096202 - Introduction to Data Analysis

Homework 2 - Exploration and Visualization

Guidelines

- Deadline is 12.12.19 at 23:55
- Submission in singles. Each student is expected to work on the notebook independently.
- Discussing the homework (but not working on them) with other students is allowed, but any student who does so must add a markdown cell at the top of the notebook explicitly naming the person with whom s/he discussed the homework.
- Submission will include a single Jupyter notebook file named '**HW2_ID.ipynb**'
- Present all supporting calculations and analyses in **code cells**. Document your code with comments and clearly specify the final answers in **markdown cells**
- Any visualization **must include** axis labels.
- A particularly thoughtful and careful work may result in bonuses.

Part 1 - Dataset

In this section you are asked to explore a given dataset, answer the following questions, and provide proper visualization to support your claim.
Download the dataset file ('HW2-data.csv') from the course's Moodle site.

The data was collected in a study during the 2005-2006 school year from two public schools in Portugal. The dataset was built from two sources: school reports, based on paper sheets (i.e. grades and absences) and questionnaires. The dataset contains the following fields:

Column Name	Description
school	Student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
sex	Student's sex (binary: 'F' - female or 'M' - male)
age	Student's age (numeric: from 15 to 22)
address	Student's home address type (binary: 'U' - urban or 'R' - rural)
famsize	Family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
Pstatus	Parent's cohabitation status (binary: 'T' - living together or 'A' - living apart)
Medu	Mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education, or 4 - higher education)
Fedu	Father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education, or 4 - higher education)
Mjob	other's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
Fjob	Father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
reason	Reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
guardian	Student's guardian (nominal: 'mother', 'father' or 'other')
traveltime	Home to school travel time (numeric: 1 -<15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 ->1 hour)
studytime	Weekly study time (numeric: 1 -<2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 ->10 hours)
failures	Number of past class failures (numeric: n if 1<=n<3, else 4)
schoolsup	Extra educational support (binary: yes or no)
famsup	Family educational support (binary: yes or no)
paid	Extra paid classes within the course subject (binary: yes or no)
activities	Extra-curricular activities (binary: yes or no)
nursery	Attended nursery school (binary: yes or no)
higher	Wants to take higher education (binary: yes or no)
internet	Internet access at home (binary: yes or no)
romantic	With a romantic relationship (binary: yes or no)
famrel	Quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
freetime	Free time after school (numeric: from 1 - very low to 5 - very high)
goout	Going out with friends (numeric: from 1 - very low to 5 - very high)
Dalc	Workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
Walc	Weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
health	Current health status (numeric: from 1 - very bad to 5 - very good)
absences	Number of school absences (numeric: from 0 to 93)
G1	First period grade (numeric: from 0 to 20)
G2	Second period grade (numeric: from 0 to 20)
G3	Final grade (numeric: from 0 to 20, output target)
class	Name of course

Questions

1. How many courses are listed in the dataset and what are their names?
2. How many students are listed in the dataset?
3. How many students are listed in each course?
4. What is the mean of the final grades in each course? Present the answer in one graph.
5. Plot the histogram of the students' ages. Describe the plot in words and explain which bin size you used and why.

   For the following questions, present your claim supported by visualization. Make sure that you choose the proper type of visualization. Briefly explain (1-2 sentences) why you chose a particular visualization.

6. Is it correct to say that students who invest more time on studying achieve higher final grades?
7. Is there any association between alcohol consumption and absences?
8. Are the courses balanced gender-wise?
9. Do students who tend to go out more consume more alcohol?
10. Is there any association between alcohol consumption and quality of family relationship?
11. Is there any association between alcohol consumption and the students' final grades?
12. Is it true that alcohol consumption is higher on weekends?
13. How many students are in each school? Show a plot that presents the number of male and female students in each school.
14. A *binge drinker* is someone who consumes a lot of alcohol on the weekends but can easily get through the week without drinking. Create a new feature named 'binge_drinker' that specifies if a student is a binge drinker or not. How many binge drinkers are in each school? which gender includes more binge drinkers?

    For the following questions, write your answers in text (no code needed).

15. Choose one of the questions above, for which you determined that *there was* some connection between two variables. Can you infer from the data that this relationship is **causal**? If so, state the causal relationship and how the data supports it. If not, suggest additional data you would collect to determine whether the relationship is causal or not.
16. Describe one type of bias that you suspect might have occurred in the collection of the provided dataset. Give examples for 2 variables that you think might be affected by this bias.

Part 2 - Simulation

In this part, your job is to simulate a N-door Monty Hall game and help a contestant in the game make a choice.

In a N-door Monty Hall game, a contestant is facing N closed doors. Behind exactly one of the doors there is a car. If the contestant opens this door, s/he gets the car, and if s/he opens any other door, s/he gets nothing.

The game is as follows:
First, the contestant picks one of the doors (assume this choice is made at random) without opening it. Then, Monty, the host, opens $N-2$ doors: all doors that the contestant **did not** pick, except one. The car is never behind any of the $N-2$ doors that Monty opens (Monty knows where the car is). That is, at this point there are two unopened doors left: the one the contestant picked and another one. Finally, Monty offers the contestant the option to switch from the door previously chosen to the other unopen door.

Your task is to simulate the probability that accepting the offer (i.e. switching doors) would win the contestant the car. To do so, you must simulate the game itself as we did in class (rather than using any math). Repeat each simulation at least 5000 times (but not more than 20000 times) before aggregating the results.

1. For $N = 4$, what is the probability that switching will get the contestant the car?
2. Simulate the probabilities that switching gets the contestant the car for $N \in \{3, 4, ..., 20\}$.
3. Make a plot of the probability of winning from a switch as a function of N. Briefly comment on the output: Is it a good idea for the contestant to switch?