HOMEWORK 1
COMPUTATIONAL METHODS FOR DATA SCIENCE
FALL SEMESTER 2022

1. **Basic Probability in Blackjack.** *(15 points)* Among all casino games, blackjack is probably the only mathematically beatable game. One key reason is that most casino games involve in independent trials, i.e. the current trial has no impact whatsoever on the probabilities of occurrence of the rest of the trials, but in blackjack, we deal with dependent trials.
The basic rule of blackjack is as follows. We consider a single deck with 52 poker cards. Among them, all number cards count their face values, all Jack, Queen and King count as 10, and Ace can count as 1 or 11. When a game starts, you first get two cards face-up, then the dealer gets one card face-up and one card face-down. Your goal is to get the total number as close as 21 but not exceed 21 (otherwise we call it bust). Once you decide to hold, the dealer is mandatory to take card if his total number is less than 17, and he stops taking card once it exceeds 17. If the dealer busts (the total number exceeds 21), you win. If not, then a comparison between your total number and dealer's decides the winner. The one who is closer to 21 wins.

   (a) A situation is called "Blackjack" if your two face-up cards has a total of 21 immediately. Given a single deck, calculate the probability that you get a blackjack.

   (b) Assume you receive one K and one 3 cards (a total number 13). Before the dealer receives his two cards, what is the probability that you will bust if you take another card?

   (c) Now the dealer receives his two cards. Given the dealer must take a card when < 17 and must hold when ≥ 17. You decide to take a card if your bust probability is smaller than dealer bust probability. Among all 10 possible face-up cards (A,2,...,10/J/Q/K) of the dealer, please explain when you will take a card if the face-up cards appear.

2. **Basic Statistics in Roulette.** *(10 points)* Roulette is a classic casino game to guess which number the ball is landed on. The player will place their bet on a game table with payout ratios and they will receive a return once they win. In this question, we assume the game table is exactly the same as the one in Lecture 01-1. Here is the payout ratio:

	Payout	Bet Amount
Red or Black	1-to-1	a multiple of 8
Odd or Even	1-to-1	a multiple of 8
1 to 18 or 19 to 36	1-to-1	a multiple of 8
Dozen (1 to 12,...)	2-to-1	a multiple of 4
Column (on the right)	2-to-1	a multiple of 4
Single Number	35-to-1	a multiple of 1

   (a) Calculate the expected value and the variance for one betting on a dozen.

   (b) Together with the results in Lecture Note 01-1, you should have all information on the expected value and variance on the game table. Assume you bet 100 dollars (the maximum bet amount per round), and the casino has a special rule on the bet amount listed in the table. Please state the strategies how you bet your 100 dollars (1) to maximize your expected value; (2) to minimize your variance.

(c) Write a simple roulette simulation and demonstrate your average winning per bet using your two strategies.

(d) Some guy called Steve says that he can beat roulette game. All he does is to keep betting 8 dollars on black, then double every time he loses. Basically he cannot lose. Please explain why it does not work in real world.

3. **Monthly Record-Breaking Temperature in California I: Matrix Calculation** *(25 points)* Climate.gov records climate statistics at individual stations in USA. In this problem, we analyze a subset of the original full data that focuses on the monthly record-breaking high temperature of California. The file CAmaxTemp.txt can be downloaded from the following link:
http://staff.stat.sinica.edu.tw/fredphoa/HW/HW1/CAmaxTemp.txt
In this dataset, the first column is the station name, the second column is the investigation period, and the last column is the yearly high-temperature record. The remaining 12 columns are the monthly high-temperature records during the investigation period that we consider in this problem, and we denote this 12×12 matrix as X.
NOTE: Please include your own iterative code how you obtain your results. DO NOT copy and paste any library or function from existing programs.

   (a) Run the LU factorization on X to obtain a lower triangular matrix L and an upper triangular matrix U.

   (b) Use Gram-Schmidt algorithm on X to obtain an orthogonal matrix Q and an upper triangular matrix R. Find the inverse of X.

   (c) Use Power Iteration method to find the largest eigenvalue-eigenvector pair of X.

   (d) Use QR factorization to find all eigenvectors with REAL eigenvalues of X. There should be 6 if you calculate correctly. (Hint: If the true eigenvalue is complex in nature, the value obtained via QR factorization does not converge).

4. **Monthly Record-Breaking Temperature in California II: PCA and SVD** *(30 points)* We continue to use the same 12×12 dataset in this problem. NOTE: Please include your own iterative code how you obtain your results. DO NOT copy and paste any library or function from existing programs.

   (a) Standardize the data and compute the variance-covariance matrix.

   (b) Find the top three principal components using power iteration. Calculate the cumulative percentage of the total eigenvalues that these three principal components cover.

   (c) Plot the data on a 3D space with three principal component axes. Provide the coordinates of the recast data.

   (d) Find all principal components with their eigenvalues using SVD.

   (e) SVD provides an extra information on U that PCA does not usually have. Is it any interpretation of this U matrix? If yes, please state it.

   (f) Conduct a rank-3 approximation (SVD version) of X.

5. **Monthly Record-Breaking Temperature in California III: ICA** *(20 points)* We continue to use the same 12×12 dataset in this problem, but we only consider the data of February, June and October, i.e. a 12×3 subdata X'. NOTE: Please include your own iterative code how you obtain your results. DO NOT copy and paste any library or function from existing programs.

(a) Explain why the monthly data is unlikely to be Gaussian.

(b) Run the three preprocessing steps of ICA on X'.

(c) Provide a graphical illustration on the transformation, like the one in Lecture 04-2.

(d) Run the Fast ICA on Kurtosis Maximization to find the three independent components.

6. **Determinant and Parallogram.** *(Bonus 10 points)* In Lecture 02-1, we know that a determinant is equal to the area of the parallelogram. Please prove it.