Lab 4: Healthy Momma, Healthy Baby

Krist Mar and Nikki Haas

12/1/2016

A Nice Introduction that Makes Us Sound Like Pros

All work and no play makes Nikki a dull girl. All work and no play makes Nikki a dull girl. All work and no play makes Nikki a dull girl. All work and no play makes Nikki a dull girl. All work and no play makes Nikki a dull girl. All work and no play makes Nikki a dull girl. All work and no play makes Nikki a dull girl. All work and no play makes Nikki a dull girl. All work and no play makes Nikki a dull girl. All work and no play makes Nikki a dull girl. All work and no play makes Nikki a dull girl. All work and no play makes Nikki a dull girl. All work and no play makes Nikki a dull girl. All work and no play makes Nikki a dull girl.

Step 1: Read in the Data

```
setwd('/Users/nicholeh/student285/w203/lab_4/Lab_4')
load('bwght_w203.RData')
desc
```

```
##    variable                           label
## 1      mage             mother's age, years
## 2     meduc            mother's educ, years
## 3    monpre       month prenatal care began
## 4     npvis total number of prenatal visits
## 5      fage             father's age, years
## 6     feduc            father's educ, years
## 7     bwght             birth weight, grams
## 8     omaps          one minute apgar score
## 9     fmaps         five minute apgar score
## 10     cigs          avg cigarettes per day
## 11    drink             avg drinks per week
## 12      lbw             =1 if bwght <= 2000
## 13     vlbw             =1 if bwght <= 1500
## 14     male                 =1 if baby male
## 15    mwhte              =1 if mother white
## 16    mblck              =1 if mother black
## 17     moth           =1 if mother is other
## 18    fwhte              =1 if father white
## 19    fblck              =1 if father black
## 20     foth           =1 if father is other
## 21   lbwght                      log(bwght)
## 22   magesq                          mage^2
## 23  npvissq                         npvis^2
```

Step 2: Exploratory Data Analysis

First, get summary statistics on each element of the dataset:

```
nrow(data)
```

```
## [1] 1832
```

```
summary(data)
```

```
##       mage            meduc            monpre           npvis
##  Min.   :16.00   Min.   : 3.00   Min.   :0.000   Min.   : 0.00
##  1st Qu.:26.00   1st Qu.:12.00   1st Qu.:1.000   1st Qu.:10.00
##  Median :29.00   Median :13.00   Median :2.000   Median :12.00
##  Mean   :29.56   Mean   :13.72   Mean   :2.122   Mean   :11.62
##  3rd Qu.:33.00   3rd Qu.:16.00   3rd Qu.:2.000   3rd Qu.:13.00
##  Max.   :44.00   Max.   :17.00   Max.   :9.000   Max.   :40.00
##                  NA's   :30      NA's   :5       NA's   :68
##       fage            feduc            bwght          omaps
##  Min.   :18.00   Min.   : 3.00   Min.   : 360   Min.   : 0.000
##  1st Qu.:28.00   1st Qu.:12.00   1st Qu.:3076   1st Qu.: 8.000
##  Median :31.00   Median :14.00   Median :3425   Median : 9.000
##  Mean   :31.92   Mean   :13.92   Mean   :3401   Mean   : 8.386
##  3rd Qu.:35.00   3rd Qu.:16.00   3rd Qu.:3770   3rd Qu.: 9.000
##  Max.   :64.00   Max.   :17.00   Max.   :5204   Max.   :10.000
##  NA's   :6       NA's   :47                     NA's   :3
##      fmaps             cigs            drink             lbw
##  Min.   : 2.000   Min.   : 0.000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.: 9.000   1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.:0.00000
##  Median : 9.000   Median : 0.000   Median :0.0000   Median :0.00000
##  Mean   : 9.004   Mean   : 1.089   Mean   :0.0198   Mean   :0.01638
##  3rd Qu.: 9.000   3rd Qu.: 0.000   3rd Qu.:0.0000   3rd Qu.:0.00000
##  Max.   :10.000   Max.   :40.000   Max.   :8.0000   Max.   :1.00000
##  NA's   :3        NA's   :110      NA's   :115
##       vlbw               male            mwhte            mblck
##  Min.   :0.000000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.000000   1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:0.0000
##  Median :0.000000   Median :1.0000   Median :1.0000   Median :0.0000
##  Mean   :0.007096   Mean   :0.5136   Mean   :0.8865   Mean   :0.0595
##  3rd Qu.:0.000000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :1.000000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##
##       moth              fwhte            fblck              foth
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:1.0000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.00000   Median :1.0000   Median :0.00000   Median :0.00000
##  Mean   :0.05404   Mean   :0.8897   Mean   :0.05841   Mean   :0.05186
##  3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.00000   Max.   :1.00000
##
##      lbwght           magesq          npvissq
##  Min.   :5.886   Min.   : 256.0   Min.   :   0.0
##  1st Qu.:8.031   1st Qu.: 676.0   1st Qu.: 100.0
##  Median :8.139   Median : 841.0   Median : 144.0
##  Mean   :8.114   Mean   : 896.4   Mean   : 148.6
##  3rd Qu.:8.235   3rd Qu.:1089.0   3rd Qu.: 169.0
##  Max.   :8.557   Max.   :1936.0   Max.   :1600.0
##                                   NA's   :68
```

Response Variables

The bwght, lbwght, omaps and fmaps variables are related to the health of the baby.

The first thing to check is if these variables are collinar. We will omit lbwghts as that is a function of bwghts.

```
library(ggplot2)
cor(data$omaps, data$fmaps, use = "complete.obs")
```

```
## [1] 0.5575238
```

```
cor(data$lbwght, data$fmaps, use = "complete.obs")
```

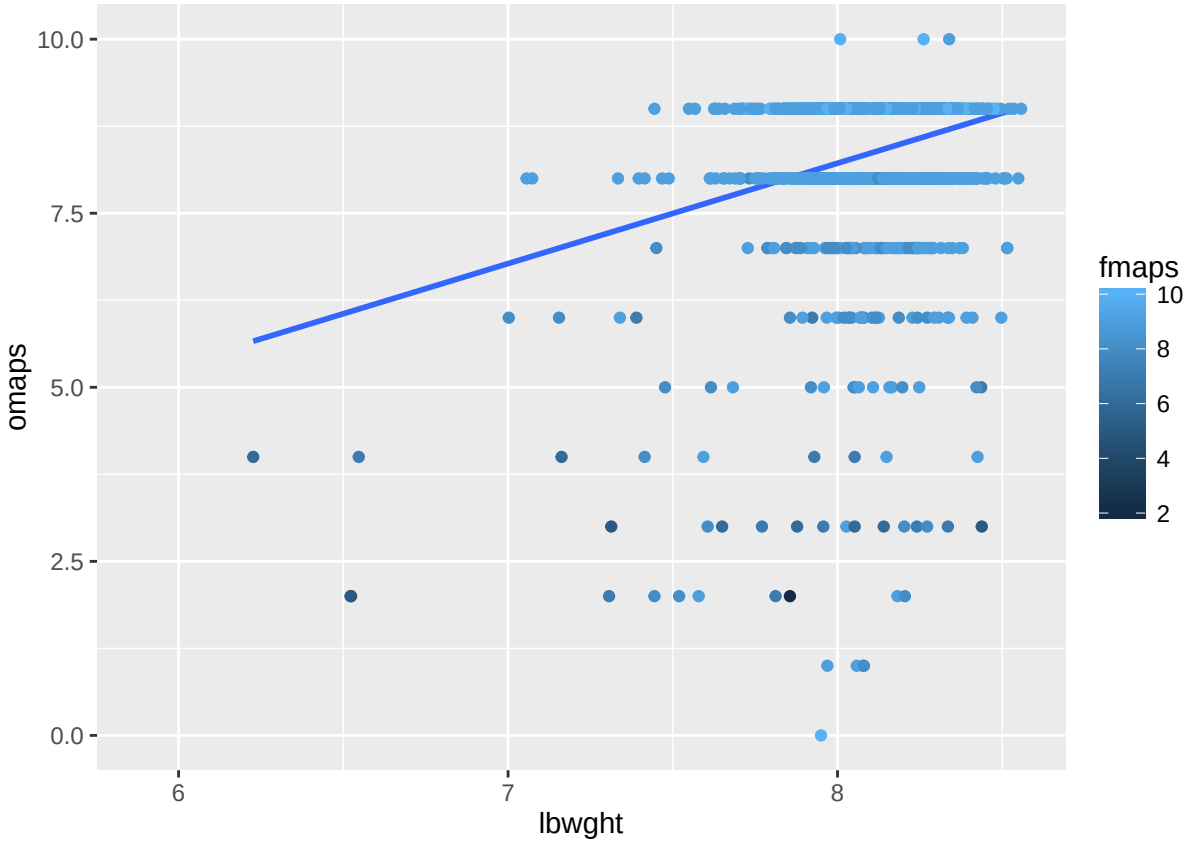
```
## [1] 0.2710456
```

```
p <- ggplot(data, aes(lbwght, omaps)) + geom_point(size = 0.25) +
  geom_smooth(method = "lm", se = FALSE) + geom_point(aes(colour = fmaps))
p
```

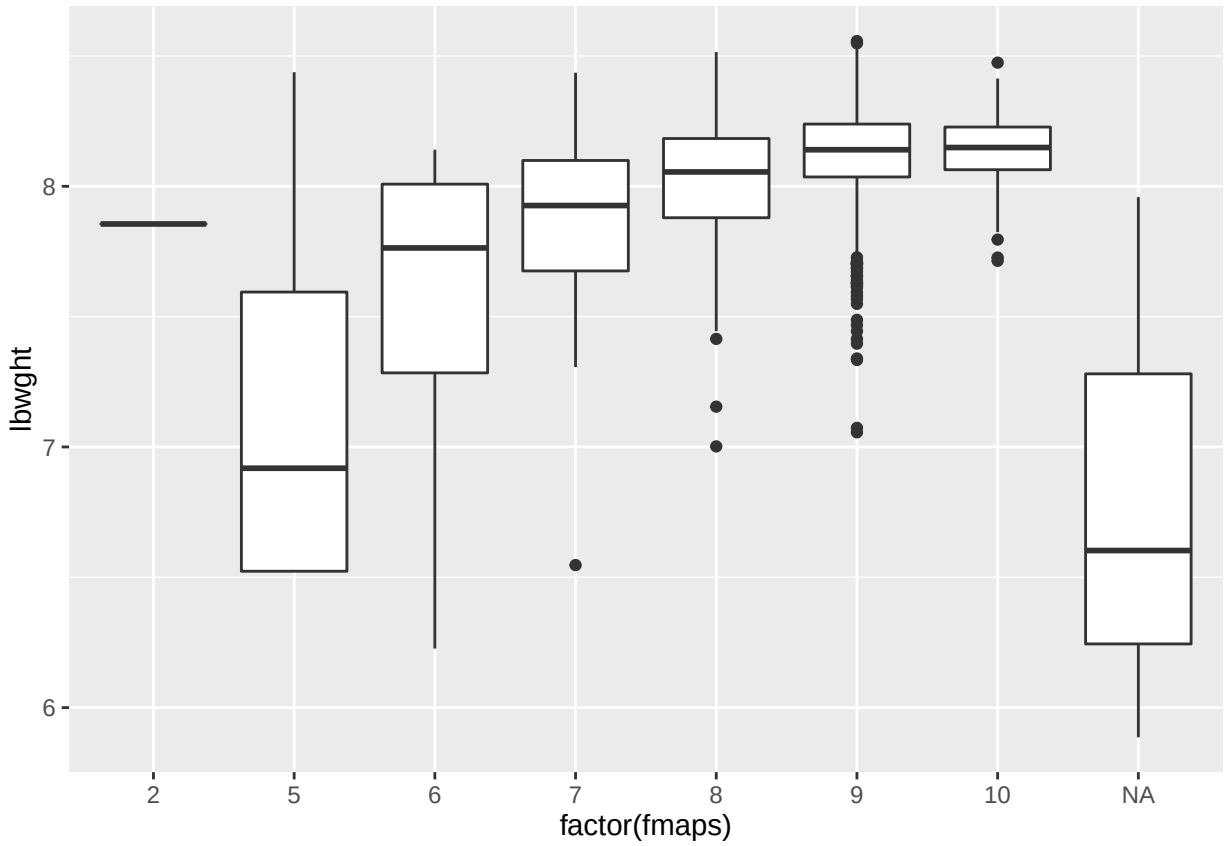
```
## Warning: Removed 3 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 3 rows containing missing values (geom_point).
```

```
## Warning: Removed 3 rows containing missing values (geom_point).
```




```
p <- ggplot(data, aes(factor(fmaps), lbwght)) + geom_boxplot()
p
```



Looking at the data, we can be reasonably assured that the response variables are related, but not collinear. It may be best to make a combined variable of `fmaps` and `omaps` such as `mapscombined = fmaps + omaps`. The difference would not make much sense as the sum; 10 - 10 and 2 - 2 are both zero, after all.

Regressors

The variables monpre and npvis are related to the prenatal care given during pregnancy. Let us review them for collinearity:

```
cor(data$npvis, data$monpre, use = "complete.obs")
```

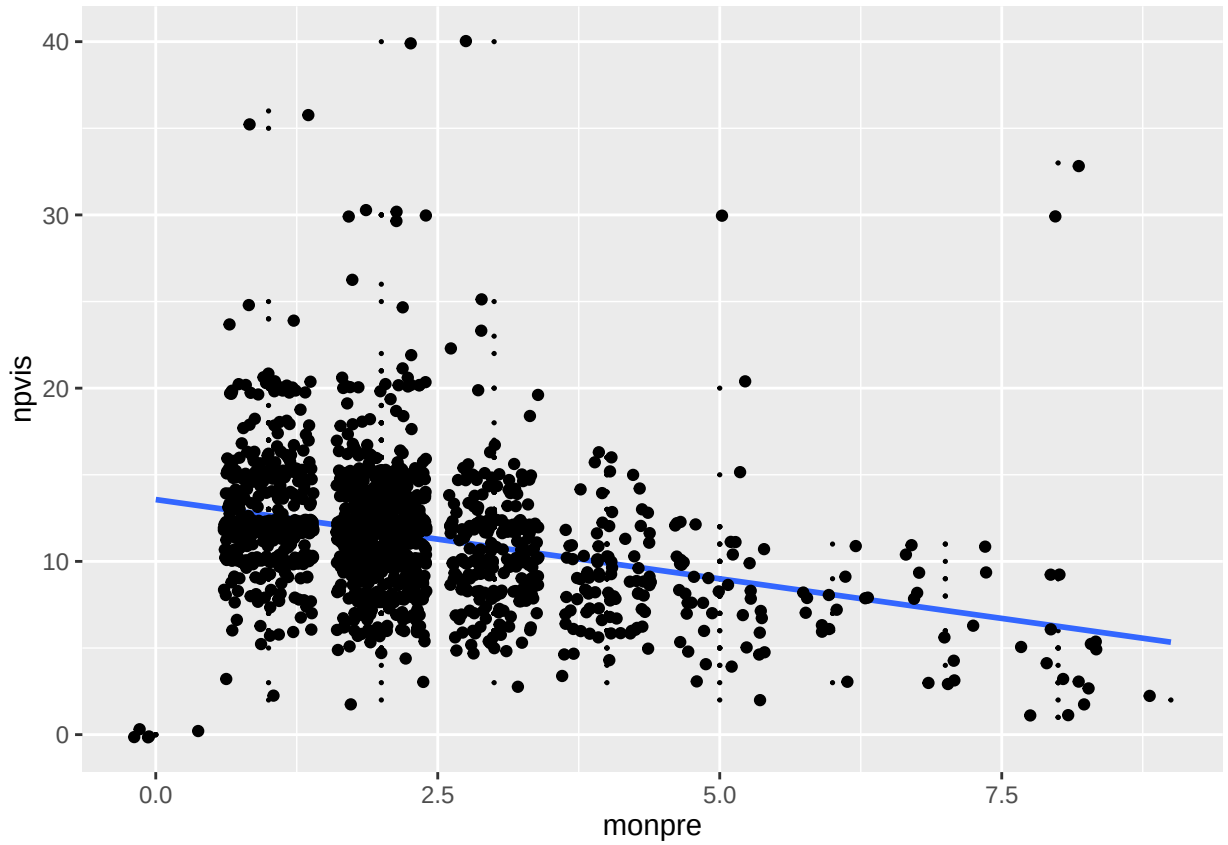
```
## [1] -0.3061006
```

```
ggplot(data, aes(monpre, npvis)) + geom_point(size = 0.25) +
  geom_smooth(method = "lm", se = FALSE) + geom_jitter()
```

```
## Warning: Removed 69 rows containing non-finite values (stat_smooth).
```

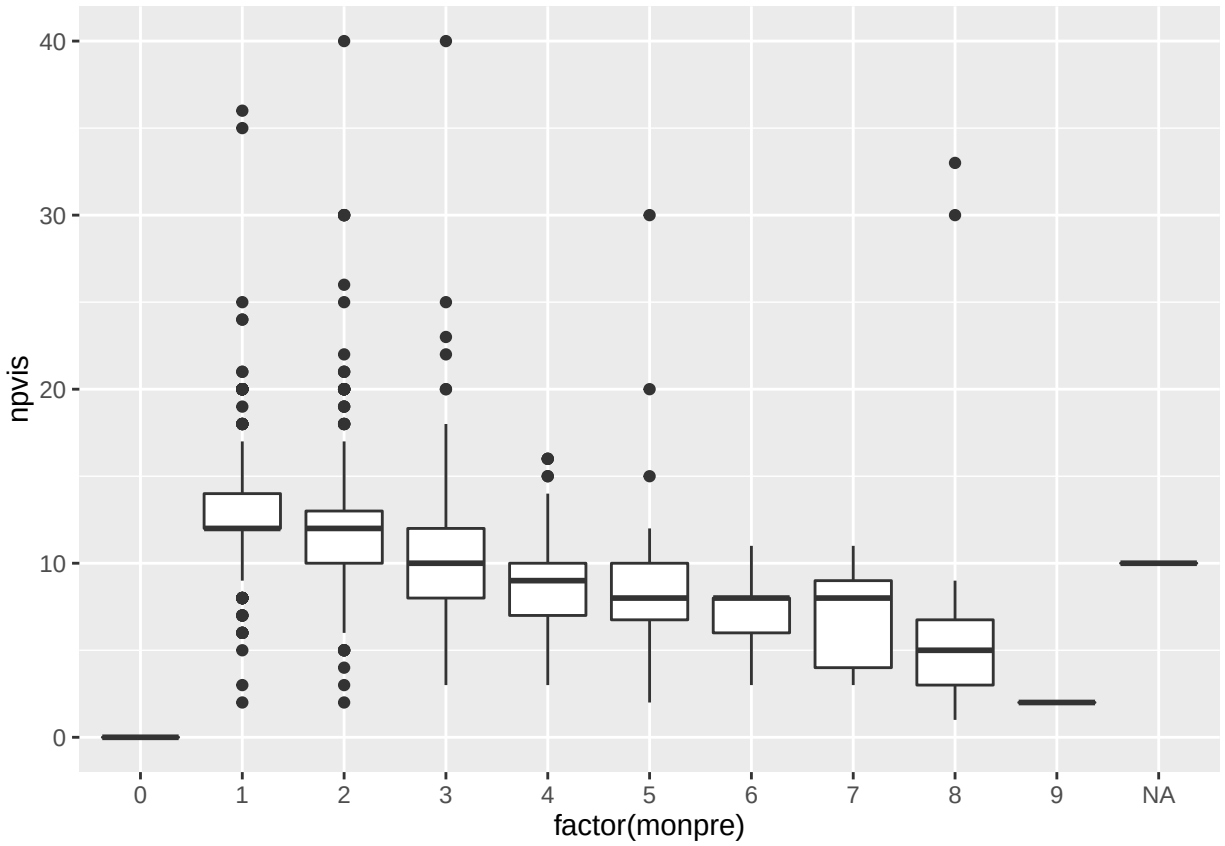
```
## Warning: Removed 69 rows containing missing values (geom_point).
```

```
## Warning: Removed 69 rows containing missing values (geom_point).
```




```
ggplot(data, aes(factor(monpre), npvis)) + geom_boxplot()
```

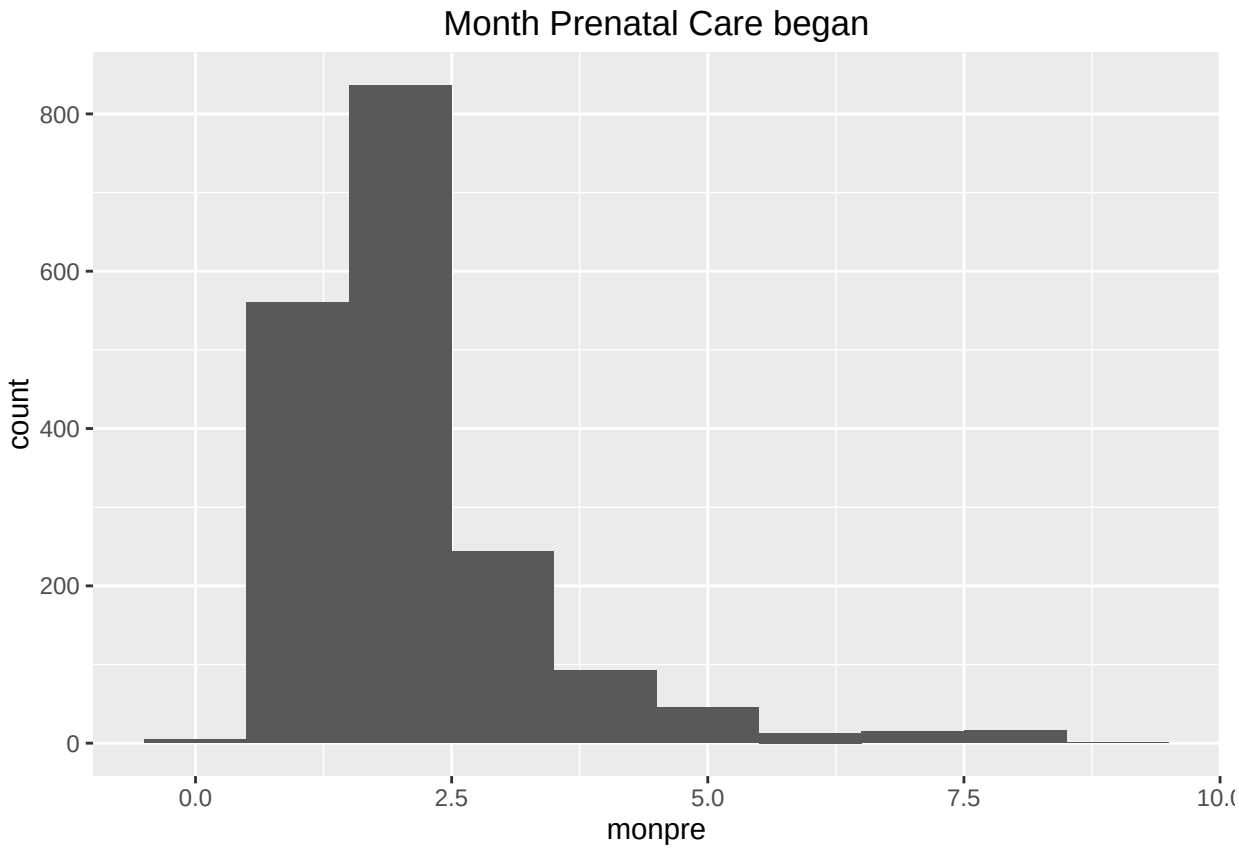
```
## Warning: Removed 68 rows containing non-finite values (stat_boxplot).
```



From this set, we can see that the data is not collinear, and indeed we can see that we might have some reporting errors. 5 mothers are listed as starting prenatal care in month 0 of their pregnancy, but they visited the doctor 0 times. These probably denote missing information or an error in reporting. Unfortunately, this data does show a definitive downward trend leading us to suspect that the number of visits is a function of month prenatal care began. This makes sense intuitively; if a mother starts prenatal care in her 2nd month of pregnancy, she has ample time for frequent doctor visits. However, if she starts her prenatal care towards the end of her pregnancy, she does not have enough time to visit the doctor as often as a woman who started in month 2.

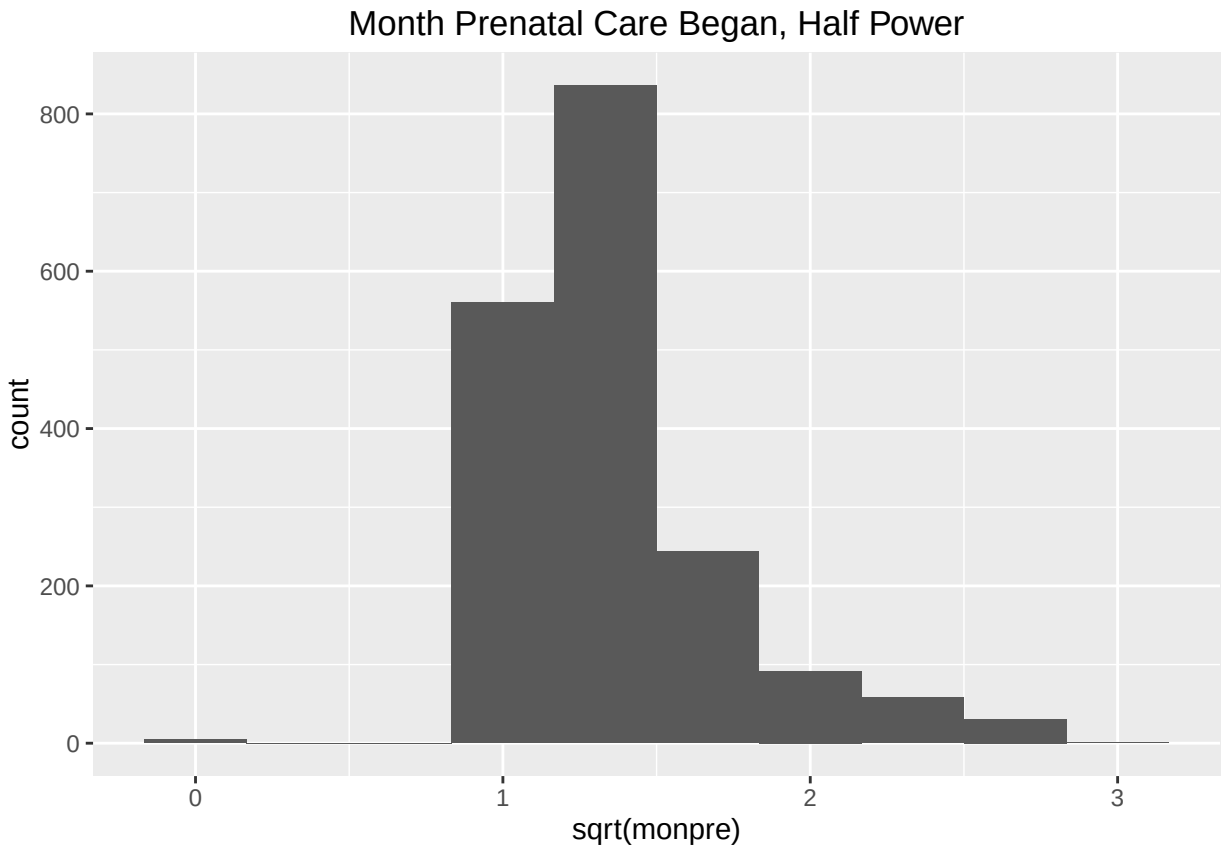
```
ggplot(data, aes(x=monpre)) + geom_histogram(aes(y = ..count..),bins = 10) +
  ggtitle("Month Prenatal Care began")
```

```
## Warning: Removed 5 rows containing non-finite values (stat_bin).
```



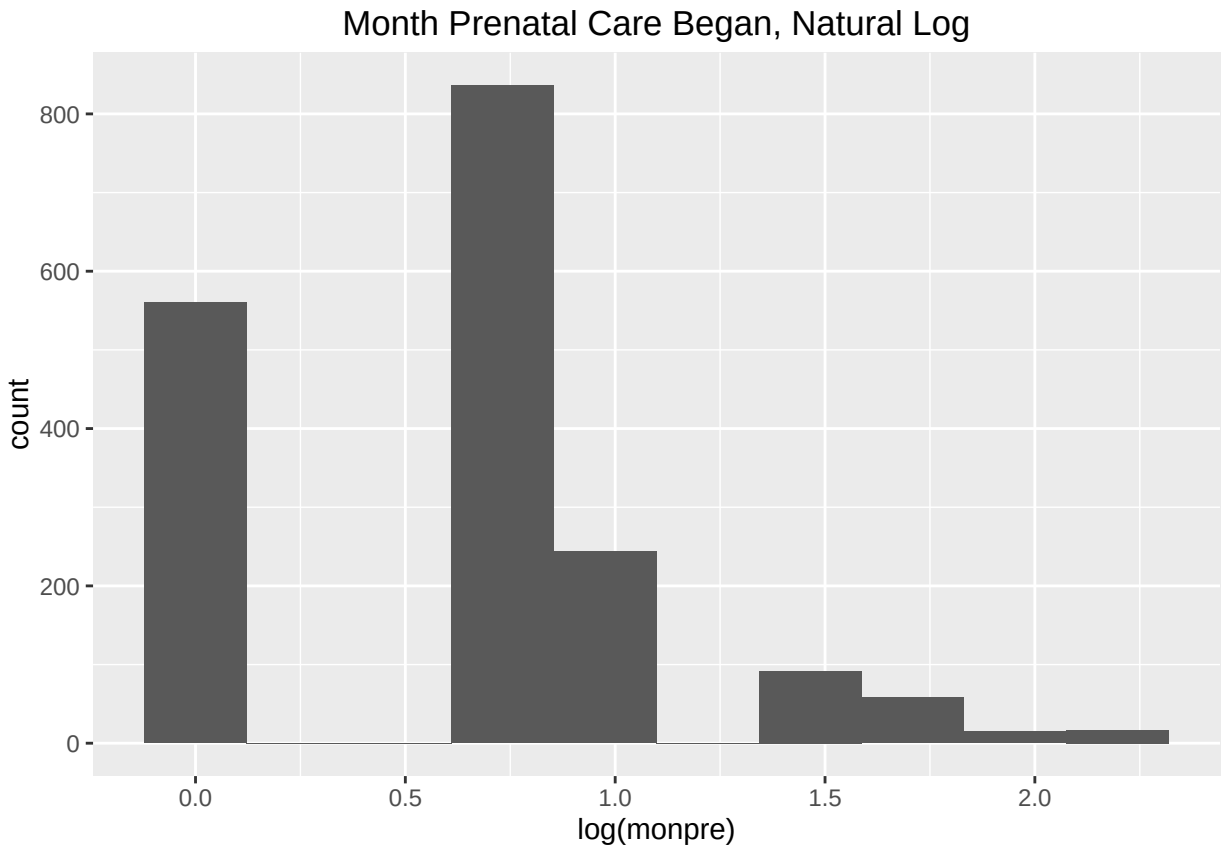

```
ggplot(data, aes(x=sqrt(monpre))) + geom_histogram(aes(y = ..count..), bins = 10) +
  ggtitle("Month Prenatal Care Began, Half Power")
```

```
## Warning: Removed 5 rows containing non-finite values (stat_bin).
```



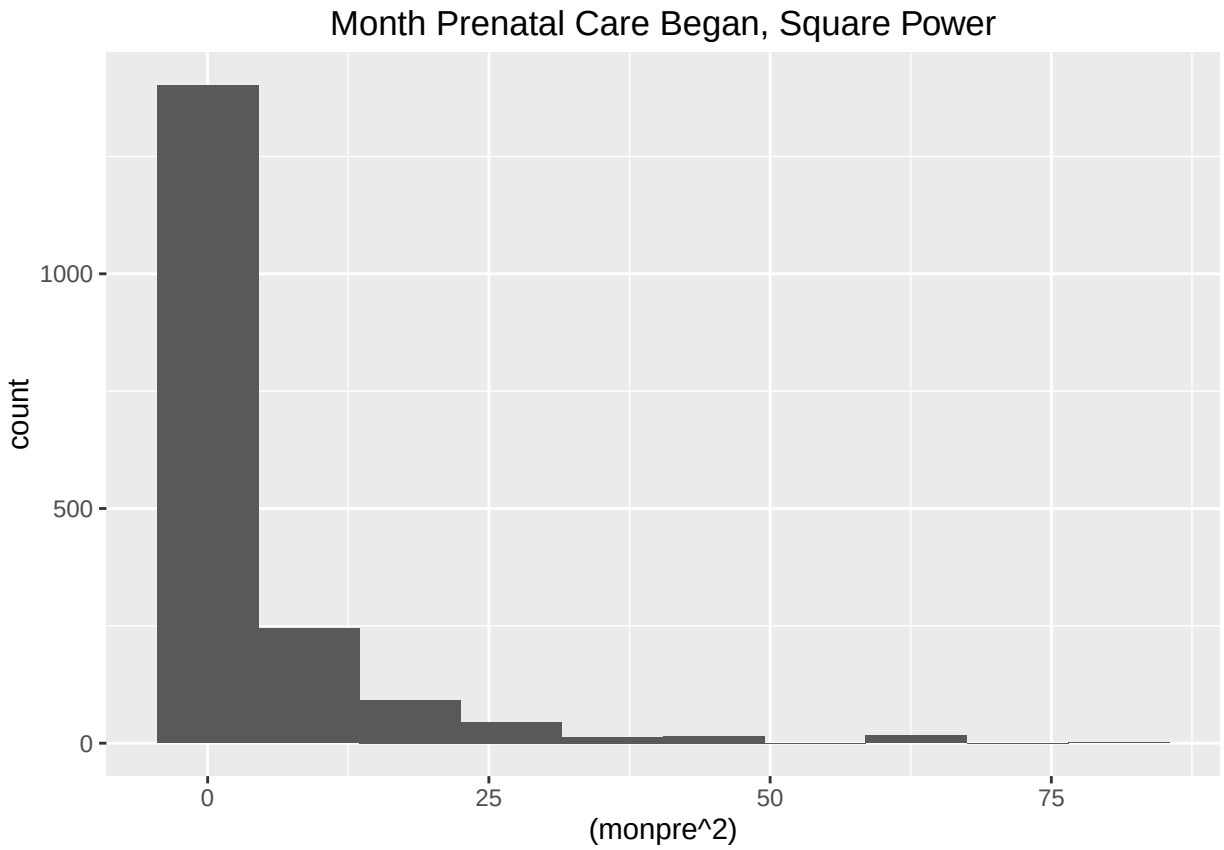

```
ggplot(data, aes(x=log(monpre))) + geom_histogram(aes(y = ..count..), bins = 10) +
  ggtitle("Month Prenatal Care Began, Natural Log")
```

```
## Warning: Removed 10 rows containing non-finite values (stat_bin).
```



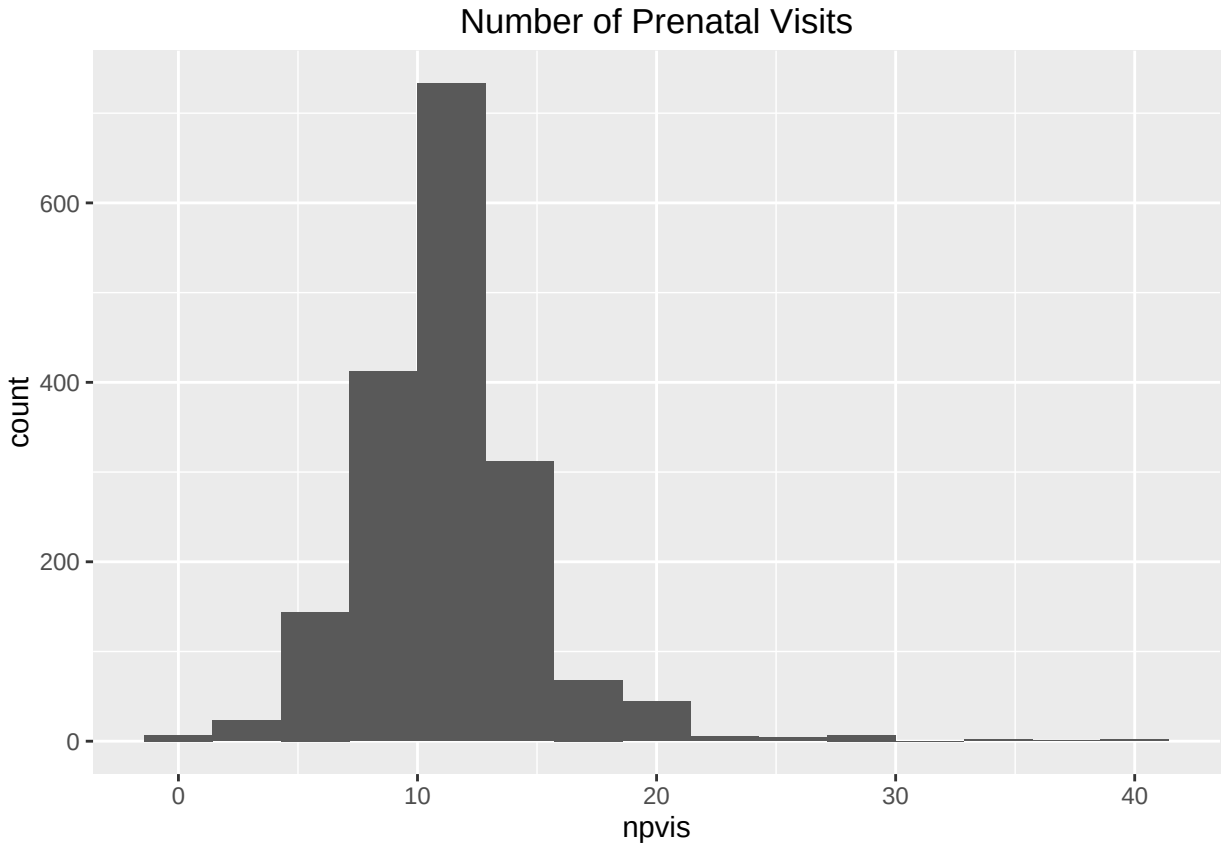

```
ggplot(data, aes(x=(monpre^2))) + geom_histogram(aes(y = ..count..), bins = 10) +
  ggtitle("Month Prenatal Care Began, Square Power")
```

```
## Warning: Removed 5 rows containing non-finite values (stat_bin).
```




```
ggplot(data, aes(x=npvis)) + geom_histogram(aes(y = ..count..), bins = 15) +
  ggtitle("Number of Prenatal Visits")
```

```
## Warning: Removed 68 rows containing non-finite values (stat_bin).
```



Look at the extreme fmops case

```
data[data$fmaps< 4,]
```

```
##      mage meduc monpre npvis fage feduc bwght omaps fmaps cigs drink lbw
## NA     NA    NA     NA    NA   NA    NA    NA    NA    NA   NA    NA  NA
## 837    32    12      2    10   40    16  2580     2     2    0     0   0
## NA.1   NA    NA     NA    NA   NA    NA    NA    NA    NA   NA    NA  NA
## NA.2   NA    NA     NA    NA   NA    NA    NA    NA    NA   NA    NA  NA
##      vlbw male mwhte mblck moth fwhte fblck foth   lbwght magesq npvissq
## NA     NA   NA    NA    NA   NA    NA    NA   NA       NA     NA      NA
## 837     0    1     1     0    0     1     0    0 7.855545   1024     100
## NA.1   NA   NA    NA    NA   NA    NA    NA   NA       NA     NA      NA
## NA.2   NA   NA    NA    NA   NA    NA    NA   NA       NA     NA      NA
```

Step 3: Modeling

Biases and Limitation

This data is extremely biased in that no still births were included in our dataset. It is a sad fact in the United States that over 2 in 1,000 births are stillbirths. Since we do not know the prenatal care data for stillbirths, we cannot completely guage how much prenatal care contributes to a child's health at birth.