Due date: Oct. 17, 2018, 11:59 PM (Arlington time). You have **two** late days — use it at as you wish. Once you run out of this quota, the penalty for late submission will be applied. You can either use your late days quota (or let the penalty be applied). Clearly indicate in your submission if you seek to use the quota.

What to turn in:

1. Your submission should include your complete code base in an archive file (`zip, tar.gz`) and `q1/, q2/,` and so on), and a very very clear README describing how to run it.
2. A brief report (typed up, submit as a PDF file, NO handwritten scanned copies) describing what you solved and implemented and known failure cases.
3. Submit your entire code and report to Blackboard.

Notes from instructor:

- Start early!
- You may ask the TA or instructor for suggestions, and discuss the problem with others (minimally). But **all parts of the submitted code must be your own**.
- Use Matlab or Python for your implementation.
- Make sure that the TA can easily run the code by plugging in our test data

Problem 1

(k-means, **55pts**) The dataset provided for this problem `NBAstats.csv` is stats from NBA players. The players are indexed by their names, and they are labeled by 5 different positions: {center (C), power forward (PF), small forward (SF), shooting guard (SG), point guard (PG)} and there are 27 attributes, e.g., age, team, games, games started, minutes played and so on (that makes total of 29 columns in the data matrix). Make sure that you standardize the data (zero-mean and standard deviation = 1) before you analyze the data.

1. **(30pts)** Write a function `cluster = mykmeans(X, k)` that clusters data $X \in \mathbb{R}^{n \times p}$ (n number of objects and p number of attributes) into k clusters.
2. **(10pts)** For this problem, use all features except team. Use **your code** to group the players into $k = \{3, 5\}$ clusters. Report the centers found for each clusters for each k, distribution of positions in each cluster and your brief observation.
3. **(5pts)** Some of the attributes are perhaps redundant in terms of Linear Algebra. Report which ones are redundant and explain why.
4. **(10pts)** For this problem, use the following set of attributes {2P%, 3P%, FT%, TRB, AST, STL, BLK} to perform k-means clustering with $k = \{3, 5\}$. Report the centers found for each clusters for each k, distribution of positions in each cluster and your brief observation.

Problem 2

(k-NN, **45pts**) The dataset provided for this problem `NBAstats.csv` is stats from 475 NBA players. The players are labeled by 5 different positions: {center (C), power forward (PF), small forward (SF), shooting guard (SG), point guard (PG)} and there are 27 attributes, e.g., age, team, games, games started, minutes played and so on. Use the first 375 players as training data and remaining 100 players as testing data. Make sure that you standardize the data (zero-mean and standard deviation = 1) before you analyze the data.

1. **(25pts)** Write a function `class = myknn(X, test, k)` that performs k-nearest neighbor (k-NN) classification where $X \in \mathbb{R}^{n \times p}$ (n number of objects and p number of attributes) is training data, `test` is testing data, and k is a user parameter.

2. **(10pts)** For this problem, use all features except team. Use your k-NN code to perform classification. Set $k = \{1, 5, 10, 30\}$ and report their accuracies and your observation.

3. **(15pts)** For this problem, use the following set of attributes {2P%, 3P%, FT%, TRB, AST, STL, BLK} to perform k-NN classification with $k = \{1, 5, 10, 30\}$. Report accuracies for each k and your observation.