Body Type Prediction - Based on okcupid.com User Information

Contents

Introduction

```
library(readr)
okc <- read_csv("https://uofi.box.com/shared/static/oy32nc373w4jqz3kummksnw6wvhfrl7a.csv",
    col_types = cols(last_online = col_datetime(format = "%Y-%m-%d-%H-%M")))
colnames(okc) <- tolower(colnames(okc))
```

Analysis

Data Cleaning

Before the analysis, we did a few data cleanings with our dataset.

```
library(tidyverse)

# remove the 10 essay related variables
okc.keep = colnames(okc)[grepl("essay", colnames(okc)) == F]
okc.clean = select(okc, okc.keep)

# remove speaks,sign, last_online and location, they don't
# seem to have relationships with body_type
okc.clean = subset(okc.clean, select = -c(speaks, sign, last_online,
    location, income))

# remove abnormal height; we define normal heights as between
# 55 and 80 inches
okc.clean = filter(okc.clean, height >= 55 & height <= 80)

# remove 'rather not say' & 'used up' in body_type category
okc.clean = filter(okc.clean, body_type != "rather not say" &
    body_type != "used up")

# get a summary of how many NAs each variables have
colnames.okc.clean = colnames(okc.clean)
summary.NAs = data.frame(colnames = colnames.okc.clean, NAs = sapply(1:length(colnames.okc.clean),
    function(i) sum(is.na(okc.clean[, i]))))
# summary.NAs remove variables with NA >= 10,000
okc.keep = summary.NAs$colnames[which(summary.NAs$NAs < 10000)]
okc.clean = select(okc.clean, okc.keep)
```

```
character.vars = lapply(okc.clean, class) == "character"
okc.clean[, character.vars] = lapply(okc.clean[, character.vars],
    as.factor)

# remove NAs in the dataset
okc.clean = na.omit(okc.clean)
dim(okc.clean)
```

```
## [1] 38374    11
```

```
summary(okc.clean)
```

```
##       age                 body_type            drinks
##  Min.   :18.00   average       :10447   desperately:  160
##  1st Qu.:26.00   fit           : 8999   not at all : 2254
##  Median :30.00   athletic      : 8286   often      : 3312
##  Mean   :32.85   thin          : 3277   rarely     : 4216
##  3rd Qu.:38.00   curvy         : 2798   socially   :28147
##  Max.   :69.00   a little extra: 1998   very often :  285
##                  (Other)       : 2569
##                              education                      ethnicity
##  graduated from college/university:16910   white                  :23673
##  graduated from masters program   : 6659   asian                  : 4231
##  working on college/university    : 4184   hispanic / latin       : 1859
##  graduated from two-year college  : 1238   black                  : 1444
##  working on masters program       : 1159   other                  : 1111
##  graduated from high school       : 1129   hispanic / latin, white:  911
##  (Other)                          : 7095   (Other)                : 5145
##      height                                  job          orientation    sex
##  Min.   :55.00   other                          : 5453   bisexual: 1634   f:15096
##  1st Qu.:66.00   student                        : 3713   gay     : 3530   m:23278
##  Median :68.00   science / tech / engineering   : 3657   straight:33210
##  Mean   :68.36   computer / hardware / software: 3551
##  3rd Qu.:71.00   sales / marketing / biz dev    : 3329
##  Max.   :80.00   artistic / musical / writer    : 3144
##                  (Other)                        :15527
##             smokes                 status
##  no            :31399   available     : 1144
##  sometimes     : 2505   married       :  172
##  trying to quit:  987   seeing someone: 1215
##  when drinking : 2151   single        :35839
##  yes           : 1332   unknown       :    4
##
##
```

```
# recategorize variable factors
smokes = ifelse(okc.clean$smokes == "yes", "yes", "no")
okc.clean$smokes = as.factor(smokes)
status = ifelse(okc.clean$status == "available" | okc.clean$status ==
    "single", "available", "not available")
okc.clean$status = as.factor(status)

okc.clean$body_type[okc.clean$body_type == "a little extra"] = "overweight"
okc.clean$body_type[okc.clean$body_type == "curvy"] = "overweight"
okc.clean$body_type[okc.clean$body_type == "full figured"] = "overweight"
```

```
okc.clean$body_type[okc.clean$body_type == "athletic"] = "fit"
okc.clean$body_type[okc.clean$body_type == "jacked"] = "fit"
okc.clean$body_type[okc.clean$body_type == "skinny"] = "thin"
okc.clean$body_type = droplevels(okc.clean$body_type)

okc.clean$ethnicity = as.character(okc.clean$ethnicity)
okc.clean$ethnicity[sapply(okc.clean$ethnicity, function(x) grepl(",",
    x)) != 0] = "mixed"
okc.clean$ethnicity = as.factor(okc.clean$ethnicity)
levels(okc.clean$ethnicity)
```

```
##  [1] "asian"            "black"            "hispanic / latin" "indian"
##  [5] "middle eastern"   "mixed"            "native american"  "other"
##  [9] "pacific islander" "white"
```

```
summary(okc.clean)
```

```
##       age             body_type             drinks
##  Min.   :18.00   average   :10447   desperately:  160
##  1st Qu.:26.00   fit       :17541   not at all : 2254
##  Median :30.00   overweight: 5918   often      : 3312
##  Mean   :32.85   thin      : 4468   rarely     : 4216
##  3rd Qu.:38.00                      socially   :28147
##  Max.   :69.00                      very often :  285
##
##                               education                ethnicity         height
##  graduated from college/university:16910   white           :23673   Min.   :55.00
##  graduated from masters program   : 6659   mixed           : 4759   1st Qu.:66.00
##  working on college/university    : 4184   asian           : 4231   Median :68.00
##  graduated from two-year college  : 1238   hispanic / latin: 1859   Mean   :68.36
##  working on masters program       : 1159   black           : 1444   3rd Qu.:71.00
##  graduated from high school       : 1129   other           : 1111   Max.   :80.00
##  (Other)                          : 7095   (Other)         : 1297
##                                job          orientation    sex       smokes
##  other                           : 5453   bisexual: 1634   f:15096   no :37042
##  student                         : 3713   gay     : 3530   m:23278   yes: 1332
##  science / tech / engineering    : 3657   straight:33210
##  computer / hardware / software: 3551
##  sales / marketing / biz dev     : 3329
##  artistic / musical / writer     : 3144
##  (Other)                         :15527
##            status
##  available    :36983
##  not available: 1391
##
##
##
##
##
```

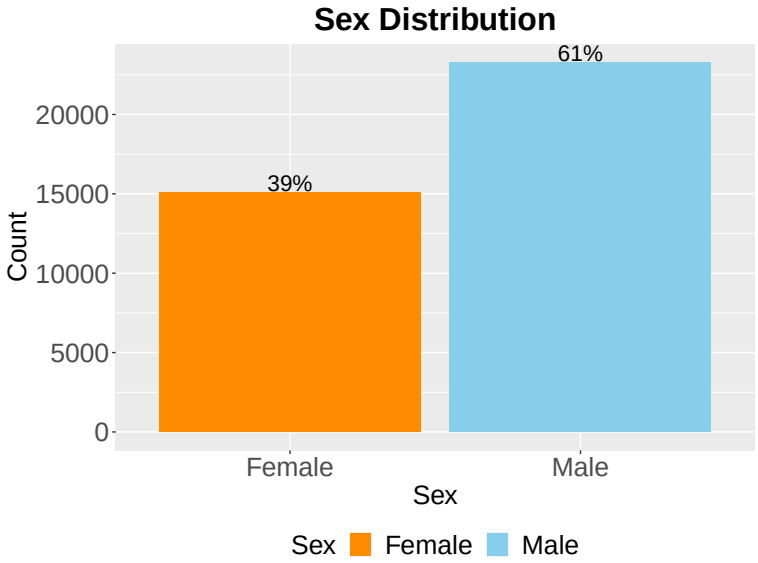
```
rm(character.vars, colnames.okc.clean, okc.keep, summary.NAs,
    smokes, status)
```

Descriptive Statistics

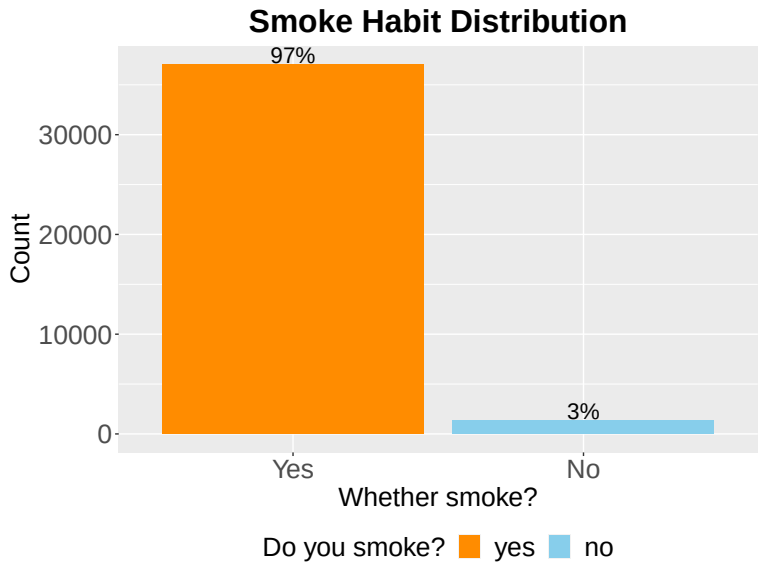
```
# Descriptive Statistics
```

Data Visualizations

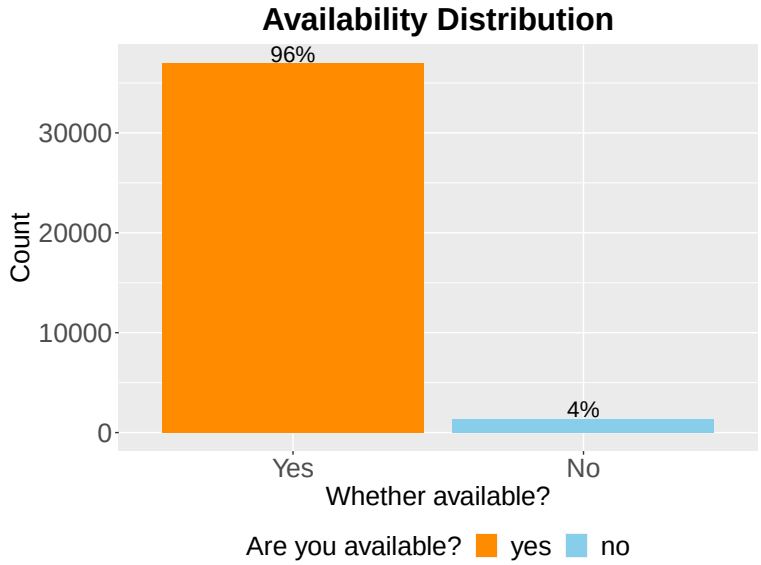
```
# Data Visualization Part 1
library(ggplot2)
library(tidyverse)
library(dplyr)

# attributes to use in the codes
size.no.title = 20
size.title = 24
size.text = 6
colors = c("darkorange","skyblue")

# Distribution of sex
sex.dist_table = okc.clean %>% group_by(sex) %>%
  dplyr::summarise(n = n())%>%
  dplyr::mutate(percent = scales::percent(n/sum(n)))

sex.dist <- ggplot(data = sex.dist_table) +
  geom_histogram(mapping = aes(x = sex, y = n,fill = sex),
                 stat = "identity", position = "identity") +
  scale_fill_manual("Sex", values = colors,
                    labels = c("Female","Male")) +
  ggtitle(label = "Sex Distribution") +
  xlab("Sex") + ylab("Count") +
  scale_x_discrete(breaks = c("f","m"),labels = c("Female","Male")) +
  geom_text(aes(x = sex, y = n, fill = sex, label = percent),
            vjust = -0.1, size =size.text, color = "black") +
    theme(legend.position="bottom",
          legend.text = element_text(size = size.no.title),
        legend.title = element_text(size = size.no.title),
        plot.title = element_text(hjust = 0.5, face = "bold", size = size.title),
        axis.text = element_text(size = size.no.title),
        axis.title = element_text(size = size.no.title))
sex.dist
```



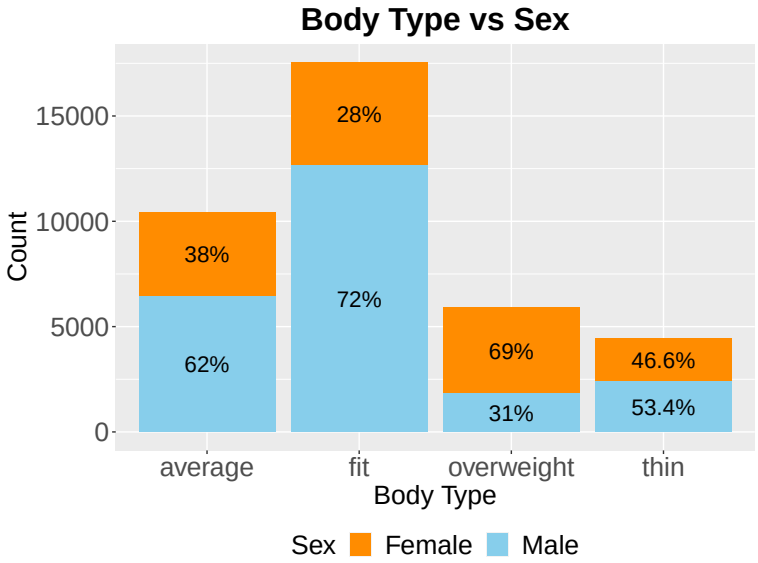

```
# Distribution of smokes
smokes.dist_table = okc.clean %>% group_by(smokes) %>%
  dplyr::summarise(n = n())%>%
  dplyr::mutate(percent = scales::percent(n/sum(n)))

smokes.dist <- ggplot(data = smokes.dist_table) +
  geom_histogram(mapping = aes(x = smokes, y = n,fill = smokes),
                 stat = "identity", position = "identity") +
  scale_fill_manual("Do you smoke?", values = colors,
                    labels = c("yes","no")) +
  ggtitle(label = "Smoke Habit Distribution") +
  xlab("Whether smoke?") + ylab("Count") +
  scale_x_discrete(breaks = c("no","yes"),labels = c("Yes","No")) +
  geom_text(aes(x = smokes, y = n, fill = smokes, label = percent),
            vjust = -0.1, size =size.text, color = "black") +
    theme(legend.position="bottom",
          legend.text = element_text(size = size.no.title),
          legend.title = element_text(size = size.no.title),
          plot.title = element_text(hjust = 0.5, face = "bold", size = size.title),
          axis.text = element_text(size = size.no.title),
          axis.title = element_text(size = size.no.title))
smokes.dist
```




```
# Distribution of status
status.dist_table = okc.clean %>% group_by(status) %>%
  dplyr::summarise(n = n())%>%
  dplyr::mutate(percent = scales::percent(n/sum(n)))

status.dist <- ggplot(data = status.dist_table) +
  geom_histogram(mapping = aes(x = status, y = n,fill = status),
                 stat = "identity", position = "identity") +
  scale_fill_manual("Are you available?", values = colors,
                    labels = c("yes","no")) +
  ggtitle(label = "Availability Distribution") +
  xlab("Whether available?") + ylab("Count") +
  scale_x_discrete(breaks = c("available","not available"),labels = c("Yes","No")) +
  geom_text(aes(x = status, y = n, fill = status, label = percent),
            vjust = -0.1, size =size.text, color = "black") +
    theme(legend.position="bottom",
          legend.text = element_text(size = size.no.title),
          legend.title = element_text(size = size.no.title),
        plot.title = element_text(hjust = 0.5, face = "bold", size = size.title),
        axis.text = element_text(size = size.no.title),
        axis.title = element_text(size = size.no.title))
status.dist
```




```
# Body type by sex
sex_table = okc.clean %>% group_by(body_type,sex) %>%
  dplyr::summarise(n = n()) %>% group_by(body_type) %>%
  dplyr::mutate(percent = scales::percent(n/sum(n)))

p.sex <- ggplot(data=sex_table) +
  geom_histogram(mapping = aes(x = body_type, y = n, fill = sex),
                 stat = "identity",position = "stack") +
  scale_fill_manual("Sex", values = colors,
                    labels = c("Female","Male")) +
  ggtitle(label = "Body Type vs Sex") +
    xlab("Body Type") + ylab("Count") +
  geom_text(aes(x = body_type, y = n, fill = sex, label = percent),
            hjust = 0.5, position = position_stack(vjust = 0.5),
            size =size.text, color = "black") +
  theme(legend.position="bottom",
        legend.text = element_text(size = size.no.title),
        legend.title = element_text(size = size.no.title),
        plot.title = element_text(hjust = 0.5, face = "bold", size = size.title),
        axis.text = element_text(size = size.no.title),
        axis.title = element_text(size = size.no.title))
p.sex
```



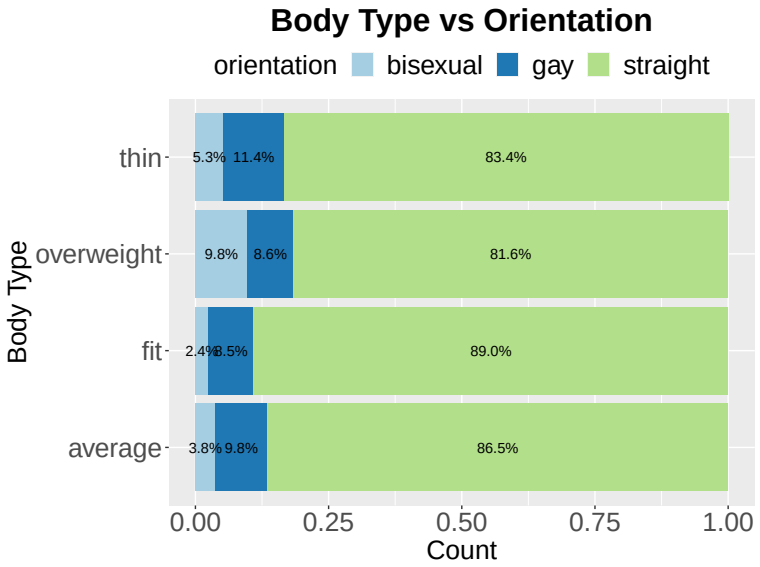
First, we would like to know how the data are distributed in the following three factors, Sex, Smoke Habit and Availability. We would like to see if the data are balanced or unbalanced within each factor. Highly unbalanced data are not informative and will not contribute too much in our prediction model.

1. Sex Distribution: We can see that female and male distribution are about 40% and 60%, respectively. This distribution is considered as balanced. So, we expect sex will provide useful information and contribute to our prediction model.
2. Smoke Habit Distribution: We can see that majority of the users smoke, which is around 97%. The data distributed within this factor is highly unbalanced and we do not expect to see much contribution of this factor in our analysis later.
3. Availability Distribution: Similar to Smoke habit, the data distribution is also very unbalanced. 96% of the users are available. So, we also do not expect this factor to provide much useful information in our analysis.

Based on the above visualizations, we are interested to see how sex is distributed within each body type and visualize if there's any relationship between sex and body type. From the "Body Type vs Sex" plot we can see that body type does have a relationship with sex, especially in "fit" and "overweight" group. We can see that in fit group, 72% of the users are male while only 28% of the users are female. On the contrary, in overweight group, around 70% of the users are female while only 30% of the users are male. This is an obvious trend that males are more confident about their body figure and females are less confident. This is also consistent with what we see in our daily life that women are less satisfied with their body figure and always want to lose more weights. We are also more certain that sex will provide useful information in our analysis and our prediction model.

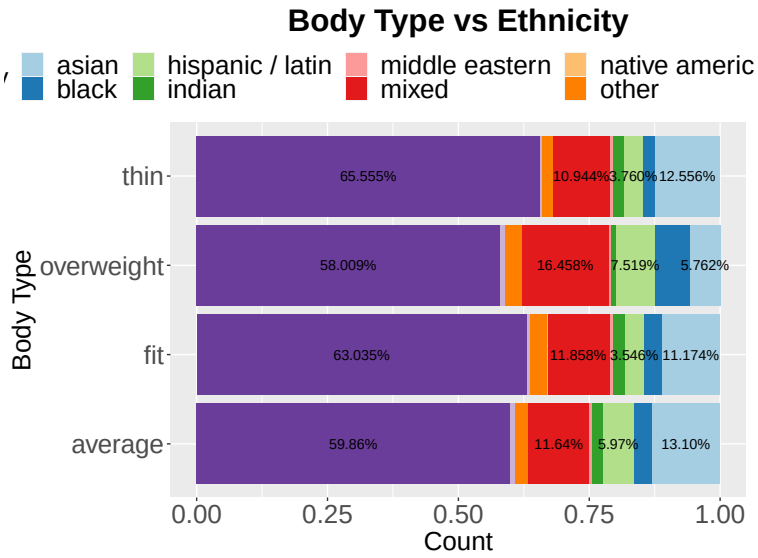
```
library(ggplot2)
library(viridis)

# attributes to use in the code
size.no.title = 20
size.no.title2 = 20
size.title = 24
size.text = 4
colors = c("darkorange","skyblue")

# Body type by orientation
orientation_table = okc.clean %>%
```

```
  group_by(body_type,orientation) %>% dplyr::summarise(n = n()) %>%
  group_by(body_type) %>% dplyr::mutate(percent = scales::percent(n/sum(n)))

p.status <-ggplot(data=orientation_table) +
  geom_bar(mapping = aes(x = body_type, y = n, fill = orientation),
           stat = "identity",position = position_fill(reverse = T)) +
  coord_flip() +
  scale_fill_brewer(palette = "Paired") +
  ggtitle(label = "Body Type vs Orientation") +
    xlab("Body Type") + ylab("Count") +
  geom_text(aes(x = body_type, y = n, fill = orientation, label = percent),
            position = position_fill(vjust = 0.5,reverse = T),
            size = size.text, color = "black") +
  theme(legend.position="top",
        legend.text = element_text(size = size.no.title2),
        legend.title = element_text(size = size.no.title2),
        plot.title = element_text(hjust = 0.5, face = "bold", size = size.title),
        axis.text = element_text(size = size.no.title),
        axis.title = element_text(size = size.no.title))
p.status
```




```
# Body type by ethnicity
body.ethnicity = table(okc.clean$ethnicity,okc.clean$body_type)
ethnicity_table = okc.clean %>%
  group_by(body_type,ethnicity) %>% dplyr::summarise(n = n()) %>%
  group_by(body_type) %>% dplyr::mutate(percent = scales::percent(n/sum(n)))

percent.eth = ethnicity_table$percent
percent.eth[c(2,4:5,7:9,12,14:15,17:19,22,24:25,27:29,32,34:35,37:39)] = ""
ethnicity_table$percent2 = percent.eth

p.ethnicity <-ggplot(data=ethnicity_table) +
  geom_bar(mapping = aes(x = body_type, y = n, fill =ethnicity),
           stat = "identity",position = "fill") +
  coord_flip() +
  scale_fill_brewer(palette = "Paired") +
```

```
  ggtitle(label = "Body Type vs Ethnicity") +
    xlab("Body Type") + ylab("Count") +
  geom_text(aes(x = body_type,y = n, fill = ethnicity, label = percent.eth),
            position = position_fill(vjust =0.5),
            size =size.text, color = "black") +
  theme(legend.position="top",
        legend.text = element_text(size = size.no.title2),
        legend.title = element_text(size = size.no.title2),
        plot.title = element_text(hjust = 0.5, face = "bold", size = size.title),
        axis.text = element_text(size = size.no.title),
        axis.title = element_text(size = size.no.title))
p.ethnicity
```



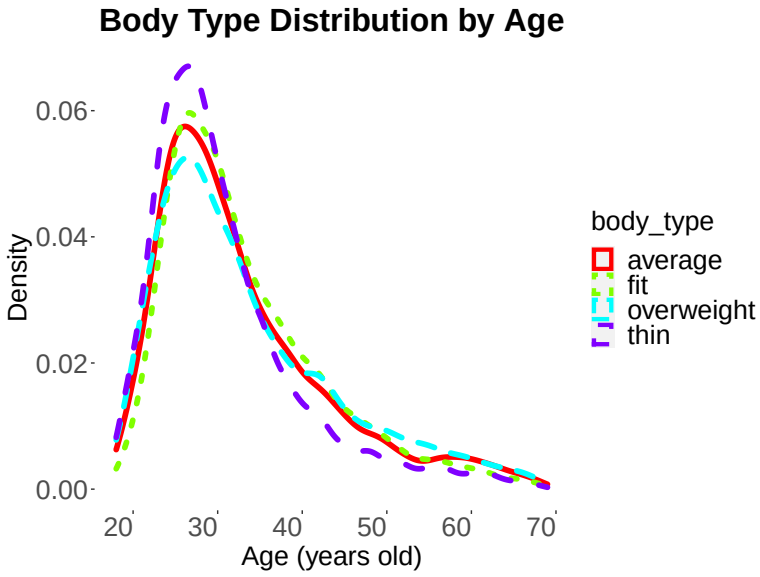
Second, we would like to take a look at two other factors, Orientation and Ethnicity, and their relationships with Body Type.

1. Body Type vs. Orientation: We do not see very obvious trend between body type and orientation. But we can still see some orientation distribution differences within different body types. For example, we can see that the bisexual group has a higher percentage (9.8%) in overweight type compared to other three body types. So, a person in bisexual group is more likely to be predicted to body type of overweight than to other three types.

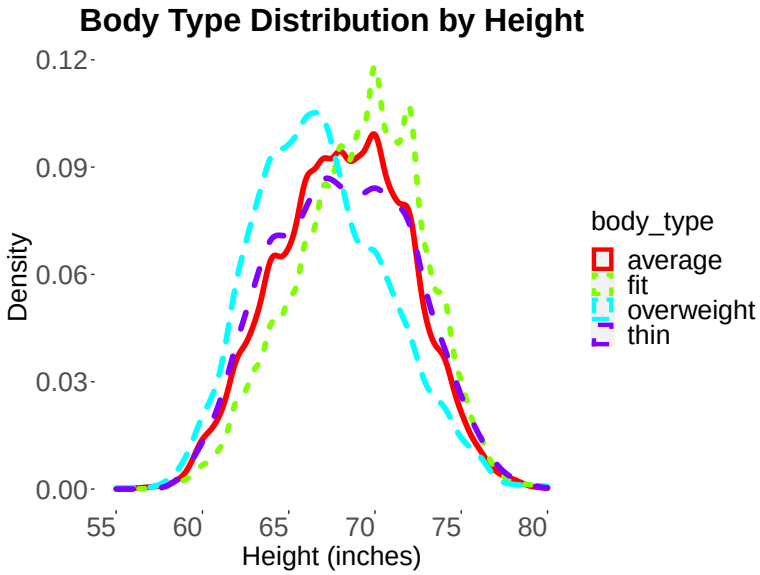
2. Body Type vs. Ethnicity: We can see that there are some body type differences within each race. For example, the percentage of Asian group in overweight type is only 5.762%, which is around half of the percentage compared to other three types. Therefore, if a user is Asian, it is less likely to predict that person to overweight than to other three types.

```
library(ggplot2)
library(viridis)
library(RColorBrewer)

size.no.title = 20
size.title = 24
size.text = 8
colors = c("darkorange","skyblue")

# body type by age
```

```
p.age <- ggplot(okc.clean, aes(x = age, color = body_type)) +
  geom_density(size = 2,aes(linetype = body_type)) +
  labs(title ="Body Type Distribution by Age", x = "Age (years old)", y = "Density") +
  scale_color_manual(values = rainbow(4)) +
  theme(axis.text.x = element_text(hjust = 1, size = size.no.title),
        axis.text.y = element_text(size = size.no.title),
        plot.title = element_text(size = size.title,face = "bold",hjust = 0.5),
        axis.title = element_text(size = size.no.title),
        legend.title = element_text(size = size.no.title),
        legend.text = element_text(size = size.no.title),
        panel.background = element_blank())

p.age
```




```
# body type by height
p.height <- ggplot(okc.clean, aes(x = height, color = body_type)) +
  geom_density(size = 2,aes(linetype = body_type)) +
  labs(title ="Body Type Distribution by Height", x = "Height (inches)", y = "Density") +
  scale_color_manual(values = rainbow(4)) +
  theme(axis.text.x = element_text(hjust = 1, size = size.no.title),
        axis.text.y = element_text(size = size.no.title),
        plot.title = element_text(size = size.title,face = "bold",hjust = 0.5),
        axis.title = element_text(size = size.no.title),
        legend.title = element_text(size = size.no.title),
        legend.text = element_text(size = size.no.title),
        panel.background = element_blank())
p.height
```



Third, we would like to see how Age and Height would affect Body Type.

1. Body Type Distribution by Age: We can see from the plot that the majority of the users in all body types are between 20 – 40 years old. Besides, we can see that the highest peak of all density lines is the around 28 years old and is the "thin" group (which is the purple line). This line dramatically goes down at around 35 years old and then become the lowest line afterwards. These trends are all consistent with our daily observations that older people tend to gain weights easier. Since we can see some different trends within different boy type group, we are expected to see the contribution of age to our model.

2. Body Type Distribution by Height: We can see that there are two peaks in the plot: one is around 66 inches with the overfit group (which is the blue line); the other is around 70 inches with the fit group (which is the green line). Following this trend, we are expected to see that using 66 inches as the base line, the taller the person is compared to the base line, the more likely he or she will be predicted in the fit group rather than the overweight group. Besides, this trend can also be explained along with Sex factor we talked about earlier: females are much more likely to classify themselves as overweight than males do. And 66 inches is more likely to be a height of a woman rather than man. Therefore, if we combine both factors, the prediction could be more accurate.