Historical MAL dataset cleanup

Edward Yu

December 07, 2018

Contents

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.

Preliminary data peaking

Load packages

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

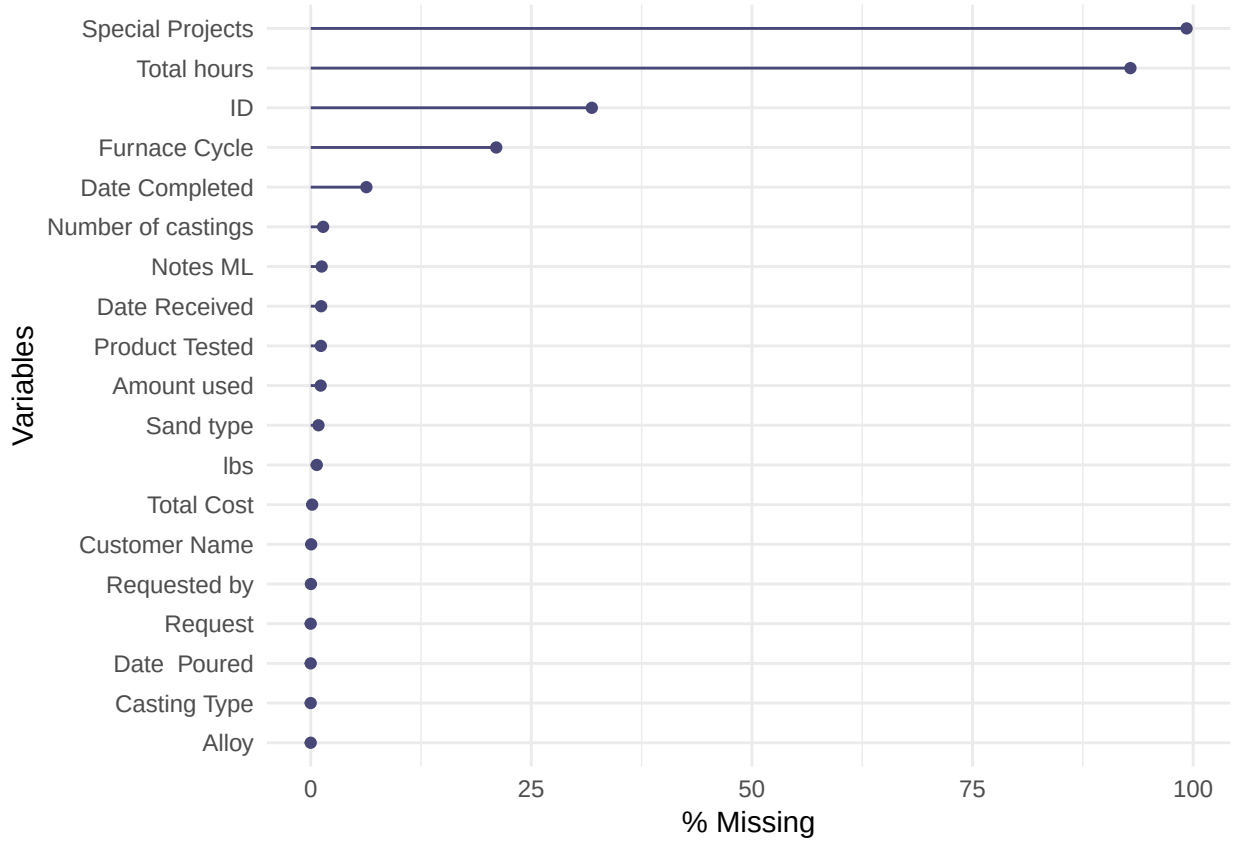
Import data

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

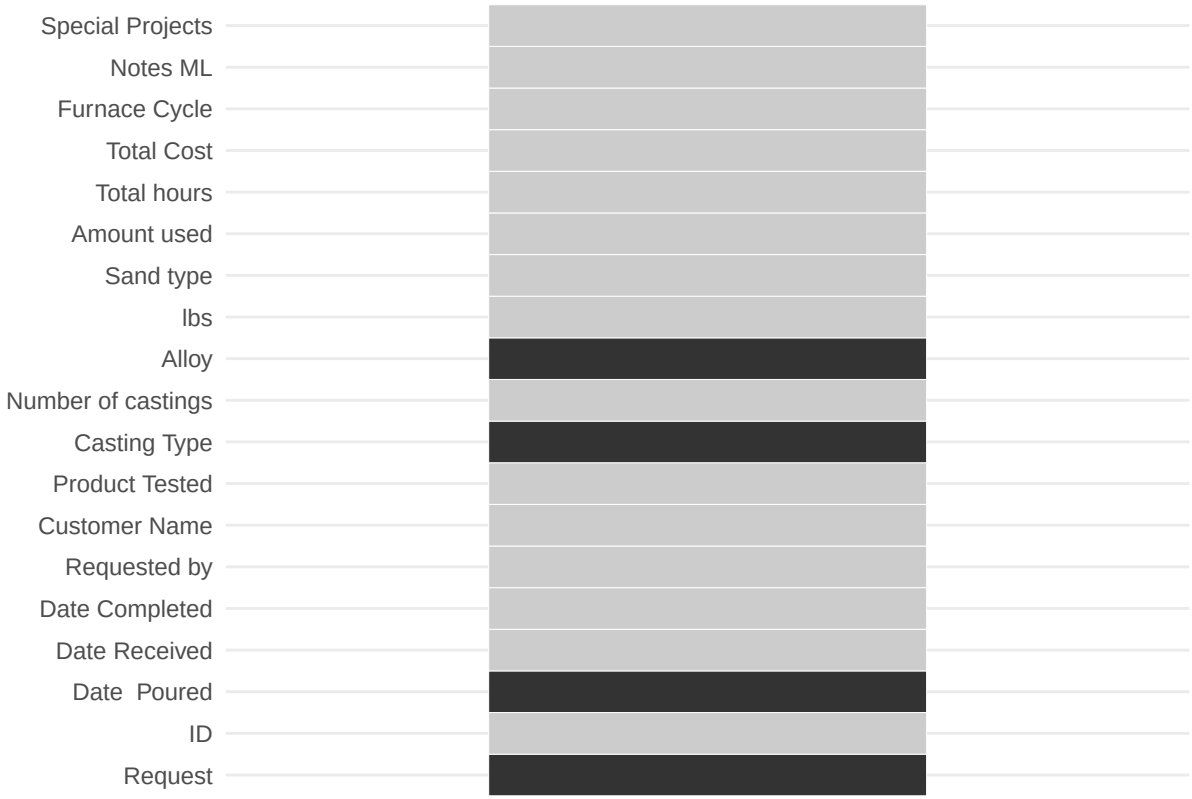
```
## Observations: 3,629
## Variables: 19
## $ Request              <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13...
## $ ID                   <int> NA, NA, NA, NA, 12201, 12194, NA, NA, NA,...
## $ `Date  Poured`       <chr> "1/5/1999", "1/6/1999", "1/7/1999", "1/8/...
## $ `Date Received`      <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)
```



It appears most variables are missing data. We'll tackle them from top to bottom: 1. Special projects 2. Notes ML 3. Furnace cycle 4