import pandas as pd

import numpy as np

import scipy.stats as sps

import matplotlib.pyplot as plt

import seaborn as sns

import scipy.stats as stats

import sklearn as skl

from sklearn import preprocessing

from sklearn.decomposition import PCA

mydata = pd.read\_csv('superfinal.csv')

len(mydata)

mydata.dtypes

mydata['lotarea'] = mydata['LTFRONT']\*mydata['LTDEPTH']

mydata['bldarea'] = mydata['BLDFRONT']\*mydata['BLDDEPTH']

mydata['bldvol'] = mydata['bldarea']\*mydata['﻿STORIES']

mydata['fullval\_la'] = mydata['FULLVAL']/mydata['lotarea']

mydata['fullval\_ba'] = mydata['FULLVAL']/mydata['bldarea']

mydata['fullval\_bv'] = mydata['FULLVAL']/mydata['bldvol']

mydata['avland\_la'] = mydata['AVLAND']/mydata['lotarea']

mydata['avland\_ba'] = mydata['AVLAND']/mydata['bldarea']

mydata['avland\_bv'] = mydata['AVLAND']/mydata['bldvol']

mydata['avtot\_la'] = mydata['AVTOT']/mydata['lotarea']

mydata['avtot\_ba'] = mydata['AVTOT']/mydata['bldarea']

mydata['avtot\_bv'] = mydata['AVTOT']/mydata['bldvol']

mydata['exland\_la'] = mydata['EXLAND'] / mydata['lotarea']

mydata['exland\_ba'] = mydata['EXLAND'] / mydata['bldarea']

mydata['exland\_bv'] = mydata['EXLAND'] / mydata['bldvol']

mydata['extot\_la'] = mydata['EXTOT'] / mydata['lotarea']

mydata['extot\_ba'] = mydata['EXTOT'] / mydata['bldarea']

mydata['extot\_bv'] = mydata['EXTOT'] / mydata['bldvol']

mydata['borough'] = mydata['BBLE'].astype(str).str[0]

mydata['borough'] = mydata['borough'].astype(int)

mydata['borough'].value\_counts()

zip3\_means = mydata.groupby('ZIP3').mean()

zip5\_means = mydata.groupby('ZIP5').mean()

taxclass\_means = mydata.groupby('TAXCLASS').mean()

borough\_means = mydata.groupby('borough').mean()

consolidated\_means\_dict = {

k: {

c: mydata[c].to\_dict()

for c in mydata.columns.values

} for k, mydata in zip(

['zip3\_means', 'zip5\_means', 'taxclass\_means', 'borough\_means'],

[zip3\_means, zip5\_means, taxclass\_means, borough\_means]

)

}

#consolidated\_means\_dict['all\_means'] = all\_means.to\_dict()

def calc\_vars(row\_data):

izip5 = row\_data['ZIP5']

izip3 = row\_data['ZIP3']

itc = row\_data['TAXCLASS']

ibo = row\_data['borough']

row\_vars = pd.Series()

row\_vars['fv\_la\_z3'] = row\_data['fullval\_la']/consolidated\_means\_dict['zip3\_means']['fullval\_la'][izip3]

row\_vars['vl\_la\_z3'] = row\_data['avland\_la']/consolidated\_means\_dict['zip3\_means']['avland\_la'][izip3]

row\_vars['vt\_la\_z3'] = row\_data['avtot\_la']/consolidated\_means\_dict['zip3\_means']['avtot\_la'][izip3]

row\_vars['xl\_la\_z3'] = row\_data['exland\_la']/consolidated\_means\_dict['zip3\_means']['exland\_la'][izip3]

row\_vars['xt\_la\_z3'] = row\_data['extot\_la']/consolidated\_means\_dict['zip3\_means']['extot\_la'][izip3]

row\_vars['fv\_la\_z5'] = row\_data['fullval\_la']/consolidated\_means\_dict['zip5\_means']['fullval\_la'][izip5]

row\_vars['vl\_la\_z5'] = row\_data['avland\_la']/consolidated\_means\_dict['zip5\_means']['avland\_la'][izip5]

row\_vars['vt\_la\_z5'] = row\_data['avtot\_la']/consolidated\_means\_dict['zip5\_means']['avtot\_la'][izip5]

row\_vars['xl\_la\_z5'] = row\_data['exland\_la']/consolidated\_means\_dict['zip5\_means']['exland\_la'][izip5]

row\_vars['xt\_la\_z5'] = row\_data['extot\_la']/consolidated\_means\_dict['zip5\_means']['extot\_la'][izip5]

row\_vars['fv\_la\_tc'] = row\_data['fullval\_la']/consolidated\_means\_dict['taxclass\_means']['fullval\_la'][itc]

row\_vars['vl\_la\_tc'] = row\_data['avland\_la']/consolidated\_means\_dict['taxclass\_means']['avland\_la'][itc]

row\_vars['vt\_la\_tc'] = row\_data['avtot\_la']/consolidated\_means\_dict['taxclass\_means']['avtot\_la'][itc]

row\_vars['xl\_la\_tc'] = row\_data['exland\_la']/consolidated\_means\_dict['taxclass\_means']['exland\_la'][itc]

row\_vars['xt\_la\_tc'] = row\_data['extot\_la']/consolidated\_means\_dict['taxclass\_means']['extot\_la'][itc]

row\_vars['fv\_la\_bo'] = row\_data['fullval\_la']/consolidated\_means\_dict['borough\_means']['fullval\_la'][ibo]

row\_vars['vl\_la\_bo'] = row\_data['avland\_la']/consolidated\_means\_dict['borough\_means']['avland\_la'][ibo]

row\_vars['vt\_la\_bo'] = row\_data['avtot\_la']/consolidated\_means\_dict['borough\_means']['avtot\_la'][ibo]

row\_vars['xl\_la\_bo'] = row\_data['exland\_la']/consolidated\_means\_dict['borough\_means']['exland\_la'][ibo]

row\_vars['xt\_la\_bo'] = row\_data['extot\_la']/consolidated\_means\_dict['borough\_means']['extot\_la'][ibo]

row\_vars['fv\_la\_none'] = row\_data['fullval\_la']

row\_vars['vl\_la\_none'] = row\_data['avland\_la']

row\_vars['vt\_la\_none'] = row\_data['avtot\_la']

row\_vars['xl\_la\_none'] = row\_data['exland\_la']

row\_vars['xt\_la\_none'] = row\_data['extot\_la']

row\_vars['fv\_ba\_z3'] = row\_data['fullval\_ba']/consolidated\_means\_dict['zip3\_means']['fullval\_ba'] [izip3]

row\_vars['vl\_ba\_z3'] = row\_data['avland\_ba']/consolidated\_means\_dict['zip3\_means']['avland\_ba'][izip3]

row\_vars['vt\_ba\_z3'] = row\_data['avtot\_ba']/consolidated\_means\_dict['zip3\_means']['avtot\_ba'][izip3]

row\_vars['xl\_ba\_z3'] = row\_data['exland\_ba']/consolidated\_means\_dict['zip3\_means']['exland\_ba'][izip3]

row\_vars['xt\_ba\_z3'] = row\_data['extot\_ba']/consolidated\_means\_dict['zip3\_means']['extot\_ba'][izip3]

row\_vars['fv\_ba\_z5'] = row\_data['fullval\_ba']/consolidated\_means\_dict['zip5\_means']['fullval\_ba'][izip5]

row\_vars['vl\_ba\_z5'] = row\_data['avland\_ba']/consolidated\_means\_dict['zip5\_means']['avland\_ba'][izip5]

row\_vars['vt\_ba\_z5'] = row\_data['avtot\_ba']/consolidated\_means\_dict['zip5\_means']['avtot\_ba'][izip5]

row\_vars['xl\_ba\_z5'] = row\_data['exland\_ba']/consolidated\_means\_dict['zip5\_means']['exland\_ba'][izip5]

row\_vars['xt\_ba\_z5'] = row\_data['extot\_ba']/consolidated\_means\_dict['zip5\_means']['extot\_ba'][izip5]

row\_vars['fv\_ba\_tc'] = row\_data['fullval\_ba']/consolidated\_means\_dict['taxclass\_means']['fullval\_ba'][itc]

row\_vars['vl\_ba\_tc'] = row\_data['avland\_ba']/consolidated\_means\_dict['taxclass\_means']['avland\_ba'][itc]

row\_vars['vt\_ba\_tc'] = row\_data['avtot\_ba']/consolidated\_means\_dict['taxclass\_means']['avtot\_ba'][itc]

row\_vars['xl\_ba\_tc'] = row\_data['exland\_ba']/consolidated\_means\_dict['taxclass\_means']['exland\_ba'][itc]

row\_vars['xt\_ba\_tc'] = row\_data['extot\_ba']/consolidated\_means\_dict['taxclass\_means']['extot\_ba'][itc]

row\_vars['fv\_ba\_bo'] = row\_data['fullval\_ba']/consolidated\_means\_dict['borough\_means']['fullval\_ba'][ibo]

row\_vars['vl\_ba\_bo'] = row\_data['avland\_ba']/consolidated\_means\_dict['borough\_means']['avland\_ba'][ibo]

row\_vars['vt\_ba\_bo'] = row\_data['avtot\_ba']/consolidated\_means\_dict['borough\_means']['avtot\_ba'][ibo]

row\_vars['xl\_ba\_bo'] = row\_data['exland\_ba']/consolidated\_means\_dict['borough\_means']['exland\_ba'][ibo]

row\_vars['xt\_ba\_bo'] = row\_data['extot\_ba']/consolidated\_means\_dict['borough\_means']['extot\_ba'][ibo]

row\_vars['fv\_ba\_none'] = row\_data['fullval\_ba']

row\_vars['vl\_ba\_none'] = row\_data['avland\_ba']

row\_vars['vt\_ba\_none'] = row\_data['avtot\_ba']

row\_vars['xl\_ba\_none'] = row\_data['exland\_ba']

row\_vars['xt\_ba\_none'] = row\_data['extot\_ba']

row\_vars['fv\_bv\_z3'] = row\_data['fullval\_bv']/consolidated\_means\_dict['zip3\_means']['fullval\_bv'][izip3]

row\_vars['vl\_bv\_z3'] = row\_data['avland\_bv']/consolidated\_means\_dict['zip3\_means']['avland\_bv'][izip3]

row\_vars['vt\_bv\_z3'] = row\_data['avtot\_bv']/consolidated\_means\_dict['zip3\_means']['avtot\_bv'][izip3]

row\_vars['xl\_bv\_z3'] = row\_data['exland\_bv']/consolidated\_means\_dict['zip3\_means']['exland\_bv'][izip3]

row\_vars['xt\_bv\_z3'] = row\_data['extot\_bv']/consolidated\_means\_dict['zip3\_means']['extot\_bv'][izip3]

row\_vars['fv\_bv\_z5'] = row\_data['fullval\_bv']/consolidated\_means\_dict['zip5\_means']['fullval\_bv'][izip5]

row\_vars['vl\_bv\_z5'] = row\_data['avland\_bv']/consolidated\_means\_dict['zip5\_means']['avland\_bv'][izip5]

row\_vars['vt\_bv\_z5'] = row\_data['avtot\_bv']/consolidated\_means\_dict['zip5\_means']['avtot\_bv'][izip5]

row\_vars['xl\_bv\_z5'] = row\_data['exland\_bv']/consolidated\_means\_dict['zip5\_means']['exland\_bv'][izip5]

row\_vars['xt\_bv\_z5'] = row\_data['extot\_bv']/consolidated\_means\_dict['zip5\_means']['extot\_bv'][izip5]

row\_vars['fv\_bv\_tc'] = row\_data['fullval\_bv']/consolidated\_means\_dict['taxclass\_means']['fullval\_bv'][itc]

row\_vars['vl\_bv\_tc'] = row\_data['avland\_bv']/consolidated\_means\_dict['taxclass\_means']['avland\_bv'][itc]

row\_vars['vt\_bv\_tc'] = row\_data['avtot\_bv']/consolidated\_means\_dict['taxclass\_means']['avtot\_bv'][itc]

row\_vars['xl\_bv\_tc'] = row\_data['exland\_bv']/consolidated\_means\_dict['taxclass\_means']['exland\_bv'][itc]

row\_vars['xt\_bv\_tc'] = row\_data['extot\_bv']/consolidated\_means\_dict['taxclass\_means']['extot\_bv'][itc]

row\_vars['fv\_bv\_bo'] = row\_data['fullval\_bv']/consolidated\_means\_dict['borough\_means']['fullval\_bv'][ibo]

row\_vars['vl\_bv\_bo'] = row\_data['avland\_bv']/consolidated\_means\_dict['borough\_means']['avland\_bv'][ibo]

row\_vars['vt\_bv\_bo'] = row\_data['avtot\_bv']/consolidated\_means\_dict['borough\_means']['avtot\_bv'][ibo]

row\_vars['xl\_bv\_bo'] = row\_data['exland\_bv']/consolidated\_means\_dict['borough\_means']['exland\_bv'][ibo]

row\_vars['xt\_bv\_bo'] = row\_data['extot\_bv']/consolidated\_means\_dict['borough\_means']['extot\_bv'][ibo]

row\_vars['fv\_bv\_none'] = row\_data['fullval\_bv']

row\_vars['vl\_bv\_none'] = row\_data['avland\_bv']

row\_vars['vt\_bv\_none'] = row\_data['avtot\_bv']

row\_vars['xl\_bv\_none'] = row\_data['exland\_bv']

row\_vars['xt\_bv\_none'] = row\_data['extot\_bv']

row\_vars['fv\_none\_z3'] = row\_data['FULLVAL']/consolidated\_means\_dict['zip3\_means']['FULLVAL'] [izip3]

row\_vars['vl\_none\_z3'] = row\_data['AVLAND']/consolidated\_means\_dict['zip3\_means']['AVLAND'][izip3]

row\_vars['vt\_none\_z3'] = row\_data['AVTOT']/consolidated\_means\_dict['zip3\_means']['AVTOT'][izip3]

row\_vars['xl\_none\_z3'] = row\_data['EXLAND']/consolidated\_means\_dict['zip3\_means']['EXLAND'][izip3]

row\_vars['xt\_none\_z3'] = row\_data['EXTOT']/consolidated\_means\_dict['zip3\_means']['EXTOT'][izip3]

row\_vars['fv\_none\_z5'] = row\_data['FULLVAL']/consolidated\_means\_dict['zip5\_means']['FULLVAL'][izip5]

row\_vars['vl\_none\_z5'] = row\_data['AVLAND']/consolidated\_means\_dict['zip5\_means']['AVLAND'][izip5]

row\_vars['vt\_none\_z5'] = row\_data['AVTOT']/consolidated\_means\_dict['zip5\_means']['AVTOT'][izip5]

row\_vars['xl\_none\_z5'] = row\_data['EXLAND']/consolidated\_means\_dict['zip5\_means']['EXLAND'][izip5]

row\_vars['xt\_none\_z5'] = row\_data['EXTOT']/consolidated\_means\_dict['zip5\_means']['EXTOT'][izip5]

row\_vars['fv\_none\_tc'] = row\_data['FULLVAL']/consolidated\_means\_dict['taxclass\_means']['FULLVAL'][itc]

row\_vars['vl\_none\_tc'] = row\_data['AVLAND']/consolidated\_means\_dict['taxclass\_means']['AVLAND'][itc]

row\_vars['vt\_none\_tc'] = row\_data['AVTOT']/consolidated\_means\_dict['taxclass\_means']['AVTOT'][itc]

row\_vars['xl\_none\_tc'] = row\_data['EXLAND']/consolidated\_means\_dict['taxclass\_means']['EXLAND'][itc]

row\_vars['xt\_none\_tc'] = row\_data['EXTOT']/consolidated\_means\_dict['taxclass\_means']['EXTOT'][itc]

row\_vars['fv\_none\_bo'] = row\_data['FULLVAL']/consolidated\_means\_dict['borough\_means']['FULLVAL'][ibo]

row\_vars['vl\_none\_bo'] = row\_data['AVLAND']/consolidated\_means\_dict['borough\_means']['AVLAND'][ibo]

row\_vars['vt\_none\_bo'] = row\_data['AVTOT']/consolidated\_means\_dict['borough\_means']['AVTOT'][ibo]

row\_vars['xl\_none\_bo'] = row\_data['EXLAND']/consolidated\_means\_dict['borough\_means']['EXLAND'][ibo]

row\_vars['xt\_none\_bo'] = row\_data['EXTOT']/consolidated\_means\_dict['borough\_means']['EXTOT'][ibo]

row\_vars['fv\_none\_none'] = row\_data['FULLVAL']

row\_vars['vl\_none\_none'] = row\_data['AVLAND']

row\_vars['vt\_none\_none'] = row\_data['AVTOT']

row\_vars['xl\_none\_none'] = row\_data['EXLAND']

row\_vars['xt\_none\_none'] = row\_data['EXTOT']

return row\_vars

#WARNING: The following command line will take ~15 hours to run.

myvars = mydata.apply(calc\_vars, axis = 1)

myvars.shape

myvars.dtypes

myvars.to\_csv('myvars\_15h.csv', encoding='utf-8')

myvars.head(5)

mydata.head(5)

%%time

expert\_var = pd.concat([mydata, myvars], axis = 1)

master\_expert\_var = expert\_var[['fv\_la\_z3','vl\_la\_z3','vt\_la\_z3','xl\_la\_z3','xt\_la\_z3','fv\_ba\_z3','vl\_ba\_z3','vt\_ba\_z3','xl\_ba\_z3','xt\_ba\_z3','fv\_bv\_z3','vl\_bv\_z3','vt\_bv\_z3','xl\_bv\_z3','xt\_bv\_z3','fv\_la\_z5','vl\_la\_z5','vt\_la\_z5','xl\_la\_z5','xt\_la\_z5','fv\_ba\_z5','vl\_ba\_z5','vt\_ba\_z5','xl\_ba\_z5','xt\_ba\_z5','fv\_bv\_z5','vl\_bv\_z5','vt\_bv\_z5','xl\_bv\_z5','xt\_bv\_z5','fv\_la\_tc','vl\_la\_tc','vt\_la\_tc','xl\_la\_tc','xt\_la\_tc','fv\_ba\_tc','vl\_ba\_tc','vt\_ba\_tc','xl\_ba\_tc','xt\_ba\_tc','fv\_bv\_tc','vl\_bv\_tc','vt\_bv\_tc','xl\_bv\_tc','xt\_bv\_tc','fv\_la\_bo','vl\_la\_bo','vt\_la\_bo','xl\_la\_bo','xt\_la\_bo','fv\_ba\_bo','vl\_ba\_bo','vt\_ba\_bo','xl\_ba\_bo','xt\_ba\_bo','fv\_bv\_bo','vl\_bv\_bo','vt\_bv\_bo','xl\_bv\_bo','xt\_bv\_bo','fullval\_la','fullval\_ba','fullval\_bv','avland\_la','avland\_ba','avland\_bv','avtot\_la','avtot\_ba','avtot\_bv','exland\_la','exland\_ba','exland\_bv','extot\_la','extot\_ba','extot\_bv']]

master\_expert\_var.dtypes

master\_expert\_var.to\_csv('Master Expert Variables 75.csv', encoding = 'utf-8')

zscale\_1 = (master\_expert\_var - master\_expert\_var.mean()) / master\_expert\_var.std()

%%time

#myvars\_zscale\_1.to\_csv('zscale\_1.csv', encoding='utf-8')

zscale\_1.dtypes

%%time

zscale\_1\_transpose = zscale\_1.transpose()

zscale\_1\_transpose.head()

zscale\_1.describe()

%%time

covariance = zscale\_1\_transpose.dot(zscale\_1)/(len(zscale\_1)-1)

print(covariance)

print(sns.heatmap(covariance))

%%time

pca = PCA(n\_components=75).fit(zscale\_1)

plt.plot(np.cumsum(pca.explained\_variance\_ratio\_))

plt.xlabel('Number of Principal Components')

plt.ylabel('Cumulative Variance Explained')

plt.grid()

plt.show()

plt.xlim(xmax=15)

plt.grid()

plt.ylim(ymax=1)

plt.show()

pca\_plot = plt.plot(np.cumsum(pca.explained\_variance\_ratio\_))

pca\_plot[0].get\_xydata()

#8 PCs explain 93.385404% of variability

PCs = pd.DataFrame(PCA(n\_components=8).fit\_transform(zscale\_1))

PCs.shape

PCs.to\_csv('Principal Components 8 cols.csv', encoding = 'utf-8')

zscale\_2 = (PCs - PCs.mean())/ PCs.std()

zscale\_2.to\_csv('PCA Z-scaled 2.csv', encoding = 'utf-8')

#For reboot

#zscale\_2 = pd.read\_csv('PCA Z-scaled 2.csv')

zscale\_2.dtypes

zscale\_2.describe()

zscale\_2\_wi = zscale\_2[['0','1','2','3','4','5','6','7']] #without index

zscale\_2\_wi.describe()

numrecords = len(zscale\_2\_wi)

Scores = pd.DataFrame(np.ones(numrecords), columns = ['Score\_1'])

Scores['Score\_2'] = np.ones(numrecords)

Scores['Score\_1'] = zscale\_2\_wi.abs().sum(axis=1)

zscale\_2\_wi\_sq = zscale\_2\_wi\*\*2

Scores['Score\_2'] = zscale\_2\_wi\_sq.abs().sum(axis=1)

Scores.to\_csv('Fraud Scores.csv', encoding = 'utf-8')

Scores.head(20)

original\_data = pd.read\_csv('NY property 1 million.csv')

scored\_original\_data = pd.concat([original\_data, Scores], axis = 1)

scored\_original\_data.to\_csv('Scored NY Property data 1 million.csv', encoding = 'utf-8')

xhigh = 2

plt.xlim(0,xhigh)

temp = Scores[Scores['Score\_1'] <= xhigh]

sns.distplot(temp['Score\_1'], bins = 100, kde = False)

xhigh = 0.1

plt.xlim(0,xhigh)

temp = Scores[Scores['Score\_2'] <= xhigh]

sns.distplot(temp['Score\_2'], bins = 100, kde = False)

xhigh = 5

plt.xlim(0,xhigh)

temp = Scores[Scores['Score\_1'] <= xhigh]

ax = sns.distplot(temp['Score\_1'], bins = 100, kde = False)

ax.set\_yscale('log')

xhigh = 5

plt.xlim(0,xhigh)

temp = Scores[Scores['Score\_2'] <= xhigh]

ax = sns.distplot(temp['Score\_2'], bins = 100, kde = False)

ax.set\_yscale('log')