Project 1 Report

Unsupervised Model on NY Property Valuations

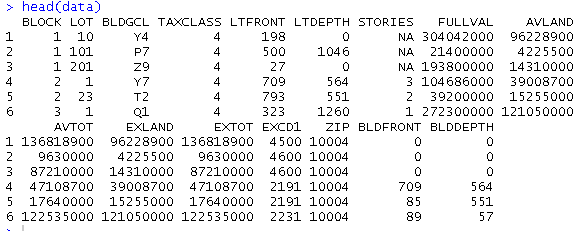
**Executive Summary**The dataset of unsupervised model on New York property valuations given in the Excel files as “property values.xlsx”. The dataset contains 1,048,575 observation (rows) and 16 entities (columns). The goal of this project is apply statistical techniques and find out which observation categorized as statistically abnormal, which lead to fraud suspicion and require further investigation on those observation. By using R statistical software to apply statistical technique and fraud detection methodology that will be thoroughly described in the next section; each observation will be analyzed and score of each observation (row) is the metric to determine if a certain observation is abnormal or not. An observation categorized as abnormal if the score lies in above 95 percentile of the entire score of dataset. By implementing this approach we found 40,224 number of observations are in the outlier which led to fraud suspicion and necessary to perform further investigation.

**Overview of the dataset**

The dataset contains 1,048,575 observation (rows) and 16 entities (columns). The list of columns are listed below:



And the snapshot of dataset are given below:



The data quality report (DQR) is attached in along with this submission

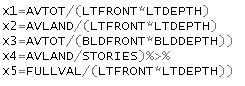
**Methodology**

The dataset contains 1,048,575 observation (rows) and 16 entities (columns). The list of entities are:





Based on these given information, 5 new variables are created for fraud analytic purpose. They are x1, x2, x3, x4, and x5:

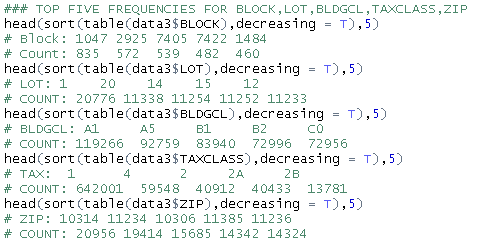


Since there are few amounts of missing value, the analysis cannot be perfectly performed before the data cleaning being employed. Based on discussion with the Professor Coggeshall, it is very important to cover as much observation as possible (around 80 percent of observations is acceptable). Data cleaning has been performed and we left with 804,471 observations, which translate into 76.72 percent of observations.

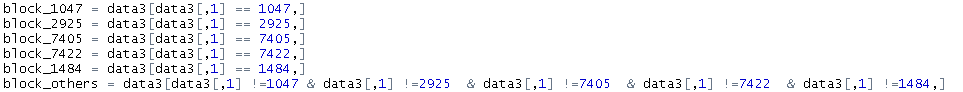
Since there are no ID number in the dataset, we used the row index as the ID. Since the row index in the CSV files is unique, we can utilize that as the ID for each observation



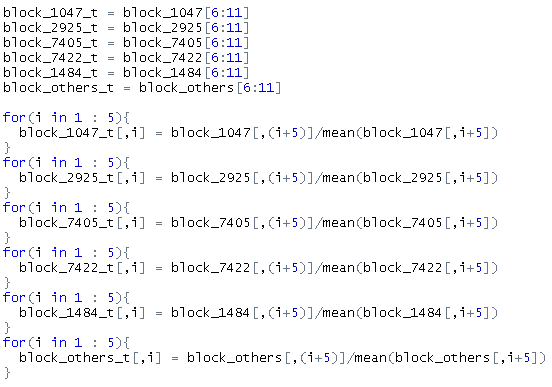
We then choose five categorical entities that we believe are necessary to do this analysis, group it into six different groups. The groups are the top 5 categorical variables with the highest frequency and the rest of the data (total = 6). The categorical entities that are selected are BLOCK, LOT, BLDGCL, TAXCLASS, and ZIP. Below are the entities that being chosen along with information and number of counts of each level:

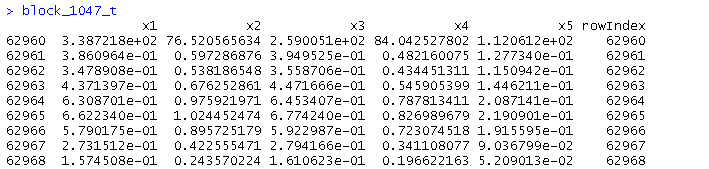


The next part is to do analysis entity by entity. In this case, BLOCK is the first entity that we would like to analyze, there are 6 different group.

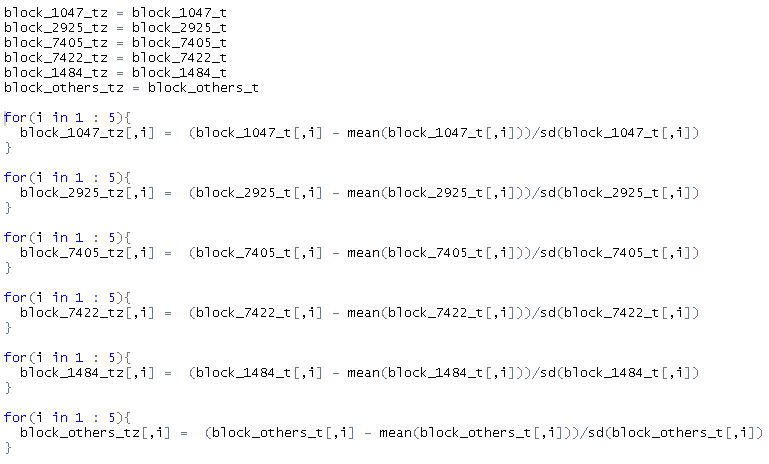


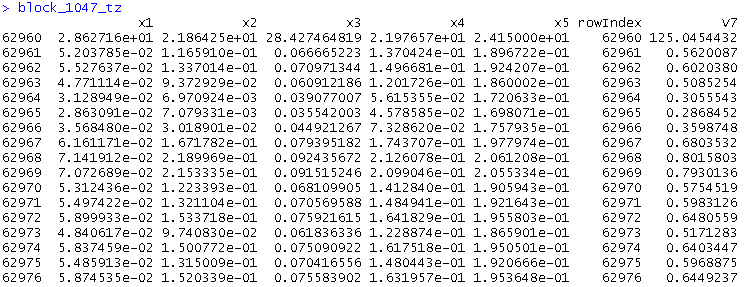
The next step is to transform each data that score (variables: x1, x2, x3, x4, x5) by dividing with the average value of each subset.



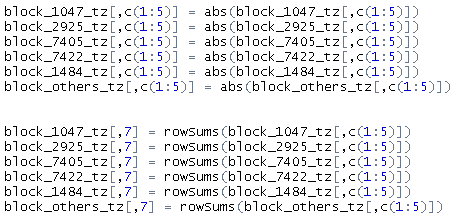


By z-scaling each score (variables: x1, x2, x3, x4, x5) based on its own subset, we can see which how many standard deviations is one score off from the mean.

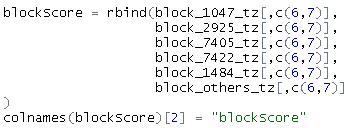




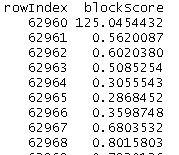
We sum all the score from variable x1 through x5 and the v7 column is the total score of each observation, based on the BLOCK grouping. We take absolute value to prevent some of the scores may be cancel out.



We then combine all the data for BLOCK section together and we get the score based on the BLOCK grouping.



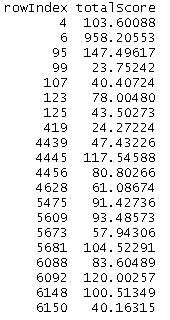
Below is some of dataset of score in the BLOCK group



We use the same methodology for the rest of selected entities, they are: LOT, BLDGCL, TAXCLASS, and ZIP. We the join them together by the “rowIndex” because the row index is unique identifier in this dataset.

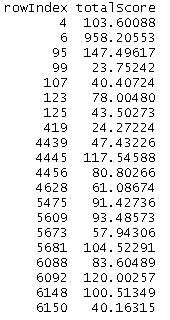


And finally we get the score of each observation:



**Model Result - Unsupervised**

The picture below are the total score of each observation. The closer the score to the value of zero means the closer the score to the mean of data set. And the greater the score, the further the score from the mean. In other words, the more the score, the more suspicious the observation is because it’s far away from “normal”



The descriptive statistic of the score is   
mean/average:



Standard deviation:



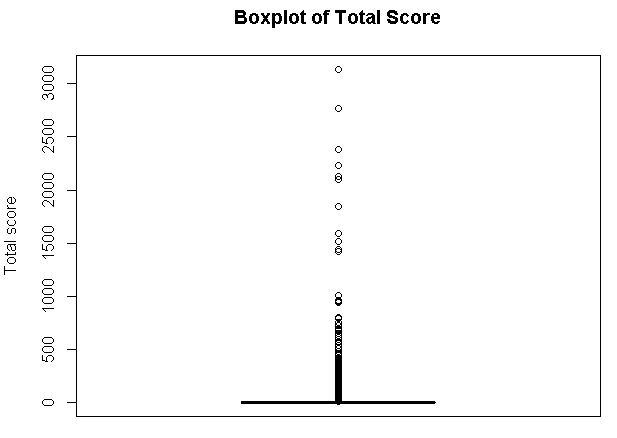
10 bins quantile of score distribution:  


To determine the outlier in this dataset, we decide find the value top 5 percent out of dataset, which we considered the score is too big to be considered normal. We looked up for the 95 percentile of dataset which is:

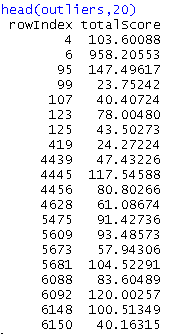


And we filtered the dataset which has the score above 12.00626  

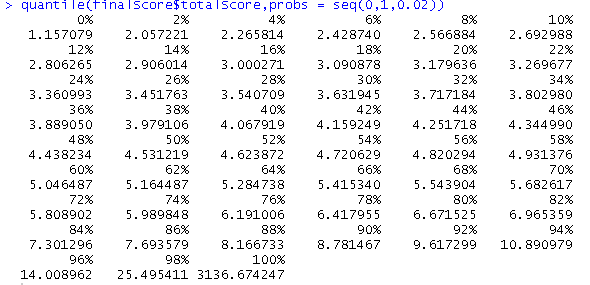

The boxplot for the total score:



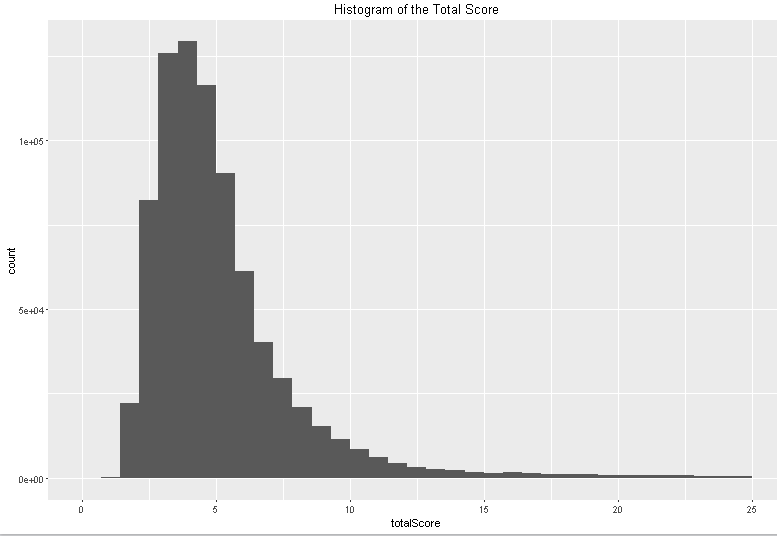
There are 40,224 observations we considered as outliers (top 5 percent of the scores), and here’s the snapshot of the outlier dataset



50 bins quantile analysis and histogram based on the quantile distribution:



Histogram of the total score:



**Discussion**

Based on the histogram plot of the scores, the distribution is right-skewed distribution. The fact that it is right skewed is considered good, because that means majority of the dataset have score that close to zero. It means that there are number of abnormal observations is considerably fewer than normal observation. For this project, we decided that top 5 percent of total score is considered as potential fraud. And out of 804,471 observations, we found 40,224 observation are potential fraud. However, it does not means that the rest of the observation is fraud-free because there are other fraud detection method that may have better detection system in this particular situation. Afterwards, this method is considered good enough in fraud detections system for NY property data because the histogram that we see shows a normal-like right skewed distribution, which is logically correct and expected based on our algorithm.