STAT 6324 Final Project

Art Tay

```r
data_full <- read.csv("heart_2020_cleaned.csv", stringsAsFactors = T)
```

Data Dictionary (From Kaggle):

HeartDisease : Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI).

BMI : Body Mass Index (BMI). (Continuous)

Smoking : Have you smoked at least 100 cigarettes in your entire life? (The answer Yes or No). (Cat - bin)

AlcoholDrinking : Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week (Cat - bin)

Stroke : (Ever told) (you had) a stroke? (Cat - bin)

PhysicalHealth : Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good? (0-30 days).

MentalHealth : Thinking about your mental health, for how many days during the past 30 days was your mental health not good? (0-30 days).

DiffWalking : Do you have serious difficulty walking or climbing stairs? (Cat - bin)

Sex : Are you male or female? (Cat - bin)

AgeCategory: Fourteen-level age category. (Cat) + (Add order?)

Race : Imputed race/ethnicity value. (Cat)

Diabetic : (Ever told) (you had) diabetes? (Cat)

PhysicalActivity : Adults who reported doing physical activity or exercise during the past 30 days other than their regular job. (Cat - bin)

GenHealth : Would you say that in general your health is… (Cat) + (Add order)

SleepTime : On average, how many hours of sleep do you get in a 24-hour period? (Numeric) + (Add Bins)

Asthma : (Ever told) (you had) asthma? (Cat - bin)

KidneyDisease : Not including kidney stones, bladder infection or incontinence, were you ever told you had kidney disease? (Cat - bin)

SkinCancer : (Ever told) (you had) skin cancer? (Cat - bin)

Percentage Plots

```r
# Code to check the class imbalance of the response.
table_0 <- data_full %>% group_by(HeartDisease) %>%
           count() %>% ungroup() %>%
           mutate(Percent = n / sum(n) * 100)
```

```r
# Function to plot the percentage of heart disease a across a given
# categorical variable.
plot_percentages <- function(cat, font = 18, ylimit = 25){
  table <- data_full %>%
           group_by(HeartDisease) %>%
           count(!!sym(cat)) %>%
           group_by(!!sym(cat)) %>%
           mutate(percent = n / sum(n) * 100) %>%
           filter(HeartDisease == "Yes")

  plot <- table %>% ggplot(
    aes(x = reorder(!!sym(cat), percent),
        y = percent, fill = percent)
  ) +
  geom_bar(stat = "identity") +
  scale_fill_gradient(low = "skyblue", high = "darkblue") +
  ylim(0, ylimit) +
  xlab(cat) +
  ylab("") +
  scale_x_discrete(guide = guide_axis(n.dodge = 2)) +
  theme_bw() +
  theme(legend.position = "none",
        text = element_text(size = font))
}
```

```r
# Demographic Plots
plot_age <- plot_percentages("AgeCategory")
plot_race <- plot_percentages("Race") + ylab("Percent Heart Disease")
plot_sex <- plot_percentages("Sex")

#Lifestyle Plots
plot_smoke <- plot_percentages("Smoking") +
  theme(axis.text.y = element_blank(), axis.ticks = element_blank())
plot_phyad <- plot_percentages("PhysicalActivity") +
  theme(axis.text.y = element_blank(), axis.ticks = element_blank())
plot_alcd <- plot_percentages("AlcoholDrinking") + ylab("Percent Heart Disease")

# Medical Issue Plots
plot_diab <- plot_percentages("Diabetic", ylimit = 40)
plot_Stroke <- plot_percentages("Stroke", ylimit = 40)
plot_kid <- plot_percentages("KidneyDisease", ylimit = 40)
plot_skin <- plot_percentages("SkinCancer", ylimit = 40)
```

```r
# Sleep Plot
sleep_table <- data_full %>% select(HeartDisease, SleepTime) %>%
               mutate(SleepTime = as.factor(SleepTime)) %>%
               group_by(HeartDisease) %>%
               count(SleepTime) %>%
               group_by(SleepTime) %>%
               mutate(percent = n / sum(n) * 100) %>%
               filter(HeartDisease == "Yes") %>%
               mutate(regular_sleep = as.factor(
                ifelse(SleepTime %in% c("6", "7", "8", "9"), 1, 0)))
```

```
plot_sleep <- sleep_table %>% ggplot(
  aes(x = SleepTime, y = percent, fill = regular_sleep)
) +
geom_bar(stat = 'identity') +
scale_fill_manual(values = c("darkblue", "gray")) +
theme_bw() +
theme(legend.position = "none", text = element_text(size = 18)) +
ylab("Percent Heart Disease")
```

Interaction Plots

```
# Function to create interaction plots between two categorical
# and the response Heart Disease.
# @param cat1 the string name of the first categorical variable.
# @param cat2 the string name of the second categorical variable.
# @return a ggplot object.
interact_plot <- function(cat1, cat2, ymin = 0, ymax = 30) {
  data <- data_full %>%
    select(HeartDisease, !!sym(cat1), !!sym(cat2)) %>%
    group_by(HeartDisease) %>%
    count(!!sym(cat2), !!sym(cat1)) %>%
    ungroup() %>%
    group_by(!!sym(cat1), !!sym(cat2)) %>%
    mutate(Percent = n / sum(n) * 100) %>%
    filter(HeartDisease == "Yes")

  plot <- data %>%
    ggplot(
      aes(x = !!sym(cat1), y = Percent,
        color = !!sym(cat2), group = !!sym(cat2))
    ) +
    geom_point() +
    geom_line() +
    scale_x_discrete(guide = guide_axis(n.dodge = 2)) +
    theme_bw() +
    theme(text = element_text(size = 18)) +
    ylab("Percent Heart Diseased") +
    ylim(ymin, ymax)

  return(plot)
}
```

```
# Uses the function to make interaction plots.
plot_Sex_Alc <- interact_plot("Sex", "AlcoholDrinking") +
  theme(legend.position = "none")

plot_Smoke_Alc <- interact_plot("Smoking", "AlcoholDrinking") +
  theme(legend.position = "none") + ylab("") +
    theme(axis.text.y = element_blank(), axis.ticks = element_blank())

plot_Alc_Kidney <- interact_plot("KidneyDisease", "AlcoholDrinking") +
    theme(axis.text.y = element_blank(), axis.ticks = element_blank()) +
    ylab("")

plot_Smoke_As <- interact_plot("Smoking", "Asthma", ymax = 20)
plot_race_sex <- interact_plot("Race", "Sex", ymax = 20)
```

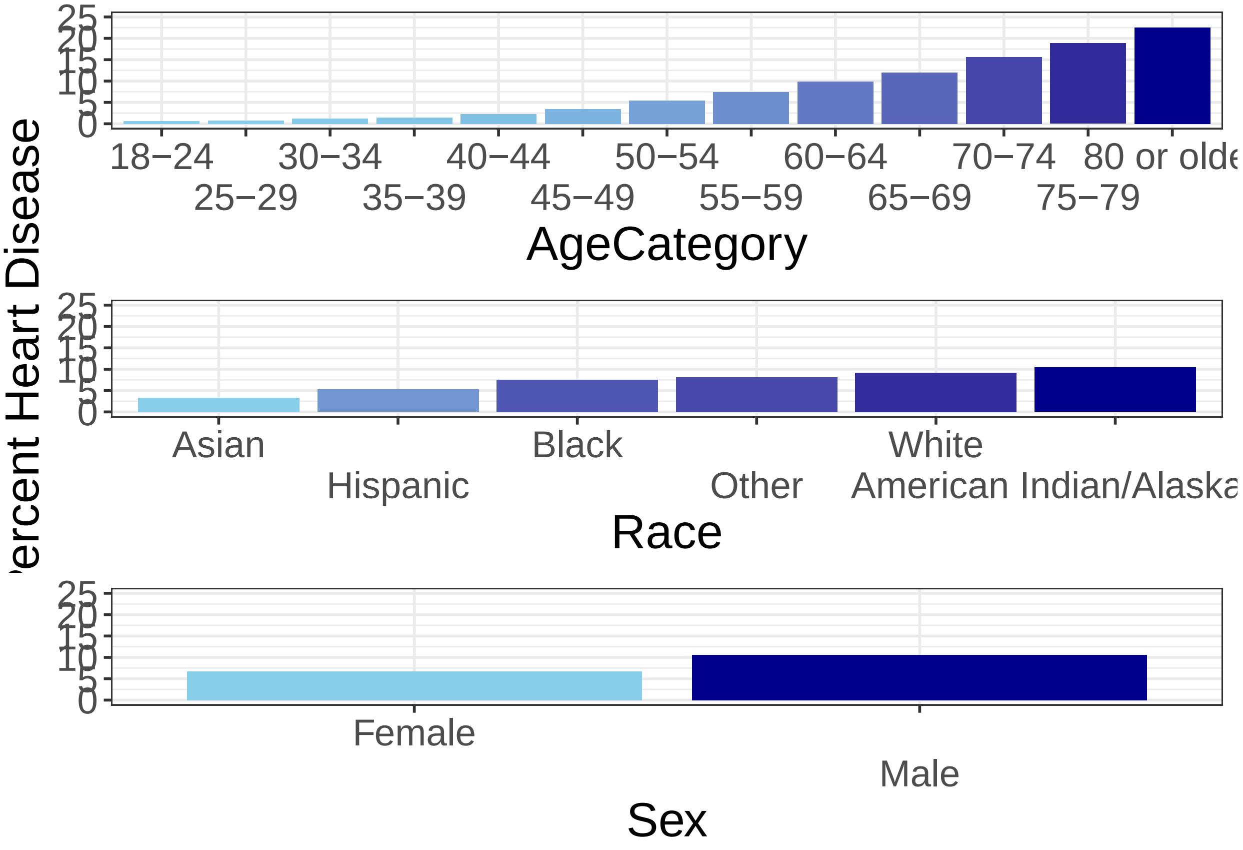
```
# Interaction plot between BMI and Physical Activity
data_bmi_phyact <- data_full %>%
  select(HeartDisease, BMI, PhysicalActivity) %>%
  group_by(HeartDisease, PhysicalActivity) %>%
  summarise(mean = mean(BMI), sd = sd(BMI))

plot_bmi_phyact <- data_bmi_phyact %>%
  ggplot(
    aes(y = mean, x = PhysicalActivity,
      color = HeartDisease, group = HeartDisease)
  ) +
  geom_point() +
  geom_line() +
  ylim(25, 35) +
  ylab("BMI") +
  theme_bw() +
  theme(text = element_text(size = 18))
```

Arranged Plots

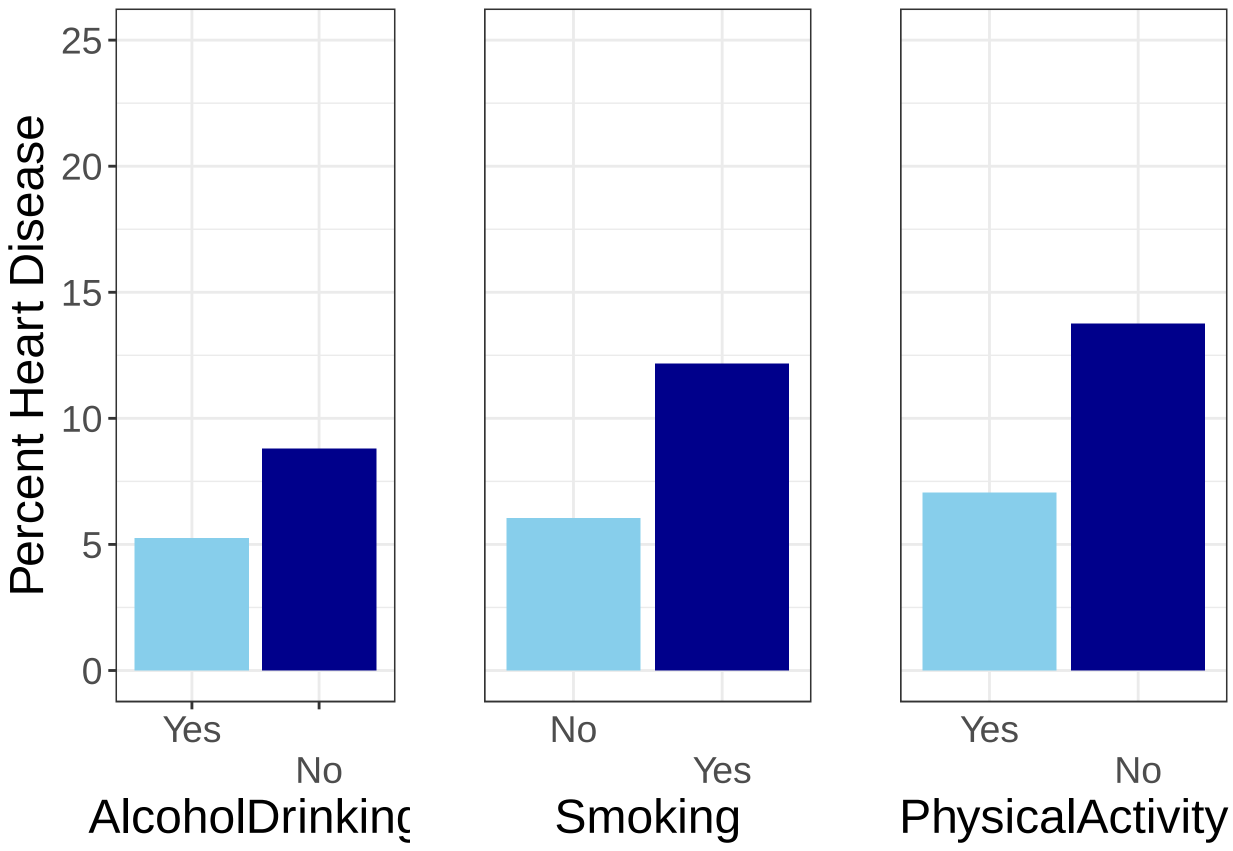
```
# Code for arranging multiple plot together in a grid for presentation.
demo_plots <- ggarrange(plot_age, plot_race, plot_sex, ncol = 1, nrow = 3)

lifestyle_plots <- ggarrange(plot_alcd, plot_smoke,
  plot_phyad, ncol = 3, nrow = 1)

med_plots <- ggarrange(plot_Stroke, plot_kid,
  plot_skin, plot_diab, ncol = 2, nrow = 2)

plots_alco <- ggarrange(plot_Sex_Alc, plot_Smoke_Alc, plot_Alc_Kidney,
  ncol = 3, nrow = 1)

plots_interactions_other <- ggarrange(plot_Smoke_As,
  plot_race_sex, plot_bmi_phyact, ncol = 1, nrow = 3)
```

Plot Check

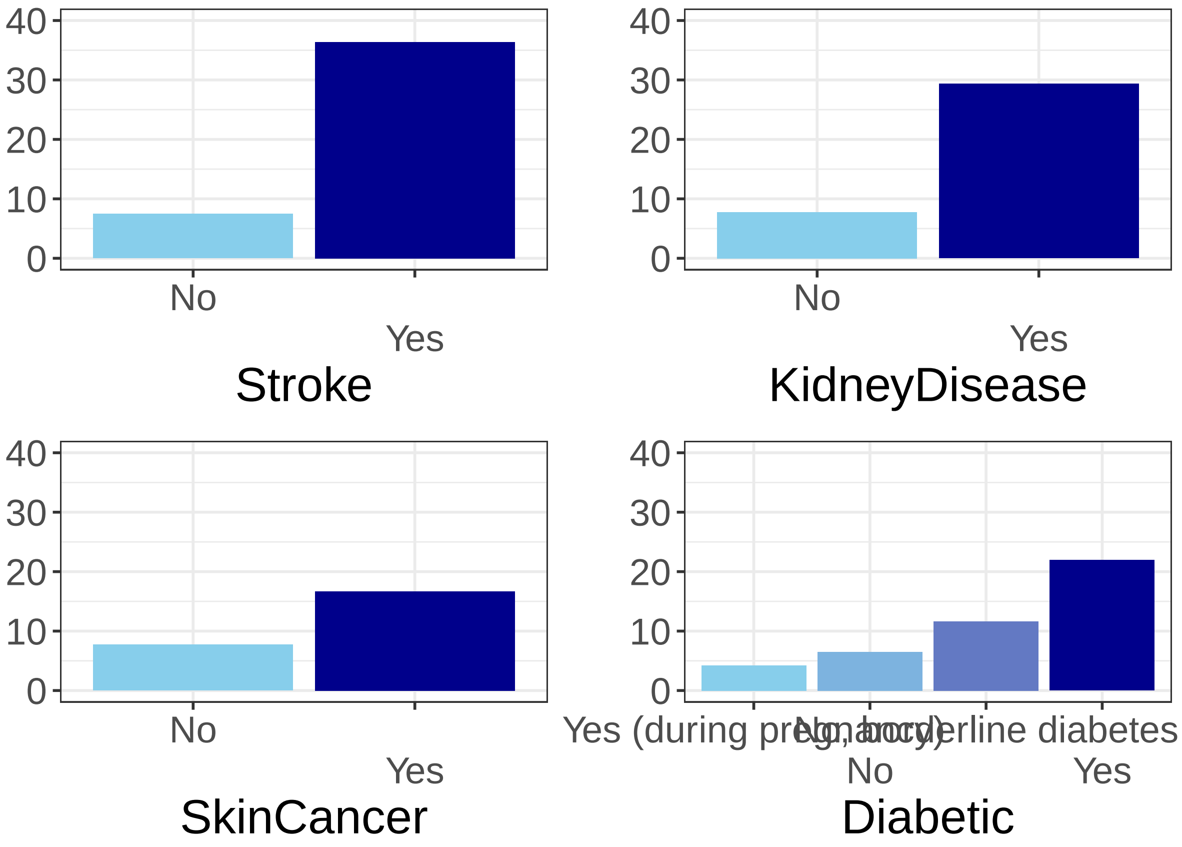
```
demo_plots
```



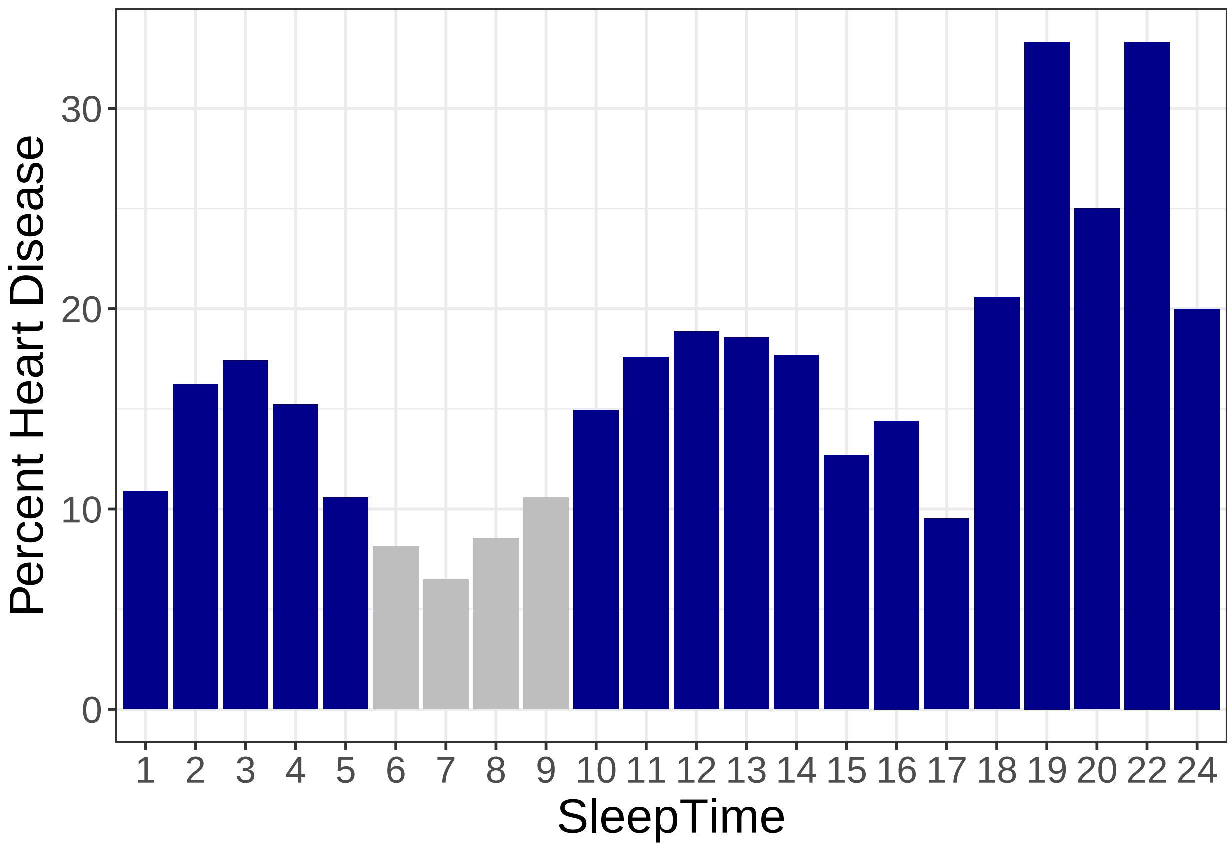
```
lifestyle_plots
```


Percent Heart Disease
25
20
15
10
5
0
Yes
No
AlcoholDrinkinç
No
Yes
Smoking
Yes
No
PhysicalActivity

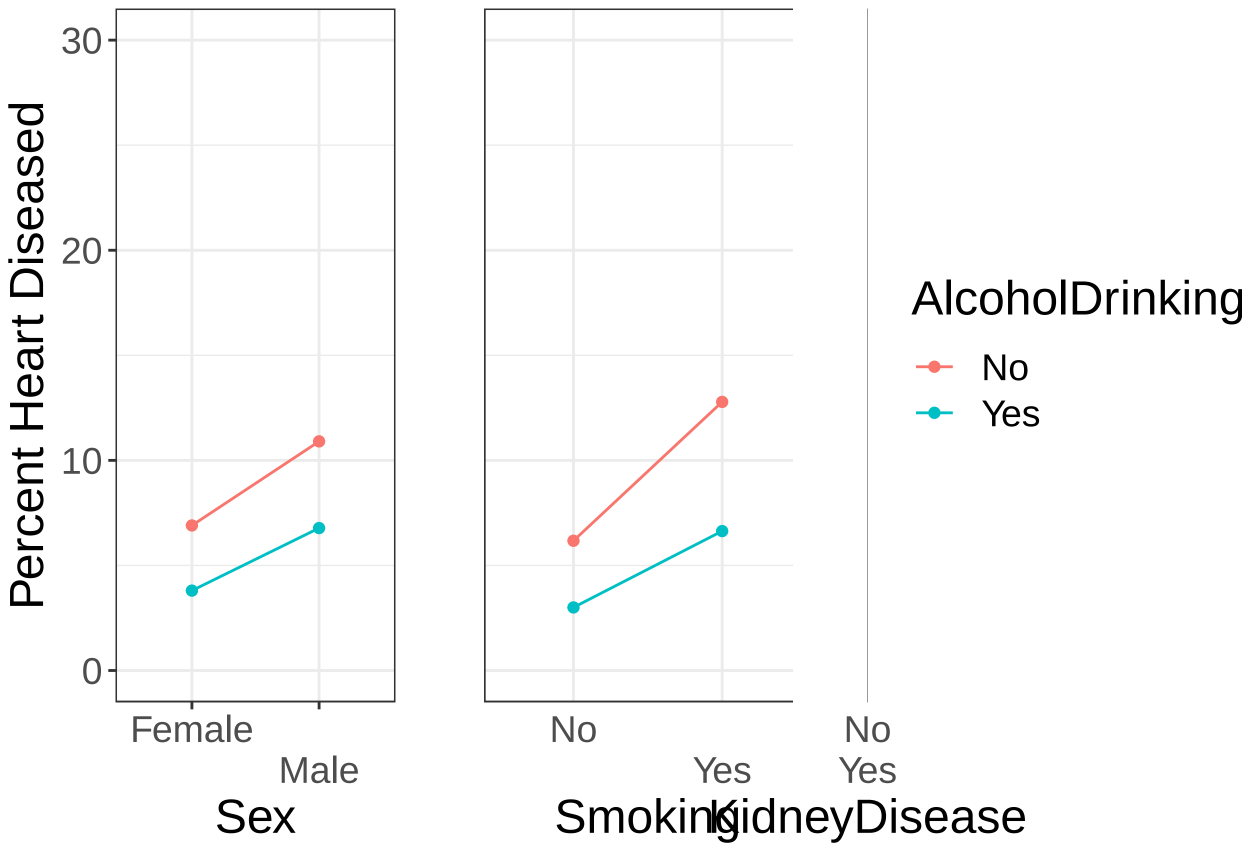
```
med_plots
```


40
30
20
10
0
No
Yes
Stroke
40
30
20
10
0
No
Yes
KidneyDisease
40
30
20
10
0
No
Yes
SkinCancer
40
30
20
10
0
Yes (during pregnancy)
No, borderline diabetes
No
Yes
Diabetic

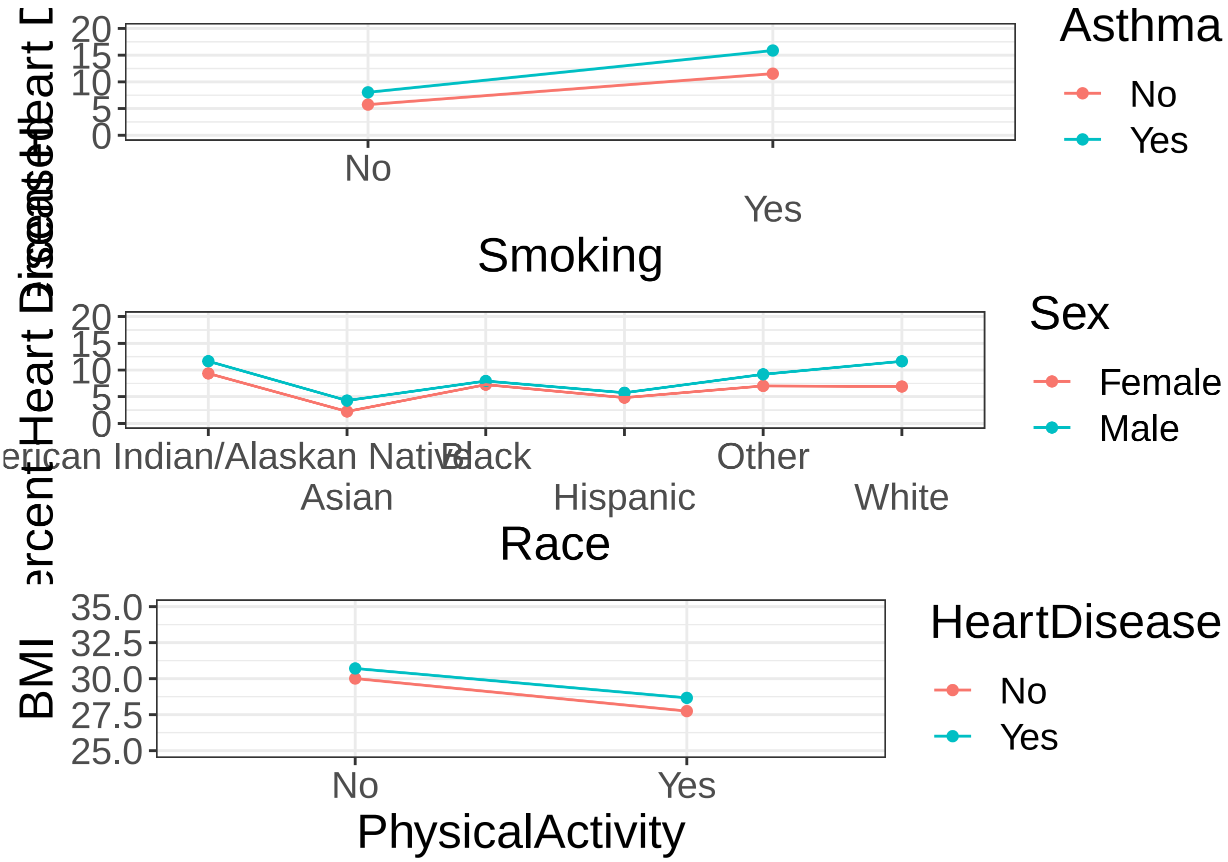
```
plot_sleep
```


Percent Heart Disease
0
10
20
30
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
22
24
SleepTime

```
plots_alco
```



```
plots_interactions_other
```


Heart D
ercent Heart Disease
20
15
10
5
0
No
Yes
Smoking
Asthma
No
Yes
20
15
10
5
0
erican Indian/Alaskan Native
Asian
Black
Hispanic
Other
White
Race
Sex
Female
Male
BMI
35.0
32.5
30.0
27.5
25.0
No
Yes
PhysicalActivity
HeartDisease
No
Yes