Project_Prctice

Safiya Geelani

2022-10-07

```
library(tidyverse)
library(ggplot2)
library(GGally)
library(ggpubr)
library(openintro)
# babies
```

Read in the data set and basic descriptive statistics

```
glimpse(babies)
```

```
## Rows: 1,236
## Columns: 8
## $ case      <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 1~
## $ bwt       <int> 120, 113, 128, 123, 108, 136, 138, 132, 120, 143, 140, 144, ~
## $ gestation <int> 284, 282, 279, NA, 282, 286, 244, 245, 289, 299, 351, 282, 2~
## $ parity    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ age       <int> 27, 33, 28, 36, 23, 25, 33, 23, 25, 30, 27, 32, 23, 36, 30, ~
## $ height    <int> 62, 64, 64, 69, 67, 62, 62, 65, 62, 66, 68, 64, 63, 61, 63, ~
## $ weight    <int> 100, 135, 115, 190, 125, 93, 178, 140, 125, 136, 120, 124, 1~
## $ smoke     <int> 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1, 1, 0, 0, 1, 1, 0, 1, ~
```

```
summary(babies)
```

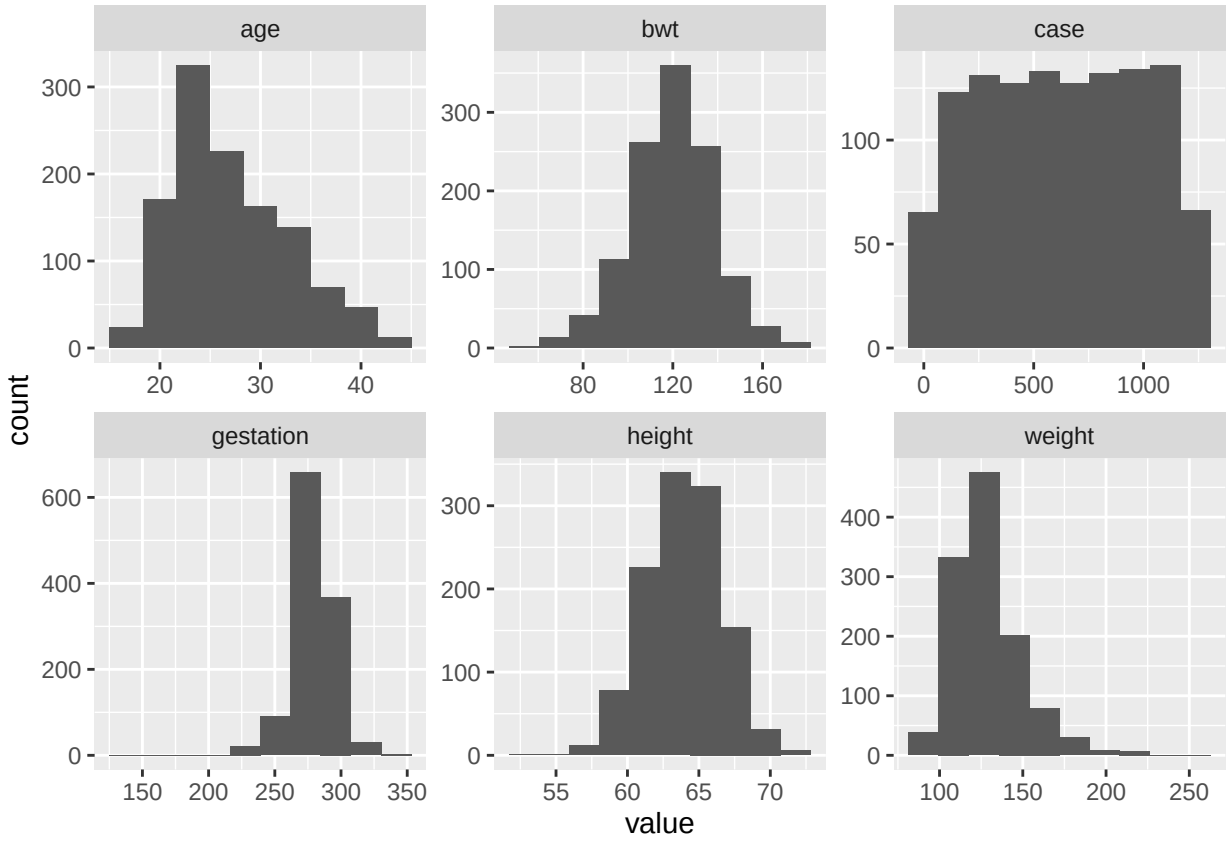
```
##       case             bwt          gestation         parity
##  Min.   :   1.0   Min.   : 55.0   Min.   :148.0   Min.   :0.0000
##  1st Qu.: 309.8   1st Qu.:108.8   1st Qu.:272.0   1st Qu.:0.0000
##  Median : 618.5   Median :120.0   Median :280.0   Median :0.0000
##  Mean   : 618.5   Mean   :119.6   Mean   :279.3   Mean   :0.2549
##  3rd Qu.: 927.2   3rd Qu.:131.0   3rd Qu.:288.0   3rd Qu.:1.0000
##  Max.   :1236.0   Max.   :176.0   Max.   :353.0   Max.   :1.0000
##                                   NA's   :13
##       age            height          weight          smoke
##  Min.   :15.00   Min.   :53.00   Min.   : 87.0   Min.   :0.0000
##  1st Qu.:23.00   1st Qu.:62.00   1st Qu.:114.8   1st Qu.:0.0000
##  Median :26.00   Median :64.00   Median :125.0   Median :0.0000
##  Mean   :27.26   Mean   :64.05   Mean   :128.6   Mean   :0.3948
##  3rd Qu.:31.00   3rd Qu.:66.00   3rd Qu.:139.0   3rd Qu.:1.0000
##  Max.   :45.00   Max.   :72.00   Max.   :250.0   Max.   :1.0000
##  NA's   :2       NA's   :22      NA's   :36      NA's   :10
```

Cleaning the data

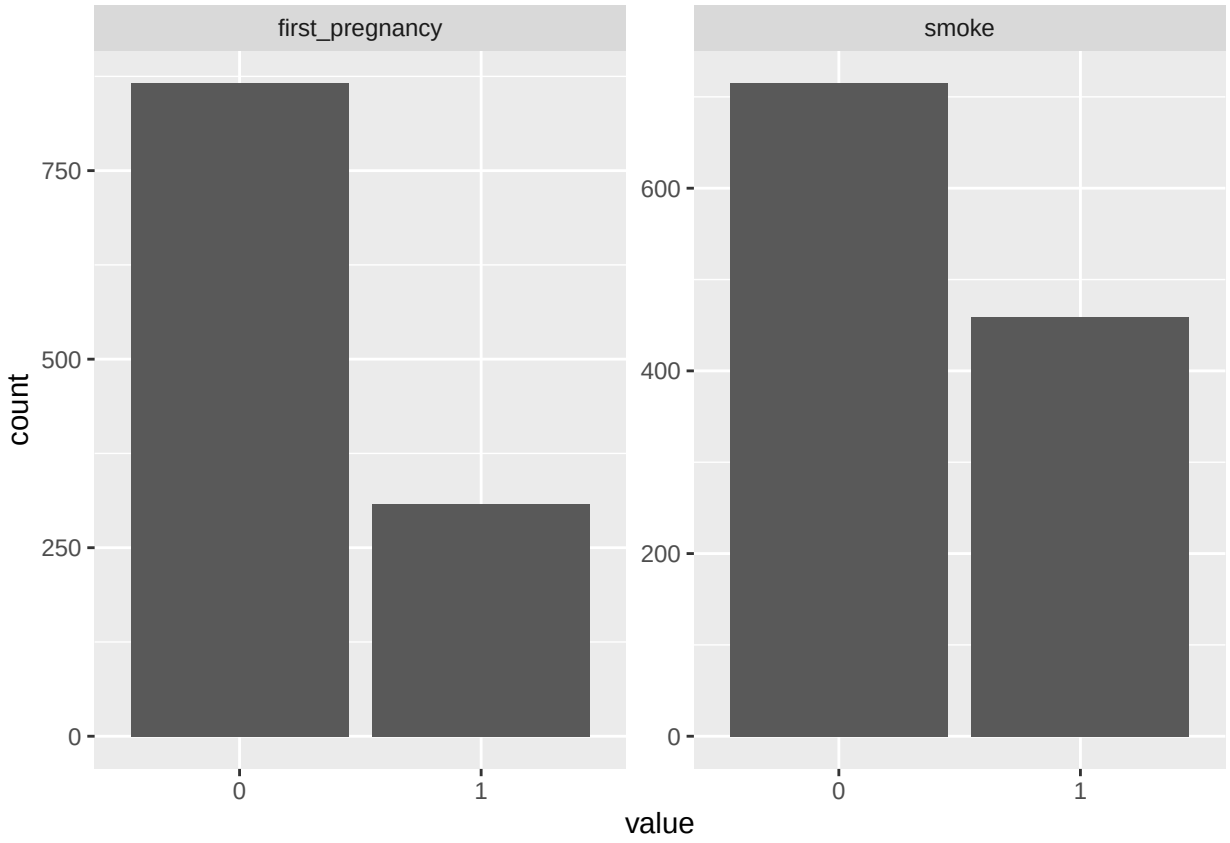
```
babies_data <- babies %>%
  na.omit() %>%
  mutate_at(vars(smoke, parity), as.factor) %>%
  rename(first_pregnancy = parity) # 0 is first pregnancy
```

Visualizing the distribution of predictors

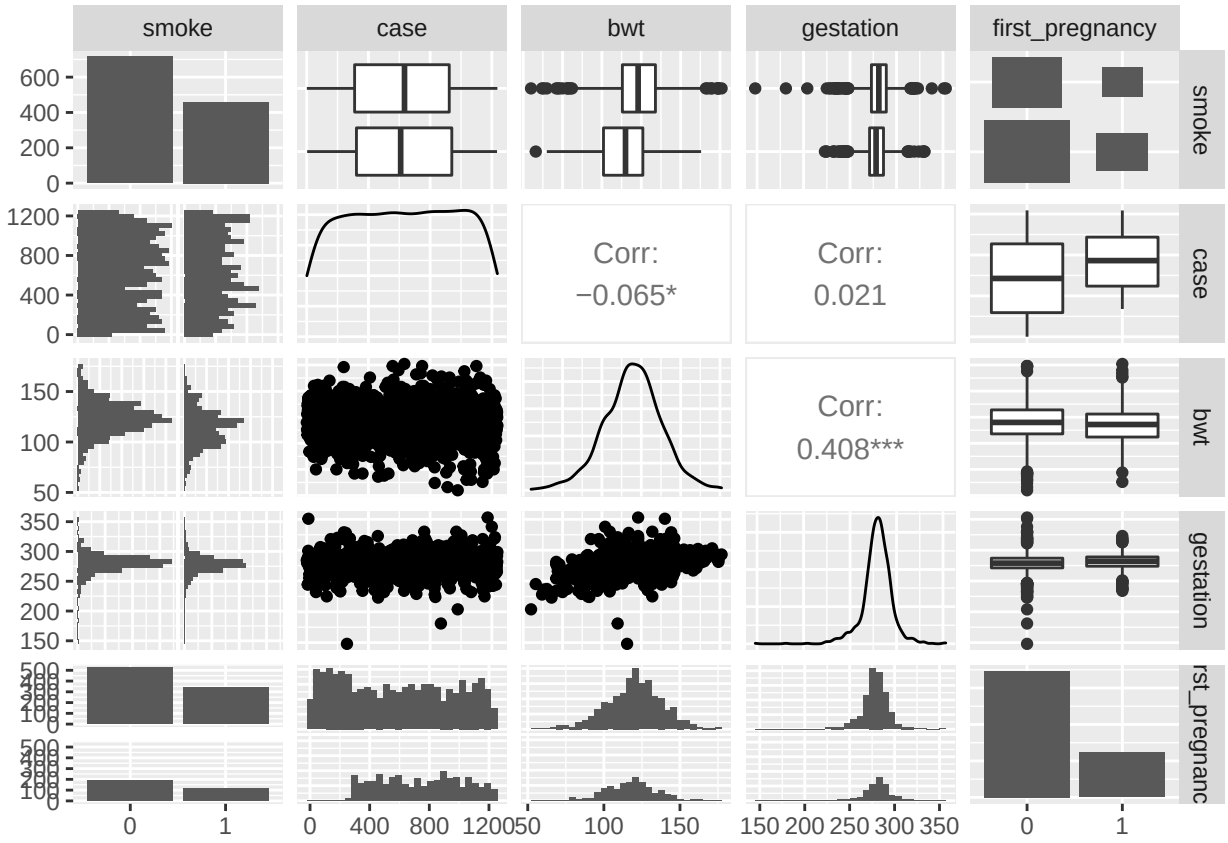
```
# Histogram for numeric variables
babies_data %>% select_if(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
  geom_histogram(bins = 10) +
  facet_wrap(~ key, scales = "free")
```



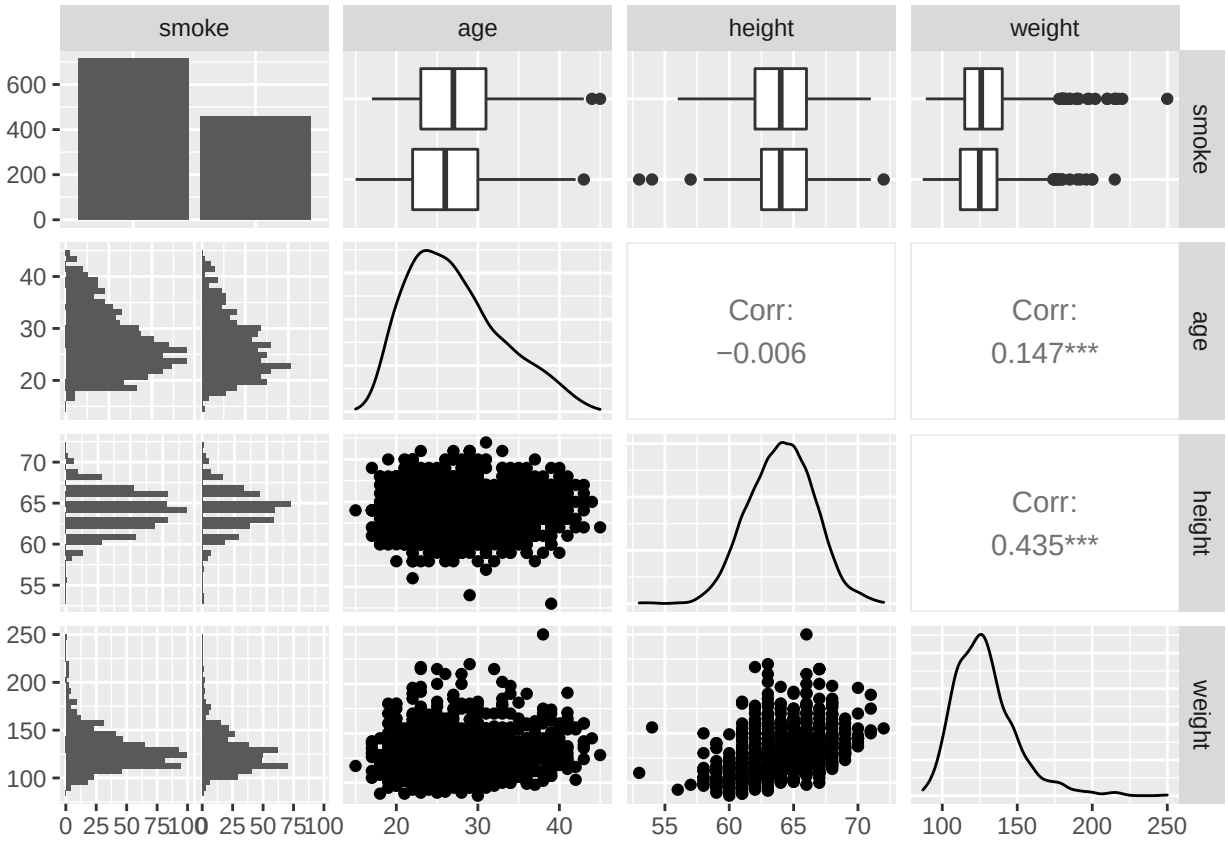

```
# Barplot for categorical variables
babies_data %>% select_if(negate(is.numeric)) %>%
  gather() %>%
  ggplot(aes(value)) +
  geom_bar() +
  facet_wrap(~ key, scales = "free")
```




```
ggpairs(babies_data[,c(8,1:4)])
```

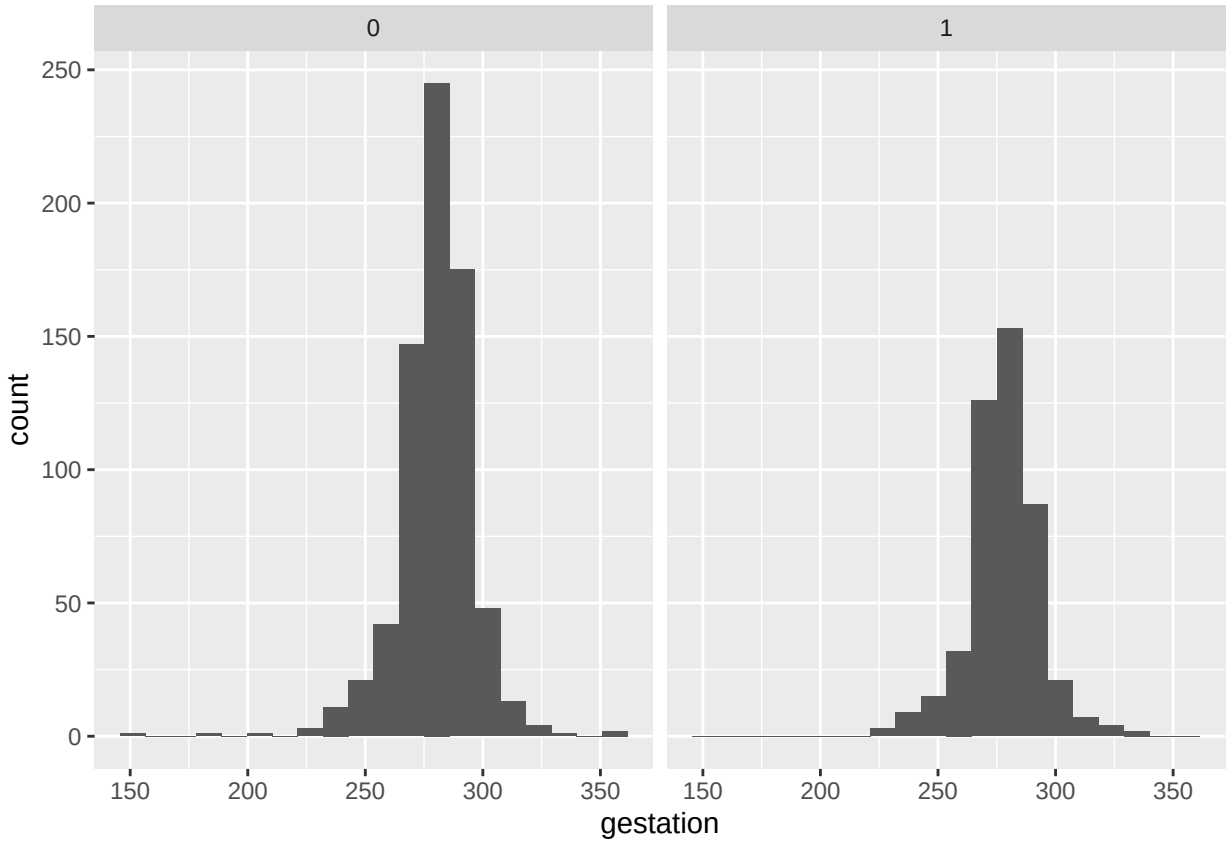



```
ggpairs(babies_data[,c(8,5:7)])
```

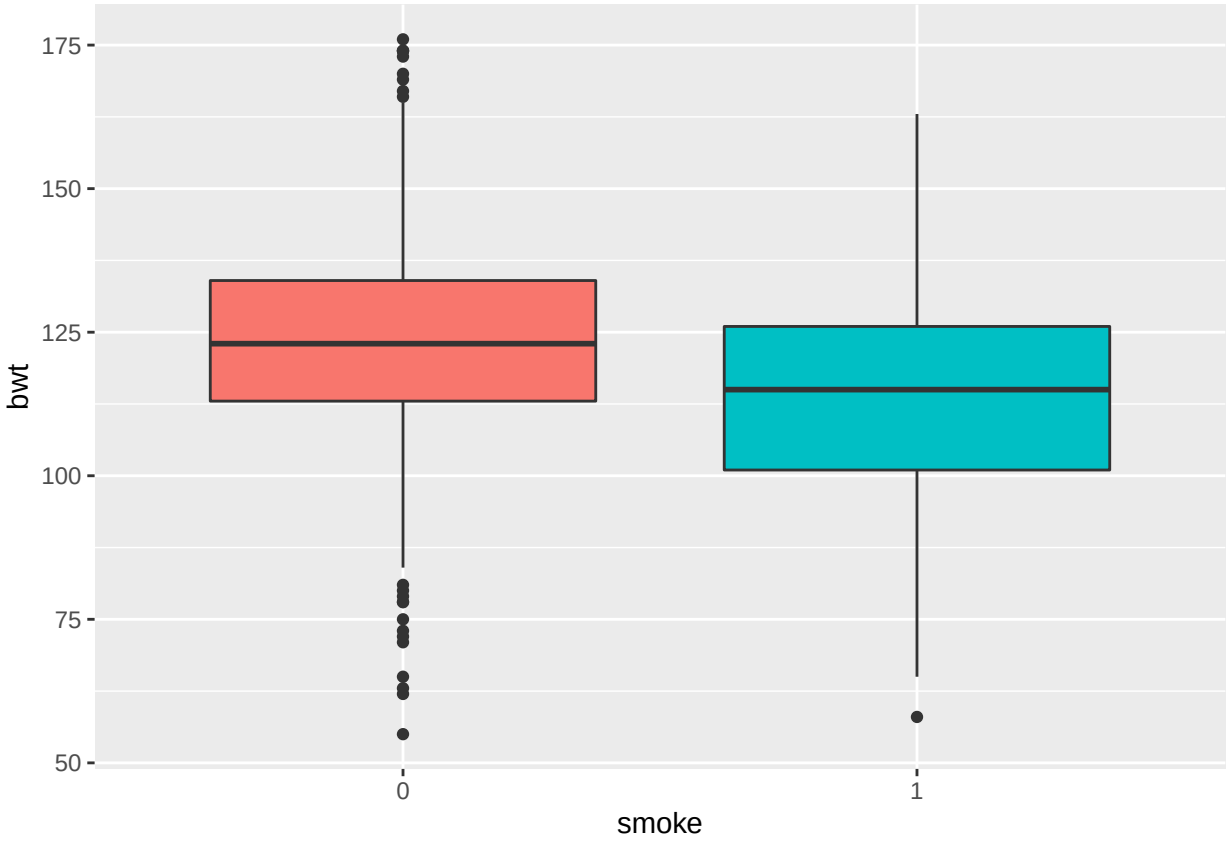


Relationship between different variables

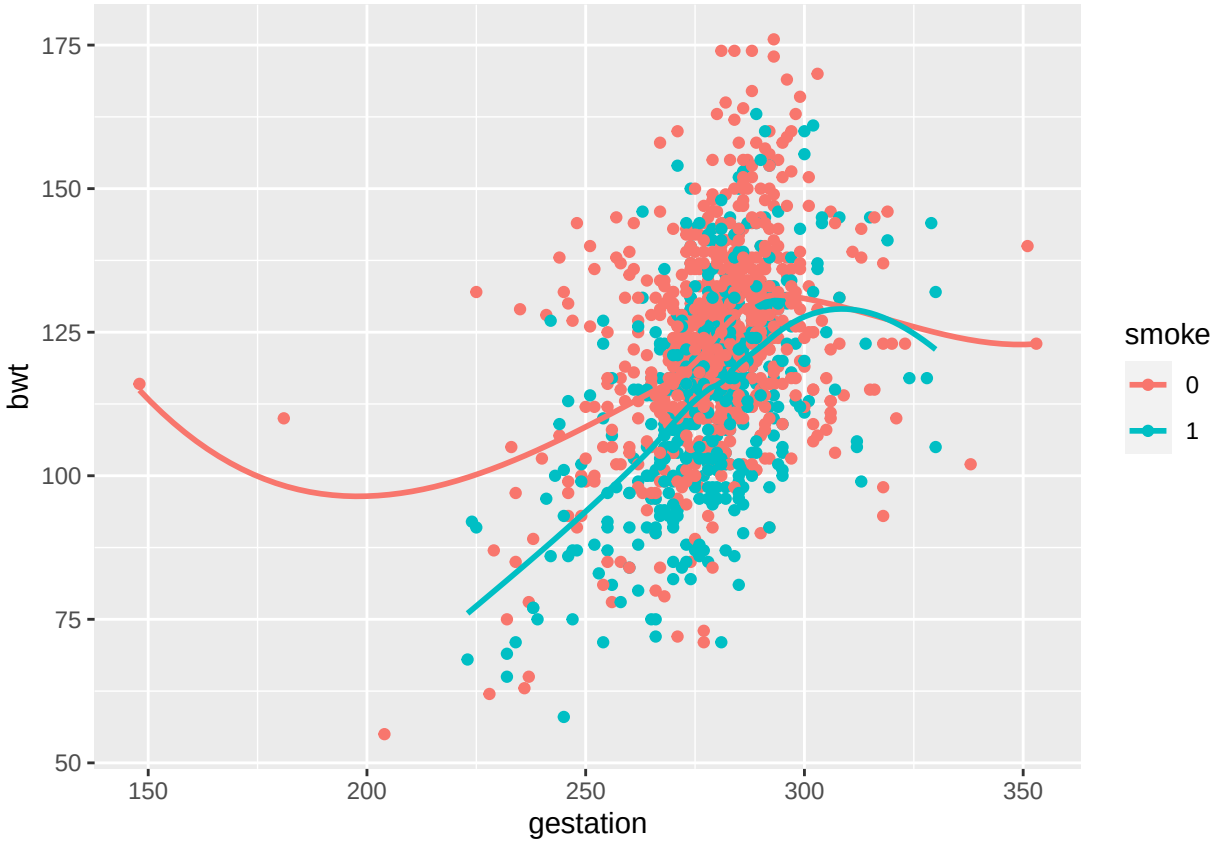
```
babies_data %>%
  ggplot() +
  geom_histogram(bins = 20, aes(x = gestation)) + facet_wrap(~smoke)
```



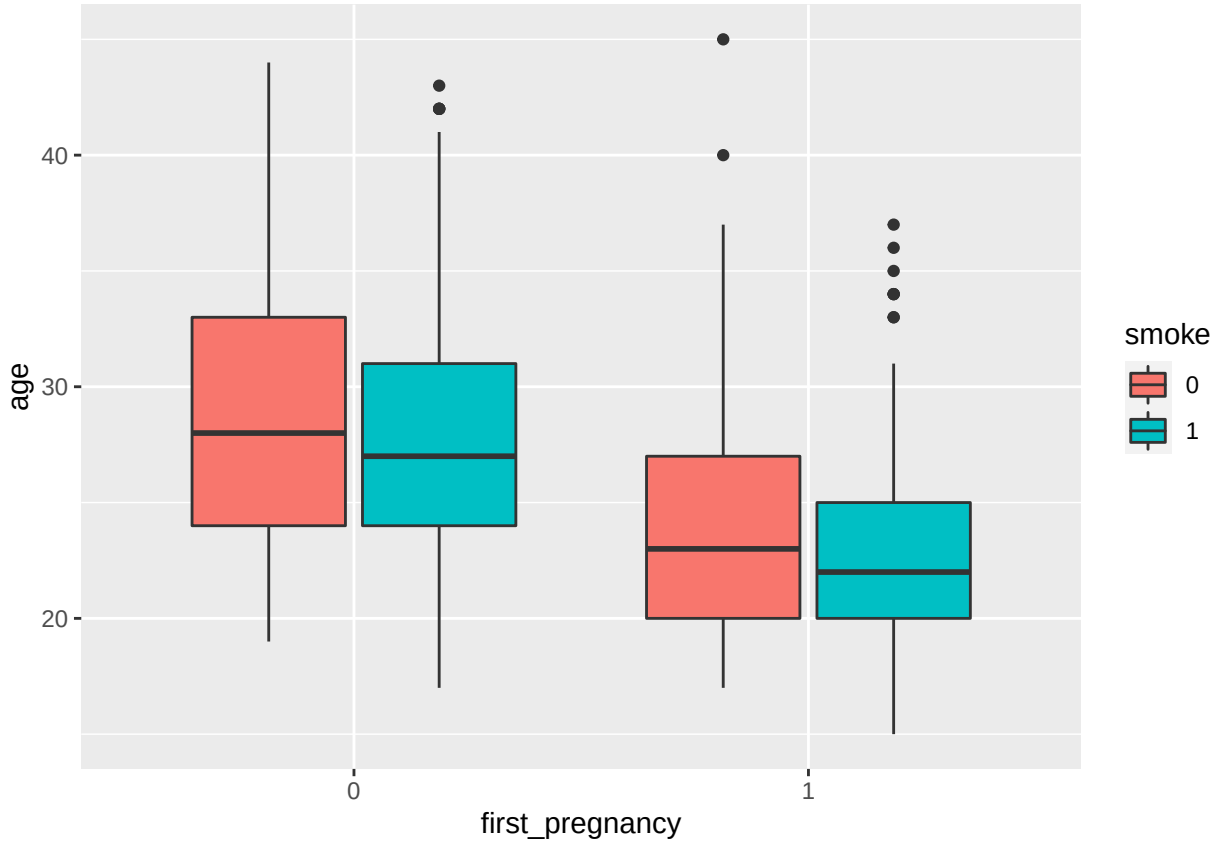

```
babies_data %>%
  ggplot(aes(x = smoke, y = bwt)) +
  geom_boxplot(aes(fill = smoke), show.legend = F)
```




```
babies_data %>%
  ggplot(aes(x = gestation, y = bwt, color = smoke)) +
  geom_point() + geom_smooth(se = F)
```




```
babies_data %>%
  ggplot(aes(x = first_pregnancy, y = age)) +
  geom_boxplot(aes(fill = smoke))
```



Research questions

1. What are the best predictors of diabetes in this dataset?
2. What is the relationship between probability of diabetes and predictor variables?

Variable selection

```
# Considering full model first
babies_full <- glm(first_pregnancy ~ ., family = "binomial", data = babies_data)
summary(babies_full)
```

```
##
## Call:
## glm(formula = first_pregnancy ~ ., family = "binomial", data = babies_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9654  -0.7422  -0.4582   0.7687   2.8914
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept) -5.0953104  2.4036240  -2.120  0.03402 *
## case         0.0014877  0.0002192   6.787 1.14e-11 ***
## bwt         -0.0118158  0.0049014  -2.411  0.01592 *
## gestation    0.0156660  0.0053444   2.931  0.00338 **
## age         -0.1853247  0.0168808 -10.978  < 2e-16 ***
## height       0.1011978  0.0348194   2.906  0.00366 **
## weight      -0.0115322  0.0044680  -2.581  0.00985 **
## smoke1      -0.3194257  0.1591969  -2.006  0.04480 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1351.3  on 1173  degrees of freedom
## Residual deviance: 1106.2  on 1166  degrees of freedom
## AIC: 1122.2
##
## Number of Fisher Scoring iterations: 5
```

```
# Final model
babies_red <- update(babies_full, ~. -case, data = babies_data)
summary(babies_red)
```

```
##
## Call:
## glm(formula = first_pregnancy ~ bwt + gestation + age + height +
##     weight + smoke, family = "binomial", data = babies_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5912  -0.8013  -0.5013   0.9340   3.0465
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.849545   2.320427  -1.659  0.09712 .
## bwt         -0.014000   0.004803  -2.915  0.00356 **
## gestation    0.015483   0.005131   3.018  0.00255 **
## age         -0.180445   0.016717 -10.794  < 2e-16 ***
## height       0.100858   0.033913   2.974  0.00294 **
## weight      -0.012044   0.004333  -2.780  0.00544 **
## smoke1      -0.311449   0.155018  -2.009  0.04452 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1351.3  on 1173  degrees of freedom
## Residual deviance: 1155.4  on 1167  degrees of freedom
## AIC: 1169.4
##
## Number of Fisher Scoring iterations: 5
```

Making data frames for doing some predictions

```
# Effect of bwt holding other variables constant
babies1 <- with(babies_data, data.frame(bwt = rep(seq(min(bwt), max(bwt),
                                                      length.out = 100),2),
                                        gestation = mean(gestation),
                                        age = mean(age),
                                        height = mean(height),
                                        weight = mean(weight),
                                        smoke = factor(rep(0:1, each = 100))))

# Effect of gestation holding other variables constant
babies2 <- with(babies_data, data.frame(bwt = mean(bwt),
                                        gestation = rep(seq(min(gestation), max(gestation),
                                                            length.out = 100), 2),
                                        age = mean(age),
                                        height = mean(height),
                                        weight = mean(weight),
                                        smoke = factor(rep(0:1, each = 100))))

# Effect of age holding other variables constant
babies3 <- with(babies_data, data.frame(bwt = mean(bwt),
                                        gestation = mean(gestation),
                                        age = rep(seq(min(age), max(age),
                                                      length.out = 100), 2),
                                        height = mean(height),
                                        weight = mean(weight),
                                        smoke = factor(rep(0:1, each = 100))))

# Effect of height holding other variables constant
babies4 <- with(babies_data, data.frame(bwt = mean(bwt),
                                        gestation = mean(gestation),
                                        age = mean(age),
                                        height = rep(seq(min(height), max(height),
                                                         length.out = 100), 2),
                                        weight = mean(weight),
                                        smoke = factor(rep(0:1, each = 100))))

# Effect of weight holding other variables constant
babies5 <- with(babies_data, data.frame(bwt = mean(bwt),
                                        gestation = mean(gestation),
                                        age = mean(age),
                                        height = mean(height),
                                        weight = rep(seq(min(weight), max(weight),
                                                         length.out = 100), 2),
                                        smoke = factor(rep(0:1, each = 100))))
```

Making predictions

```
babies_1 <- cbind(babies1, predict(babies_red, newdata = babies1, type = "link", se = TRUE))
babies_1 <- within(babies_1, {
  pred_prob <- plogis(fit)
  lower <- plogis(fit - (1.96 * se.fit))
  upper <- plogis(fit + (1.96 * se.fit))})

babies_2 <- cbind(babies2, predict(babies_red, newdata = babies2, type = "link", se = TRUE))
babies_2 <- within(babies_2, {
  pred_prob <- plogis(fit)
  lower <- plogis(fit - (1.96 * se.fit))
  upper <- plogis(fit + (1.96 * se.fit))})

babies_3 <- cbind(babies3, predict(babies_red, newdata = babies3, type = "link", se = TRUE))
babies_3 <- within(babies_3, {
  pred_prob <- plogis(fit)
  lower <- plogis(fit - (1.96 * se.fit))
  upper <- plogis(fit + (1.96 * se.fit))})

babies_4 <- cbind(babies4, predict(babies_red, newdata = babies4, type = "link", se = TRUE))
babies_4 <- within(babies_4, {
  pred_prob <- plogis(fit)
  lower <- plogis(fit - (1.96 * se.fit))
  upper <- plogis(fit + (1.96 * se.fit))})

babies_5 <- cbind(babies5, predict(babies_red, newdata = babies5, type = "link", se = TRUE))
babies_5 <- within(babies_5, {
  pred_prob <- plogis(fit)
  lower <- plogis(fit - (1.96 * se.fit))
  upper <- plogis(fit + (1.96 * se.fit))})
```

Plotting the predictions

```
a <- ggplot(babies_1, aes(x = bwt, y = pred_prob)) +
  geom_ribbon(aes(ymin = lower, ymax = upper, fill = smoke), alpha = 0.10) +
  geom_line(aes(colour = smoke), size = 1) + theme_light() +
  xlab("birth_wt") + ylab("Pregnancy_pp") #pp means predicted probabilities

b <- ggplot(babies_2, aes(x = gestation, y = pred_prob)) +
  geom_ribbon(aes(ymin = lower, ymax = upper, fill = smoke), alpha = 0.10) +
  geom_line(aes(colour = smoke), size = 1) +  theme_light() +
  ylab("Pregnancy_pp")

c <- ggplot(babies_3, aes(x = age, y = pred_prob)) +
  geom_ribbon(aes(ymin = lower, ymax = upper, fill = smoke), alpha = 0.10) +
  geom_line(aes(colour = smoke), size = 1) +  theme_light() +
  ylab("Pregnancy_pp")

d <- ggplot(babies_4, aes(x = height, y = pred_prob)) +
  geom_ribbon(aes(ymin = lower, ymax = upper, fill = smoke), alpha = 0.10) +
  geom_line(aes(colour = smoke), size = 1) +  theme_light() +
```

```
  ylab("Pregnancy_pp")

e <- ggplot(babies_5, aes(x = weight, y = pred_prob)) +
  geom_ribbon(aes(ymin = lower, ymax = upper, fill = smoke), alpha = 0.10) +
  geom_line(aes(colour = smoke), size = 1) +  theme_light() +
  ylab("Pregnancy_pp")

ggarrange(b, c, d, e, a, 0, ncol = 2, nrow = 3)
```