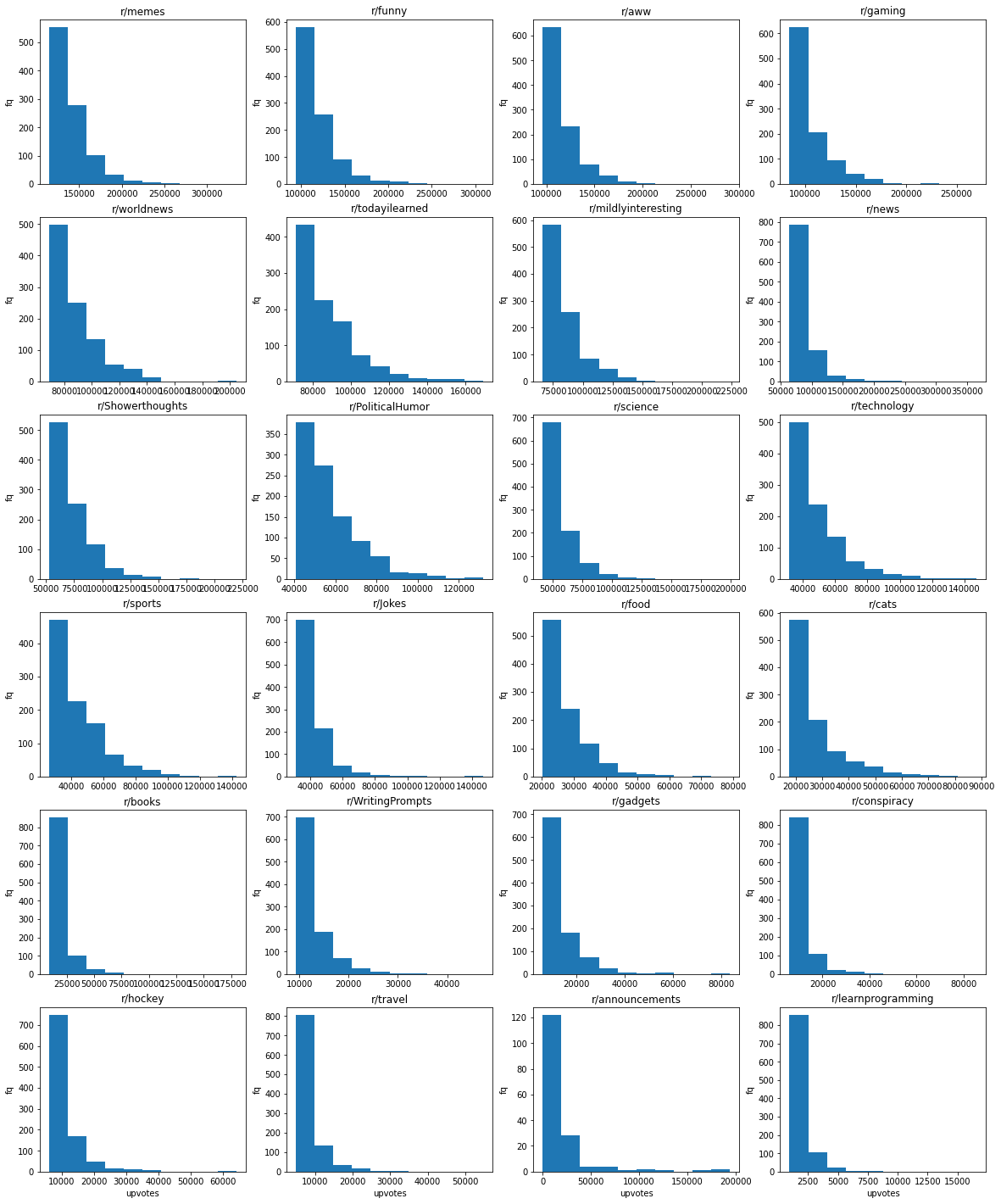
• **2.1.8 (3)** What is the conditional probability of an author having a top

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)

1. 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

**A.**

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)

**A.**

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?

1. 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

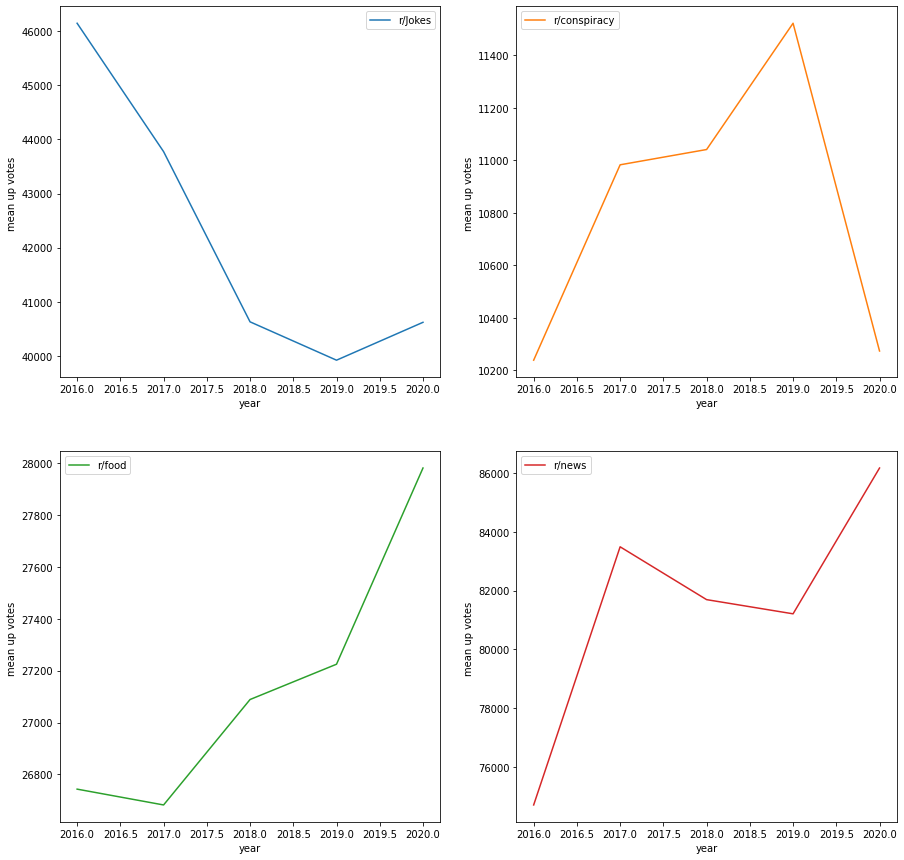
2016-2020 for the following subreddits: r/Jokes, r/food,r/conspiracy, and

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.

**A.**

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!**)

1. “downs” and “num\_reports” have no data i.e the mean and Variance of “downs” is 0 and “num\_reports” has mean and variance as “Nan”, so these are clearly not going to have an impact on our target variable “Upvotes”.

• **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.

The two binary variables that clearly not going to be useful are ‘media\_only’ and ‘is\_crosspostable’ as they all have only one unique value which certainly doesn’t impact our target variable.

• **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.

1. As both subreddit\_id and subreddit\_name\_prefixed are just the identifiers to uniquely identify a post in a subreddit.

When we compare two posts, one with the least number of upvotes and one with the highest number of upvotes in a subreddit, they both have the same subreddit\_id and subreddit\_name\_prefixed which would not help in any predictive analysis of per-post upvotes.

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

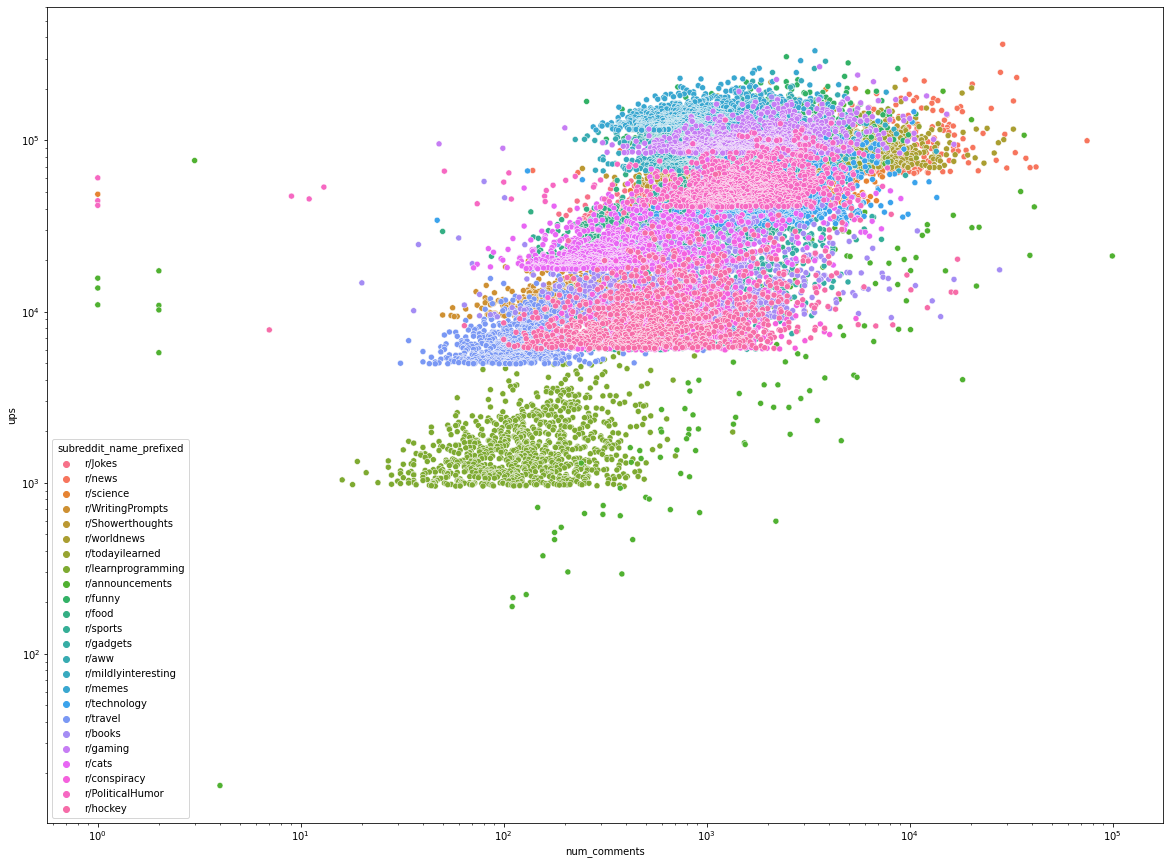
analysis of per-post upvotes.

1. As every link of a post is unique and no posts share a common permalink and this would not help us 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.

1. Num\_comments is positively correlated with upvotes with a correlation coefficient of 0.33

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

1. Score has the strongest correlation with ups with a correlation coefficient of 0.99

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

1. Is\_self has the weakest correlation with coefficient -0.30481653929867814

**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

1. RMSE : 32288.660237022763

So 32289 is the expected error to occur when we try to predict a the number of upvotes for a given post, which means that the predicted value by our model can be off by 32289 votes for a given post.

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

subreddit\_name\_prefixed actually do?

1. OneHotEncoder creates a binary transform of “subreddit\_name\_prefixed” because our linear model cannot take the inputs as strings, these categorical values have to be transformed to numerical values and this is where OneHotEncoder helps solve this problem.

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

are doing that to subreddit\_name\_prefixed?

1. drop=”first” would drop the first categorical value in the column ” subreddit\_name\_prefixed” .

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

using log?

1. As we have zeros in our data, and log of zero is undefined. To handle this, we need to add 1 to the outcome variable.

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

that?

1. StandardScaler standardizes our data meaning all the values are subtracted with the mean of the columns to which the value belongs, and the result is divided by the standard deviation of that column.

We want to standardize our data to maintain internal consistency.

And in our data, many of the features have large differences in their ranges and are of different units.

• **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.

Chart, scatter chart

Description automatically generated

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

the data as the true number of upvotes changes?

1. Data fits well as we can see the difference between the y-test and the predicted value seem to concentrate toward the left side of the graph, which suggests that our model predicts posts with less than 50,000 upvotes really good.

But, as the number of upvotes goes above 50,000 our model tends to fail.

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

**A.** New RMSE with logged independent variables :

0.332433787220992

• **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,

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

Chart, scatter chart

Description automatically generated

The blue points represent the new plot and the orange represents the old plot, as we can see in the figure the new points are now a little concentrated towards the left when compared to the old ones.

**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 analysis above, or any other exploratory analysis you

choose to do.

1. Two reduce the noise in the dataset we removed the outliers in the dataset that is first 25 percentile and top 25 percentile.

So this step has helped us to improve the quality of the dataset so that models can train good quality datasets.

The model that we implemented was Random forest this model uses ensembled techniques which is the reason why the RMSE has improved significantly.

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

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

1. The RMSE was improved by 7549.991558994174 the most important step was reducing the noise as the model was trained on the more refined dataset.

We came to know this after removing the outliers and again implementing linear regression which improved the RMSE significantly.