

NPFL142.C4DHI – tutorial #1

Cluster computing, exploring *Titanic* data and searching *Migrant stories*

Info on Cluster Computing.....	2
Exercises with <i>Titanic</i> dataset.....	3
Exercise 1.1 – Getting a data set.....	3
Exercise 1.2 – Loading a data set and showing its structure.....	3
Exercise 1.3 – Exploration of Sex of Titanic passengers.....	4
Exercise 1.4 – Exploration of survival of Titanic passengers based on Sex.....	5
Exercises with <i>Migrant stories</i> dataset.....	6
Exercise 2.1 – Getting a data set.....	6
Exercise 2.2 – Loading a data set and showing its structure.....	7
Exercise 2.3 – Directions of migration.....	8
Exercise 2.4 – Vocabulary of migrants.....	9

Info on Cluster Computing

We will be computing on the [Artificial Intelligence Cluster](#) administered by ÚFAL MFF UK. Namely, we will be using RStudio provided by the [JupyterLab Notebook](#) installed at AIC. You have already tried out to log in <https://aic.ufal.mff.cuni.cz/jlab> when you did the [homework assignment #0](#).

An identical setup is an advantage of a common computing environment. We will not have to troubleshoot problems that occur during system installation on local devices. All required libraries will be pre-installed, allowing us to focus directly on the code details. It will also be very convenient to copy data and codes from the lecturer's directories to your home directory.

At the end of the course, you will have a total of 5 directories in your home directory. Each of them will be for one lecture+lab session (i.e. 5 sessions in total). There will be data and codes in the directories and you can download their contents to your local drive using File Manager in RStudio (Files&Plots desktop > More > Export). We will follow the following convention for R code names in the directories:

- `dataset.l.R` # R code presented during a **l**ecture, e.g. `titanic.l.R`
- `dataset.t.R` # R code presented during a **t**utorial (i.e. during a lab session)
- `dataset.h.R` # R codes for the **h**omework assignments

Exercises with *Titanic* dataset

Data description – [Titanic - Machine Learning from Disaster](#). We will be working with the `train.csv` file downloaded from the Kaggle web site and renamed as `titanic.csv`.

Exercise 1.1 – Getting a data set

In RStudio move to `Terminal` window

- Use `cd` command to move to your home directory
- Create a new directory `mkdir -p 1/titanic`
- Move to `titanic` directory `cd 1/titanic`
- Get the *Titanic* dataset `cp ~hladka/1/titanic/titanic.csv ./`
- List the directory contents `ls`

Exercise 1.2 – Loading a data set and showing its structure

In RStudio create a blank R script

- Move to Files&Plots desktop
- In Files manager
 - move to `1/titanic` directory
 - use More in the menu to run Set as working directory
 - use New Blank File in the menu to create a blank R Script and name it `titanic.t.R`. Then the script is open in the Code editor window (upper-left window) and you can add the `commands` listed below to the script.

We suppose using [tidyverse](#) package.

```
library(tidyverse)
```

Load the *Titanic* dataset into your R environment and look at its structure.

```
dataset <- read_csv("titanic.csv")
print(dataset)
# A tibble: 891 × 12
  PassengerId Survived Pclass Name      Sex      Age SibSp Parch Ticket  Fare Cabin
  <dbl>      <dbl> <dbl> <chr>   <chr>   <dbl> <dbl> <dbl> <chr>  <dbl> <chr>
1         1         0     3 Braun... male     22     1     0 A/5 2...  7.25 NA
2         2         1     1 Cumin... fema...  38     1     0 PC 17... 71.3  C85
3         3         1     3 Heikk... fema...  26     0     0 STON/...  7.92 NA
4         4         1     1 Futre... fema...  35     1     0 113803 53.1  C123
5         5         0     3 Allen... male     35     0     0 373450  8.05 NA
. . .
```

Check the number of examples, i.e. passengers onboard.

```
nrow(dataset)          # number of rows in the tibble
[1] 891
```

Check the number of attributes.

```
ncol(dataset)           # number of columns in the tibble
[1] 12
```

Check the attribute names.

```
colnames(dataset)
 [1] "PassengerId" "Survived"   "Pclass"     "Name"       "Sex"
 [6] "Age"         "SibSp"      "Parch"      "Ticket"     "Fare"
[11] "Cabin"       "Embarked"
```

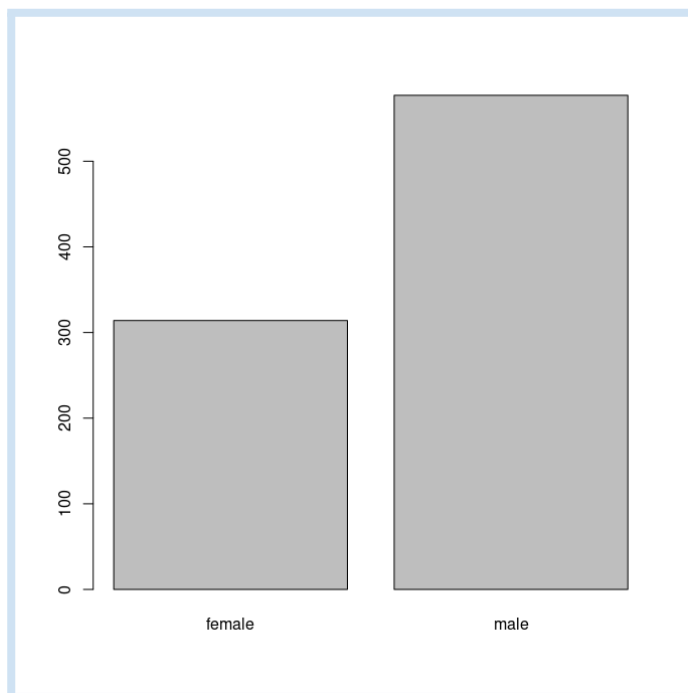
Exercise 1.3 – Exploration of Sex of Titanic passengers

Explore the distribution of values of a given attribute.

```
table(dataset$Sex)           # frequency table for Sex attribute
female  male
   314   577
```

Visualize the distribution.

```
barplot(table(dataset$Sex))
```



Calculate the proportion of men and women.

```
N <- nrow(dataset)
table(dataset$Sex)/N
  female  male
0.352413 0.647587
```

Proportion in rounded percentage.

```
round(table(dataset$Sex)/N * 100, 1)
```

```
female  male
35.2    64.8
```

Exercise 1.4 – Exploration of survival of Titanic passengers based on Sex

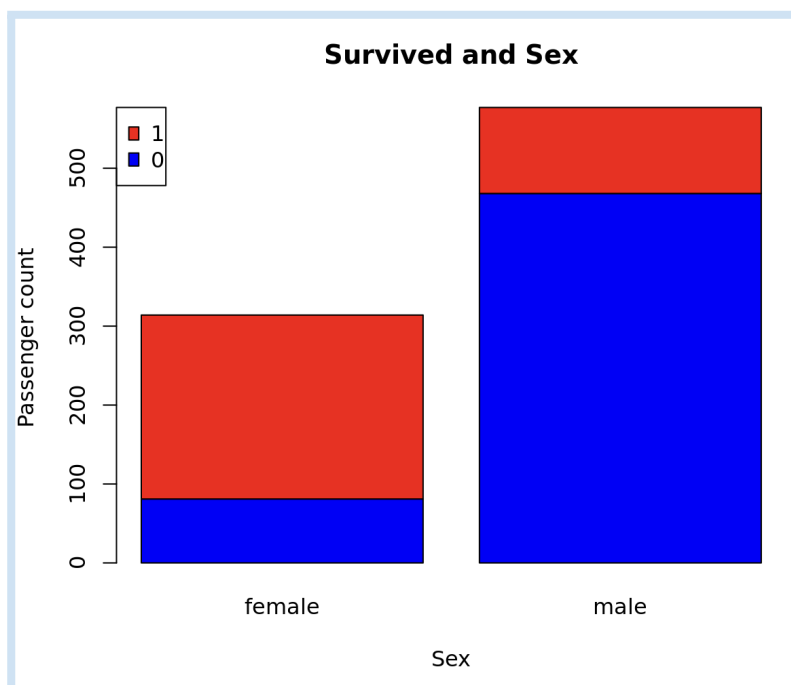
Research Question: What did survival depend on during the sinking of the Titanic?

Explore the distribution of values of the Survived and Sex attributes.

```
survived.sex <- table(dataset$Survived, dataset$Sex)
survived.sex                                     # contingency table for Survived and Sex attributes
  female male
0      81  468
1     233  109
```

Visualize the contingency table.

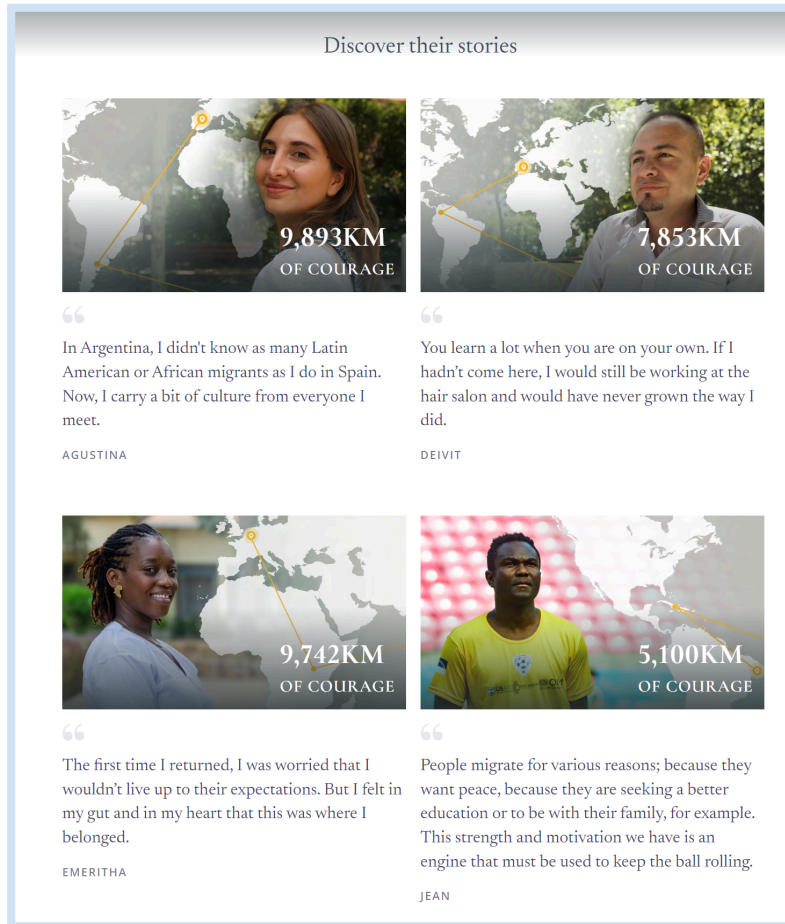
```
barplot(survived.sex,
        main = "Survived vs. Sex",
        xlab = "Sex",
        ylab = "Passenger count",
        col = c("blue", "red"),
        legend.text = TRUE,
        args.legend = list(x = "topleft")
)
```



Exercises with *Migrant stories* dataset

Data description – “I am a migrant” is a campaign created by the International Organization for Migration (IOM, <https://www.iamamigrant.org>) to promote diversity, inclusion and fight xenophobia and divisive narratives on migration. The platform features first-hand accounts from people on the move. The stories are written in English, and we do not know how they were collected.

Discover their stories



Name	Distance (KM)	Quote
AGUSTINA	9,893KM	In Argentina, I didn't know as many Latin American or African migrants as I do in Spain. Now, I carry a bit of culture from everyone I meet.
DEVIT	7,853KM	You learn a lot when you are on your own. If I hadn't come here, I would still be working at the hair salon and would have never grown the way I did.
EMERITHA	9,742KM	The first time I returned, I was worried that I wouldn't live up to their expectations. But I felt in my gut and in my heart that this was where I belonged.
JEAN	5,100KM	People migrate for various reasons; because they want peace, because they are seeking a better education or to be with their family, for example. This strength and motivation we have is an engine that must be used to keep the ball rolling.

Exercise 2.1 – Getting a data set

Some stories are available as a dataset published in the [LINDAT/CLARIAH-CZ repository](#) and they are searchable in [TEITOK](#).

Migrant Stories

Please use the following text to cite this item or export to a predefined format: BIBTEX CMDI

Hájek, Martin; Mirovský, Jiří and Hladká, Barbora, 2022, *Migrant Stories*, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (UFAL), Faculty of Mathematics and Physics, Charles University, <http://hdl.handle.net/11234/1-4818>.

This resource is also integrated in following services: TEITOK Share: f t LINDAT / CLARIAH-CZ

Authors	Hájek, Martin ; Mirovský, Jiří and Hladká, Barbora
Item identifier	http://hdl.handle.net/11234/1-4818
Project URL	https://ufal.mff.cuni.cz/courses/npfl134/migrant-stories
Date issued	2022-10-22
Type	corpus, text
Size	1017 entries
Language(s)	English
Description	Migrant Stories is a corpus of 1017 short biographic narratives of migrants supplemented with meta information about countries of origin/destination, the migrant gender, GDP per capita of the respective countries, etc. The corpus has been compiled as a teaching material for data analysis.
Publisher	Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics (UFAL)
Acknowledgement	4EU+ European University Alliance Project code: 2021_F3_10 Project name: SWitCH: Crash Course on Data Analytics for Students of Social Studies and Humanities
Subject(s)	corpus

In RStudio move to **Terminal** window

- Use `cd` command to move to your home directory
- Create a new directory `mkdir -p 1/migrants`
- Move to `titanic` directory `cd 1/migrants`
- Get the *Titanic* dataset `cp ~hladka/1/migrants/migrants.tsv ./`
- List the directory contents `ls`

Exercise 2.2 – Loading a data set and showing its structure

In RStudio create a blank R script

- Move to Files&Plots desktop
- In Files manager
 - move to `1/migrants` directory
 - use More in the menu to run Set as working directory
 - use New Blank File in the menu to create a blank R Script and name it `migrants.t.R`. Then the script is open in the Code editor window (upper-left window) and you can add the **commands** listed below to the script.

Load the *Migrant stories* dataset into your R environment and look at its structure.

```
dataset <- read_tsv("migrants.tsv")  
print(dataset)
```

Exercise 2.3 – Directions of migration

Research Question: What are directions of migration?

Sort the origin countries (country+or attribute) by the number of migrants who left the country in a decreasing way.

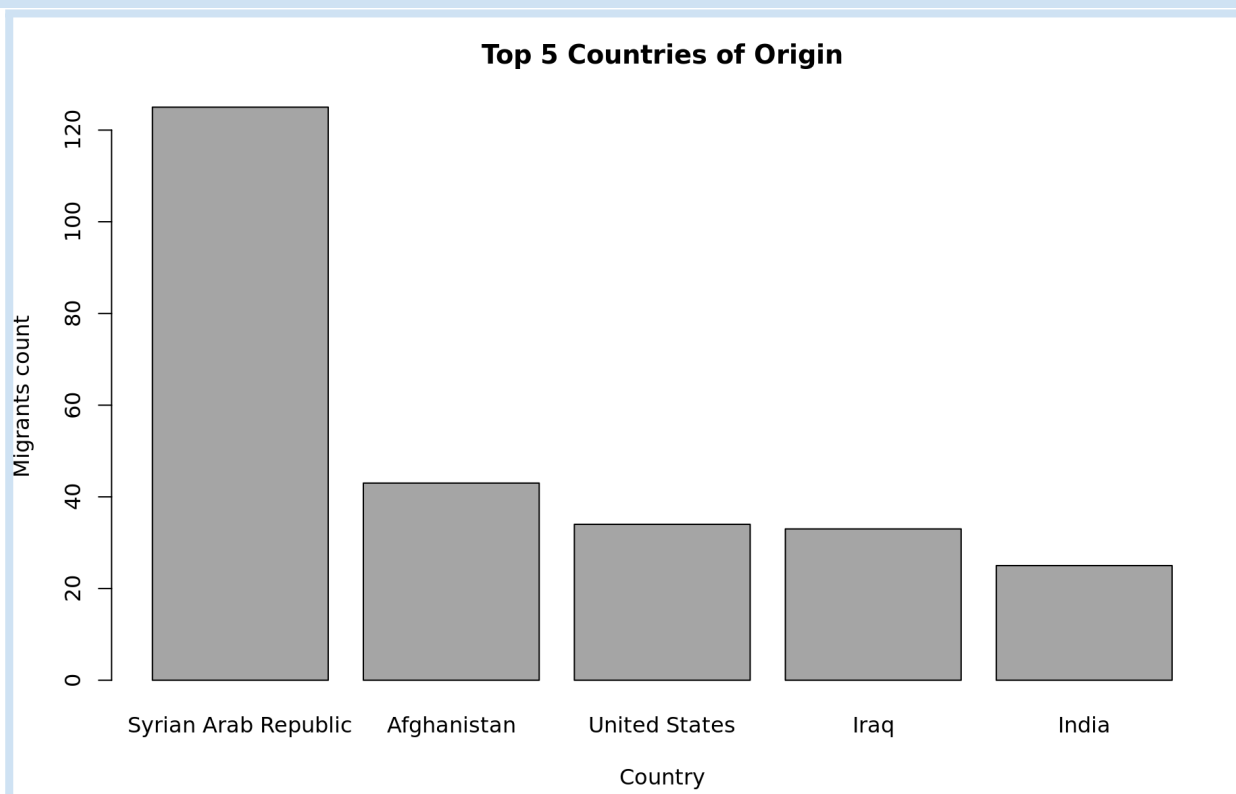
```
countries.or <- table(dataset$country_or)  
sorted.countries.or <- sort(countries.or, decreasing=T)
```

Select TOP 3 origin countries.

```
n <- 3 # TOP 3  
sorted.countries.or[1:n]  
Syrian Arab Republic      Afghanistan      United States  
          125              43              34
```

Visualize TOP n origin countries.

```
n <- 5  
barplot(sorted.countries.or[1:n],  
        ylab="Migrants count",  
        xlab="Country", width=2,  
        main="Top 5 Countries of Origin",  
        col="darkgrey",  
        cex.names=1.0)
```



Exercise 2.4 – Vocabulary of migrants

Research Question: How do migrants describe their stories?

We suppose using [stringr](#) package.

```
library(stringr)
```

Read e.g. the first story.

```
dataset$story[1]
```

We are interested in how many times a string occurs in a document. In programming, functions allow us to repeat a sequence of code without having to write the code over again. We write a function to count the number of occurrences for a string in a set of documents.

```
string_counter <- function(texts, str) {  
  # texts - set of documents  
  # str - string  
  counts <- NULL  
  for(i in 1:length(texts)){  
    counts[i] <- str_count(texts[i], str)  
  }  
  return(counts)  
}
```

How many times does *famil* string occur in each Migrant story?

```
string <- "famil"  
counts <- string_counter(dataset$story, string)  
table(counts)
```

A histogram is a graphical representation of the distribution of values of a variable using a bar plot with columns (bins) of the same width (i.e. the width of the intervals), while the height of the bins means the frequency of the variable in the given interval. Plot a histogram showing frequency distribution of *famil* occurrences.

```
hist(counts, breaks = 10,  
      main = "Histogram for family occurrences",  
      xlab = "famil count",  
      ylab = "Migrant count"  
)
```

Histogram for famil occurrences

