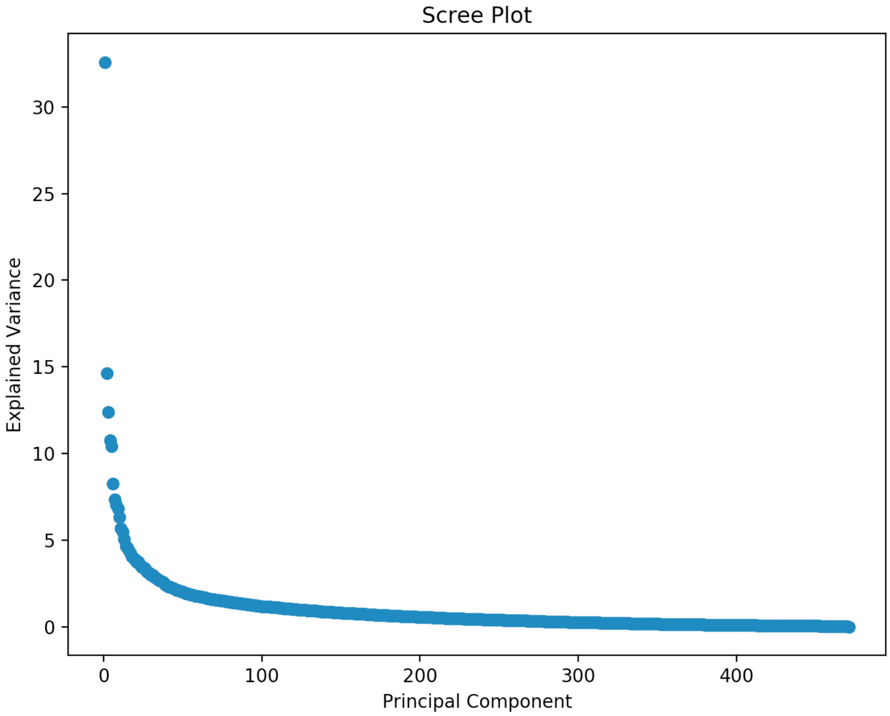
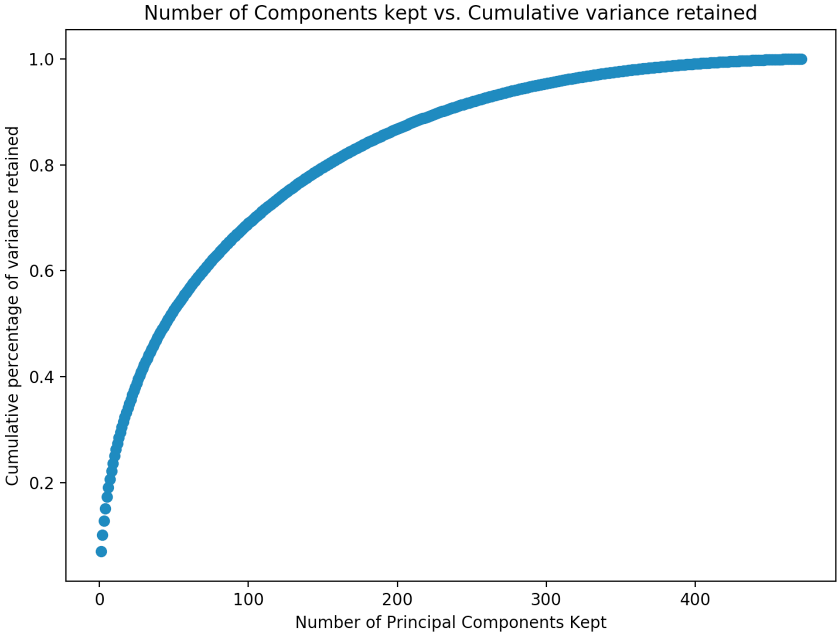
16791: Hw 3

Q1a

1)



2)

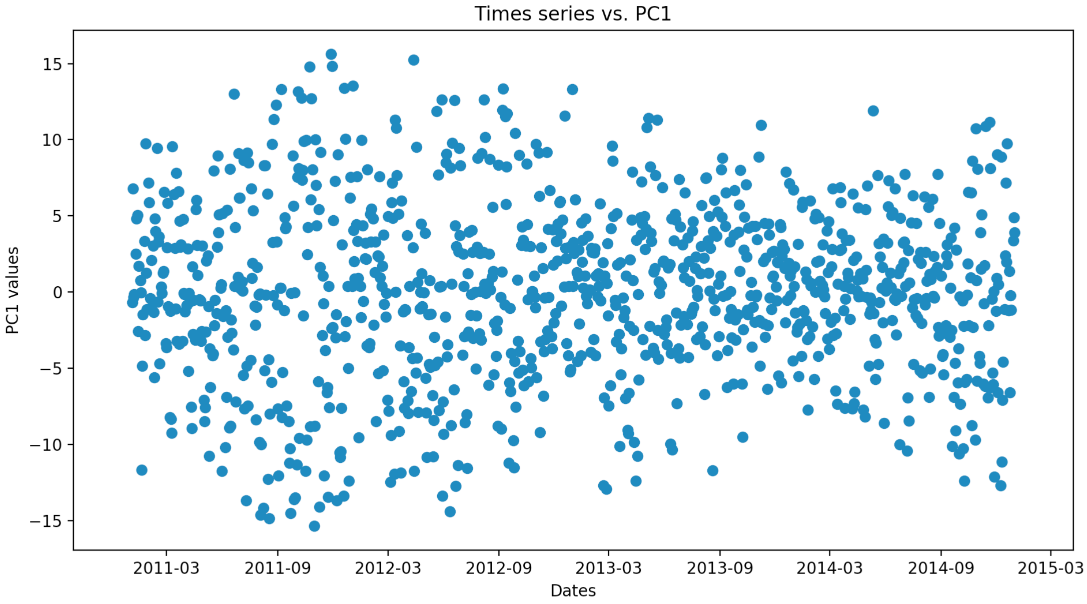


3) 153 components must be retained in order to capture at least 80% of the total variance

4) Only 10.15% of the total variance is kept if only the top 2 PCA components were retained, which means 89.8% of the total variance will be discarded. The total variance is 465, so the reconstruction error would be 89.85%*465 = 417.8.

Q1b

1)

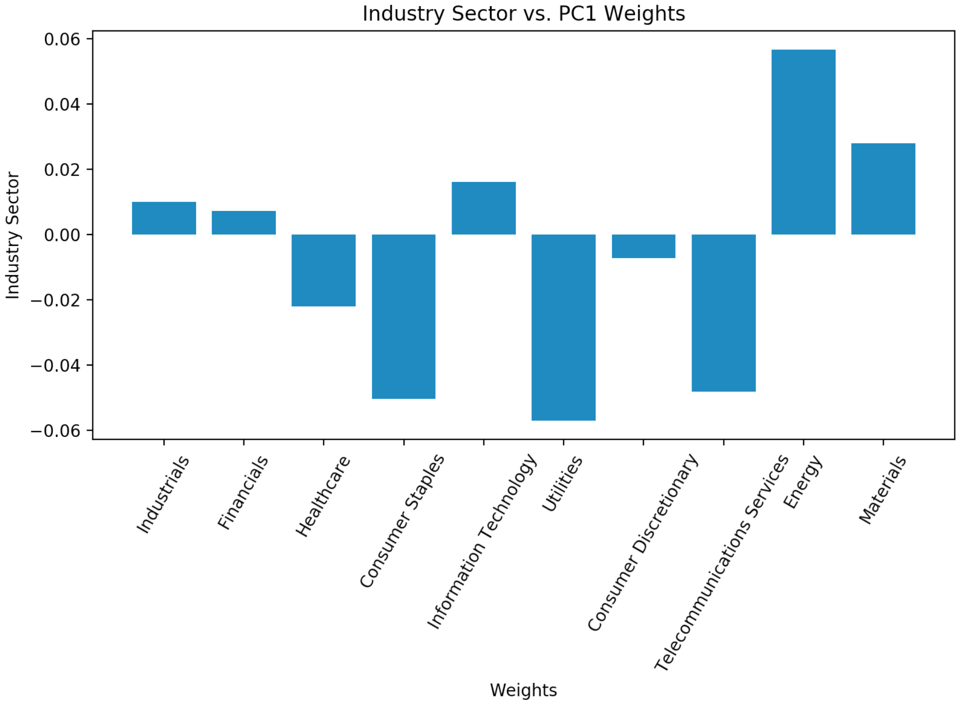


Changes in PC1 value seem to vary the most between 2011 to 2012.

The date with the lowest value for this component is 2011-10-31. On that day there was a snowstorm that smashed records in Northeastern states. More than three million customers would find themselves without power and with the prospect of enduring several more days without it. Towns were buried in dense snowfalls, closing down streets. Less people probably went to stores for shopping.

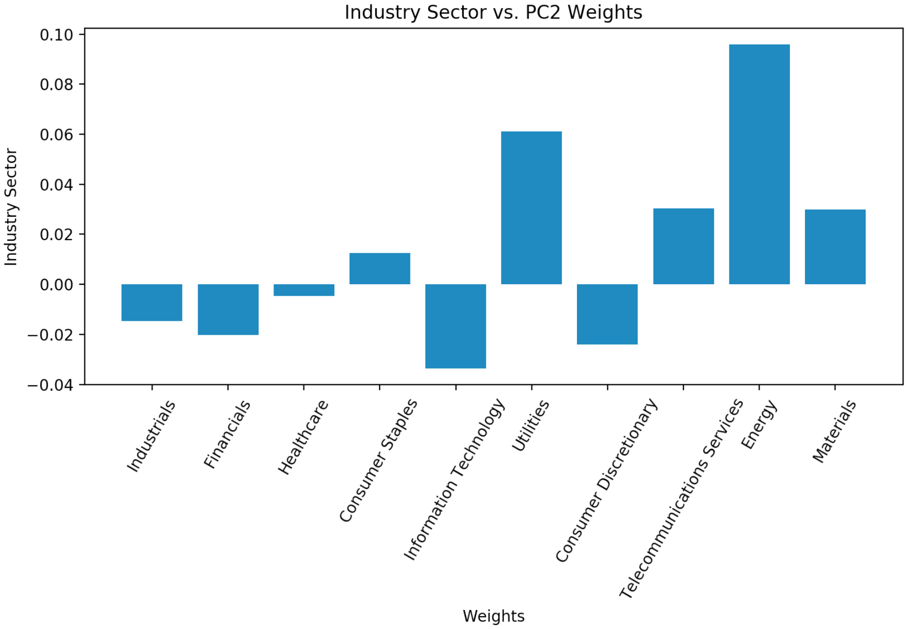
2) Extracted weights from pca.components_

3)



This could be showing the overall trends for each sector of stock prices, such that energy and materials sectors tend to have a rise in stock prices and healthcare, consumer staples, utilities, and telecommunication services tend to have a fall in stock prices.

4)



This plot distinguishes utilities, telecommunication services, energy, and materials from the rest, perhaps suggesting some common relationship between them. Perhaps all of these have the greatest changes in stock prices in terms of overall magnitude.

5) We could use PC1 to track overall market tendencies, since PC1 seems less biased towards specific subgroups of sectors. PC1 also, by definition, tends to encompass most of the explained variances since PCs are sorted based on eigenvalues.

Q2

a) Read documentation

b) See code

The top three features with the greatest correlation coefficients are Abdomen (corr=0.813), Chest (corr= 0.7026), and Hip (corr= 0.6250).

c)
Using univariate feature selection…

- Best set of features for size 1: Abdomen
- Best set of features for size 2: Chest and Abdomen
- Best set of features for size 3: Chest, Abdomen, and Hip

An exhaustive subset search searches for the best subsets of the variables for predicting a target variable in linear regression, using an efficient branch-and-bound algorithm. Univariate feature selection works by selecting the best features based on univariate statistical tests. In other words, univariate feature selection examines each feature individually to determine the strength of the relationship of the feature with the response variable. However, univariate feature selection does not consider the interaction between features, which the exhaustive subset search implicitly accounts. Thus, the exhaustive subset search, with appropriate limits, is more likely to derive the optimal set of features.

d)

Neck , Abdomen, and Wrist were included in model using 10-fold cross-validation RFECV in python.