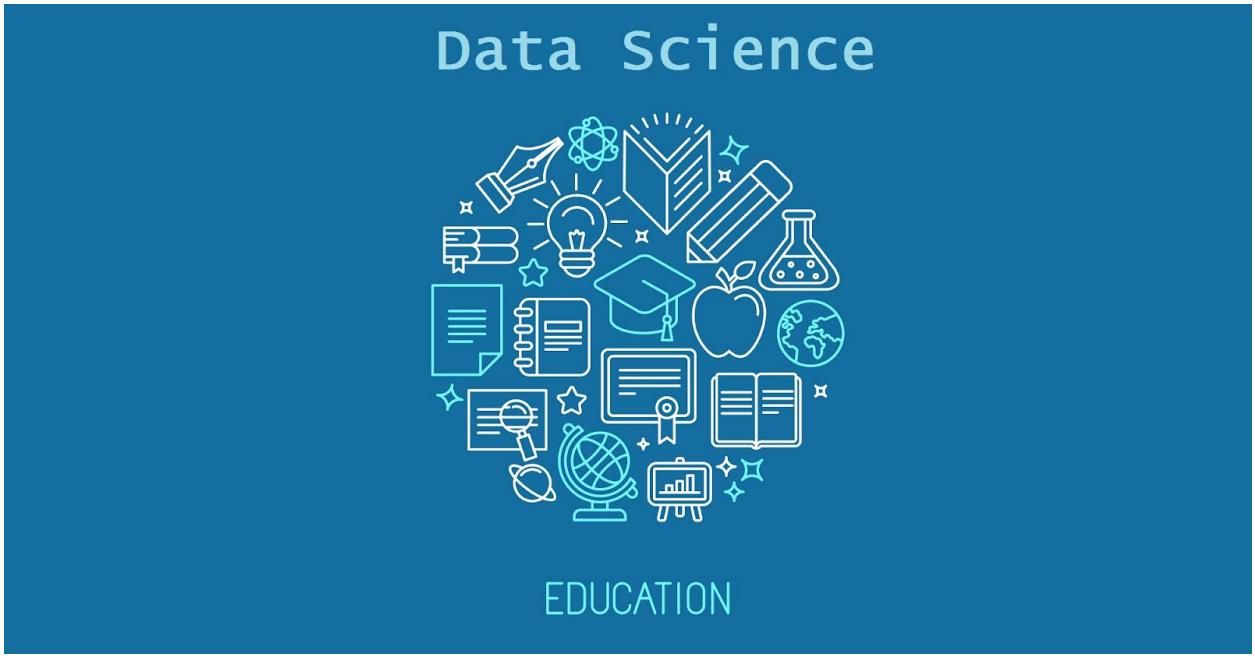
COMP 395: Geometrical Optimization

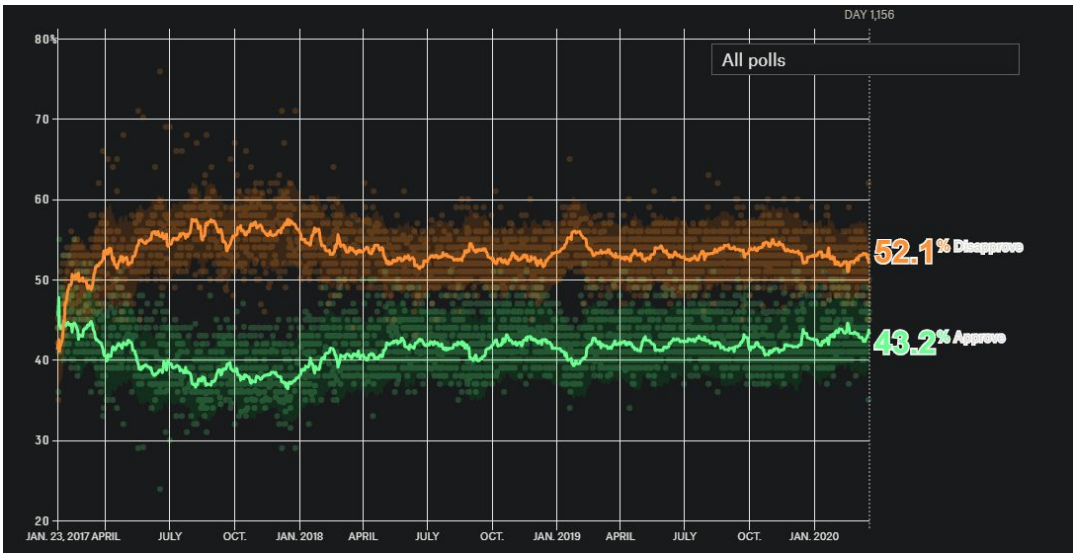
Final Project

Dylan Morison

USING DATA SCIENCE TECHNIQUES TO ANALYZE TRUMP'S APPROVAL RATINGS



It may seem trivial to deduce Trump's disapproval rating over his time in office by simply subtracting his approval rating from 1 and treating the resulting number as his disapproval rating at any given time. However by simply examining the intuitive graph supplied by: https://projects.fivethirtyeight.com/trump-approval-ratings/



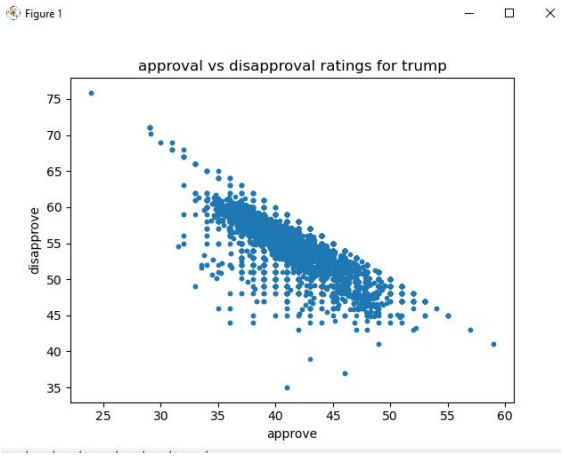
one can observe Trump's approval rating and disapproval rating never actually add up to 1. Say Trump's approval rating is X and his disapproval rating is y, one can observe X + y is very close to 1, but *never* actually equals 1. This leads to the assertion that approval and disapproval are essentially a basic percentage. If half of the country approves of him one would expect the other half to not approve of him. As such, it seems to me it would be useful to use various data science techniques to analyze the relationship between X and y. Doing so will (hopefully) allow us to determine if the relationship between approval and disapproval is a simple percentage, or if there is something more complex and subtle going on. (All data used in my models is supplied by https://projects.fivethirtyeight.com). projects.fivethirtyeight.com's data scientists use local polynomial regression to create the smooth curve shown above. The data scientists in charge of projects.fivethirtyeight.com use local polynomial regression to create the elegant graph displayed above. As they describe in the following link:

https://fivethirtyeight.com/features/how-were-tracking-donald-trumps-approval-ratings/

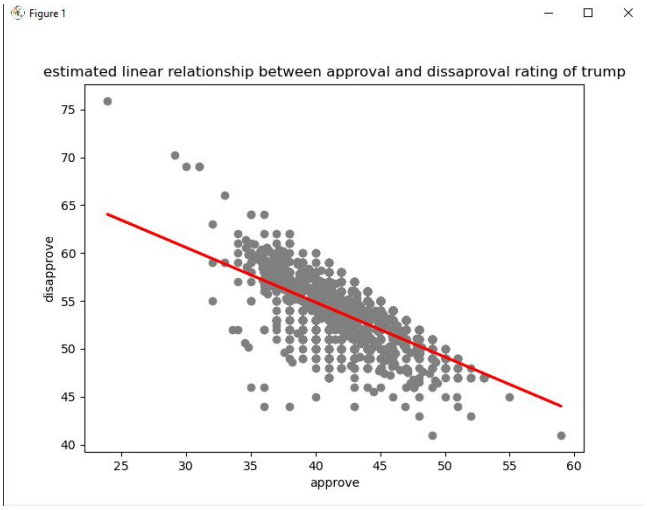
They try to fit the data as well as they can by smoothing just the right amount as to describe the data as well as they possibly can. This actually reminded me of our first homework assignment where we were tasked with a very similar problem. Anyway, the first technique we will analyze is linear regression.

Linear Regression

First we will apply linear regression to get some insight into the relationship between X and y. Here is the preliminary data:

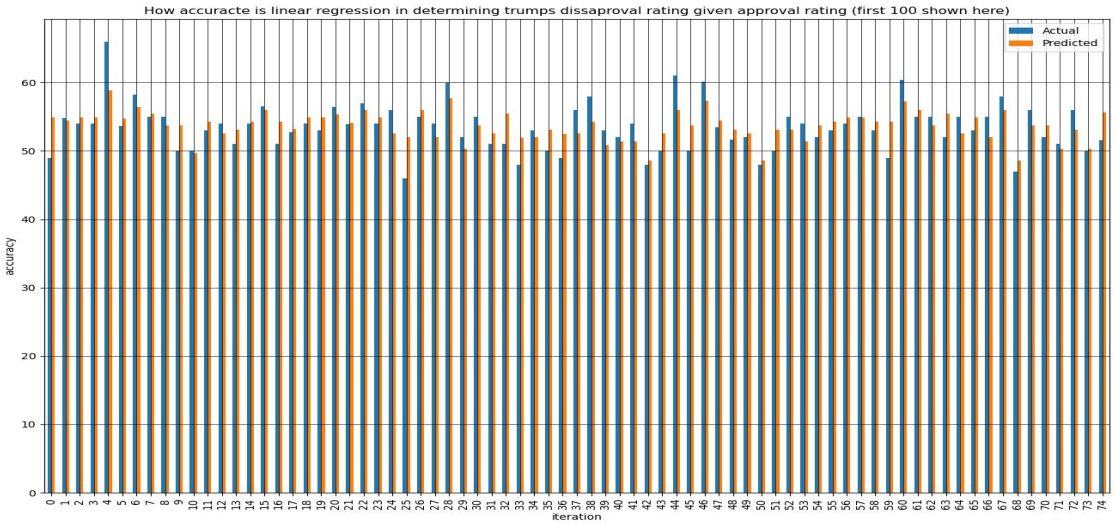


After applying linear regression we obtain the following estimated relationship between X and y:



Now what we really want to know is if we can train a linear regression model to determine how well **y** can be predicted given **X**. In other words, given Trump's **approval** rating at any given time, what is his **disapproval** rating? Is y = 1 - X, or something else? What we find is y is never precisely equal to 1 - X. Sometimes y gets pretty close to being equal to (1 - x), but

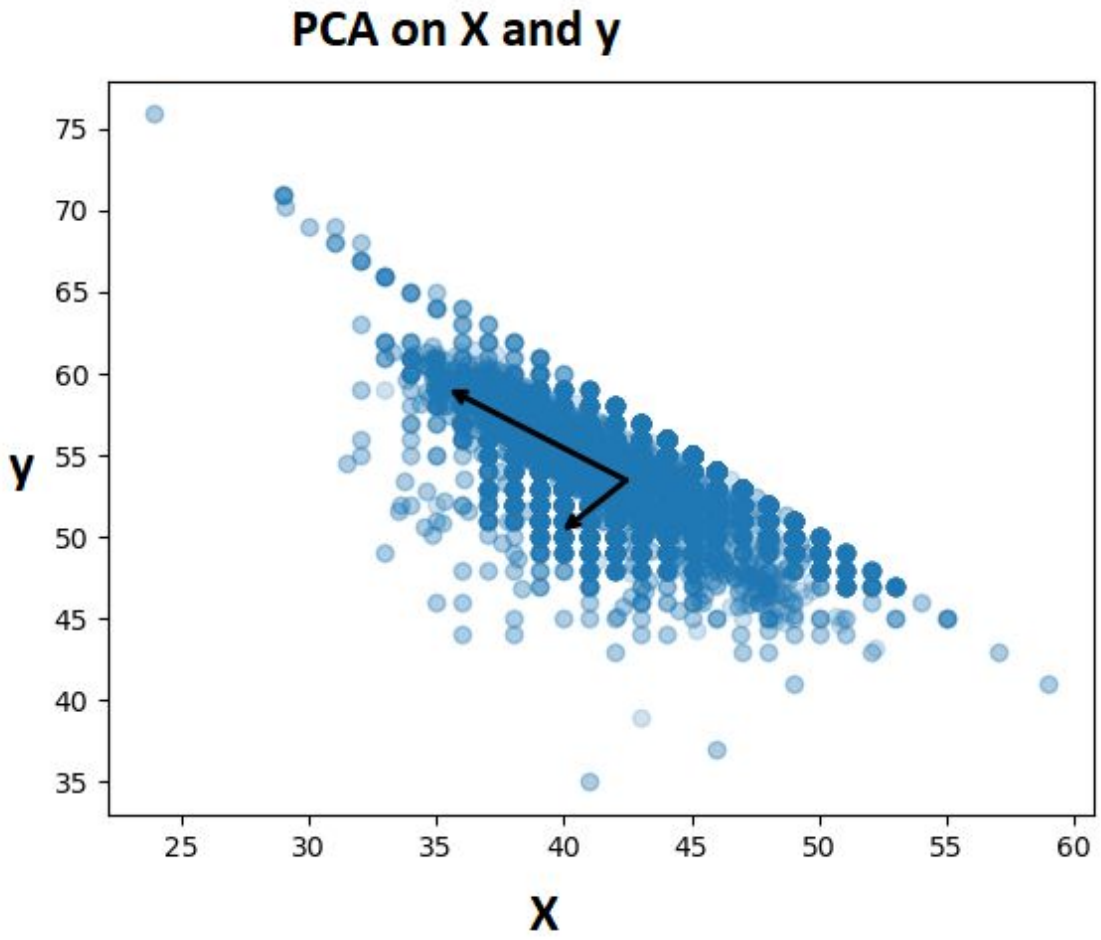
often y is farther away. The key observation in this case is the fact that they are *never* equal.



I also have attached actual vs predicted values of y here:
https://drive.google.com/file/d/1Fa-zbucFWHKaWfX-rjfyzfwkN4mgRjQf/view

It can be observed from the bar graph and from simply comparing all the values in the link above that linear regression performs similar to how we would expect, as some of the predicted values are exceptionally close to the actual values, while others are not so accurately predicted. Ultimately, it seems predicted is pretty accurate when compared to actual values. Moreso, the accuracy of predicted versus actual values is a good initial indication that given the approval rating at a given time, one could make a very accurate estimation of disapproval rating.

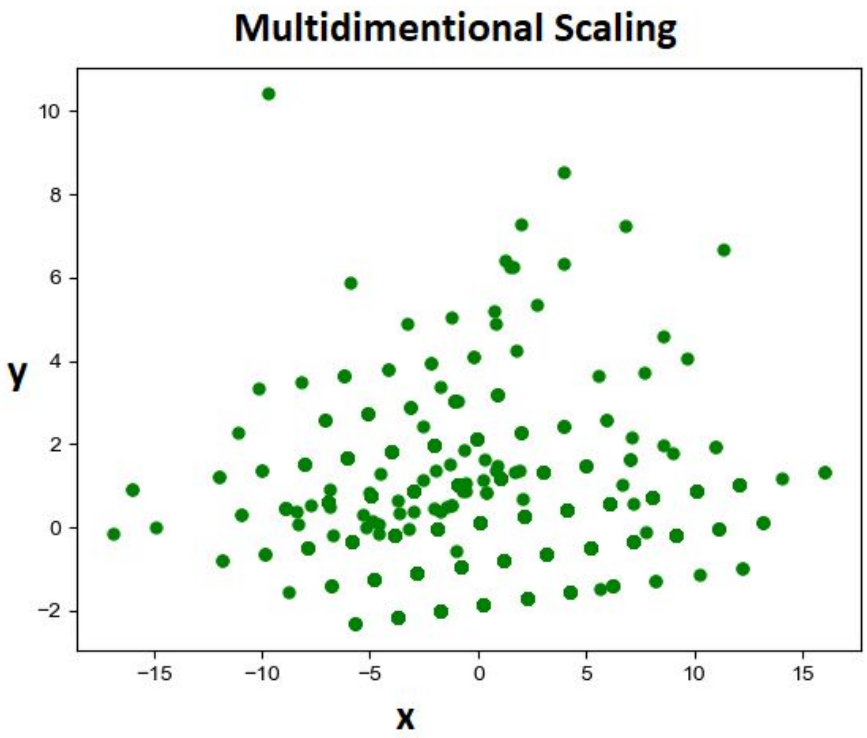
Principal Component Analysis (PCA)

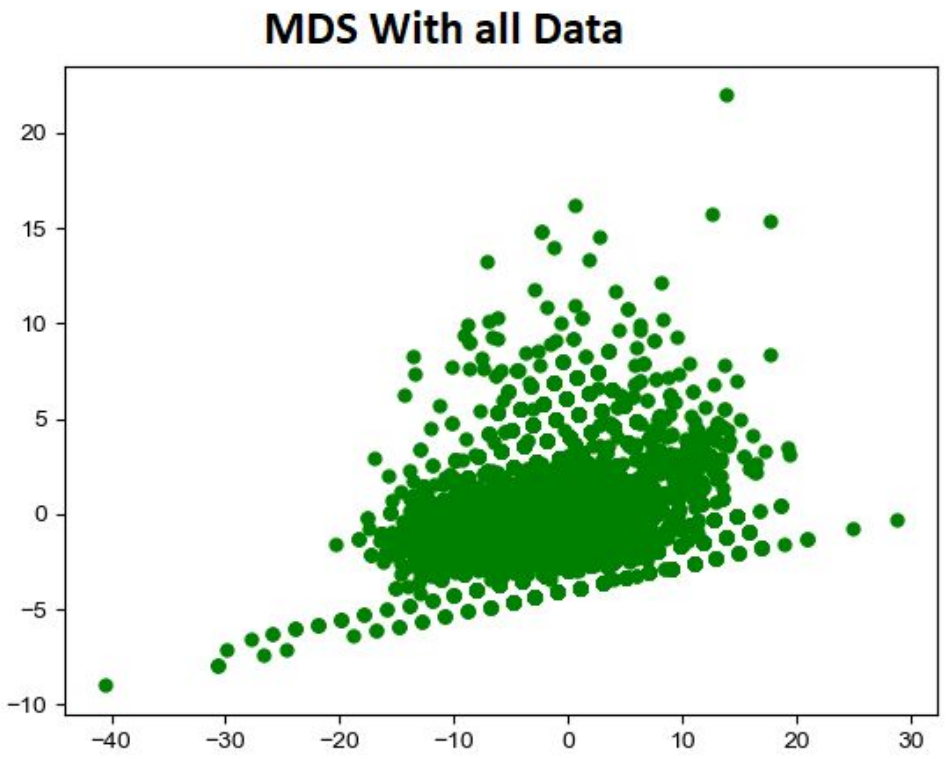


When PCA is implemented on our data set we see underlying intrinsic directions in which Trump's approval rating and disapproval rating seems to gravitate towards. First and foremost, PCA uncovers that there is more variance in the data set as Trump's disapproval rating, or y, increases. In other words a larger y is correlated with greater inconsistency in the relationship between X and y. I hypothesize this is the case because Trump has done many different things as president that have angered people, however the people who

support him from the beginning, on average, tend to continue supporting him. Additionally, his disapproval rating is always larger than his approval rating.

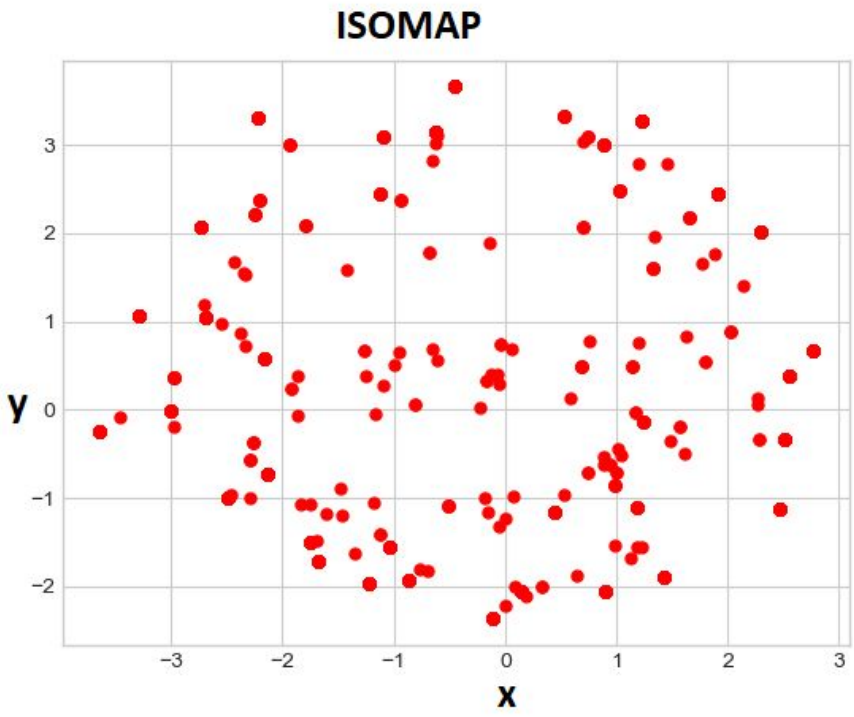
Multidimensional Scaling (MDS)



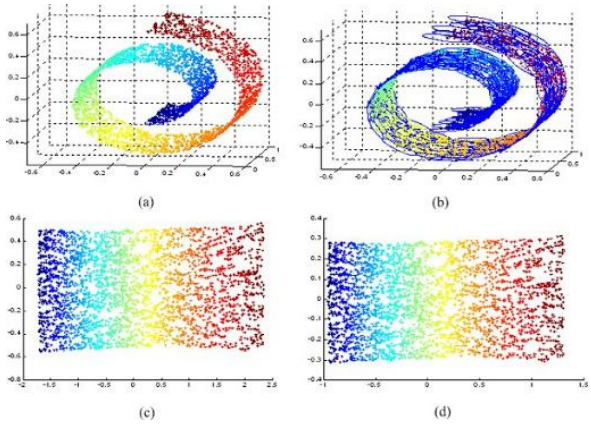


Above are two different Multidimensional Scaling graphs, the first with about 10% of the data, the second with all the data. What becomes immediately apparent is that there is little variation in the data set. Of the 11,000 rows of data in the data set there exists around a couple dozen outliers, which makes logical sense given that, again, we are essentially working with pseudo-percentages as our data. I would be worried if the full MDS graph were not as orderly and symmetric as it appears due to the inherent symmetry of the original data. Interestingly, I am not able to detect quite as much symmetrical appearance in the following ISOMAP graph as I was in the preceding MDS graph. This is both expected and unexpected as the Isomap graphs I have examined in research articles have also been harder to interpret then MDS graphs. Most often there are additional visual images added to the graph to give the reader a quick understanding of what different portions of output represent. In my case it does appear there is some similarity to how far apart my

isomap results are spread out.

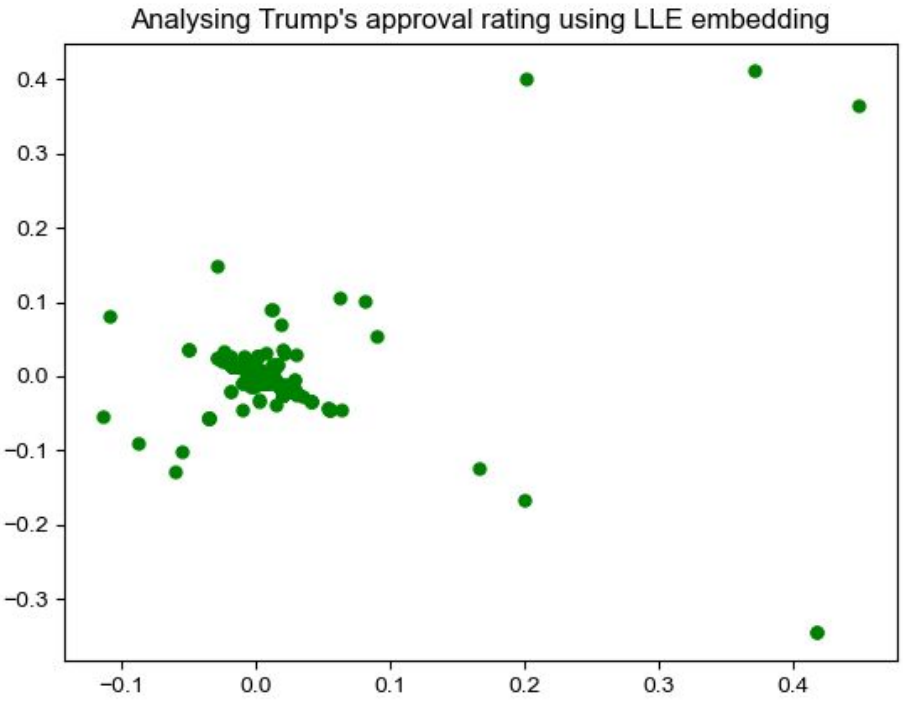


Here an underlying symmetry is still apparent, but compared to MDS all my data points are more equally spaced apart rather than inhabiting a shared closeness. Another factor to consider here is again the fact that my original data set is two dimensional and somewhat linear. It seems to me the usage of a 3-dimensional dataset would have given some more interesting graphical results, as some of the most interesting applications of Isomap I've read about are projecting a 3-dimensional object into 2 dimensional space. The most prominent example of this is the swiss roll we used in class:



Combine the fact that my data set is of size 11,000 and vast quantities of my data exhibits inherit "closeness", I observe what seems to be a reassuring ISOMAP output. This is exhibited by the scale of our x-axis and y-axis; the x-axis ranges from -3 to 3, while the y-axis only ranges from -2, 2. The fact that multiple points are overlapping in the center of the graph combined with how many points are evenly and closely spaced between each other is in my mind an indication that the euclidean distances of the edge matrix are generally small.

Locally Linear Embedding

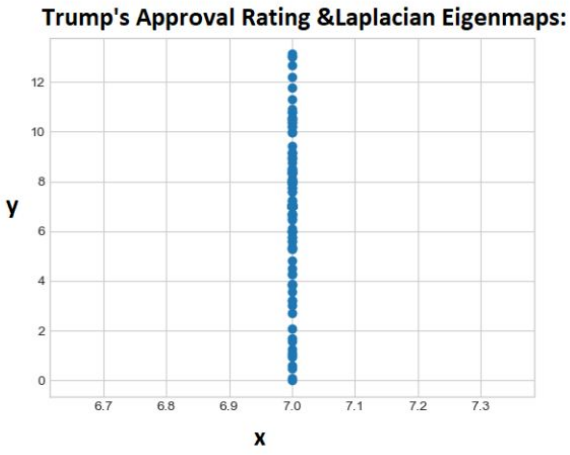


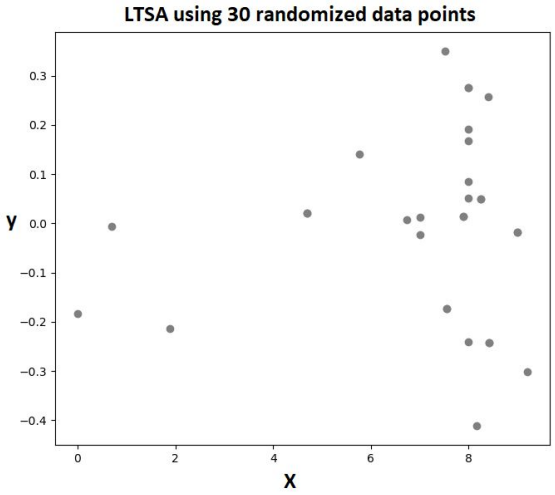
My LLE results clearly show the underlying structure of my data by preserving the closeness of a vast majority of rows in my data matrix. We know Trump's approval and disapproval rating are inherently related in terms of percentages (after all, it wouldn't make sense for approval to be at 90% and disapproval to also be at 90%). We can also observe from the very first figure I introduced that Trump's approval rating usually lays within a range of 35 and 45, while his disapproval rating lays within a range of 50 to 60. Given the data set consists of more than eleven thousand rows, I would be worried if LLE outputted a

graph which did not preserve the similarity between such a large dataset. We can observe there are only about 15 small outliers and about 6 very large large outliers.

Local Tangent Space Alignment and Laplacian Eigenmaps

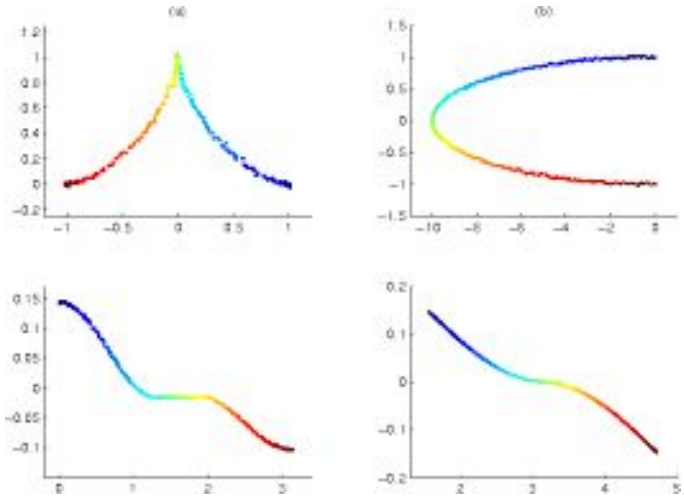
The two methods I seem to struggle with the most are Local Tangent Space Alignment and Laplacian Eigenmaps. Up to this point my implementations have outputted results that either make sense given my mathematical background of these algorithms, or preserve an underlying relationship my data set possesses. However for a variety of reasons I am less confident with my mathematical understanding of Local Tangent Space Alignment and Laplacian Eigenmaps, and as a result I feel my results do not reflect my data set as well as I was hoping they would. That being said, below are my LTSA and LE outputs.



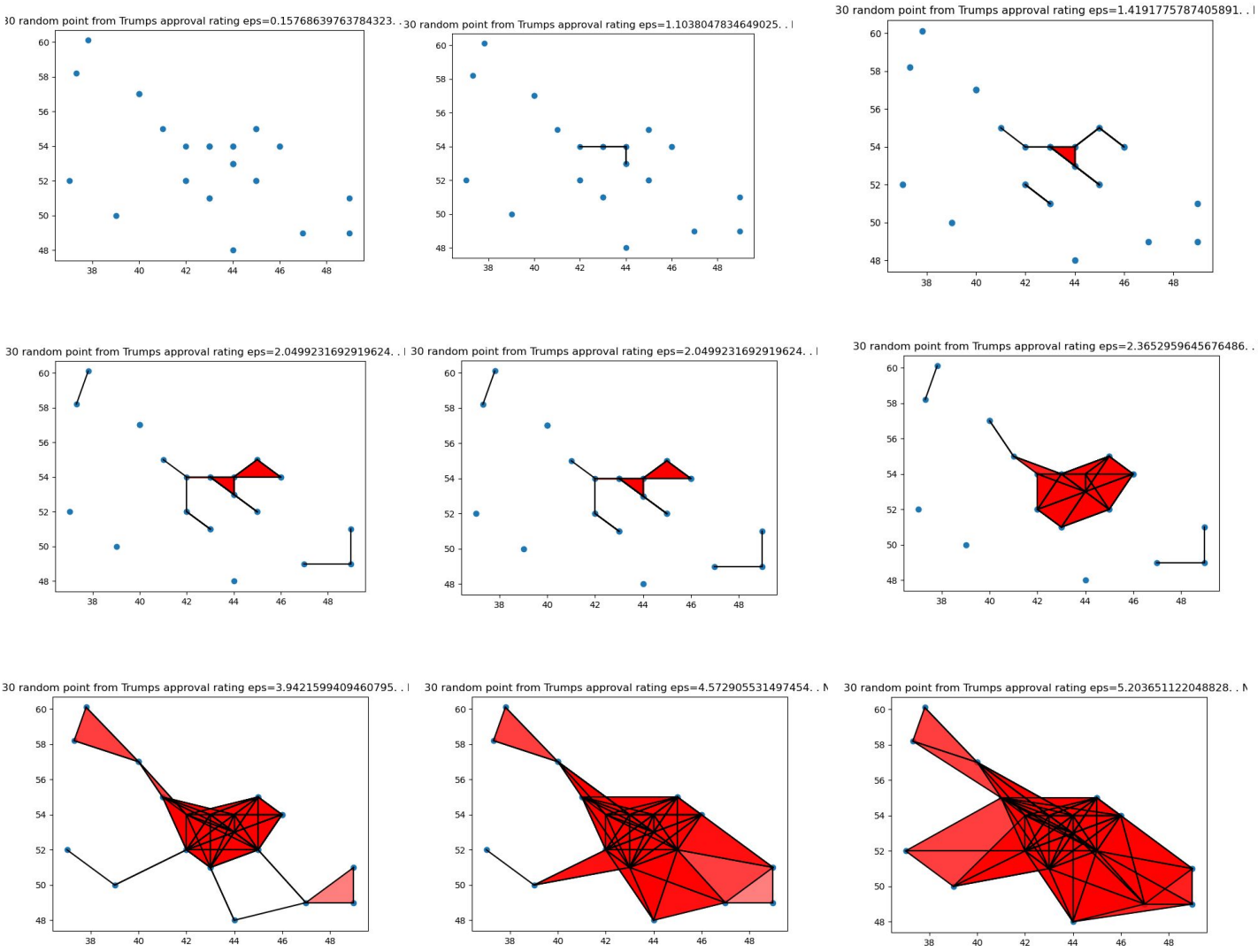


Laplacian Eigenmaps preserves a linear relationship between approval and disapproval ratings while Local Tangent Space Alignment preserves what seems to be a somewhat random, equally spaced relationship between approval and disapproval ratings. The two results could not be more different. My hypothesis is that the randomness that is occurring

is a result of the fact that my original dataset only has two dimensions. From in class lectures, scholarly reading, and spending quite a bit of time examining figures of LTSA implementations (as shown below), I would have expected to obtain output displaying the inherent linearity that exists by nature of the dataset I chose, similar to the figure below:



Persistent Homology



Above are 9 out of 54 figures I hand picked to represent my results of Persistent Homology. My Persistent Homology results exemplify a similar relationship between data points as my LLE results. After running the algorithm on many different randomized portions of my data I found I was always left with similar graphs as shown above. In the middle we see a clumped group of points that connect first, followed by 6 or 7 outliers more spread out. I would have predicted for these outliers to be much more rare than they are given the fact the data set is of size *11,000.* However each iteration with 30 random points almost always

created similar outliers. Persistent Homology seems to be showing that although the data set is pretty standard and clean, consisting of natural numbers between 1 and 99, there is still noticeable variation in the topological features. I find this fascinating because it really allows us to observe features in the data no other model seemed to exemplify.