Programming Assignment #1 - Data manipulation and Some Regression

PART 1.1 - Understanding APIs

1.1.1 30 API calls are required to collect the submissions for 3 subreddits
1.1.2 Reddit only displays 1000 items. So we have to set the limit of 1000 for every subreddit.
1.1.3 For 25 subreddits - 8.33 mins (approx)
For 500 subreddits - 166.66 mins (approx)

PART 1.2 - Thinking about your sample

1.2.1 No
1.2.2 The top posts only represent the most upvoted posts in that subreddit which doesn't include -
a) Recent trending posts which haven't made it to the top of all time yet. E.g. r/politics became r/jokes for a couple of days to protest against the mods.
b) Downvoted/Controversial posts which although aren't the most upvoted but are part of the subreddit nevertheless.

There isn't necessarily a sample bias here since we aim to analyze and predict what gets a post the most upvotes for which the most upvoted posts are the right sample.

PART 2.1 Univariate Descriptive Analysis

2.1.1 The names of different subreddits are as follows:

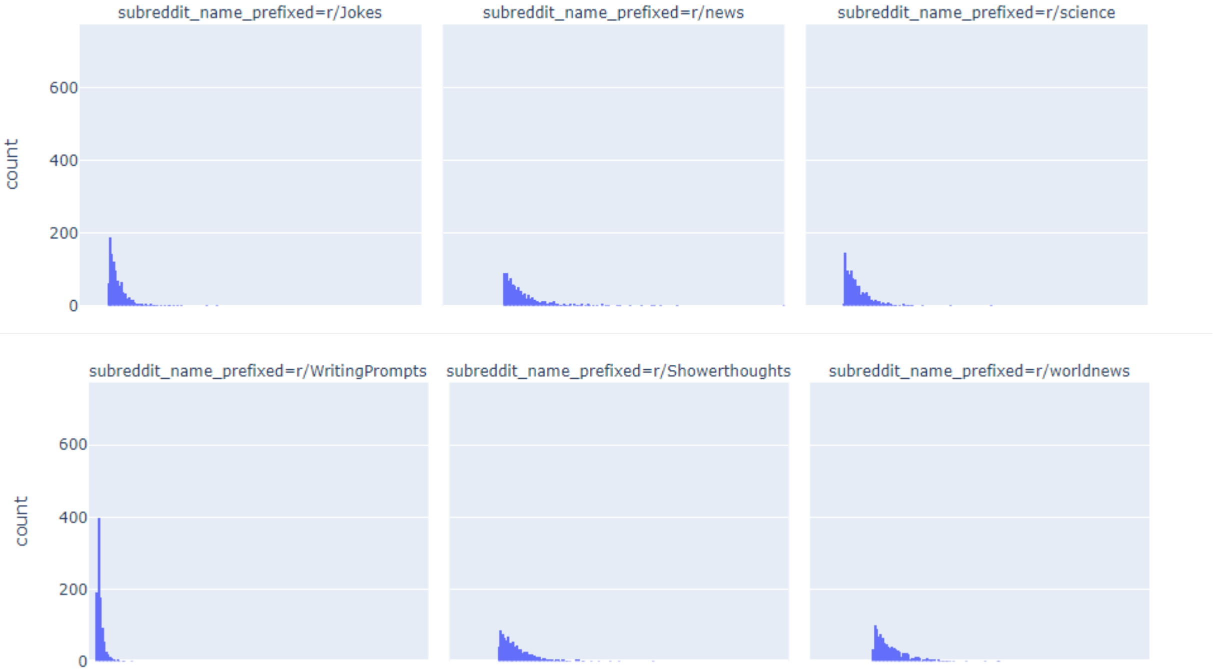
1. r/Jokes
2. r/news
3. r/science
4. r/WritingPrompts
5. r/Showerthoughts
6. r/worldnews
7. r/todayilearned
8. r/learnprogramming
9. r/announcements
10. r/funny
11. r/food
12. r/sports
13. r/gadgets
14. r/aww
15. r/mildlyinteresting
16. r/memes
17. r/technology

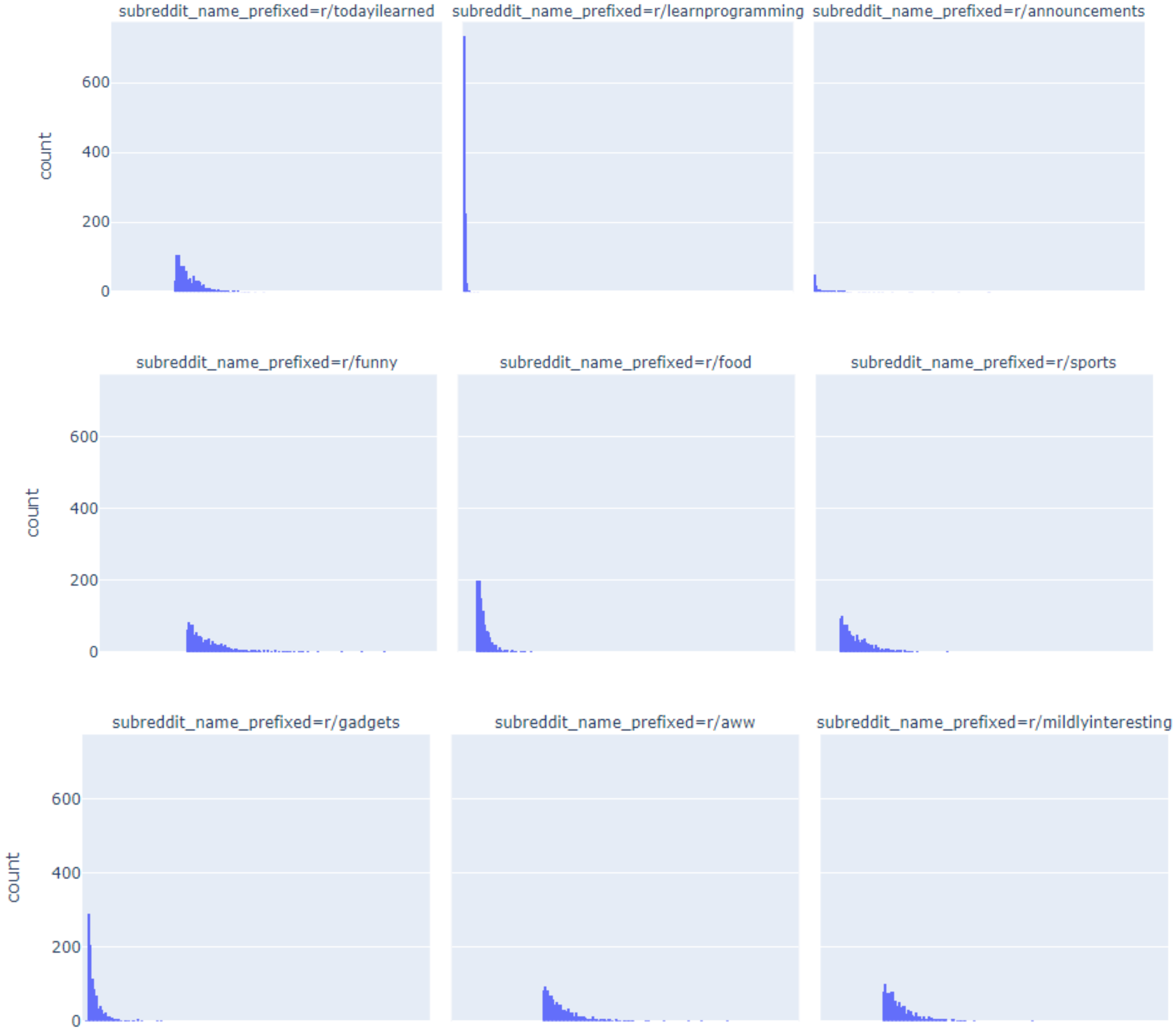
18. r/travel
19. r/books
20. r/gaming
21. r/cats
22. r/conspiracy
23. r/PoliticalHumor
24. r/hockey

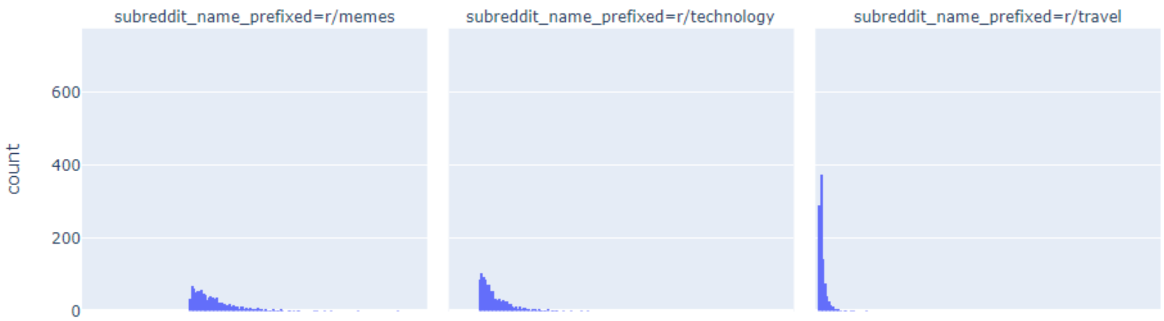
2.1.2 The reddit authors that have more than one unique subreddits in given data is 570.
2.1.3 The mean number of upvotes for posts in r/Jokes is 41057.7813440321.
2.1.4 The variance of the number of votes in r/news is 123497482.9039488
2.1.5 The standard deviation of the number of upvotes received across the entire dataset is 43102.48447371037.
2.1.6 The variance of upvotes is the square of the standard deviation of the number of upvotes.
2.1.7 The subreddit that has the third-highest median number of upvotes is r/aww.
2.1.8 Total number of authors in r/news and r/worldnews = 66
Total number of authors in r/worldnews = 654
Conditional probability = 66/654 = 0.100

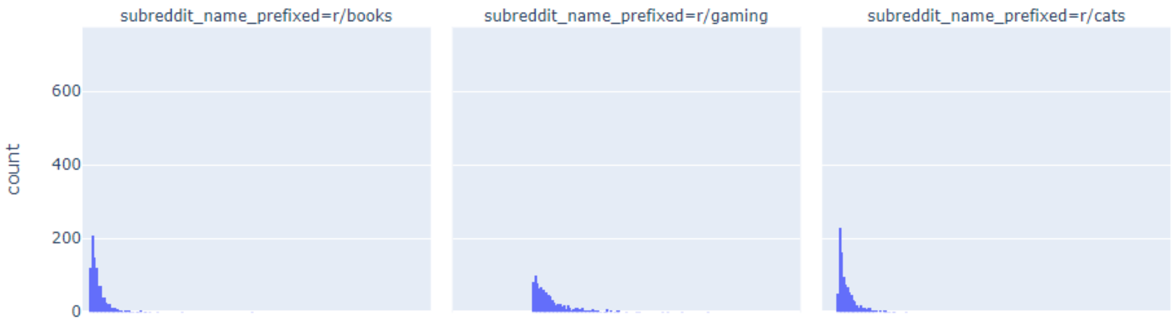
PART 2.2 Plotting

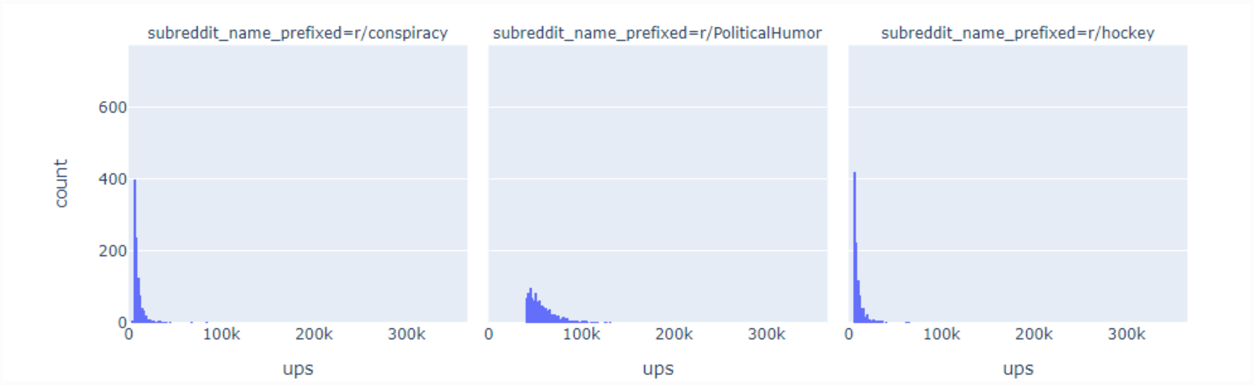
2.2.1




subreddit_name_prefixed=r/todayilearned
subreddit_name_prefixed=r/learnprogramming
subreddit_name_prefixed=r/announcements
count
600
400
200
0
subreddit_name_prefixed=r/funny
subreddit_name_prefixed=r/food
subreddit_name_prefixed=r/sports
count
600
400
200
0
subreddit_name_prefixed=r/gadgets
subreddit_name_prefixed=r/aww
subreddit_name_prefixed=r/mildlyinteresting
count
600
400
200
0







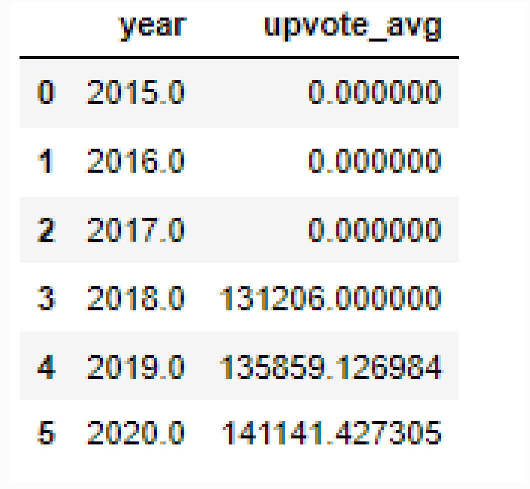
2.2.2 The least popular subreddit is r/learnprogramming

2.2.3 For r/news, the percentage is approximately 83%. For r/science, the percentage is approximately 98% For r/worldnews, the percentage is approximately 78%.

2.2.4 For r/news, the percentage is approximately 73%. For r/science, the percentage is approximately 13% For r/worldnews, the percentage is approximately 97%.

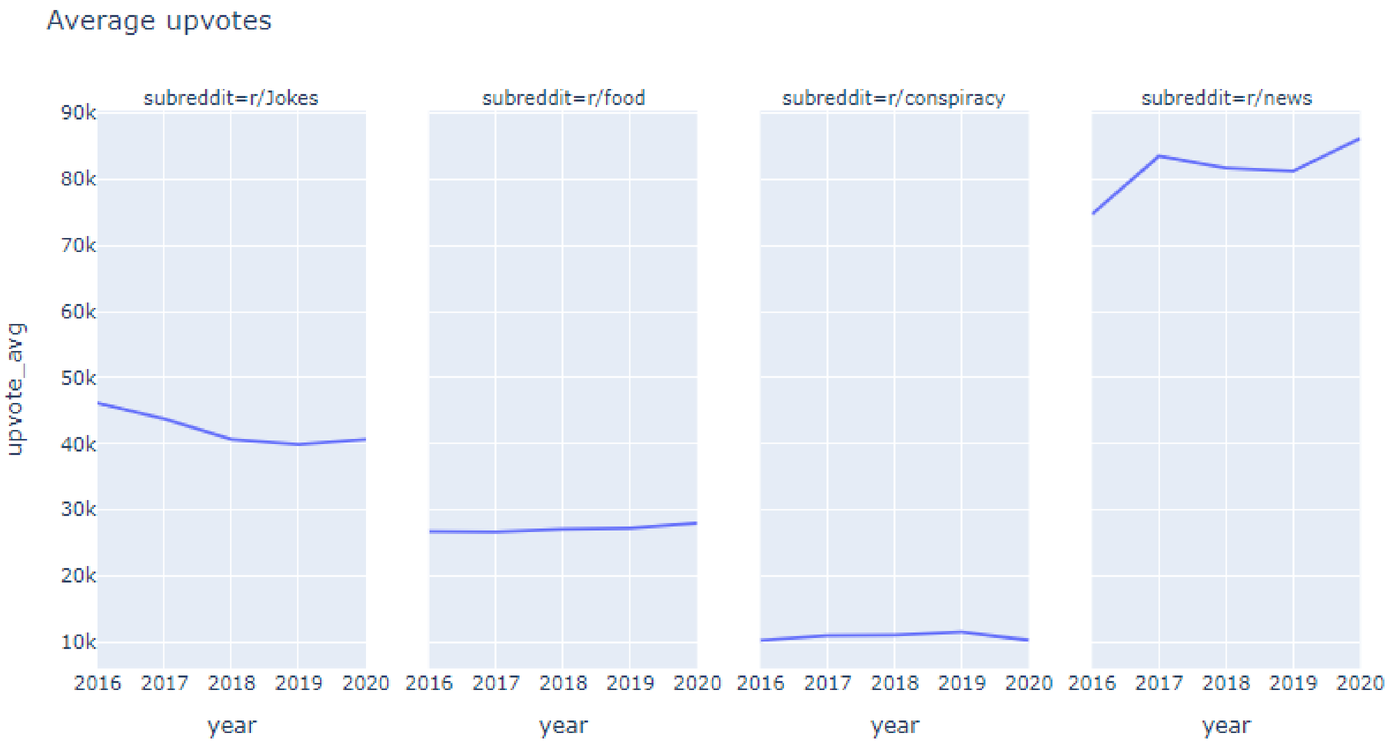
2.2.5 35 posts were sent in 2010

2.2.6 Following is the given table



	year	upvote_avg
0	2015.0	0.000000
1	2016.0	0.000000
2	2017.0	0.000000
3	2018.0	131206.000000
4	2019.0	135859.126984
5	2020.0	141141.427305

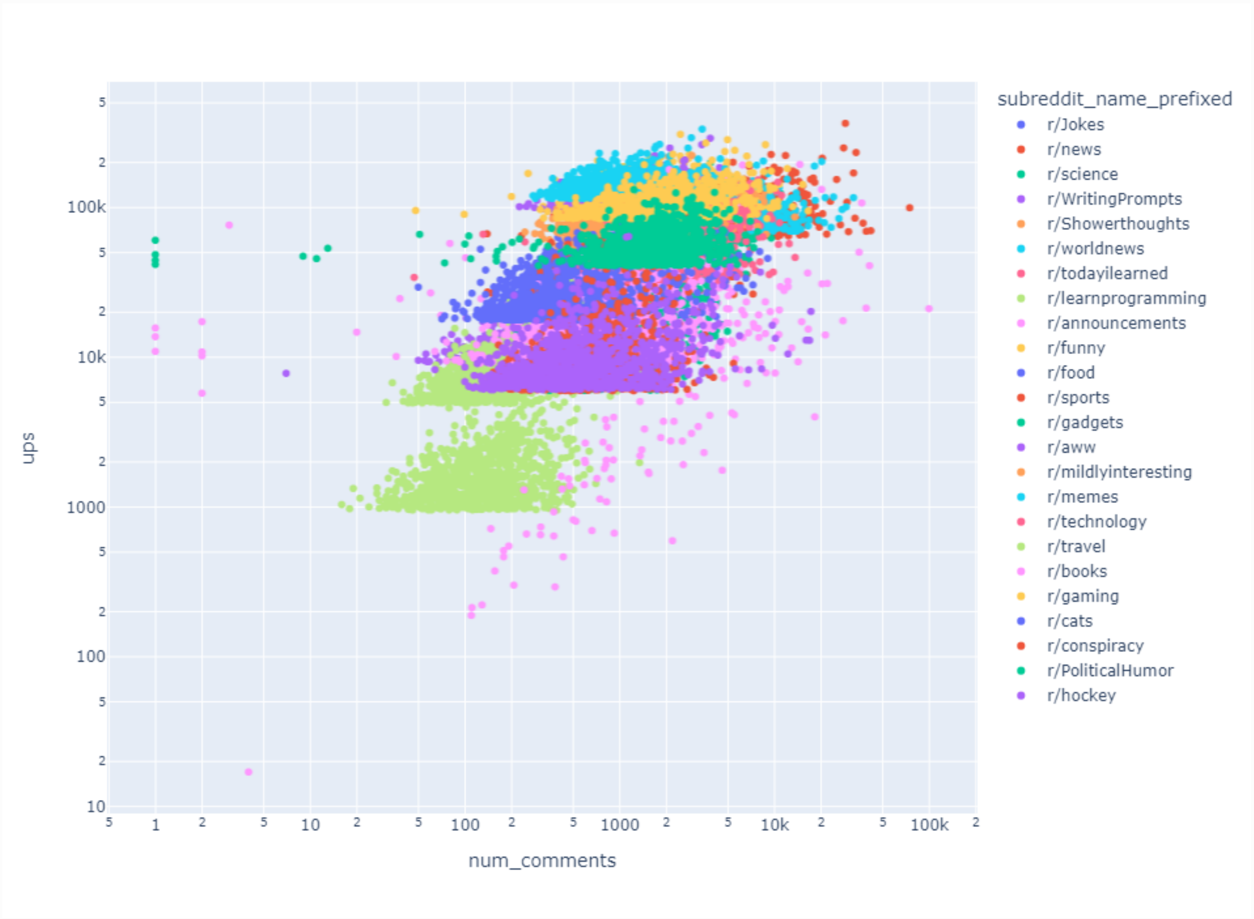
2.2.7 Resulting plot is as follows:



2.2.8 The most 'up and coming' among four different subreddits is r/news since its graph has the steepest(+ve) recent slope.

PART 2.3 - Data Cleaning & Regression-related Analyses

2.3.1 The 2 continuous variables that are not useful for the analysis are num_reports and downs because all rows have the same values.
2.3.2 The binary variables which are not useful for the analysis are is_crosspostable and media_only because all rows have the same values.
2.3.3 It is not useful to use subreddit_id and subreddit_name_prefixed because they both will act as a unique key to point to a particular subreddit. Every subreddit name will have the same subreddit id, thus any one of these is sufficient to get the result.
2.3.4 The permalink is basically a URL of that post with the title of the post. There is a separate column for 'Title' which could be used directly when needed.
2.3.5 The following plot is as follows:

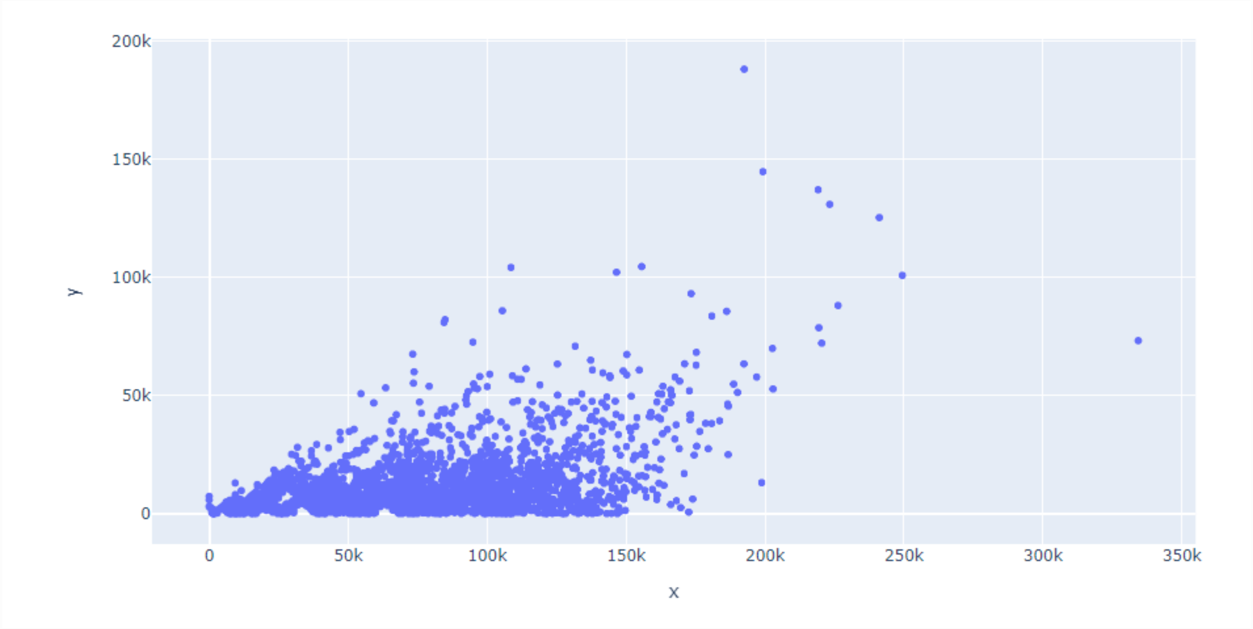


2.3.6 The num_comments and upvotes are directly proportional to each other.
2.3.7 The strongest positive correlation with ups is with num_crossposts.
2.3.8 The weakest positive correlation with ups is with created_utc.

PART 3.1 Regression Basics

3.1.1 RMSE = 0.16434081889892896
RMSE is the average error we should expect in our prediction. For the number of upvotes, 10^{RMSE} = 1.46 is the expected error.

3.1.2 In one-hot encoding, binary style of categorizing follows. In this, every categorical value is converted into a new column which is then assigned either 0 or 1. Applying one-hot encoding on ‘subreddit_name_prefixed’ will create different columns for every value with binary vectors.
3.1.3 The drop = “first” is used on ‘subreddit_name_prefixed’ to make sure that there are no reference columns and the remaining new columns can become linearly independent.
3.1.3 1 is added to the outcome variable before using log is to avoid log x approaching negative infinity as x reaches 0
3.1.4 StandardScaler is used for scaling purposes. After scaling the data, we could say that the data is all fair now and can be used on a single scaling platform. For example - we have 2 columns: age and salary. Age is a double-digit number, but the salary could be anything. So to make sure that every feature is scaled on the same page, StandardScaler is used.
3.1.5 The scatterplot is as follows:



3.1.6 The above scatterplot suggests that the average difference on the y-axis certainly grows with an increase in upvotes on the x-axis
3.1.7 The new RMSE with the logged independent variables is 0.1573237419417155.
3.1.8 The old value of RMSE is greater than the new value of RMSE. This is because taking a log transform reduced the skewness in our features.

PART 3.2 Interpreting Regression Coefficients

3.2.1 The strongest positive predictor of upvotes is ‘num_comments’. The standard deviation corresponds to 0.225871.
3.2.2 The strongest negative predictor of upvotes is the subreddit(r/learnprogramming). The standard deviation corresponds to 333707849788.58252.

PART 3.3 Attempting to improve our predictions

3.3.1 To improve our predictions we did 2 things - we changed our model to Random forest and we dropped the feature with the most trivial regression coefficient (‘over_18’). Since ‘over_18’ had the smallest regression coeff a change in its value had the least impact on our prediction. We noticed that removing this feature reduced the prediction accuracy on the training set but slightly improved the test set accuracy which hints towards an overfit.
3.3.2 Our RMSE improved from 0.1573237419417155
to 0.15654547170665237 which is a reduction/improvement of ~0.5%