Data Science II: Final Project- RMD and Knitted Files

Read in data:

```
flu = read.csv("severe_flu.csv")
head(flu)
```

```
##   id age gender race smoking height weight  bmi diabetes hypertension SBP LDL
## 1  1  59      0    1       1  162.7   73.2 27.6        0            0 120  95
## 2  2  54      1    1       1  169.9   73.6 25.5        1            1 133  87
## 3  3  55      1    3       1  175.4   86.3 28.1        0            0 123 139
## 4  4  59      0    1       0  169.5   77.3 26.9        0            0 121 126
## 5  5  62      1    1       0  168.7   84.9 29.8        1            0 122 107
## 6  6  64      1    1       0  170.2   75.7 26.1        0            1 132  99
##   severe_flu
## 1          0
## 2          0
## 3          0
## 4          1
## 5          1
## 6          0
```

Libraries:

```
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
library(tidymodels)
```

```
## -- Attaching packages --------------------------------------- tidymodels 1.3.0 --
```

```
## v broom        1.0.7     v rsample      1.2.1
## v dials        1.4.0     v tibble       3.2.1
## v dplyr        1.1.4     v tidyr        1.3.1
## v infer        1.0.7     v tune         1.3.0
## v modeldata    1.4.0     v workflows    1.2.0
## v parsnip      1.3.0     v workflowsets 1.1.0
## v purrr        1.0.4     v yardstick    1.3.2
## v recipes      1.1.1
```

```
## -- Conflicts ---------------------------------------- tidymodels_conflicts() --
## x purrr::discard()         masks scales::discard()
## x dplyr::filter()          masks stats::filter()
## x dplyr::lag()             masks stats::lag()
## x purrr::lift()            masks caret::lift()
## x yardstick::precision()   masks caret::precision()
## x yardstick::recall()      masks caret::recall()
## x yardstick::sensitivity() masks caret::sensitivity()
## x yardstick::specificity() masks caret::specificity()
## x recipes::step()          masks stats::step()
```

```
library(splines)
library(mgcv)
```

```
## Loading required package: nlme

##
## Attaching package: 'nlme'

## The following object is masked from 'package:dplyr':
##
##     collapse

## This is mgcv 1.9-1. For overview type 'help("mgcv-package")'.
```

```
library(pdp)
```

```
##
## Attaching package: 'pdp'

## The following object is masked from 'package:purrr':
##
##     partial
```

```
library(earth)
```

```
## Loading required package: Formula

## Loading required package: plotmo

## Loading required package: plotrix

##
## Attaching package: 'plotrix'

## The following object is masked from 'package:scales':
##
##     rescale
```

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v readr     2.1.5
## v lubridate 1.9.3     v stringr   1.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x readr::col_factor() masks scales::col_factor()
## x nlme::collapse()    masks dplyr::collapse()
## x purrr::discard()    masks scales::discard()
## x dplyr::filter()     masks stats::filter()
## x stringr::fixed()    masks recipes::fixed()
## x dplyr::lag()        masks stats::lag()
## x purrr::lift()       masks caret::lift()
## x pdp::partial()      masks purrr::partial()
## x readr::spec()       masks yardstick::spec()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(ggplot2)
library(bayesQR)
library(dplyr)
```

Factors

```r
flu <- flu %>%
  mutate(
    gender = factor(gender, levels = c(0, 1), labels = c("Female", "Male")),
    race = factor(race, levels = c(1, 2, 3, 4), labels = c("White", "Asian", "Black", "Hispanic")),
    smoking = factor(smoking, levels = c(0, 1, 2), labels = c("Never smoked", "Former smoker", "Current
    diabetes = factor(diabetes, levels = c(0, 1), labels = c("No", "Yes")),
    hypertension = factor(hypertension, levels = c(0, 1), labels = c("No", "Yes")),
    severe_flu = factor(severe_flu, levels = c(0, 1), labels = c("No", "Yes"))
  )
```

Exploratory analysis:

```r
# observe first couple of rows
head(flu)
```

```
##   id age gender  race       smoking height weight  bmi diabetes hypertension
## 1  1  59 Female White Former smoker  162.7   73.2 27.6       No           No
## 2  2  54   Male White Former smoker  169.9   73.6 25.5      Yes          Yes
## 3  3  55   Male Black Former smoker  175.4   86.3 28.1       No           No
## 4  4  59 Female White  Never smoked  169.5   77.3 26.9       No           No
## 5  5  62   Male White  Never smoked  168.7   84.9 29.8      Yes           No
## 6  6  64   Male White  Never smoked  170.2   75.7 26.1       No          Yes
```

```
##   SBP LDL severe_flu
## 1 120  95         No
## 2 133  87         No
## 3 123 139         No
## 4 121 126        Yes
## 5 122 107        Yes
## 6 132  99         No
```

```
str(flu)
```

```
## 'data.frame':    1000 obs. of  13 variables:
##  $ id          : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ age         : int  59 54 55 59 62 64 64 62 67 66 ...
##  $ gender      : Factor w/ 2 levels "Female","Male": 1 2 2 1 2 2 1 2 1 2 ...
##  $ race        : Factor w/ 4 levels "White","Asian",..: 1 1 3 1 1 1 1 4 1 1 ...
##  $ smoking     : Factor w/ 3 levels "Never smoked",..: 2 2 2 1 1 1 1 2 1 1 ...
##  $ height      : num  163 170 175 170 169 ...
##  $ weight      : num  73.2 73.6 86.3 77.3 84.9 75.7 89.2 81.9 68.3 76.3 ...
##  $ bmi         : num  27.6 25.5 28.1 26.9 29.8 26.1 28.9 27.7 24.3 25.5 ...
##  $ diabetes    : Factor w/ 2 levels "No","Yes": 1 2 1 1 2 1 1 1 1 1 ...
##  $ hypertension: Factor w/ 2 levels "No","Yes": 1 2 1 1 1 2 1 1 2 2 ...
##  $ SBP         : int  120 133 123 121 122 132 122 119 138 135 ...
##  $ LDL         : int  95 87 139 126 107 99 99 123 97 111 ...
##  $ severe_flu  : Factor w/ 2 levels "No","Yes": 1 1 1 2 2 1 1 2 1 1 ...
```

```
summary(flu)
```

```
##        id              age           gender          race
##  Min.   :   1.0   Min.   :46.00   Female:522   White   :656
##  1st Qu.: 250.8   1st Qu.:57.00   Male  :478   Asian   : 64
##  Median : 500.5   Median :60.00                Black   :184
##  Mean   : 500.5   Mean   :60.08                Hispanic: 96
##  3rd Qu.: 750.2   3rd Qu.:63.00
##  Max.   :1000.0   Max.   :72.00
##            smoking        height          weight            bmi
##  Never smoked  :584   Min.   :151.5   Min.   : 59.10   Min.   :20.10
##  Former smoker :313   1st Qu.:165.2   1st Qu.: 75.10   1st Qu.:25.90
##  Current smoker:103   Median :169.7   Median : 80.10   Median :27.70
##                       Mean   :169.7   Mean   : 80.03   Mean   :27.86
##                       3rd Qu.:174.0   3rd Qu.: 84.80   3rd Qu.:29.60
##                       Max.   :191.9   Max.   :103.70   Max.   :36.70
##  diabetes  hypertension      SBP             LDL        severe_flu
##  No :855   No :536      Min.   :108.0   Min.   : 41.0   No :747
##  Yes:145   Yes:464      1st Qu.:124.0   1st Qu.: 98.0   Yes:253
##                         Median :130.0   Median :111.0
##                         Mean   :129.9   Mean   :110.5
##                         3rd Qu.:135.0   3rd Qu.:123.0
##                         Max.   :154.0   Max.   :174.0
```

```
#checking for missing
colSums(is.na(flu))
```

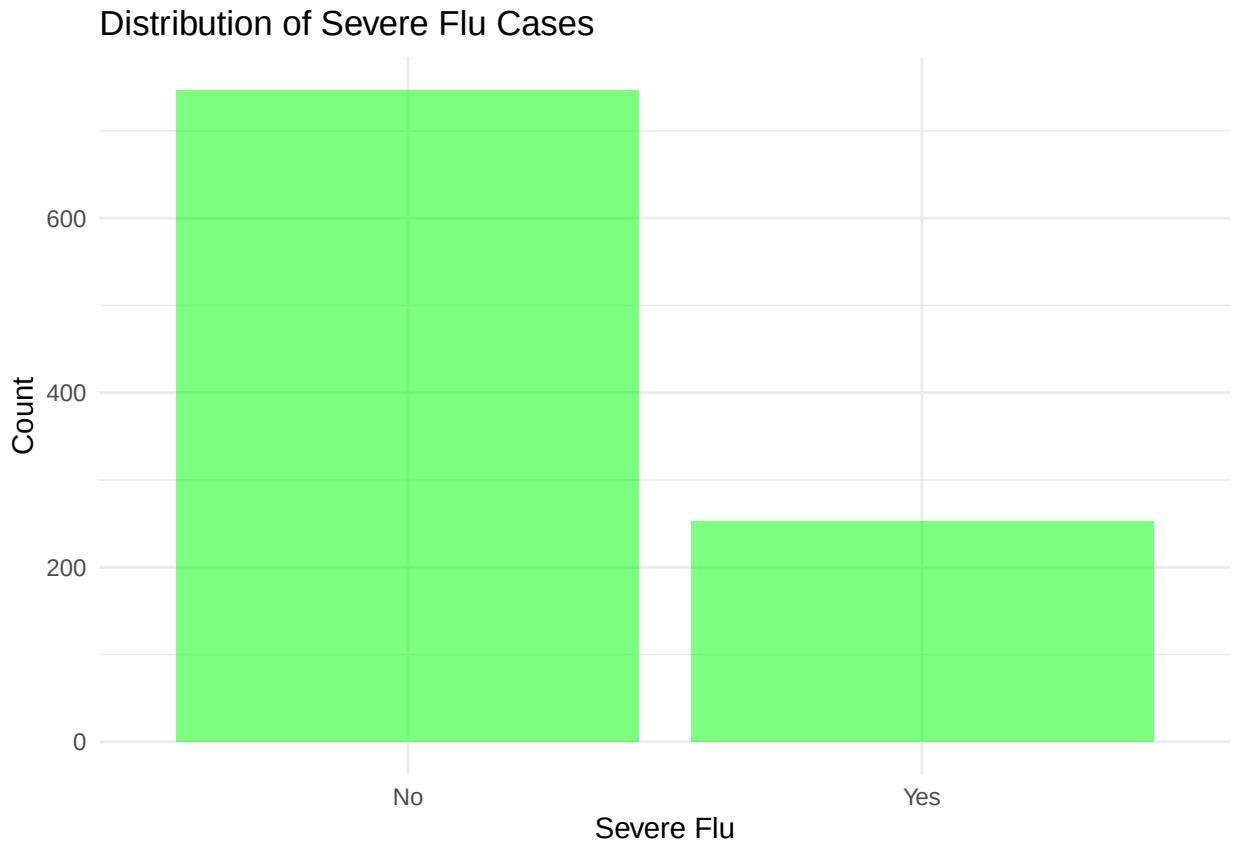
```
##           id          age       gender         race       smoking       height
##            0            0            0            0            0            0
##       weight          bmi     diabetes hypertension          SBP          LDL
##            0            0            0            0            0            0
##   severe_flu
##            0
```

```
# checking for duplicates
sum(duplicated(flu))
```

```
## [1] 0
```

Bar Plot for distribution of severe flu

```
ggplot(flu, aes(x = severe_flu)) +
  geom_bar(fill = "green", alpha = 0.5) +
  theme_minimal() +
  labs(title = "Distribution of Severe Flu Cases",
       x = "Severe Flu",
       y = "Count")
```



Summarizing continous variables

```
summary(flu[, c("age", "height", "weight", "bmi", "SBP", "LDL")])
```

```
##       age            height          weight            bmi
##  Min.   :46.00   Min.   :151.5   Min.   : 59.10   Min.   :20.10
##  1st Qu.:57.00   1st Qu.:165.2   1st Qu.: 75.10   1st Qu.:25.90
##  Median :60.00   Median :169.7   Median : 80.10   Median :27.70
##  Mean   :60.08   Mean   :169.7   Mean   : 80.03   Mean   :27.86
##  3rd Qu.:63.00   3rd Qu.:174.0   3rd Qu.: 84.80   3rd Qu.:29.60
##  Max.   :72.00   Max.   :191.9   Max.   :103.70   Max.   :36.70
##       SBP             LDL
##  Min.   :108.0   Min.   : 41.0
##  1st Qu.:124.0   1st Qu.: 98.0
##  Median :130.0   Median :111.0
##  Mean   :129.9   Mean   :110.5
##  3rd Qu.:135.0   3rd Qu.:123.0
##  Max.   :154.0   Max.   :174.0
```

Summarizing categorical variables distribution among those with severe fu vs. without severe flu

```
# For gender
flu %>%
  group_by(severe_flu, gender) %>%
  summarize(count = n(), .groups = "drop") %>%
  group_by(severe_flu) %>%
  mutate(proportion = count / sum(count))
```

```
## # A tibble: 4 x 4
## # Groups:   severe_flu [2]
##   severe_flu gender count proportion
##   <fct>      <fct>  <int>      <dbl>
## 1 No         Female   400      0.535
## 2 No         Male     347      0.465
## 3 Yes        Female   122      0.482
## 4 Yes        Male     131      0.518
```

```
# For race
flu %>%
  group_by(severe_flu, race) %>%
  summarize(count = n(), .groups = "drop") %>%
  group_by(severe_flu) %>%
  mutate(proportion = count / sum(count))
```

```
## # A tibble: 8 x 4
## # Groups:   severe_flu [2]
##   severe_flu race     count proportion
##   <fct>      <fct>    <int>      <dbl>
```

```
## 1 No          White      492     0.659
## 2 No          Asian       48     0.0643
## 3 No          Black      143     0.191
## 4 No          Hispanic    64     0.0857
## 5 Yes         White      164     0.648
## 6 Yes         Asian       16     0.0632
## 7 Yes         Black       41     0.162
## 8 Yes         Hispanic    32     0.126
```

```
# For smoking
flu %>%
  group_by(severe_flu, smoking) %>%
  summarize(count = n(), .groups = "drop") %>%
  group_by(severe_flu) %>%
  mutate(proportion = count / sum(count))
```

```
## # A tibble: 6 x 4
## # Groups:   severe_flu [2]
##   severe_flu smoking        count proportion
##   <fct>      <fct>          <int>      <dbl>
## 1 No         Never smoked     439     0.588
## 2 No         Former smoker    243     0.325
## 3 No         Current smoker    65     0.0870
## 4 Yes        Never smoked     145     0.573
## 5 Yes        Former smoker     70     0.277
## 6 Yes        Current smoker    38     0.150
```

```
# For diabetes
flu %>%
  group_by(severe_flu, diabetes) %>%
  summarize(count = n(), .groups = "drop") %>%
  group_by(severe_flu) %>%
  mutate(proportion = count / sum(count))
```

```
## # A tibble: 4 x 4
## # Groups:   severe_flu [2]
##   severe_flu diabetes count proportion
##   <fct>      <fct>    <int>      <dbl>
## 1 No         No         654      0.876
## 2 No         Yes         93      0.124
## 3 Yes        No         201      0.794
## 4 Yes        Yes         52      0.206
```

```
# For hypertension
flu %>%
  group_by(severe_flu, hypertension) %>%
  summarize(count = n(), .groups = "drop") %>%
  group_by(severe_flu) %>%
  mutate(proportion = count / sum(count))
```

```
## # A tibble: 4 x 4
## # Groups:   severe_flu [2]
```

```
##   severe_flu hypertension count proportion
##   <fct>      <fct>        <int>      <dbl>
## 1 No         No             409      0.548
## 2 No         Yes            338      0.452
## 3 Yes        No             127      0.502
## 4 Yes        Yes            126      0.498
```

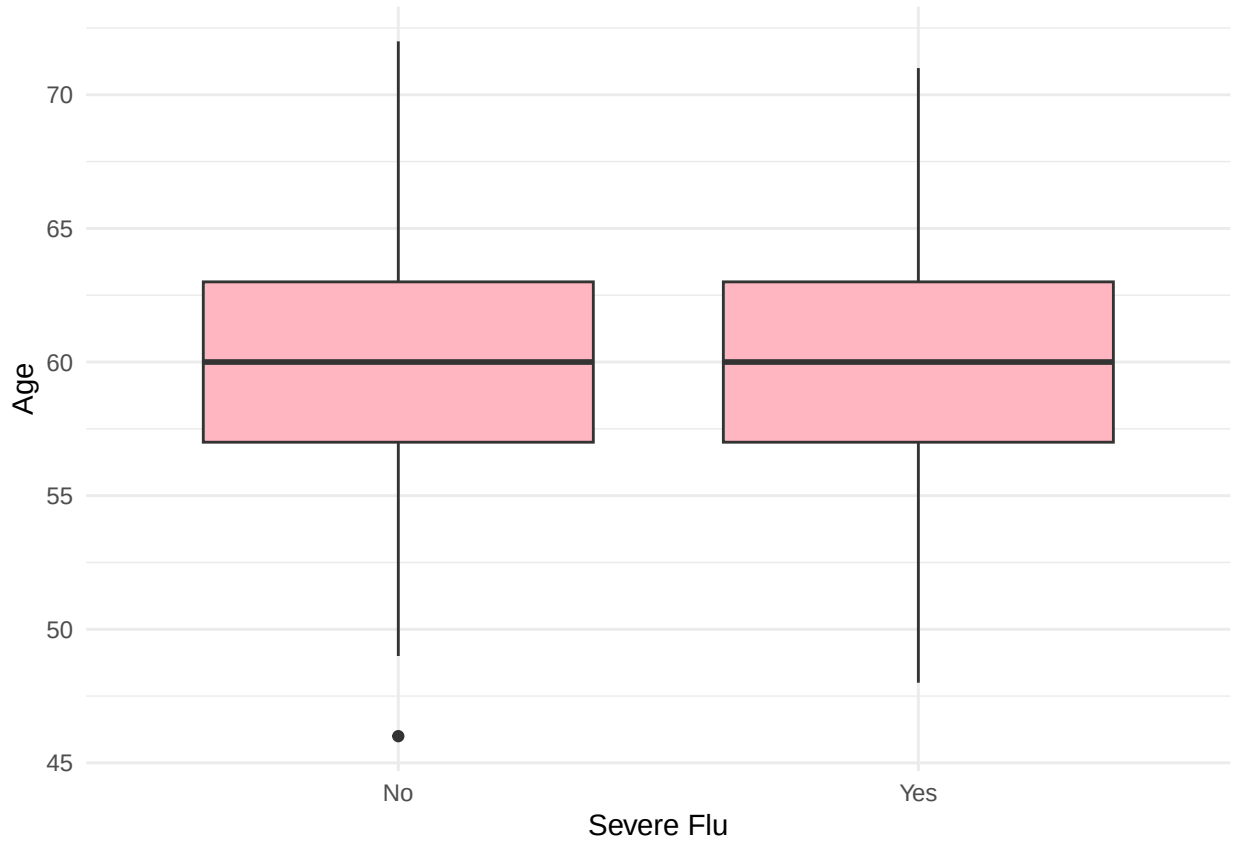
Assessing correlation among continuous variables

```
cor(flu[, c("age", "height", "weight", "bmi", "SBP", "LDL")])
```

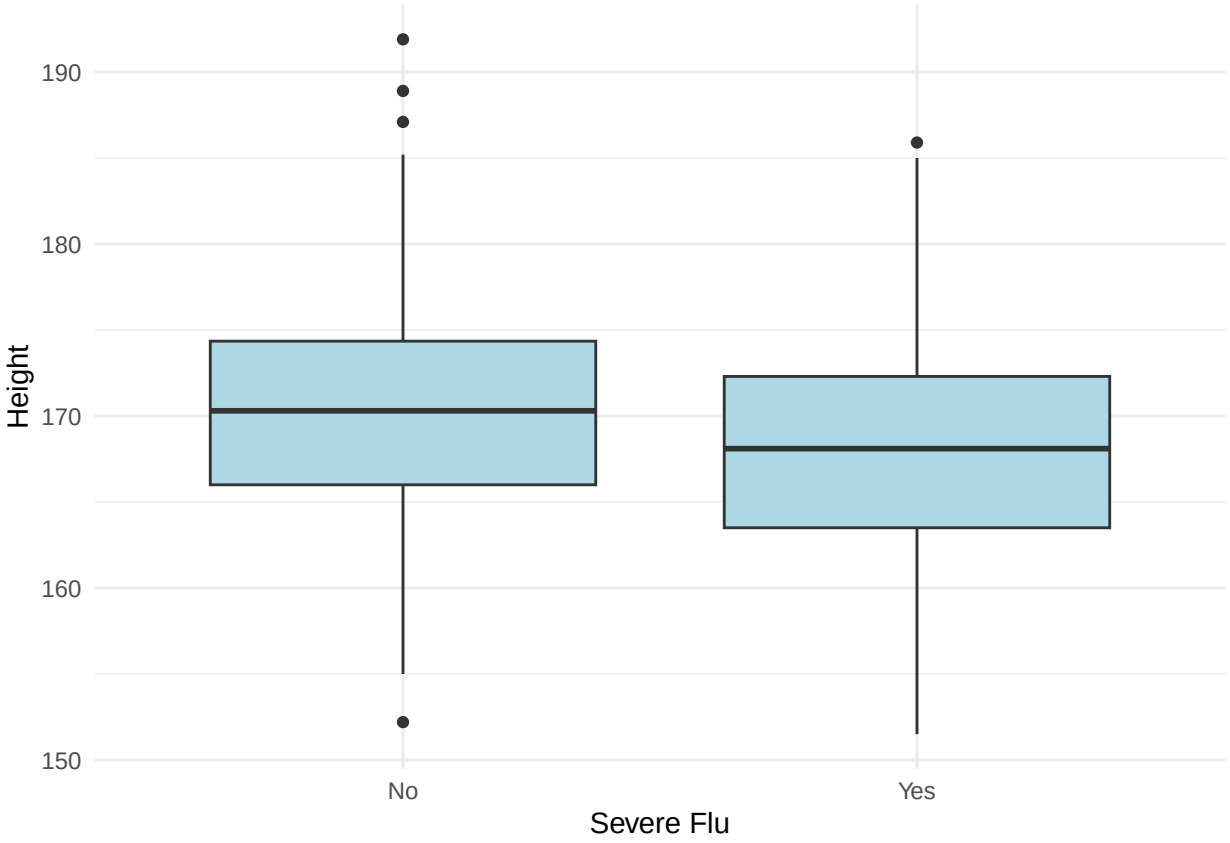
```
##                age      height       weight          bmi          SBP
## age     1.00000000  0.01794456 -0.029940909 -0.04158528  0.44027512
## height  0.01794456  1.00000000  0.267637393 -0.48933944  0.03143295
## weight -0.02994091  0.26763739  1.000000000  0.70666669 -0.01929686
## bmi    -0.04158528 -0.48933944  0.706666689  1.00000000 -0.04120141
## SBP     0.44027512  0.03143295 -0.019296863 -0.04120141  1.00000000
## LDL     0.20742590  0.01832110 -0.001534474 -0.01566285  0.24444416
##                 LDL
## age     0.207425901
## height  0.018321100
## weight -0.001534474
## bmi    -0.015662850
## SBP     0.244444156
## LDL     1.000000000
```

Asses relationship between severe flu and continous variables

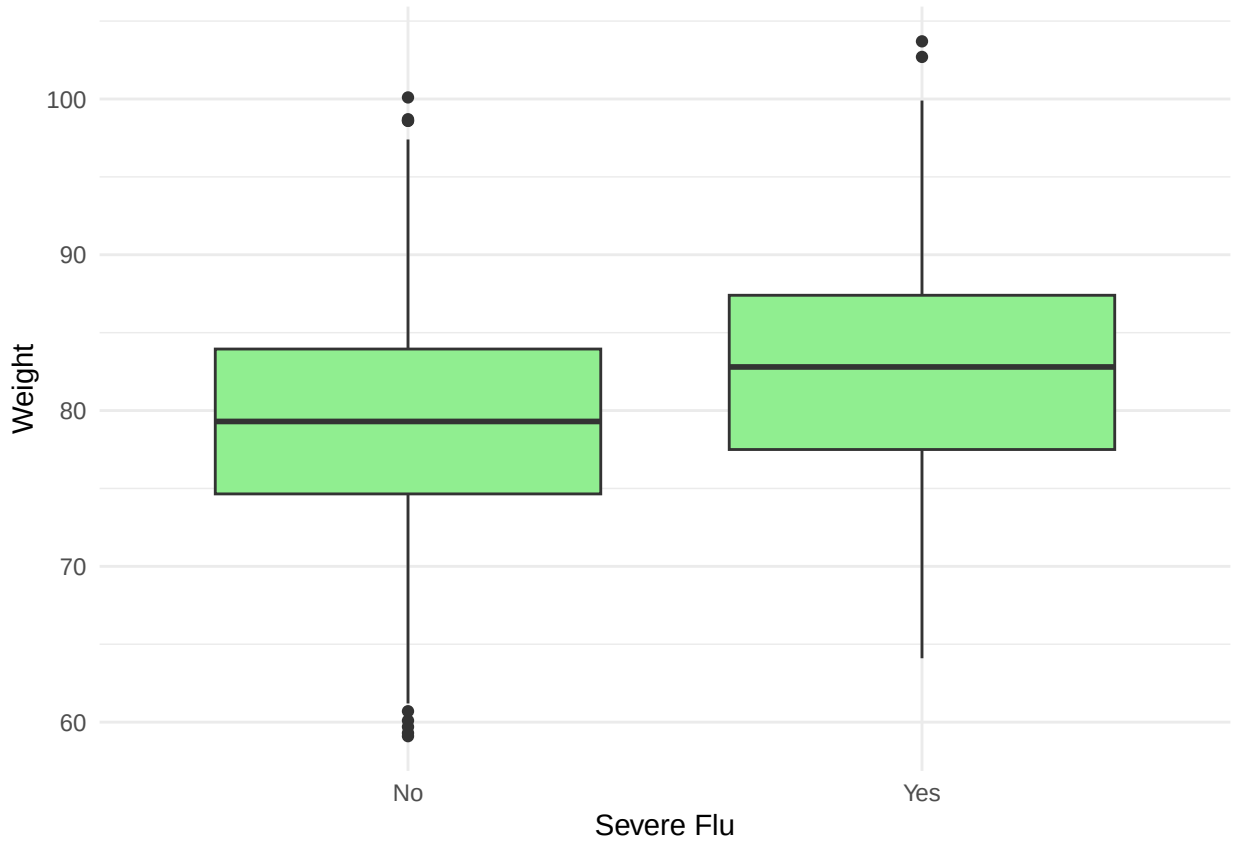
```
ggplot(flu, aes(x = severe_flu, y = age)) +
  geom_boxplot(fill = "lightpink") +
  theme_minimal() +
  labs(x = "Severe Flu", y = "Age")
```



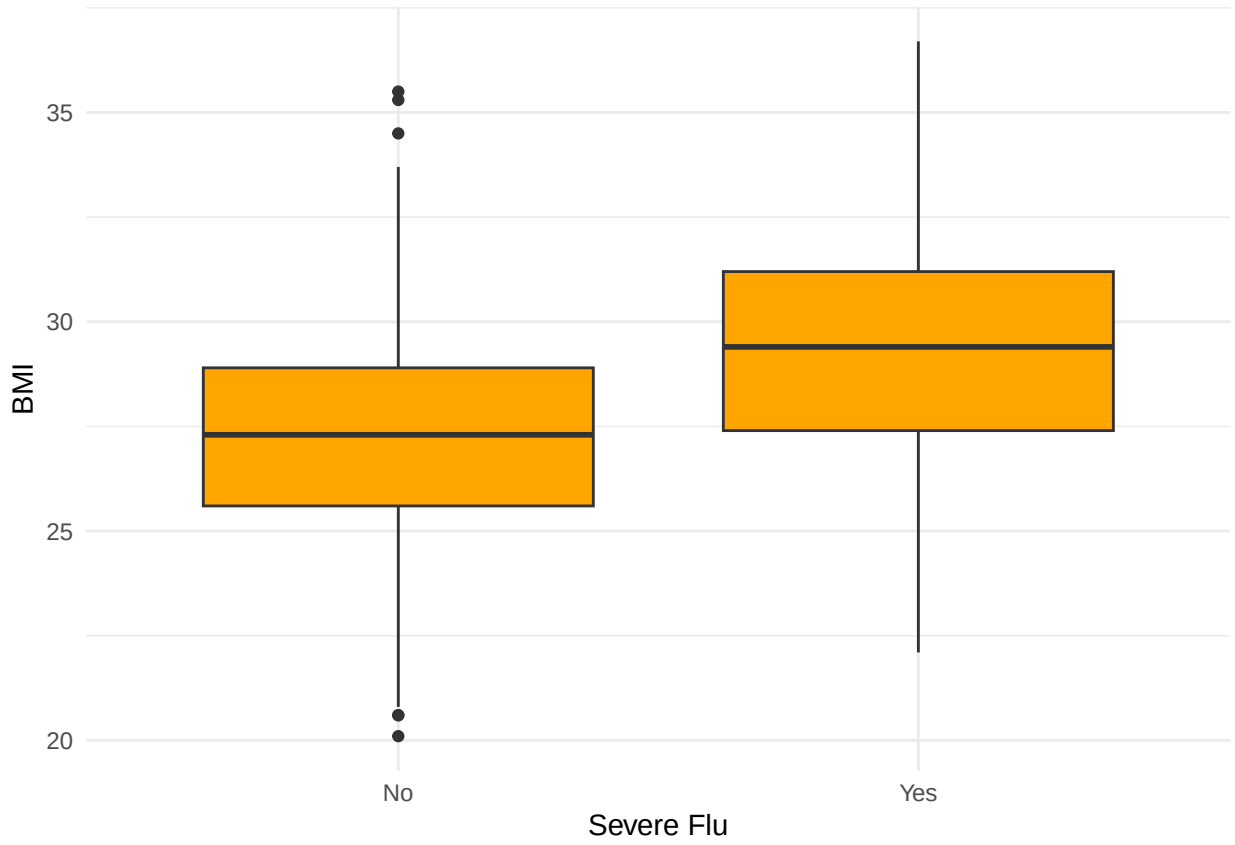

```
ggplot(flu, aes(x = severe_flu, y = height)) +
  geom_boxplot(fill = "lightblue") +
  theme_minimal() +
  labs(x = "Severe Flu", y = "Height")
```



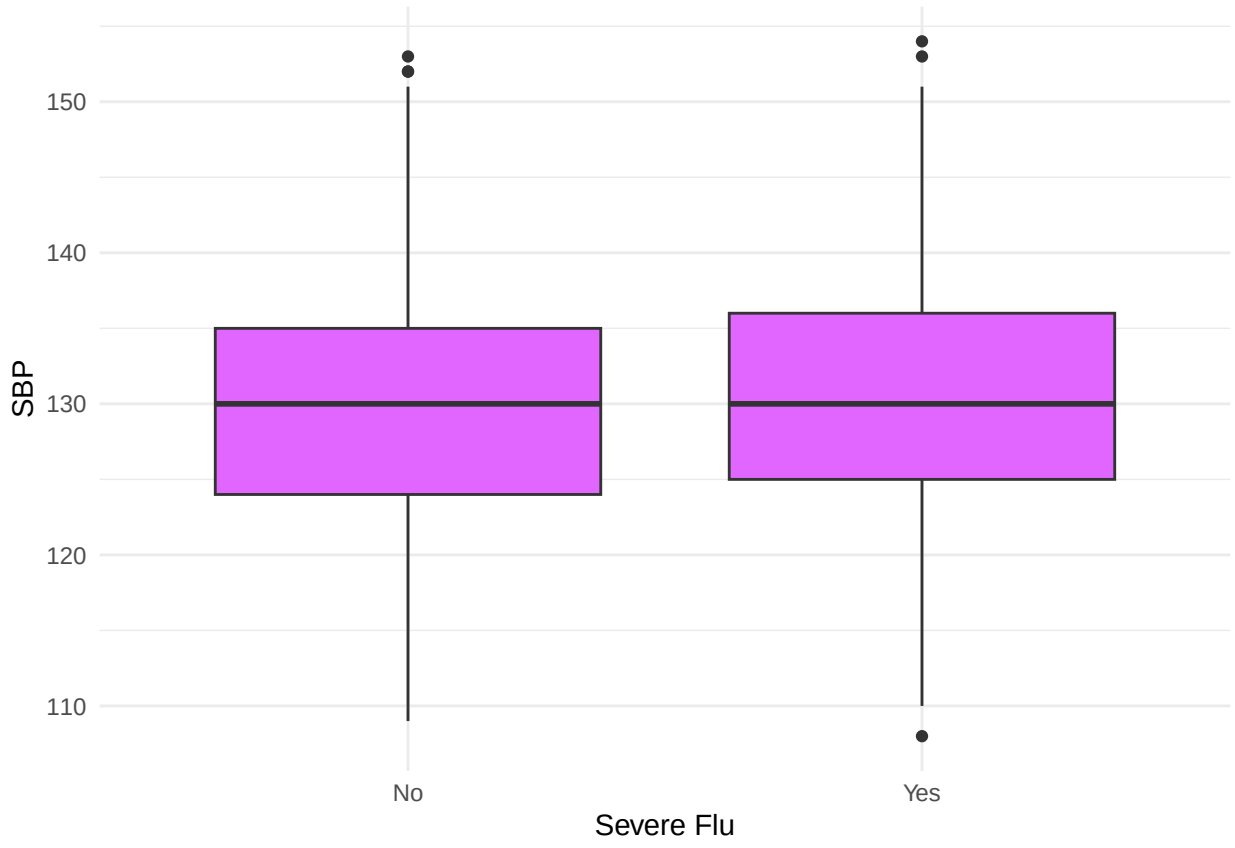

```
ggplot(flu, aes(x = severe_flu, y = weight)) +
  geom_boxplot(fill = "lightgreen") +
  theme_minimal() +
  labs(x = "Severe Flu", y = "Weight")
```



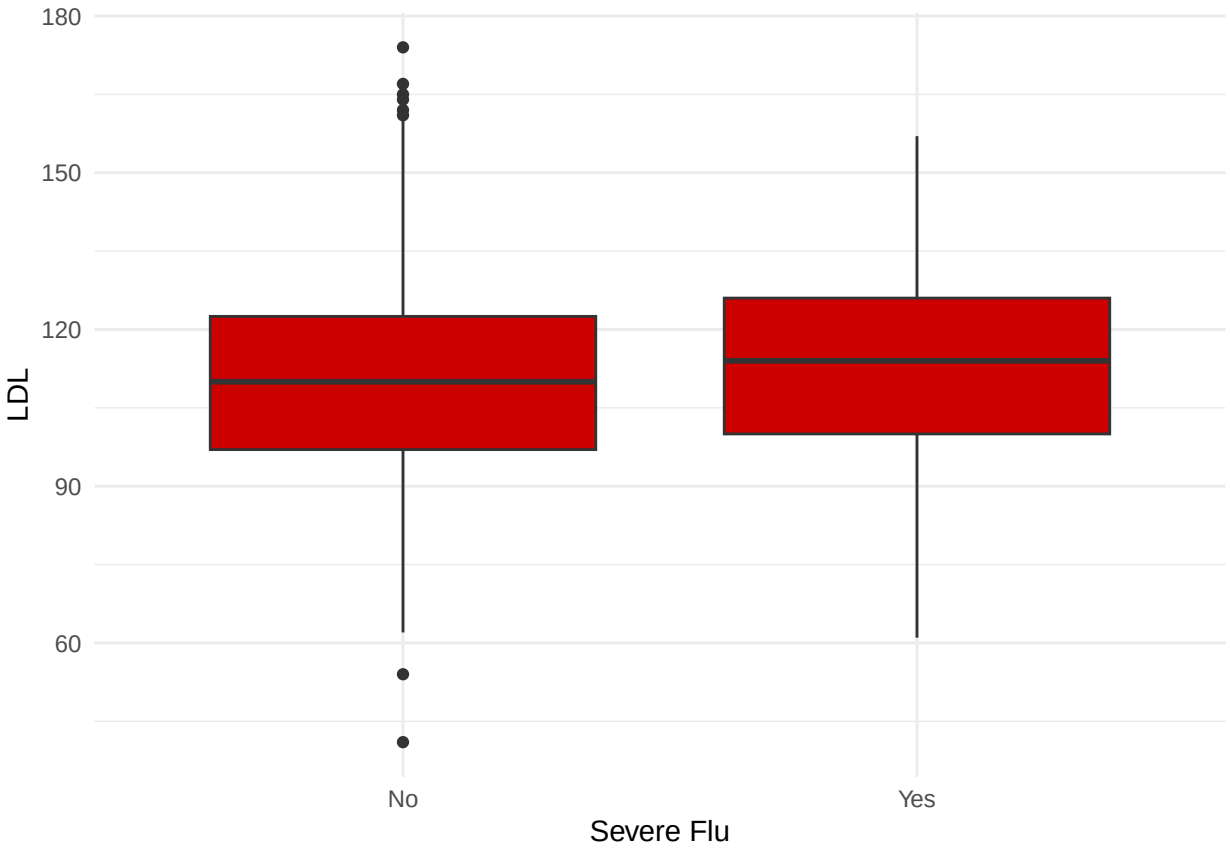

```
ggplot(flu, aes(x = severe_flu, y = bmi)) +
  geom_boxplot(fill = "orange") +
  theme_minimal() +
  labs(x = "Severe Flu", y = "BMI")
```




```
ggplot(flu, aes(x = severe_flu, y = SBP)) +
  geom_boxplot(fill = "mediumorchid1") +
  theme_minimal() +
  labs(x = "Severe Flu", y = "SBP")
```




```
ggplot(flu, aes(x = severe_flu, y = LDL)) +
  geom_boxplot(fill = "red3") +
  theme_minimal() +
  labs(x = "Severe Flu", y = "LDL")
```



Part 1: Evaluating whether boosting and SVM provide superior predictive performance compared to simpler models.

Part 2: Developing a predictive risk score (i.e., the predicted probability) that quantifies the chance of experiencing severe flu based on individual participant characteristics.

Part 3: Identifying key demographic and clinical factors that predict the risk of severe flu and assessing how these factors influence the risk.