Question 1 (Total 20 marks)

The following set of files contain information on the passes made during two of the matches during the FIFA 2018 world cup:

1. `BRA-BEL-FIFA-2018.xlsx`: the Brazil-Belgium quarter-final.
2. `FRA-ARG-FIFA-2018.xlsx`: the France-Argentina second round tie.

Each sheet denotes the number of passes made from one player to another. For instance, cell E2 in the "BRA" sheet of `BRA-BEL-FIFA-2018.xlsx` displays the value 2. This indicates that during the game, 2 passes were made from Alisson (jersey number 1) to Thiago Silva (jersey number 2). Similarly, cell F3 indicates that 10 passes were made from Thiago Silva to Miranda (jersey 3). *Note that the pass matrix is not symmetric.*

Part I (10 marks)

Let us focus on the Brazil team, from the Brazil-Belgium game. Our goal is to write a function that simulates a sequence of passes (of a given length), starting from a particular player. In order to do this, treat each row in the pass matrix as an un-normalised probability mass function and `sample( )` from that row. In other words, if you divide a row by the total number of passes that that player made in the game, you will obtain the proportion of times that player would pass to a particular team-mate. This "pmf" can be sampled from.

The function should be called `rpass_seq`, and it should have the following signature:

```
rpass_seq(init_player, n_passes, pass_matrix)
```

- `init_player` is the jersey number of the player who starts with the ball.
- `n_passes` is the length of the pass sequence (the number of passes to simulate).
- `pass_matrix` is the pass matrix of the particular team.

You may decide on the class and type and format of the above arguments; but they should all be there, with those names. The output of the function call should be a vector of jersey numbers, showing the sequence of passes. For instance:

```
rpass_seq(init_player = "1", n_passes= 5, bra_mat)
```

```
## [1] "1"  "2"  "10" "12" "9"  "17"
```

This says the sequence of passes was

$$1 \to 2 \to 10 \to 12 \to 9 \to 17$$

Run your function 1000 times for pass sequences of length 5, starting from Alisson, and create a bar chart showing the proportion of times the sequence ended with a particular player.

Part II (7 marks)

An analyst on the team comes to you and says that she would like to condense the pass matrix to one that contains only positional information rather than specific players. Convert

each of the four pass matrices such that they show this information. You will need the following information:

```
# gk:  goalkeeper, def: defender
# mid: midfielder, fwd: forward
bra_pos <- list(gk=1, def=c(2,3,12,22), mid=c(11, 15,17,19),
                fwd=c(9, 10))
bel_pos <- list(gk=1, def=c(2,4,5,15), mid=c(6,7,8,22),
                fwd=c(9, 10))
fra_pos <- list(gk=1, def=c(2,4,5,21), mid=c(6, 13, 14),
                fwd=c(7, 9, 10))
arg_pos <- list(gk=12, def=c(2,3,14,16,17), mid=c(7,11,15,22),
                fwd=c(10))
```

Here is what the converted Argentina matrix looks like:

```
##     gk def mid fwd
## gk   0  22   4   0
## def 10 104  96  20
## mid  2  63  29  18
## fwd  0  19  13   0
```

Store your four matrices in a list called `q1_matrices`.

Part III (3 marks)

The following wikipedia page contains information about the 2018 world cup.

https://en.wikipedia.org/wiki/2018_FIFA_World_Cup

Write a few lines of code that will extract the following table from there. Store the dataframe in `q1_money`.

Position	Amount (million USD)	Amount (million USD)
Position	Per team	Total
Champions	38	38
Runners-up	28	28
Third place	24	24
Fourth place	22	22
5th–8th place (quarter-finals)	16	64
9th–16th place (round of 16)	12	96
17th–32nd place (group stage)	8	128
Total	Total	400

Question 2 (Total 15 marks)

The file `CDNOW2.csv` contains information on a cohort of customers who purchased compact discs (music records) from an online store. This cohort of customers all made their first purchase in the first 12 weeks of 1994. This cohort of 23,570 customers were tracked until June of the following year - a sum total of 78 weeks. Read the data into R as `cdnow`.

Each row in the dataset is a transaction made by a customer, i.e. a purchase from the store. The columns in the dataset are:

- `cid`: the customer id.
- `week_num`: the week at which this transaction was performed (from 1 – 78).
- `qty`: The number of CDs bought during this transaction.
- `amt`: The amount of money spent in this transaction.

Here are your tasks:

1. Find the customers who purchased CDs in at least two consecutive weeks, and keep only those customers **and only those particular rows**. Store your tibble as `q2_consec`.
2. Week number 1 corresponds to the first Monday of 1993. The remaining week numbers correspond to subsequent Mondays. Add a column of class Date, that contains the date corresponding to each week number.
3. The plot below displays the total quantity of CDs bought in each of the 78 weeks. Add separate 5-week moving average lines for the first 12 weeks and subsequent 66 weeks.
4. Now suppose that a new hypothetical customer has made $X = 5$ purchases. The company would like to know how the chances of him making a 6th purchase change as the weeks progress. Interpret (or refine) this question as best you can, and use the data to provide the company with some insight. Please explain your findings clearly, and include a plot if you can.