library(readr)  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ ggplot2 3.3.6 ✔ dplyr 1.0.10  
## ✔ tibble 3.1.8 ✔ stringr 1.4.1   
## ✔ tidyr 1.2.1 ✔ forcats 0.5.2   
## ✔ purrr 0.3.5   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(skimr)

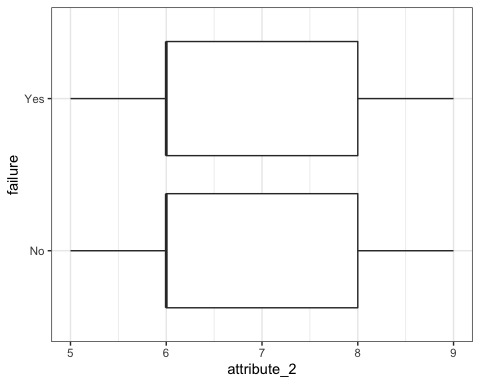
train <- read\_csv("train.csv")

## Rows: 26570 Columns: 26  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (4): product\_code, attribute\_0, attribute\_1, failure  
## dbl (22): id, loading, attribute\_2, attribute\_3, measurement\_0, measurement\_...  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

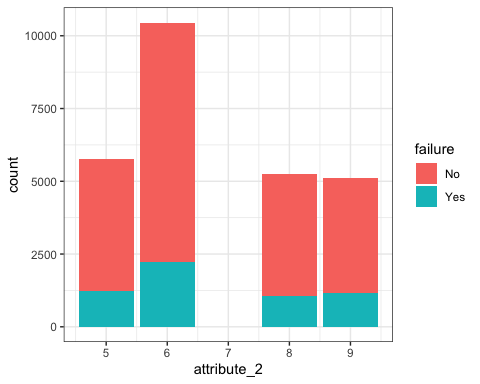
#str(train)  
#summary(train)  
#skim(train)

train = train %>% mutate(attribute\_0 = as\_factor(attribute\_0)) %>%  
 mutate(attribute\_1 = as\_factor(attribute\_1))

ggplot(train, aes(x = attribute\_2, y = failure)) +   
 geom\_boxplot() +   
  
 theme\_bw()



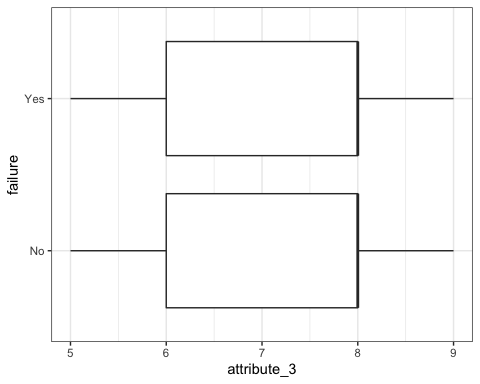
ggplot(train, aes(x=attribute\_2, fill = failure)) + geom\_bar() + theme\_bw()



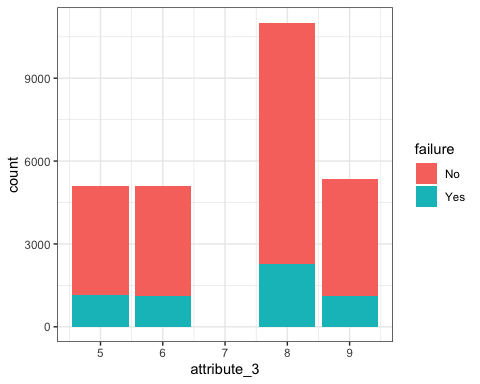
t1 = table(train$failure, train$attribute\_2) #create a table object  
prop.table(t1, margin = 2 ) #crosstab with proportions

##   
## 5 6 8 9  
## No 0.7883781 0.7878527 0.7996190 0.7727451  
## Yes 0.2116219 0.2121473 0.2003810 0.2272549

ggplot(train, aes(x = attribute\_3, y = failure)) +   
 geom\_boxplot() +   
  
 theme\_bw()



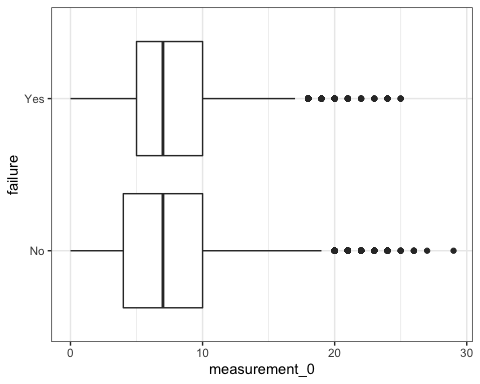
ggplot(train, aes(x=attribute\_3, fill = failure)) + geom\_bar() + theme\_bw()



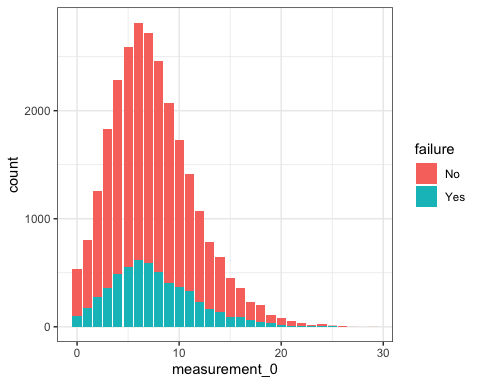
t2 = table(train$failure, train$attribute\_3) #create a table object  
prop.table(t2, margin = 2 ) #crosstab with proportions

##   
## 5 6 8 9  
## No 0.7727451 0.7824726 0.7937358 0.7930002  
## Yes 0.2272549 0.2175274 0.2062642 0.2069998

ggplot(train, aes(x = measurement\_0, y = failure)) +   
 geom\_boxplot() +   
  
 theme\_bw()



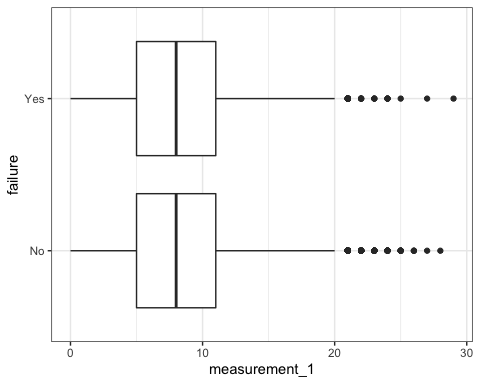
ggplot(train, aes(x=measurement\_0, fill = failure)) + geom\_bar() + theme\_bw()



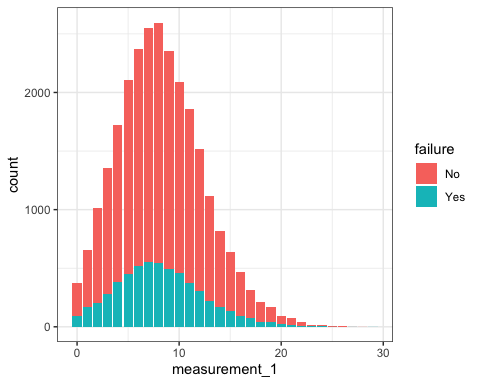
t3 = table(train$failure, train$measurement\_0) #create a table object  
prop.table(t3, margin = 2 ) #crosstab with proportions

##   
## 0 1 2 3 4 5 6  
## No 0.8159851 0.7834793 0.7800797 0.8018610 0.7864035 0.7876209 0.7819602  
## Yes 0.1840149 0.2165207 0.2199203 0.1981390 0.2135965 0.2123791 0.2180398  
##   
## 7 8 9 10 11 12 13  
## No 0.7844796 0.7928368 0.8037699 0.7884058 0.7682495 0.7877095 0.7890724  
## Yes 0.2155204 0.2071632 0.1962301 0.2115942 0.2317505 0.2122905 0.2109276  
##   
## 14 15 16 17 18 19 20  
## No 0.7839506 0.8004386 0.7520891 0.7423581 0.7839196 0.6991150 0.7848101  
## Yes 0.2160494 0.1995614 0.2479109 0.2576419 0.2160804 0.3008850 0.2151899  
##   
## 21 22 23 24 25 26 27  
## No 0.8113208 0.8684211 0.7647059 0.7083333 0.5714286 1.0000000 1.0000000  
## Yes 0.1886792 0.1315789 0.2352941 0.2916667 0.4285714 0.0000000 0.0000000  
##   
## 29  
## No 1.0000000  
## Yes 0.0000000

ggplot(train, aes(x = measurement\_1, y = failure)) +   
 geom\_boxplot() +   
  
 theme\_bw()



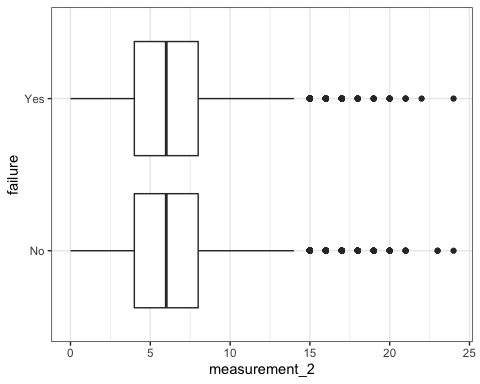
ggplot(train, aes(x=measurement\_1, fill = failure)) + geom\_bar() + theme\_bw()



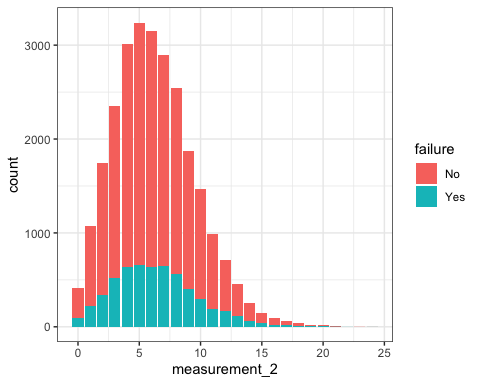
t4 = table(train$failure, train$measurement\_1) #create a table object  
prop.table(t4, margin = 2 ) #crosstab with proportions

##   
## 0 1 2 3 4 5 6  
## No 0.7671958 0.7411945 0.8031496 0.7930781 0.7766821 0.7864955 0.7827368  
## Yes 0.2328042 0.2588055 0.1968504 0.2069219 0.2233179 0.2135045 0.2172632  
##   
## 7 8 9 10 11 12 13  
## No 0.7825235 0.7893724 0.7899873 0.7793202 0.7973046 0.8000000 0.8057296  
## Yes 0.2174765 0.2106276 0.2100127 0.2206798 0.2026954 0.2000000 0.1942704  
##   
## 14 15 16 17 18 19 20  
## No 0.7892157 0.7927786 0.8004246 0.7634069 0.8018868 0.7810651 0.7659574  
## Yes 0.2107843 0.2072214 0.1995754 0.2365931 0.1981132 0.2189349 0.2340426  
##   
## 21 22 23 24 25 26 27  
## No 0.7692308 0.8750000 0.8235294 0.7333333 0.8750000 1.0000000 0.5000000  
## Yes 0.2307692 0.1250000 0.1764706 0.2666667 0.1250000 0.0000000 0.5000000  
##   
## 28 29  
## No 1.0000000 0.0000000  
## Yes 0.0000000 1.0000000

ggplot(train, aes(x = measurement\_2, y = failure)) +   
 geom\_boxplot() +   
  
 theme\_bw()



ggplot(train, aes(x=measurement\_2, fill = failure)) + geom\_bar() + theme\_bw()



t5 = table(train$failure, train$measurement\_2) #create a table object  
prop.table(t5, margin = 2 ) #crosstab with proportions

##   
## 0 1 2 3 4 5 6  
## No 0.7799043 0.7973856 0.8044597 0.7808511 0.7871846 0.7960506 0.7984152  
## Yes 0.2200957 0.2026144 0.1955403 0.2191489 0.2128154 0.2039494 0.2015848  
##   
## 7 8 9 10 11 12 13  
## No 0.7752420 0.7793944 0.7856000 0.7972696 0.8030457 0.7717087 0.7522124  
## Yes 0.2247580 0.2206056 0.2144000 0.2027304 0.1969543 0.2282913 0.2477876  
##   
## 14 15 16 17 18 19 20  
## No 0.7713178 0.7414966 0.7472527 0.7014925 0.7567568 0.6842105 0.6250000  
## Yes 0.2286822 0.2585034 0.2527473 0.2985075 0.2432432 0.3157895 0.3750000  
##   
## 21 22 23 24  
## No 0.6250000 0.0000000 1.0000000 0.5000000  
## Yes 0.3750000 1.0000000 0.0000000 0.5000000

library(mice)

##   
## Attaching package: 'mice'

## The following object is masked from 'package:stats':  
##   
## filter

## The following objects are masked from 'package:base':  
##   
## cbind, rbind

set.seed(1234) #sets seed for random number generator  
imp\_measurement\_3 = mice(train, m=5, method='pmm', printFlag=FALSE)

## Warning: Number of logged events: 402

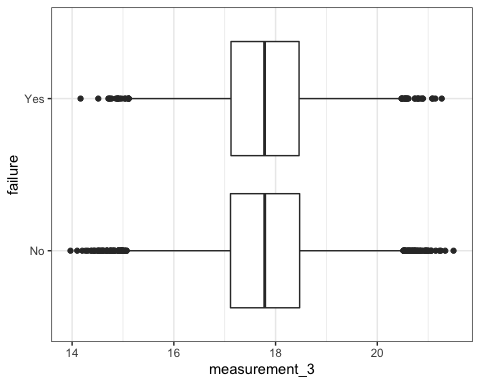
#m is the number of imputations, 5 is a reasonable value as a default  
#pmm is "predictive mean matching" = imputation method for numeric data  
#printFlag reduces amount of output  
summary(imp\_measurement\_3)

## Class: mids  
## Number of multiple imputations: 5   
## Imputation methods:  
## id product\_code loading attribute\_0 attribute\_1   
## "" "" "pmm" "" ""   
## attribute\_2 attribute\_3 measurement\_0 measurement\_1 measurement\_2   
## "" "" "" "" ""   
## measurement\_3 measurement\_4 measurement\_5 measurement\_6 measurement\_7   
## "pmm" "pmm" "pmm" "pmm" "pmm"   
## measurement\_8 measurement\_9 measurement\_10 measurement\_11 measurement\_12   
## "pmm" "pmm" "pmm" "pmm" "pmm"   
## measurement\_13 measurement\_14 measurement\_15 measurement\_16 measurement\_17   
## "pmm" "pmm" "pmm" "pmm" "pmm"   
## failure   
## ""   
## PredictorMatrix:  
## id product\_code loading attribute\_0 attribute\_1 attribute\_2  
## id 0 0 1 1 1 1  
## product\_code 1 0 1 1 1 1  
## loading 1 0 0 1 1 1  
## attribute\_0 1 0 1 0 1 1  
## attribute\_1 1 0 1 1 0 1  
## attribute\_2 1 0 1 1 1 0  
## attribute\_3 measurement\_0 measurement\_1 measurement\_2  
## id 1 1 1 1  
## product\_code 1 1 1 1  
## loading 1 1 1 1  
## attribute\_0 1 1 1 1  
## attribute\_1 1 1 1 1  
## attribute\_2 1 1 1 1  
## measurement\_3 measurement\_4 measurement\_5 measurement\_6  
## id 1 1 1 1  
## product\_code 1 1 1 1  
## loading 1 1 1 1  
## attribute\_0 1 1 1 1  
## attribute\_1 1 1 1 1  
## attribute\_2 1 1 1 1  
## measurement\_7 measurement\_8 measurement\_9 measurement\_10  
## id 1 1 1 1  
## product\_code 1 1 1 1  
## loading 1 1 1 1  
## attribute\_0 1 1 1 1  
## attribute\_1 1 1 1 1  
## attribute\_2 1 1 1 1  
## measurement\_11 measurement\_12 measurement\_13 measurement\_14  
## id 1 1 1 1  
## product\_code 1 1 1 1  
## loading 1 1 1 1  
## attribute\_0 1 1 1 1  
## attribute\_1 1 1 1 1  
## attribute\_2 1 1 1 1  
## measurement\_15 measurement\_16 measurement\_17 failure  
## id 1 1 1 0  
## product\_code 1 1 1 0  
## loading 1 1 1 0  
## attribute\_0 1 1 1 0  
## attribute\_1 1 1 1 0  
## attribute\_2 1 1 1 0  
## Number of logged events: 402   
## it im dep meth out  
## 1 0 0 constant product\_code  
## 2 0 0 constant failure  
## 3 1 1 loading pmm attribute\_3  
## 4 1 1 measurement\_3 pmm attribute\_3  
## 5 1 1 measurement\_4 pmm attribute\_3  
## 6 1 1 measurement\_5 pmm attribute\_3

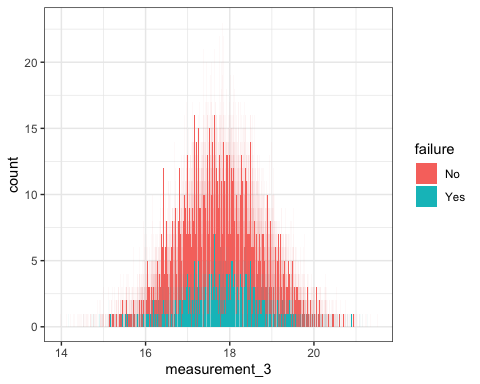
train = complete(imp\_measurement\_3)   
summary(train)

## id product\_code loading attribute\_0   
## Min. : 0 Length:26570 Min. : 33.16 material\_7:21320   
## 1st Qu.: 6642 Class :character 1st Qu.: 99.95 material\_5: 5250   
## Median :13284 Mode :character Median :122.36   
## Mean :13284 Mean :127.81   
## 3rd Qu.:19927 3rd Qu.:149.15   
## Max. :26569 Max. :385.86   
## attribute\_1 attribute\_2 attribute\_3 measurement\_0   
## material\_8:10865 Min. :5.000 Min. :5.00 Min. : 0.000   
## material\_5:10362 1st Qu.:6.000 1st Qu.:6.00 1st Qu.: 4.000   
## material\_6: 5343 Median :6.000 Median :8.00 Median : 7.000   
## Mean :6.754 Mean :7.24 Mean : 7.416   
## 3rd Qu.:8.000 3rd Qu.:8.00 3rd Qu.:10.000   
## Max. :9.000 Max. :9.00 Max. :29.000   
## measurement\_1 measurement\_2 measurement\_3 measurement\_4   
## Min. : 0.000 Min. : 0.000 Min. :13.97 Min. : 8.008   
## 1st Qu.: 5.000 1st Qu.: 4.000 1st Qu.:17.12 1st Qu.:11.054   
## Median : 8.000 Median : 6.000 Median :17.79 Median :11.734   
## Mean : 8.233 Mean : 6.257 Mean :17.79 Mean :11.733   
## 3rd Qu.:11.000 3rd Qu.: 8.000 3rd Qu.:18.47 3rd Qu.:12.412   
## Max. :29.000 Max. :24.000 Max. :21.50 Max. :16.484   
## measurement\_5 measurement\_6 measurement\_7 measurement\_8   
## Min. :12.07 Min. :12.71 Min. : 7.968 Min. :15.22   
## 1st Qu.:16.44 1st Qu.:16.84 1st Qu.:11.048 1st Qu.:18.34   
## Median :17.13 Median :17.51 Median :11.713 Median :19.02   
## Mean :17.13 Mean :17.51 Mean :11.718 Mean :19.02   
## 3rd Qu.:17.81 3rd Qu.:18.18 3rd Qu.:12.392 3rd Qu.:19.70   
## Max. :21.43 Max. :21.54 Max. :15.419 Max. :23.81   
## measurement\_9 measurement\_10 measurement\_11 measurement\_12   
## Min. : 7.537 Min. : 9.323 Min. :12.46 Min. : 5.167   
## 1st Qu.:10.757 1st Qu.:15.208 1st Qu.:18.17 1st Qu.:10.702   
## Median :11.432 Median :16.125 Median :19.21 Median :11.719   
## Mean :11.431 Mean :16.118 Mean :19.17 Mean :11.702   
## 3rd Qu.:12.103 3rd Qu.:17.026 3rd Qu.:20.20 3rd Qu.:12.708   
## Max. :15.412 Max. :22.479 Max. :25.64 Max. :17.663   
## measurement\_13 measurement\_14 measurement\_15 measurement\_16   
## Min. :10.89 Min. : 9.14 Min. : 9.104 Min. : 9.701   
## 1st Qu.:14.89 1st Qu.:15.06 1st Qu.:13.958 1st Qu.:15.270   
## Median :15.63 Median :16.04 Median :14.973 Median :16.438   
## Mean :15.65 Mean :16.05 Mean :15.001 Mean :16.464   
## 3rd Qu.:16.37 3rd Qu.:17.08 3rd Qu.:16.028 3rd Qu.:17.629   
## Max. :22.71 Max. :22.30 Max. :21.626 Max. :24.094   
## measurement\_17 failure   
## Min. : 196.8 Length:26570   
## 1st Qu.: 618.9 Class :character   
## Median : 700.8 Mode :character   
## Mean : 701.2   
## 3rd Qu.: 784.1   
## Max. :1312.8

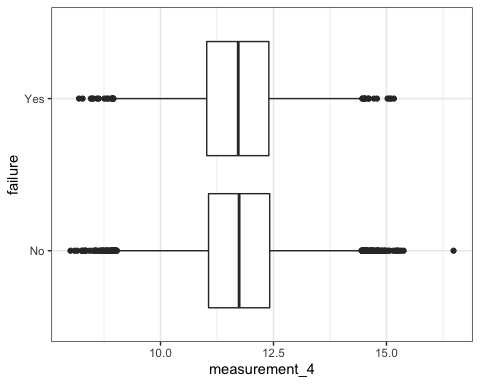
ggplot(train, aes(x = measurement\_3, y = failure)) +   
 geom\_boxplot() +   
  
 theme\_bw()



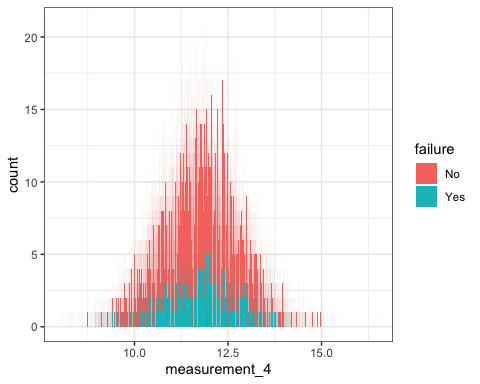
ggplot(train, aes(x=measurement\_3, fill = failure)) + geom\_bar() + theme\_bw()



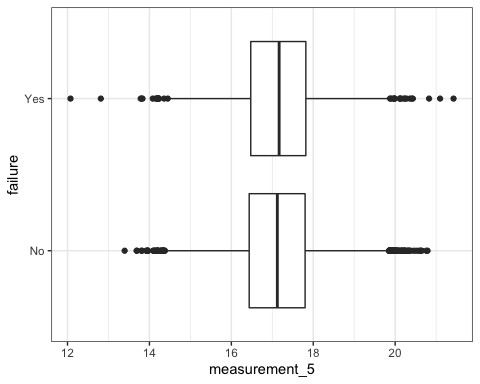
ggplot(train, aes(x = measurement\_4, y = failure)) +   
 geom\_boxplot() +   
  
 theme\_bw()



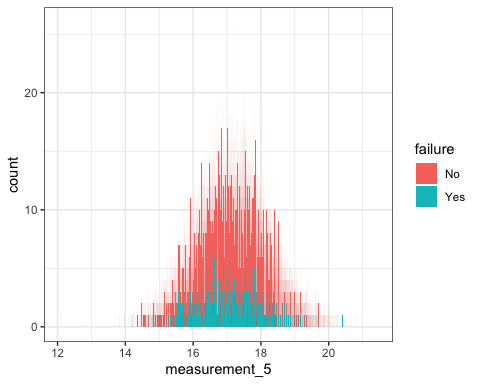
ggplot(train, aes(x=measurement\_4, fill = failure)) + geom\_bar() + theme\_bw()



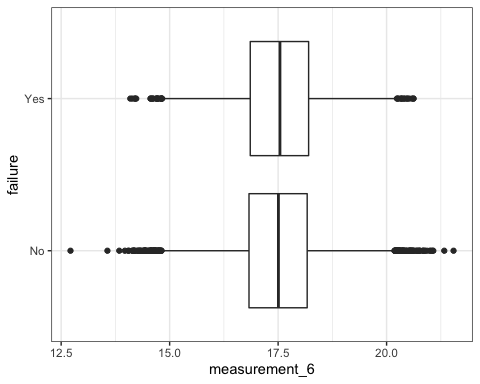
ggplot(train, aes(x = measurement\_5, y = failure)) +   
 geom\_boxplot() +   
  
 theme\_bw()



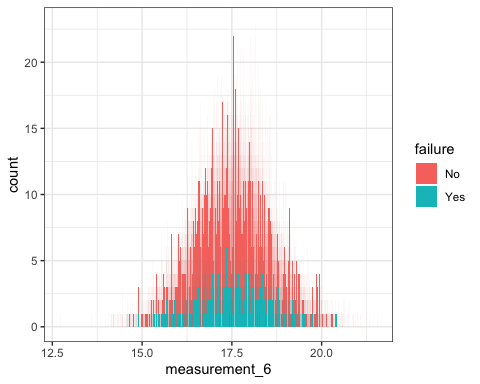
ggplot(train, aes(x=measurement\_5, fill = failure)) + geom\_bar() + theme\_bw()



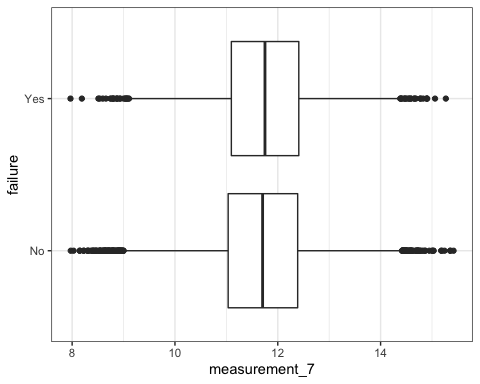
ggplot(train, aes(x = measurement\_6, y = failure)) +   
 geom\_boxplot() +   
  
 theme\_bw()



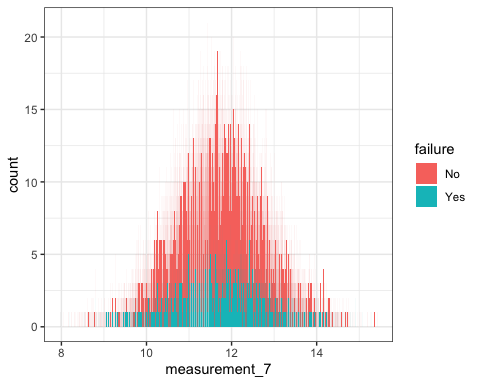
ggplot(train, aes(x=measurement\_6, fill = failure)) + geom\_bar() + theme\_bw()



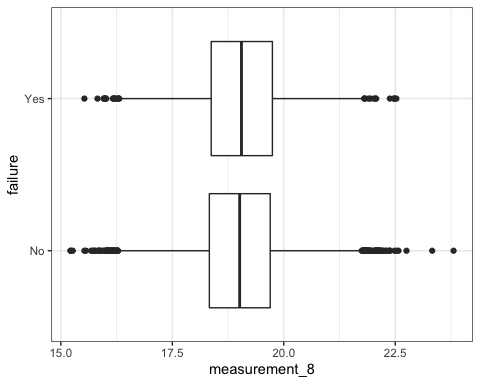
ggplot(train, aes(x = measurement\_7, y = failure)) +   
 geom\_boxplot() +   
  
 theme\_bw()



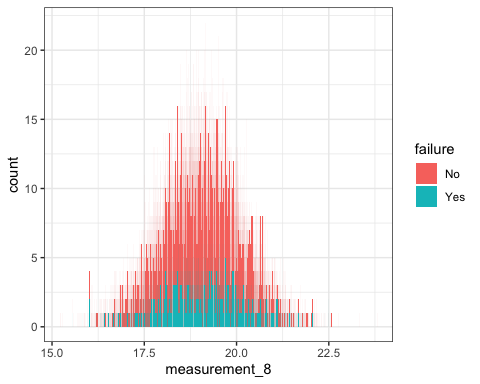
ggplot(train, aes(x=measurement\_7, fill = failure)) + geom\_bar() + theme\_bw()



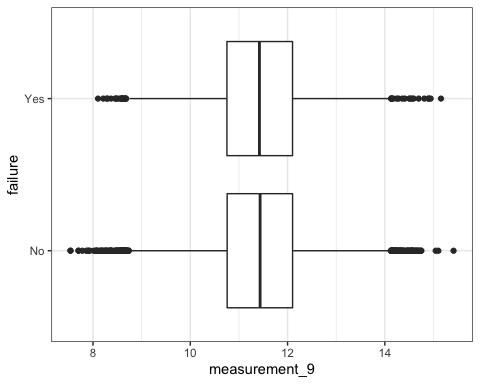
ggplot(train, aes(x = measurement\_8, y = failure)) +   
 geom\_boxplot() +   
  
 theme\_bw()



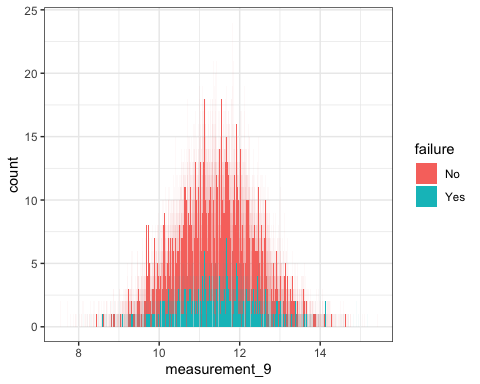
ggplot(train, aes(x=measurement\_8, fill = failure)) + geom\_bar() + theme\_bw()



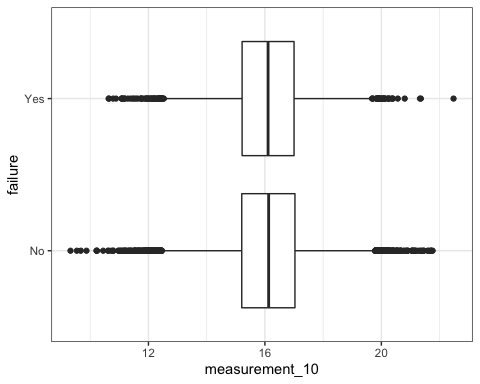
ggplot(train, aes(x = measurement\_9, y = failure)) +   
 geom\_boxplot() +   
  
 theme\_bw()



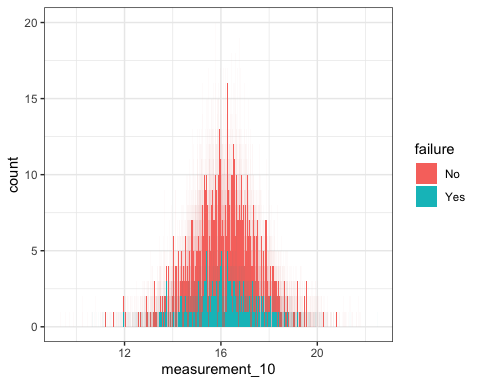
ggplot(train, aes(x=measurement\_9, fill = failure)) + geom\_bar() + theme\_bw()



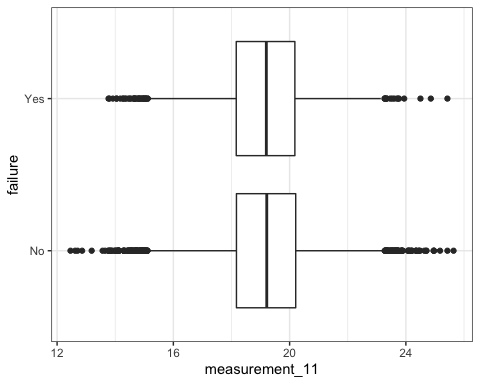
ggplot(train, aes(x = measurement\_10, y = failure)) +   
 geom\_boxplot() +   
  
 theme\_bw()



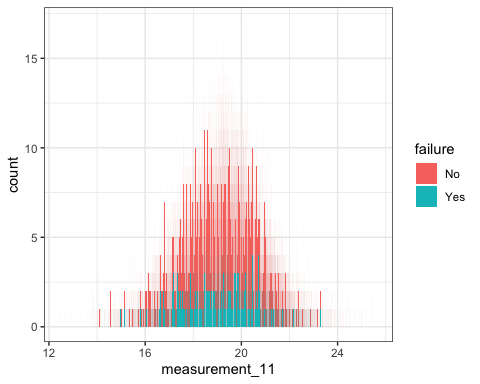
ggplot(train, aes(x=measurement\_10, fill = failure)) + geom\_bar() + theme\_bw()



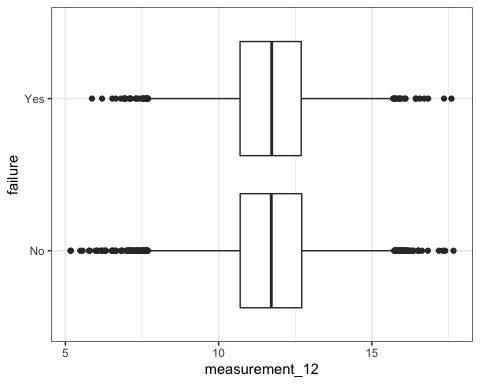
ggplot(train, aes(x = measurement\_11, y = failure)) +   
 geom\_boxplot() +   
  
 theme\_bw()



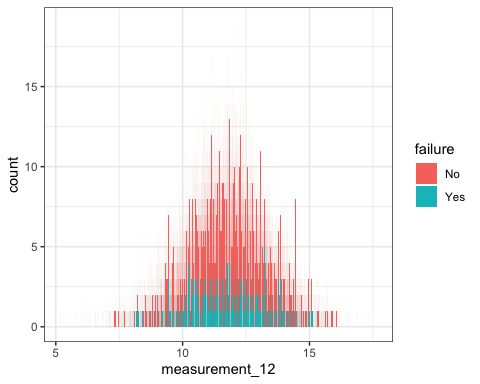
ggplot(train, aes(x=measurement\_11, fill = failure)) + geom\_bar() + theme\_bw()



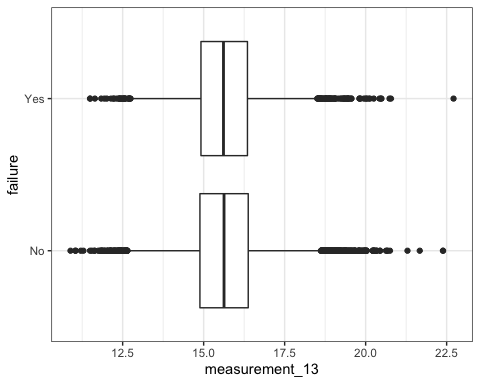
ggplot(train, aes(x = measurement\_12, y = failure)) +   
 geom\_boxplot() +   
  
 theme\_bw()



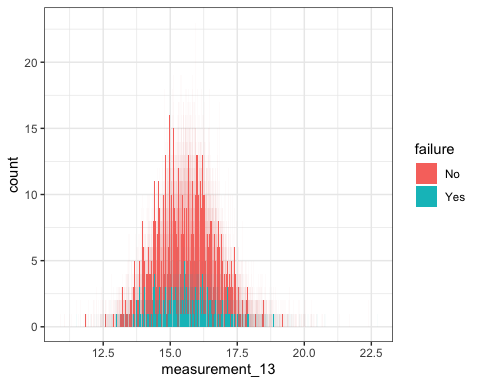
ggplot(train, aes(x=measurement\_12, fill = failure)) + geom\_bar() + theme\_bw()



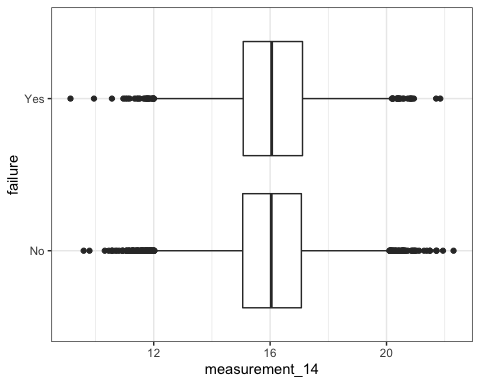
ggplot(train, aes(x = measurement\_13, y = failure)) +   
 geom\_boxplot() +   
  
 theme\_bw()



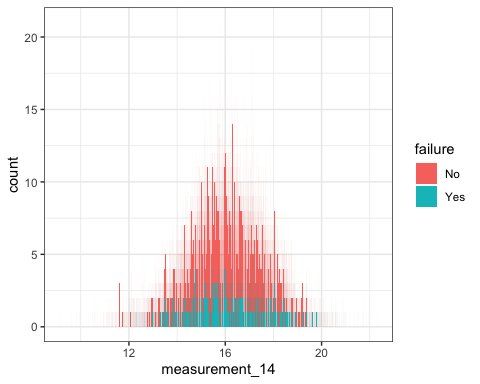
ggplot(train, aes(x=measurement\_13, fill = failure)) + geom\_bar() + theme\_bw()



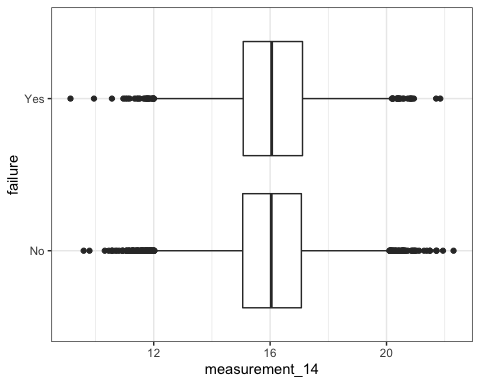
ggplot(train, aes(x = measurement\_14, y = failure)) +   
 geom\_boxplot() +   
  
 theme\_bw()



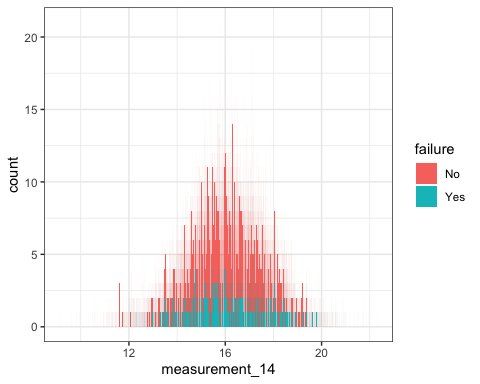
ggplot(train, aes(x=measurement\_14, fill = failure)) + geom\_bar() + theme\_bw()



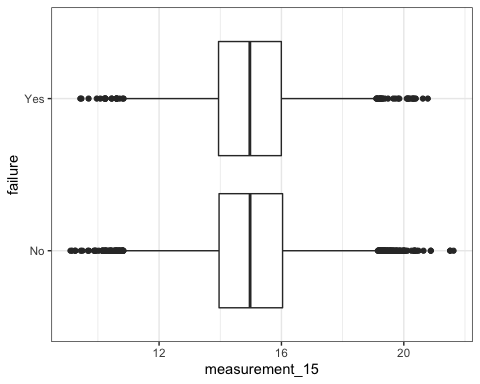
ggplot(train, aes(x = measurement\_14, y = failure)) +   
 geom\_boxplot() +   
  
 theme\_bw()



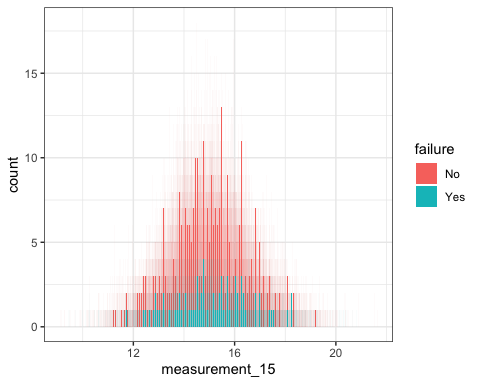
ggplot(train, aes(x=measurement\_14, fill = failure)) + geom\_bar() + theme\_bw()



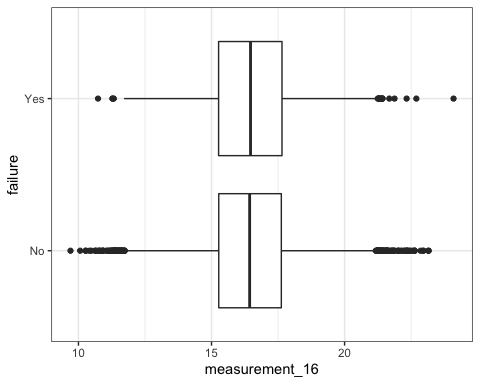
ggplot(train, aes(x = measurement\_15, y = failure)) +   
 geom\_boxplot() +   
  
 theme\_bw()



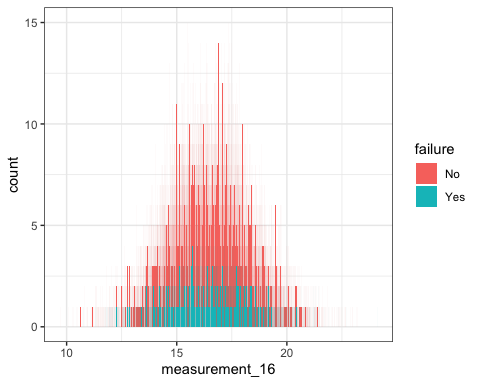
ggplot(train, aes(x=measurement\_15, fill = failure)) + geom\_bar() + theme\_bw()



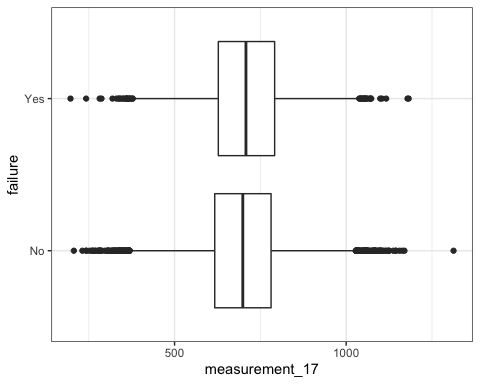
ggplot(train, aes(x = measurement\_16, y = failure)) +   
 geom\_boxplot() +   
  
 theme\_bw()



ggplot(train, aes(x=measurement\_16, fill = failure)) + geom\_bar() + theme\_bw()



ggplot(train, aes(x = measurement\_17, y = failure)) +   
 geom\_boxplot() +   
  
 theme\_bw()



ggplot(train, aes(x=measurement\_17, fill = failure)) + geom\_bar() + theme\_bw()

