Exploring User Retention on Community Visits

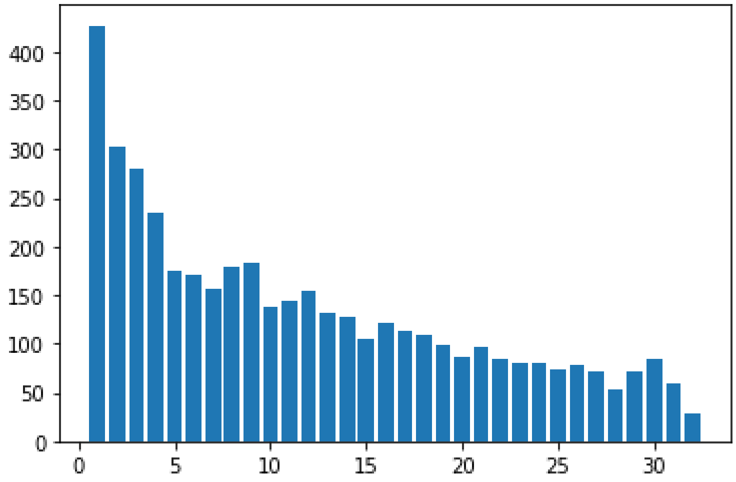
Alac Wong Just now · 4 min read

By Alac Wong

User retention is an important statistic used in many businesses to improve and drive their product to the next level such as Spotify wrapped. By understanding their users, companies like Amazon, Spotify, Meta rise to the top. Today we will look at some Reddit data and see if we can find some behavioural patterns in the data and if we can find some correlation between the actions a user takes and how long they stay on the platform. Some other questions we want to explore are, can we create a model that calculates how long a user will stay on our platform, and how homogenous are users of the same subreddit.

Processing the Data

The dataset I used was a metadata dataset which about 1 GB in size. This dataset contained information about the user, when a comment was posted and which subreddit was posted. The first question I wondered, how long do users actually stay on the platform? We define this as we define as t_n minus t_0 where t_n is the time of their last active comment t_0 is the time of the first active comment. To do this, I aggregated the user’s activity together on a downsample of the dataset and took that difference. Here is the distribution of how long users stay on the platform.

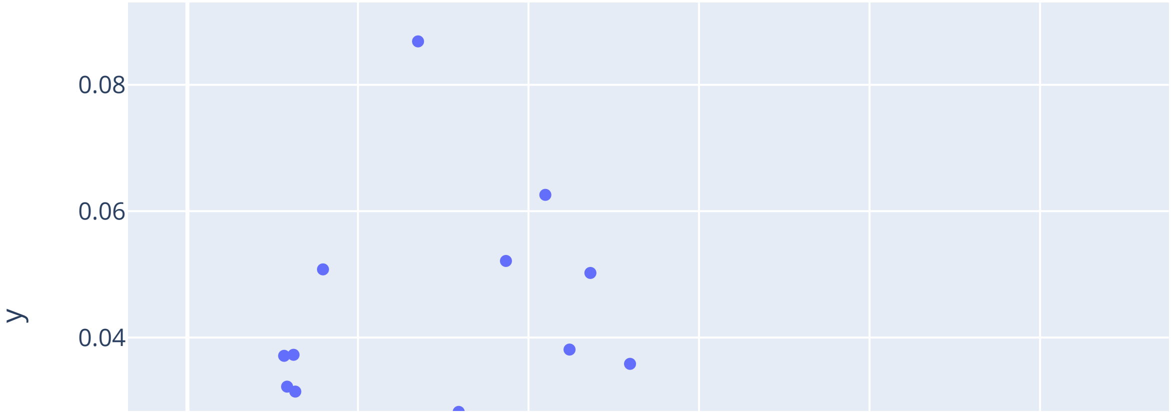


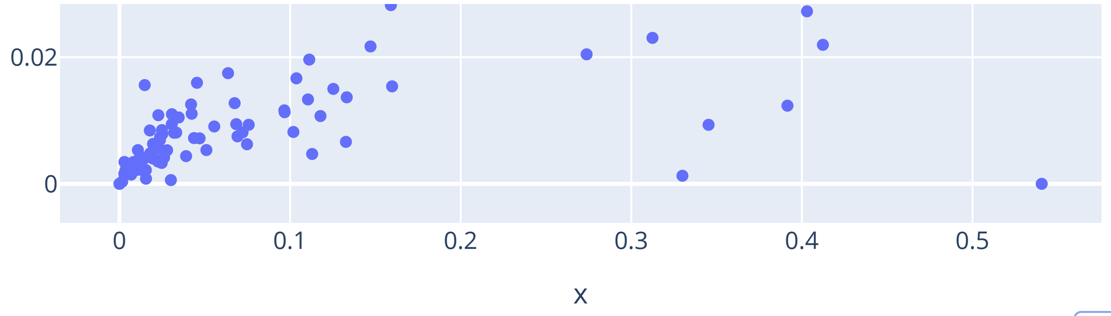
Retention (months) vs Users

As we can see exponential decrease in user activity in the earlier months, however the user loss becomes very steady towards the later months, suggesting that the longer the user stays, the less likely they are to leave.

Community Embeddings

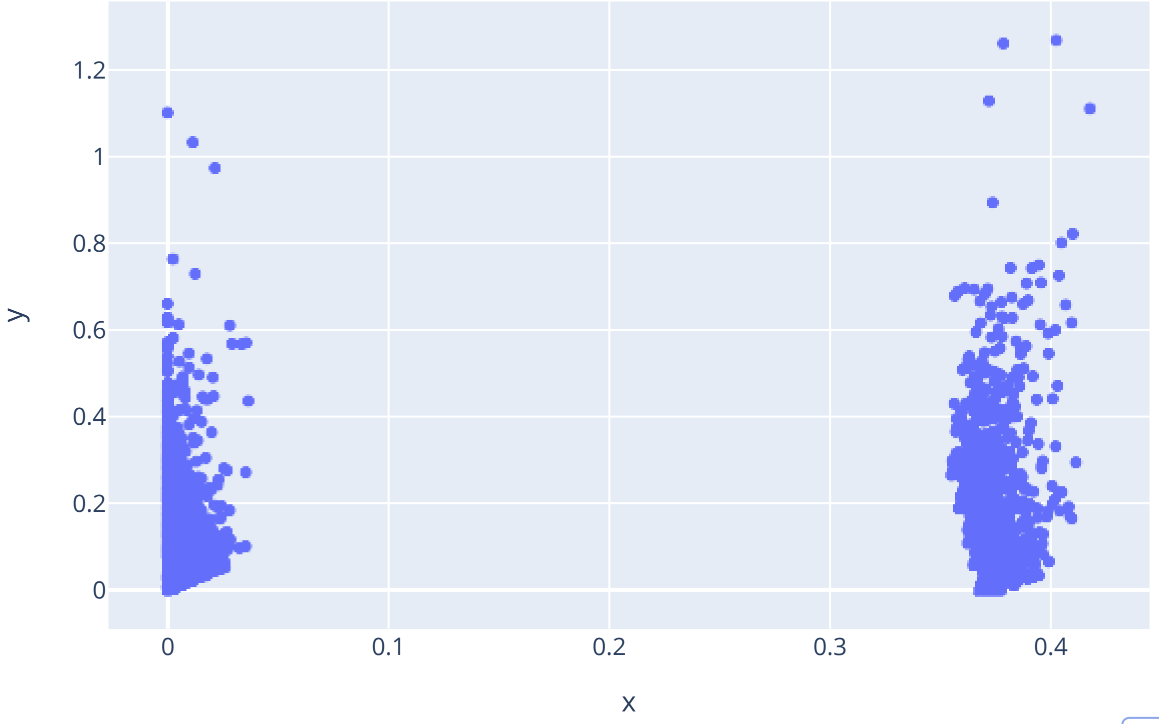
One of the things I was particularly interested in was which subreddits a user visits, and how they impacts their user retention. Here we want to visualize similar subreddits/users. To model this I constructed an n x d matrix where each row represents a user, and each column represents a subreddit. If a user i commented in subreddit j, then matrix[i][j] = 1, otherwise it is 0.





EDIT CHART

Subreddits embedded onto User Space



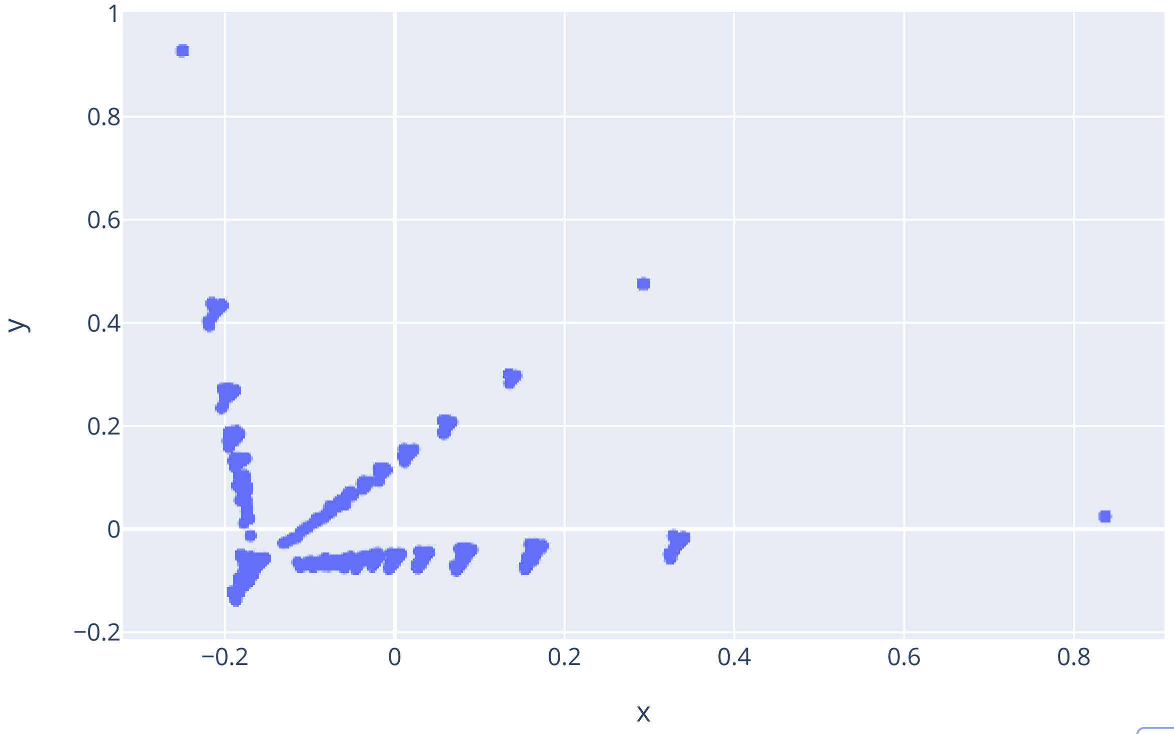
EDIT CHART

Users embedded onto Subreddit space

We used matrix factorization (a technique used in collaborative filtering algorithms) to embed our high dimensional matrix into a 2 dimensional space. The first plot shows subreddits embedded on to a user space. Although there is some pattern in the plot it looks somewhat random. It turns out that are far away are in fact subreddits with tons of users, while the points that are clustered together on the left are subreddits that have fewer users. This plot did not seem to encode subreddit similarity, as it seemed to be dictated by user cardinality. However the second plot seems to be split into 2 clusters.

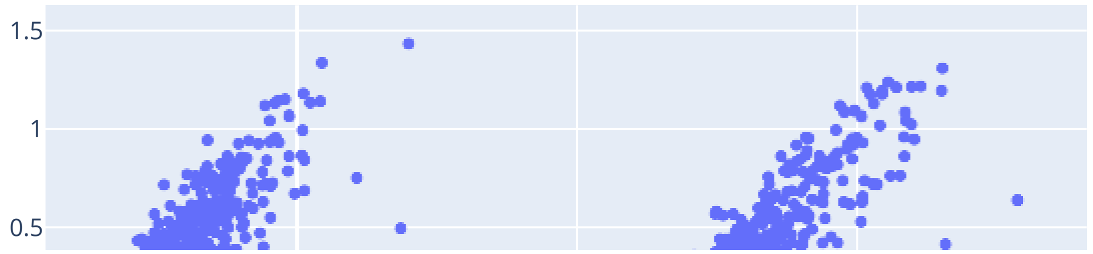
This suggests subreddits are not homogenous as user activity seems to converge together which I assume is regression to the mean. However in the second chart, we can see that similar user activity does cluster together.

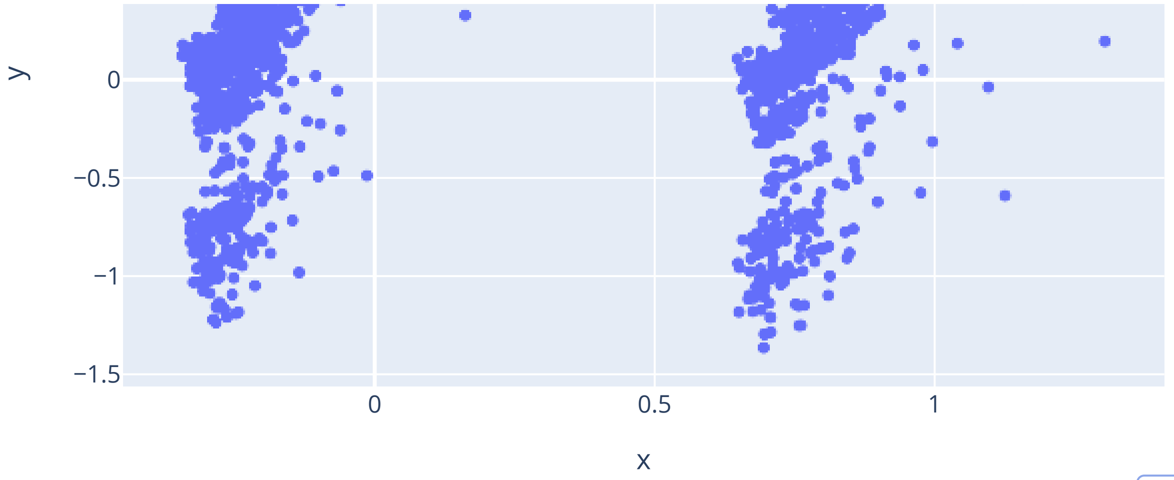
The next thing I did was use a weighted matrix (iditf). If a user visits multiple subreddits, we give less weight to those subreddits. Therefore if user i visited subreddit j, matrix[i][j] = 1/ number of subreddits he has commented. When visualizing using the matrix factorization algorithm, we do not see much of a difference, however using PCA, the results are surprising.



EDIT CHART

Weighted Matrix PCA





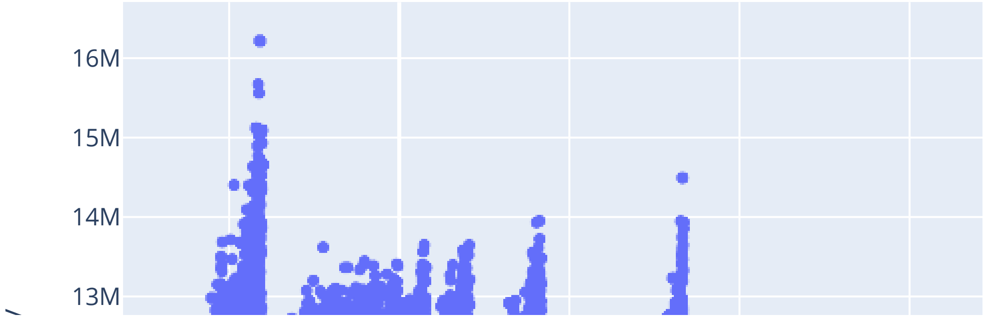
EDIT CHART

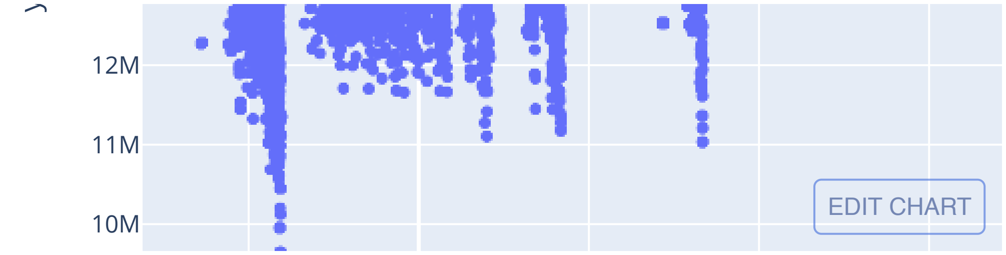
Normal Matrix PCA

With the normal matrix, the same 2 cluster pattern is observed again, however with the weighted matrix, we see a recursive fractal pattern in the second graph. You may be curious what those 2 clusters actually represent in the Normal Matrix. It turns out right cluster is users that use askreddit, and the left cluster is those who do not. Since the matrix is not weighted, it seems askreddit skews the data. However with the non weighted matrix, this skew is alleviated and we can observe the true relationships between users. For example user's that uses memes, cluster together with users that use dankmemes.

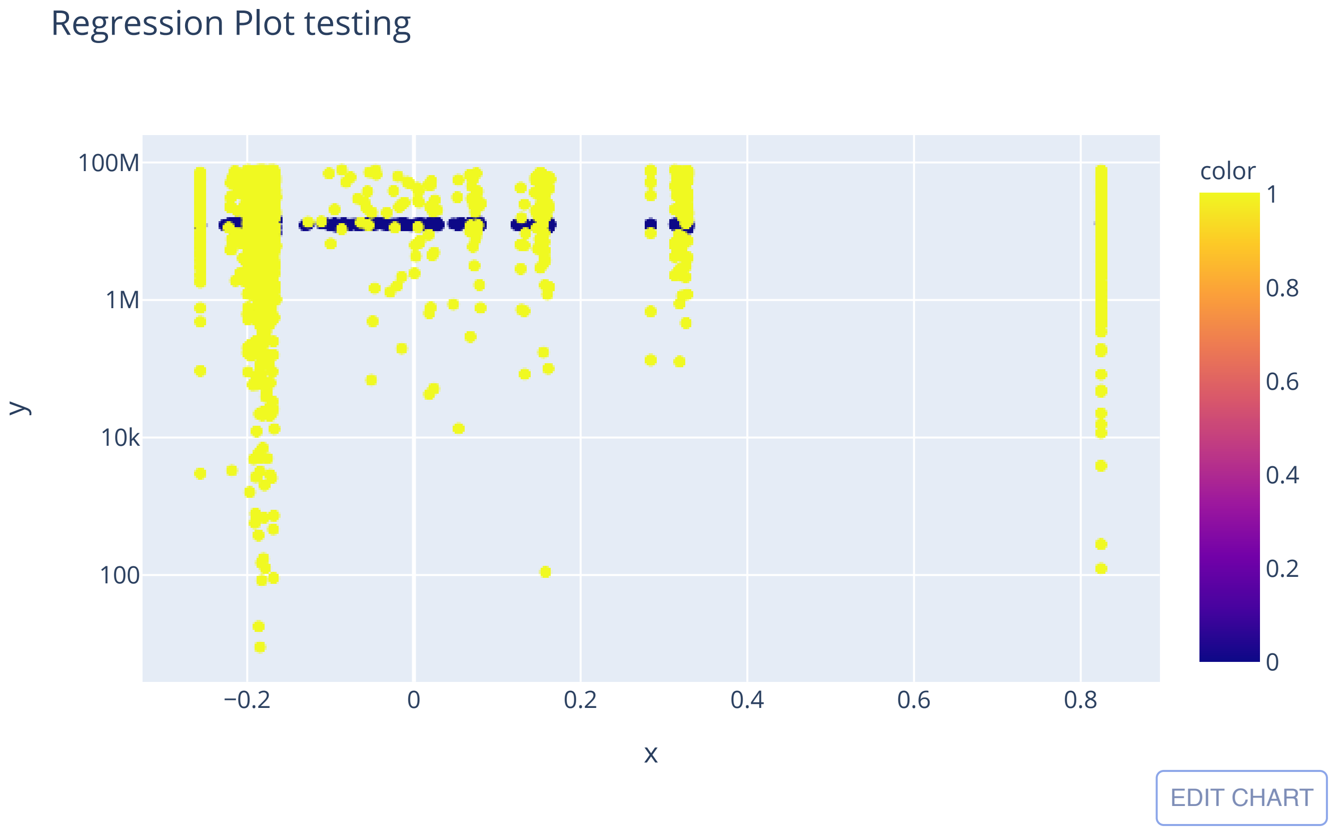
Regression

The next thing I tried to do was build a model to represent user retention. Using the matrices as training data, and user retention values as labels, I used linear regression as our model. To visualize this, I used PCA to push down the weighted matrix into 1 dimension (x axis) and then used user retention as the y axis.



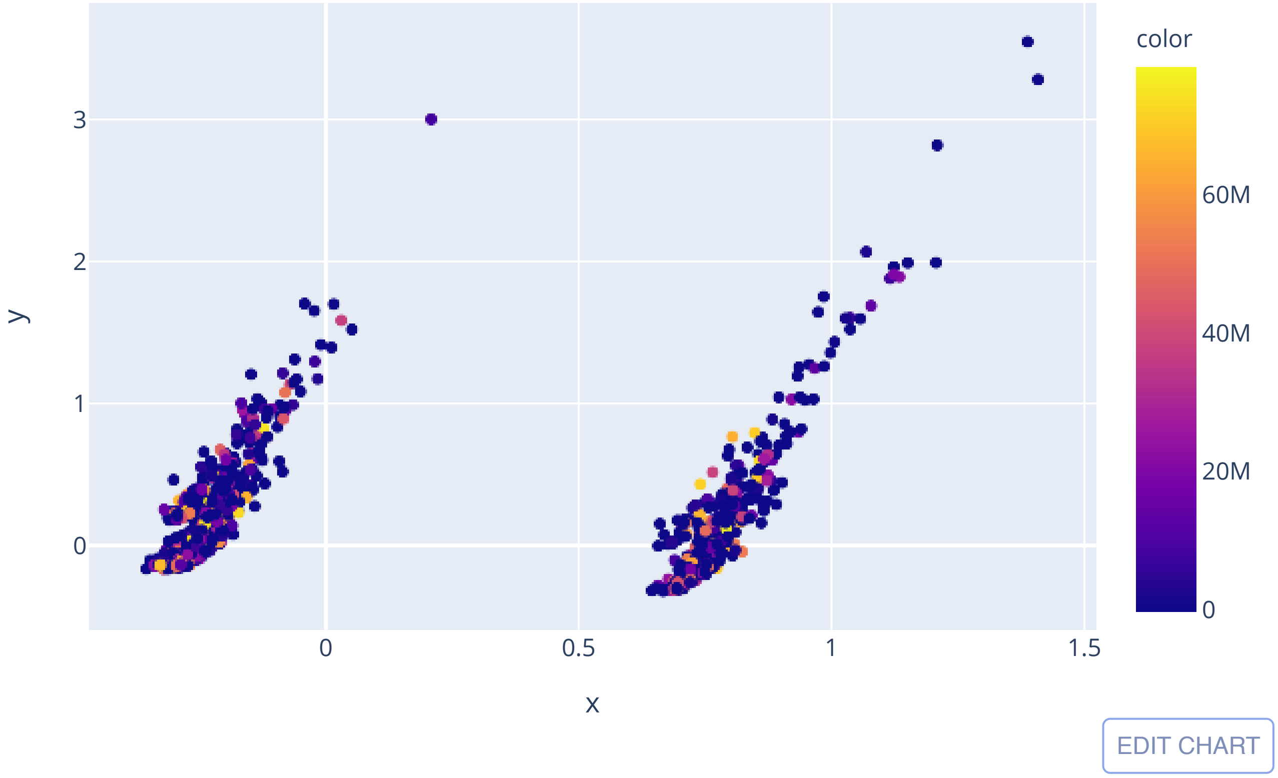


Linear Regression Plot

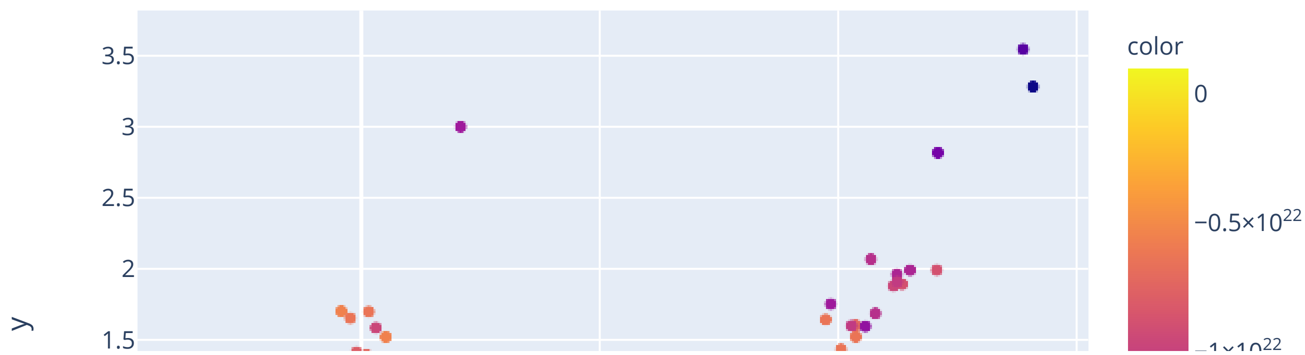


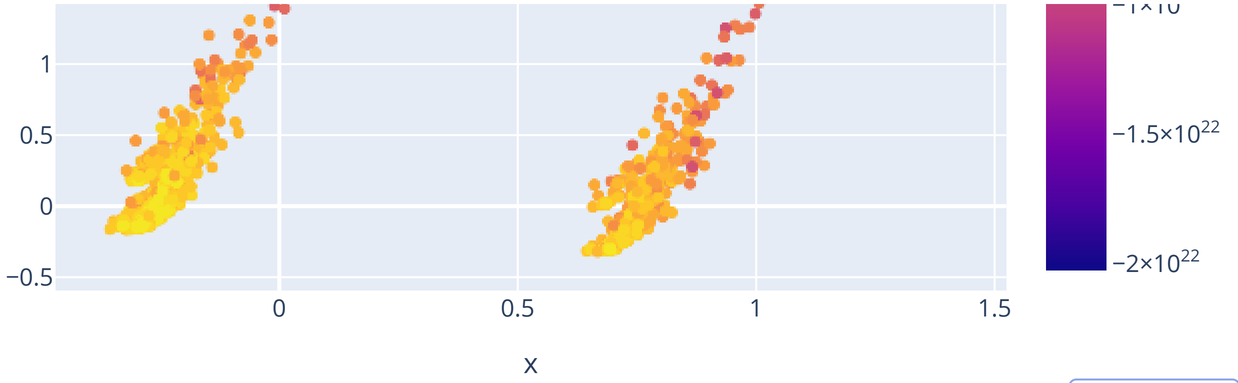
Linear Regression Plot compared to actual

The first plot is a scatterplot of our linear regression. There is definitely some variance between values as we are compressing 92 dimensions into 1, however there is even more variance between in the second plot, where the yellow points represent the actual user retention. It seems there is a lot of normal variance between user retention, the types of subreddits they visit isn't enough information to reliably compute how long they stay on the platform. Our linear regression plot suffers heavily from regression to the mean, where all our outputs seems to take the mean of user retention levels who visit the same subreddits so it can't capture the intricacies of the whole distribution. A way to alleviate this, would be to feed in more data, perhaps post frequency, current user retention, or perhaps read history(this is harder to get, only reddit has it), which will allow us to capture more intricacies, or we could try to use a more complex model to capture more variance.



User retention encoded as color





EDIT CHART

User retention predictions encoded as color

Here is our regression plot mapped onto our data visualizations. As you can see, the actual results, have a huge amount of variance ranging throughout the entire color scale. However for our prediction plot, all our points fall around the same value meaning our current model does not model user retention well.

Data Science