Missing Data

Data Quality:

Alfonso R. Reyes

April 2019

Finding and filling missing data

Example of filling missing data

Well Tests

Well test data: numeric variables

```
#>  [1] "INSTRUMENTATION_TYPE"          "CALC_SEQ_NO"
#>  [3] "CALC_DC_SEQ_NO"                "DURATION_HRS"
#>  [5] "CHOKE_SIZE"                    "Z_SEP_PRESS_PSIG"
#>  [7] "Z_SEP_PRESS_BARG"              "Z_SEP_TEMP_C"
#>  [9] "Z_SEP_TEMP_F"                  "WH_PRESS_PSIG"
#> [11] "WH_PRESS_BARG"                 "ANNULUS_PRESS_PSIG"
#> [13] "ANNULUS_PRESS_BARG"            "WH_USC_PRESS_PSIG"
#> [15] "WH_USC_PRESS_BARG"             "WH_USC_TEMP_C"
#> [17] "WH_USC_TEMP_F"                 "WH_DSC_PRESS_PSIG"
#> [19] "WH_DSC_PRESS_BARG"             "GL_CHOKE_SIZE"
#> [21] "GL_RATE_SCFPERDAY"             "GL_RATE_SM3PERDAY"
#> [23] "GL_RATE_MSCFPERDAY"            "Z_TOTAL_GAS_SCFPERDAY"
#> [25] "Z_TOTAL_GAS_SM3PERDAY"         "Z_TOTAL_GAS_MSCFPERDAY"
#> [27] "TOT_WATER_RATE_ADJ_BBLSPERDAY" "TOT_WATER_RATE_ADJ_M3PERDAY"
#> [29] "NET_COND_RATE_ADJ_BBLSPERDAY"  "NET_COND_RATE_ADJ_SM3PERDAY"
#> [31] "GAS_RATE_ADJ_SCFPERDAY"        "GAS_RATE_ADJ_SM3PERDAY"
#> [33] "GAS_RATE_ADJ_MSCFPERDAY"       "NET_OIL_RATE_ADJ_BBLSPERDAY"
#> [35] "NET_OIL_RATE_ADJ_SM3PERDAY"    "LIQUID_RATE_ADJ_BBLSPERDAY"
#> [37] "LIQUID_RATE_ADJ_SM3PERDAY"     "GOR_SCFPERBBL"
#> [39] "Z_FGOR_SCFPERBBL"              "WOR"
#> [41] "Z_IGLR_SCFPERBBL"              "Z_TGLR_SCFPERBBL"
#> [43] "Z_GUF_SCFPERBBL"               "WATERCUT_PCT"
#> [45] "DRY_WET_GAS_RATIO"             "WGR_BBLPERSCF"
#> [47] "CGR_BBLPERSCF"                 "RESULT_NO"

#>  [1] "PRODUCTION_DAY"     "CODE"                "OBJECT_START_DATE"
#>  [4] "OBJECT_END_DATE"    "WELL_CLASS"          "WELL_TYPE"
#>  [7] "PROD_METHOD"        "PUMP_TYPE"           "WELL_METER_FREQ"
#> [10] "ALLOC_FLAG"         "COND_AS_OIL_FLAG"    "CHOKE_UOM"
```

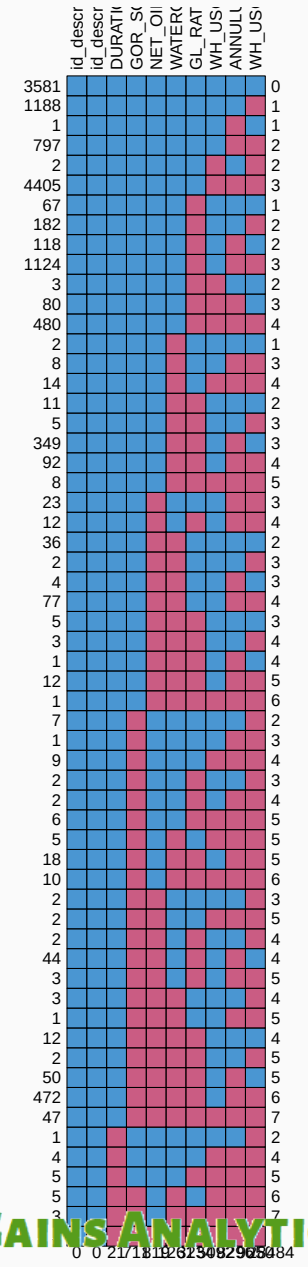
Look for missing values

In this case we are using the statistical package **mice** to fill in for the missing values.

```
library(tidyverse)
library(mice)
library(lattice)
```

```
# plot the missing data in a matrix by variables
md_pattern <- md.pattern(df, rotate.names = TRUE)
```

Look for missing values



Missing values per variable

```
# number of complete cases: no variable value is missing in that observation
sum(complete.cases(sel_numdf))
```

```
#> [1] 3581
```

```
sel_numdf %>%
    skim_to_list() %>%
    .[["numeric"]] %>%
    as.data.frame(sapply(as.numeric)) %>%
    dplyr::select(variable, missing, complete)
```

```
#>                     variable missing complete
#> 1          ANNULUS_PRESS_PSIG    8296     5118
#> 2                DURATION_HRS      21    13393
#> 3           GL_RATE_SM3PERDAY    3234    10180
#> 4               GOR_SCFPERBBL     711    12703
#> 5 NET_OIL_RATE_ADJ_SM3PERDAY     819    12595
#> 6                WATERCUT_PCT    1261    12153
#> 7           WH_USC_PRESS_PSIG    5092     8322
#> 8               WH_USC_TEMP_F    9050     4364
```

How well the fill data fits in