Homework 3

Important: To submit your homework, please create your GitHub repository on the AU-R-Programming organization by Monday (Oct. 5th 2020) at 2pm and use this repository to work on this assignment. Your final submission will be done via Canvas in html format (in which you will specify the name of the corresponding GitHub repository) and is due by Friday, Oct. 16th 2020 at 11.59pm (no late work is accepted). The version submitted on Canvas will have to correspond to the last version on the GitHub repository.

To start, create a **private** GitHub repository and start its name with **HW3**. This project **must** be done using GitHub and respect the following requirements:

(1) All commit messages must be reasonably `clear` and `meaningful`.

(2) Include the code in the output necessary to replicate your results.

Problem 1: COVID-19

The “covid.csv” file contains data on Covid-19 cases in a province of Italy. For the following exercises, only provide first rows or elements in the output if the results are too large.

1. Correctly format the dates in “Date.of.birth”, “First.day.of.symptoms”, “Date.of.outcome” and “Date.of.diagnosis”.
2. Format “Hospitalization.type”, “Symptoms” and “Outcome” as factors and “Epidemiological.link…Notes” as character.
3. Create a three-dimensional table reporting the three factors from the previous question. Hint: use the function `table()` with the variables of interest.
4. Using the table object created in the previous question, subset it in order to create the following table.

	Disease in progress	Healed
Home isolation	64	153
Intensive care	0	1

5. Order the data based on the date of diagnosis (from first to most recent).
6. Add a column that reports whether or not a case was asymptomatic AND in home isolation. Name the observation “Home_Asymptomatic” if the conditions apply and “Non_Home_Asymptomatic” if not and then produce a bar plot of this new variable.
7. Count the number of cases of people born after 1981 and that have healed.
8. Count the number of cases that are asymptomatic OR in home isolation (but not both) AND were born before 1982.
9. Create a new dataset including only the rows where “Epidemiological.link…Notes” includes the words “contact” OR “symptom” (or both). Hint: you can use the `grep()` function and `tolower()`.

10. In the previous dataset add a column reporting the age (in years, therefore in integer format) of each patient as of October 2nd, 2020. Save this dataset into a .csv file and make it available on your GitHub repository for this assignment.
11. Produce a pie chart for the type of hospitalization for cases born between 1960 and 1980.

Problem 2: Salary

The "adult.csv" dataset was downloaded from the UCI data repository (https://archive.ics.uci.edu/ml/datasets/Adult) and collects census information paired with a variable determining whether an individual's salary is larger or smaller than (or equal to) \$50,000. To do so, we will use logistic regression which can be generally represented as follows:

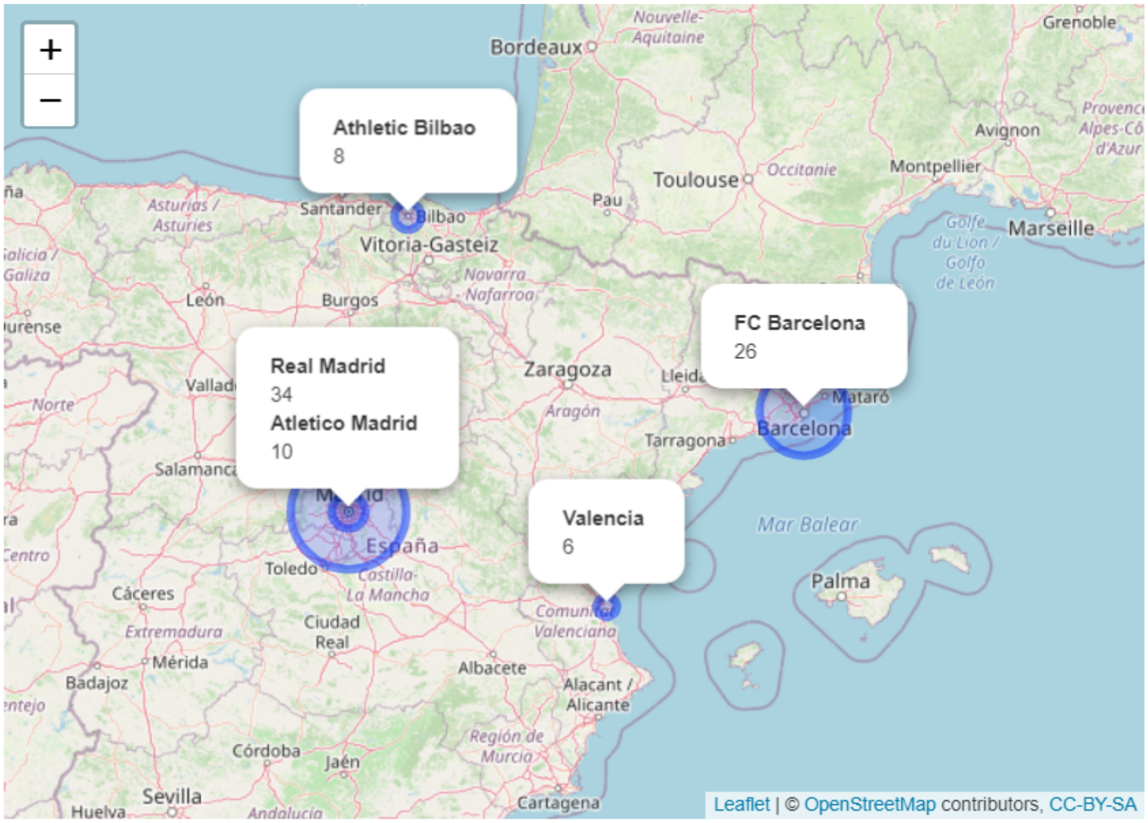
$$y = g(\mathbf{X}\beta)$$

where y is a dichotomous variable with two levels, $g(\cdot)$ is a known function, $\mathbf{X}$ is the matrix/dataframe containing the independent (census) variables and β is the coefficient vector determining how each variable contributes to determining y. Our goal is to estimate the coefficient vector β and check which coefficients are statistically significant.

1. Load the data and check the formatting of the variables.
2. Rename the last column (currently called "NA") containing the dichotomous salary information. Assign the name `salary` to it.
3. The values in `salary` have a space in front of them (e.g. " >50K"): remove the space from all values (hint: you can use `substring()`).
4. Again, in the variable `salary` , replace ">50K" with the value 1 and "<=50K" with 0. Make sure to format it as a factor with two levels.
5. Use the `glm()` function to estimate this logistic model. Only specify the arguments `formula = salary ~ .` , `data` and `family = binomial` . Save the result of the estimation in an object called `fit` .
6. Using the information in `fit` (hint: you can use the functions `coef()` and `summary()` to extract information), create a dataframe collecting:
   1. The names of the variables (names of the rows of the dataframe);
   2. The value of the coefficients (first column);
   3. A logical vector stating which coefficients are positive (second column);
   4. The p-values (third column). .
7. Subset the dataframe created in the previous question to only show the rows where the p-values are strictly smaller than 0.05. Knowing that (i) the remaining rows are statistically significant and that (ii) positive coefficients contribute to increase the probability of a salary larger than \$50,000 (the opposite for negative values), comment on those variables that negatively contribute to salary.

Bonus Problem (10 points): Map

Using the `leaflet` and `htmltools` libraries (or others of your choice), create a simple map to represent the historical champions of La Liga. More specifically, the goal of this problem is to obtain the map below or something similar to it:



The data needed for this graph is represented below:

Teams	Champions	City
Real Madrid	34	Madrid
FC Barcelona	26	Barcelona
Atlético Madrid	10	Madrid
Athletic Bilbao	8	Bilbao
Valencia	6	Valencia