Big Data Analytics – Homework 2

Trisha P Malhotra

Considering data (places\_rated.data) from the Places Rated Almanac, by Richard Boyer and David Savageau:

Form a data matrix X whose dimensions are NxP, where N are the number of cities (329) and P are the number of features (9).

Normalize X by transforming each feature column to zero mean and unit standard deviation.

Perform PCA and calculate explained variance ratios, loading vectors etc.

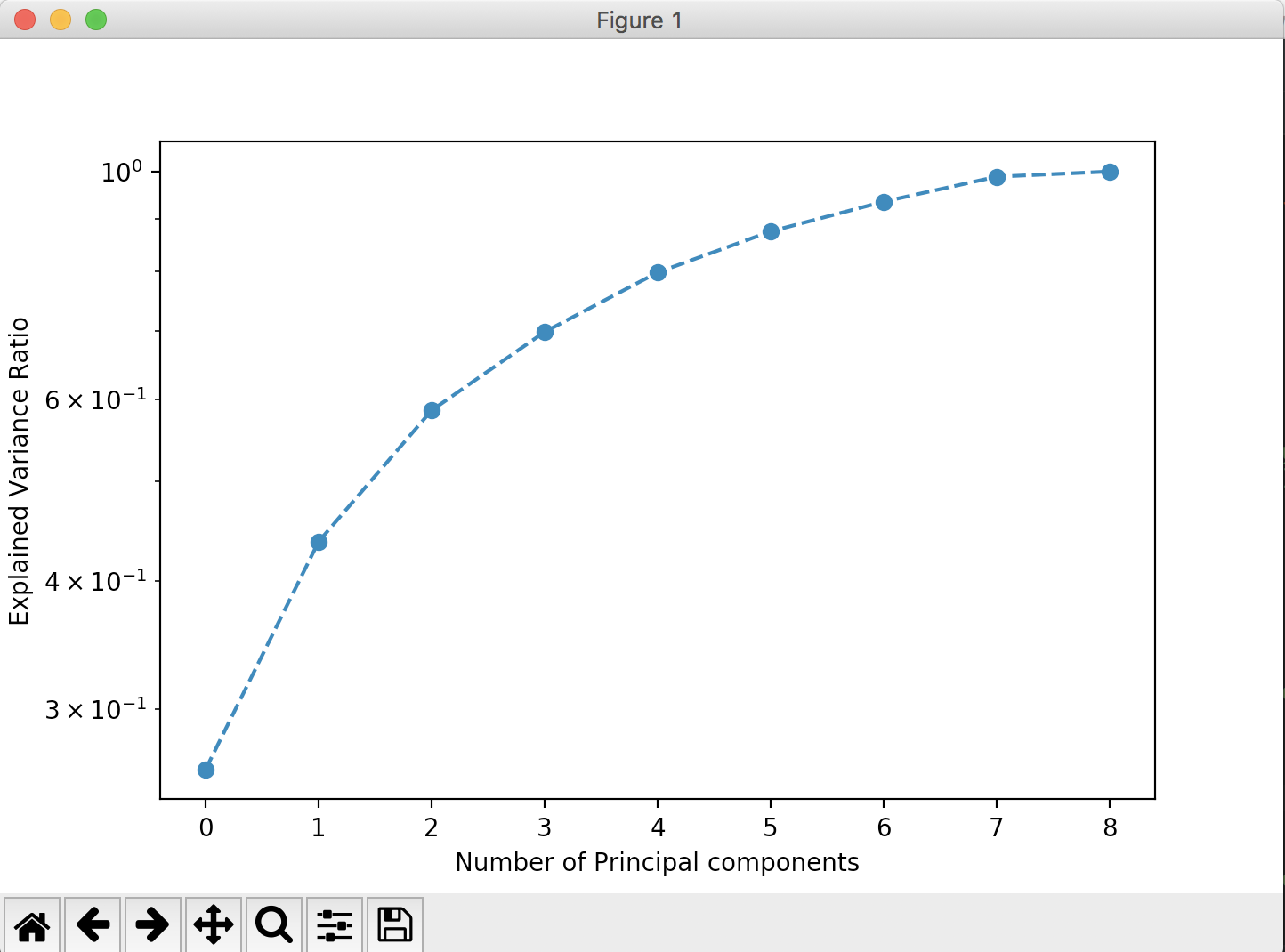
Ans: Answer in program: bda-hw2.py

Now answer the following questions

1. Plot explained variance ratio as a function of number of principal components. What are minimum number of principal components needed to explain at least 80% of the variance in the data.

Ans:

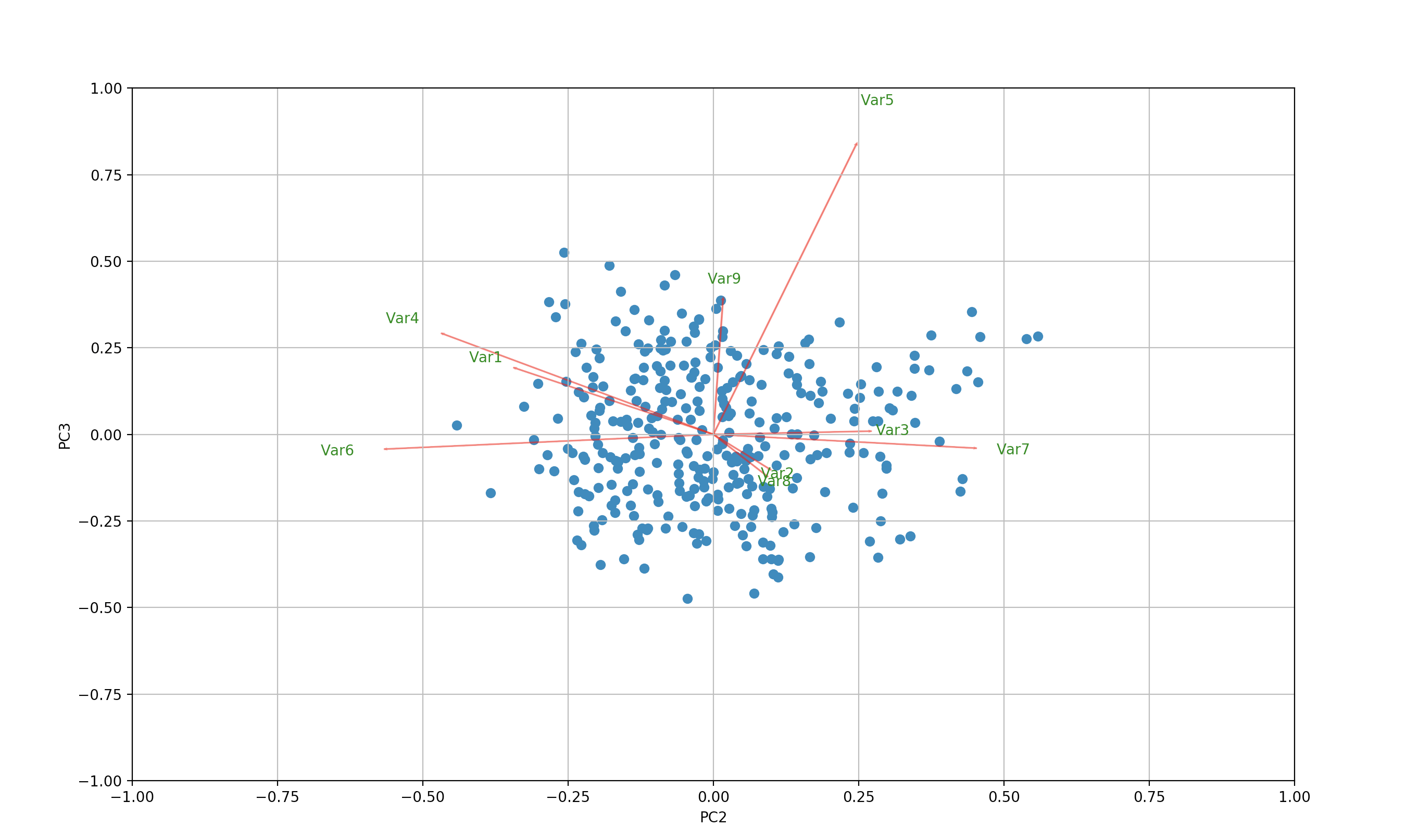
* As per the plot below:
* The minimum number of Principal Components to obtain at least 80% of the variance in data is: **6**, which gives **87.45**63319283 % to be precise.
* Having **5** principal components gives: **79.78**35901642 % of the total variance.

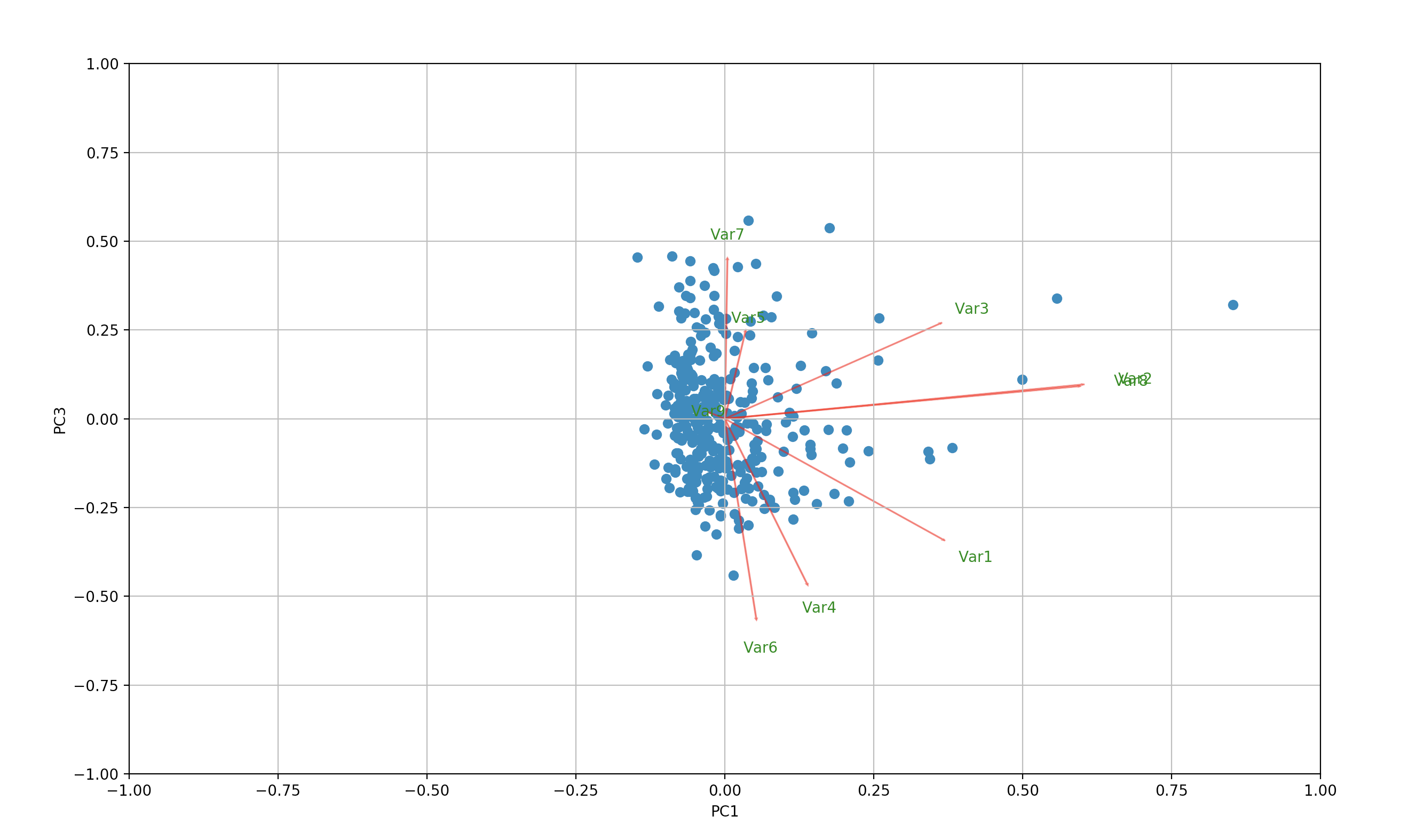


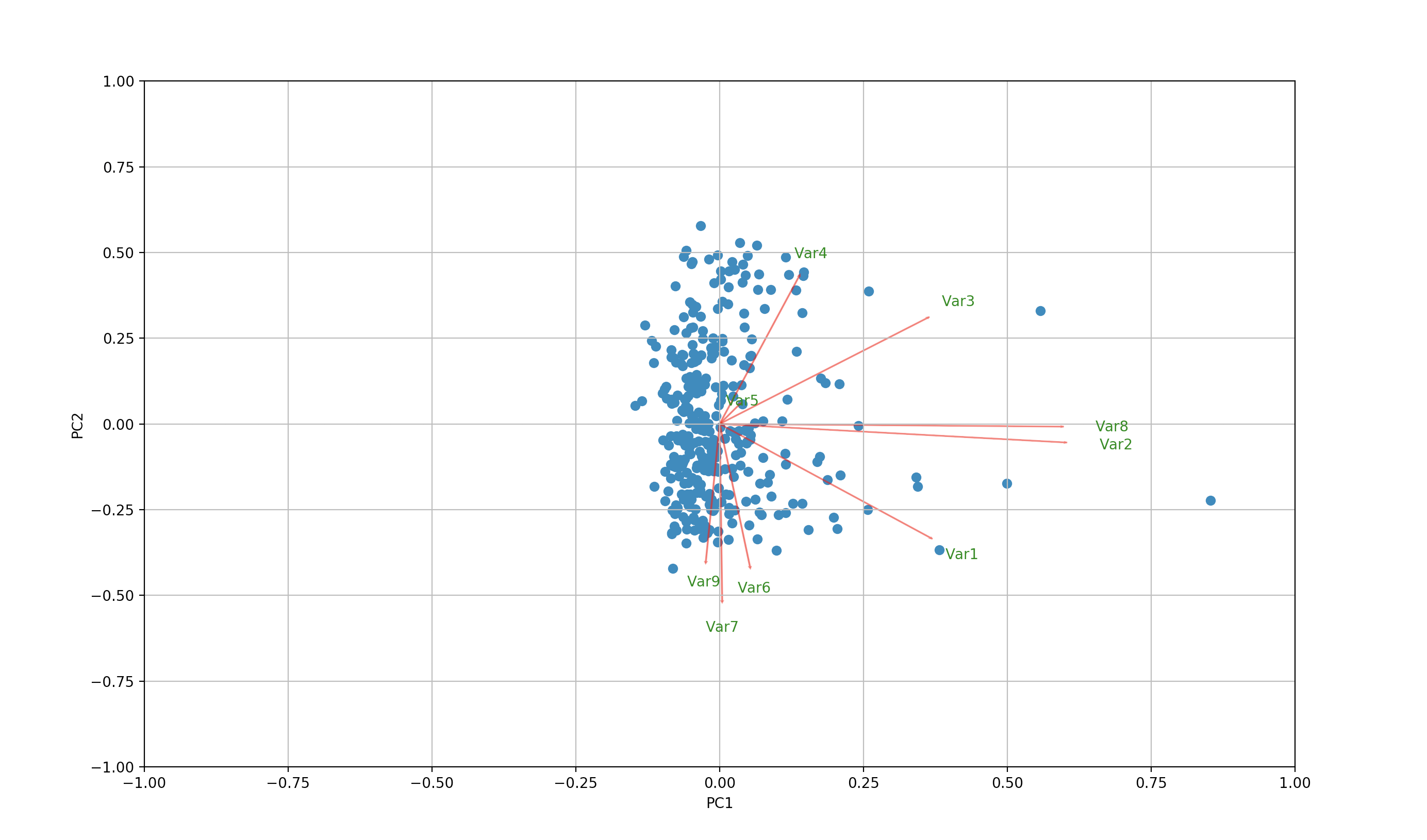
2. List the loading vectors for the first 3 principal components and interpret them.

3. Transform the original data into the principal component space. Plot the transformed data in PCA1-PCA2, PCA1-PCA3 and PCA2-PCA3 space, i.e. a biplot like we discussed in class. Plot the attribute axes on the plot (note the plot may get a bit messy if you are using full city names in the plot.)

Ans.





Screen%20Shot%202018-02-11%20at%2011.18.26%20PM.png

4. Discuss the plots. Identify any unusual cities (i.e. outliers in these plots)

=========================================X===================================