Programming Assignment #1 – Data Manipulation and Some Regression

Part 1.1 - Understanding APIs (5 points)

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

API can return data up to 25 most recent posts but we could extend this limit to 100 per one playing. so we need 30 API calls to collect the submissions.

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

to get top 1000 posts from each of the 3 subreddits

• **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*

2 seconds are needed to request one API. If we need collecting 1000 posts from 25 different subreddit s, we have to call 250 numbers of API. Therefore, we need about 8 and half minutes(500 seconds). Fu rthermore, we need about 17 minutes(1000seconds) to collect 1000 posts from 500 different subreddi ts.

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?

NO

• **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?

The top 1000 posts are just famous posts that are recently posted or viewed many times, and it canno t represent ALL possts

Part 2.1 - Univariate descriptive analyses (13points)

• **2.1.1 (1)** What are the names (subreddit_name_prefixed) of the 25 different subreddits that are in part2_data.csv?

['r/Jokes' 'r/news' 'r/science' 'r/WritingPrompts' 'r/Showerthoughts'
'r/worldnews' 'r/todayilearned' 'r/learnprogramming' 'r/announcements'
'r/funny' 'r/food' 'r/sports' 'r/gadgets' 'r/aww' 'r/mildlyinteresting'
'r/memes' 'r/technology' 'r/travel' 'r/books' 'r/gaming' 'r/cats'
'r/conspiracy' 'r/PoliticalHumor' 'r/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**

• **2.1.3 (1)** What is the mean number of upvotes (ups) for posts in r/Jokes? **41057.7813440321**

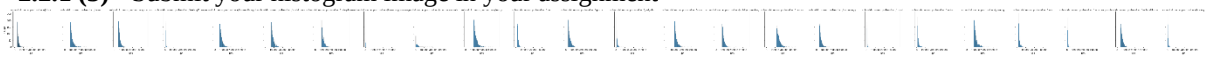
• **2.1.4 (1)** What is the variance of the number of upvotes in r/news? **600707867.6203133**

• **2.1.5 (2)** What is the standard deviation of the number of upvotes received across the entire dataset? **43102.4844737104**

• **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 standard deviation of the number of upvotes is the square root of the variance of upvotes.**

• **2.1.7 (1)** Which subreddit had the third highest median number of upvotes? **109811.0**
• **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? **0.2012072434607646**

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)

learnprogramming. The graph which goes to the right of the x-axis means that upvotes for posts are high. On the other hand, the graph which goes high of the y-axis means the counts of ups are high. 'learningprogramming' ups leans to the left, and it has the lowest value so we can say that it's the least popular subreddit.

• **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)

84%, 98%, 79%

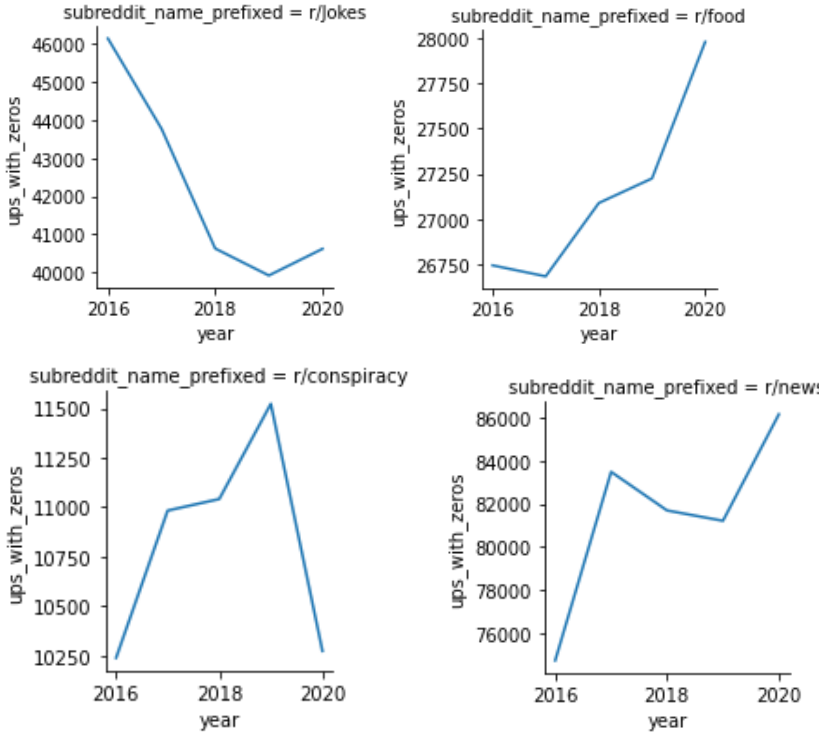
• **2.2.4 (2)** - **Approximately (within 1-2 percentage points)** what is the probability that a post on each of the three subreddits plotted below has more than 70,000 upvotes? (Give answers for each subreddit)

71%, 13%, 93%

• **2.2.5 (1)** - How many posts in the dataset were sent in 2010? **35**
• **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!

	year	ups_with_zeros
183	2015	0.000000
111	2016	0.000000
39	2017	0.000000
87	2018	131206.000000
63	2019	135859.126984
15	2020	141141.427305
135	2021	138620.820225
159	2022	127305.750000

• **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.

The most "up and coming" subreddit will mean the one which got the highest ups rising rate compar ed to the recent past.

For Jokes, they got 40000 ups in 2019 and 41000 in 2020. It means the recent ups rising rate is about 1.03%(41000/40000).
For food, they got about 27000 ups in 2019 and 28000 in 2020. It means the recent ups rising rate is a bout 1.04%.
For conspiracy, they got about 11500 ups in 2019 and 10250 in 2020. It means the recent ups rising r ate is about 0.89%.
For news, they got about 82000 ups in 2019 and 86000 in 2020. It means the recent ups rising rate is about 1.05%.
The highest ups rising rate was r/news's, so we could tell r/news is most "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!**)

downs, num_reports => downs feature has too many zero, and num_reports feature has too many NaN values.

• **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.

is_crosspostable, media_only => is_crosspostable, media_only features only have 'False' values.

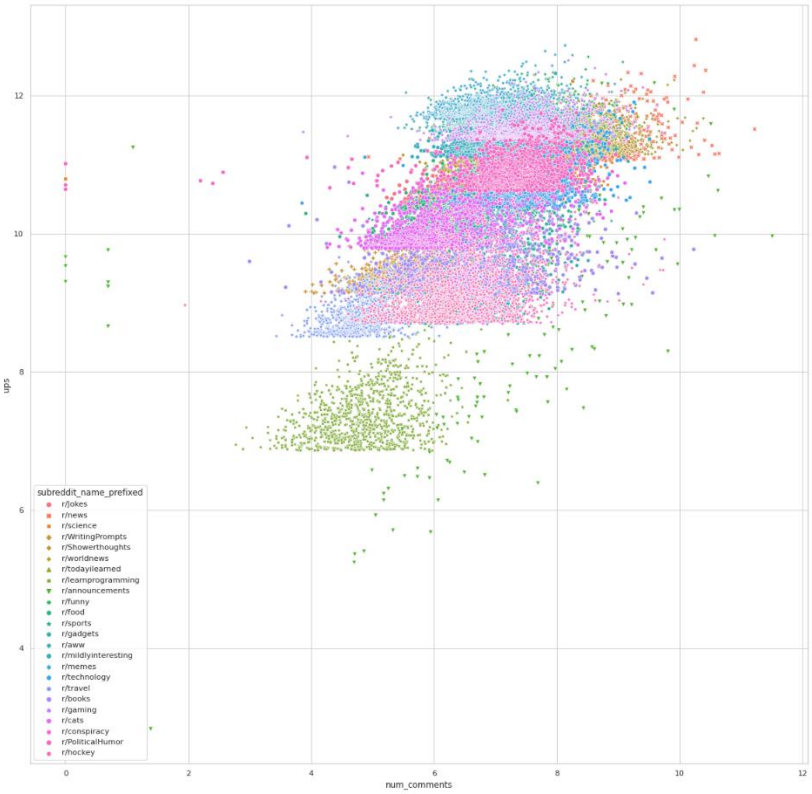
• **2.3.3 (2)** - Explain why we it is not useful to use *both* subreddit_id and subreddit_name_prefixed in any predictive analysis of per-post upvotes.

subreddit_id and subreddit_name_prefixed features have identical meaning. If we use both features, it means we use duplicated features.

• **2.3.4 (2)** - Explain why it is not useful to use permalink in any predictive analysis of per-post upvotes.

permalink feature doesn't have any numerical values, and it is unrelated with per-post upvotes. It only contains the hyperlink of the data, so it would be not helpful in 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.

It has positive relationship, which means when the num_comments values increase, the ups values also increase.

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

num_crossposts has the strongest positive correlation with ups, 0.5379816522109334.

• **2.3.8 (2)** - Which of these has the weakest positive correlation with ups? **created_utc has the weakest positive correlation with ups, 0.16547147438976492**

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

Our RMSE is 0.37. It is the low expected error on the test data, and the number of upvotes are well

predicted.

• **3.1.2 (2)** - What did the whole one-hot encoding thing on subreddit_name_prefixed actually do?
It change data class convert categorical variables into another dataframe. For example,

subreddit_name_prefixed
r/Jokes
…
r/hockey'

Subreddit is converted into

r/Jokes	0	1	…	0
…				
r/hockey	1	0	…	0

• **3.1.3 (1)** - What does the argument drop = "first" do for us when we are doing that to subreddit_name_prefixed?

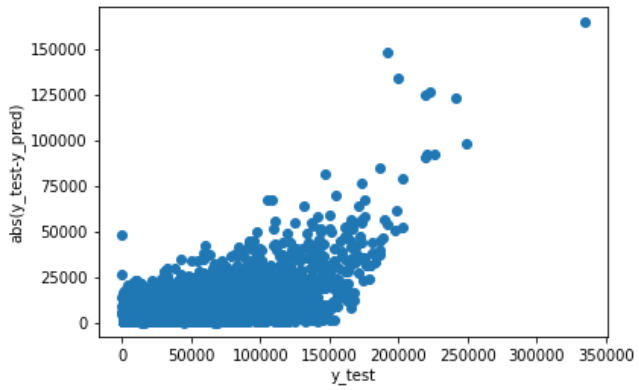
When we make linear regression model, we should make features one less. If there are too many parameters, the probabilities of failing in predicting will increase. To be specific, it is because if every variables are zero, it means the last category is '1'.

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

The reason for using log(x+1) transformation is to avoid log(x) approaching negative infinity as x approaches zero.

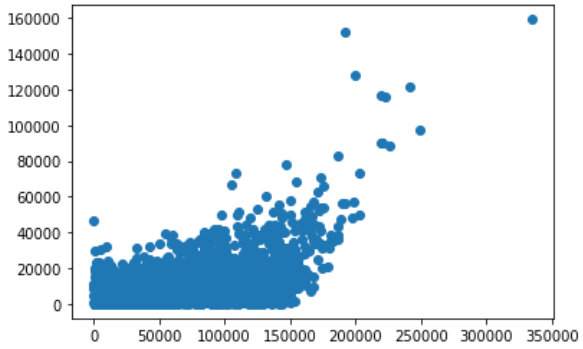
• **3.1.4 (3)** - What does the StandardScaler do? Why do we want to do that?
StandardScalar preprocess our data through mean and bias. In our code, we have to do that in order to standardization our continuous variables.

• **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?
The more linear the scatter plot is, the higher the accuracy is. Our graph is an upward-right graph generally.
• **3.1.7 (3)** - What is the new RMSE with the logged independent variables?
0.35327319741137153
• **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.



The new one has less error than old RMSE, so new RMSE is better one.

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?
The strongest positive predictor of upvotes is num_comments. The value of it is 0.653350.

• **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?
The strongest negative predictor of upvotes is r/learningprogramming, and its values is -0.547933