1) Consider data set 1 (ds1.csv). The data set comprises features (the Five xs) along with three sequences that may or may not be generated from the features (3 ys).

a) Describe the data set in a few sentences. E.g. What are the distributions of each feature? Summary statistics?

This dataset consists of 100,000 rows and has 8 columns. Judging by the column names, there seem to be 5 predictors and 3 target columns. Another thing that stood out to me is that ‘x4’ is missing. This might just be a naming error, or it might be the case that having x4 in the dataset would help us generate a more accurate predictive model for the 3 ‘y’ columns. Another possibility is that x4 wasn’t as good a predictor and was removed during feature selection or was combined with one of the other variables.

Here are the summary statistics for all columns –

	x1	x2	x3	x5	x6	ya	yb	yc
variance	8.257	4.003	3.073	1.004	21.752	339.979	0.592	0.001
skewness	-0.008	0.0	-0.004	2.04	0.004	0.311	-0.524	0.72
count	100000	100000	100000	100000	100000	100000	100000	100000
mean	5.011	-3.006	2.501	0.999	0.001	3.828	2.112	0.0
std	2.874	2.001	1.753	1.002	4.664	18.439	0.769	0.031
min	0.0	-12.499	-3.489	0.0	-13.885	-64.022	-0.524	-0.543
25%	2.536	-4.354	1.19	0.286	-2.612	-8.998	1.58	-0.002
50%	5.022	-3.003	2.504	0.691	-0.001	2.667	2.231	0.0
75%	7.486	-1.649	3.802	1.387	2.622	15.58	2.733	0.002
max	10.0	6.09	8.679	15.103	13.925	107.714	3.841	0.818

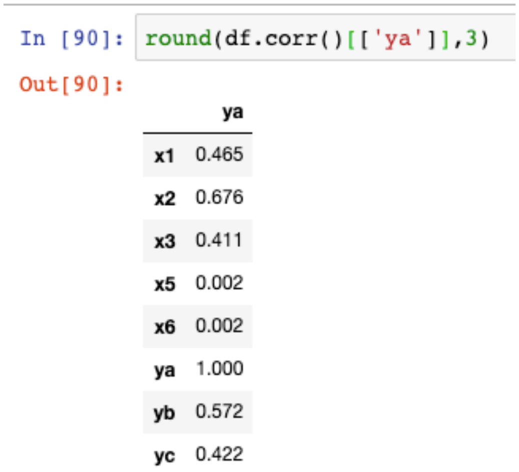
Looking at the skewness for each of the columns, columns x1, x2, x3, x6 seem like they are almost exactly normally distributed. If you look at their variances, x5, which is the only predictor column not normally distributed, has a variance of 1. This tells me that there might have been some preprocessing performed on all of the ‘x’ columns to improve model performance to better predict the ‘y’ columns.

All of the target columns (the ‘y’ columns) have non-standard distributions. ‘yc’ seems to have very small values, and ‘ya’ has the biggest values.

b) Try to come up with a predictive model, e.g. y = f(x_1 , … , x_n) for each y sequence. Describe your model and how you came up with them. What (if any) are the predictive variables? How good would you say each of your models is?

a. ya –

I first looked at what columns correlate the best with ‘ya’ -



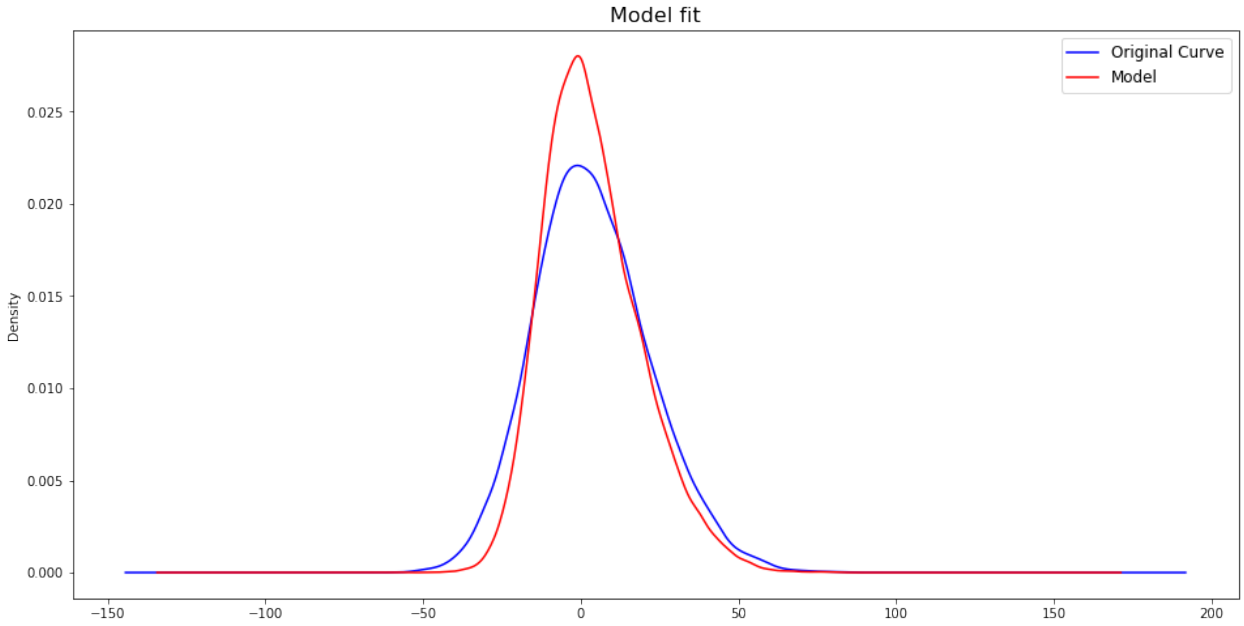
```
In [90]: round(df.corr()[['ya']],3)
Out[90]:
        ya
x1   0.465
x2   0.676
x3   0.411
x5   0.002
x6   0.002
ya   1.000
yb   0.572
yc   0.422
```

It seems clear from the above values that x1, x2, and x3 correlate the best with ‘ya’ among the predictive variables. I then used a multivariate linear model using columns ‘x1’, ‘x2’, and ‘x3’ in Python, and I got an RMSE of 10.48. As the mean is just 3.48, this model wasn’t as accurate.

To further improve this model, I used polynomial regression with x1, x2, x3. I found that degree 2 gave the most optimum RMSE value, which was 10.02. After this point, I started reaching diminishing returns. Thus, the final equation that obtained from this model is as follows –

```
ya = x1*3.0142 + x2*3.688 + x3*0.8538 + (x1^2)*0.0998 +
x1*x2*0.5066 + x1*x3*0.0172 + (x2^2)*0.0024 + x2*x3*0.007 +
(x3^2)*(-0.0003)
```

This is the curve obtained by applying this equation to the three predictors –



The model is fairly accurate, with a R squared score of 0.705.

b. yb –

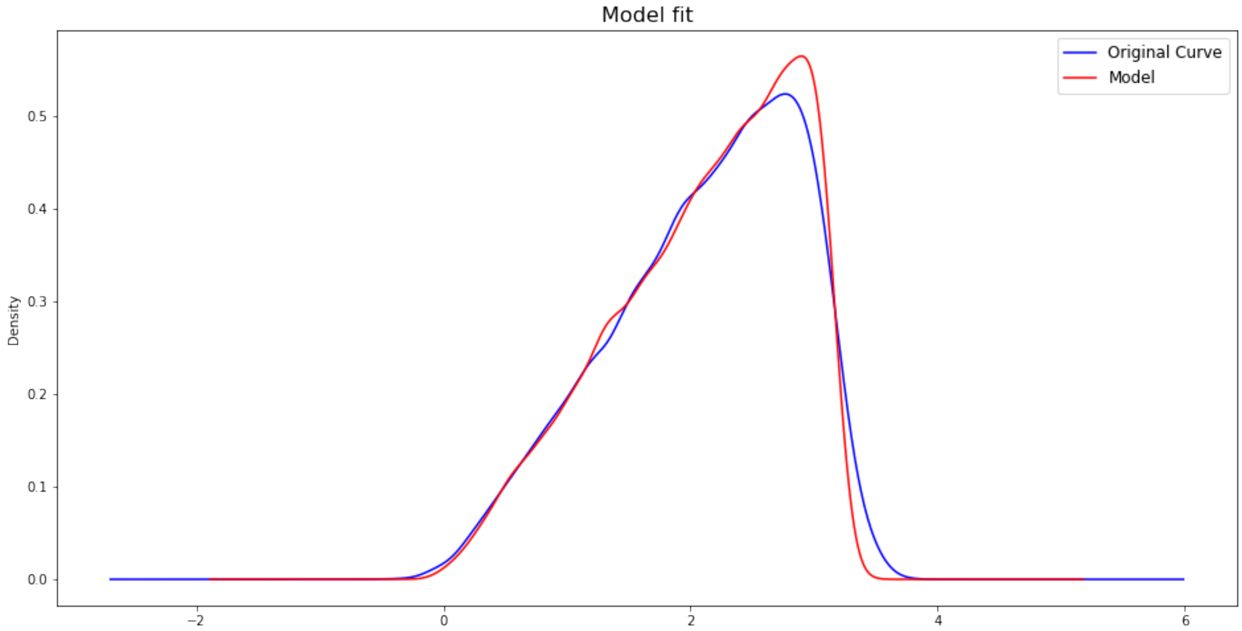
Random Forest Regression is a popular ensemble machine learning model which builds upon decision trees and combines outputs from several decision trees to come up with the best fit.

I found that the Random Forest Regressor with 10 estimators, 2 minimum sample splits, and the 'best' splitter as parameter values was the most accurate model to calculate yb, given x1, x2, x3, x5, and x6. I used the scikit-learn package in Python to come up with this model. I explored other models such as linear regression as well, from which I got an RMSE value of 0.24. Using this Random Forest Regressor, I got an RMSE value of 0.2149, which is much smaller than the mean value of yb (2.112) or the median value of yb (2.231). Thus, I feel that this is the best model for this variable. Here are the feature importances for each of the predictors –

x1	0.9463
x2	0.0135
x3	0.0134
x5	0.0133

x6	0.0135

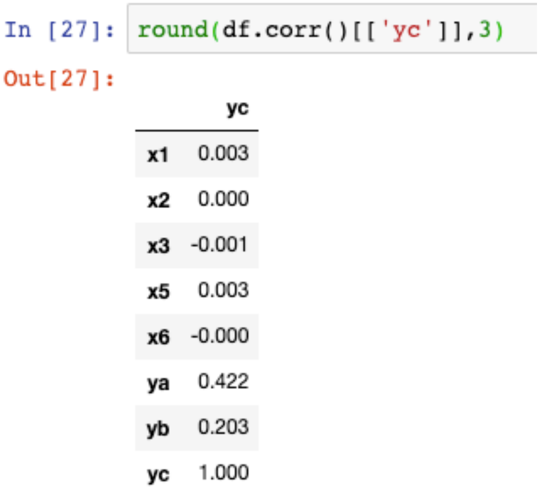
Here is the best fit curve for the same –



This model is very accurate, with an R squared value of 0.922.

c. yc –

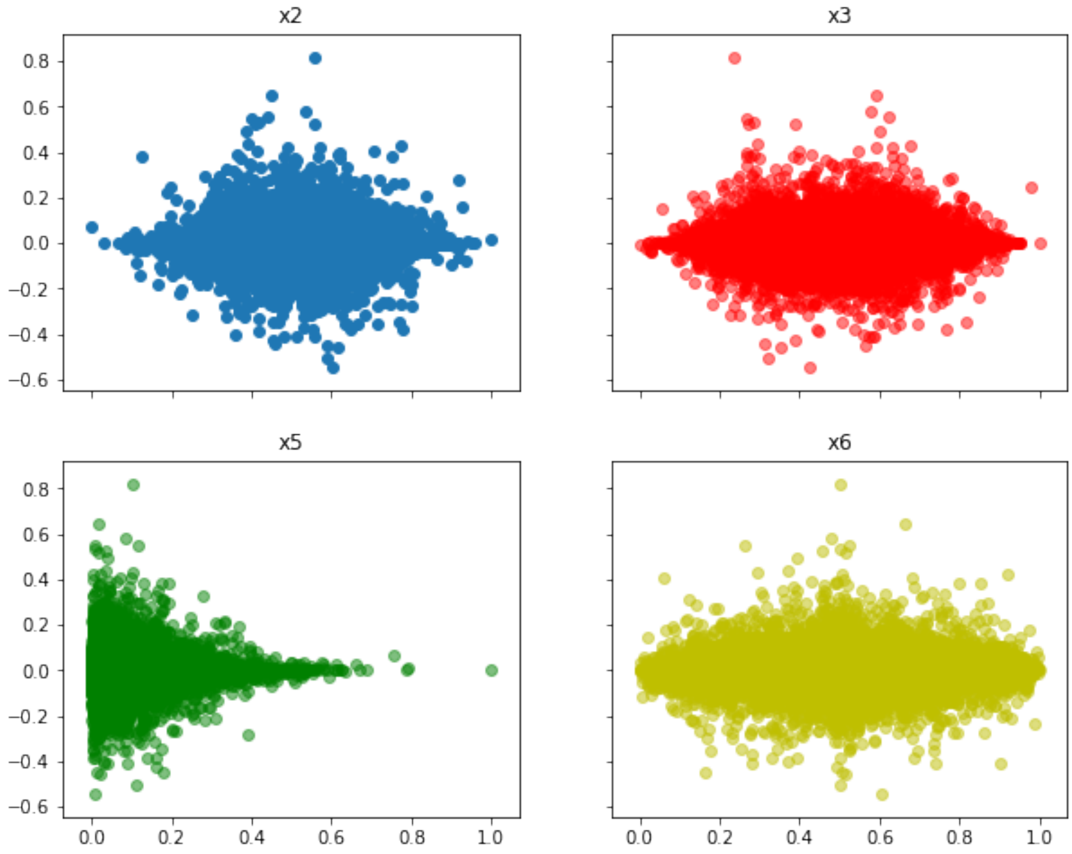
yc has the lowest correlation among the three target variables with any of the predictors.



```
In [27]: round(df.corr()[['yc']],3)
Out[27]:
```

	yc
x1	0.003
x2	0.000
x3	-0.001
x5	0.003
x6	-0.000
ya	0.422
yb	0.203
yc	1.000

Comparing ‘yc’ with the rest of the predictors resulted in some interesting visualizations –

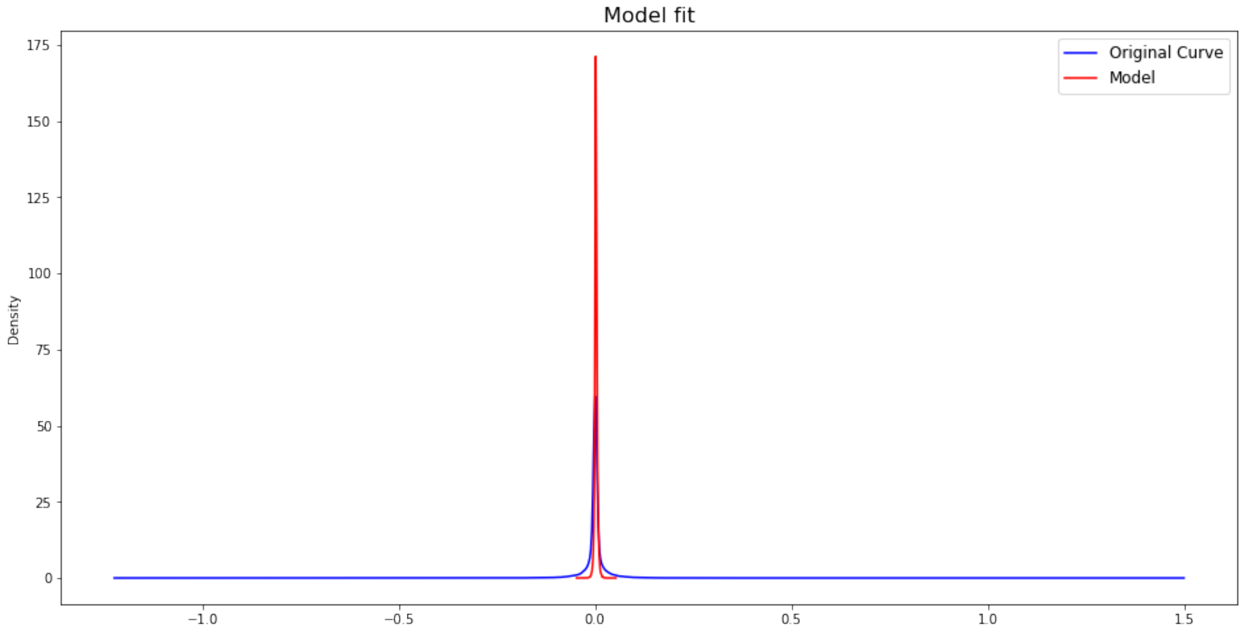


There wasn’t an easy solution to this problem. I used the support vector regressor to find the best fit for this model. I used GridSearch to find the best parameters for this model, and I came up with the following –

```
SVR(C=1.5, cache_size=200, coef0=0.0, degree=3, epsilon=0.1,

gamma=1e-07, kernel='poly', max_iter=-1, shrinking=True,
tol=0.001, verbose=False)
```

Here is the model fit for the same –



The R squared value for this model was the lowest, at -0.015. Thus, this model isn’t a good indicator for this column.

2) Consider data set 2 (ds2.csv). The data asset comprises a set of observations.

a) Describe the data set in a few sentences.

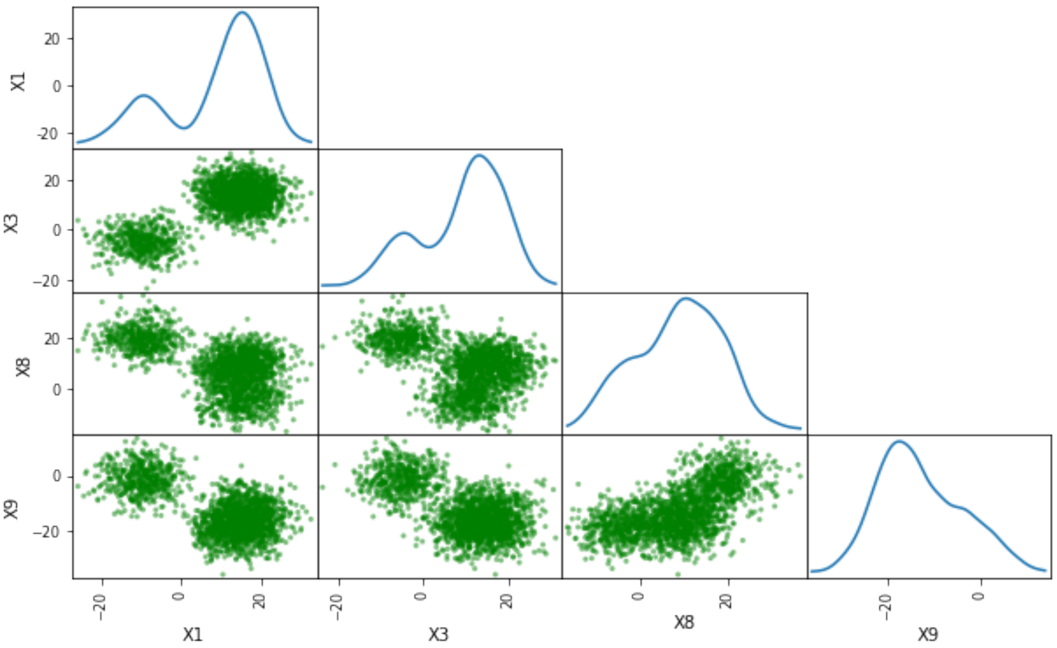
This dataset contains 10 columns and 2000 observations. Judging by their names, it seems like none of the columns are target variables, and thus this is an unlabeled dataset.

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
variance	143.32	44.29	97.25	112.89	77.45	236.41	108.7	97.64	78.87	211.24
skewness	-0.81	-0.11	-0.59	0.0	0.12	-0.1	-0.03	-0.2	0.43	0.13
count	2000.0	2000.0	2000.0	2000.0	2000.0	2000.0	2000.0	2000.0	2000.0	2000.0
mean	8.68	11.72	9.25	-2.68	2.77	0.08	8.2	8.71	-12.86	-1.34
std	11.97	6.66	9.86	10.63	8.8	15.38	10.43	9.88	8.88	14.53
min	-25.82	-8.5	-23.67	-29.43	-22.03	-35.26	-21.43	-16.81	-36.07	-36.47
25%	0.23	7.16	2.65	-10.65	-4.1	-14.0	-0.81	1.48	-19.43	-13.22
50%	12.75	11.9	11.42	-2.63	2.48	1.5	8.53	9.63	-14.42	-2.09
75%	17.36	16.28	16.5	5.34	9.66	14.05	17.14	16.08	-6.53	10.56
max	32.27	32.91	31.23	26.42	29.31	31.73	32.08	36.85	13.55	32.64

Looking at the summary statistics, it seems like some of these columns have non-standard distributions. Exceptions to this include X4 and X7 which seem almost perfectly normally distributed and X2, X5, X6, X10 have low skewness as well.

While observing the mean and average values for some of the other columns, there doesn’t seem to be a clear pattern emerging. I will try to visualize these columns to get a better understanding.

b) How would you visualize this data set? Can you make an interesting visualization?

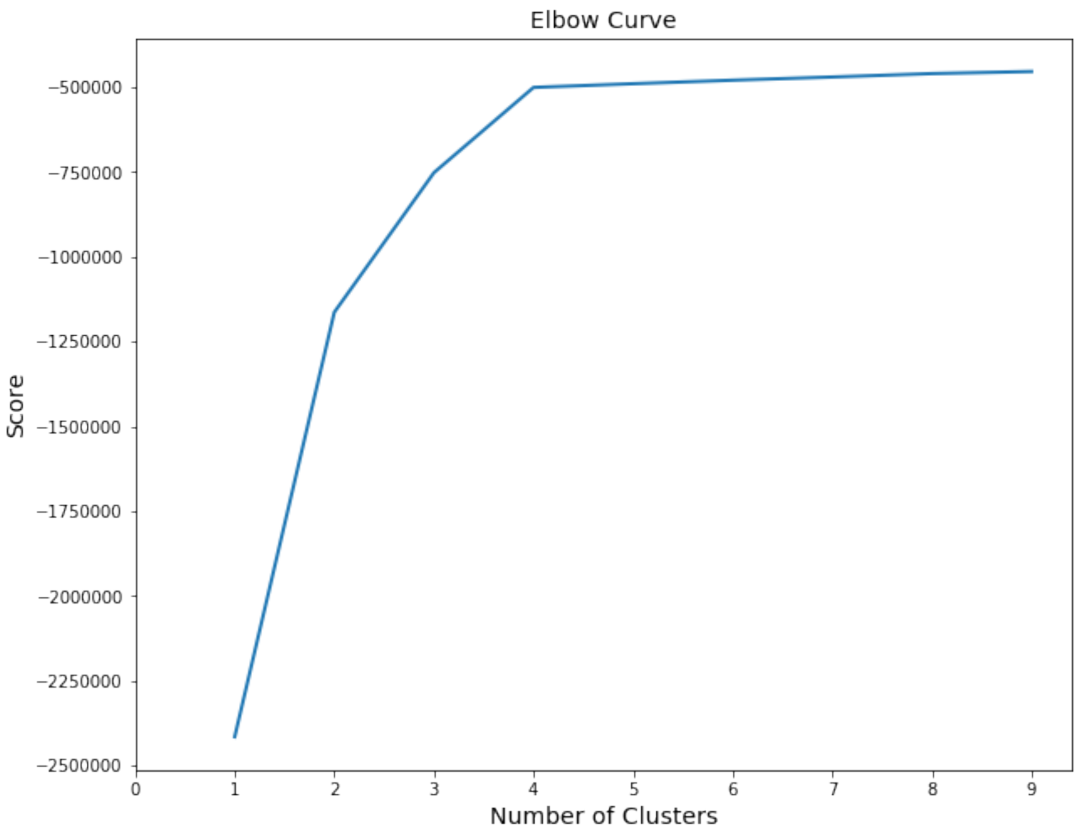


First of all, by looking at the density plots, you can determine that most of these columns have two peaks, that is there are local maxima's in each of these observations apart from the maximum value. This lends credence to the point made later about this dataset actually being something that is a concatenation of distinct datasets.

If you look at the scatter plots for the different columns, it again seems apparent that there are at least two or three different sets of data present in this dataset, as you can see more than one distinct blobs forming.

c) Someone suggests that the observations are really from multiple different files and were accidentally joined into one larger data set. Does anything about the data set suggest this? If so, how many different sources/file do you think there are?

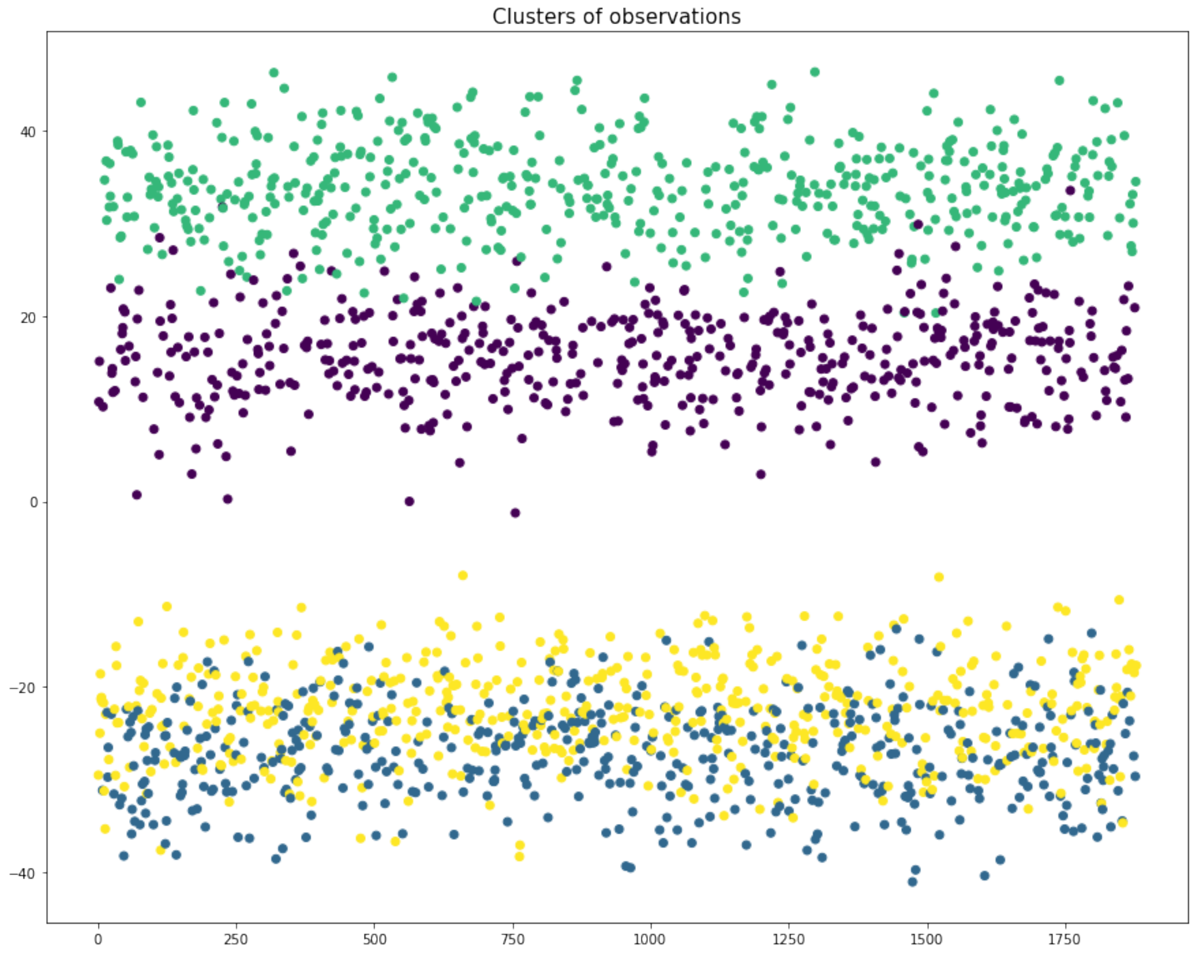
I implemented K Means clustering to find out how many clusters there might be. I used the 'elbow method' to find the optimum number of clusters. Here is the elbow curve –



It seems clear from this graph that there are 4 clusters in this dataset. This means that we can separate this dataset into 4 parts, and it is likely that these 4 parts came from different file sources, as the ‘Score’ variable in the above curve almost completely flattens after 4.

d) Bonus points: If you think there are more than one source in ds2, can you assign each observation to the right source (based on the number of sources you identified in 2c)?

I have stratified the data into four different sources. Here is a visualization of the different data sources –



I have attached the original dataset to the email with an extra column named ‘source’ which indicates the source that that particular observation belongs to. I have also separated the four sources and attached them as four different sheets in the same Excel file.