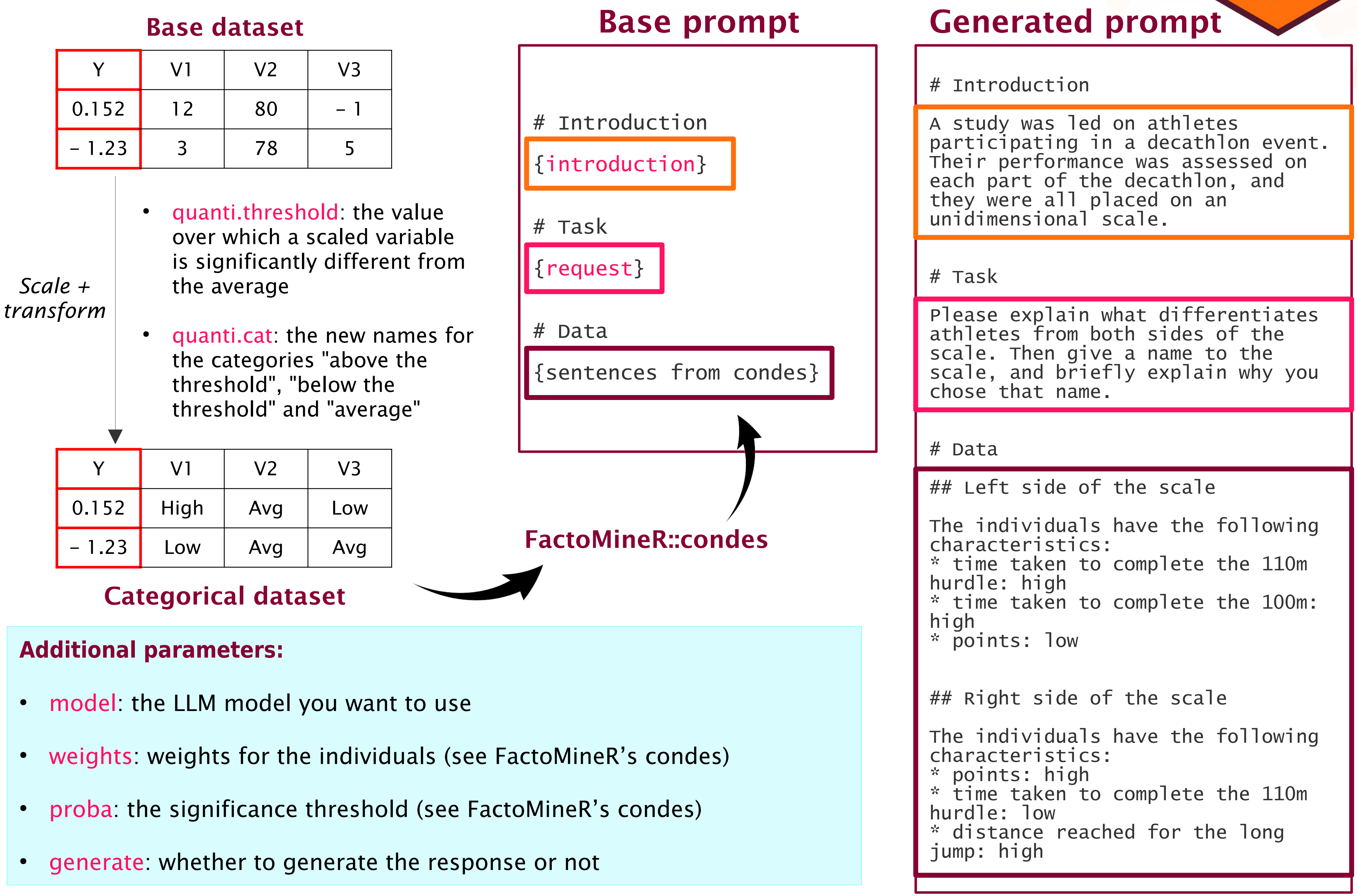
NaileR :: nail_condes

behind the scenes



Base dataset

Y	V1	V2	V3
0.152	12	80	– 1
– 1.23	3	78	5

Scale + transform

- quanti.threshold: the value over which a scaled variable is significantly different from the average
- quanti.cat: the new names for the categories "above the threshold", "below the threshold" and "average"

Y	V1	V2	V3
0.152	High	Avg	Low
– 1.23	Low	Avg	Avg

Categorical dataset

FactoMineR::condes

Base prompt

```
# Introduction
{introduction}

# Task
{request}

# Data
{sentences from condes}
```

Generated prompt

```
# Introduction
A study was led on athletes
participating in a decathlon event.
Their performance was assessed on
each part of the decathlon, and
they were all placed on an
unidimensional scale.

# Task
Please explain what differentiates
athletes from both sides of the
scale. Then give a name to the
scale, and briefly explain why you
chose that name.

# Data
## Left side of the scale

The individuals have the following
characteristics:
* time taken to complete the 110m
hurdle: high
* time taken to complete the 100m:
high
* points: low

## Right side of the scale

The individuals have the following
characteristics:
* points: high
* time taken to complete the 110m
hurdle: low
* distance reached for the long
jump: high
```

Additional parameters:

- model: the LLM model you want to use
- weights: weights for the individuals (see FactoMineR's condes)
- proba: the significance threshold (see FactoMineR's condes)
- generate: whether to generate the response or not

NaileR 1.1.0 – Nel HERVÉ, Sébastien LÊ – Updated 2024-07

NaileR :: nail_condes

tips and tricks

```
# Introduction

A study was led on athletes
participating in a decathlon event.
Their performance was assessed on
each part of the decathlon, and
they were all placed on an
unidimensional scale.


# Task

Please explain what differentiates
athletes from both sides of the
scale. Then give a name to the
scale, and briefly explain why you
chose that name.


# Data

## Left side of the scale

The athletes have the following
characteristics:
* time taken to complete the 110m
hurdle: high
* time taken to complete the 100m:
high
* points: low


## Right side of the scale

The athletes have the following
characteristics:
* points: high
* time taken to complete the 110m
hurdle: low
* distance reached for the long
jump: high
```

The introduction

- Give context for the data
- Explain that the individuals were placed on a scale
- Do not hesitate to repeat important words several times

The request

- Use clear words: explain, describe, summarize…
- Use the same words as in the introduction
- When there are multiple instructions, separate them into different sentences or with commas

More tips

- Set the generate parameter to FALSE to check your prompt and see if it needs editing
- If the variable names are shortened, rename them into long, complete and unambiguous names
- Adapt the quanti.cat to your variables
- Play around with the quanti.threshold and/or proba parameters to fine-tune your analysis
- Do not hesitate to store the prompt in a variable and replace parts of it (with gsub for example) if necessary

NaileR :: nail_condes

use case

Input

```r
library(NaileR)
library(FactoMineR)

data(decathlon)

res_pca_deca = PCA(decathlon, quanti.sup = 11:12,
quali.sup = 13, graph = F)

names(decathlon) <- c('Time taken to complete the
100m', 'Distance reached for the long jump',
'Distance reached for the shot put', 'Height reached
for the high jump',  'Time taken to complete the
400m', 'Time taken to complete the 110m hurdle',
'Distance reached for the discus', 'Height reached
for the pole vault', 'Distance reached for the
javeline', 'Time taken to complete the 1500 m',
'Rank/Counter-performance indicator', 'Points',
'Competition')

deca_work = cbind(res_pca_deca$ind$coord[,1],
decathlon)

intro = "A study was led on athletes participating
in a decathlon event. Their performance was assessed
on each part of the decathlon, and they were all
placed on an unidimensional scale."
requ = "Please explain what differentiates athletes
from both sides of the scale. Then give a name to
the scale, and briefly explain why you chose that
name."

ppt = nail_condes(deca_work, num.var = 1,
quanti.threshold = 1, quanti.cat = c('High', 'Low',
'Average'), introduction = intro, request = requ,
generate = F)

ppt = gsub("individuals", "athletes", ppt)

res_llm = ollamar::generate('llama3', ppt, output =
'df')
cat(res_llm$response)
```

Output

```
Based on the data, athletes on the left side of the
scale tend to have lower values in various events
such as distance reached in the discus and long
jump, higher times taken to complete the 400m, 110m
hurdle, and 100m, and a low overall performance
score. On the other hand, athletes on the right side
of the scale tend to have higher values in these
same events, indicating better performances.

I would name this scale the "Decathlon Performance
Index" (DPI). I chose this name because it
accurately reflects the unidimensional nature of the
scale, which measures overall performance across
various decathlon events. The DPI provides a
comprehensive view of an athlete’s performance,
allowing for easy comparison and ranking between
athletes.
```

Notes:

- variable "Rank" was renamed "Rank/Counter-performance indicator" to help the LLM understand that a high rank means a poor performance
- quanti.cat was changed to "high" and "low" to be clearer for both time and distance
- the customized prompt was sent to ollamar’s generate function, with the output set to "df" to get the result as a data frame