In 2000, Enron was one of the largest companies in the United States. By 2002, it had collapsed into bankruptcy due to widespread corporate fraud. In the resulting Federal investigation, a significant amount of typically confidential information entered into the public record, including tens of thousands of emails and detailed financial data for top executives. In this project, you will play detective, and put your new skills to use by building a person of interest identifier based on financial and email data made public as a result of the Enron scandal. To assist you in your detective work, we've combined this data with a hand-generated list of persons of interest in the fraud case, which means individuals who were indicted, reached a settlement or plea deal with the government, or testified in exchange for prosecution immunity.

Followings are all the features available in the dataset

Main features

'poi', 'salary', 'deferral\_payments', 'total\_payments', 'loan\_advances', 'bonus', estricted\_stock\_deferred', 'deferred\_income', 'total\_stock\_value', 'expenses', 'exercised\_stock\_options', 'other', 'long\_term\_incentive', 'restricted\_stock', 'director\_fees' to\_messages', 'email\_address', 'from\_poi\_to\_this\_person', 'from\_messages', 'from\_this\_person\_to\_poi', 'shared\_receipt\_with\_poi'

From above features following important feature list is selected in order to detect any fraud if any. All the selected features are numerical.

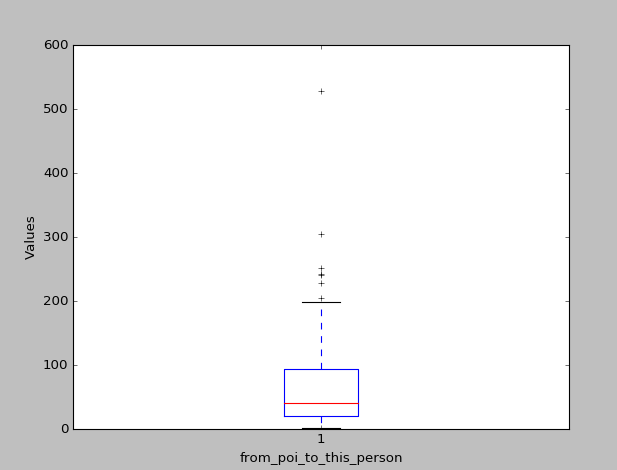
features\_list = ['poi','salary', 'deferral\_payments', 'total\_payments', 'loan\_advances', 'bonus', 'deferred\_income', 'total\_stock\_value','expenses', 'long\_term\_incentive', 'to\_messages', from\_poi\_to\_this\_person', 'from\_messages', 'from\_this\_person\_to\_poi', 'shared\_receipt\_with\_poi']

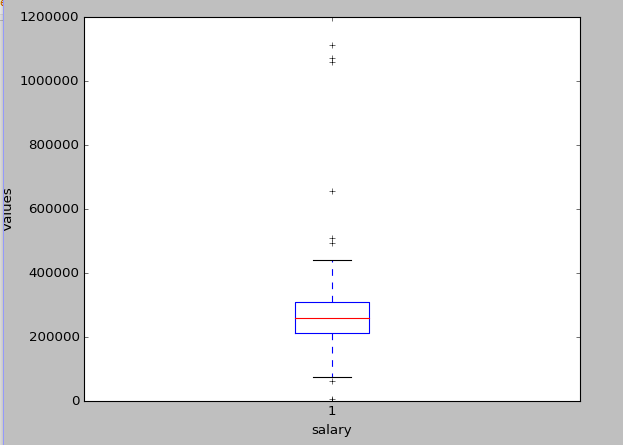
Remove outliers

The totals of all the values (salary, bonuse, to\_emails, etc.) for individual persons were shown in the main dictionary as ‘TOTAL’. The ‘TOTAL’ field deviate all the data because the total values are very high compare to other values. Therefore, the ‘TOTAL’ field remove form the dictionary using following code.

data\_dict.pop('TOTAL', None)

The function ‘create\_boxplot()’ was return to plot boxplots for each feature selected. Based on the boxplots results, there are more potential outliers. However, these outliers maybe valid data points in order to identify a fraud if any. Therefore, these data points were kept in the data dictionary.





Fracture selection

Feature extraction

PCA: Unsupervised Dimentiality reduction. Find features most explaining labels

regularization