Rachel's EDA

```
data <- read.csv("data/EDA.csv")
```

```
head(data)
```

```
##   X ATTEND BFACIL  BMI CIG_0 DBWT DLMP_MM DLMP_YY DMAR DOB_MM DOB_TT DOB_WK
## 1 1      1      1 26.9     0 2920       8    2017    2      4    704      1
## 2 2      4      2 27.6     0 3030       7    2017    1      4   2024      1
## 3 3      1      1 22.3     0 3460      11    2017   NA      8   1740      6
## 4 4      1      1 28.6     0 2735       4    2018    2     12    948      6
## 5 5      1      1 44.1     0 3345       6    2017    2      4   1510      4
## 6 6      3      1 30.2     0 3487      10    2017    1      8    130      4
##   DOB_YY DWgt_R FAGECOMB FEDUC FHISPX FRACE15 FRACE31 FRACE6 ILLB_R ILOP_R
## 1   2018    175       26     3      0       2       2      2     36    888
## 2   2018    152       22     4      1       1       1      1    888    888
## 3   2018    167       26     3      1       1       1      1     87    888
## 4   2018    183       44     1      1       1       1      1    999    999
## 5   2018    290       32     4      0       2       2      2     28    888
## 6   2018    197       32     7      0       1       1      1     41     45
##   ILP_R IMP_SEX IP_GON LD_INDL MAGER MAGE_IMPFLG MAR_IMP MBSTATE_REC MEDUC
## 1    36      NA      N       N    23          NA      NA           1     3
## 2   888      NA      N       N    27          NA      NA           1     6
## 3    87      NA      N       N    26          NA      NA           1     3
## 4   999      NA      N       N    39          NA      NA           2     1
## 5    28      NA      N       N    31          NA      NA           1     3
## 6    41      NA      N       N    32          NA      NA           1     7
##   MHISPX MM_AICU MRACE15 MRACE31 MRACEIMP MRAVE6 MTRAN M_Ht_In NO_INFEC
## 1      0       N       2       2       NA      2     N      63        1
## 2      1       N       1       1       NA      1     N      63        1
## 3      0       N       1       1       NA      1     N      64        1
## 4      1       N       1       1       NA      1     N      66        1
## 5      0       N       2       2       NA      2     N      65        1
## 6      0       N       1       1       NA      1     N      61        1
##   NO_MMORB NO_RISKS PAY PAY_REC PRECARE PREVIS PRIORDEAD PRIORLIVE PRIORTERM
## 1        1        1   1       1       1     13         0         1         0
## 2        1        1   3       3       1     13         0         0         0
## 3        1        1   2       2       1     12         0         1         0
## 4        1        0   1       1       2     16         0         2         2
## 5        1        1   1       1       1     10         0         1         0
## 6        1        1   2       2       1     13         0         1         2
##   PWgt_R RDMETH_REC RESTATUS RF_CESAR RF_CESARN SEX WTGAIN pregnancy.length
## 1    152          1        1        N         0   M     23               8
## 2    156          1        1        N         0   F      0               9
## 3    130          1        2        N         0   F     37               9
## 4    177          4        2        Y         2   F      6               8
## 5    265          1        1        N         0   M     25              10
```

```
## 6    160          1        1        N         0   M     37               10
##   WT.percent.gain CIG_0_BIN
## 1      0.15131579         0
## 2      0.00000000         0
## 3      0.28461538         0
## 4      0.03389831         0
## 5      0.09433962         0
## 6      0.23125000         0
```

```
nrow(data)
```

```
## [1] 3000
```

```
# dropping datapoints with missing valus
colnames(data)
```

```
##  [1] "X"                "ATTEND"           "BFACIL"           "BMI"
##  [5] "CIG_0"            "DBWT"             "DLMP_MM"          "DLMP_YY"
##  [9] "DMAR"             "DOB_MM"           "DOB_TT"           "DOB_WK"
## [13] "DOB_YY"           "DWgt_R"           "FAGECOMB"         "FEDUC"
## [17] "FHISPX"           "FRACE15"          "FRACE31"          "FRACE6"
## [21] "ILLB_R"           "ILOP_R"           "ILP_R"            "IMP_SEX"
## [25] "IP_GON"           "LD_INDL"          "MAGER"            "MAGE_IMPFLG"
## [29] "MAR_IMP"          "MBSTATE_REC"      "MEDUC"            "MHISPX"
## [33] "MM_AICU"          "MRACE15"          "MRACE31"          "MRACEIMP"
## [37] "MRAVE6"           "MTRAN"            "M_Ht_In"          "NO_INFEC"
## [41] "NO_MMORB"         "NO_RISKS"         "PAY"              "PAY_REC"
## [45] "PRECARE"          "PREVIS"           "PRIORDEAD"        "PRIORLIVE"
## [49] "PRIORTERM"        "PWgt_R"           "RDMETH_REC"       "RESTATUS"
## [53] "RF_CESAR"         "RF_CESARN"        "SEX"              "WTGAIN"
## [57] "pregnancy.length" "WT.percent.gain"  "CIG_0_BIN"
```

```
data <- subset(data, PRECARE!=99 & CIG_0!=99 & BMI!=99.9 & PWgt_R!=999  & PREVIS!=99 & WTGAIN!=99 & DBW
```

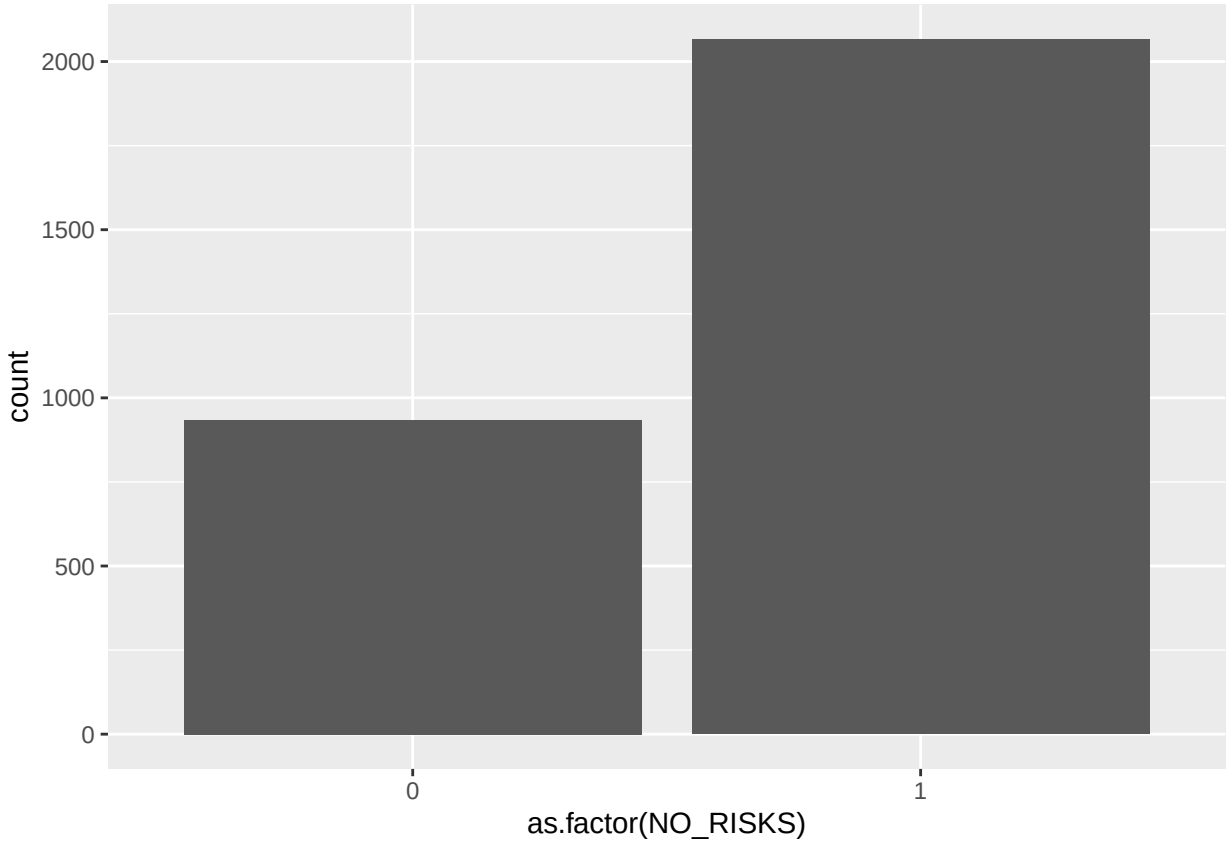
```
data$pregnancy.length <- 12*(data$DOB_YY - data$DLMP_YY) + (data$DOB_MM - data$DLMP_MM)
data$WT.percent.gain <- data$WTGAIN / data$PWgt_R
```

```
# binarize CIG_0
data$CIG_0_BIN <- ifelse(data$CIG_0 > 0,1,0)
```

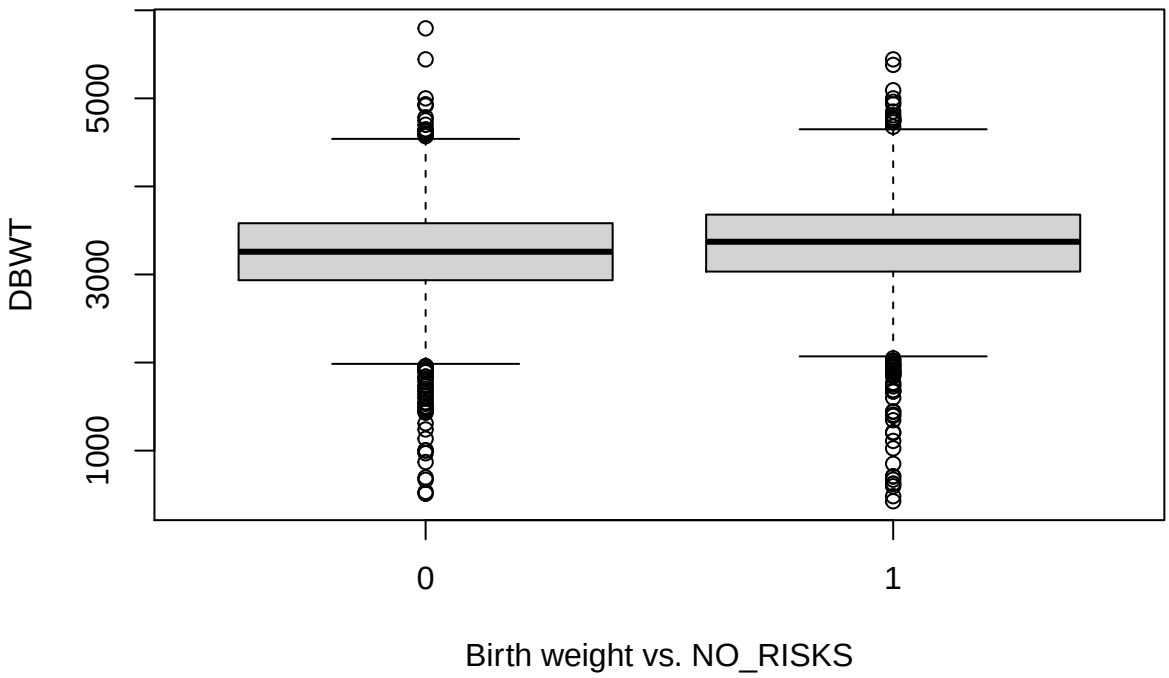
EDA NO_RISKS (9) MRAVE6[sic] — Mother's race recode 6 FRACE6 — Father's race recode 6 (9) MEDUC (9) FEDUC (9)

```
library(ggplot2)
```

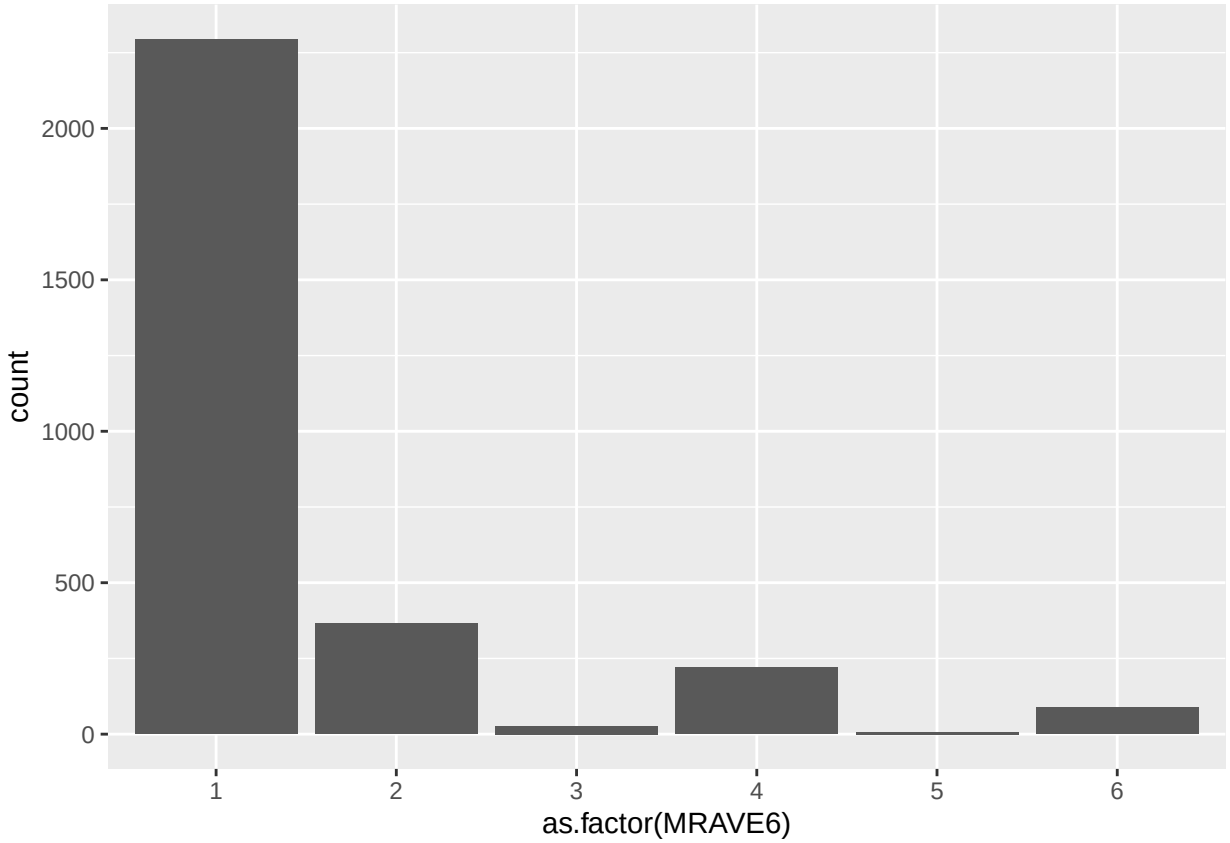
```
ggplot(data, aes(x=as.factor(NO_RISKS))) + geom_bar()
```



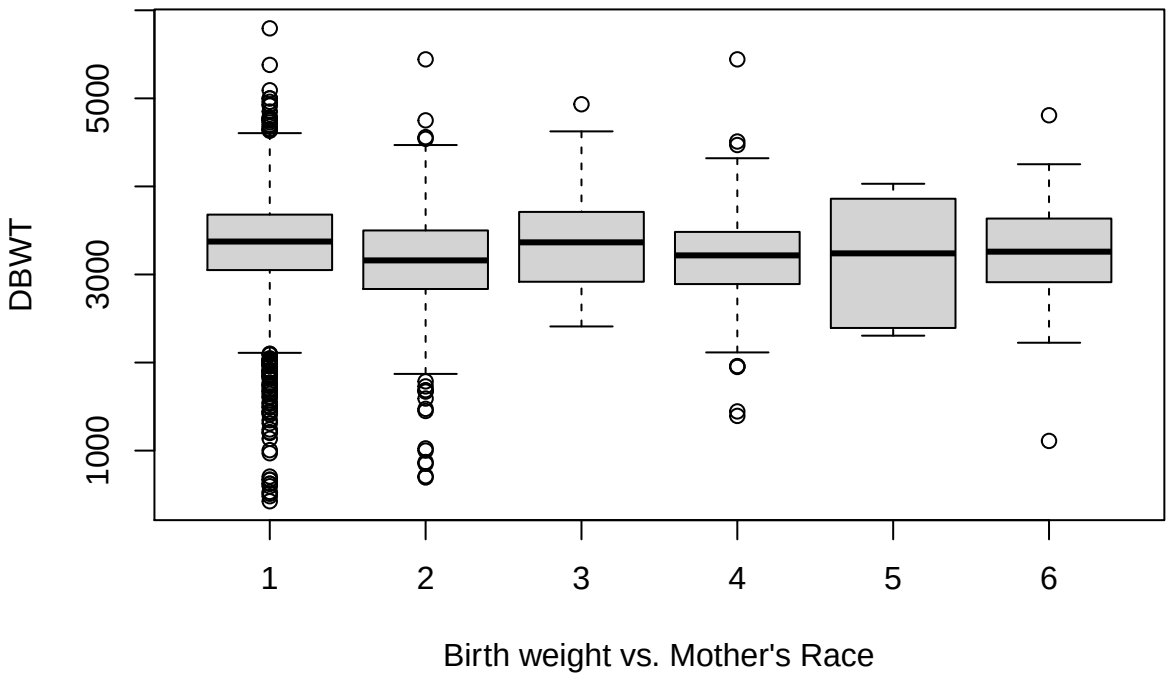

```
boxplot(DBWT ~ as.factor(NO_RISKS), data, xlab = "Birth weight vs. NO_RISKS")
```



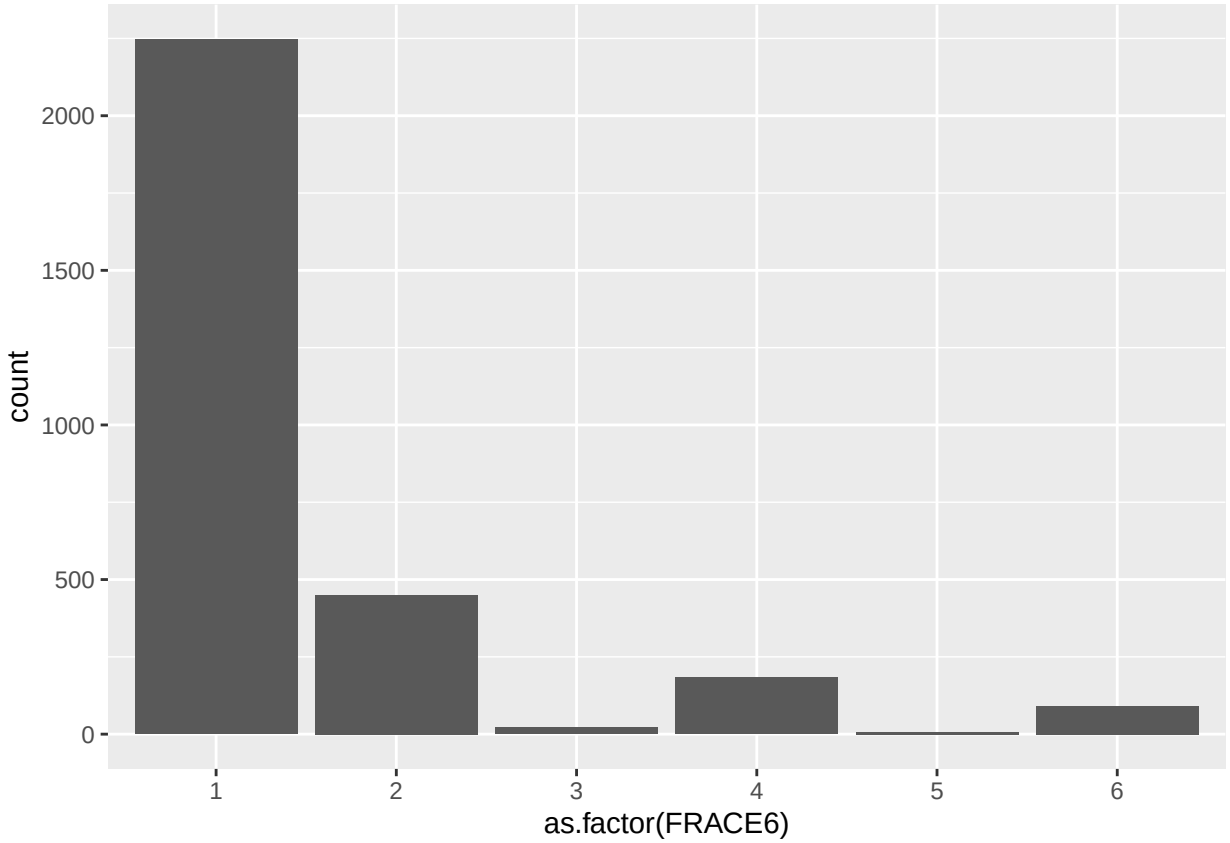

```
ggplot(data, aes(x=as.factor(MRAVE6))) + geom_bar()
```



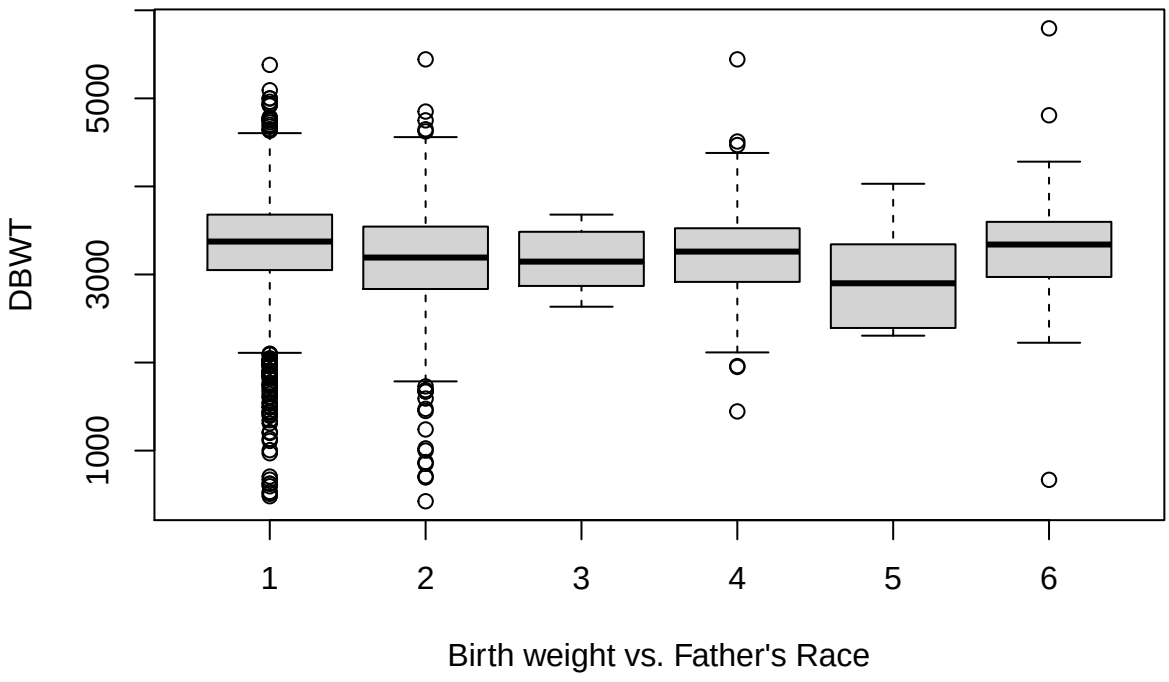

```
boxplot(DBWT ~ as.factor(MRAVE6), data, xlab = "Birth weight vs. Mother's Race")
```



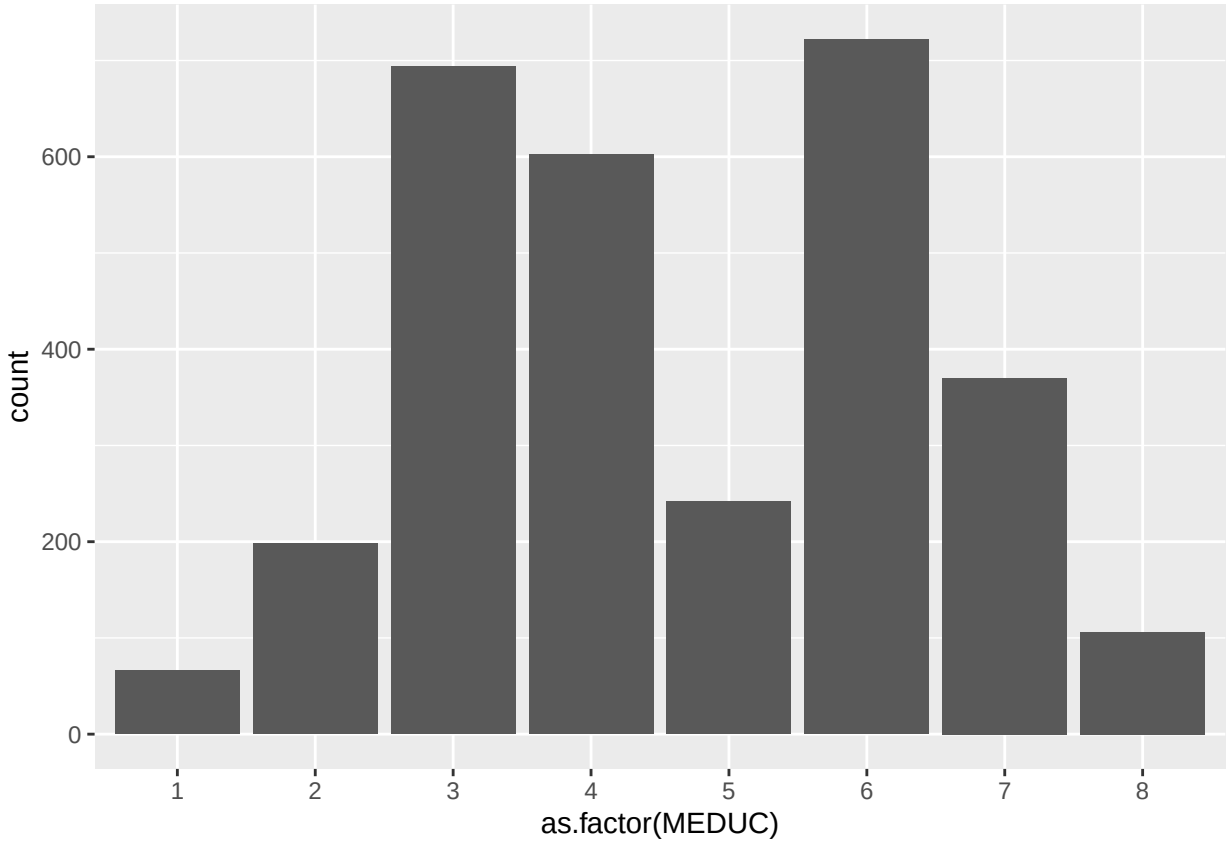

```
ggplot(data, aes(x=as.factor(FRACE6))) + geom_bar()
```



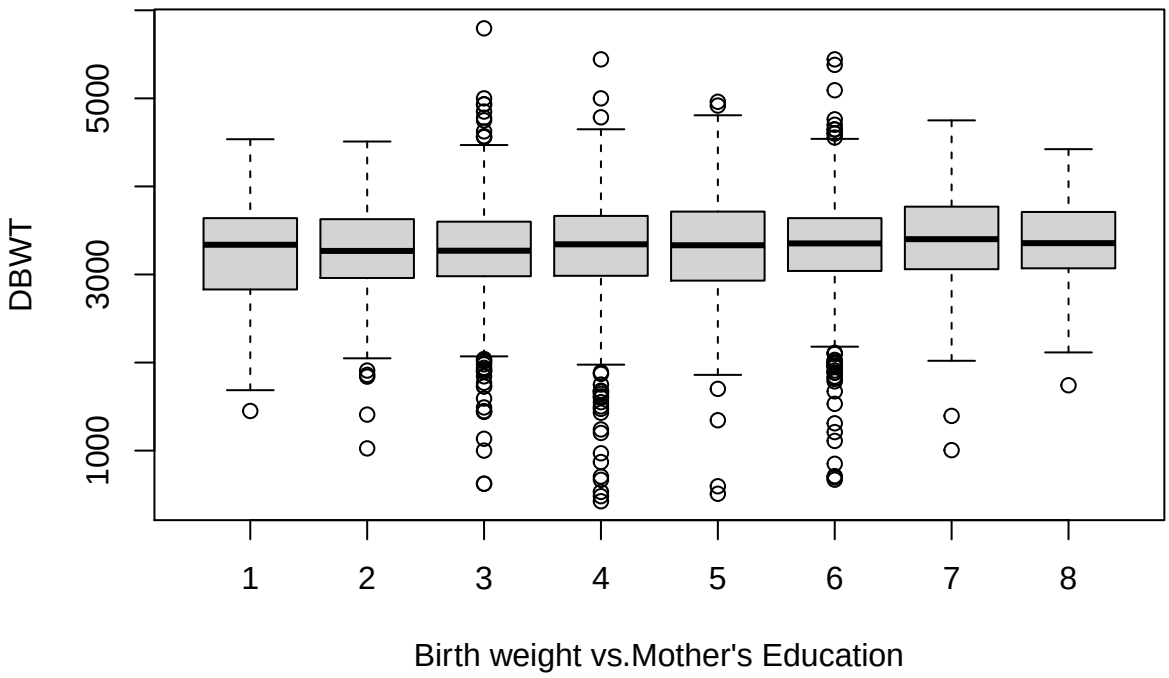

```
boxplot(DBWT ~ as.factor(FRACE6), data, xlab = "Birth weight vs. Father's Race")
```



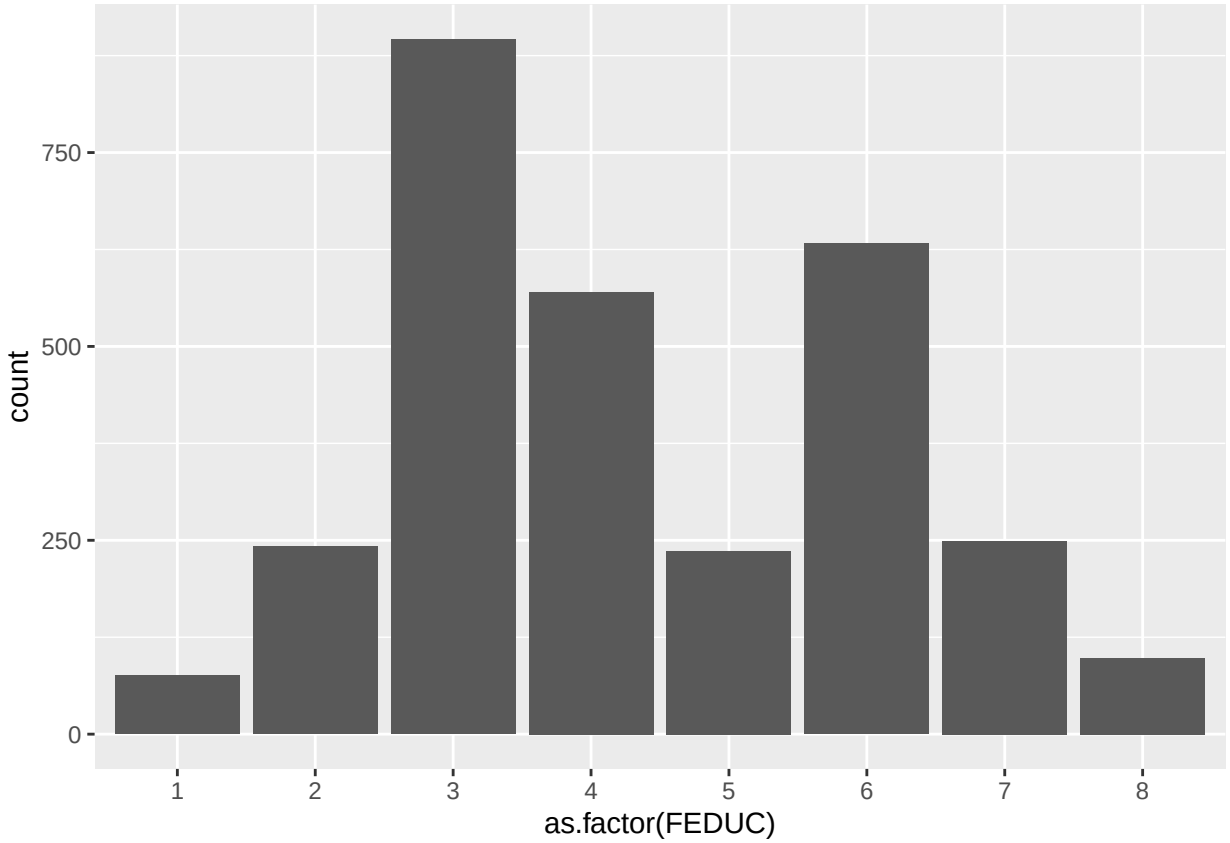

```
ggplot(data, aes(x=as.factor(MEDUC))) + geom_bar()
```



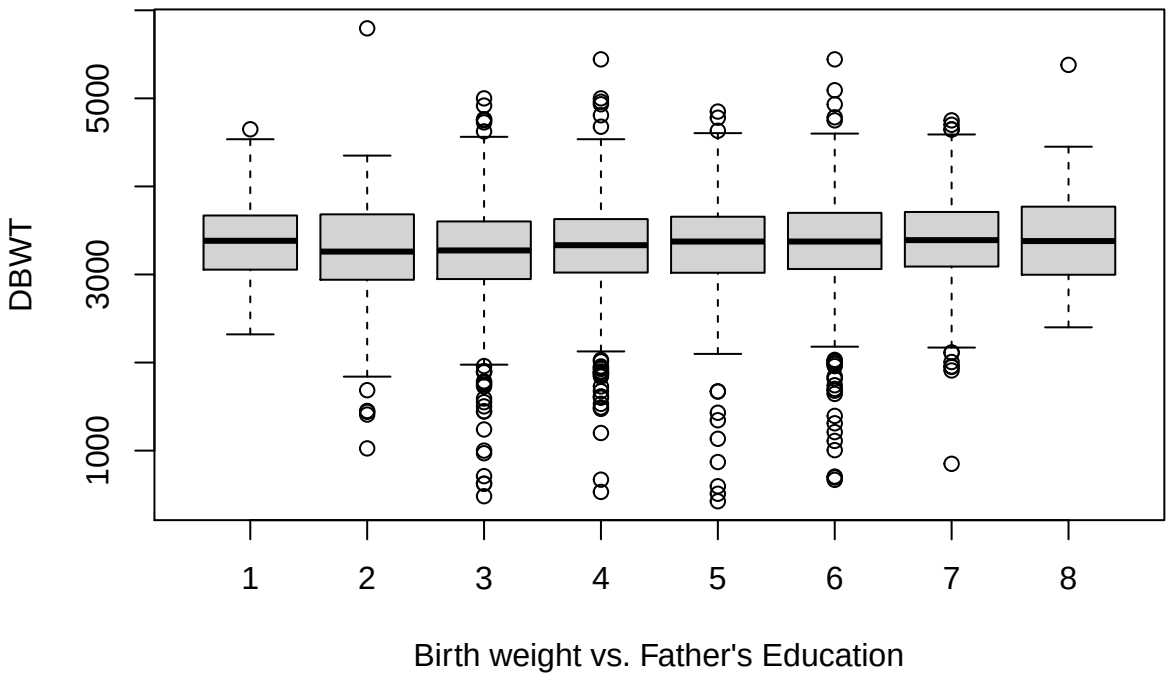

```
boxplot(DBWT ~ as.factor(MEDUC), data, xlab = "Birth weight vs.Mother's Education")
```



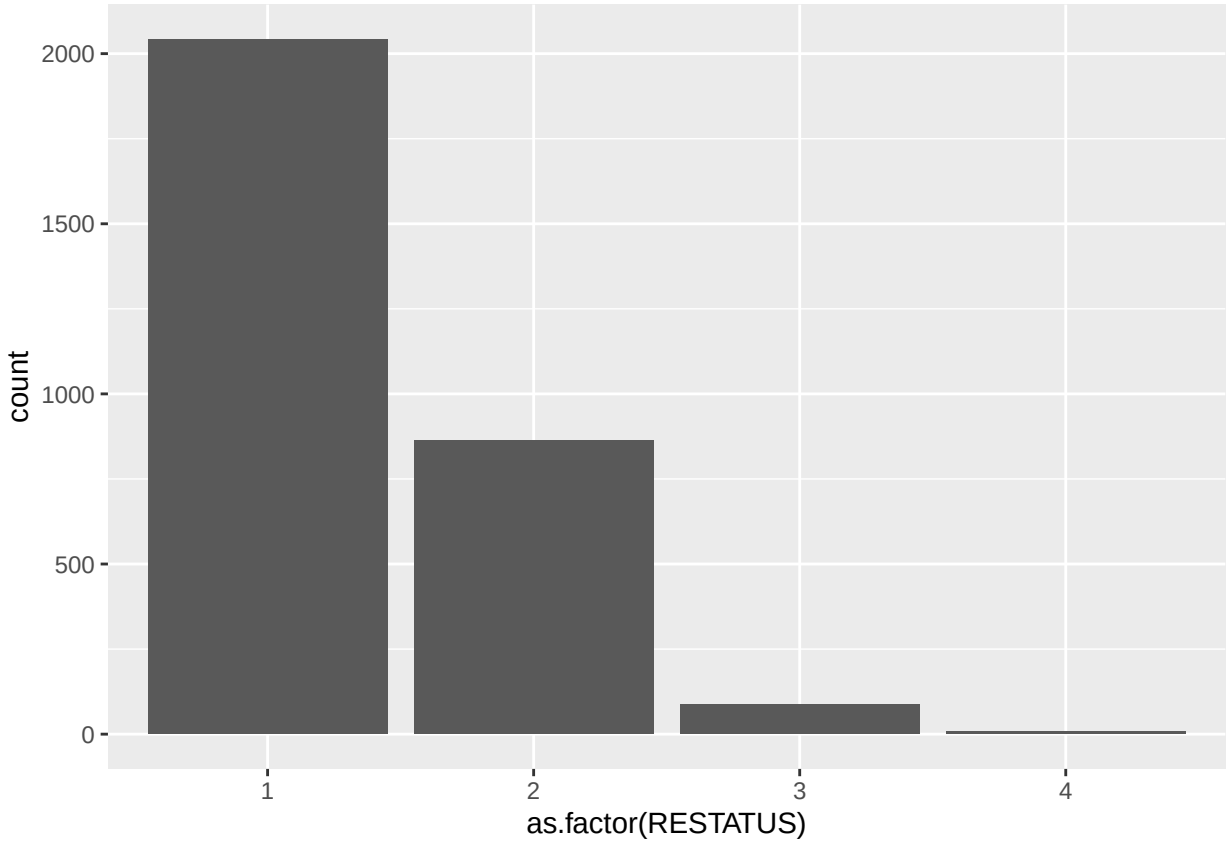

```
ggplot(data, aes(x=as.factor(FEDUC))) + geom_bar()
```



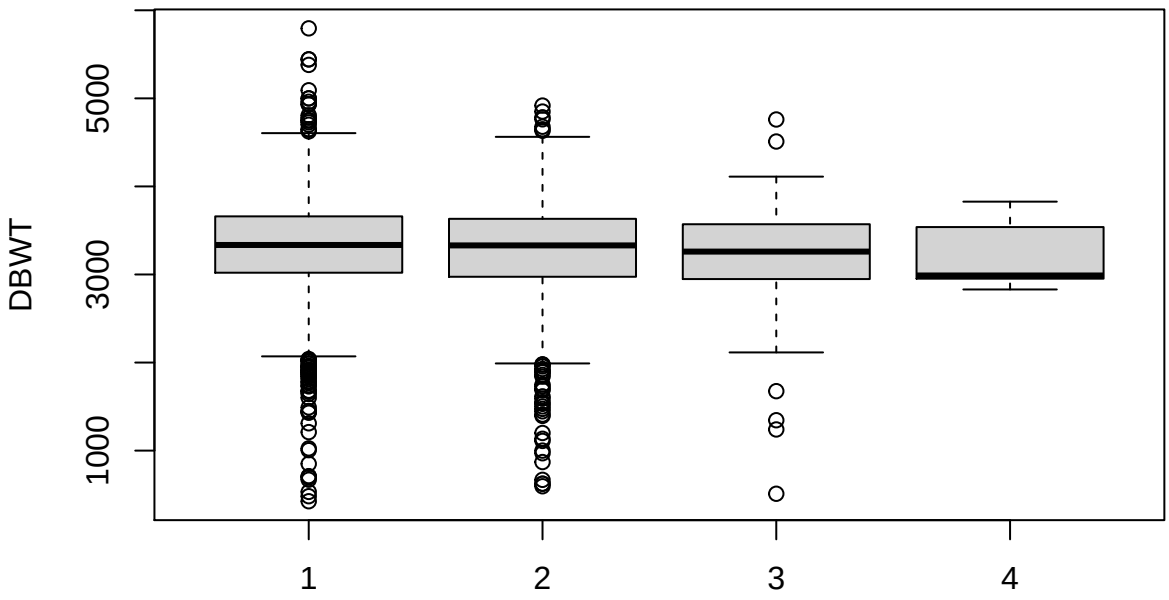

```
boxplot(DBWT ~ as.factor(FEDUC), data, xlab = "Birth weight vs. Father's Education")
```




```
ggplot(data, aes(x=as.factor(RESTATUS))) + geom_bar()
```

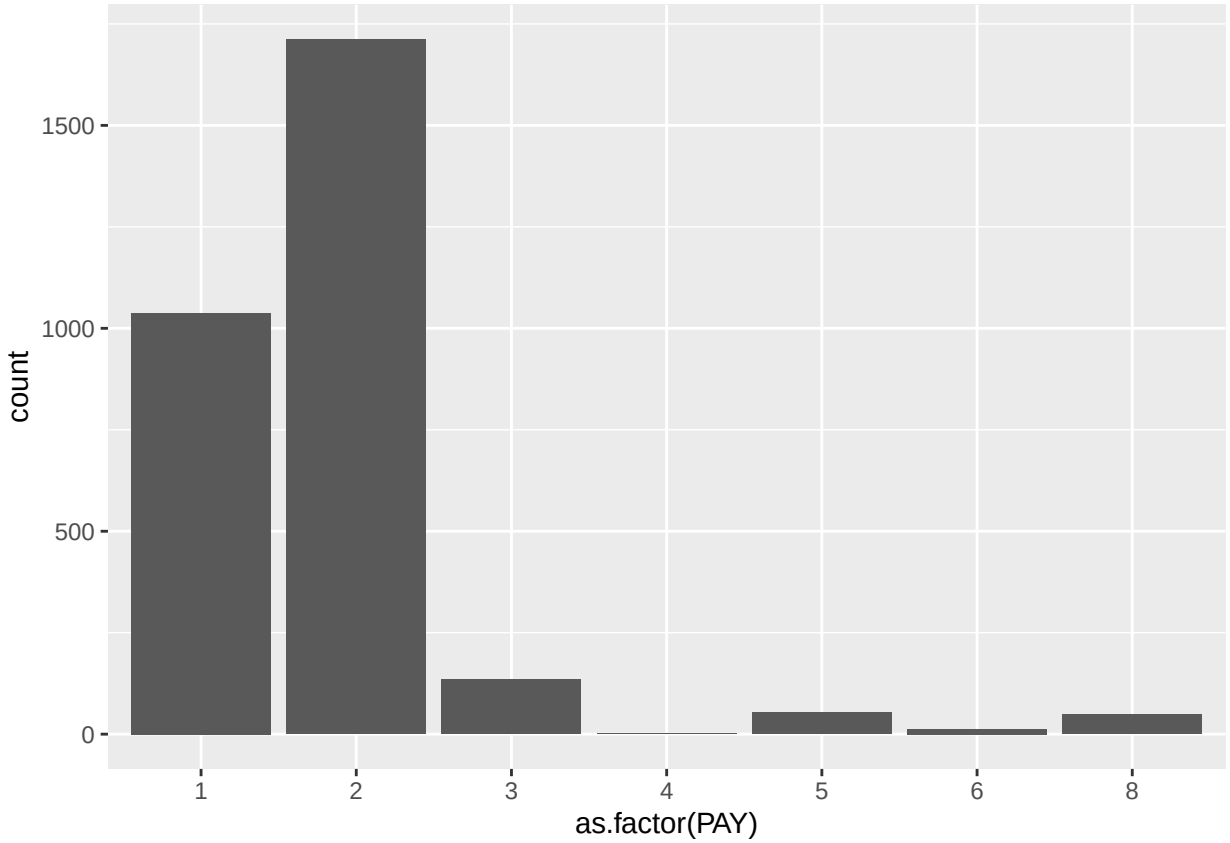



```
boxplot(DBWT ~ as.factor(RESTATUS), data, xlab = "Birth weight vs. Residential Status")
```

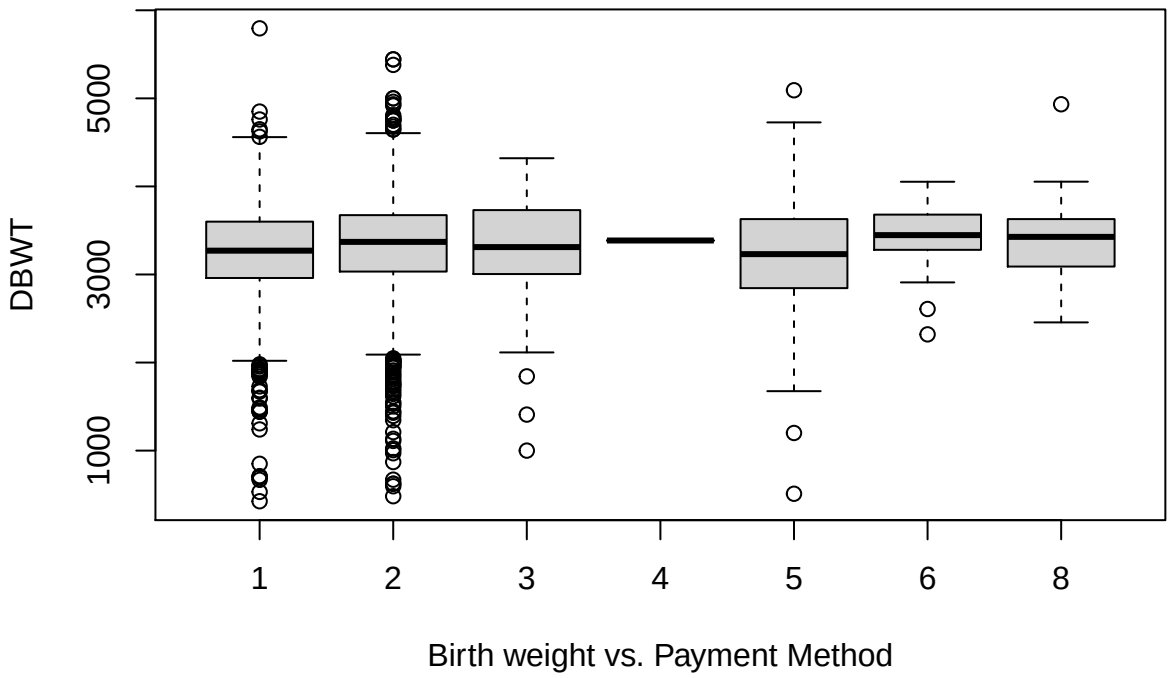


Birth weight vs. Residential Status

```
ggplot(data, aes(x=as.factor(PAY))) + geom_bar()
```

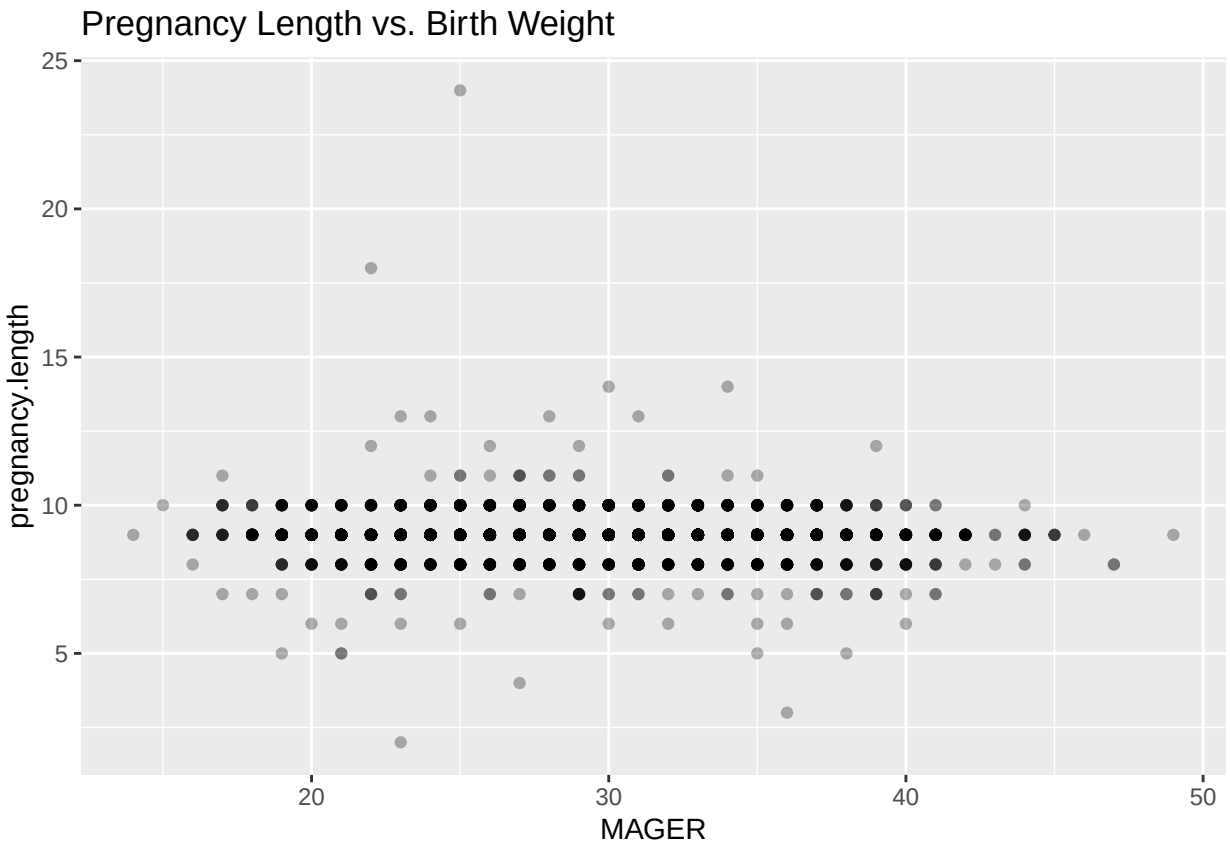



```
boxplot(DBWT ~ as.factor(PAY), data, xlab = "Birth weight vs. Payment Method")
```



MAGER & pregnancy length

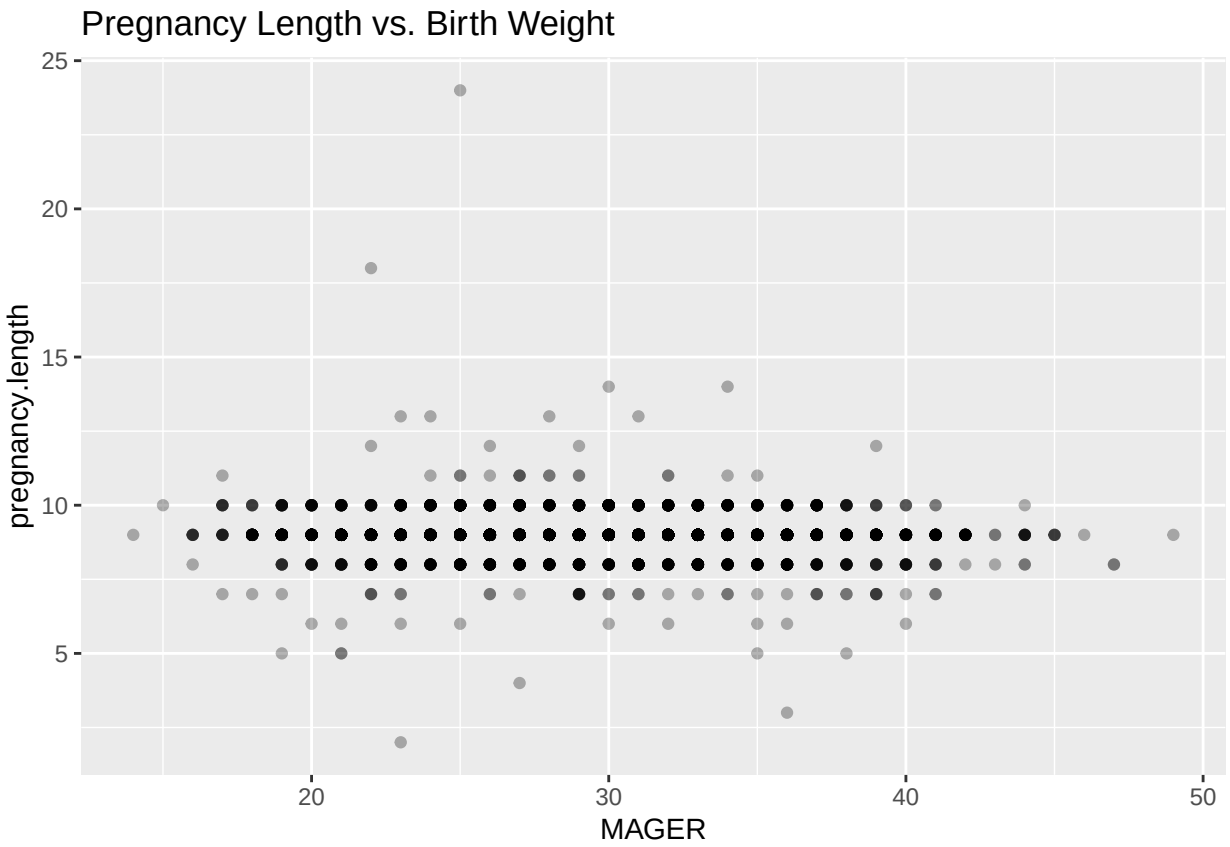
```
ggplot(data, aes(x = MAGER, y = pregnancy.length)) +
  geom_point(alpha = 0.3) + labs(title = "Pregnancy Length vs. Birth Weight") #+
```




```
# theme(
 #  plot.title = element_text(size = 5, hjust = 0.5),
  # text = element_text(size = 5),
   #legend.position = "bottom"
 #)
```

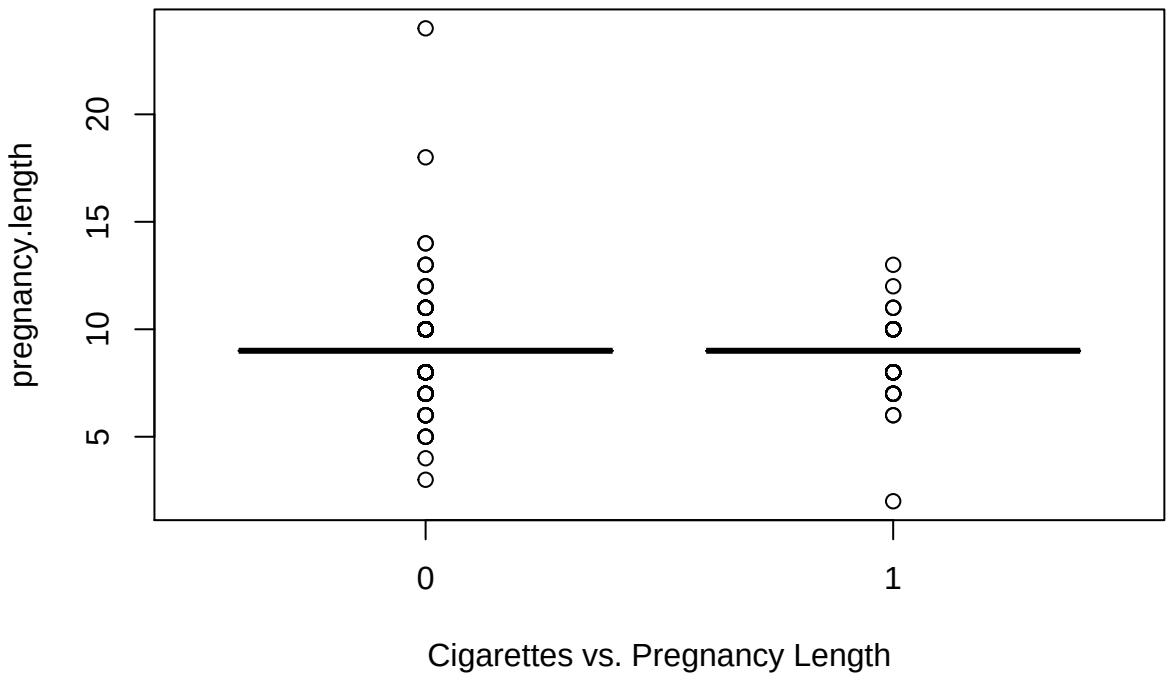
PRECARE & pregnancy length

```
ggplot(data, aes(x = MAGER, y = pregnancy.length)) +
  geom_point(alpha = 0.3) + labs(title = "Pregnancy Length vs. Birth Weight")
```



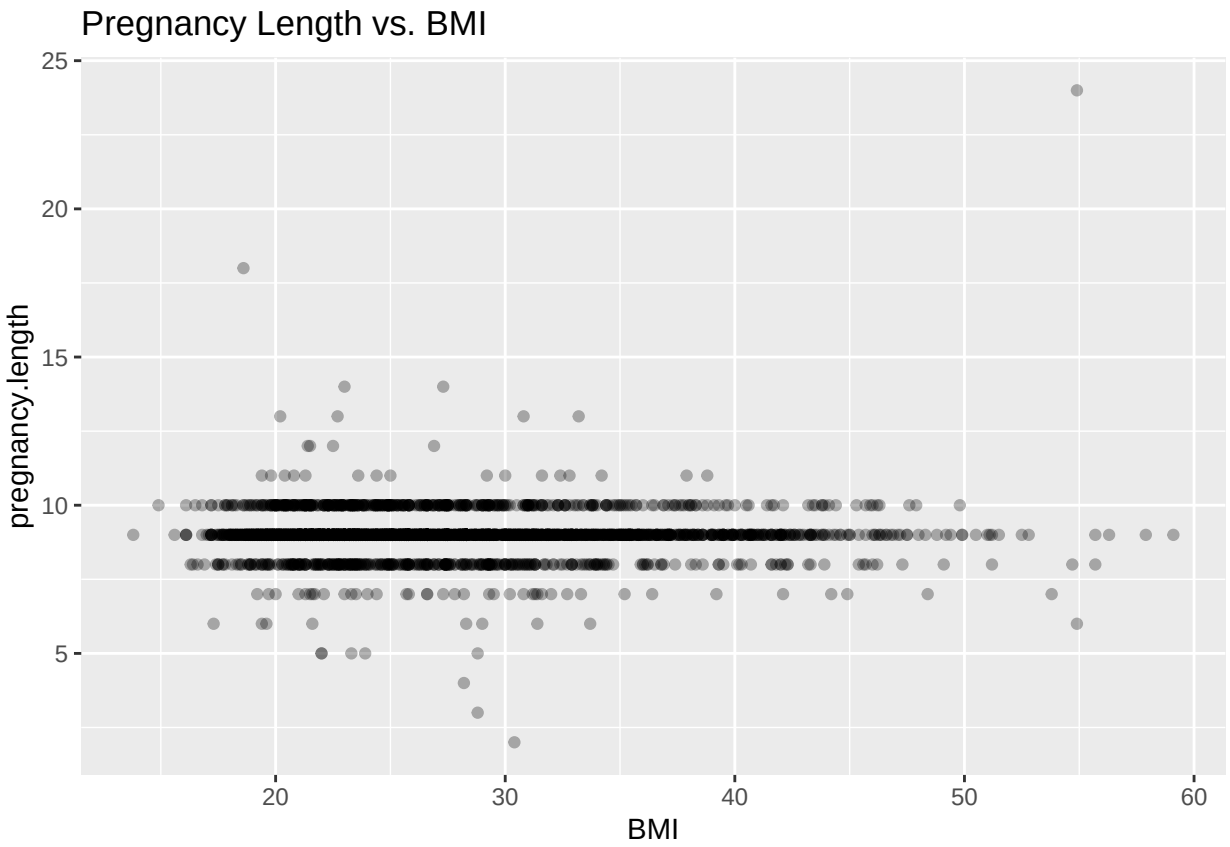
CIG_0 and pregnancy length

```
boxplot(pregnancy.length ~ as.factor(CIG_0_BIN), data, xlab = "Cigarettes vs. Pregnancy Length")
```



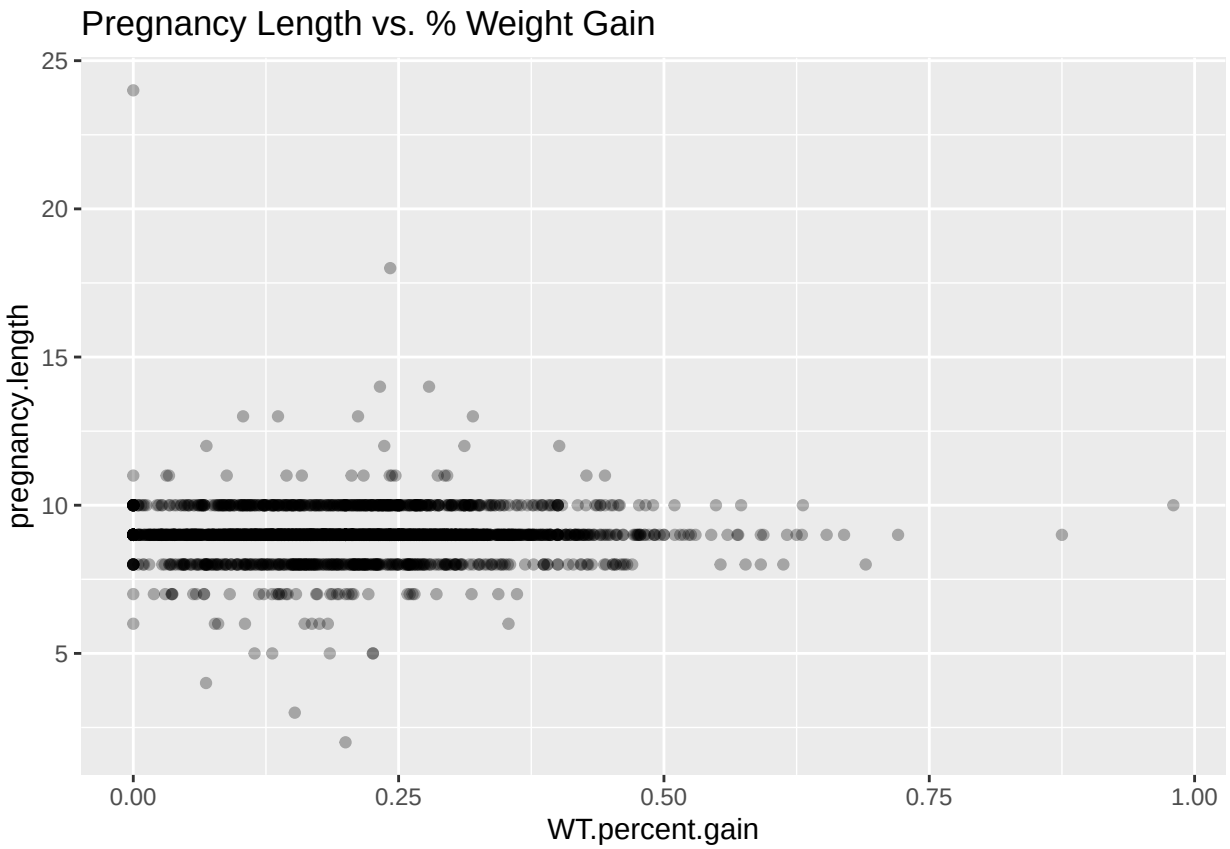
BMI and pregnancy length

```
ggplot(data, aes(x = BMI, y = pregnancy.length)) +
  geom_point(alpha = 0.3) + labs(title = "Pregnancy Length vs. BMI")
```



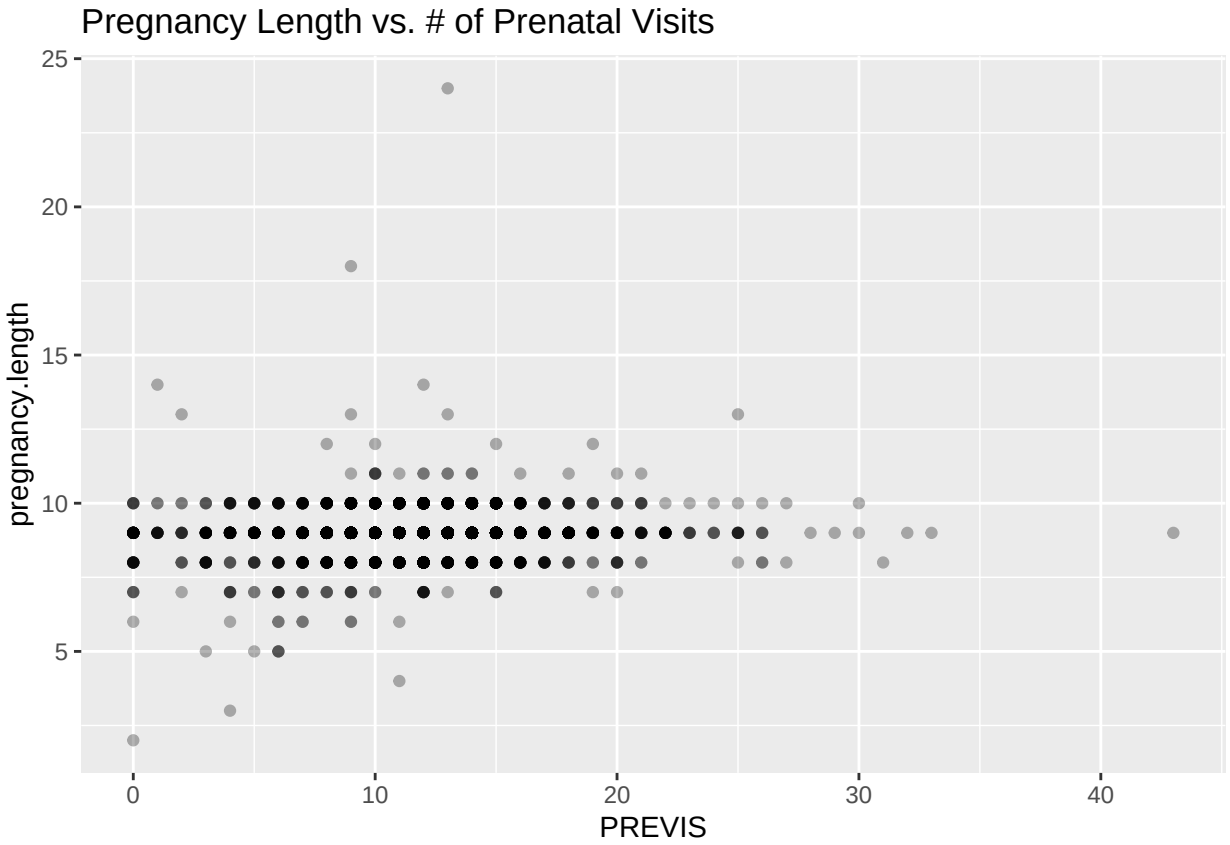
weight gain & preg length

```
ggplot(data, aes(x = WT.percent.gain, y = pregnancy.length)) +
  geom_point(alpha = 0.3) + labs(title = "Pregnancy Length vs. % Weight Gain")
```



PREVIS & pregnancy length

```
ggplot(data, aes(x = PREVIS, y = pregnancy.length)) +
  geom_point(alpha = 0.3) + labs(title = "Pregnancy Length vs. # of Prenatal Visits")
```



MEDUC & PAY

FEDUC & PAY

NO_RISKS & MRAVE6

NO_RISKS & FRACE6