STA130 Capstone - Star

Final code

Yanfei Zhu (Harry) Yukan Zou (Kibitzer) Crystal Wang (Crystal) Yingying Zhang (Maggie)

```
library(tidyverse)
library(rhdf5)
```

1. Getting data

```
# load in data
header <- h5ls("STA130_APOGEE.h5")
header
```

```
##    group       name       otype  dclass          dim
## 0      /       al_h H5I_DATASET   FLOAT        99705
## 1      /   al_h_err H5I_DATASET   FLOAT        99705
## 2      /        c_h H5I_DATASET   FLOAT        99705
## 3      /    c_h_err H5I_DATASET   FLOAT        99705
## 4      /       ca_h H5I_DATASET   FLOAT        99705
## 5      /   ca_h_err H5I_DATASET   FLOAT        99705
## 6      /       fe_h H5I_DATASET   FLOAT        99705
## 7      /   fe_h_err H5I_DATASET   FLOAT        99705
## 8      /       logg H5I_DATASET   FLOAT        99705
## 9      /   logg_err H5I_DATASET   FLOAT        99705
## 10     /       mg_h H5I_DATASET   FLOAT        99705
## 11     /   mg_h_err H5I_DATASET   FLOAT        99705
## 12     /        n_h H5I_DATASET   FLOAT        99705
## 13     /    n_h_err H5I_DATASET   FLOAT        99705
## 14     /        o_h H5I_DATASET   FLOAT        99705
## 15     /    o_h_err H5I_DATASET   FLOAT        99705
## 16     /        snr H5I_DATASET   FLOAT        99705
## 17     /    spectra H5I_DATASET   FLOAT 7514 x 99705
## 18     /    star_id H5I_DATASET INTEGER        99705
## 19     /       teff H5I_DATASET   FLOAT        99705
## 20     /   teff_err H5I_DATASET   FLOAT        99705
## 21     / wavelength H5I_DATASET   FLOAT         7514
```

```
wavelength <- "STA130_APOGEE.h5" %>%
  h5read("wavelength") %>% as_tibble()
head(wavelength)
```

```
## # A tibble: 6 x 1
##    value
##    <dbl>
## 1 15152.
## 2 15152.
## 3 15153.
## 4 15153.
## 5 15153.
## 6 15153.
```

```
spectra <- "STA130_APOGEE.h5" %>%
  h5read("spectra", index=list(NULL, 1:100)) %>% t() %>% as_tibble()
head(spectra)
```

```
## # A tibble: 6 x 7,514
##      V1    V2    V3    V4    V5    V6    V7    V8    V9   V10   V11   V12   V13
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 0.958 0.964 0.971 0.974 0.979 0.983 0.984 0.982 0.979 0.976 0.976 0.975 0.973
## 2 1.00  1.00  1.01  1.02  1.03  1.03  1.03  1.03  1.03  1.03  1.03  1.02  1.02
## 3 1.01  1.00  1.01  1.02  1.03  1.03  1.03  1.03  1.03  1.03  1.02  1.02  1.02
## 4 0.773 0.843 0.942 1.01  1.04  1.05  1.06  1.06  1.06  1.05  1.04  1.02  1.01
## 5 0.984 0.987 1.01  1.03  1.04  1.03  1.03  1.05  1.04  1.03  1.03  1.02  1.02
## 6 0.922 0.924 0.947 0.970 0.981 0.988 0.996 0.997 0.991 0.988 0.978 0.962 0.957
## # ... with 7,501 more variables: V14 <dbl>, V15 <dbl>, V16 <dbl>, V17 <dbl>,
## #   V18 <dbl>, V19 <dbl>, V20 <dbl>, V21 <dbl>, V22 <dbl>, V23 <dbl>,
## #   V24 <dbl>, V25 <dbl>, V26 <dbl>, V27 <dbl>, V28 <dbl>, V29 <dbl>,
## #   V30 <dbl>, V31 <dbl>, V32 <dbl>, V33 <dbl>, V34 <dbl>, V35 <dbl>,
## #   V36 <dbl>, V37 <dbl>, V38 <dbl>, V39 <dbl>, V40 <dbl>, V41 <dbl>,
## #   V42 <dbl>, V43 <dbl>, V44 <dbl>, V45 <dbl>, V46 <dbl>, V47 <dbl>,
## #   V48 <dbl>, V49 <dbl>, V50 <dbl>, V51 <dbl>, V52 <dbl>, V53 <dbl>, ...
```

```
snr <- "STA130_APOGEE.h5" %>%
  h5read("snr") %>% as_tibble()
head(snr)
```

```
## # A tibble: 6 x 1
##   value
##   <dbl>
## 1  283.
## 2  529.
## 3  533.
## 4  852.
## 5  173.
## 6  492.
```

```
star_id <- "STA130_APOGEE.h5" %>%
  h5read("star_id", bit64conversion='bit64') %>%
  as_tibble()
head(star_id)
```

```
## # A tibble: 6 x 1
##         x
##   <int64>
## 1   4.e17
## 2   4.e17
## 3   4.e17
## 4   4.e17
## 5   4.e17
## 6   5 e17
```

```
teff <- "STA130_APOGEE.h5" %>%
  h5read("teff") %>% as_tibble()
head(teff)
```

```
## # A tibble: 6 x 1
##   value
##   <dbl>
## 1 5031.
## 2 4976.
## 3 4982.
## 4 4074.
## 5 4757.
## 6 4669.
```

```
logg <- "STA130_APOGEE.h5" %>%
  h5read("logg") %>% as_tibble()
head(logg)
```

```
## # A tibble: 6 x 1
##   value
##   <dbl>
## 1  3.46
## 2  2.48
## 3  2.53
## 4  1.28
## 5  2.58
## 6  2.53
```

```
fe_h <- "STA130_APOGEE.h5" %>%
  h5read("fe_h") %>% as_tibble()
head(fe_h)
```

```
## # A tibble: 6 x 1
##     value
##     <dbl>
## 1 -0.160
## 2 -0.431
## 3 -0.427
## 4 -0.283
## 5 -0.0651
## 6 -0.135
```

```
al_h <- "STA130_APOGEE.h5" %>%
  h5read("al_h") %>% as_tibble()
head(al_h)
```

```
## # A tibble: 6 x 1
##     value
##     <dbl>
## 1 -0.146
## 2 -0.323
## 3 -0.300
## 4 -0.301
## 5 -0.0539
## 6 -0.0966
```

```
c_h <- "STA130_APOGEE.h5" %>%
  h5read("c_h") %>% as_tibble()
head(c_h)
```

```
## # A tibble: 6 x 1
##    value
##    <dbl>
## 1 -0.186
## 2 -0.503
## 3 -0.463
## 4 -0.358
## 5 -0.198
## 6 -0.214
```

```
ca_h <- "STA130_APOGEE.h5" %>%
  h5read("ca_h") %>% as_tibble()
head(ca_h)
```

```
## # A tibble: 6 x 1
##     value
##     <dbl>
## 1 -0.122
## 2 -0.345
## 3 -0.363
## 4 -0.281
## 5 -0.0255
## 6 -0.133
```

```
mg_h <- "STA130_APOGEE.h5" %>%
  h5read("mg_h") %>% as_tibble()
head(mg_h)
```

```
## # A tibble: 6 x 1
##     value
##     <dbl>
## 1 -0.0682
## 2 -0.318
## 3 -0.326
## 4 -0.195
## 5 -0.0858
## 6 -0.100
```

```
n_h <- "STA130_APOGEE.h5" %>%
  h5read("n_h") %>% as_tibble()
head(n_h)
```

```
## # A tibble: 6 x 1
##     value
##     <dbl>
## 1 -0.0761
## 2 -0.198
## 3 -0.239
## 4 -0.0369
## 5  0.136
## 6  0.161
```

```
o_h <- "STA130_APOGEE.h5" %>%
  h5read("o_h") %>% as_tibble()
head(o_h)
```

```
## # A tibble: 6 x 1
##     value
##     <dbl>
## 1 -0.0466
## 2 -0.318
## 3 -0.342
## 4 -0.185
## 5 -0.0824
## 6 -0.0716
```

```
fe_h_err <- "STA130_APOGEE.h5" %>%
  h5read("fe_h_err") %>% as_tibble()
head(fe_h_err)
```

```
## # A tibble: 6 x 1
##     value
##     <dbl>
## 1 0.00689
## 2 0.00754
## 3 0.00752
## 4 0.00888
## 5 0.00806
## 6 0.00746
```

```
al_h_err <- "STA130_APOGEE.h5" %>%
  h5read("al_h_err") %>% as_tibble()
head(al_h)
```

```
## # A tibble: 6 x 1
##     value
##     <dbl>
## 1 -0.146
## 2 -0.323
## 3 -0.300
## 4 -0.301
## 5 -0.0539
## 6 -0.0966
```

```
c_h_err <- "STA130_APOGEE.h5" %>%
  h5read("c_h_err") %>% as_tibble()
head(c_h)
```

```
## # A tibble: 6 x 1
##    value
##    <dbl>
## 1 -0.186
## 2 -0.503
## 3 -0.463
## 4 -0.358
## 5 -0.198
## 6 -0.214
```

```
ca_h_err <- "STA130_APOGEE.h5" %>%
  h5read("ca_h_err") %>% as_tibble()
head(ca_h_err)
```

```
## # A tibble: 6 x 1
##    value
##    <dbl>
## 1 0.0138
## 2 0.0164
## 3 0.0163
## 4 0.0136
## 5 0.0143
## 6 0.0130
```

```
mg_h_err <- "STA130_APOGEE.h5" %>%
  h5read("mg_h_err") %>% as_tibble()
head(mg_h_err)
```

```
## # A tibble: 6 x 1
##    value
##    <dbl>
## 1 0.0113
## 2 0.0130
## 3 0.0130
## 4 0.0116
## 5 0.0118
## 6 0.0109
```

```
n_h_err <- "STA130_APOGEE.h5" %>%
  h5read("n_h_err") %>% as_tibble()
head(n_h_err)
```

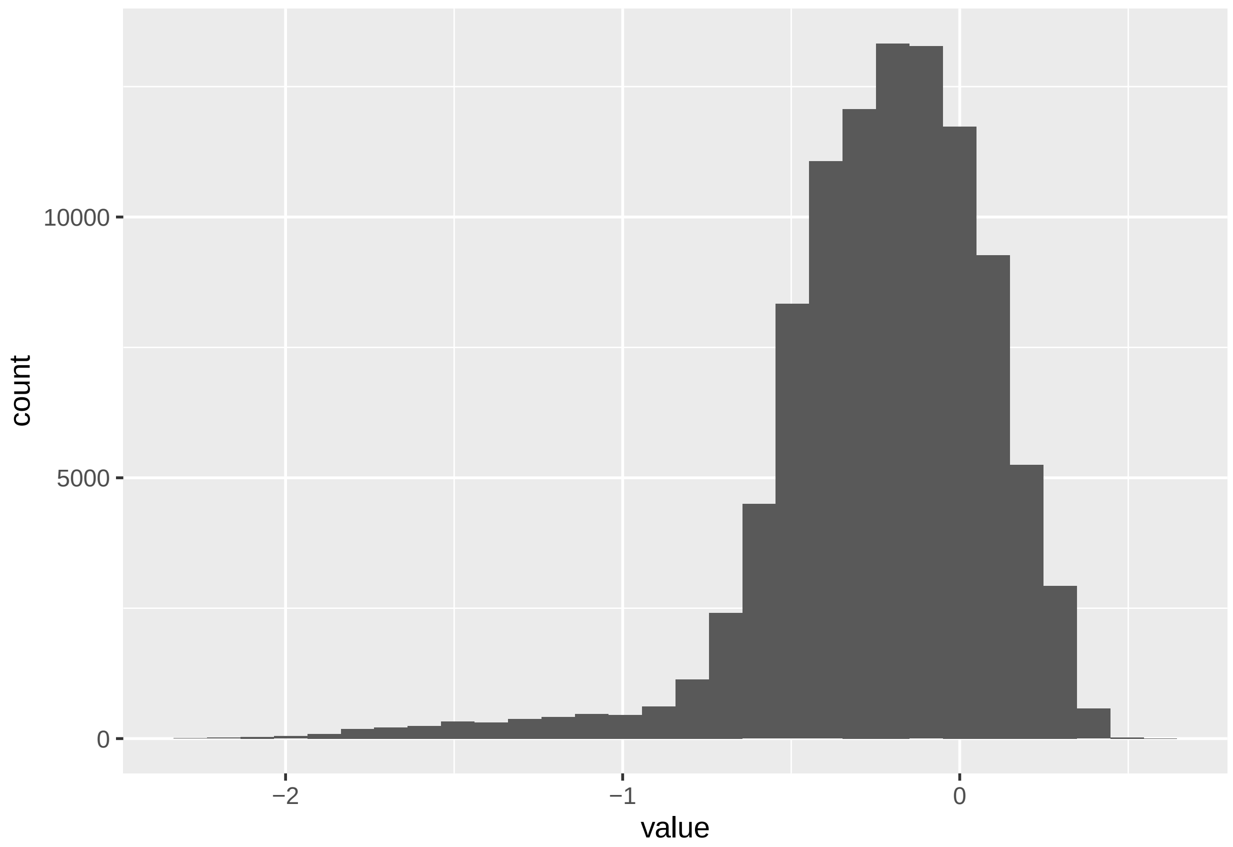
```
## # A tibble: 6 x 1
##    value
##    <dbl>
## 1 0.0186
## 2 0.0214
## 3 0.0214
## 4 0.0122
## 5 0.0164
## 6 0.0146
```

```
o_h_err <- "STA130_APOGEE.h5" %>%
  h5read("o_h_err") %>% as_tibble()
head(o_h_err)
```

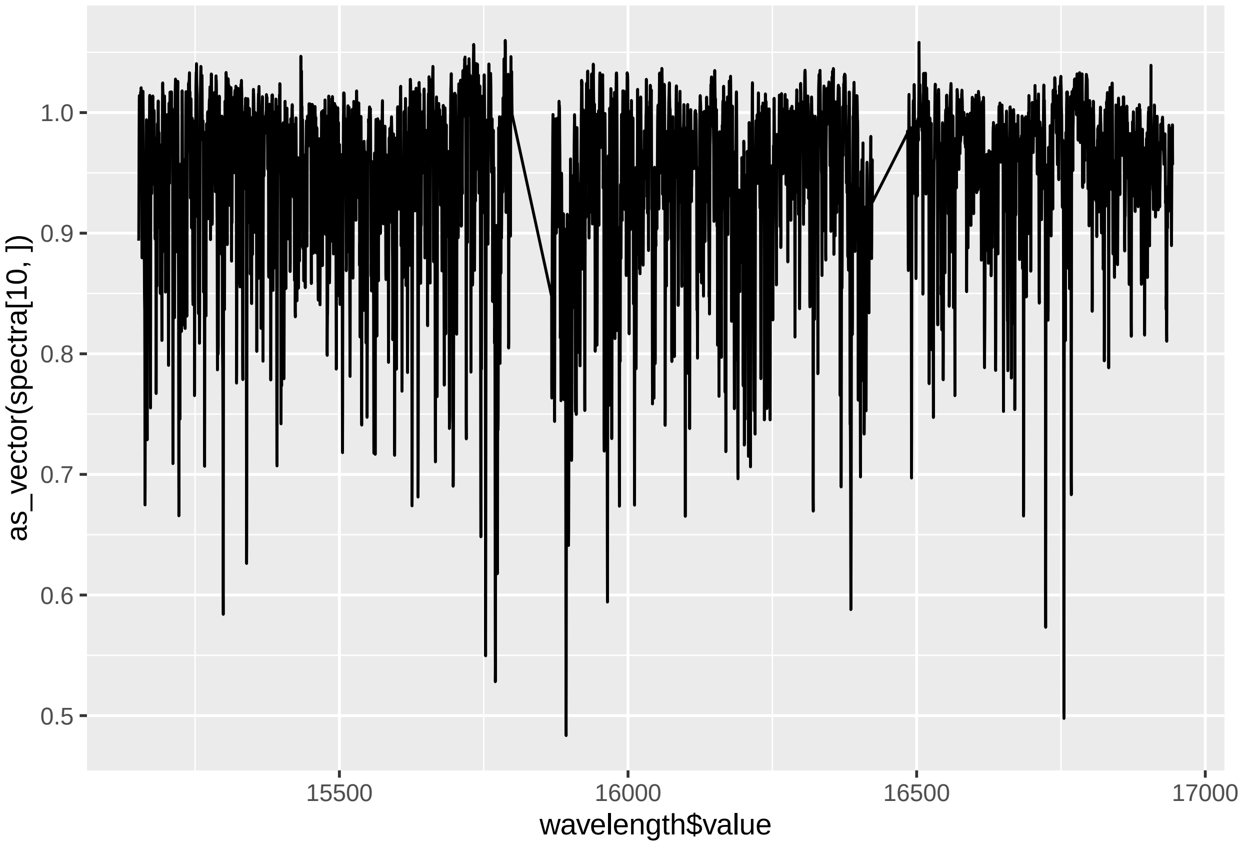
```
## # A tibble: 6 x 1
##     value
##     <dbl>
## 1 0.0234
## 2 0.0253
## 3 0.0254
## 4 0.00980
## 5 0.0176
## 6 0.0145
```

2. Try Given Example Graphs

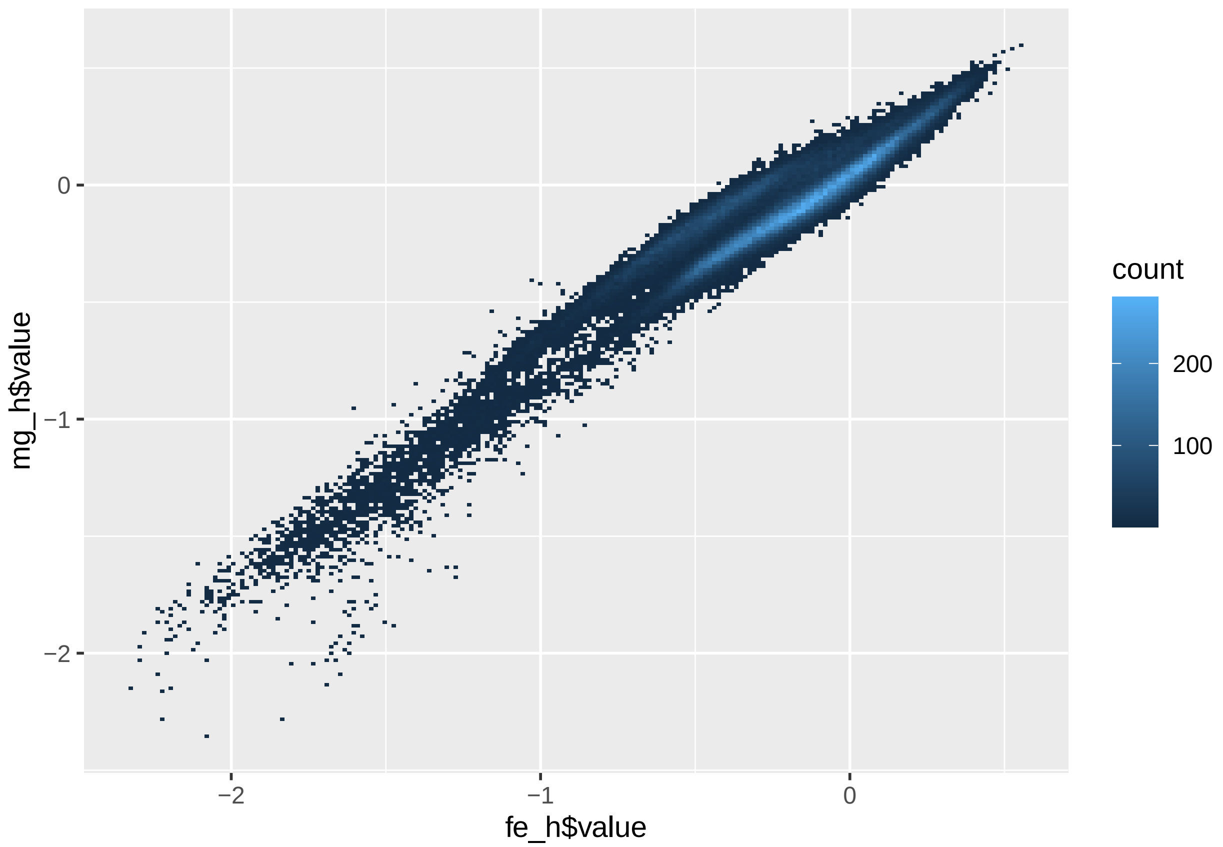
```
ggplot(data=fe_h) + aes(x=value) + geom_histogram()
```



```
ggplot() + aes(x=wavelength$value, y=as_vector(spectra[10,])) + geom_line()
```



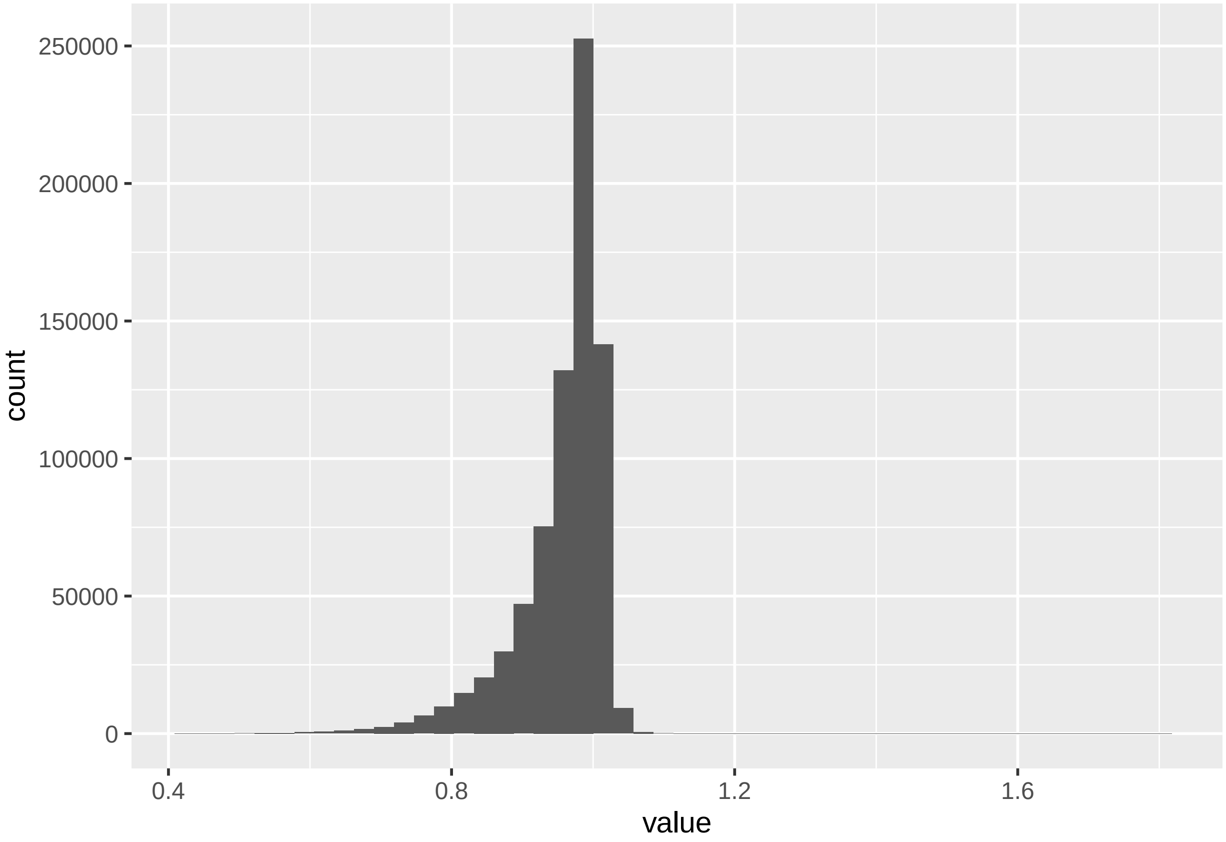
```
ggplot() + aes(x=fe_h$value, y=mg_h$value) + geom_bin_2d(bins=200)
```



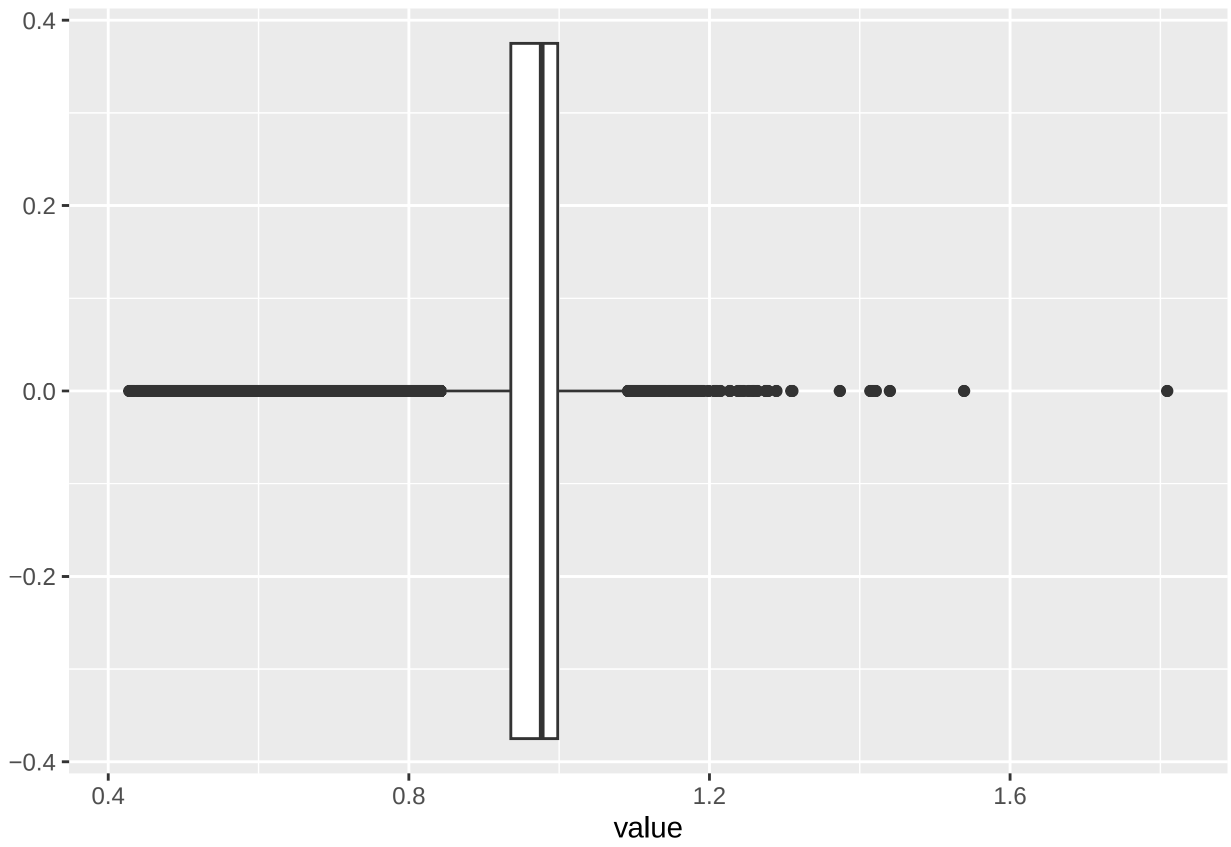
3. Question 1

Visualization Visualize data using histogram and boxplot.

```
value <- as_vector(spectra)

ggplot() + aes(x = value) + geom_histogram(bins = 50) +
  coord_cartesian(xlim = c(min(value) - 0.01, max(value) + 0.01))
```



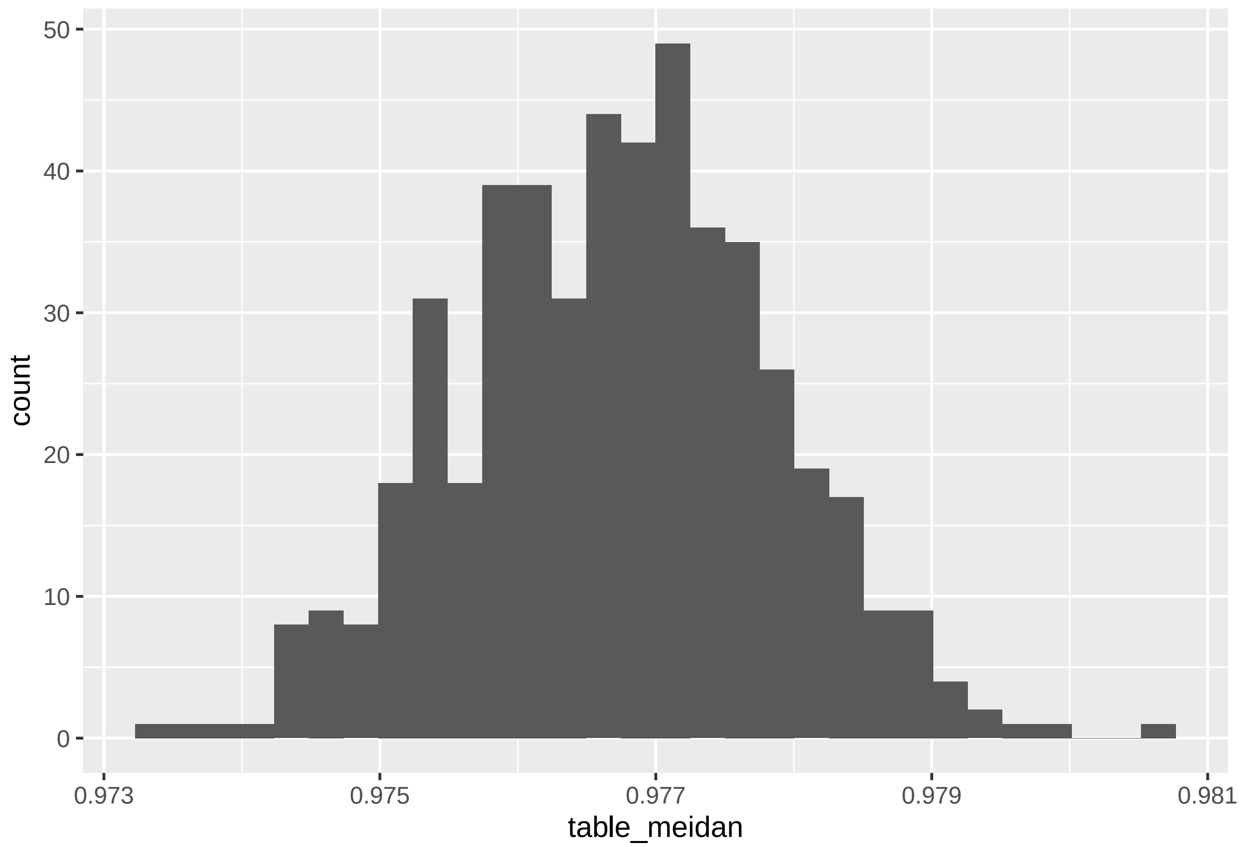
```
ggplot() + aes(x = value) + geom_boxplot() +
  coord_cartesian(xlim = c(min(value) - 0.01, max(value) + 0.01))
```



Null & Alternative Hypotheses > NuLL Hypotheses: The median of all star spectra is 0.9764 Angstroms. > Alternative Hypotheses: The median of all star is not 0.9764 Angstroms.

Simulation

```
student_num_last2 = 981
set.seed(student_num_last2 + 2)


N = 500
table_meidan <- 1:N

for (i in 1:N){
  simulated_x <- spectra[sample(nrow(spectra), size = 100, replace = TRUE), ]

  number <- median(as_vector(simulated_x))

  table_meidan[i] <- number
}

data <- as.data.frame(table_meidan)
ggplot(data = data, aes(x = table_meidan)) + geom_histogram()
```



```
p_1 <- data %>% filter(table_meidan < 0.9764)


total_possibility <- (nrow(p_1)/N)

sided2_p <- total_possibility * 2

sided2_p
```

```
## [1] 0.756
```

4. Question 2

```
o_h <- o_h %>% rowid_to_column()
fe_h <- fe_h %>% rowid_to_column()
OF = data.frame(Rowid = o_h$rowid, O_value = o_h$value, Fe_value = fe_h$value)
```

```
'NA' %in% OF
```

```
## [1] FALSE
```

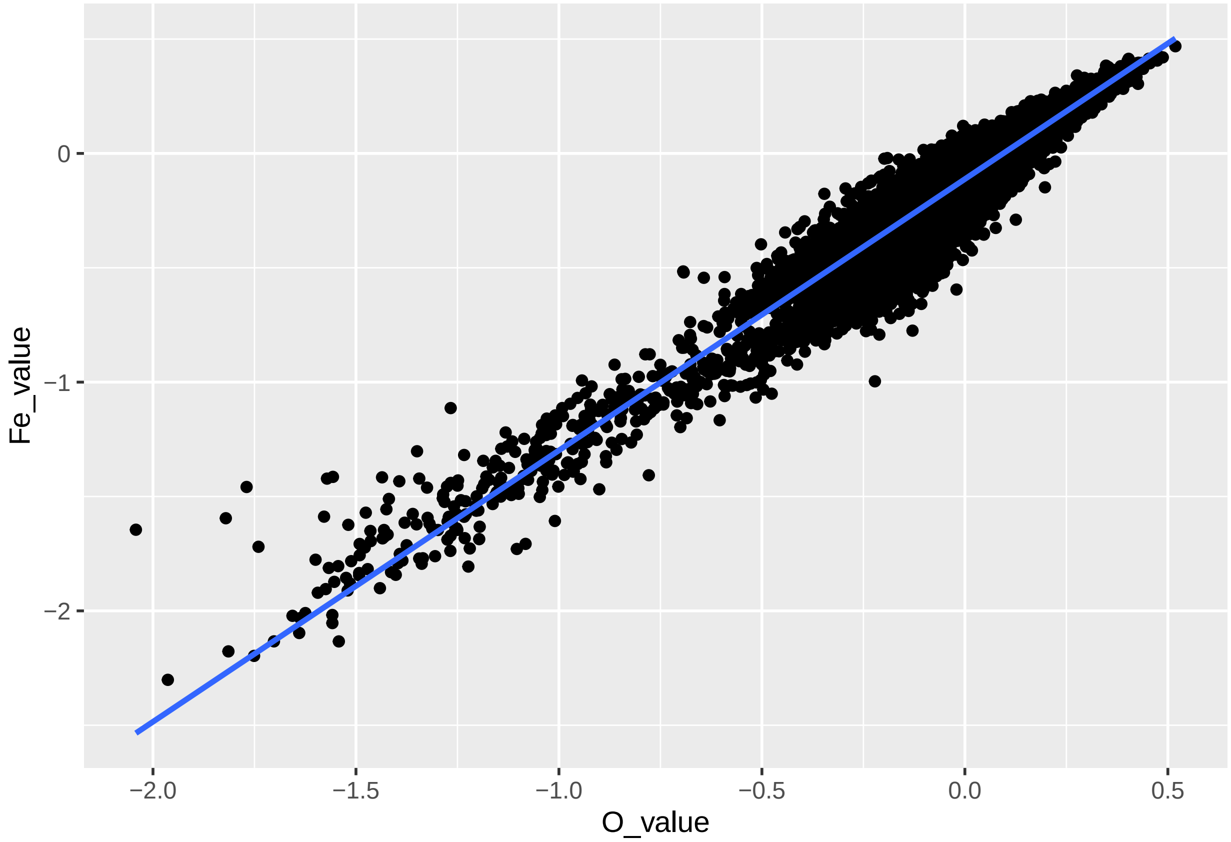
```
set.seed(140)
rn <- sample(2:10, 1)
OF_small <- filter(OF, Rowid%%rn == 0)
cor(OF_small$O_value, OF_small$Fe_value)
```

```
## [1] 0.945917
```

```
o_f_mod <- lm(OF_small$Fe_value ~ OF_small$O_value)
summary(o_f_mod)
```

```
##
## Call:
## lm(formula = OF_small$Fe_value ~ OF_small$O_value)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.62139 -0.05355  0.02732  0.06718  0.88945
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -0.112078   0.001148  -97.62   <2e-16 ***
## OF_small$O_value  1.186386   0.004075  291.12   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.107 on 9968 degrees of freedom
## Multiple R-squared:  0.8948, Adjusted R-squared:  0.8947
## F-statistic: 8.475e+04 on 1 and 9968 DF,  p-value: < 2.2e-16
```

```
OF_small %>% ggplot(aes(x = O_value, y = Fe_value)) + geom_point() + geom_smooth(method=lm, se=FALSE)
```



```
set.seed(130) # use this seed to make our analysis reproducible
nrow = nrow(OF_small)
training_ind <- sample(1:nrow, size = round(0.8 * nrow))
test_data <- OF_small %>% filter(Rowid %in% training_ind)
training_data <- OF_small %>% filter(!Rowid %in% training_ind)
y_train <- training_data$Fe_value
y_test <- test_data$Fe_value
```

```
mod_train <- lm(Fe_value ~ O_value, data = training_data)
yFe_mod_test <- predict(mod_train, newdata = test_data)
mod_test_RMSE <- sqrt(mean(y_test - yFe_mod_test)^2)
yFe_mod_train <- predict(mod_train, newdata = training_data)
mod_train_RMSE <- sqrt(mean((y_train - yFe_mod_train)^2))

tibble(model = 'slr',
       rmse_train = mod_train_RMSE,
       rmse_test = mod_test_RMSE,
       ratio = mod_train_RMSE/mod_test_RMSE)
```

```
## # A tibble: 1 x 4
##   model rmse_train rmse_test ratio
##   <chr>      <dbl>     <dbl> <dbl>
## 1 slr        0.108    0.0282  3.82
```

5. Question 3

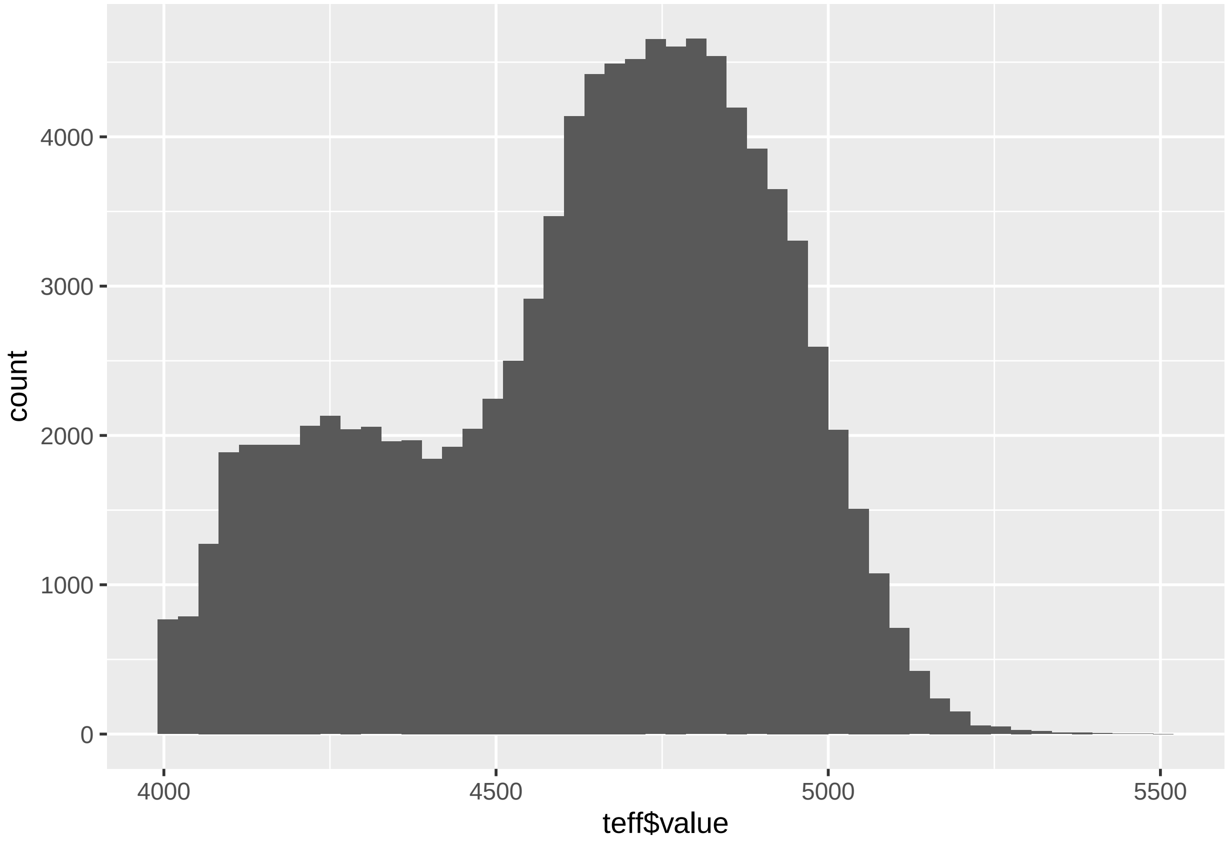
Check Result

```
mean(teff$value)
```

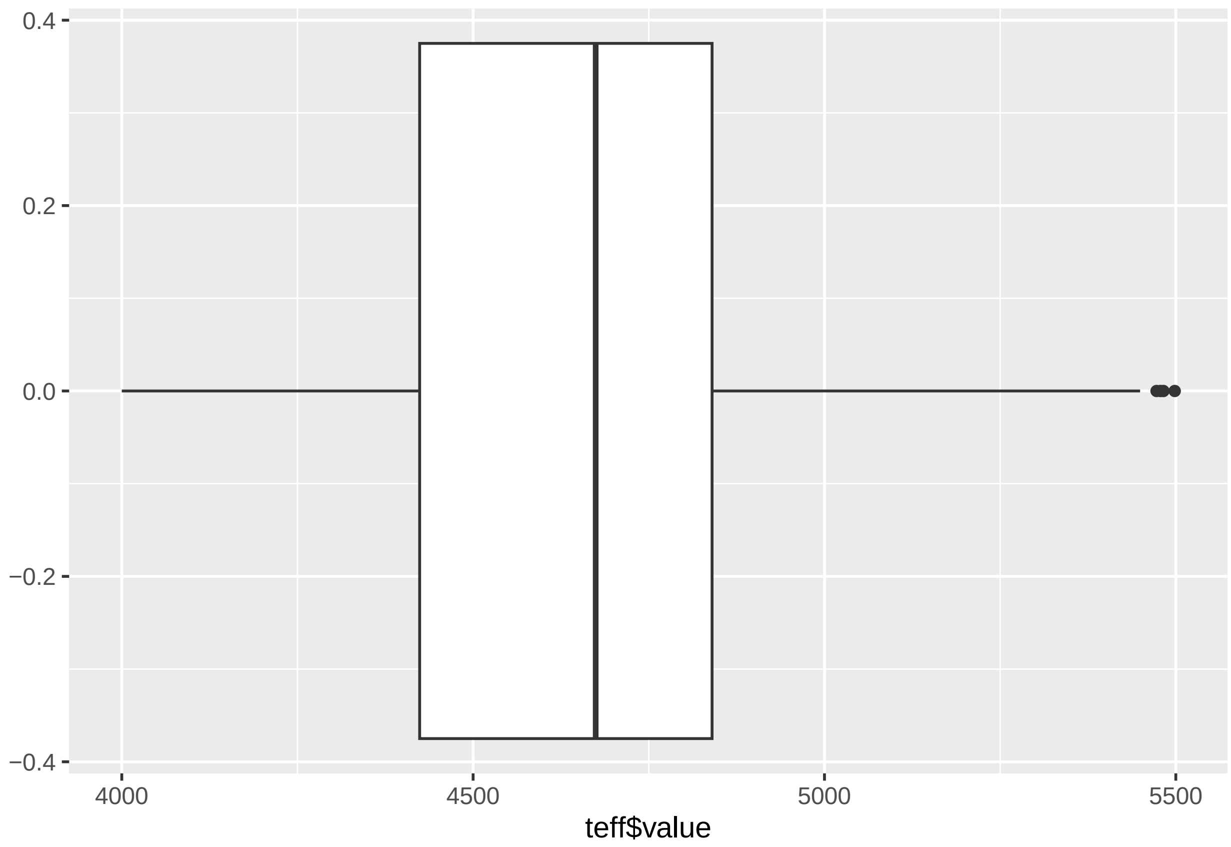
```
## [1] 4627.759
```

Visulization

```
ggplot() + aes(x = teff$value) + geom_histogram(bins = 50)
```



```
ggplot() + aes(x = teff$value) + geom_boxplot()
```



Null & Alternative Hypotheses > NuLL Hypotheses: The "effective temperature" of all stars is 4600 on average, measured in Kelvin. > Alternative Hypotheses: The "effective temperature" of all stars is not 4600 on average, measured in Kelvin.

Testing

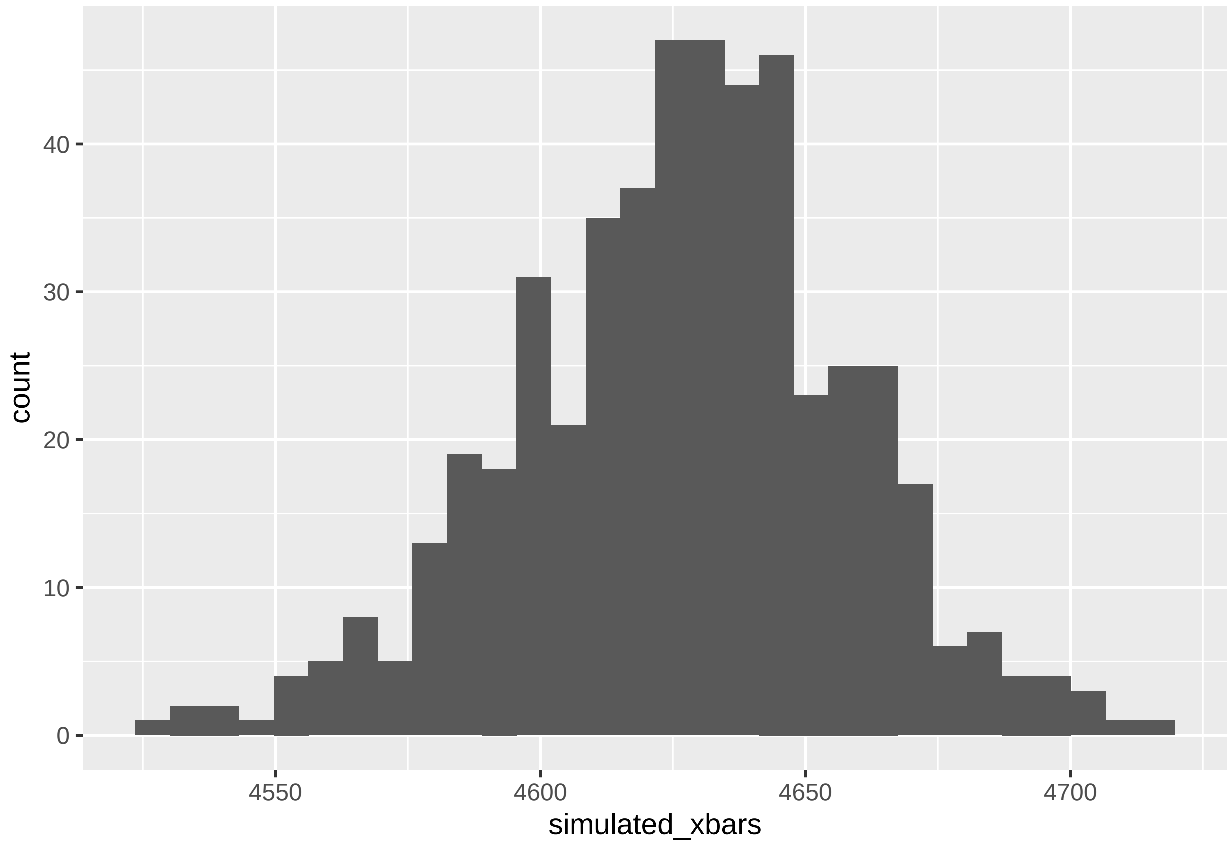
```
all_value <- (teff$value)
glimpse(all_value)
```

```
##  num [1:99705] 5031 4976 4982 4074 4757 ...
```

```
mean(all_value)
```

```
## [1] 4627.759
```

```
student_num_last2 = 778
set.seed(student_num_last2 + 2)  # REQUIRED so the result is reproducible!

# Code your answer here
N = 502
simulated_xbars <- 1:N
for (i in 1:N){
  simulated_x <- sample(all_value, 80,replace = TRUE)
  simulated_xbar <- mean(simulated_x)
  simulated_xbars[i] <- simulated_xbar
}

data <- as.data.frame(simulated_xbars)
ggplot(data = data, aes(x = simulated_xbars)) + geom_histogram()
```



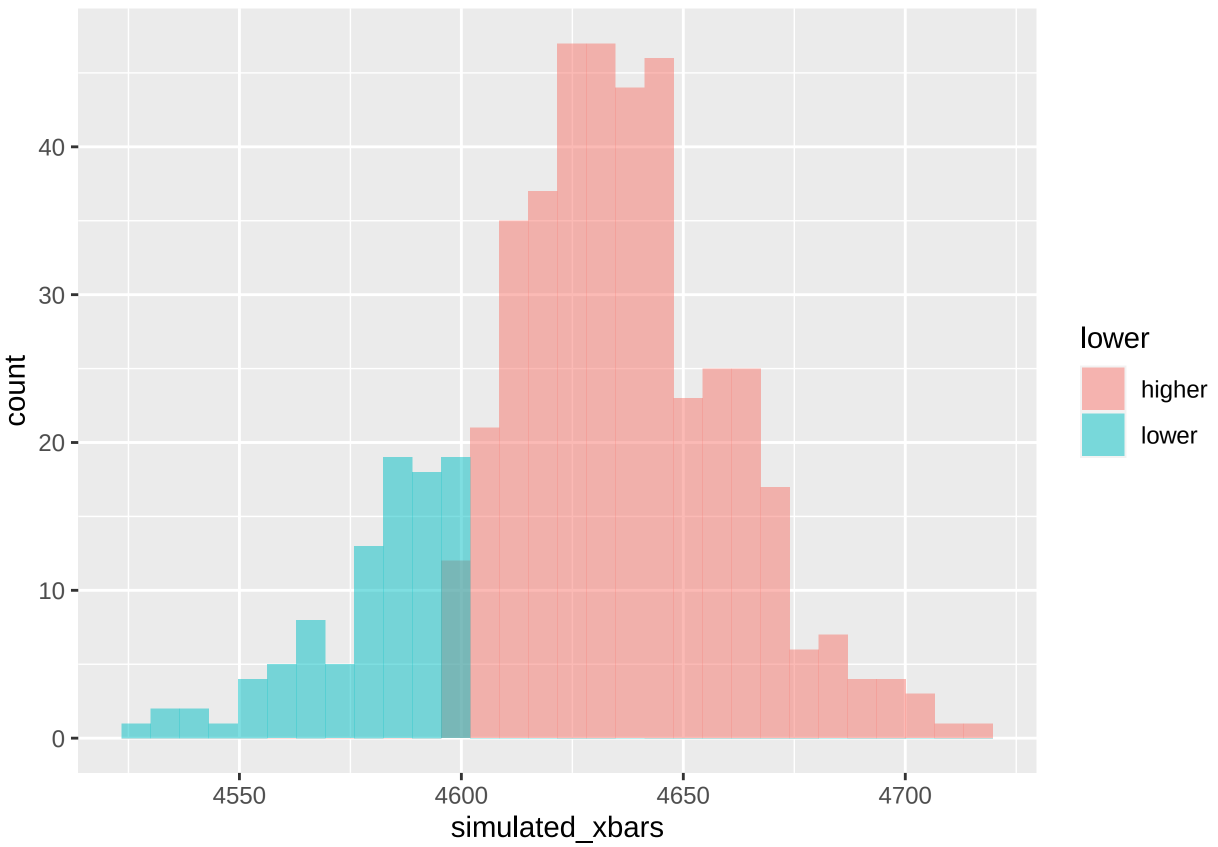
```
p_1 <- data %>% filter(simulated_xbars < 4600)
nrow(p_1)
```

```
## [1] 97
```

```
total_possibility <- (nrow(p_1)/N)

sided2_p <- total_possibility * 2

sided2_p
```

```
## [1] 0.3864542
```

```
lower_higher_4600 <- mutate(data, lower=ifelse(simulated_xbars < 4600, yes='lower', no='higher'))

lower_higher_4600 %>% ggplot() + aes(x=simulated_xbars, fill=lower) + geom_histogram(position="identity
```


40
30
20
10
0
count
4550
4600
4650
4700
simulated_xbars
lower
higher
lower