```r
library(ggplot2)
library(car)
```

```
## Loading required package: carData
```

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:car':
##
##     recode
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

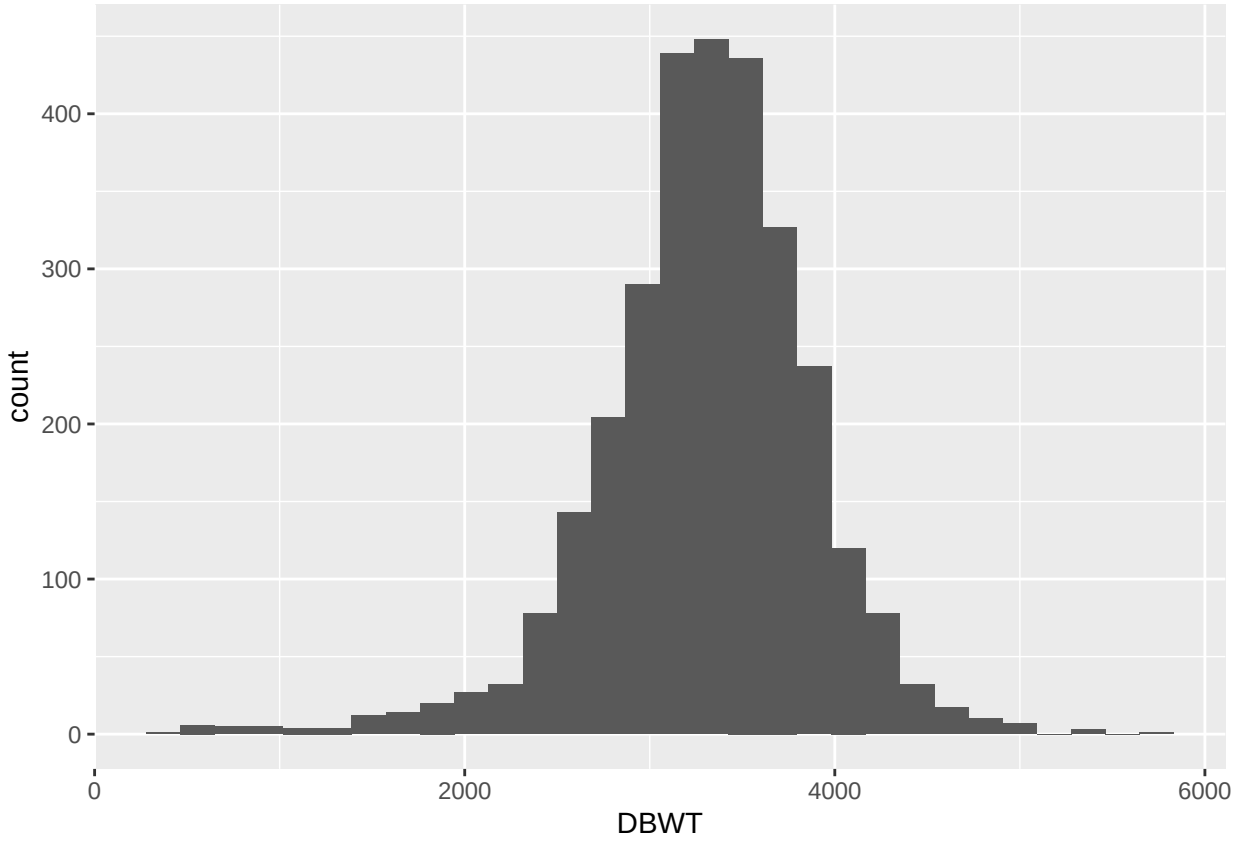
```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
EDA_df <- read.csv("data/EDA.csv")
EDA_df$CIG_0_BIN <- factor(EDA_df$CIG_0_BIN)
EDA_df$PRECARE <- factor(EDA_df$PRECARE)
EDA_df$SEX <- factor(EDA_df$SEX)
EDA_df$RESTATUS <- factor(EDA_df$RESTATUS)
EDA_df$PAY <- factor(EDA_df$PAY)
EDA_df$NO_RISKS <- factor(EDA_df$NO_RISKS)
EDA_df$MRAVE6 <- factor(EDA_df$MRAVE6)
EDA_df$FRACE6  <- factor(EDA_df$FRACE6)
EDA_df$MEDUC <- factor(EDA_df$MEDUC)
EDA_df$FEDUC <- factor(EDA_df$FEDUC)
```

Response variable:

```r
# response variable
ggplot(EDA_df, aes(x = DBWT)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

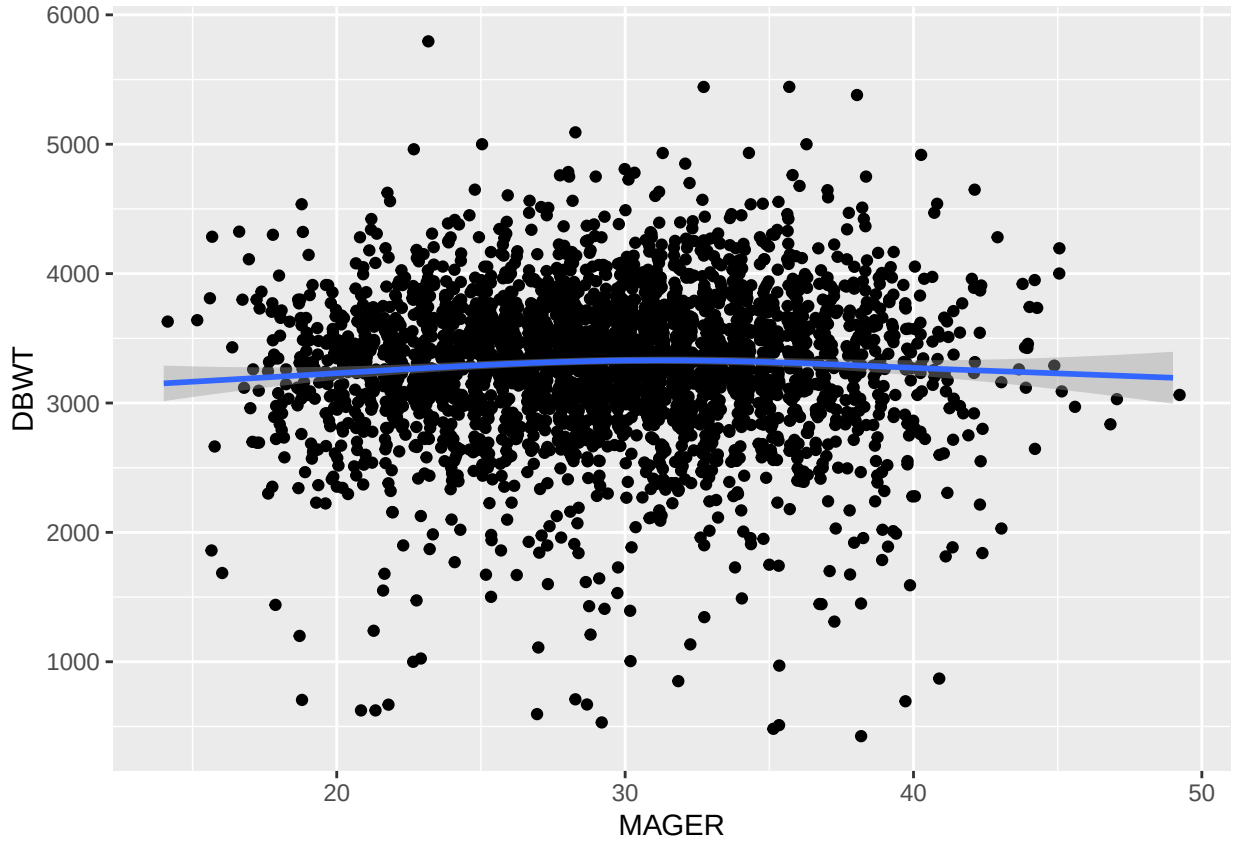



```
# Measure of symmetry
DBWT_sym = (quantile(EDA_df$DBWT, 0.75) - median(EDA_df$DBWT)) /
  (median(EDA_df$DBWT) - quantile(EDA_df$DBWT, 0.25))
```

MAGER

```
ggplot(EDA_df, aes(x = MAGER, y = DBWT)) +
  geom_point(position = "jitter") +
  geom_smooth()
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



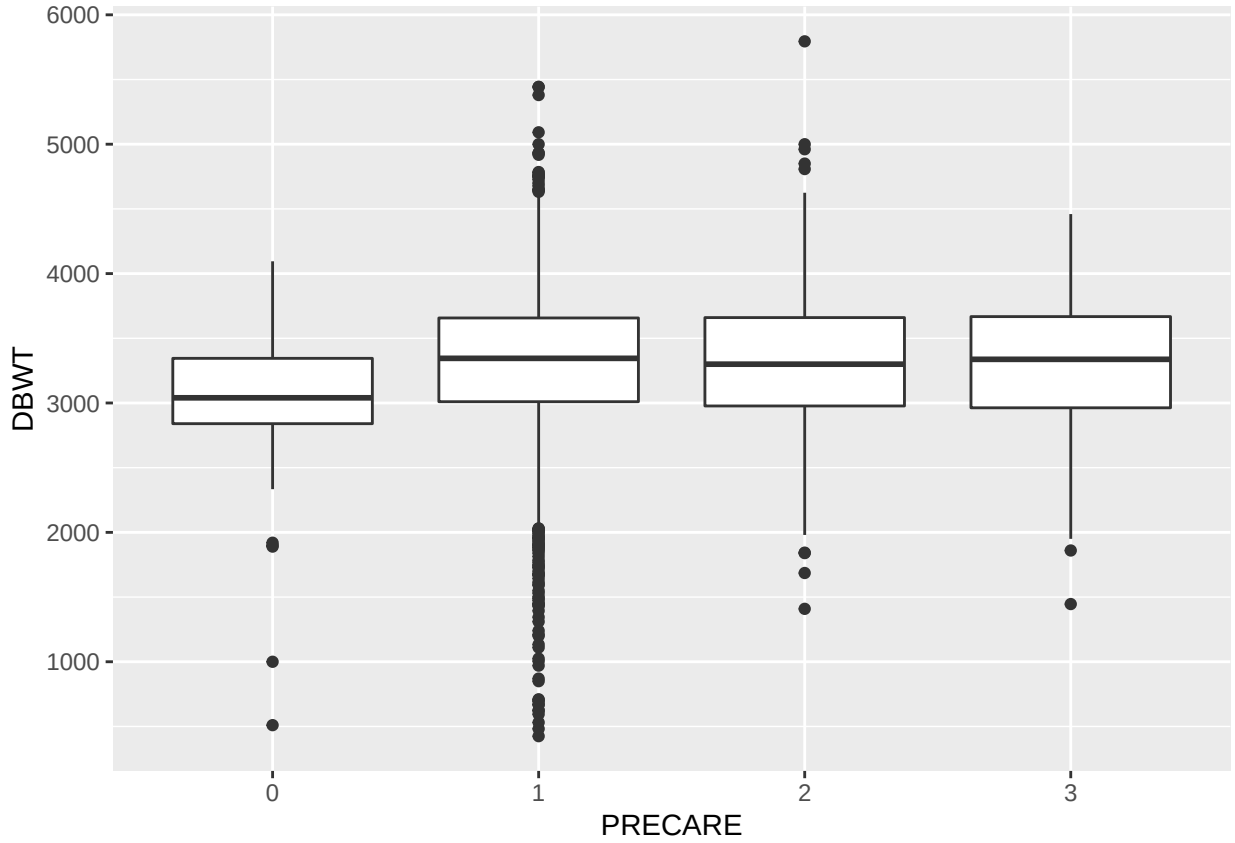

```
ggplot(EDA_df, aes(x = MAGER, y = DBWT)) +
  geom_point(position = "jitter", aes(colour = CIG_0_BIN), alpha = 0.5) +
  geom_smooth(method = 'lm', aes(colour = CIG_0_BIN))
```

```
## `geom_smooth()` using formula 'y ~ x'
```

Slightly negative slope for smoking mothers between `MAGER` and `DBWT`.

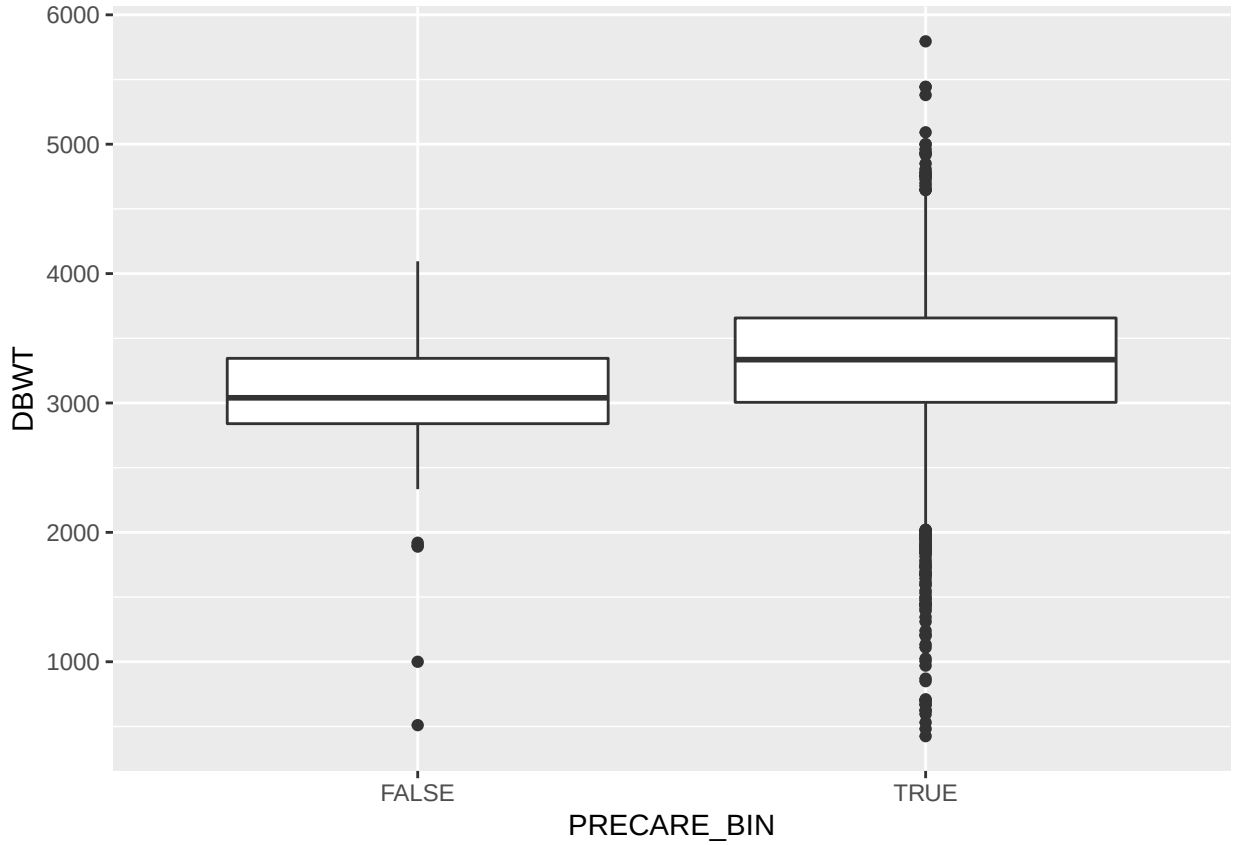
PRECARE

```
ggplot(EDA_df, aes(x = PRECARE, y = DBWT)) +
  geom_boxplot()
```



Observing that the most important difference lies between `PRECARE` of 0 and the other status of `PRECARE`, we binarize `PRECARE` for downstream analysis.

```
# Binarizing PRECARE
EDA_df$PRECARE_BIN <- ifelse(EDA_df$PRECARE == 0, FALSE, TRUE)
ggplot(EDA_df, aes(x = PRECARE_BIN, y = DBWT)) +
  geom_boxplot()
```

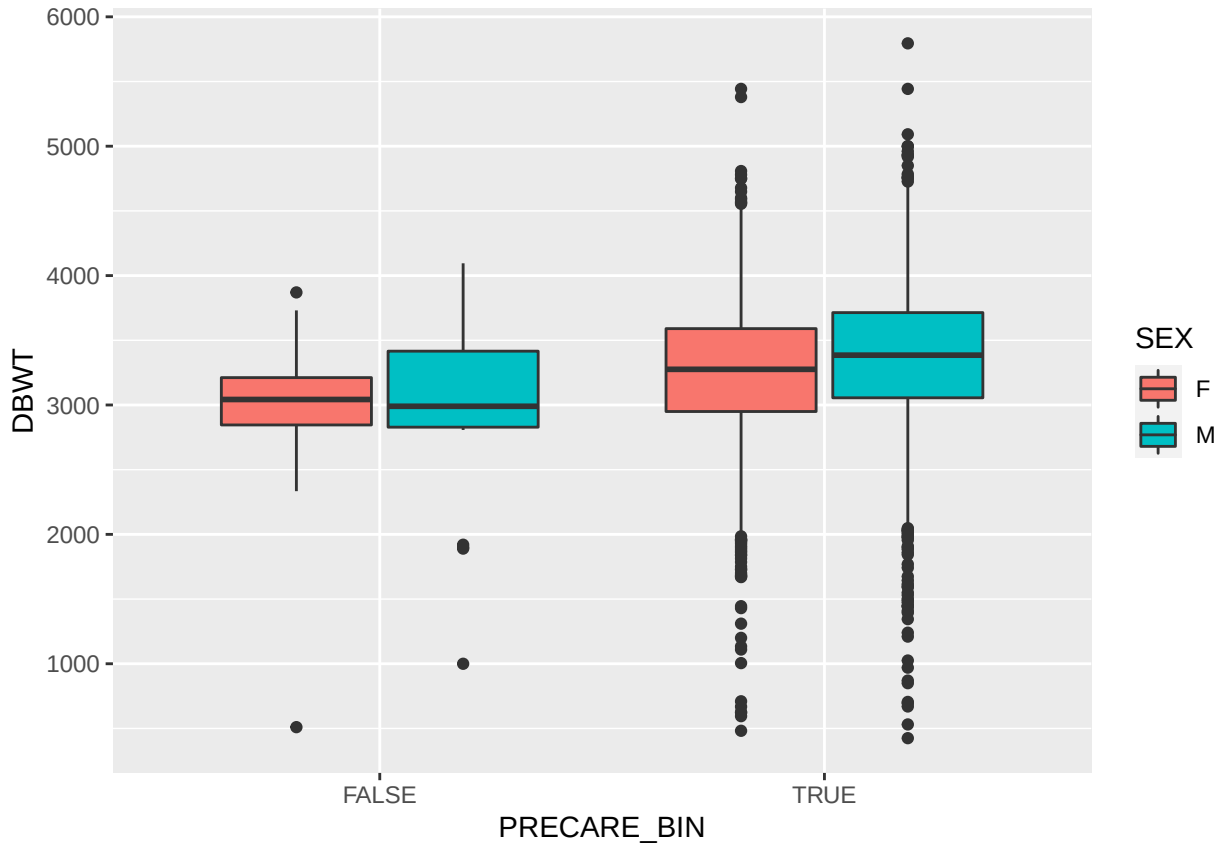



```
EDA_df %>% count(PRECARE)
```

```
##   PRECARE    n
## 1       0   41
## 2       1 2435
## 3       2  414
## 4       3  110
```

No PRECARE has significant DBWT.

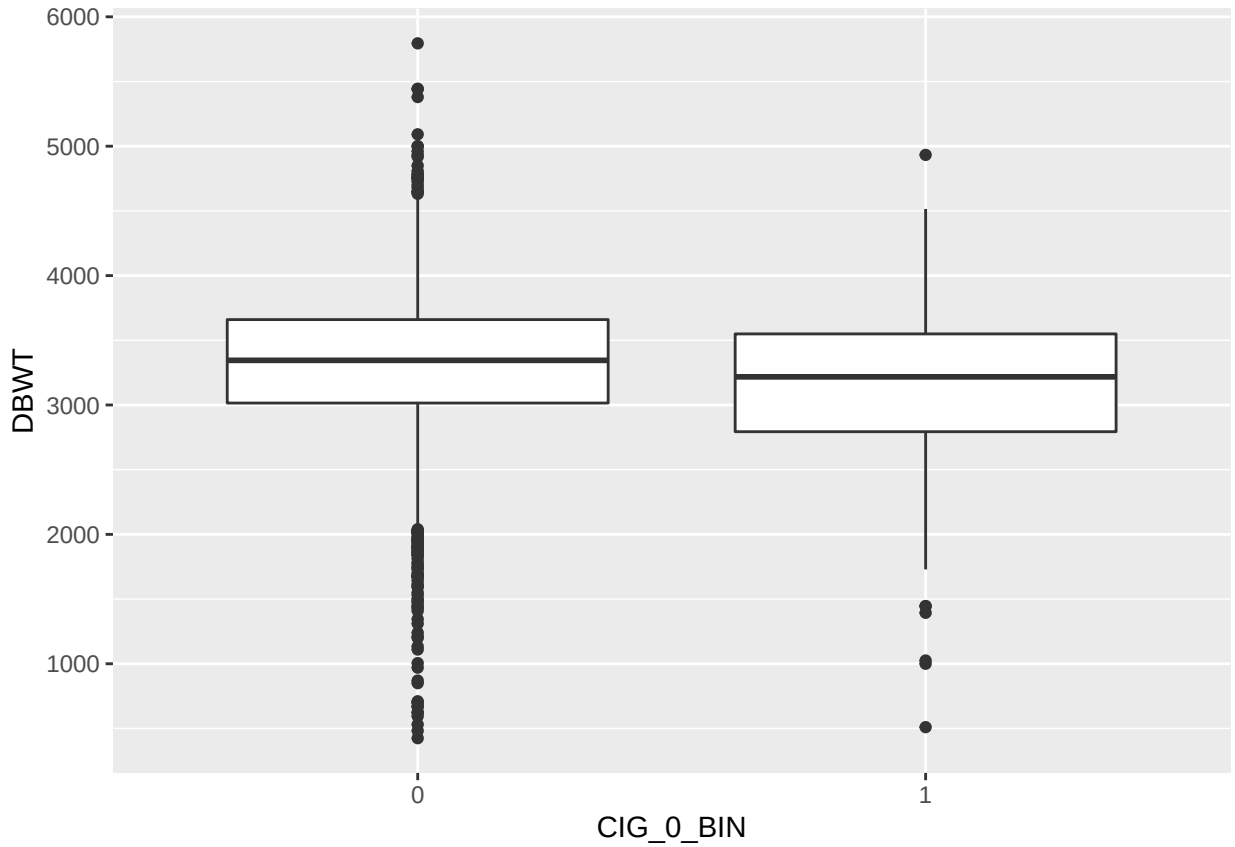
```
ggplot(EDA_df, aes(x = PRECARE_BIN, y = DBWT)) +
  geom_boxplot(aes(fill = SEX))
```



`SEX` matters more for higher `PRECARE` (start prenatal care later).

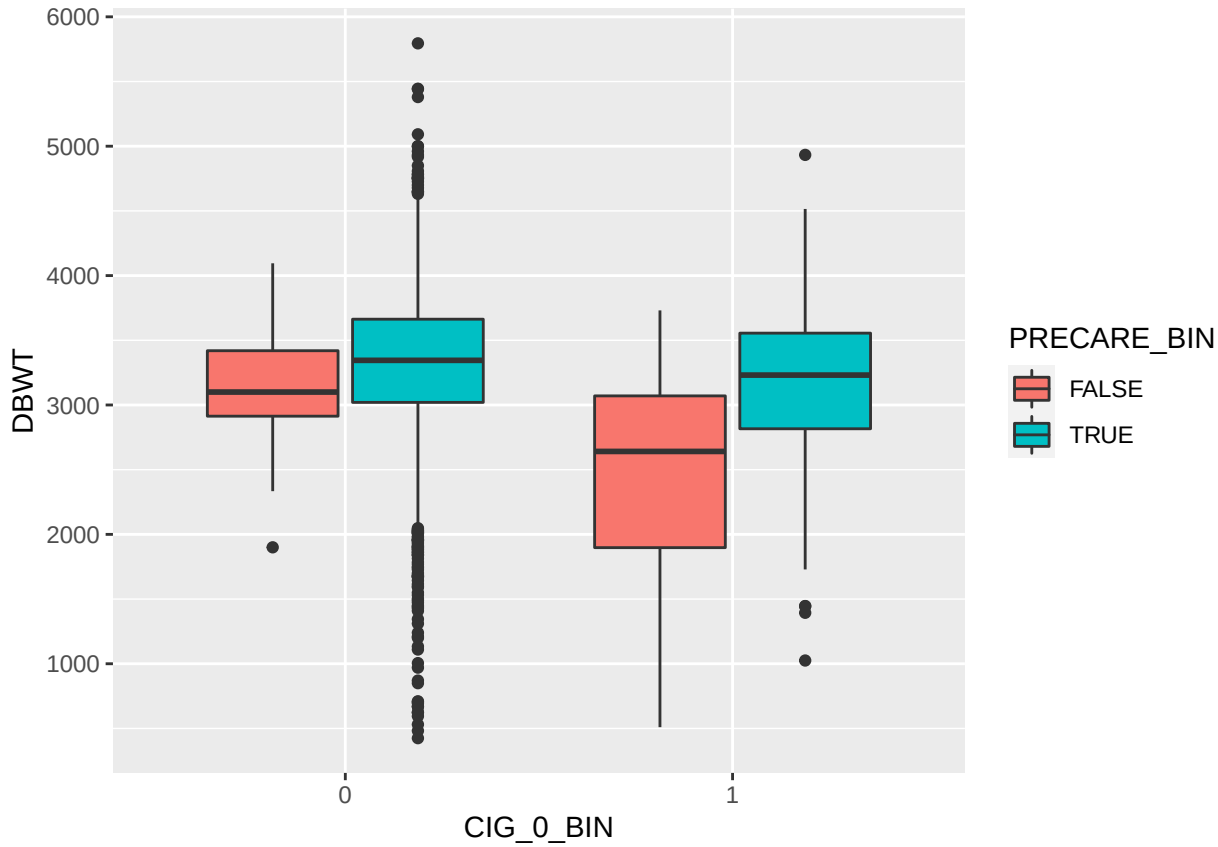
CIG_0_BIN

```
ggplot(EDA_df, aes(x = CIG_0_BIN, y = DBWT)) +
  geom_boxplot()
```



No smoking leads to higher DBWT.

```
ggplot(EDA_df, aes(x = CIG_0_BIN, y = DBWT)) +
  geom_boxplot(aes(fill = PRECARE_BIN))
```



PRECARE difference is more obvious in smoking mothers. But it might due to the relative smaller number of smoking mothers.

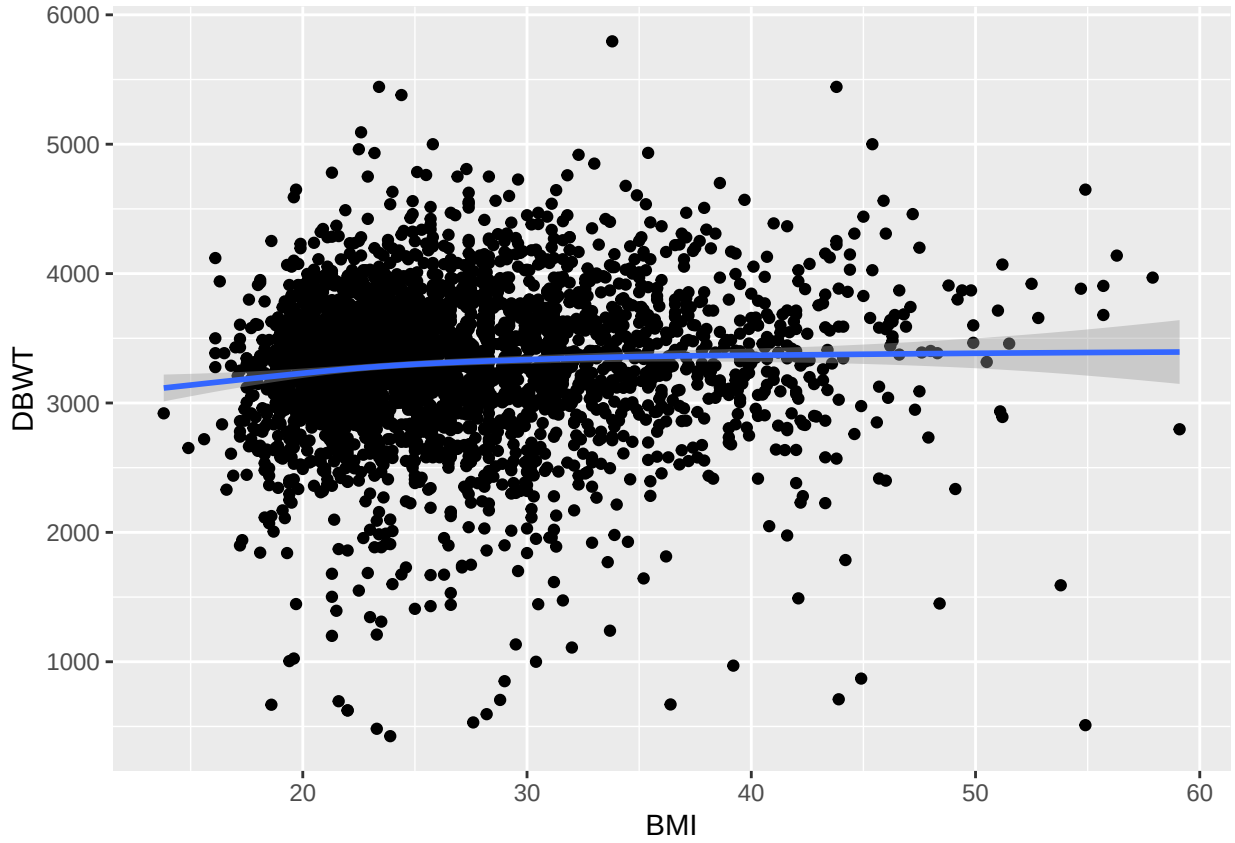
```
EDA_df %>% count(CIG_0_BIN)
```

```
##   CIG_0_BIN    n
## 1         0 2768
## 2         1  232
```

BMI:

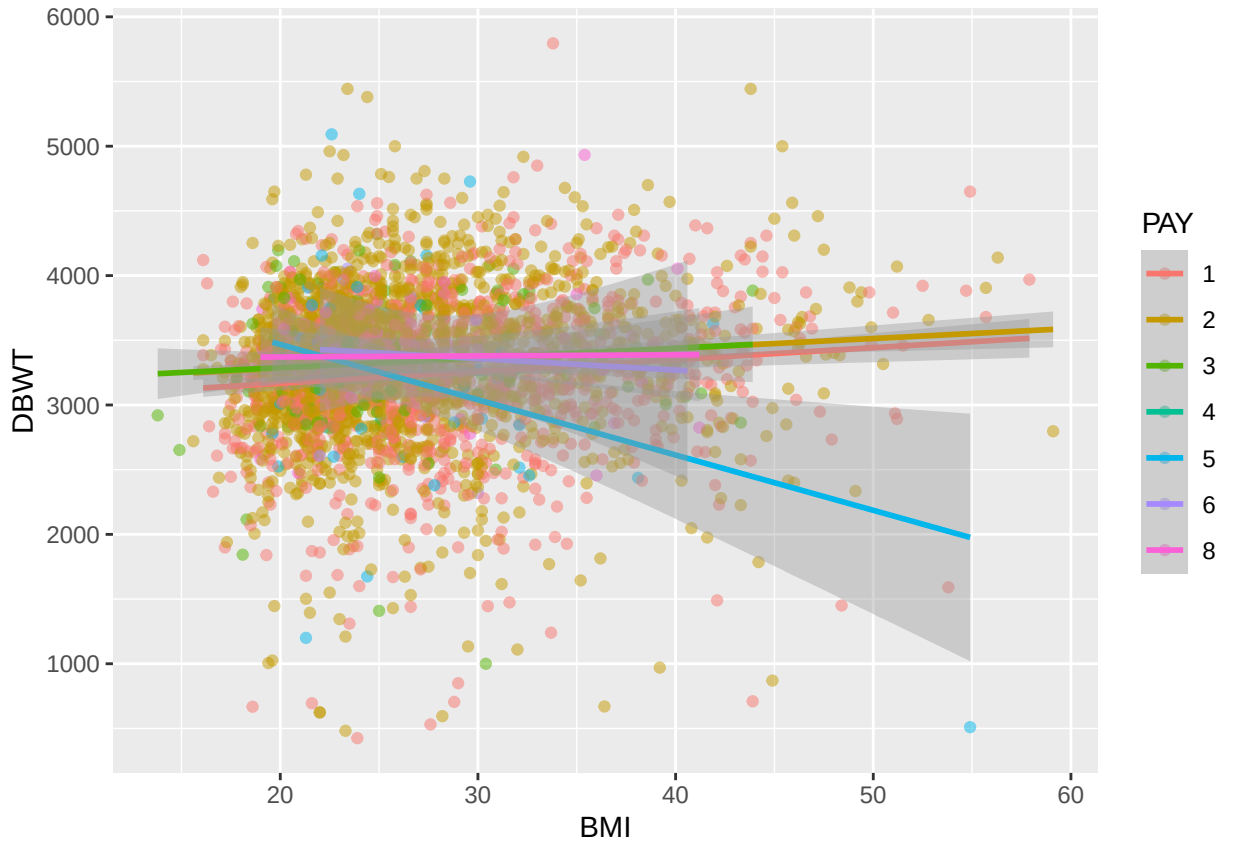
```
ggplot(EDA_df, aes(x = BMI, y = DBWT)) +
  geom_point() +
  geom_smooth()
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



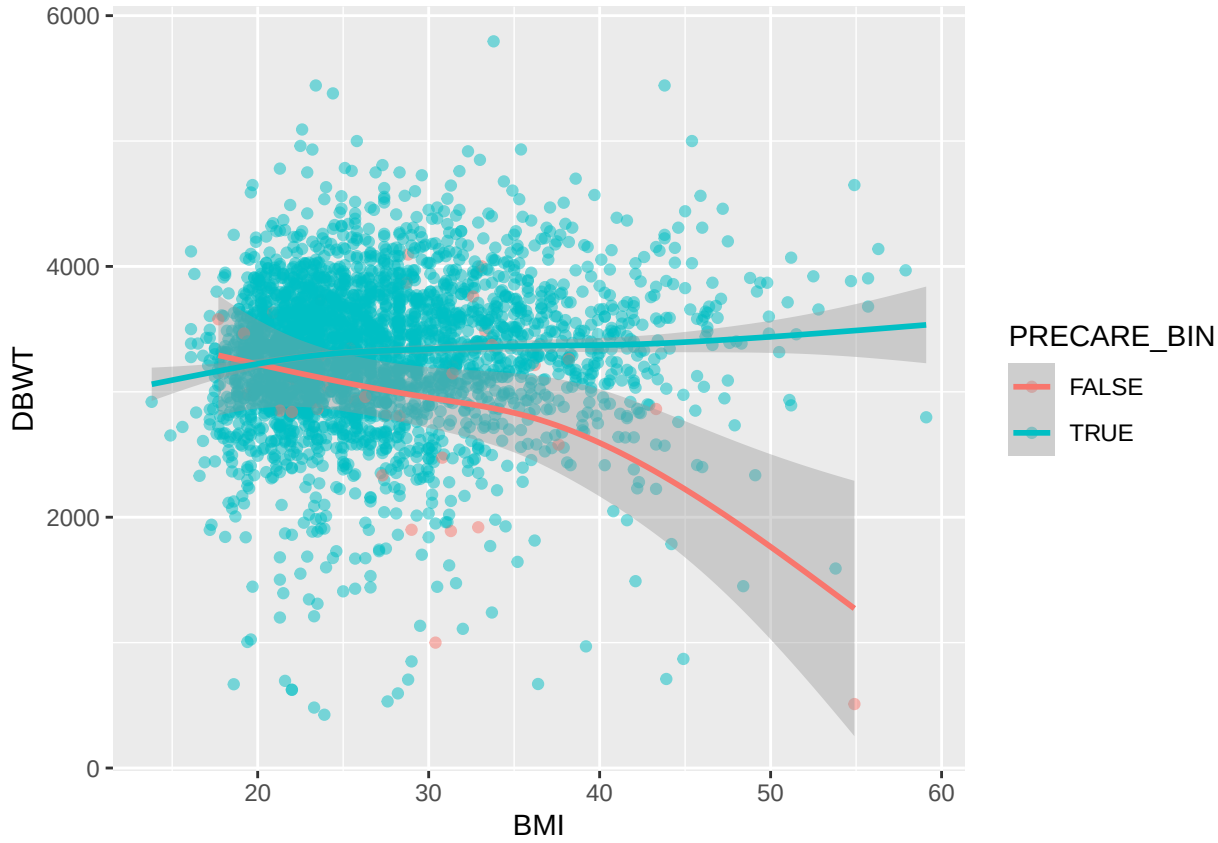

```
ggplot(EDA_df, aes(x = BMI, y = DBWT)) +
  geom_point(aes(colour = PAY), alpha = 0.5) +
  geom_smooth(method = 'lm', aes(colour = PAY))
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
ggplot(EDA_df, aes(x = BMI, y = DBWT)) +
  geom_point(aes(colour = PRECARE_BIN), alpha = 0.5) +
  geom_smooth(aes(colour = PRECARE_BIN))
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



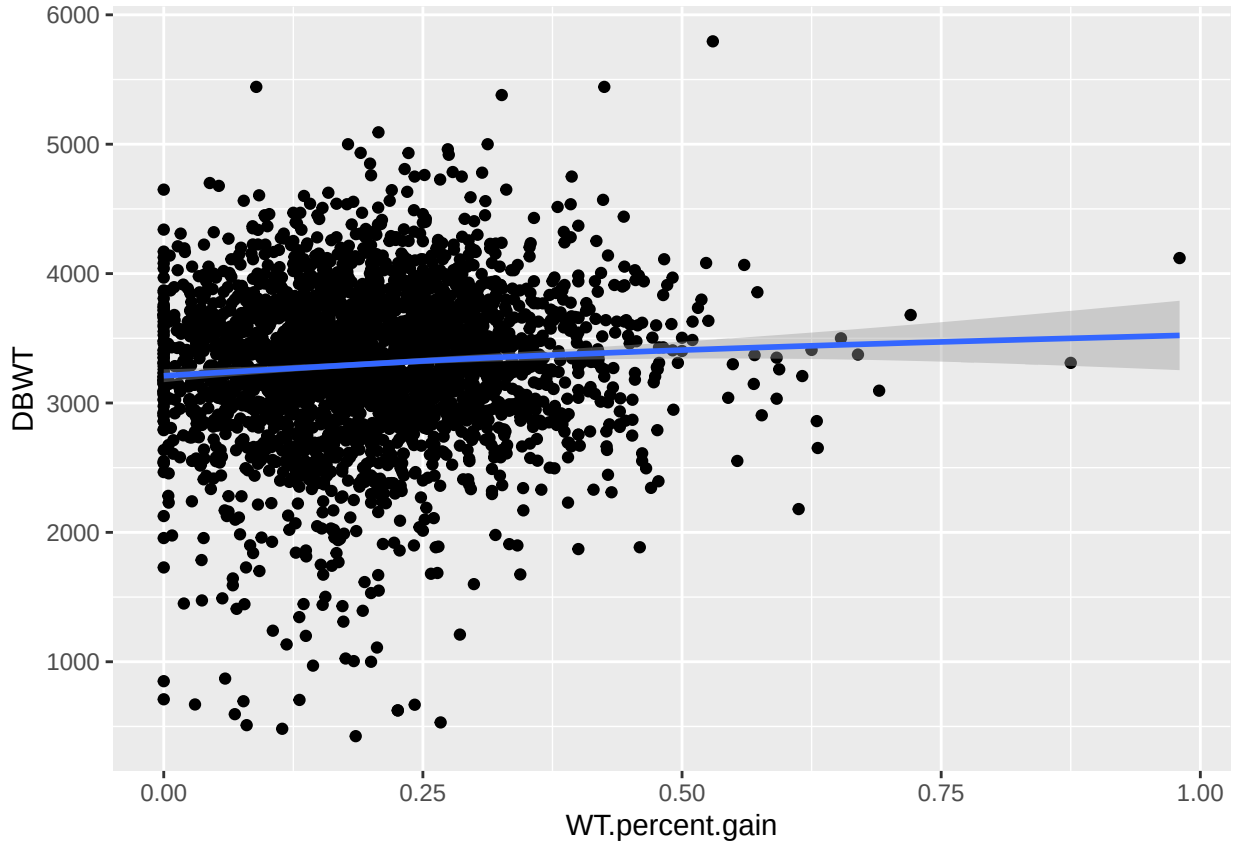

```
EDA_df %>% count(PAY)
```

```
##   PAY    n
## 1   1 1038
## 2   2 1712
## 3   3  134
## 4   4    1
## 5   5   53
## 6   6   13
## 7   8   49
```

WTGAIN.percentage:

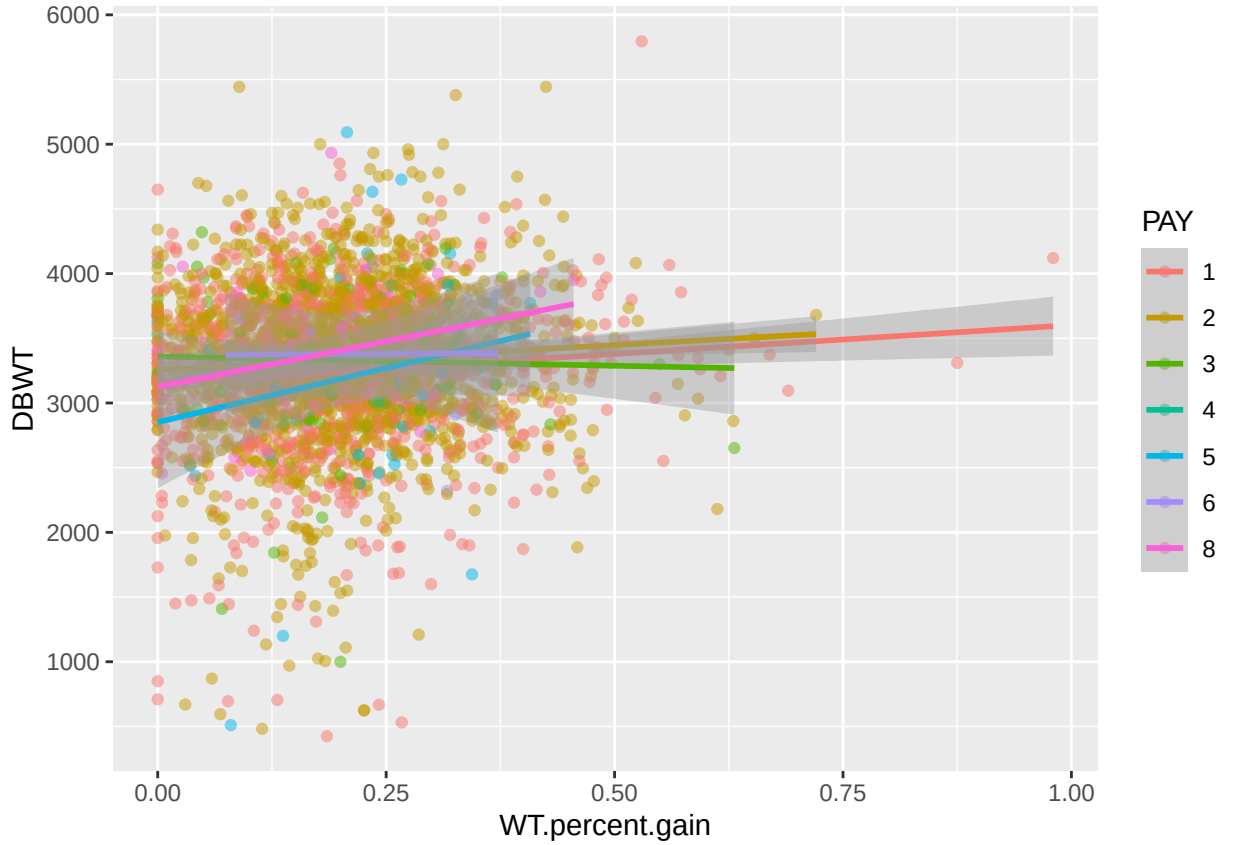
```
ggplot(EDA_df, aes(x = WT.percent.gain, y = DBWT)) +
  geom_point() +
  geom_smooth()
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



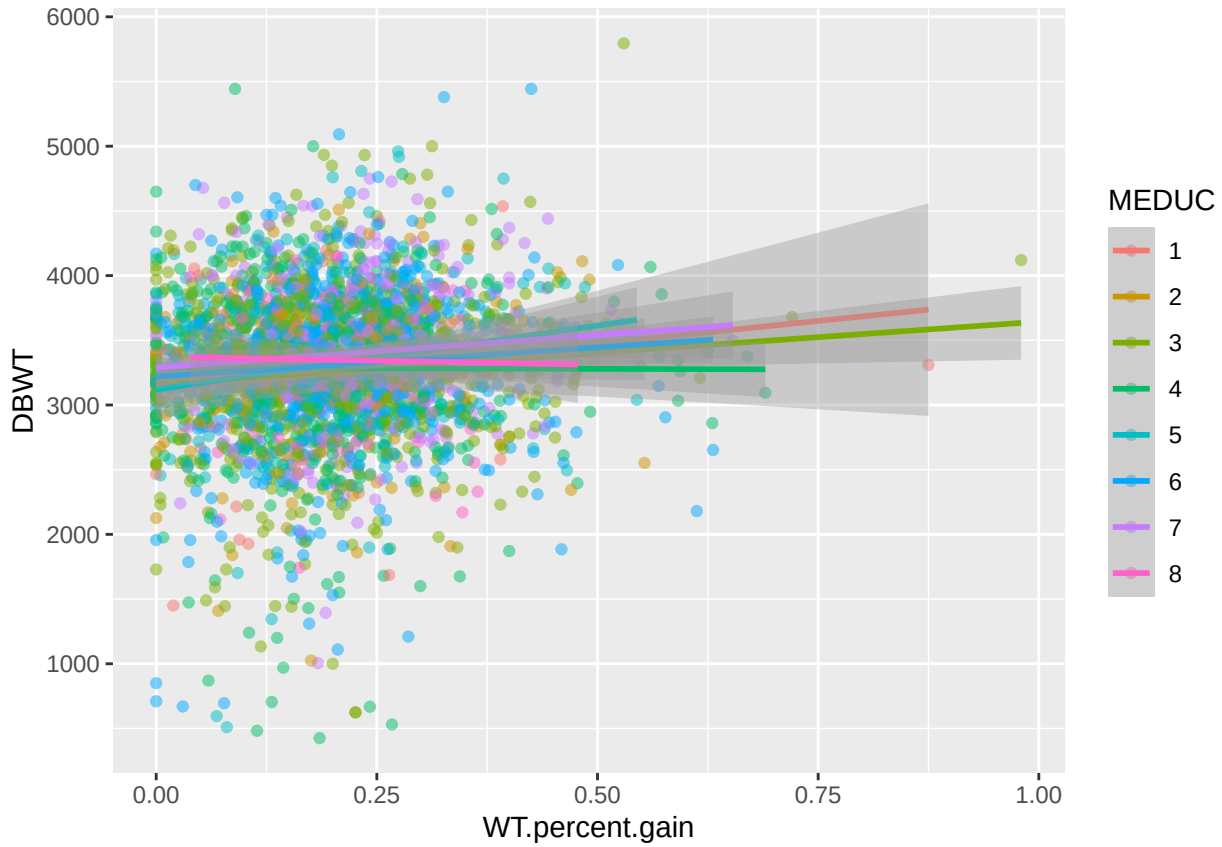

```
ggplot(EDA_df, aes(x = WT.percent.gain, y = DBWT)) +
  geom_point(aes(colour = PAY), alpha = 0.5) +
  geom_smooth(method = 'lm', aes(colour = PAY))
```

```
## `geom_smooth()` using formula 'y ~ x'
```



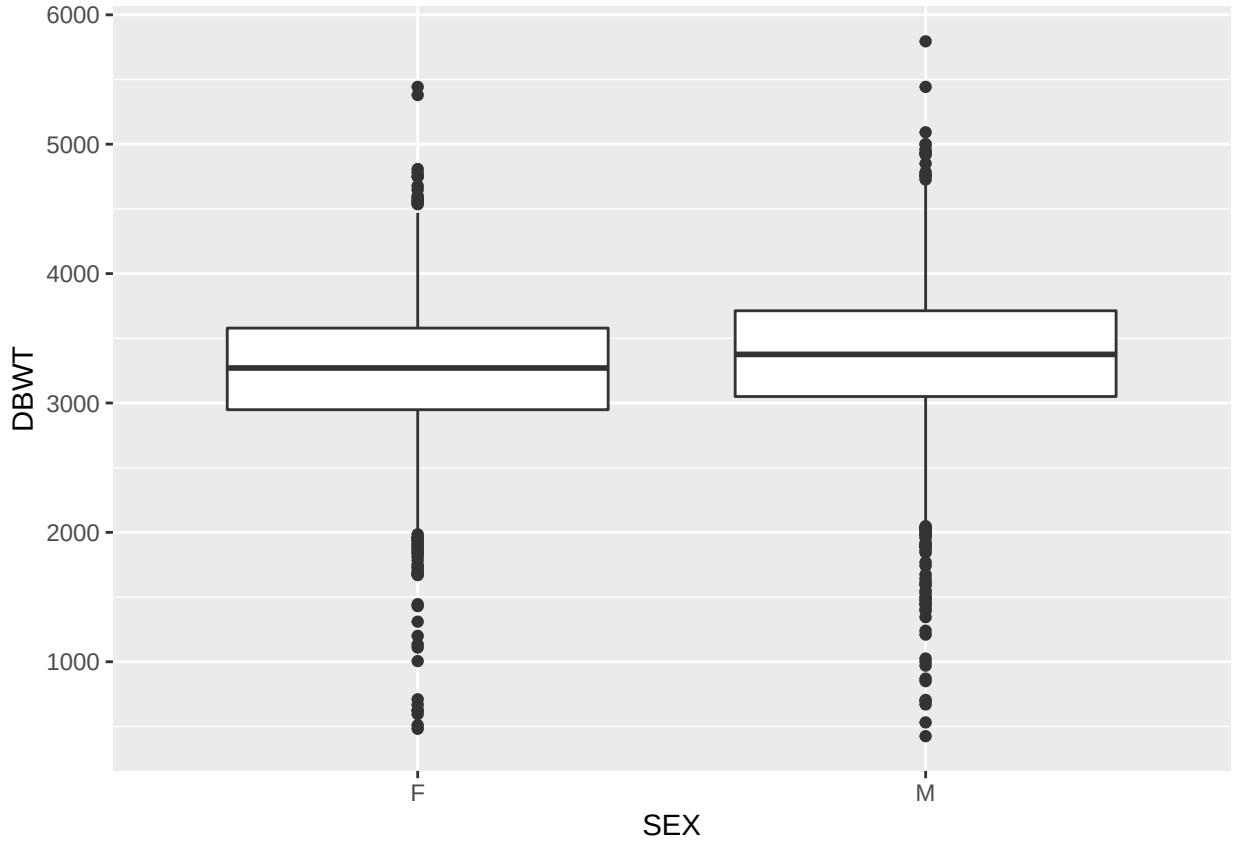

```
ggplot(EDA_df, aes(x = WT.percent.gain, y = DBWT)) +
  geom_point(aes(colour = MEDUC), alpha = 0.5) +
  geom_smooth(method = "lm",aes(colour = MEDUC))
```

```
## `geom_smooth()` using formula 'y ~ x'
```

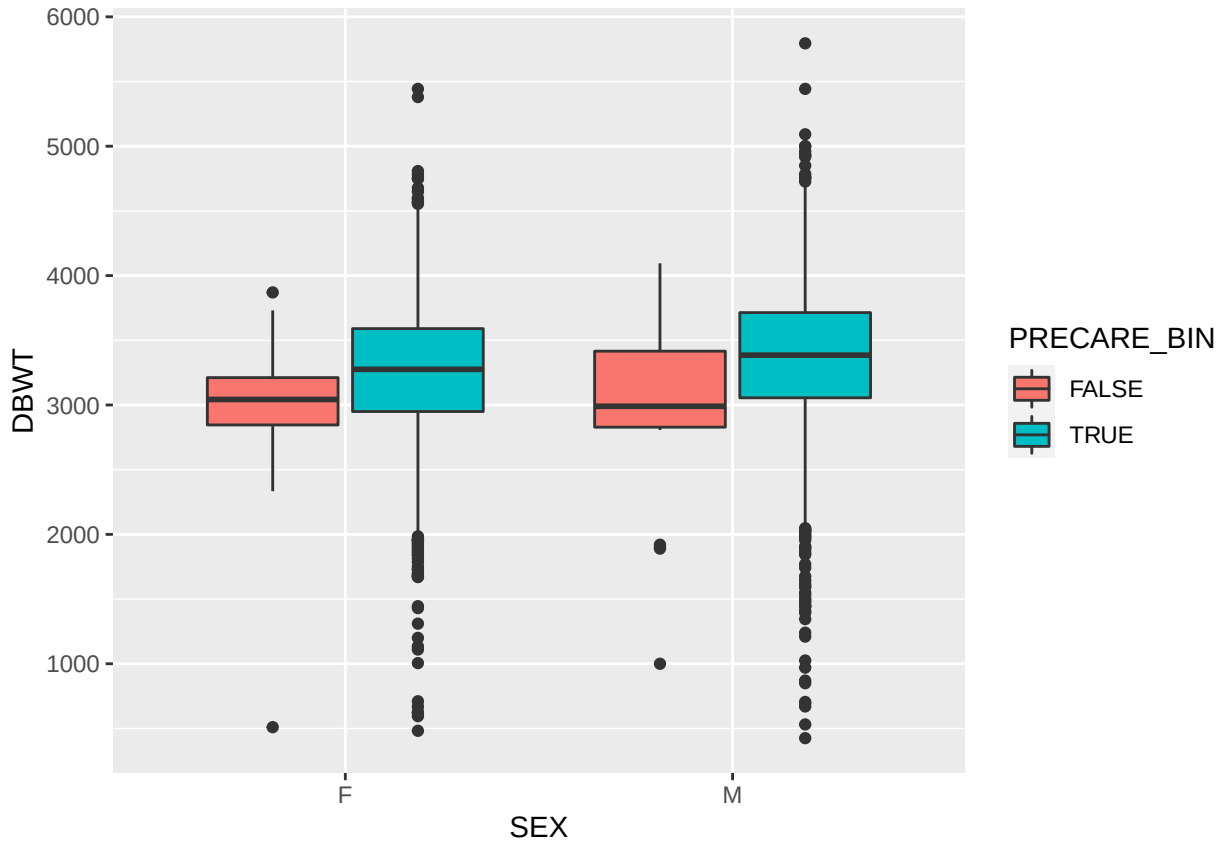


SEX

```
ggplot(EDA_df, aes(x = SEX, y = DBWT)) +
  geom_boxplot()
```




```
ggplot(EDA_df, aes(x = SEX, y = DBWT)) +
  geom_boxplot(aes(fill = PRECARE_BIN))
```

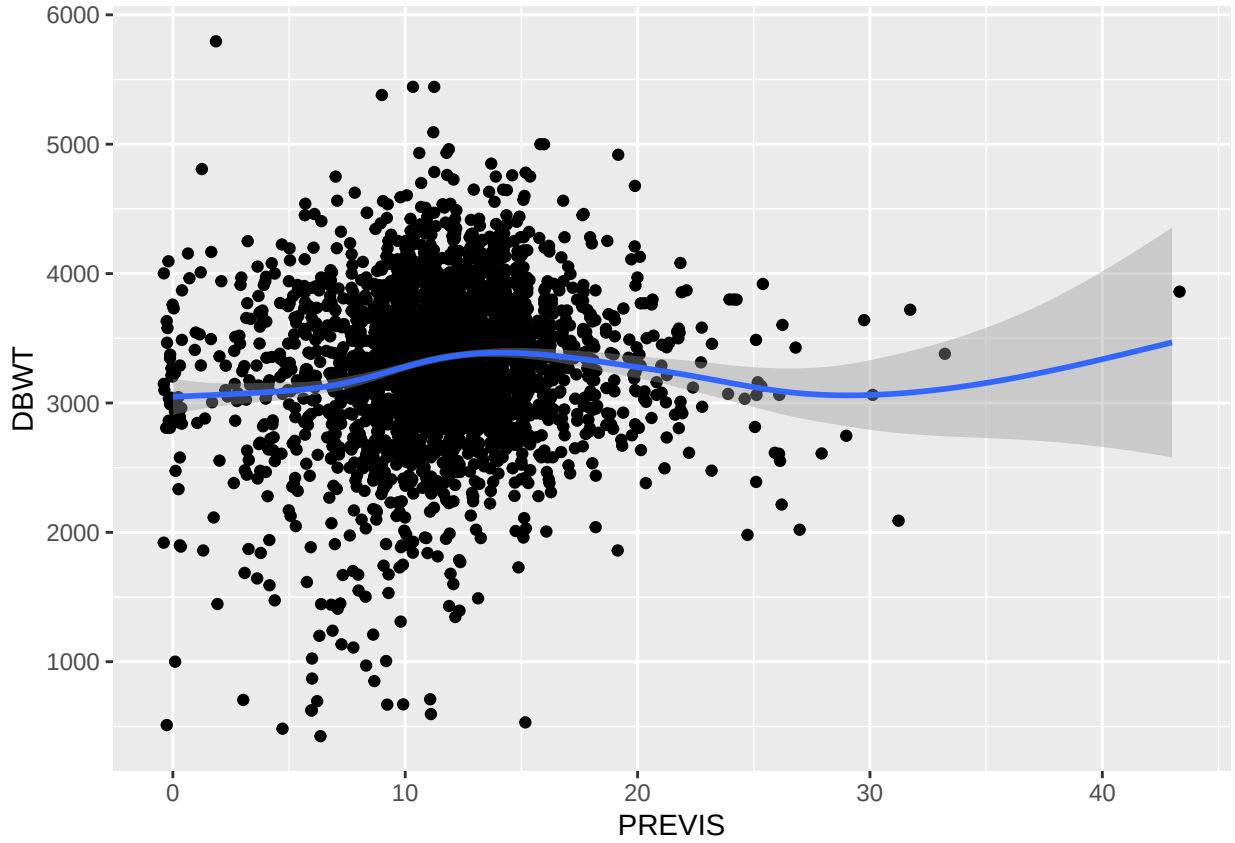


PRECARE matters more in male babies.

PREVIS

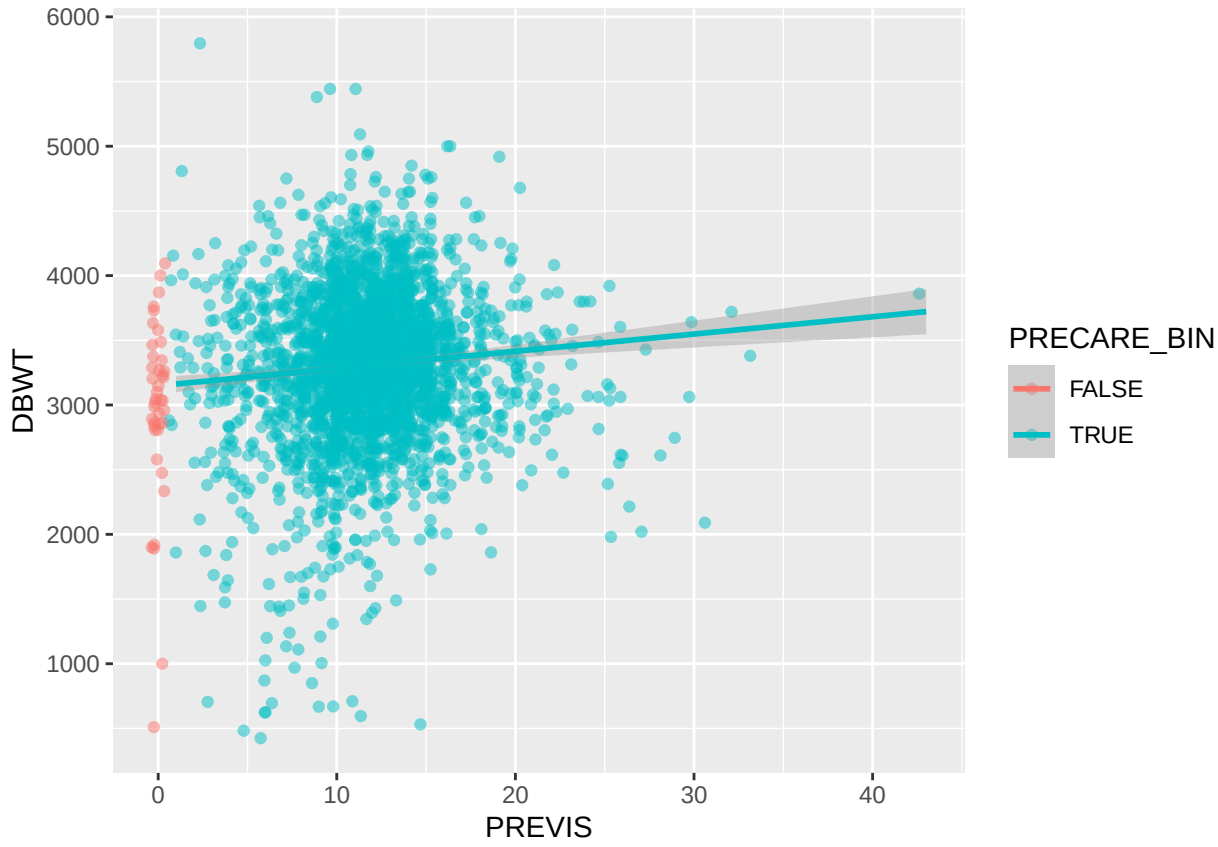
```
ggplot(EDA_df, aes(x = PREVIS, y = DBWT)) +
  geom_point(position = "jitter") +
  geom_smooth()
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




```
ggplot(EDA_df, aes(x = PREVIS, y = DBWT)) +
  geom_point(position = "jitter", aes(colour = PRECARE_BIN), alpha = 0.5) +
  geom_smooth(method = "lm", aes(colour = PRECARE_BIN))
```

```
## `geom_smooth()` using formula 'y ~ x'
```

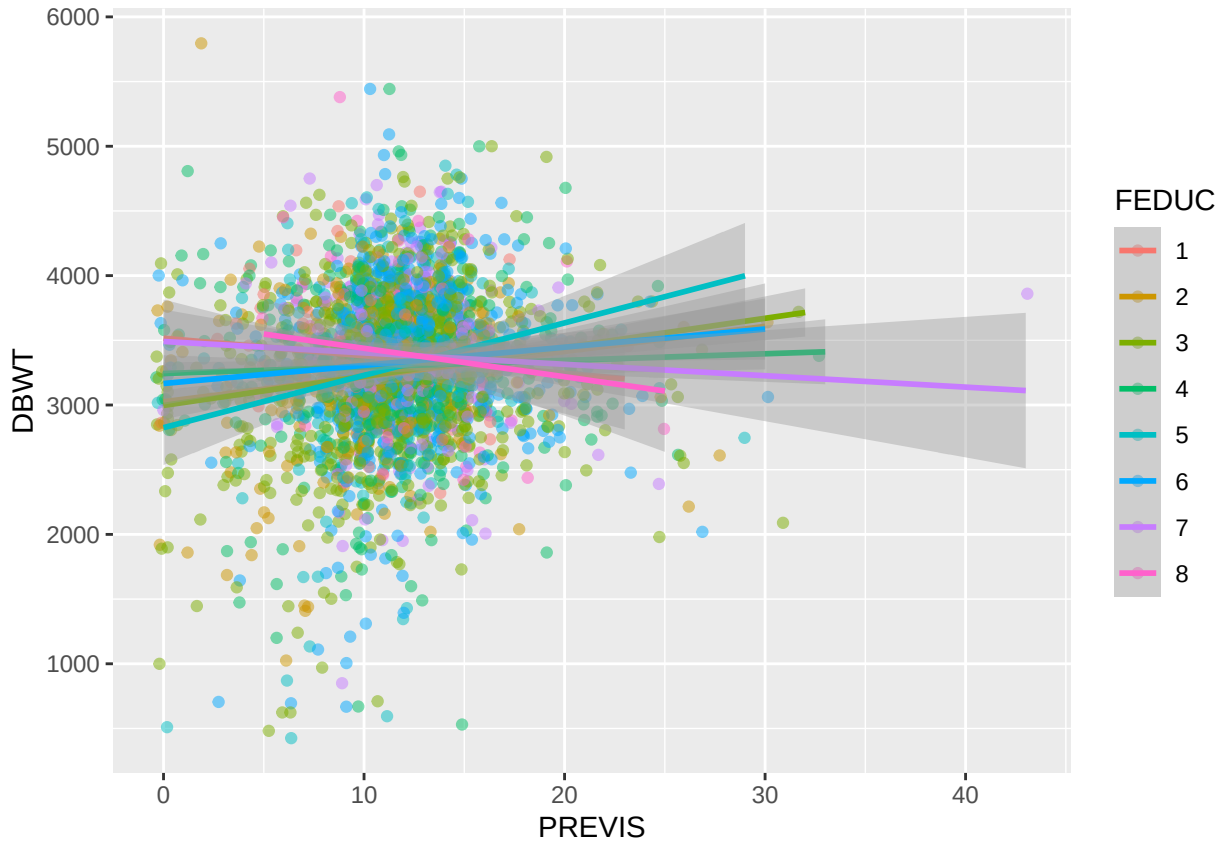



```
ggplot(EDA_df, aes(x = PREVIS, y = DBWT)) +
  geom_point(position = "jitter", aes(colour = PAY), alpha = 0.5) +
  geom_smooth(method = "lm", aes(colour = PAY))
```

```
## `geom_smooth()` using formula 'y ~ x'
```

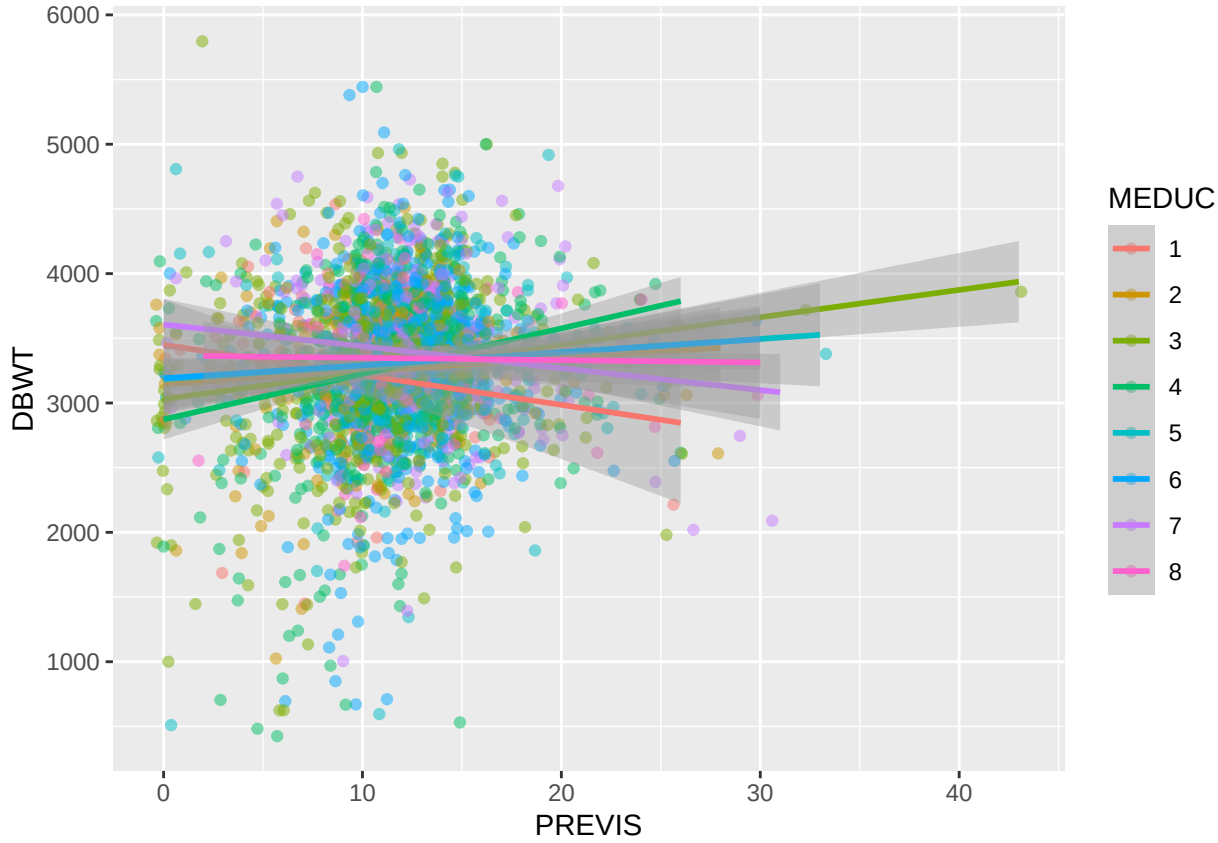

```
ggplot(EDA_df, aes(x = PREVIS, y = DBWT)) +
  geom_point(position = "jitter", aes(colour = FEDUC), alpha = 0.5) +
  geom_smooth(method = "lm", aes(colour = FEDUC))
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
ggplot(EDA_df, aes(x = PREVIS, y = DBWT)) +
  geom_point(position = "jitter", aes(colour = MEDUC), alpha = 0.5) +
  geom_smooth(method = "lm", aes(colour = MEDUC))
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
EDA_df %>% count(FEDUC)
```

```
##   FEDUC   n
## 1     1  76
## 2     2 242
## 3     3 896
## 4     4 570
## 5     5 236
## 6     6 633
## 7     7 249
## 8     8  98
```

```
EDA_df %>% count(MEDUC)
```

```
##   MEDUC   n
## 1     1  66
## 2     2 198
## 3     3 694
## 4     4 602
## 5     5 242
## 6     6 722
## 7     7 370
## 8     8 106
```