3_4_Solutions

October 15, 2018

```
In [38]: options(repr.matrix.max.cols=8, repr.matrix.max.rows=5)
```

```
In [28]: path<-"https://raw.githubusercontent.com/nmeraihi/data/master/"
```

1 Question 1

1.1 a)

Importer les données `qc_hommes_2.csv` à partir du répertoire data github dans un *data frame* df

```
In [29]: df<-read.csv(paste(path,"qc_hommes_2.csv",sep = ""), sep=",")
```

```
In [30]: head(df)
```

age	lx
0 an	100000
1 an	99501
2 ans	99483
3 ans	99467
4 ans	99454
5 ans	99442

```
In [31]: tail(df)
```

	age	lx
106	105 ans	96
107	106 ans	51
108	107 ans	26
109	108 ans	13
110	109 ans	6
111	110 ans et plus	3

1.2 b)

Dans la colonne age, garder seulement la partie numérique. Vous devriez alors obtenir age={0,1,2 ...}

```
In [153]: df$age<-gsub("ans", "", df$age)
          df$age<-gsub("an", "", df$age)
```

1.3 c)

À ce df, ajouter une nouvelle colonne `dx` (nombre de décès entre l'âge x et x+n). Donc dx est le nombre de décès qui surviennent dans chaque intervalle d'âge au sein d'une cohorte initiale de 100 000 naissances vivantes à l'âge 0.

$$d_x = l_x - l_{x+1}$$

```
In [154]: a<-df[-nrow(df), 2]-df[-1, 2]
          a<-c(a, a[length(a)])
          df$dx<-a
```

1.4 d)

Calculer qx (quotient de mortalité entre l'âge x et x+n). Donc qx est probabilité qu'un individu d'âge x décède avant d'atteindre l'âge x+n.

$$q_x = \frac{d_x}{l_x}$$

```
In [155]: df$qx<-round(df$dx/df$lx,5)
```

```
In [156]: head(df)
```

age	lx	dx	qx
0	100000	499	0.00499
1	99501	18	0.00018
2	99483	16	0.00016
3	99467	13	0.00013
4	99454	12	0.00012
5	99442	11	0.00011

1.5 e)

Maintenant que vous avez toutes les données, on peut calculer la probabilité qu'un individu d'âge x survive jusqu'à l'âge x+n.

$$tP_x = \frac{l_{x+t}}{l_x}$$

Calculer la probabilité qu'un individu de 22 ans survive les trois prochaines années

```
In [159]: age<-22
          t<-3
          p<-df[age+1+t,2]/df[age+1,2]
          p
```

0.998192831903079

```
In [160]: library(formattable)
```

```
In [161]: percent(p)
```

99.82%

2 Question 2

```
In [162]: Id=c(1,2,3,4)
          Age=c(14,12,15,10)
          Sex=c('F','M','M','F')
          Code=c('a','b','c','d')
          df1=data.frame(Id,Age)
          df2=data.frame(Id,Sex,Code)
```

Avec les données suivantes;

```
In [163]: df1
```

Id	Age
1	14
2	12
3	15
4	10

```
In [164]: df2
```

Id	Sex	Code
1	F	a
2	M	b
3	M	c
4	F	d

Créer un *data frame* `M` qui fait une jointure de `df1` et `df2`

```
In [165]: M=merge(df1,df2,by='Id')
          M
```

Id	Age	Sex	Code
1	14	F	a
2	12	M	b
3	15	M	c
4	10	F	d

3 Question 3

Selon un journaliste de la CNBC, le prix de l'action de Apple (AAPL) est très corrélé avec le prix de l'action de Boeing Co (BA).

Calculer la corrélation des prix Adj Close **mensuels** de ces deux compagnies sur la période allant du 2016-11-01 au 2017-10-01.

Indice: créer deux vecteur avec les valeurs des prix. Vous pouvez importer les données à partir de finance yahoo dans la section *Historical Data* avec les dates et périodes indiquées ci-haut.

```
In [2]: path<-"https://raw.githubusercontent.com/nmeraihi/data/master/"
```

```
In [170]: df_app <-read.csv(paste(path,"AAPL_month.csv",sep = ""), header = T)
```

```
In [176]: df_ba <-read.csv(paste(path,"BA_month.csv",sep = ""), header = T)

In [179]: a<-cbind(df_app$Adj.Close,df_ba$Adj.Close)

In [180]: colnames(a)<-c("Apple", "Boeing")

In [181]: rownames(a)<-seq(as.Date("2016/11/1"), by = "month", length.out = 12)

In [182]: cor(a)
```

	Apple	Boeing
Apple	1.000000	0.872264
Boeing	0.872264	1.000000

4 Question 4

Créer un *data frame* avec les données HackerRank-Developer-Survey. Dans ces données, sont une série de réponse que les développeurs de HackerRank ont répondu suite à un sondage ayant pour but de comprendre les l'intérêt des femmes envers l'informatique.

```
In [1]: library(dplyr, warn.conflicts = F)

In [3]: values <- read.csv(paste(path,"HackerRank-Developer-Survey-2018-Values.csv",sep = ""), h

In [4]: head(values)
```

RespondentID	StartDate	EndDate	CountryNumeric2	q1AgeBeginCoding	q2Age
6464453728	10/19/17 11:51	10/20/17 12:05	South Korea	16 - 20 years old	18 - 24 years
6478031510	10/26/17 6:18	10/26/17 7:49	Ukraine	16 - 20 years old	25 - 34 years
6464392829	10/19/17 10:44	10/19/17 10:56	Malaysia	11 - 15 years old	12 - 18 years
6481629912	10/27/17 1:51	10/27/17 2:05	Curaçao	11 - 15 years old	12 - 18 years
6488385057	10/31/17 11:46	10/31/17 11:59		16 - 20 years old	25 - 34 years
6463843138	10/19/17 3:02	10/19/17 3:18	United States	41 - 50 years old	35 - 44 years

4.1 a)

En utilisanrtt le package `dplyr`, faites un petit tableau qui donne la proportion des hommes et des femmes dans ce *dataset*.

Utilisez la variable `q3Gender`

```
In [34]: values_2<-values %>%
             group_by(q3Gender) %>%
                 filter(q3Gender %in% c('Male','Female'))%>%
                     count()

In [35]: values_2$n<-(values_2$n/ sum(values_2$n)) * 100
         values_2
```

q3Gender	n
Female	16.55688
Male	83.44312

4.2 b)

En utilisanrtt le package `dplyr`, faites un tableau qui donne la proportion des hommes et des femmes en les séparant par le fait qu'ils soient étudiants ou non.

Utilisez les variables `q3Gender`, `is_student` et `q8Student`

```
In [36]: values$is_student <- ifelse(values$q8Student == '','Yes','No')
```

```
In [37]: values %>% group_by(q3Gender, is_student) %>%
             filter(q3Gender %in% c('Male','Female')) %>%
             count() %>%
             ungroup() %>%
             group_by(is_student) %>%
             mutate(n = (n/ sum(n)) * 100)
```

q3Gender	is_student	n
Female	No	20.82685
Female	Yes	13.55364
Male	No	79.17315
Male	Yes	86.44636

4.3 c)

Dressez un tableau qui donne le nombre de répondants par pays (utilisez la variable `CountryNumeric2`)

```
In [13]: values %>% group_by(CountryNumeric2) %>% count() %>% head()
```

CountryNumeric2	n
	3991
Afghanistan	3
Albania	8
Algeria	22
American Samoa	1
Andorra	1

4.4 d)

Faites un tableau qui donne le nombre de répondants en les classant par le diplôme obtenu.

Utilisez la variable `q4Education`

```
In [26]: values %>%
                     filter(!is.na(q4Education))%>%
                     group_by(q4Education)%>%
                     summarise(Total = n())%>%
                     arrange(desc(Total))  %>%
                     mutate(q4Education = reorder(q4Education,Total)) %>%
                     head(10)
```

q4Education	Total
College graduate	12010
Post graduate degree (Masters, PhD)	6030
Some college	2499
Some post graduate work (Masters, PhD)	2493
High school graduate	1289
Some high school	316
#NULL!	305
Vocational training (like bootcamp)	148

4.5 e)

Faites un tableau qui donne le nombre de développeurs par catégorie d'âge.
Utilisez la variable `q1AgeBeginCoding`

```
In [27]: values %>%
                    filter(!is.na(q1AgeBeginCoding)) %>%
                    group_by(q1AgeBeginCoding) %>%
                    summarise(Total = n()) %>%
                    arrange(desc(Total)) %>%
                    mutate(q1AgeBeginCoding = reorder(q1AgeBeginCoding,Total)) %>%
                    head(10)
```

q1AgeBeginCoding	Total
16 - 20 years old	14293
11 - 15 years old	5264
21 - 25 years old	3626
5 - 10 years old	933
26 - 30 years old	642
31 - 35 years old	193
36 - 40 years old	67
41 - 50 years old	34
#NULL!	30
50+ years or older	8