Descriptive Statistics Example

Arvon Clemons; Crystal Zang

9/1/2020

Loading R packages

- psych: we are using the summary statistics fucntions provided in this package
- dplyr: dataframe manipulation
- ggplot2: visualization

```
#require function combines the "installation" and "library" process of loading an R package
require(psych)
require(ggplot2)
require(dplyr)
```

Note that the `message = FALSE, warning=FALSE` parameter was added to the code chunk to prevent printing warning and messages when loading the packages.

Data Description

```
load("surgery_data.RData") #save this data in the same working directory of the rmd file, i.e. in the s

#save(surgery_data, file="surgery_data.RData")
glimpse(surgery_data) #previw dataframe, provides data type of each variable: numeric (dbl meaning doub
```

```
## Rows: 32,001
## Columns: 25
## $ ahrq_ccs            <chr> "<Other>", "<Other>", "<Other>", "<Other>", "<O...
## $ age                 <dbl> 67.8, 39.5, 56.5, 71.0, 56.3, 57.7, 56.6, 64.2,...
## $ gender              <chr> "M", "F", "F", "M", "M", "F", "M", "F", "M", "F...
## $ race                <chr> "Caucasian", "Caucasian", "Caucasian", "Caucasi...
## $ asa_status          <chr> "I-II", "I-II", "I-II", "III", "I-II", "I-II", ...
## $ bmi                 <dbl> 28.04, 37.85, 19.56, 32.22, 24.32, 40.30, 64.57...
## $ baseline_cancer     <chr> "No", "No", "No", "No", "Yes", "No", "No", "No"...
## $ baseline_cvd        <chr> "Yes", "Yes", "No", "Yes", "No", "Yes", "Yes", ...
## $ baseline_dementia   <chr> "No", "No", "No", "No", "No", "No", "No", "No",...
## $ baseline_diabetes   <chr> "No", "No", "No", "No", "No", "No", "Yes", "No"...
## $ baseline_digestive  <chr> "Yes", "No", "No", "No", "No", "No", "No", "No"...
## $ baseline_osteoart   <chr> "No", "No", "No", "No", "No", "No", "No", "No",...
## $ baseline_psych      <chr> "No", "No", "No", "No", "No", "Yes", "No", "No"...
## $ baseline_pulmonary  <chr> "No", "No", "No", "No", "No", "No", "No", "No",...
## $ baseline_charlson   <dbl> 0, 0, 0, 0, 0, 0, 2, 0, 1, 2, 0, 1, 0, 0, 0, 0,...
## $ mortality_rsi       <dbl> -0.63, -0.63, -0.49, -1.38, 0.00, -0.77, -0.36,...
## $ complication_rsi    <dbl> -0.26, -0.26, 0.00, -1.15, 0.00, -0.84, -1.34, ...
## $ ccsmort30rate       <dbl> 0.0042508, 0.0042508, 0.0042508, 0.0042508, 0.0...
## $ ccscomplicationrate <dbl> 0.07226355, 0.07226355, 0.07226355, 0.07226355,...
## $ hour                <dbl> 9.03, 18.48, 7.88, 8.80, 12.20, 7.67, 9.53, 7.5...
## $ dow                 <chr> "Mon", "Wed", "Fri", "Wed", "Thu", "Thu", "Tue"...
## $ month               <chr> "Nov", "Sep", "Aug", "Jun", "Aug", "Dec", "Apr"...
## $ moonphase           <chr> "Full Moon", "New Moon", "Full Moon", "Last Qua...
## $ mort30              <chr> "No", "No", "No", "No", "No", "No", "No", "No",...
## $ complication        <chr> "No", "No", "No", "No", "No", "No", "No", "Yes"...
```

```
# summary statistics
describe(surgery_data)
```

```
##                     vars     n  mean    sd median trimmed   mad   min   max
## ahrq_ccs*              1 32001 11.21  6.66  10.00   11.07  7.41  1.00 23.00
## age                    2 31999 57.66 15.04  58.60   58.22 14.83  1.00 90.00
## gender*                3 31998  1.46  0.50   1.00    1.45  0.00  1.00  2.00
## race*                  4 31521  1.92  0.39   2.00    1.97  0.00  1.00  3.00
## asa_status*            5 31993  1.49  0.56   1.00    1.45  0.00  1.00  3.00
## bmi                    6 28711 29.45  7.27  28.19   28.70  5.92  2.15 92.59
## baseline_cancer*       7 32001  1.34  0.47   1.00    1.30  0.00  1.00  2.00
## baseline_cvd*          8 32001  1.51  0.50   2.00    1.51  0.00  1.00  2.00
## baseline_dementia*     9 32001  1.01  0.09   1.00    1.00  0.00  1.00  2.00
## baseline_diabetes*    10 32001  1.13  0.34   1.00    1.04  0.00  1.00  2.00
## baseline_digestive*   11 32001  1.22  0.41   1.00    1.15  0.00  1.00  2.00
## baseline_osteoart*    12 32001  1.18  0.38   1.00    1.10  0.00  1.00  2.00
## baseline_psych*       13 32001  1.09  0.29   1.00    1.00  0.00  1.00  2.00
## baseline_pulmonary*   14 32001  1.11  0.31   1.00    1.01  0.00  1.00  2.00
## baseline_charlson     15 32001  1.18  1.88   0.00    0.78  0.00  0.00 13.00
## mortality_rsi         16 32001 -0.53  1.04  -0.30   -0.49  0.74 -4.40  4.86
## complication_rsi      17 32001 -0.41  1.20  -0.27   -0.43  0.46 -4.72 13.30
## ccsmort30rate         18 32001  0.00  0.00   0.00    0.00  0.00  0.00  0.02
## ccscomplicationrate   19 32001  0.13  0.09   0.11    0.12  0.06  0.02  0.47
## hour                  20 32001 10.38  2.92   9.65   10.08  3.14  6.00 19.00
## dow*                  21 32001  3.01  1.41   3.00    3.01  1.48  1.00  5.00
## month*                22 32001  6.62  3.52   7.00    6.63  4.45  1.00 12.00
## moonphase*            23 32001  2.48  1.11   2.00    2.48  1.48  1.00  4.00
## mort30*               24 32001  1.00  0.07   1.00    1.00  0.00  1.00  2.00
## complication*         25 32001  1.13  0.34   1.00    1.04  0.00  1.00  2.00
##                     range  skew kurtosis   se
## ahrq_ccs*           22.00  0.13    -1.15 0.04
## age                 89.00 -0.37     0.00 0.08
## gender*              1.00  0.15    -1.98 0.00
## race*                2.00 -0.72     2.96 0.00
## asa_status*          2.00  0.58    -0.70 0.00
## bmi                 90.44  1.54     5.15 0.04
## baseline_cancer*     1.00  0.66    -1.56 0.00
## baseline_cvd*        1.00 -0.02    -2.00 0.00
## baseline_dementia*   1.00 11.37   127.24 0.00
## baseline_diabetes*   1.00  2.20     2.83 0.00
## baseline_digestive*  1.00  1.35    -0.17 0.00
## baseline_osteoart*   1.00  1.68     0.81 0.00
## baseline_psych*      1.00  2.85     6.10 0.00
## baseline_pulmonary*  1.00  2.51     4.28 0.00
## baseline_charlson   13.00  2.48     6.88 0.01
## mortality_rsi        9.26 -0.14     1.05 0.01
## complication_rsi    18.02  1.75    12.10 0.01
## ccsmort30rate         0.02  1.54     1.50 0.00
## ccscomplicationrate   0.45  1.45     2.69 0.00
## hour                 13.00  0.63    -0.76 0.02
## dow*                  4.00 -0.01    -1.32 0.01
## month*               11.00 -0.05    -1.23 0.02
## moonphase*            3.00  0.02    -1.35 0.01
## mort30*               1.00 15.13   226.88 0.00
## complication*         1.00  2.16     2.66 0.00
```

```
#Check for NAs in gender variable
anyNA(surgery_data$gender)
```

```
## [1] TRUE
```

```
#remove observations with missing 'gender' values using "!" operator and is.na() function
gender_comp <- surgery_data[!is.na(surgery_data$gender), ]

#Create vectors of BMI based on gender; omit NAs
female_bmi <- gender_comp$bmi[gender_comp["gender"] == "F"]
male_bmi <- gender_comp$bmi[gender_comp["gender"] == "M"]

# Calculate mean BMI
# since there may be NA values, use na.rm = T to remove any possible NAs when calculating the mean
mean(female_bmi, na.rm = T); mean(male_bmi, na.rm = T)
```

```
## [1] 29.80188
```

```
## [1] 29.04185
```

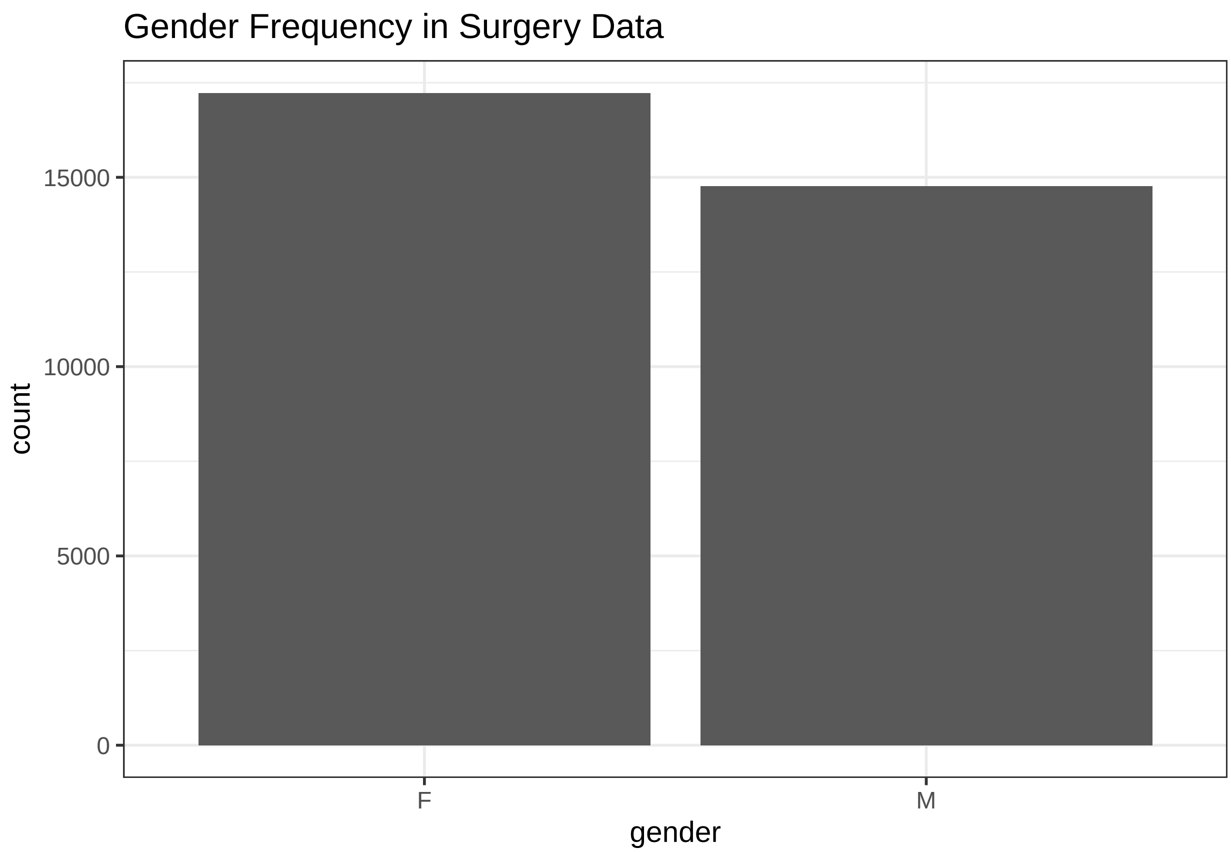
```
#Total numbers of obese or non-obese participants by gender
nonobeseMale <- sum(male_bmi <= 30, na.rm = T)
nonobeseFemale <- sum(female_bmi <= 30, na.rm = T)

obeseMale <- sum(male_bmi > 30, na.rm = T)
obeseFemale <- sum(female_bmi > 30, na.rm = T)

# Calculate proportion of participants who are over 30 BMI
propMale <- obeseMale / (obeseMale + nonobeseMale)
propFemale <- obeseFemale / (obeseFemale + nonobeseFemale)
```

Visualization

You can also embed plots, for example:

Bar plot for categorical variables

```
#manually omit NA values in the bar plot using "subset" function
ggplot(data = subset(surgery_data, !is.na(gender)), aes(x = gender))+
  geom_bar() +  #can manually change binwidth
  labs(title = "Gender Frequency in Surgery Data", #label axes
       x = "gender") +
  theme_bw() #make the plot looks pretty
```

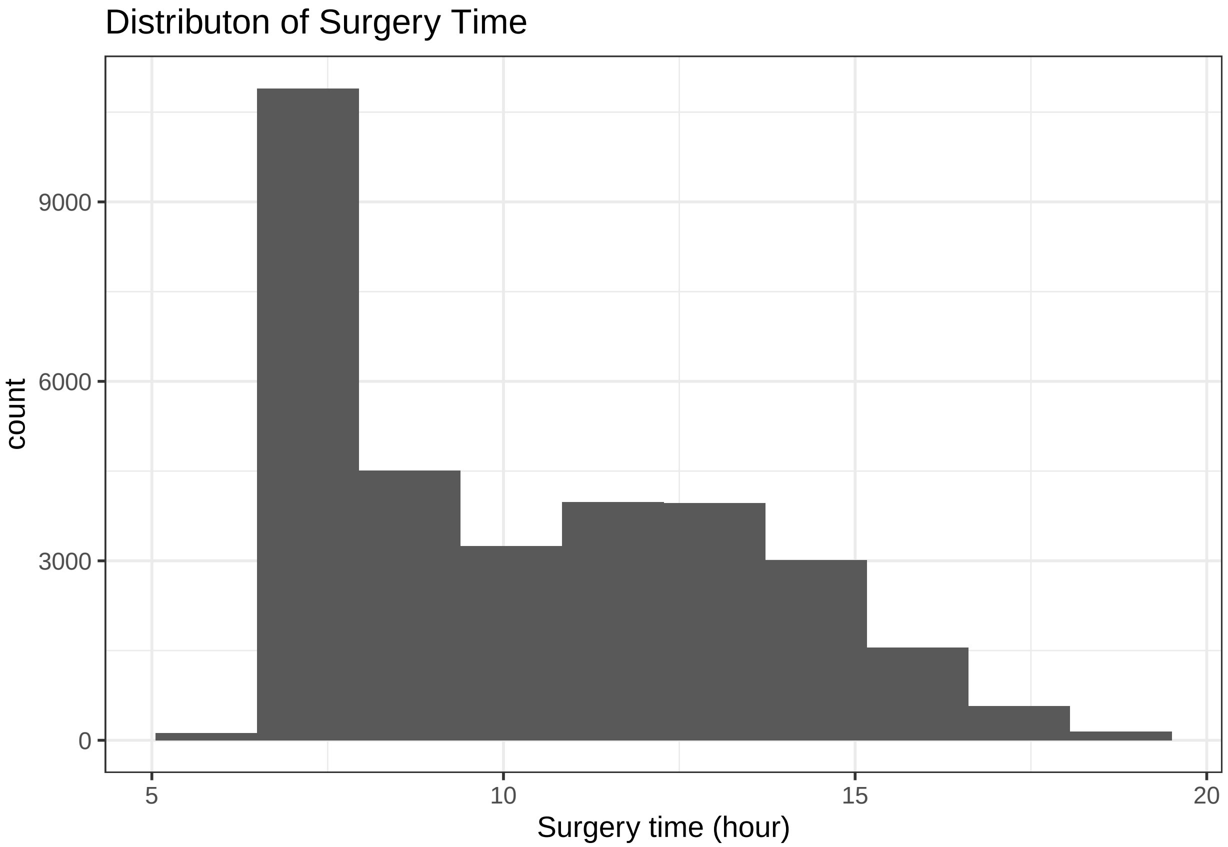


Histogram plot for numeric variables

```
summary(surgery_data$hour) #summmary statistics for 'hour' variable
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00    7.65    9.65   10.38   12.72   19.00
```

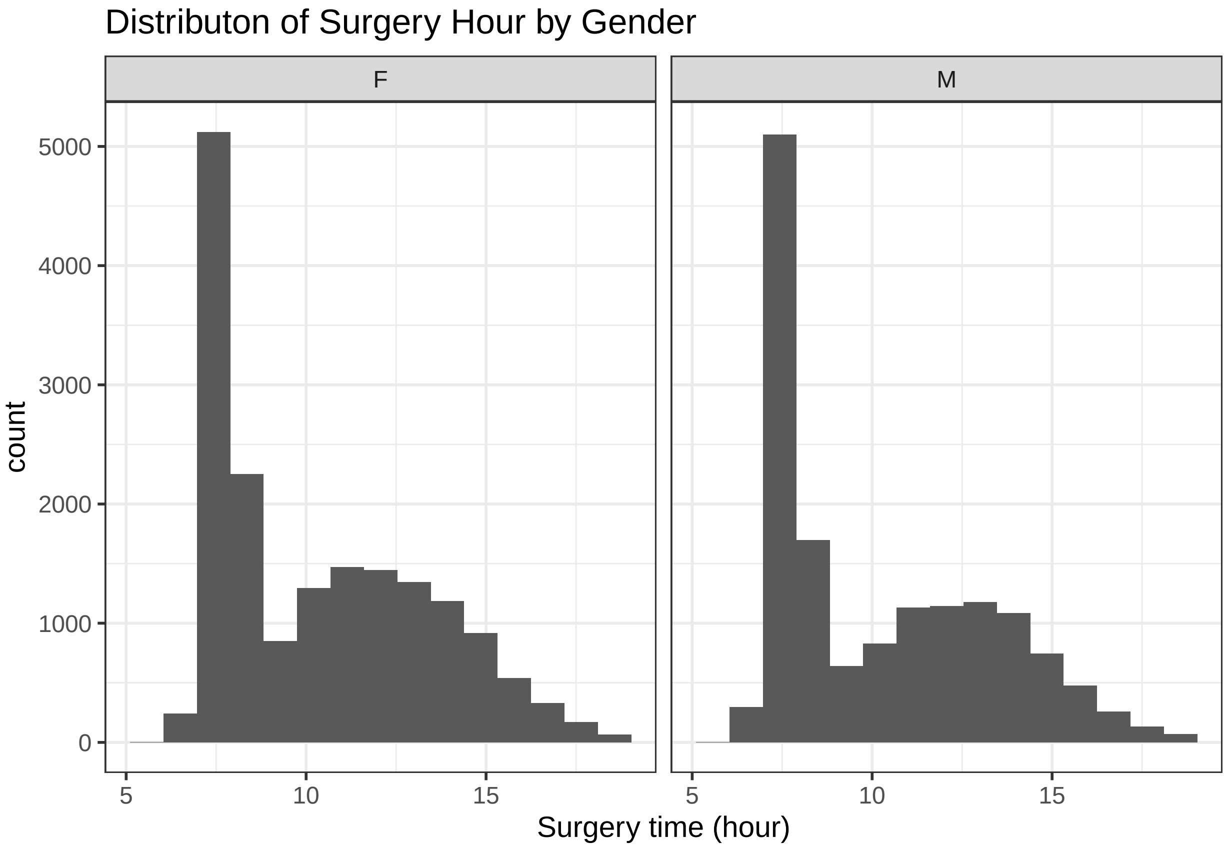
```
hour_mean <- mean(surgery_data$hour) #mean
hour_sd <- sd(surgery_data$hour) #standard deviation

#create a new variable "age_z" in the dataframe, which is the z-score of the age variable
surgery_data$hour_z <- (surgery_data$hour - hour_mean)/hour_sd

#NA values are automatically omitted in the histogram
ggplot(data = surgery_data, aes(x = hour))+
  geom_histogram(bins = 10) +  #can manually change the number of bins, now we have 10 bins
  labs(title = "Distributon of Surgery Time",
       x = "Surgery time (hour)") +
  theme_bw() #make the plot looks pretty
```



Note that we don't have NA values in the hour variable. If there are NAs, use "na.rm=T" argument in the mean and sd calculation. Ex. "hour_mean <- mean(surgery_data$hour, na.rm=T)"

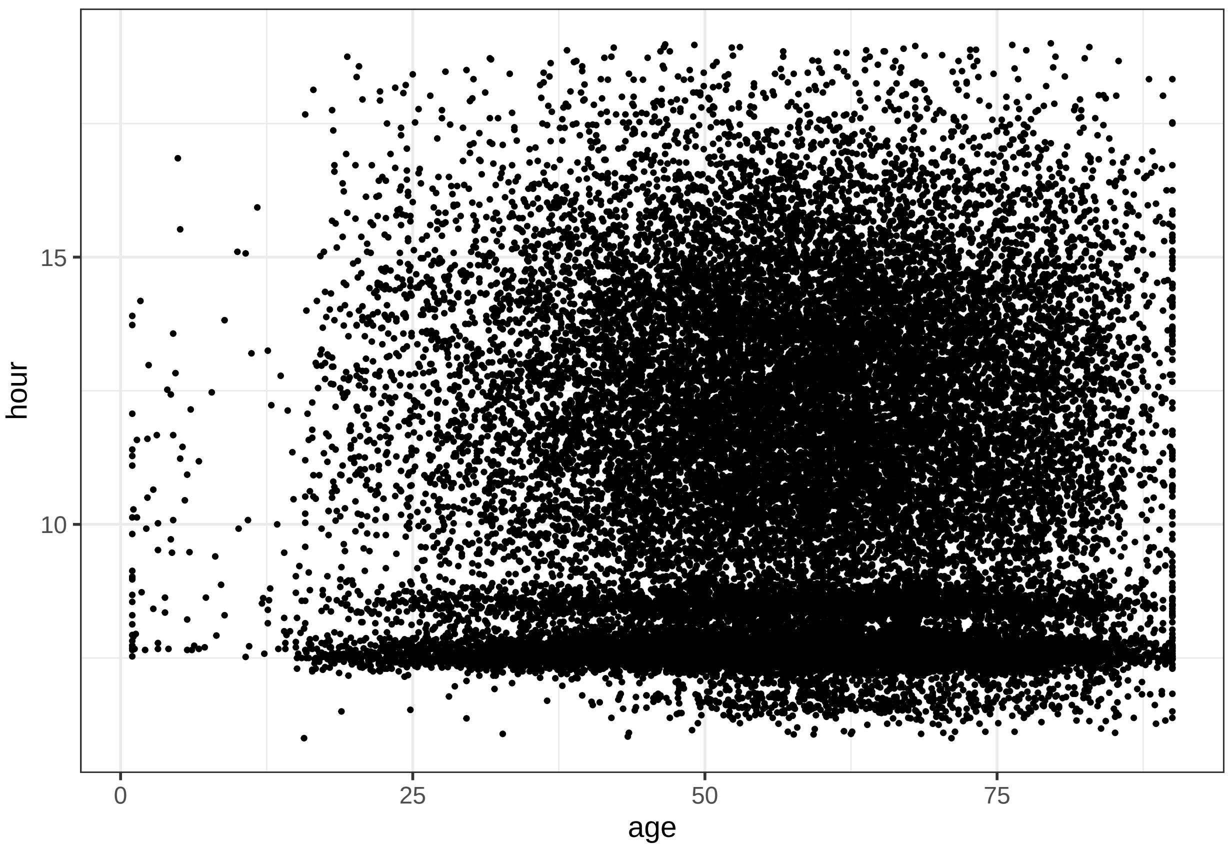
Side-by-side Plot for A Numeric Variable by Categories

```r
ggplot(data = subset(surgery_data, !is.na(gender)), aes(x = hour))+
  geom_histogram(bins = 15) +  #can manually change the number of bins, now we have 15 bins
  labs(title = "Distributon of Surgery Hour by Gender",
       x = "Surgery time (hour)") +
  facet_wrap(~gender) + #provides side by side plot by gender
  theme_bw() #make the plot looks pretty
```



Scatter Plot for Two Numeric Variables

```
ggplot(surgery_data, aes(x = age, y = hour))+
  geom_point(size = 0.5) +#can adjust the size of the point
   theme_bw()
```



##Session Information

```
sessionInfo()
```

```
## R version 3.6.3 (2020-02-29)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 7 x64 (build 7601) Service Pack 1
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] dplyr_0.8.5   ggplot2_3.3.0 psych_2.0.7
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.4.6     compiler_3.6.3   pillar_1.4.3     tools_3.6.3
##  [5] digest_0.6.25    evaluate_0.14    lifecycle_0.2.0  tibble_3.0.0
##  [9] gtable_0.3.0     nlme_3.1-144     lattice_0.20-38  pkgconfig_2.0.3
## [13] rlang_0.4.5      cli_2.0.2        yaml_2.2.1       parallel_3.6.3
## [17] xfun_0.13        withr_2.1.2      stringr_1.4.0    knitr_1.28
## [21] vctrs_0.2.4      grid_3.6.3       tidyselect_1.0.0 glue_1.4.0
## [25] R6_2.4.1         fansi_0.4.1      rmarkdown_2.1    farver_2.0.3
## [29] purrr_0.3.3      magrittr_1.5     scales_1.1.0     htmltools_0.4.0
## [33] ellipsis_0.3.0   assertthat_0.2.1 mnormt_2.0.2     colorspace_1.4-1
## [37] labeling_0.3     utf8_1.1.4       stringi_1.4.6    munsell_0.5.0
## [41] tmvnsim_1.0-2    crayon_1.3.4
```