Fachbereich AIHI
Universität Salzburg
Univ.-Prof. Dr. Wolfgang Trutschnig

Salzburg, April 2024
SS 23/24

Übungsblatt 04 zu 'Applied AI Using R'

From now on, some of the exercises sheets will be in English.

Main objective of the current sheet is analyze some standard datasets by applying functions we have already discussed in the course. Notice that exercises like these prepare you for participating in projects and for internships in companies.

Aufgabe 17.
In this exercise we work again with the `flights` dataset of the `nycflights13` package.

1. For each carrier, compute the average departure delay, the standard deviation of the departure delay and the number of flights. Which carrier would you prefer?

2. For each departure airport, calculate the percentage of flights that have a negative departure delay (i.e. that left early).

3. Those negative departure delays skew our computed average and allow carriers to "cheat". Repeat part 1., but first set all negative departure delays to zero (hint: use `ifelse()`). Which carrier would you prefer now?

4. The columns `dep_time`, `sched_dep_time`, `arr_time` and `sched_arr_time` have a weird format. Use `?flights` to look it up in the documentation. Write a function that takes this time format and returns the number of minutes since midnight. Use your function to transform all four `*_time` columns in the dataset to the new format.

5. The `NA` values in the dataset correspond to cancelled flights. For each minute of the day, calculate the proportion of cancelled flights (hint: use the new time format from part 4.). Would you rather choose to fly before lunch or after lunch?

Aufgabe 18.
For this exercise we need to obtain a publicly available dataset containing leaked passwords. Use the `2017 master password list` column of the following spreadsheet (yes, this seems to be real data):
`https://docs.google.com/spreadsheets/d/1cz7TDhm0ebVpySqbTvrHrD3WpxeyE4hLZtifWSnoNTQ/edit#gid=16`

1. Download the data and import it to `R`. We only need the columns `rank`, `Password`, `category` and `offline crack (sec)`.

2. Look up the documentation for the `stringr` package (part of the `tidyverse`). Create new columns `has_uppercase`, `has_lowercase`, `has_numbers`, `has_symbols` that indicate if the password contains uppercase letters, lowercase letters, numbers or non-alpha-numeric symbols. Also create a column `len` that contains the length of each password.

3. Consider the following password types: Lowercase only, lowercase + uppercase, lowercase + uppercase + numbers. For each combination of password length and password type compute the average offline crack time. Make a plot using `geom_tile()`.

4. For each combination of category and length compute the percentage of "relatively strong" passwords (containing lowercase, uppercase and numbers).

5. Are "relatively strong" passwords on average longer than passwords that contain just lowercase letters?

Aufgabe 19.
We analyze de world happiness index:

1. Download the data for Figure 2.1 from `https://worldhappiness.report/ed/2024/#appendices-and-data` and load it into `R` (hint: use the `readxl::read_xls()` function).

2. Download the data from `https://ourworldindata.org/grapher/continents-according-to-o tab=table` and load it into `R`.

3. For each country in the world happiness index add the corresponding region. There are seven contries that have no corresponding entry. Add these regions manually.

4. How many "Entities" in the `continents` dataframe have no happiness index?

5. Create a boxplot depicting the distribution of the happiness index for each region.

6. Download the data from `https://ourworldindata.org/grapher/economic-inequality-gini- tab=table` and load it into `R`.

7. For each country in the world happiness index add the Gini coefficient. Make a scatter plot to show if there is a relation between the happiness index and the Gini coefficient.