Amount of donors pledge varied over years comparing to donation amount. In the first year pledge vs donation amount is more consistent than the next two years. While total pledge amount significantly increased -for the year it is more than doubled comparing to previous year- , donation amount didn’t increased in parallel to pledge. This change would deviate the prediction of the donation amount for the current year.

After screening the distribution of the donation amount it shows that %10 of the donors donated up to $1K, %5 of the donors donated between $1K to $1M and the rest of them didn’t participate in any donation(Outliers such as $4M, $10M are excluded).

For Linear Regression Method, the strongest 11 variables were chosen when visualizing the correlation level of the features is being applied. I have applied tried-and-error method in order to decide on the number of variables. I approached in three ways to use input data. First of all, I have used the all dataset as an input and found the r^2 scores for train data as %49 and test data as %49. Then, I dropped the outliers that bigger than amount $1M and found the r^2 scores for train data as %53 and test data as %10. Lastly, I dropped the rows having zero donation amount and found the r^2 scores for train data as %44 and test data as %48. We have bias issue for this problem. To increase the model prediction score, we could add new features or could group data set based on donation amount such as up to $1k, $1k to $100k and so on, and run the model for each group.

To solve Classification Model, I followed the same Methodology of Linear Regression Model. I chose the strongest 6 variables. In each step of model overfitted. To solve the overfitting issue, I dropped the features by applying tried-and-error method but, it didn’t worked as expected. Most probably, we need to add new features to develop the model.