CSCI 720 Big Data Analytics
Assignment 2

We will consider data (places_rated.data) from the Places Rated Almanac, by Richard Boyer and David Savageau, copyrighted and published by Rand McNally. The nine rating criteria used by Places Rated Almanac are:

- Climate & Terrain
- Housing
- Health Care & Environment
- Crime
- Transportation
- Education
- The Arts
- Recreation
- Economics

For all but two of the above criteria, the higher the score, the better. For Housing and Crime, the lower the score the better. The scores are computed using the following component statistics for each criterion (see the Places Rated Almanac for details):

- **Climate & Terrain**: very hot and very cold months, seasonal temperature variation, heating- and cooling-degree days, freezing days, zero-degree days, ninety-degree days.
- **Housing**: utility bills, property taxes, mortgage payments.
- **Health Care & Environment**: per capita physicians, teaching hospitals, medical schools, cardiac rehabilitation centers, comprehensive cancer treatment centers, hospices, insurance/hospitalization costs index, flouridation of drinking water, air pollution.
- **Crime**: violent crime rate, property crime rate.
- **Transportation**: daily commute, public transportation, Interstate highways, air service, passenger rail service.
- **Education**: pupil/teacher ratio in the public K-12 system, effort index in K-12, accademic options in higher education.
- **The Arts**: museums, fine arts and public radio stations, public television stations, universities offering a degree or degrees in the arts, symphony orchestras, theatres, opera companies, dance companies, public libraries.
- **Recreation**: good restaurants, public golf courses, certified lanes for tenpin bowling, movie theatres, zoos, aquariums, family theme parks, sanctioned automobile race tracks, pari-mutuel betting attractions, major- and minor- league professional sports teams, NCAA Division I football and basketball teams, miles of ocean or Great Lakes coastline, inland water, national forests, national parks, or national wildlife refuges, Consolidated Metropolitan Statistical Area access.
- **Economics**: average household income adjusted for taxes and living costs, income growth, job growth.

In addition latitude and longitude, population and state and case number are also given, but we will ignore these and only consider the 9 features mentioned above.

You will use principal components analysis to identify the major components of variation in the ratings amongst cities. You will write some code to do the following

- 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. You may use scikit-learn's PCA functions (sklearn.decomposition.PCA).

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

Include any code with your submission. If you are using a Jupyter notebook, include your notebook file as well as a pdf of your notebook with plots etc..