Data cleaning

Kristy Robledo

04/06/2020

## Read in data

trt<-read\_sas("T:\\ECHO\\Clean\\trt.sas7bdat") %>%  
 clean\_names() %>%  
 remove\_constant()  
  
glimpse(trt)

## Rows: 3,788  
## Columns: 19  
## $ studyeventoid <chr> "SE\_ECH\_CYC1", "SE\_ECH\_CYC1", "SE\_ECH\_CYC1", "S...  
## $ studyeventrepeatkey <chr> "1", "1", "2", "3", "4", "1", "1", "1", "2", "3...  
## $ subjectkey <chr> "SS\_ECH0001", "SS\_ECH0001", "SS\_ECH0001", "SS\_E...  
## $ study\_subject\_id <chr> "ECH0001", "ECH0001", "ECH0001", "ECH0001", "EC...  
## $ secondary\_id <chr> "D-H", "D-H", "D-H", "D-H", "D-H", "D-H", "D-H"...  
## $ studyoid <chr> "S\_ECH61459", "S\_ECH61459", "S\_ECH61459", "S\_EC...  
## $ metadataversionoid <chr> "v1.0.0-S\_ECH61459", "v1.0.0-S\_ECH61459", "v1.0...  
## $ formoid <chr> "F\_ECH\_TRT\_100", "F\_ECH\_TRT\_100", "F\_ECH\_TRT\_10...  
## $ itemgrouprepeatkey <chr> "1", "2", "1", "1", "1", "1", "2", "3", "1", "1...  
## $ exttreat <dbl> 2, 1, 1, 2, 2, 1, 2, NA, 2, 2, 2, 1, 2, 2, 2, 1...  
## $ extfdat <date> 2015-07-03, 2015-07-04, 2015-07-04, 2015-07-10...  
## $ extdose <dbl> 162, 710, 710, 162, 162, 710, 162, NA, 162, 162...  
## $ extmodt <dbl> 0, 0, 0, 0, 0, 0, 0, NA, 0, 0, 0, 0, 0, 0, 0, 0...  
## $ extmodr <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...  
## $ extmodro <chr> "", "", "", "", "", "", "", "", "", "", "", "",...  
## $ extdelt <dbl> 0, 2, 2, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 2, 2,...  
## $ extdelr <dbl> NA, 9, 9, NA, NA, NA, NA, NA, NA, 9, NA, NA, NA...  
## $ extdelro <chr> "", "reaction to paclitaxel", "allergic reactio...  
## $ extrto <chr> "", "", "", "", "", "", "", "", "", "", "", "",...

## Patients summmary

trt %>%  
 tabyl(subjectkey) ->pids  
  
nrow(pids)

## [1] 222

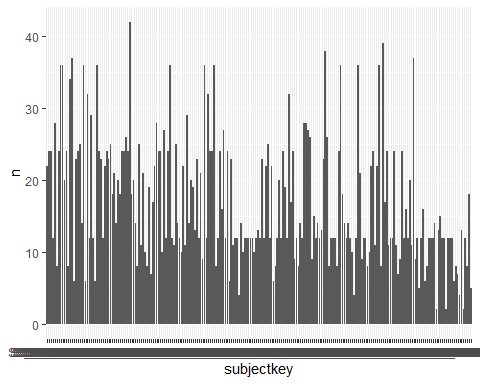
min(pids$n)

## [1] 2

max(pids$n)

## [1] 42

pids %>%  
 ggplot(aes(x=subjectkey, y=n)) +  
 geom\_bar(stat = "identity")



## Drugs

Types of drugs

trt %>%  
 tabyl(exttreat)

## exttreat n percent valid\_percent  
## 1 1464 0.386483633 0.387301587  
## 2 2088 0.551214361 0.552380952  
## 3 96 0.025343189 0.025396825  
## 4 8 0.002111932 0.002116402  
## 5 28 0.007391763 0.007407407  
## 9 96 0.025343189 0.025396825  
## NA 8 0.002111932 NA

Patients with no drug name

trt %>%  
 filter(is.na(exttreat)) %>%  
 select(subjectkey, exttreat, extfdat)

## # A tibble: 8 x 3  
## subjectkey exttreat extfdat   
## <chr> <dbl> <date>   
## 1 SS\_ECH0001 NA NA   
## 2 SS\_ECH0014 NA 2016-06-20  
## 3 SS\_ECH0014 NA 2016-06-20  
## 4 SS\_ECH0014 NA 2016-07-11  
## 5 SS\_ECH0054 NA 2017-05-02  
## 6 SS\_ECH0060 NA 2017-06-14  
## 7 SS\_ECH0118 NA NA   
## 8 SS\_ECH0118 NA NA

## Dates

Find missing dates

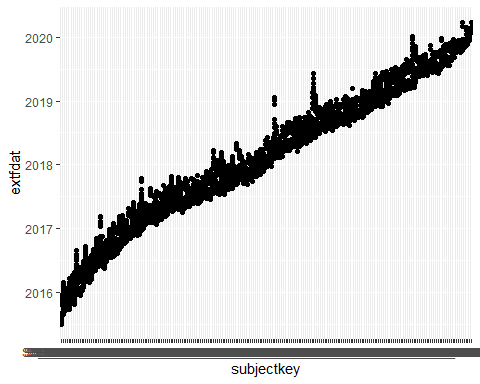
trt %>%  
 filter(is.na(extfdat)) %>%  
 select(subjectkey, exttreat, extfdat)

## # A tibble: 4 x 3  
## subjectkey exttreat extfdat   
## <chr> <dbl> <date>   
## 1 SS\_ECH0001 NA NA   
## 2 SS\_ECH0118 NA NA   
## 3 SS\_ECH0118 NA NA   
## 4 SS\_ECH0200 3 NA

Plot dates to check

trt %>%  
 ggplot(aes(x=subjectkey, y=extfdat)) +  
 geom\_point()

## Warning: Removed 4 rows containing missing values (geom\_point).



## Doses

Find missing doses

trt %>%  
 filter(is.na(extdose)) %>%  
 select(subjectkey, exttreat, extdose)

## # A tibble: 13 x 3  
## subjectkey exttreat extdose  
## <chr> <dbl> <dbl>  
## 1 SS\_ECH0001 NA NA  
## 2 SS\_ECH0038 1 NA  
## 3 SS\_ECH0038 2 NA  
## 4 SS\_ECH0038 1 NA  
## 5 SS\_ECH0038 2 NA  
## 6 SS\_ECH0039 2 NA  
## 7 SS\_ECH0039 2 NA  
## 8 SS\_ECH0151 1 NA  
## 9 SS\_ECH0151 2 NA  
## 10 SS\_ECH0154 2 NA  
## 11 SS\_ECH0159 2 NA  
## 12 SS\_ECH0200 3 NA  
## 13 SS\_ECH0046 2 NA

Summarise doses by drug type

trt %>%  
 filter(is.na(extdose)==FALSE) %>%  
 group\_by(exttreat) %>%  
 summarise(n=n(),   
 mean(extdose),   
 median=median(extdose),  
 min=min(extdose),   
 max=max(extdose))

## `summarise()` ungrouping output (override with `.groups` argument)

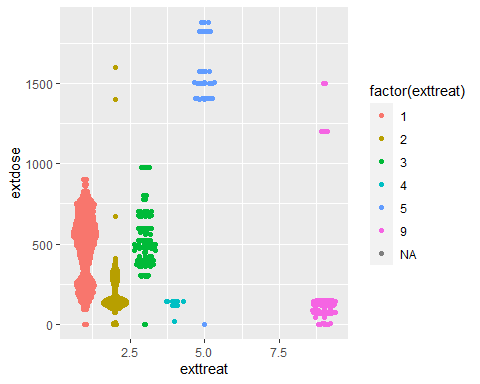
## # A tibble: 7 x 6  
## exttreat n `mean(extdose)` median min max  
## <dbl> <int> <dbl> <dbl> <dbl> <dbl>  
## 1 1 1461 490. 530 0 900.  
## 2 2 2080 179. 144 0 1600   
## 3 3 95 552. 500 0 975   
## 4 4 8 120. 140 17 145   
## 5 5 28 1527. 1502. 0 1881   
## 6 9 96 406. 126 0 1500   
## 7 NA 7 150. 135 0 490

Plot with a bit of noise - I prefer beeswarm to jitter!

trt %>%  
 ggplot(aes(x=exttreat, y=extdose, colour=factor(exttreat))) +  
 geom\_point()+  
 geom\_quasirandom()

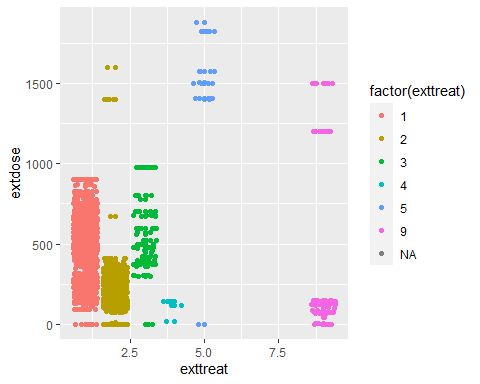
## Warning: Removed 20 rows containing missing values (position\_quasirandom).

## Warning: Removed 20 rows containing missing values (geom\_point).



trt %>%  
 ggplot(aes(x=exttreat, y=extdose, colour=factor(exttreat))) +  
 geom\_point()+  
 geom\_jitter()

## Warning: Removed 20 rows containing missing values (geom\_point).  
  
## Warning: Removed 20 rows containing missing values (geom\_point).



## long

Make your data ‘longer’

trt %>%  
 select(exttreat, extdose) %>%  
 pivot\_longer(everything(), names\_to="column", values\_to="data") ->long

## wide

This also performs a function at the same time!

trt %>%  
 select(subjectkey, exttreat, extdose) %>%  
 pivot\_wider(names\_from = exttreat,   
 values\_from = extdose,   
 names\_prefix="drug",  
 values\_fn=mean) ->meandoses

## Recode factors levels

glimpse(trt)

## Rows: 3,788  
## Columns: 19  
## $ studyeventoid <chr> "SE\_ECH\_CYC1", "SE\_ECH\_CYC1", "SE\_ECH\_CYC1", "S...  
## $ studyeventrepeatkey <chr> "1", "1", "2", "3", "4", "1", "1", "1", "2", "3...  
## $ subjectkey <chr> "SS\_ECH0001", "SS\_ECH0001", "SS\_ECH0001", "SS\_E...  
## $ study\_subject\_id <chr> "ECH0001", "ECH0001", "ECH0001", "ECH0001", "EC...  
## $ secondary\_id <chr> "D-H", "D-H", "D-H", "D-H", "D-H", "D-H", "D-H"...  
## $ studyoid <chr> "S\_ECH61459", "S\_ECH61459", "S\_ECH61459", "S\_EC...  
## $ metadataversionoid <chr> "v1.0.0-S\_ECH61459", "v1.0.0-S\_ECH61459", "v1.0...  
## $ formoid <chr> "F\_ECH\_TRT\_100", "F\_ECH\_TRT\_100", "F\_ECH\_TRT\_10...  
## $ itemgrouprepeatkey <chr> "1", "2", "1", "1", "1", "1", "2", "3", "1", "1...  
## $ exttreat <dbl> 2, 1, 1, 2, 2, 1, 2, NA, 2, 2, 2, 1, 2, 2, 2, 1...  
## $ extfdat <date> 2015-07-03, 2015-07-04, 2015-07-04, 2015-07-10...  
## $ extdose <dbl> 162, 710, 710, 162, 162, 710, 162, NA, 162, 162...  
## $ extmodt <dbl> 0, 0, 0, 0, 0, 0, 0, NA, 0, 0, 0, 0, 0, 0, 0, 0...  
## $ extmodr <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...  
## $ extmodro <chr> "", "", "", "", "", "", "", "", "", "", "", "",...  
## $ extdelt <dbl> 0, 2, 2, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 2, 2,...  
## $ extdelr <dbl> NA, 9, 9, NA, NA, NA, NA, NA, NA, 9, NA, NA, NA...  
## $ extdelro <chr> "", "reaction to paclitaxel", "allergic reactio...  
## $ extrto <chr> "", "", "", "", "", "", "", "", "", "", "", "",...

trt %>%  
 mutate(drug = recode(exttreat, `1`="Drugname", `2`="Drugname2")) -> newtrt

## Warning: Unreplaced values treated as NA as .x is not compatible. Please specify  
## replacements exhaustively or supply .default

newtrt %>%  
 tabyl(drug, exttreat)

## drug 1 2 3 4 5 9 NA\_  
## Drugname 1464 0 0 0 0 0 0  
## Drugname2 0 2088 0 0 0 0 0  
## <NA> 0 0 96 8 28 96 8

Change continous to categorical

trt %>%  
 mutate(catdose = ifelse(extdose>=100,"High", "Low")) ->catdat

## Change case of character variable

toupper and tolower functions

trt %>%  
 mutate(newreason = toupper(extdelro)) ->uppercase

## Excel error on read in of dataset

df<-read\_xlsx("Y:\\Statistics\\Kristy\\for tony\\Laima\\Data\\PAVANI CADMS 20171031.xlsx", sheet = "FinalData")

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting numeric in AI8224 / R8224C35: got '?2006'

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting numeric in AI8234 / R8234C35: got '?2006'