Fraud Analytics Project 1 Unsupervised learning on New York Property valuation

**Prepared by: Rupantar Rana**

**MS Business Analytics Student Marshall School of Business University of Southern California** [**rupantar@marshall.usc.edu**](mailto:rupantar@marshall.usc.edu)

MARCH 2016

**TABLE OF CONTENTS**

Introduction and Motivation ....................................................................................... Data source identification and description………………………………………………………………….………...

Data Cleaning…………………………………………………………………………

Building Variables…………………………………………………………………….

Building Entity Levels ……………………………………………………………….

Procedure 1: Principal components regression………………………………….

Procedure 2: Mahalanobis Distance ……………………………………………

Evaluate Results ............................................................................................................

Executive Summary ……………………………………………………....................

**Project Introduction and Motivation**

The objective of this project is to build unsupervised model on the NY property evaluation data to identify properties that have been fraudulently evaluated. The data set consists of 1048575 rows and 26 features. The dataset contains the property valuations for the fiscal year 2010/2011.

Initial feature set:

BBLE

BLOCK

LOT

EASEMENT

OWNER

BLDGCL

TAXCLASS

LTFRONT

LTDEPTH

STORIES

FULLVAL

AVLAND

AVTOT

EXLAND

EXTOT

EXCD1

STADDR

ZIP

EXMPTCL

BLDFRONT

BLDDEPTH

AVLAND2

AVTOT2

EXLAND2

EXTOT2

EXCD2

PERIOD

YEAR

VALTYPE

**Variable Description / Data Dictionary**

**BBLE**

Length 11 alphanumeric

Concatenation of AV\_BORO, AV\_BLOCK, AV\_LOT, AV\_EASEMENT, descriptions of which follow.

**BLOCK**

Length 5 numeric

VALID BLOCK RANGES BY BORO

MANHATTAN 1 TO 2,255

BRONX 2,260 TO 5,958

BROOKLYN 1 TO 8,955

QUEENS 1 TO 16,350

STATEN ISLAND 1 TO 8,050

**LOT**

Length 4 numeric

UNIQUE # WITHIN BORO/BLOCK.

**EASE**

Length 1 alpha

IS A FIELD THAT IS USED TO DESCRIBE EASEMENT.

SPACE Indicates the lot has no Easement.

'A' Indicates the portion of the Lot that has an Air Easement

'B' Indicates Non-Air Rights.

'E' Indicates the portion of the lot that has a Land Easement

'F' THRU 'M' Are duplicates of 'E'.

'N' Indicates Non-Transit Easement

'P' Indicates Piers.

R' Indicates Railroads.

'S' Indicates Street

'U' Indicates U.S. Government

**YEAR**

4 Length 4 Numeric

Four-digit year of the file. For example: if the year4 = 2001

the current values are for the Fiscal year 2001/2002 assessments.

The Tentative and Final value contain the predicted values for the 2002/2003

fiscal year.

**TAX-CLASS**

Length 2 Character

Current Property Tax Class Code (NYS Classification)

VALID VALUES -

TAX CLASS 1 = 1-3 UNIT RESIDENCES

TAX CLASS 1A = 1-3 STORY CONDOMINIUMS

ORIGINALLY A CONDO

TAX CLASS 1B = RESIDENTIAL VACANT LAND

TAX CLASS 1C = 1-3 UNIT CONDOMINUMS

ORIGINALLY TAX CLASS 1

TAX CLASS 1D = SELECT BUNGALOW COLONIES

TAX CLASS 2 = APARTMENTS

TAX CLASS 2A = APARTMENTS WITH 4-6 UNITS

TAX CLASS 2B = APARTMENTS WITH 7-10 UNITS

TAX CLASS 2C = COOPS/CONDOS WITH 2-10 UNITS

TAX CLASS 3 = UTILITIES (EXCEPT CEILING RR)

TAX CLASS 4A = UTILITIES - CEILING RAILROADS

TAX CLASS 4 = ALL OTHERS

**OWNER**

Length 21 Character

The Owner’s Name.

**ZIP**

Length 5 numeric (no decimals)

Postal Zip code of the property

**STADDR**

Length 21 Character

The street address

**LTFRONT**

DEC Length 7 Numeric (9999.99)

Lot Frontage in feet.

**LOTDEP**

DEC Length 7 Numeric (9999.99)

Lot Depth in feet.

**BLDFRONT**

DEC Length 7 Numeric (9999.99)

Building Frontage in feet.

**BLDDEPTH**-DEC Length 7 Numeric (9999.99)

Lot Depth in feet.

**MARKET VALUES**

**AVLAND**

FULLVAL-LAND

Length 11 numeric (no decimals)

If not zero, Current year’s total market value of the land

**AVTOT**

FULLVAL-TOTAL

Length 11 numeric (no decimals)

If not zero, Current year’s total market value

**FULLVAL**

Length 11 numeric (no decimals)

If not zero, New Total Market Value of property

**Data Cleaning**

Step 1:

Remove the features that have no predictive power. These features remained constant throughout the data set.

Variables Removed:

Period

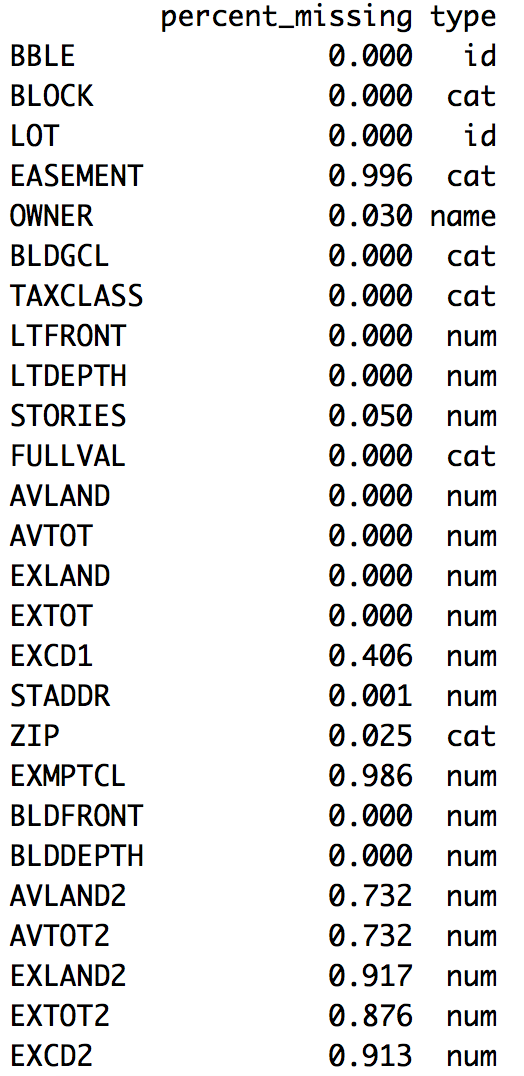
Year

Valtype

We removed these variables by assigning them Null values.

Step 2:

Calculate the % of data that is missing for each of the feature:



We then use 70 % as the threshold for removing variables that more than 70% missing values. The following variables were removed from the dataset.

EASEMENT

EXMPTCL

EXCD2

EXLAND2

EXTOT2

AVLAND2

AVTOT2

**Building Variables:**

Using the existing variables in the dataset we start to build new variables that include the following.

LTAREA = LTFRONT \* LTDEPTH

BLDAREA = BLDFRONT \* BLDDEPTH

Full value per LtArea = Fullvalue / LtArea

Full value per Bld Area = Fullvalue / BldArea

AvtotPerAvLand = Avtotal / Avland

**ENTITY LEVELS:**

Entity levels used for each of these variables include:

BLDGCL

TAXCLASS

STORIES

ZIP

For each of the entities we calculate the mean and append it to the dataset then we divide the feature by the mean to build entity level variables.

The new dataset has 44 variables with 20 numerical variables that can be used for model building. These variables include:

**Variables name : ValPerLtArea**

ValPerLtAreaByBLDGCL

ValPerLtAreaByTaxcls

ValPerLtAreaByZipcls

ValPerLtAreaBystoriescls

**Variable name : ValPerBldArea**

ValPerBldAreaByBLDGCL

ValPerBldAreaByTaxcls

ValPerBldAreaByZipcls

ValPerBldAreaBystoriescls

**Variable name : AvtotPerAvLand**

AvtotPerAvLandByBLDGCL

AvtotPerAvLandByTaxclass

AvtotPerAvLandByZipcls

AvtotPerAvLandByStories

**Variable name : LTAREA**

LTAREAByBLDGCL

LTAREAByTaxcls

LTAREAByZip

LTAREAByStories

**Variable name : BLDAREA**

BLDAREAByBLDGCL

BLDAREAByTaxcls

BLDAREAByZip

BLDAREAByStories

We use these 20 new variables to build our model.

**Model Building**

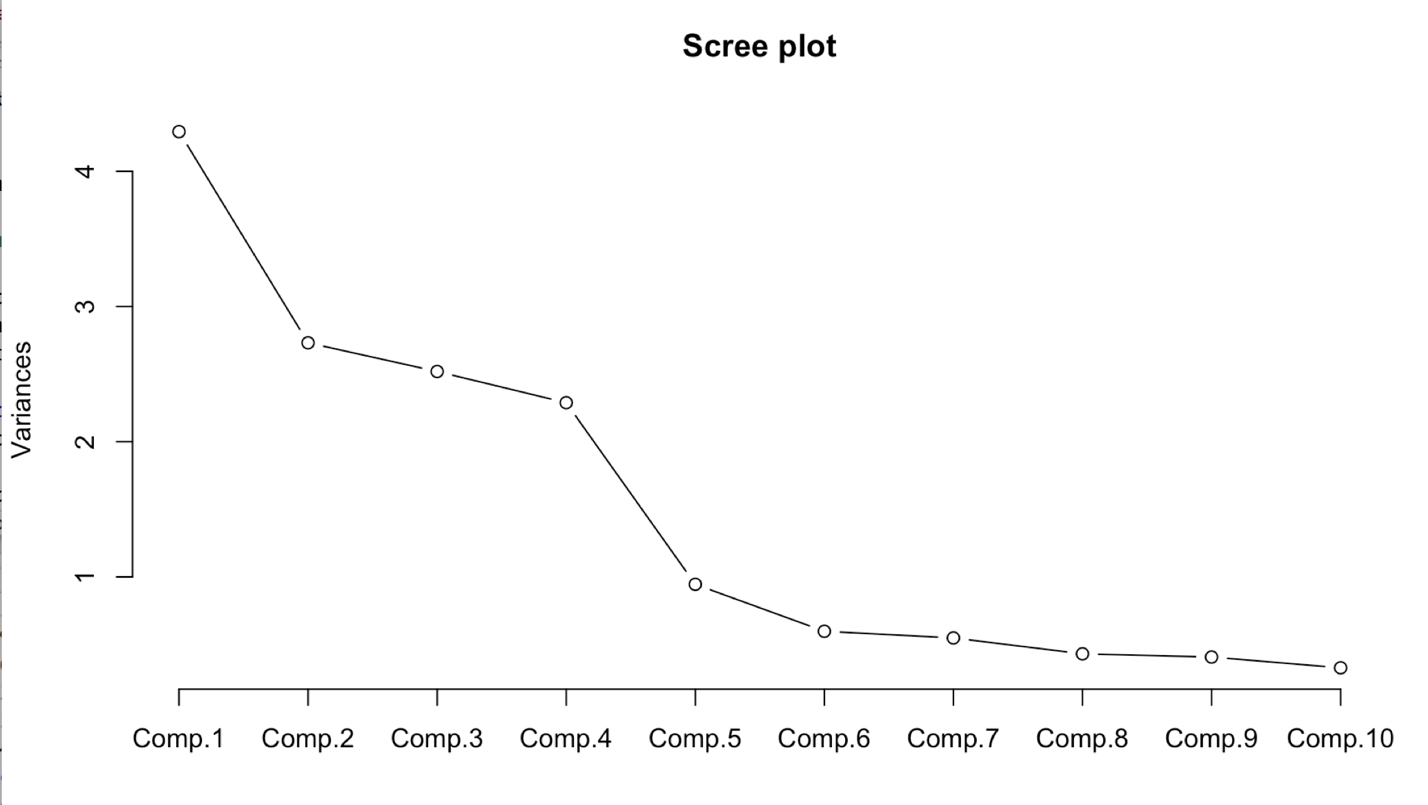
**Principal Components Model**

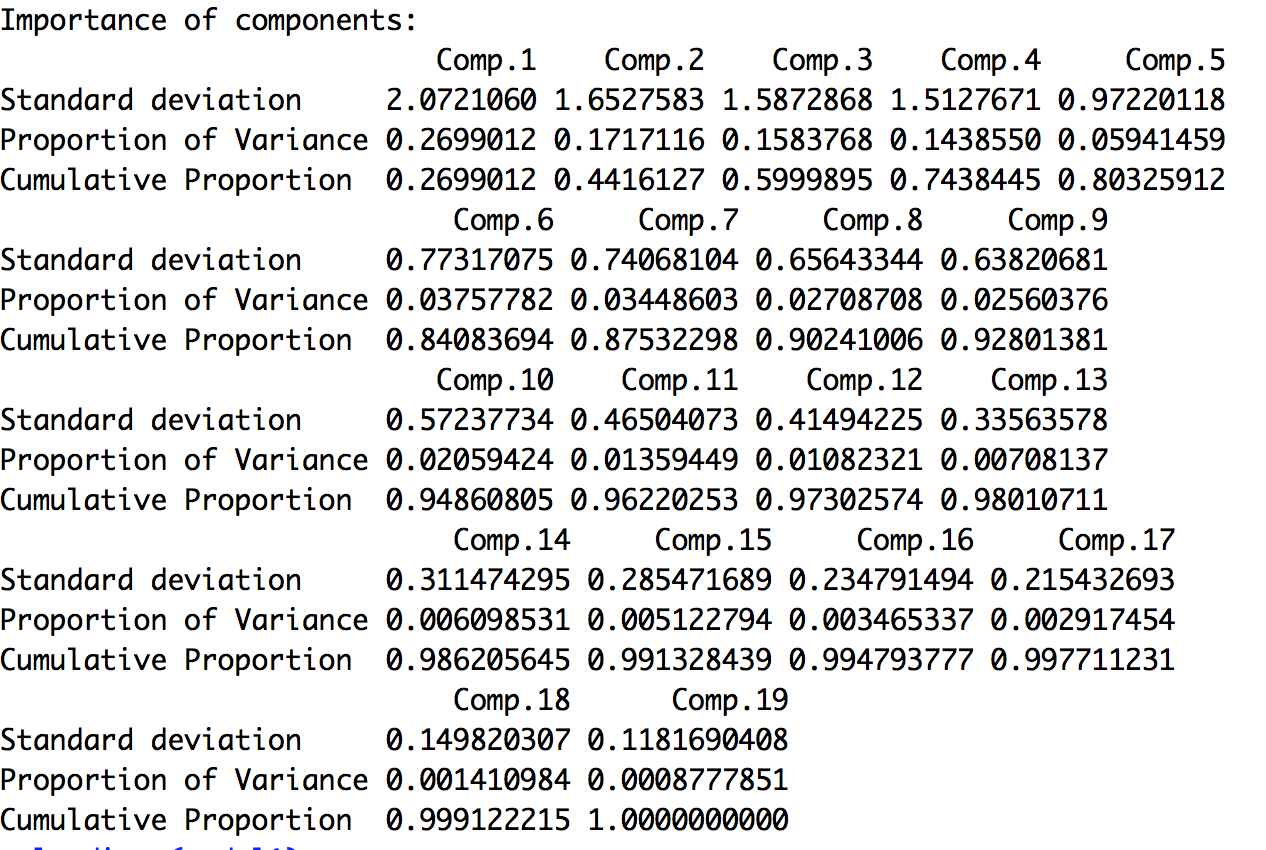
**Scaling**

We use the scale function in R to scale the features so that the it becomes easier for us to calculate the covariance matrix when we build the PCA model. We also removing ‘NA’ and ‘Inf’ valued rows.

Now we have 8,78,471 rows and 20 columns of scaled data to build the PCA model.

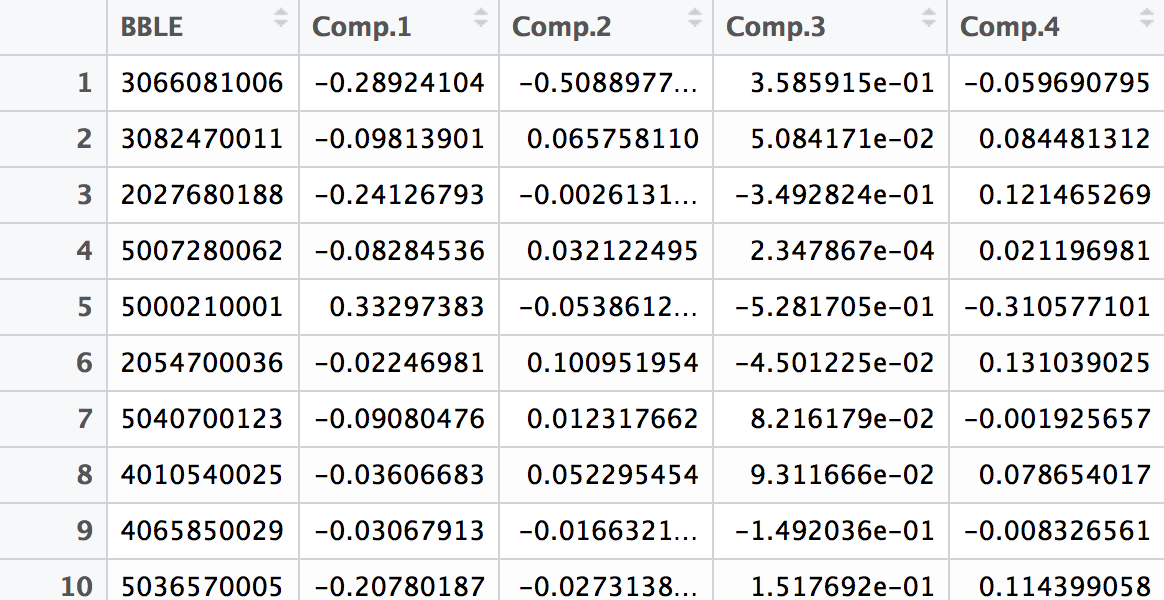
**Principal components Analysis**

****

****

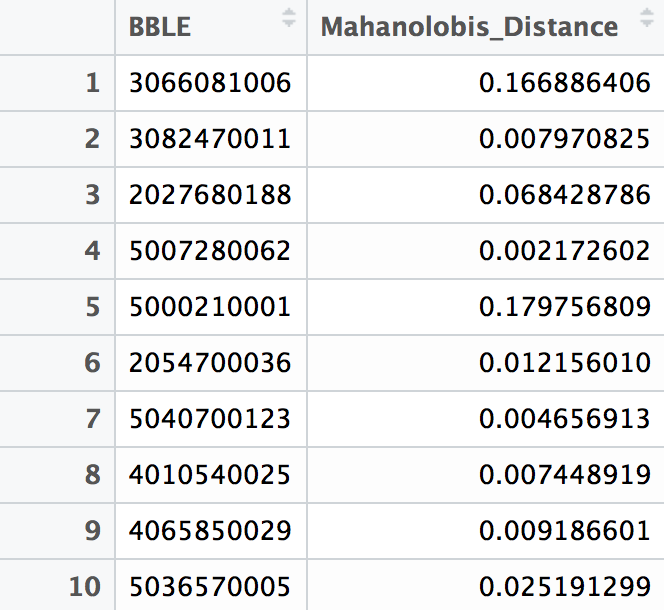
From the above figure we find that almost 80% of the variation in the data is explained by the first 5 components. For the most important components we choose the components that have eigen values (square of the standard deviation) greater than 1. Based on the eigen values, we choose the first 4 components for outlier detection. Now we have successfully reduced the dimension of the dataset from 20 features to 4 principal components.

**Principal components dataset**

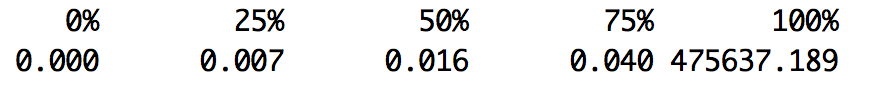


**Anomaly score: Mahalanobis distance**

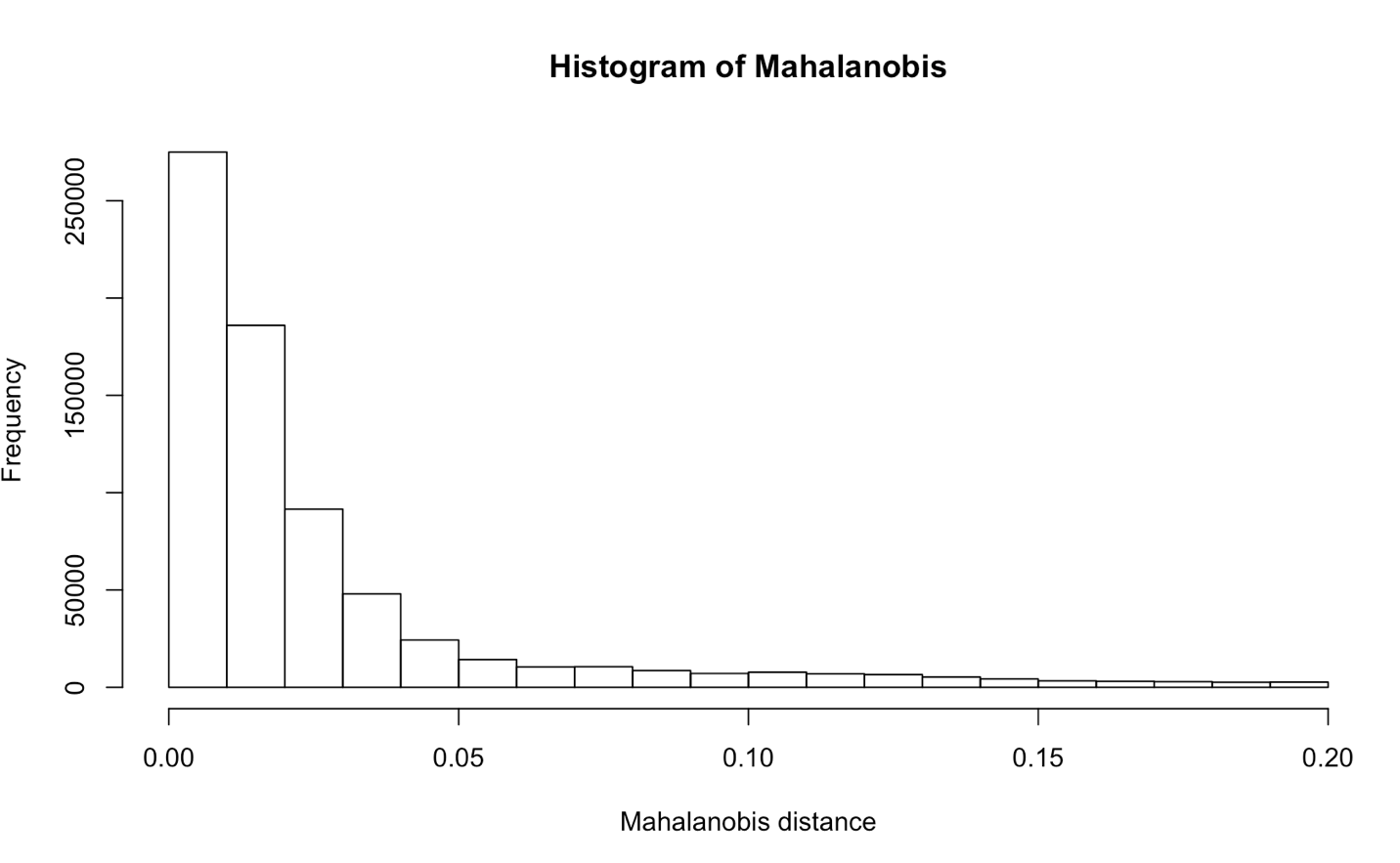
The multivariate model uses the mahalanobis distance to calculate the anomaly score for each observation in the dataset. The mahalanobis distance calculation is done based on the 4 principal component features derived earlier. After calculation of the mahalanobis distance, the BBLE was sorted based on the distance metric that was calculated.

****

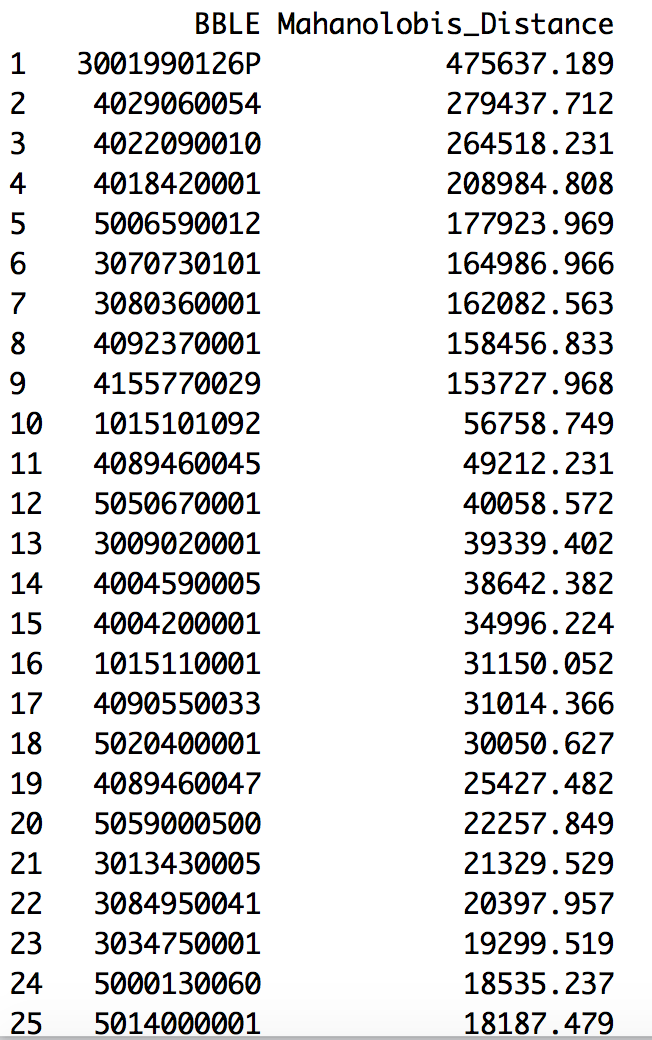
**Mahanalobis Quantiles**

****

Please note here that 75 % of the mahalanobis distance is below 0.04. There are about 111 data points that have mahalanobis distance greater than 1000. The maximum distance is for the property owned by the Port of New York.

****

**Mahanobilis outliers**

****

Executive Summary

Feature engineering was used to create new entity level features that were used as the input feature set for the Principal components analysis. Out of the PCAs calculated for each of the features only 4 PCAs with maximum proportion of variation explained were chosen. For Anomaly detection, Mahalanobis distance was calculated using the 4 PCAs. Data set clearly showed around 111 outliers based on the mahalanobis distance. Although the method used above is based on classical statistics, machine learning based models such as autoencoders could be used for outlier detection.