Historical MAL dataset cleanup

Edward Yu

December 07, 2018

Contents

Introduction

From Marek:

If you are interested I have a small project for R, which is very useful. It has to do with history records of MAL. Here is some basic info:

- Goal: Clean up and consolidate dataset to enable easy searching of past melt records
- Tasks:
  - Mostly working with strings removing duplicates (E. Yu, YU, Yu Edward . . .)
  - Removing empty records
  - Missing data
  - Multiple variables in one column

There might be other things to do but I have not spent much time looking at the dataset. We could also pull some basic stats on usage, costs, repeats etc. I don't know your skill level, but it is relatively simple project and I am estimating it would take me about 8 hrs of work. Actual coding, if you know what to use, could be done in less than 1 hour but that requires proficiency in typing and in R.

I just noticed that sand for this year should have all been W410, excel incremented the name by 1 each time. I think I might be adding information about individual tests from this year incrementally as it comes in and since it is only several rows, perhaps you can delete the entire set of rows from this year, if that makes things easier on your end.

Load & peak data

Load packages

```
if(!require(tidyverse)) install.packages("tidyverse")
library(tidyverse)
if(!require(naniar)) install.packages("naniar")
library(naniar) # gg_miss
```

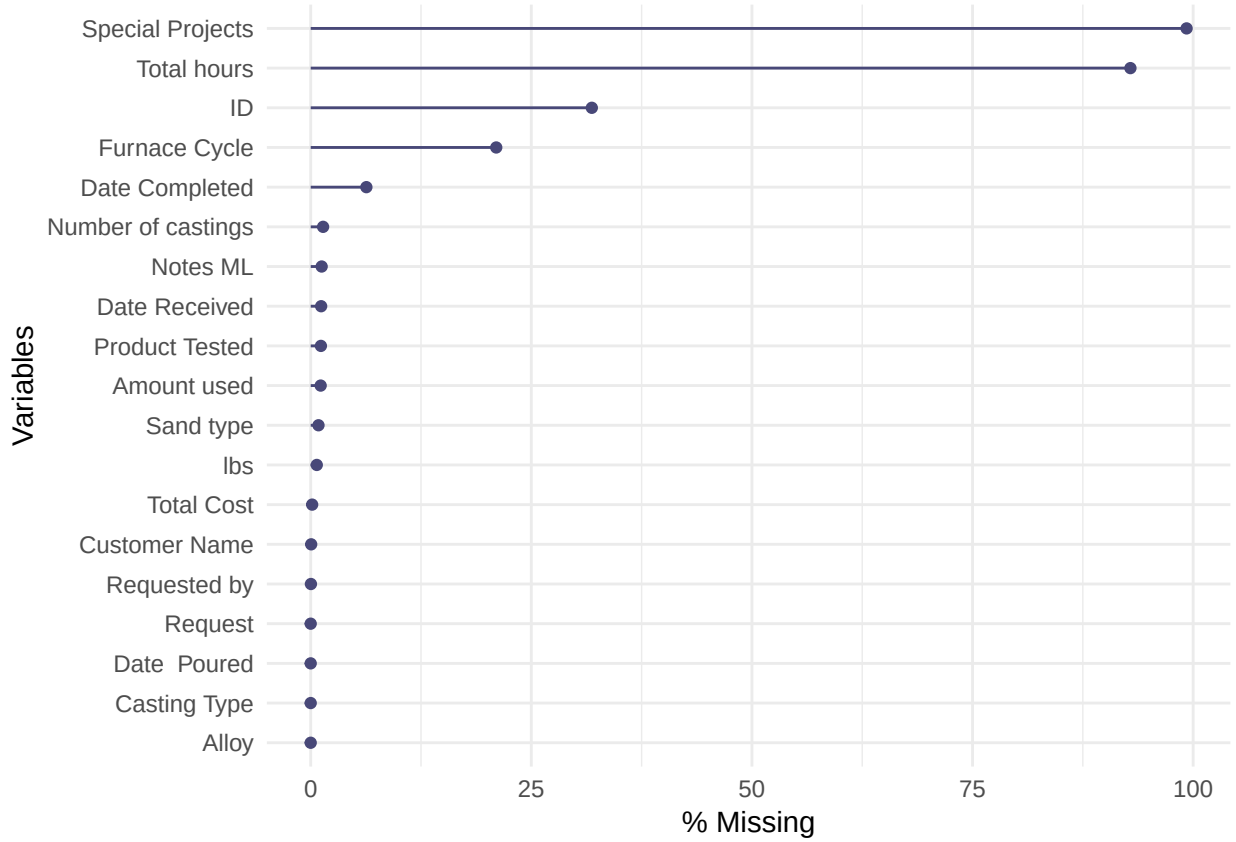
Import data

```
setwd("~/R/historical.MAL")
x <- read_csv("data/History.csv")
glimpse(x)
```

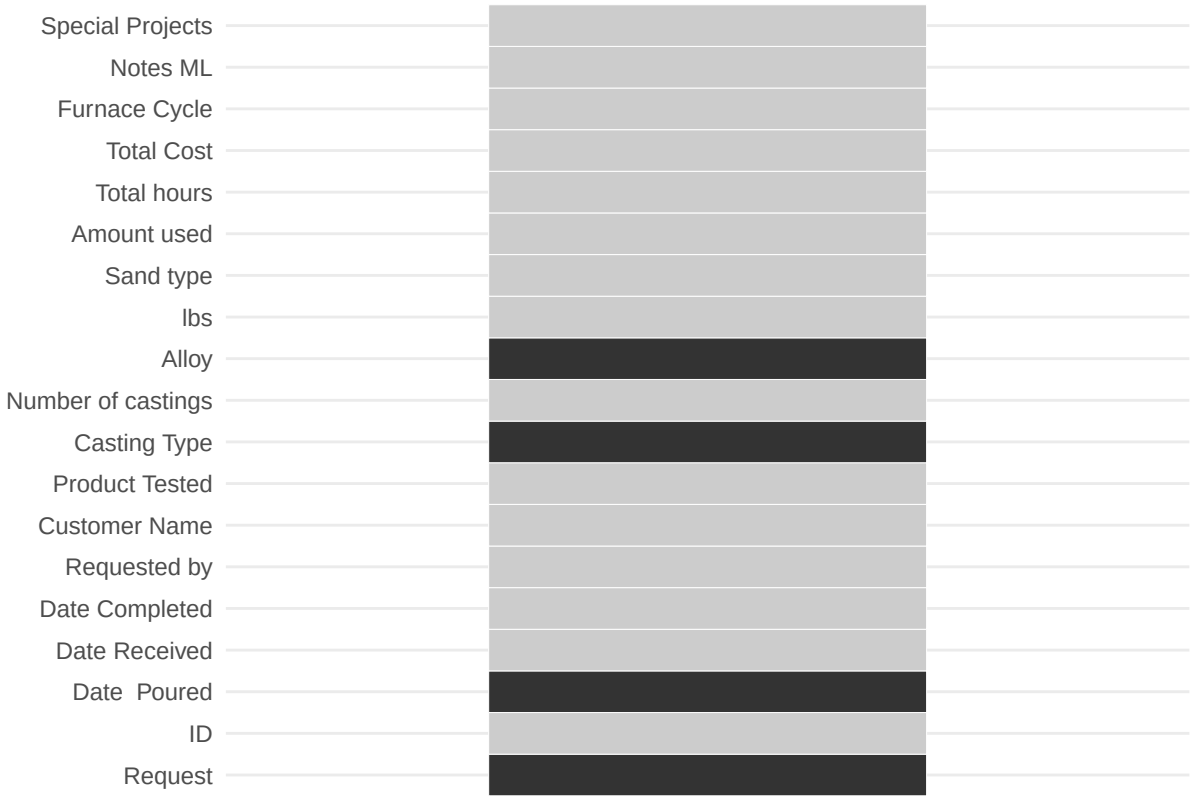
```
## Observations: 3,629
## Variables: 19
## $ Request              <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13...
## $ ID                   <int> NA, NA, NA, NA, 12201, 12194, NA, NA, NA,...
## $ `Date  Poured`       <chr> "1/5/1999", "1/6/1999", "1/7/1999", "1/8/...
## $ `Date Received`      <chr> "1/4/1999", "1/4/1999", "1/4/1999", "1/4/...
## $ `Date Completed`     <chr> "1/13/1999", "1/13/1999", "1/13/1999", "1...
## $ `Requested by`       <chr> "18", "CLINGERMAN,M.", "CLINGERMAN, M.", ...
## $ `Customer Name`      <chr> "TS&D", "TS&D", "TS&D", "TS&D", "BRILLION...
## $ `Product Tested`     <chr> "ISOCURE", "ISOCURE", "ISOCURE", "ISOCURE...
## $ `Casting Type`       <chr> "STEPCONE", "STEPCONE", "EROSION WEDGE", ...
## $ `Number of castings` <int> 8, 8, 8, 8, 3, 1, 8, 10, 8, 4, 10, 8, 2, ...
## $ Alloy                <chr> "GRAY IRON", "GRAY IRON", "GRAY IRON", "G...
## $ lbs                  <int> 250, 250, 600, 600, 90, 90, 160, 30, 20, ...
## $ `Sand type`          <chr> "TECHNISAND 1L-5W", "TECHNISAND 1L-5W", "...
## $ `Amount used`        <int> 840, 840, 1680, 1680, 270, 210, 640, 240,...
## $ `Total hours`        <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ `Total Cost`         <dbl> 1300, 1300, 2210, 2210, 862, 715, 2080, 8...
## $ `Furnace Cycle`      <chr> "W68", "W69", "W70, W71", "W72, W73", "W7...
## $ `Notes ML`           <chr> "TEST NEW BASE RESIN WITH STEPCONE CASTIN...
## $ `Special Projects`   <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
```

Peak missing values

```
gg_miss_var(x, show_pct = T)
```




```
gg_miss_which(x)
```



Outline of actions to take

Rename variables to be all lowercase with no spaces. Seems the most important variables are casting type and allow type, as these are the only with zero missing values.

- **Request**: should have 3,629 levels
- **Notes ML**: n/a
- **Date poured**: fill missing values
- **Date completed**: fill missing values, perhaps create new column calculating days to complete from date received/completed
- **Date received**: fill missing values, for some reason there are less dates received than dates completed
- **Furnace cycle**: need to come up with new way to ID new lining and cycles
- **ID**: not even sure what this refers to
- **Customer name**: fill missing values, will require some renaming/matching
- **Casting type**: n/a
- **Requested by**: fill missing values, will require some renaming/matching
- **Total cost**: fill missing values, perhaps determine how it is calculated to automate the calculation
- **Product tested**: fill missing values, will require some renaming/matching
- **Amount used**: unsure what amount this is talking about
- **Sand type**: fill missing values, will require some renaming/matching
- **lbs**: fill missing values
- **Alloy**: n/a
- **Number of castings**: some n/a values, fill in with rounded averages
- **Total hours**: many n/a values, should be calculated automatically based on number of castings, casting type, etc

- **Special projects**: most values are missing, unsure of importance of this field, should likely merge with comments or remove entirely

We have many missing datapoints, fields that aren't inuitive, some useless fields, fields that need added, etc. We'll start with the most simple and move on.

Cleaning

Rename columns

Convert column names to lower case, replace spaces with periods.

```
# names <- colnames(x) %>%
#   tolower()
names <- tolower(colnames(x))     # convert to lowercase
names <- gsub("  ", " ", names)   # remove double spaces
names <- gsub(" ", "\\.", names) # replace space with .
colnames(x) <- names
colnames(x)
```

```
##  [1] "request"            "id"                 "date.poured"
##  [4] "date.received"      "date.completed"     "requested.by"
##  [7] "customer.name"      "product.tested"     "casting.type"
## [10] "number.of.castings" "alloy"              "lbs"
## [13] "sand.type"          "amount.used"        "total.hours"
## [16] "total.cost"         "furnace.cycle"      "notes.ml"
## [19] "special.projects"
```

$request

There is a duplicate entry somewhere.

```
which(duplicated(x$request)==TRUE)
```

```
## [1] 3611
```

```
as.data.frame(t(x[3609:3611,]))
```

	V1	V2	V3
request	3609	3610	3610
id	NA	NA	NA
date.poured	9/13/2017	unknown	6/28/2018
date.received	NA	NA	NA
date.completed	NA	NA	NA
requested.by	VIVAS	unknown	unknown
customer.name	ASK	unknown	ASK
product.tested	COATINGS	unknown	unknown
casting.type	STEP CONES	unknown	unknown
number.of.castings	NA	NA	NA
alloy	GRAY IRON	unknown	Aluminum
lbs	NA	NA	NA
sand.type	NA	unknown	W410
amount.used	NA	NA	NA

	V1	V2	V3
total.hours	NA	unknown	unknown
total.cost	0	NA	0
furnace.cycle	S1	S2	S3
notes.ml	NA	NA	NA
special.projects	NA	NA	NA

The first entry appears to have been made in error until we see the furnace cycle was incremented. Probably shouldn't remove, will simply re-assign all request variables to equal row numbers.

```
x <- x %>%
  mutate(request = seq(1:nrow(x)))
```