DATA 106 - Assignment 1

Jillian Morrison

September 2, 2019

General rules

- For some questions, the needed methods may not have been covered in class. For them, please do some research to solve them.
- You must show your work in order to get points. Providing correct answers without supporting codes or intermediate steps does not receive full credit.
- You must submit both the R file as a .R file and the Assignment file as a PDF. For the Assignment file include the code, the output and explanations (if necesssary).

Questions

1. a. Create 2 data frames, buildings (first data frame) and data (second data frame)

```
##   location      name
## 1        1 building1
## 2        2 building2
## 3        3 building3
```

```
##   survey location efficiency
## 1      1        1         51
## 2      1        2         64
## 3      1        3         70
## 4      2        2         71
## 5      2        3         80
## 6      2        1         58
```

Notice that the 2 dataframes have the variable location in common. Merge the two dataframes by this variable. Name the resulting dataframe COW_Buildings

b. Rename the location variable in the 'building' dataset as "Location.ID". Call this new dataset 'buildings_2'

c. Merge the datasets buildings_2 and data. Call this new dataframe NewCOWbuildings

d. Explain the difference between inner join, outer join, right join, left join and cross join.

2. Refer to the table below:

```
Gender <- c("Female","Female","Male","Male")
Restaurant <- c("Yes","No","Yes","No")
Count <- c(220, 780, 400, 600)
DiningSurvey <- data.frame(Gender, Restaurant, Count)
DiningSurvey
```

```
##   Gender Restaurant Count
## 1 Female        Yes   220
## 2 Female         No   780
## 3   Male        Yes   400
## 4   Male         No   600
```

a. Check if any row has count more than 400

b. Append the new variable Flavour to the DiningSurvey dataset.

```
Flavour <- c("Yes", "No", "Yes", NA)
```

c. Use the "is.na()" argument to find missing Flavour data by Gender. Hint(Use the table() function to tabulate the variables is.na(Flavour) and Gender)

3. Consider the RentalUnits Dataset

```
RentalUnits <- matrix(c(45,37,34,10,15,12,24,18,19),ncol=3,byrow=TRUE)
colnames(RentalUnits) <- c("Section1","Section2","Section3")
rownames(RentalUnits) <- c("Rented","Vacant","Reserved")
RentalUnits <- as.table(RentalUnits)
RentalUnits
```

```
##          Section1 Section2 Section3
## Rented         45       37       34
## Vacant         10       15       12
## Reserved       24       18       19
```

a. Use the margin.table() or rowSums() function to find the amount of Occupancy summed over Sections.

b. Find the amount of Units summed by Section.

c. Use the "prop.table()" function to create a basic table of proportions.

d. Find row percentages, and column percentages.

e. Use "summary()" to perform a Chi-Square Test of Independence, of the "RentalUnits" variables. Describe what the Chi- Square test of indendence does (You do not need to go into details).

4. Consider the url 'https://statbel.fgov.be/en/themes/population/structure-population'

I have extracted all the information in table 'Structure of Population' of Belgium to a dataframe called "M". You will need to install the package called rvest.

```
#install.packages('rvest')
library('rvest')
```

```
## Loading required package: xml2
```

```
url='https://statbel.fgov.be/en/themes/population/structure-population'
TAB=read_html(url)%>%html_nodes('td')%>%html_text()
NAMES=read_html(url)%>%html_nodes('th')%>%html_text()

M_ <- as.numeric(gsub(",","",unlist(TAB)))
```

```
## Warning: NAs introduced by coercion
```

```
M=data.frame(matrix(M_,ncol=7,byrow=T))
M=cbind(NAMES[9:23],M)
names(M)=NAMES[1:8]
M
```

```
##             Place of residence Population on 1st January 2018
## 1                      Belgium                       11376070
## 2      Brussels-Capital Region                        1198726
## 3               Flemish Region                        6552967
## 4               Walloon Region                        3624377
## 5    German-speaking Community                             77
## 6          Province of Antwerp                        1847486
## 7          Province of Limburg                            871
## 8    Province of East Flanders                        1505053
## 9  Province of Flemish Brabant                        1138489
## 10   Province of West Flanders                        1191059
## 11 Province of Walloon Brabant                            401
## 12         Province of Hainaut                        1341645
## 13           Province of Liège                        1105326
## 14      Province of Luxembourg                            283
## 15           Province of Namur                            493
##    Natural balance Internal migration balance
## 1                7                          0
## 2                8                        -15
## 3              939                         12
## 4               -2                          3
## 5               60                         79
## 6                2                       -448
## 7              -49                        180
## 8              225                          4
## 9              373                          5
## 10              -2                          3
## 11             100                          2
## 12              -2                          2
## 13            -476                       -522
## 14             124                        311
## 15            -278                        221
##    International migration balance Statistical adjustment Total growth
## 1                               50                     -2           55
## 2                               17                   -730           10
## 3                               25                     -1           36
## 4                                8                    -24            9
## 5                              208                     -5          342
## 6                               NA                   -478           11
## 7                                3                   -131            3
## 8                                6                   -279           10
## 9                                3                   -254            8
## 10                               4                   -103            5
## 11                             652                    -57            2
## 12                               2                    268            3
## 13                               3                   -260            2
## 14                             965                     11            1
## 15                               1                     14            1
##    Population on 1st January 2019
## 1                         11431406
## 2                          1208542
## 3                          6589069
## 4                          3633795
## 5                               78
## 6                          1857986
## 7                              874
## 8                          1515064
## 9                          1146175
## 10                         1195796
## 11                             404
## 12                         1344241
## 13                         1106992
## 14                             285
## 15                             494
```

```
#######NOTE#########
##Header cells - contains header information (created with the <th> element)
##Standard cells - contains data (created with the <td> element)

##These can be found in the page source see: https://smallbusiness.chron.com/see-html-code-46954.html
###################
```

a. Create a scatterplot of Total Growth on the y axis and Population on 1st January 2019 on the x axis. Be sure to add axis and column names. Add a linear regression line to the plot (see http://www.sthda.com/english/wiki/scatter-plots-r-base-graphs)

b. Remove the outlier from part a and remake the plot.(Hint: look for ways to remove a specific element from a dataframe) Also add a linear regression line to the plot.

c. Describe what you see with and without the outlier.

d. Go to the bottom of page 2, you will see “Warning: NAs introduced by coercion”. Which element of the table was “coerced” into being missing (i.e. NA). How would you replace the NA with the correct value?