Introduction to Machine Learning in R (NPFL054)

Easy HW – Data analysis
Contact: Barbora Hladká (hladka@ufal.mff.cuni.cz)

Write an R code that reads the Titanic data file and finds anwers to the questions 1-3. Here is a link to download the file: `https://ufal.mff.cuni.cz/~hladka/2021/docs/train.csv`.

The `train.csv` file contains data for 891 of the real Titanic passengers. Each row represents one person. The columns describe different attributes about the persons including whether they survived, their passenger-class, their name, their sex, their age, the number of their siblings and spouse aboard the Titanic, the number of their parents and children aboard the Titanic, their ticked id, the fare they paid, their cabin number and their port of embarkation.

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	1	0	3	Braund, Mr. Owen Harris	male	22.00	1	0	A/5 21171	7.2500		S
2	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38.00	1	0	PC 17599	71.2833	C85	C
3	3	1	3	Heikkinen, Miss. Laina	female	26.00	0	0	STON/O2. 3101282	7.9250		S
4	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.00	1	0	113803	53.1000	C123	S
5	5	0	3	Allen, Mr. William Henry	male	35.00	0	0	373450	8.0500		S
6	6	0	3	Moran, Mr. James	male	NA	0	0	330877	8.4583		Q
7	7	0	1	McCarthy, Mr. Timothy J	male	54.00	0	0	17463	51.8625	E46	S
8	8	0	3	Palsson, Master. Gosta Leonard	male	2.00	3	1	349909	21.0750		S
9	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.00	0	2	347742	11.1333		S
10	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.00	1	0	237736	30.0708		C

Questions

1. For each attribute print the number of missing values. For `Age` attribute replace the missing values with its median and visualize its values using histogram. Describe the histogram.

2. Calculate the conditional probability that a person survives given their `Pclass` and `Sex`:

   $\Pr(\text{Survived} = 1 | \text{Sex} = female, \text{Pclass} = 1)$ $\quad$ $\Pr(\text{Survived} = 1 | \text{Sex} = female, \text{Pclass} = 2)$
   $\Pr(\text{Survived} = 1 | \text{Sex} = female, \text{Pclass} = 3)$ $\quad$ $\Pr(\text{Survived} = 1 | \text{Sex} = male, \text{Pclass} = 1)$
   $\Pr(\text{Survived} = 1 | \text{Sex} = male, \text{Pclass} = 2)$ $\quad$ $\Pr(\text{Survived} = 1 | \text{Sex} = male, \text{Pclass} = 3)$

3. Produce side-by-side boxplots of the age of the male and female passangers. For each `Sex` value, draw a point for the average age in the corresponding boxplot. Describe the boxplots.

Presentation

- Create a 10 min presentation.

- Present your answers. If you want to highlight something in your R code, please do it.

- Explain your answers clearly so that your audience understands your method well.

- Submit your presentation and R code to `hladka@ufal.mff.cuni.cz`.