ITA 04 – Assignment – Day 3

1. Consider the data set **occupationalStatus** in the datasets package.

(a) What is the probability of a son having the same occupational status as his father? [Hint: investigate what diag(x) does if x is a matrix.]

```
library(datasets)
data(occupationalStatus)
transition_mat <- as.matrix(occupationalStatus) / colSums(occupationalStatus)
prob_same_status <- sum(diag(transition_mat))
prob_same_status
```

b)Renormalize the data so that each row sums to 1. In the new data set the ith row represents the conditional distribution of a son's occupational status given that his father has occupational status i.

```
 renorm_data <- occupationalStatus / rowSums(occupationalStatus)

renorm_data
```

c) What is the probability that a son has occupational status between 1 and 3, given that his father has status 1?

What if the father has occupational status 8?

```
status 1

prob_1_to_3_given_1 <- sum(renorm_data[1, 1:3]

prob_1_to_3_given_1
```

2. Create the following data frame, subsequently invert Gender for all individuals.

a) Name Age Height Weight Gender

Name	Age	Height	Weight	Gender
Alex	25	177	57	M
Lilly	31	163	69	M
Mark	23	190	83	F

b) Create the below data frame

Name	Working
Alex	Yes
Lilly	No
Mark	No

c) Add the data frame column-wise to the previous one.

How many rows and columns does the new data frame have?

sol:

```
import pandas as pd

df1 = pd.DataFrame({

    'Name': ['Alex', 'Lilly', 'Mark'],

    'Age': [25, 31, 23],

    'Height': [177, 163, 190],

    'Weight': [57, 69, 83],

    'Gender': ['M', 'M', 'F']

})

df1['Gender'] = df1['Gender'].apply(lambda x: 'F' if x == 'M' else 'M')
```

3. A student recorded his/her scores on weekly R programming quizzes that were marked out of a possible 10 points. His/Herscores were as follows:

8, 5, 8, 5, 7, 6, 7, 7, 5, 7, 5, 5, 6, 6, 9, 8, 9, 7, 9, 9, 6, 8, 6, 6, 7
What is the mode of his/her scores on the weekly R programming quizzes?

sol:

the mode of a dataset is the value that appears most fequently,5 and 7 both appears 5 times.

5 and 7 are mode

4. Construct the following data frame.

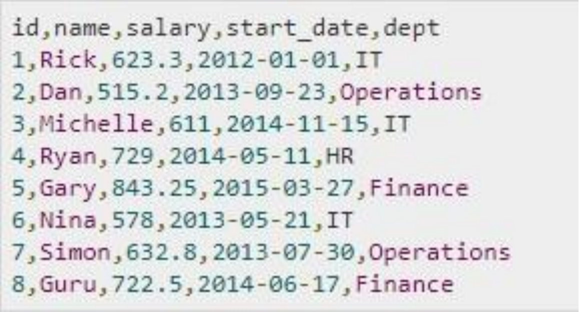
Countries	population_in_million	gdp_per_capita		
A	100	2000		
B	200	7000	C 120	15000

a) Write appropriate R code and reshape the above data frame from wide data format to long data format.
b) Write R code and reshape from long to wide data format.

sol:

```
library(tidyr)

df <- data.frame(Countries = c("A", "B", "C"),

                 population_in_million = c(100, 200, 120),

                 gdp_per_capita = c(2000, 7000, 15000))

df_long <- gather(df, key = "variable", value = "value", -Countries)
```

5. Consider the following data present. Create this file using windows notepad . Save the file as **input.csv** using the save As All files(*.*) option in notepad.



```
id,name,salary,start_date,dept
1,Rick,623.3,2012-01-01,IT
2,Dan,515.2,2013-09-23,Operations
3,Michelle,611,2014-11-15,IT
4,Ryan,729,2014-05-11,HR
5,Gary,843.25,2015-03-27,Finance
6,Nina,578,2013-05-21,IT
7,Simon,632.8,2013-07-30,Operations
8,Guru,722.5,2014-06-17,Finance
```

i. Use appropriate R commands to read input.csv file.

ii. Analyze the CSV File and compute the following.

   a. Get the maximum salary
   b. Get the details of the person with max salary
   c. Get all the people working in IT department
   d. Get the persons in IT department whose salary is greater than 600
   e. Get the people who joined on or after 2014

iii. Get the people who joined on or after 2014 and write the output onto a file called output.csv