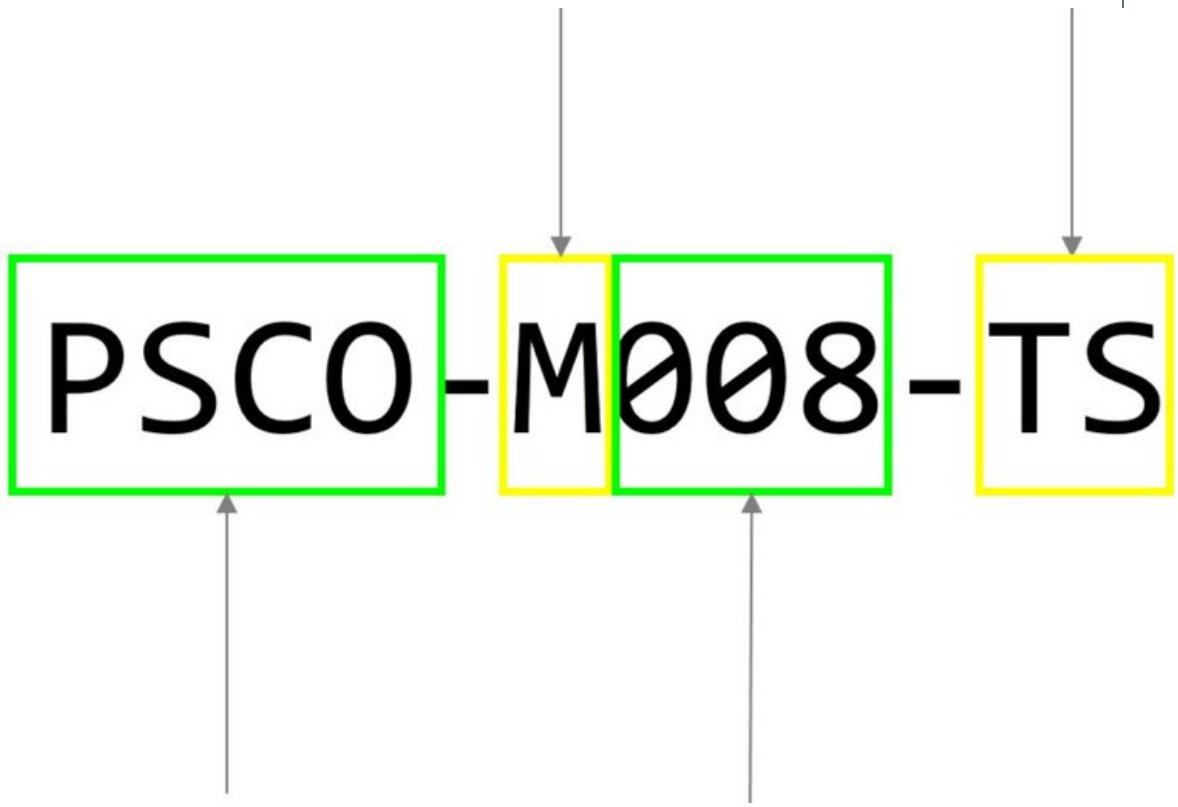

Data Science for Petroleum Engineering - Part 5.2: Finding and filling missing data

Published on August 22, 2017

Alfonso R. Reyes
Petroleum Engineering Data Scientist
32 articles

NOTE. You can find the **PDF** version of the R markdown notebook in GitHub at this link. The reproducible R markdown **notebook** (*.Rmd*) itself is here. Both are **full** versions of this LinkedIn article. For the time being, LinkedIn publishing does not support markdown which would make sharing scientific and engineering documents much easier.

Mistyped data

One of the challenges in cleaning up well data is having uniform and standard well names. This becomes important at the time of classification, ranking and selection. An example of this is when looking for the top 20 oil producers or wells with higher watercut or GOR or wells with highest or lowest gas injection rate (gas lift wells). If a well name is not correct you may encounter repeated occurrences of the well, a wrong classification of the well, or a well that should have received attention but did not. Besides, how good an

the plot of summary.

One of the first things to do, if we are folowing a well name standard, is fin
wells in the raw data file are compliant. One way of doing it is comparing t
with a pattern. In R there are several functions that use patterns for name ve
correction. We will be using a few: `**grep**`, `**grepl**`, `**gsub**`, and couple mor
package `**stringr**`.

Let's start then defining the pattern of a well name.

Pattern detection

If we take a look at the well name in the picture at the top we see that the naming should follow these rules:

- the first 4 alphabetic characters represent the abbreviation of the field
- then, there is dash
- after the dash comes one character for the platform
- then 3 digits, from 000 to 999 that represent the well number
- then a dash, and finally
- two alphabetic characters for the completion type

So, there is a total of 10 significant identifiers plus 2 dashes.

If we use `regular expressions` or `regex` in its simplest form the wells should follow this pattern:

```
PSCO-[M,Q,R,S][0-9][0-9][0-9]-[T,L,S]S
```

Applying the pattern

If we apply the pattern over the raw data set:

```
# using a template to find which well names do not follow a pattern
myXl[!grepl("PSCO-[M,O,P,,Q,R,S][0-9][0-9][0-9]-[T,L,S]S", myXl$Wellname), ]
```

which yields a list of the wells named incorrectly:

This is much better than visually inspecting them in a spreadsheet, isn't it?

What are the type of offences?

- Incorrect well number: `PSCO-M0007-TS`, `PSCO-M0026-TS`
- Platform omitted: `PSCO-027-TS`
- Platform in lowercase: `PSCO-r015-LS`, `PSCO-m016-LS`
- Incorrect field name: `PiSCO-R009-SS`, `PISCO-R027-LS`
- Incorrect completion type: `PSCO-R022-T`, `PSCO-Q019-L`, `PSCO-Q001-S`
- Extra spaces in the name: `PSCO-S019 -LS`

More about regular expressions

stringr - regular expressions

regular expressions in R

Regular Expressions Cheatsheet

Alfonso R. Reyes
Petroleum Engineering Data Scientist