Problem Set 10

David Rügamer, Julia Terhart, Philipp Kopper

22 June 2020

Resources

1) Read chapters 10-12 on data wrangling in R4DS.
2) Optionally, complete the DataCamp Units provided on the Moodle page for this topic.
3) Accept the invitation to the assignment of this problem set: https://classroom.github.com/a/mbhvyPEa

Application

In the repository of the assignment, you find a `.csv` file. The data set assesses the association of the temperature on the day of launch and the fact that there was thermal distress. The file has two columns of significance: the temperature in degrees Fahrenheit on the day of the launch and a dichotomous feature that indicates if there was thermal distress during the launch.

a) The `.csv` format of the file `shuttle.csv` is a bit odd. Read it in using the `readr` package.

There is an additional data set with details on the launches in a separate `.csv`. This data set indicates the executive in charge of the respective launch.

b) Read in the data set `names.csv` using the `readr` package and add it to the `shuttle` data set.

The data set is very messy. The strings are not very standardised and do not 100%ly correspond to the three executives "John Miller", "Aaron Smith" and "Newt Montgomery". Either there are typos or the name is reported very informally.

c) Use the `stringr` package to (as much as possible!) convert the strings to the intended three executives. Hint: Maybe think of the use of `regex` and look for matches. Note that this task can be very demanding, so don't waste all of your time if you do not get there.

Transfer

For this problem, you will again work with the `pammtools` package (like in PS7). We work with the `daily` data set. The `daily` data set reflects the nutrition protocols of ICU patients. The patients have a unique id (`CombinedID`) which can be mapped to the data set `patient` which indicates more information on the patients. For each patient, multiple days have been reported. We have information on the `caloriesPercentage`, the percentage of administered calories w.r.t. the recommended amount, and the `proteinGproKG`, the grams of administered protein per Kg body weight.

```
library(pammtools)
head(daily)
```

```
## # A tibble: 6 x 4
##   CombinedID Study_Day caloriesPercentage proteinGproKG
##        <int>     <int>              <dbl>         <dbl>
## 1       1110         1               0            0
## 2       1110         2               0            0
## 3       1110         3               4.05         0
## 4       1110         4              35.1          0.259
## 5       1110         5              77.2          0.647
## 6       1110         6              17.3          0
```

Make use of the `dplyr` package throughout this problem.

a) Make use of `dplyr` to create a `tibble` that does not distinguish between the study days anymore but only reports the average calories and protein for each patient. Compute the correlation between the two features. Report the patients with the highest values of each. Your `tibble` or `data.frame` should have two rows.

b) Look for these patients in the `patient` data set. Report their BMI and their `PatientDied` status. (Note: 0 indicates that the patient survived the time in the ICU.)

c) Analyse the `patient`data set using `dplyr`.

- Report the proportion of deaths in the ICU for each year.
- Create a data set for these patients who survived. How many patients did so?
- Also, create a data frame for the patients who died.
- Compare the two data sets.

Which features seem to be the most different? Report some significant differences. Formulate some hypotheses from the exploratory analysis which could be investigated later on.

d) Fit a linear regression which intends to model `PatientDied` using all relevant features. How does this correspond to the previous task? What is the benefit of this compared to the exploratory analysis?

e) Use `ggplot2` to examine the association of the `BMI` and the `ApacheIIscore` for men and women separately. Hint: `facet`.

```r
sessionInfo()
```

```
## R version 4.0.0 (2020-04-24)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Catalina 10.15.4
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRblas.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] pammtools_0.2.3
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.4.6        pillar_1.4.4        compiler_4.0.0
##  [4] iterators_1.0.12    pec_2019.11.03      tools_4.0.0
##  [7] digest_0.6.25       evaluate_0.14       lifecycle_0.2.0
## [10] tibble_3.0.1        checkmate_2.0.0     gtable_0.3.0
## [13] nlme_3.1-148        lattice_0.20-41     mgcv_1.8-31
## [16] pkgconfig_2.0.3     rlang_0.4.6         foreach_1.5.0
## [19] Matrix_1.2-18       cli_2.0.2           yaml_2.2.1
## [22] prodlim_2019.11.13  mvtnorm_1.1-0       xfun_0.14
## [25] dplyr_1.0.0         stringr_1.4.0       knitr_1.28
## [28] generics_0.0.2      vctrs_0.3.0         grid_4.0.0
## [31] tidyselect_1.1.0    glue_1.4.1          R6_2.4.1
## [34] timereg_1.9.5       fansi_0.4.1         survival_3.1-12
## [37] rmarkdown_2.2       lava_1.6.7          Formula_1.2-3
## [40] tidyr_1.1.0         purrr_0.3.4         ggplot2_3.3.1
## [43] magrittr_1.5        codetools_0.2-16    backports_1.1.7
## [46] scales_1.1.1        ellipsis_0.3.1      htmltools_0.4.0
## [49] splines_4.0.0       assertthat_0.2.1    colorspace_1.4-1
## [52] numDeriv_2016.8-1.1 utf8_1.1.4          stringi_1.4.6
## [55] lazyeval_0.2.2      munsell_0.5.0       crayon_1.3.4
```

You can hand in this problem set by the 6th of July to receive feedback.