Research Skills: Programming with R

Assignment 2

This graded set of homework assignments must be handed in on Canvas before June 8th, 21:00. It tests your mastery of Worksheets 1 to 6. You will be asked to write functions and apply them repeatedly, to clean and tidy data sets, and to fit and evaluate classification models.

It will be graded as follows:

- 0.5 point each for Questions 1 through 7
- 1.5 point each for Questions 8 through 10
- 1.0 point in total for overall code organisation & style
- 1.0 point in total for complying with the instructions below

The guidelines for overall code organisation & style can be found in the slides for Class 4. Note that to receive full marks for this aspect you will have to make use of the `%>%` operator where applicable.

Questions 1 through 7 can be graded semi-automatically. All correct solutions will receive full marks, and any deviations from the requested answers, down to misspellings, will receive 0 points. For Questions 8 through 10, partial solutions will receive partial points, and the efficiency and succinctness of your answers will matter.

All questions are independent except for Question 10; copy the data set before modifying it, and start afresh with the original each time. Other instructions:

- solve all the questions in a single R script
- use `Programming_with_R_2020_BLOCK4_A4-script_template.R`, from Canvas, as the basis of this script
- load the data exactly as shown in this demo; do not adapt the relative paths
- use any function from 'base R', `dplyr`, `tidyr`, `ggplot2`, `caret`, and no other packages
- name your script `Assignment2.R`
- include your name and u-number at the top of your script
- store your solutions to Questions 1 - 7 in the objects described

This is an individual assignment: You may discuss it with your fellow students in general terms but do not share code. Evidence of plagiarism will be referred to the Exam Committee. Good luck!

Data set information

This assignment uses three real (but slightly simplified) data sets that deal with language: `modality`, `AoA`, and `concreteness`. The first dataset, `modality`, has modality exclusivity norms for 400 random nouns, for which participants provided perceptual strength ratings (from 0 very weak to 5 very strong) across five sensory modalities: hearing, taste, touch, smell, and vision (Lynott and Connell 2013). The second dataset, `AoA`, has the age of acquisition of the meaning of words and expressions, that is, it indicates at which age a child learns for the first time a word and its meaning (Brysbaert and Biemiller 2017). The third dataset, `concreteness`, has (i) concreteness ratings (1 not concrete to 5 very concrete), that is, how concrete a certain word is perceived, (ii) its frequency by million, that is, how many times the word appears for a corpus of a million words), and (iii) its dominant position in a sentence (e.g., noun, adjective, etc) for 40 thousand generally known English words (Brysbaert, Warriner, and Kuperman 2014).

Question 1.

Create a copy of `modality` dataset and replace each continuous modality score for a binary score. You should replace a score with "high" if the score is larger than the mean of its modality, and "low" otherwise. For example, since the mean of auditory modality is 2.15, the first value, 1.41 should be replaced by "low". Create this object with a meaningful name initially, then store the dataset in an object called `answer1`.

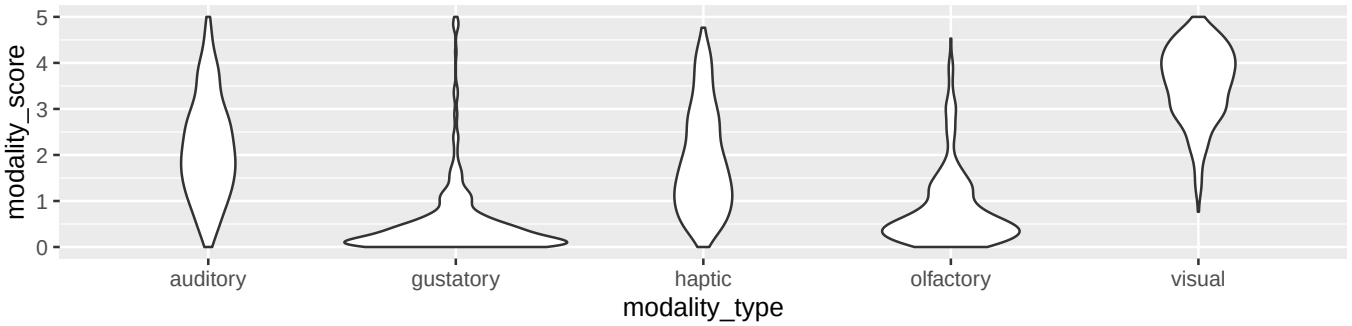
Question 2.

Create a copy of modality and add an `exclusivity` column with the ratio between the dominant modality (the highest score) and the sum of all modality scores (e.g., for academy, exclusivity = 2.705882/(1.41176471 + 0.05882353 0.4705882 + 0.11764706 + 2.705882)). Create this object with a meaningful name initially, then store the modified data frame in an object called `answer2`.

Question 3.

To create a plot with the code shown below, the `modality` dataframe must be re-shaped first. Create the appropriately re-shaped dataframe with a meaningful name initially, then store it in an object called `answer3`. This object should deliver the plot as shown, using the exact code shown below.

```
ggplot(answer3, aes(x = modality_type, y = modality_score)) +
  geom_violin()
```



(Note: For this question, you thus only need to store the re-shaped data set in your `answer3` object; using this `answer3`, it should be possible to produce the plot shown without any alterations to the provided plot code at all.)

Question 4.

The `AoA` data set includes the age of acquisition for different meanings of each word. Summarize this data set so that we have only the smallest age of acquisition for each meaning, sort by AoA (from younger to older age). Create this object with a meaningful name initially, then store it in an object called `answer4`.

Question 5.

Create a new wider dataset that includes the information provided by the `modality` and `concreteness` datasets. Keep only the rows where all the information from the two datasets is available. Create this object with a meaningful name initially, and then store the new data set in an object called `answer5`.

Question 6.

Using the `concreteness` data set and `train()`, fit a `"knn"` model using 3-fold cross validation, optimising accuracy. It should predict whether a word dominant position is a noun or not based on all other variables except for the identity of the word, `Word`); try values for `k` of 3 and 5. Use `set.seed(1)` before fitting this model. Create it with a meaningful name initially, then store it in an object called `answer6`.

(Note: For this question, you do not need to split `concreteness` into a train and test set.)

Question 7.

Using the same dataset, `concreteness`, and `train()`, fit a logistic regression model using 3-fold cross validation, optimising accuracy. It should predict whether a word dominant position is a noun or not based on `concreteness`, `freq_by_million` and their interaction. Make sure that the output of the logistic regression model takes as a reference "Noun"; otherwise stick to the defaults. Use `set.seed(1)` before fitting this model. Create it with a meaningful name initially, then store it in an object called `answer7`.

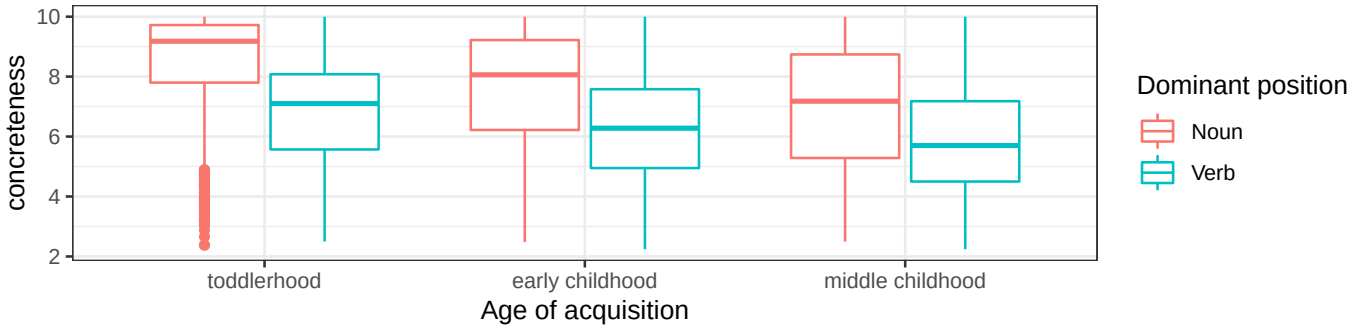
(Note: For this question, you do not need to split `concreteness` into a train and test set. **Ignore the warnings**)

Question 8.

Using the `AoA` and `concreteness` data sets together, create a boxplot showing how the age of acquisition is affected by concreteness and dominant position. Rather than the raw features use the following transformations:

- For age of acquisition, bin it in three categories "toddlerhood" (<=3 years old), "early childhood" (>3 and < 7 years old), "middle childhood" (>=7) (pay attention to show them in chronological order in the plot),
- for concreteness, multiply it by 2, so that the maximum value is 10 (rather than 5), and
- for dominant position, display only the `"Verb"` and `"Noun"` positions.

(The actual labels on the axes do not matter; see one possible plot below for inspiration, but note that you do not need to replicate this plot's aesthetics exactly). Store the plot as an object named `answer8`.



Question 9.

For the following exercise we use the dataset `word_complete_data.csv` which merges all the information from previous exercises:

```
word_complete_data <- read.csv("input/word_complete_data.csv", stringsAsFactors = FALSE)
```

Create a copy of the `word_complete_data` data set with the following modifications:

- remove all the rows with invalid dominant positions ("Unclassified" or variants of missing values),
- for numeric columns, replace missing values with the average value of the column; consider 0 in `freq_by_million` as a missing value,
- `freq_by_million` gets replaced by `log_freq`: to do so first divide `freq_by_million` by one million and then apply the `log()` function,
- change the name of the columns `auditory`, `gustatory` and `olfactory` adding a `_mod` at the end (e.g., `auditory_mod`, `gustatory_mod`, etc),
- replace `AoA` by a column called `by_early_childhood` with `"yes"` if AoA <7 otherwise `"no"`,

The first few rows of a correct solution should look like this:

```
##        word auditory_mod gustatory_mod   haptic olfactory_mod visual_mod
## 1         a     2.148771     0.5279396 1.866581     0.8270425   3.541226
## 2 abandoned     2.148771     0.5279396 1.866581     0.8270425   3.541226
## 3   ability     2.148771     0.5279396 1.866581     0.8270425   3.541226
##   exclusitivity concreteness dominant_position    log_freq by_early_childhood
## 1     0.4337218         1.46           Article  0.04035469               yes
## 2     0.4337218         2.52              Verb -7.29488943               yes
## 3     0.4337218         1.81              Noun -6.92693810               yes
```

Question 10.

For this question we will use `answer9`, if you were unable to solve `answer9`, read the file `word_numeric_data.csv` and work with it instead. Remove the columns `word` and `dominant_position` from `answer9`, and split it into a train and test set. 60% of the observations should be in the train set, 40% in the test set; the `by_early_childhood` variable should be balanced between the splits. Fit a `"knn"` model predicting `by_early_childhood` on the basis of all other variables, optimizing *recall* and centering and scaling all explanatory variables. Use 5-fold cross validation, and test the default values of k.

Then create a `confusionMatrix()` for the test set, and for your final answer, extract only "Precision" and "Recall" from it and store in a named vector (use these metrics as the names) and store this vector as `answer10`. Use `set.seed(1)` at the very start of your solution.

References

Brysbaert, Marc, and Andrew Biemiller. 2017. "Test-Based Age-of-Acquisition Norms for 44 Thousand English Word Meanings." *Behavior Research Methods* 49 (4): 1520–3. https://doi.org/10/gbsdc8.

Brysbaert, Marc, Amy Beth Warriner, and Victor Kuperman. 2014. "Concreteness Ratings for 40 Thousand Generally Known English Word Lemmas." *Behavior Research Methods* 46 (3): 904–11. https://doi.org/10/gd6hzk.

Lynott, Dermot, and Louise Connell. 2013. "Modality Exclusivity Norms for 400 Nouns: The Relationship Between Perceptual Experience and Surface Word Form." *Behavior Research Methods* 45 (2): 516–26. https://doi.org/10/f42jkd.