Lab 4: Does Prenatal Care Improve Infant Health?

w203: Statistics for Data Science

November 23, 2016

Introduction

This is a group lab. You may work in teams of 2 or 3.

The file bwght_w203.RData contains data from the National Center for Health Statistics and from birth certificates. Your team has been hired by a health advocacy group to study this data and help them understand whether prenatal care improves health outcomes for newborn infants.

The file includes a birthweight variable. Additonally, the one- and five-minute APGAR scores are included. These are measures of the well being of infants just after birth.

Variable descriptions are provided as follows.

```
#Load Libraries
library(gridExtra)
library(ggplot2)
library(stargazer)
```

```
##
## Please cite as:

##  Hlavac, Marek (2015). stargazer: Well-Formatted Regression and Summary Statistics Tables.

##  R package version 5.2. http://CRAN.R-project.org/package=stargazer
```

```
library(lmtest)
```

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(sandwich)
library(car)
library(corrplot)
```

```
# setwd('C:/Users/mullar/Google Drive/Berkeley/W203/Lab_4/')
# setwd('C:/Users/Rob/Google Drive/Berkeley/W203/Lab_4/')
setwd('/Users/robmulla/Google Drive/Berkeley/W203/Lab_4')
getwd()
```

```
## [1] "/Users/robmulla/Google Drive/Berkeley/W203/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
```

Assignment

Prepare a report addressing the question of whether prenatal care improves newborn health outcomes.

A successful submission will include

1. A brief introduction
2. A model building process, supported by exploratory analysis. Your EDA should be interspersed with, and support, your modeling decisions. In particular, you should use exploratory techniques to address

Exploritory Analysis

Summary

```
summary(data)
```

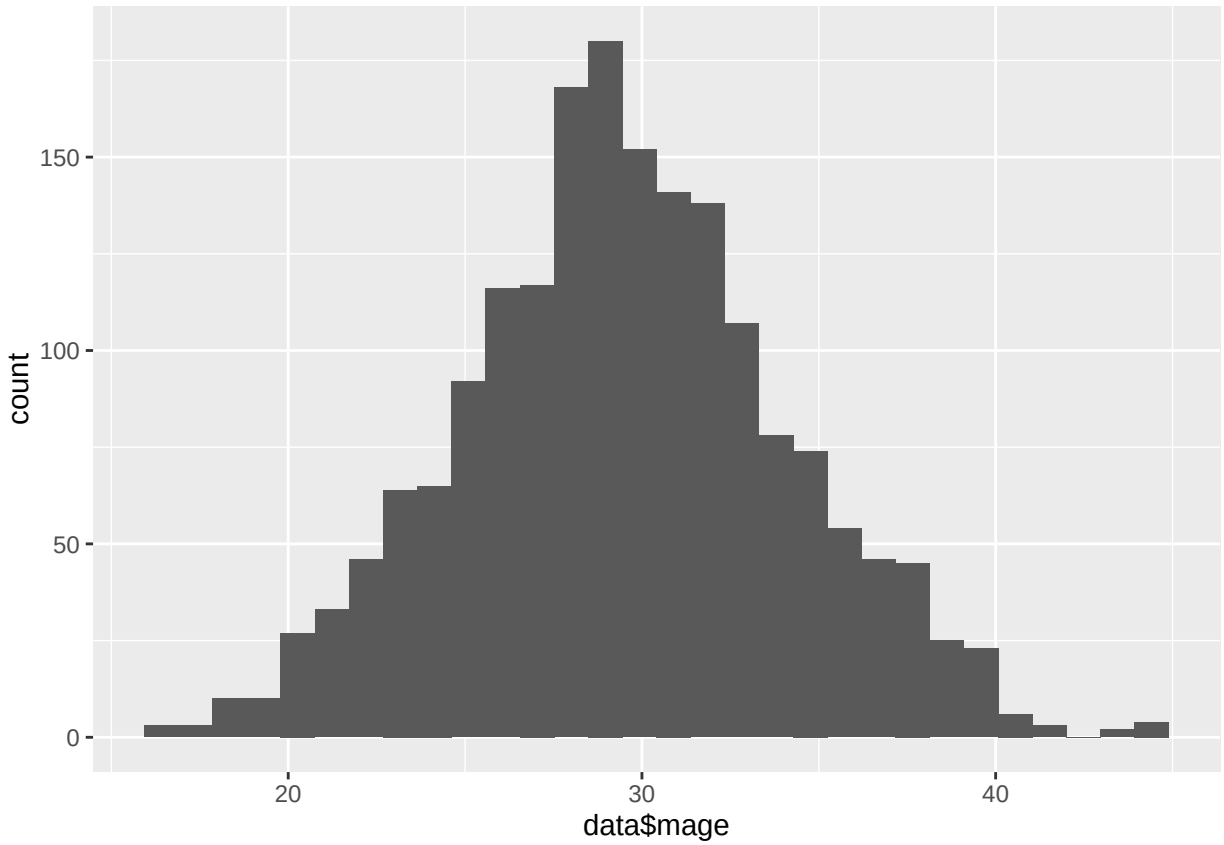
```
##       mage           meduc           monpre          npvis
##  Min.   :16.00   Min.   : 3.00   Min.   :0.000   Min.   : 0.00
##  1st Qu.:26.00   1st Qu.:12.00   1st Qu.:1.000   1st Qu.:10.00
##  Median :29.00   Median :13.00   Median :2.000   Median :12.00
```

```
##  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
```

Histograms

```
qplot(data$mage)
```

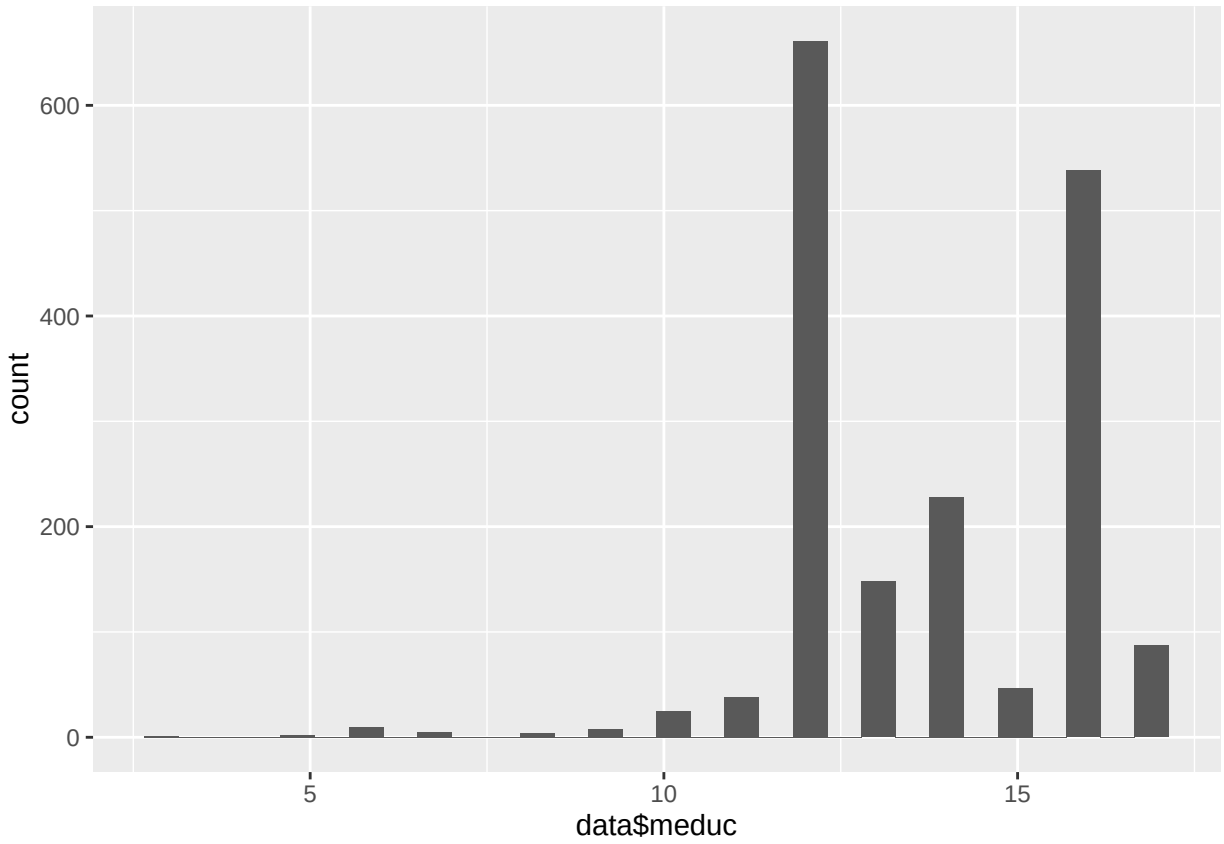
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
qplot(data$meduc)
```

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

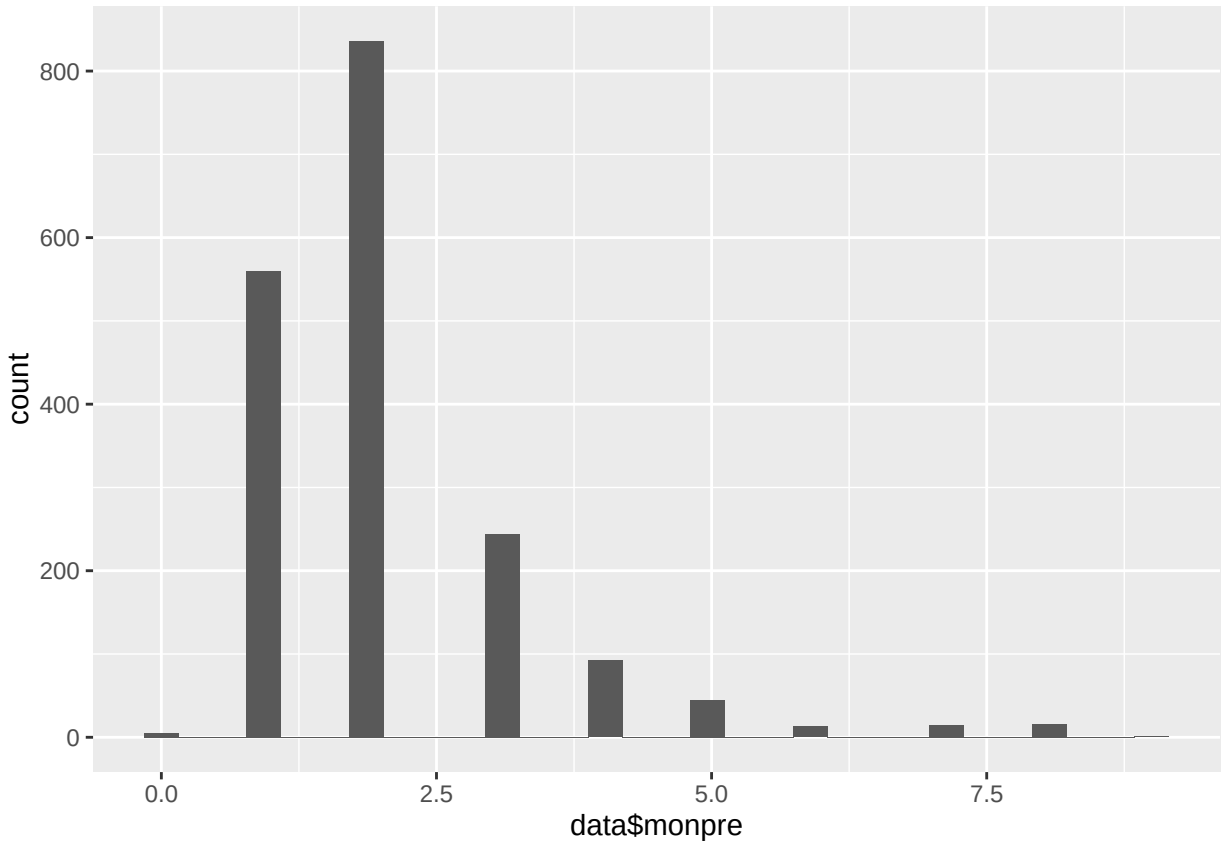
```
## Warning: Removed 30 rows containing non-finite values (stat_bin).
```




```
qplot(data$monpre)
```

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

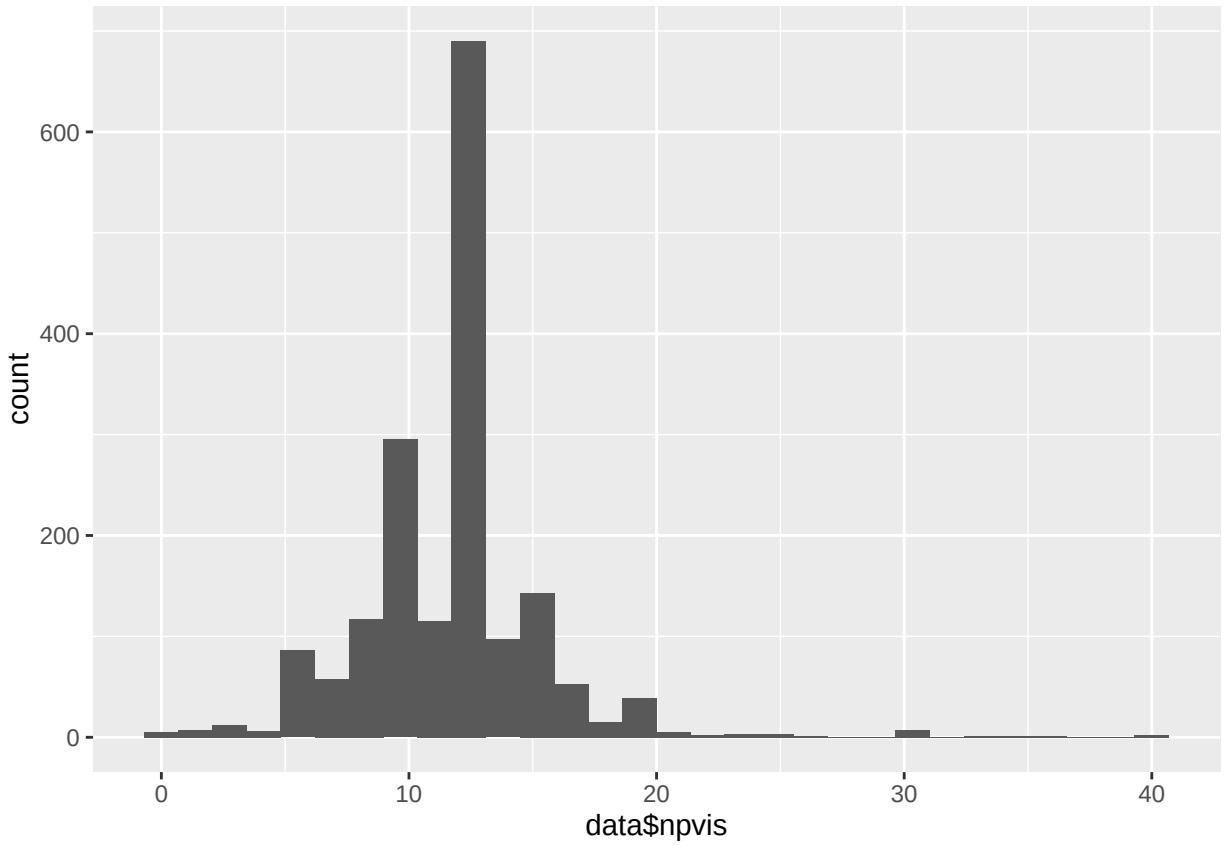
```
## Warning: Removed 5 rows containing non-finite values (stat_bin).
```




```
qplot(data$npvis)
```

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

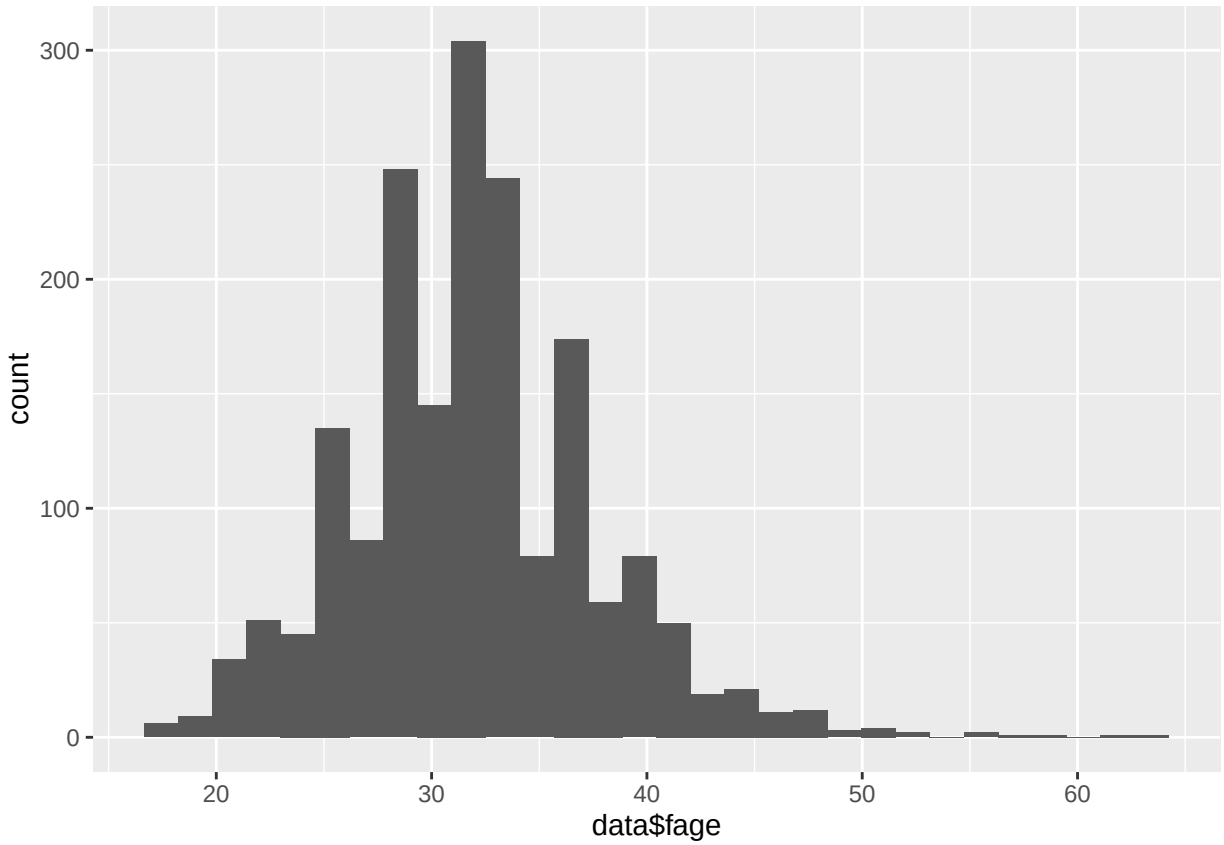
```
## Warning: Removed 68 rows containing non-finite values (stat_bin).
```




```
qplot(data$fage)
```

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

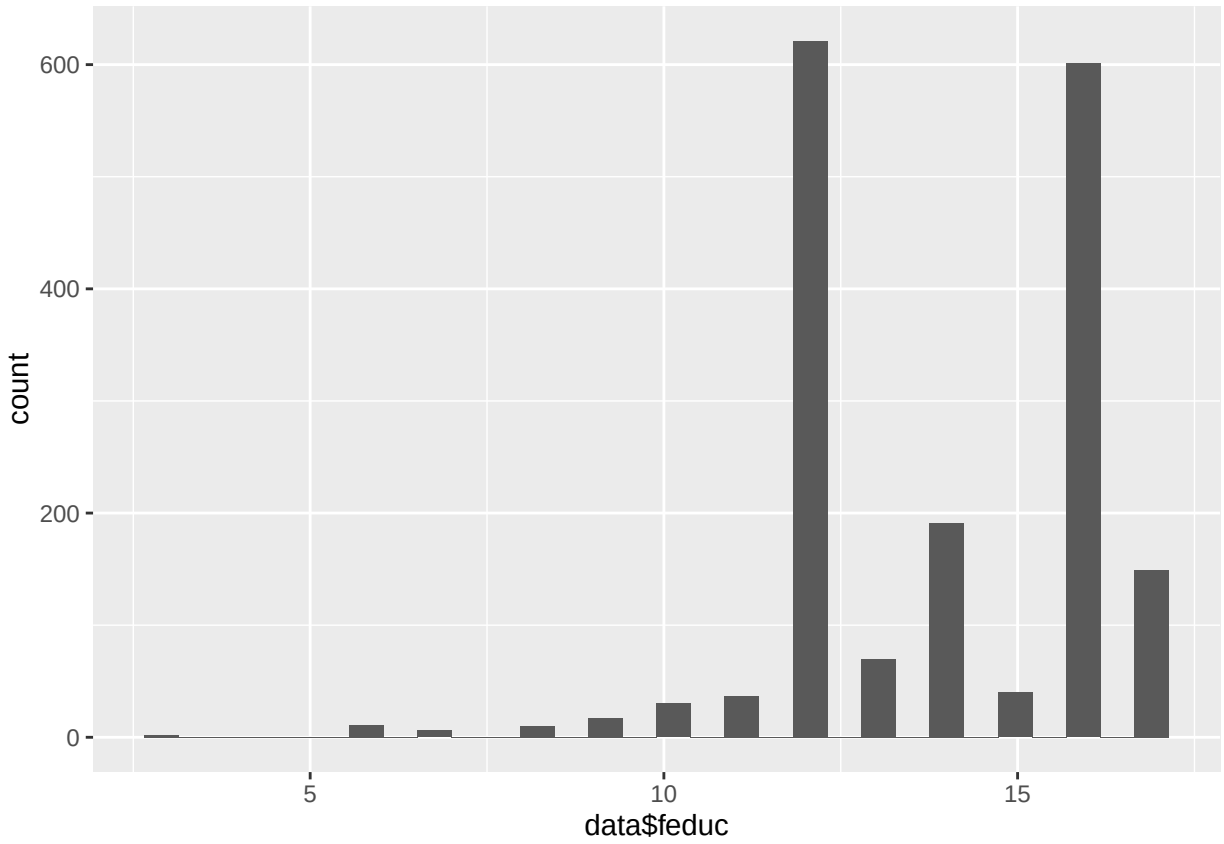
```
## Warning: Removed 6 rows containing non-finite values (stat_bin).
```




```
qplot(data$feduc)
```

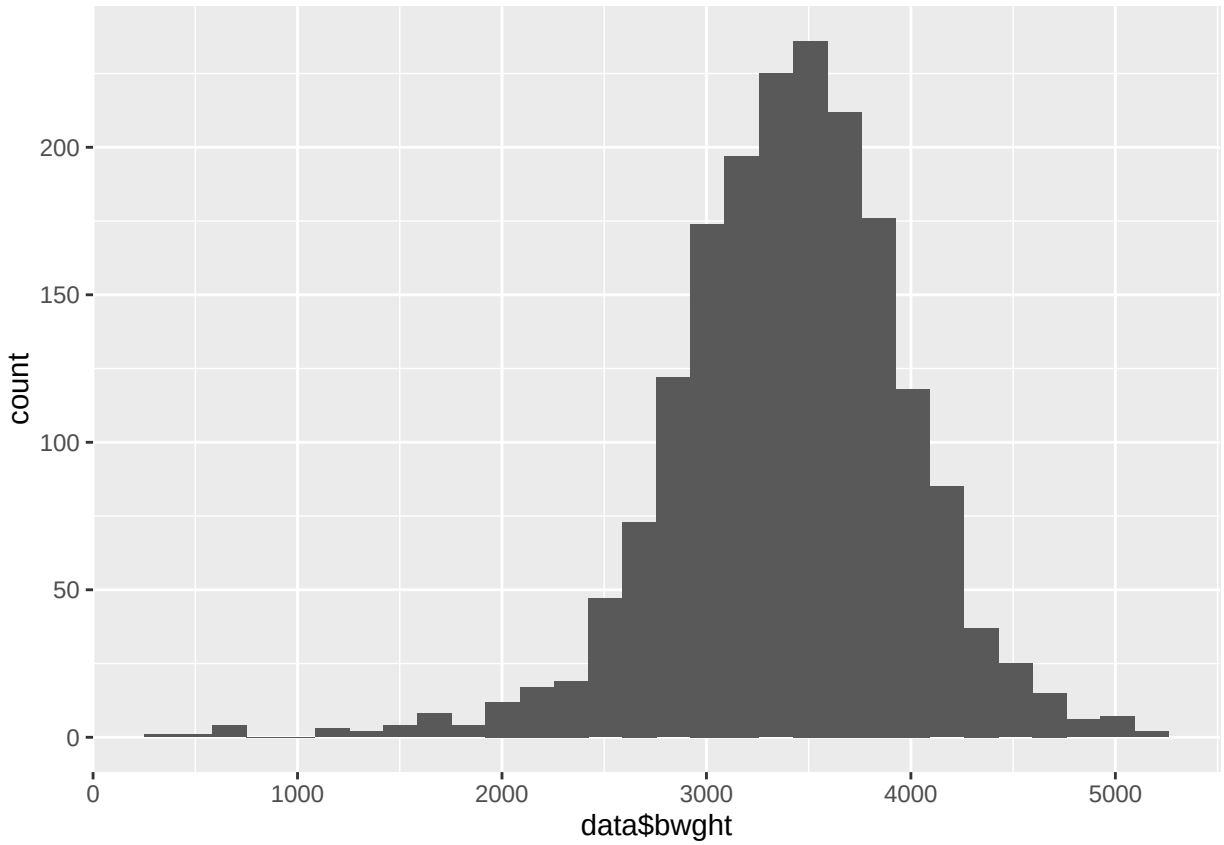
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 47 rows containing non-finite values (stat_bin).
```




```
qplot(data$bwght)
```

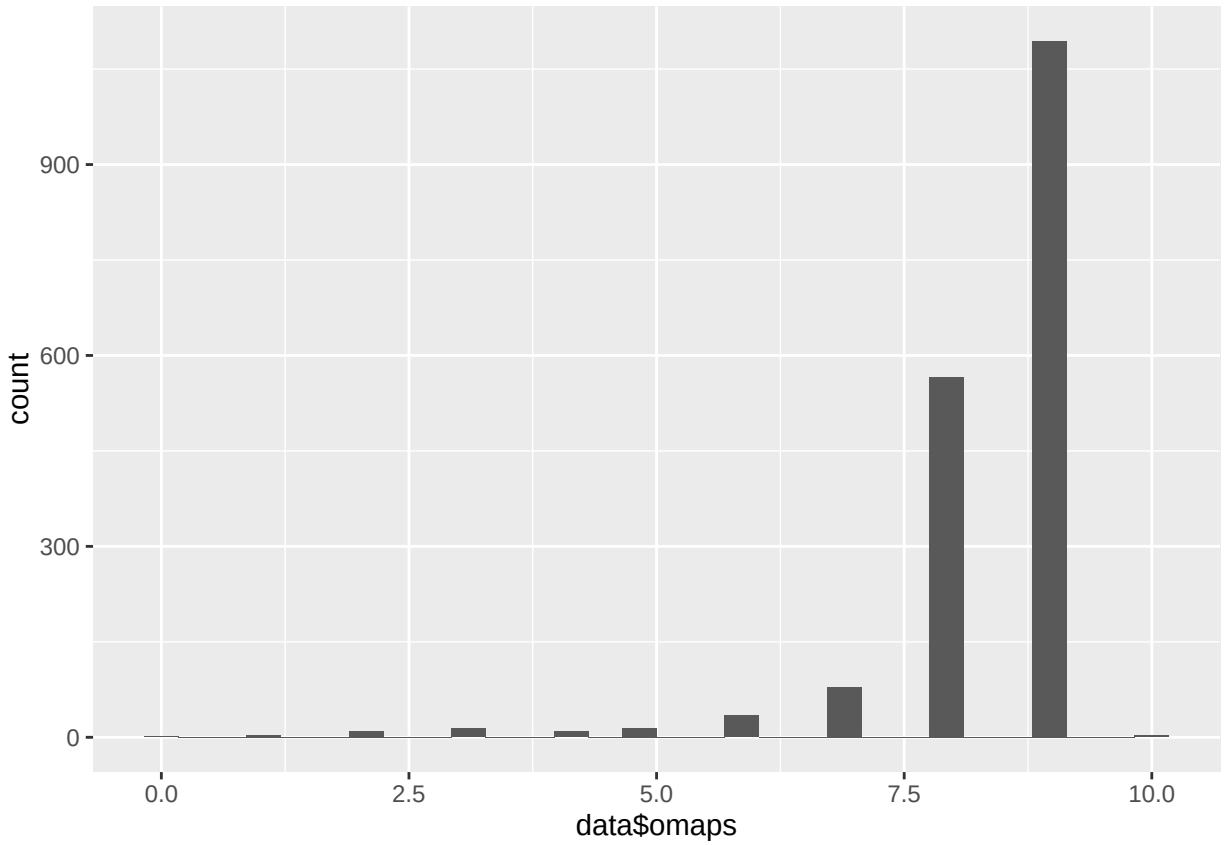
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
qplot(data$omaps)
```

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

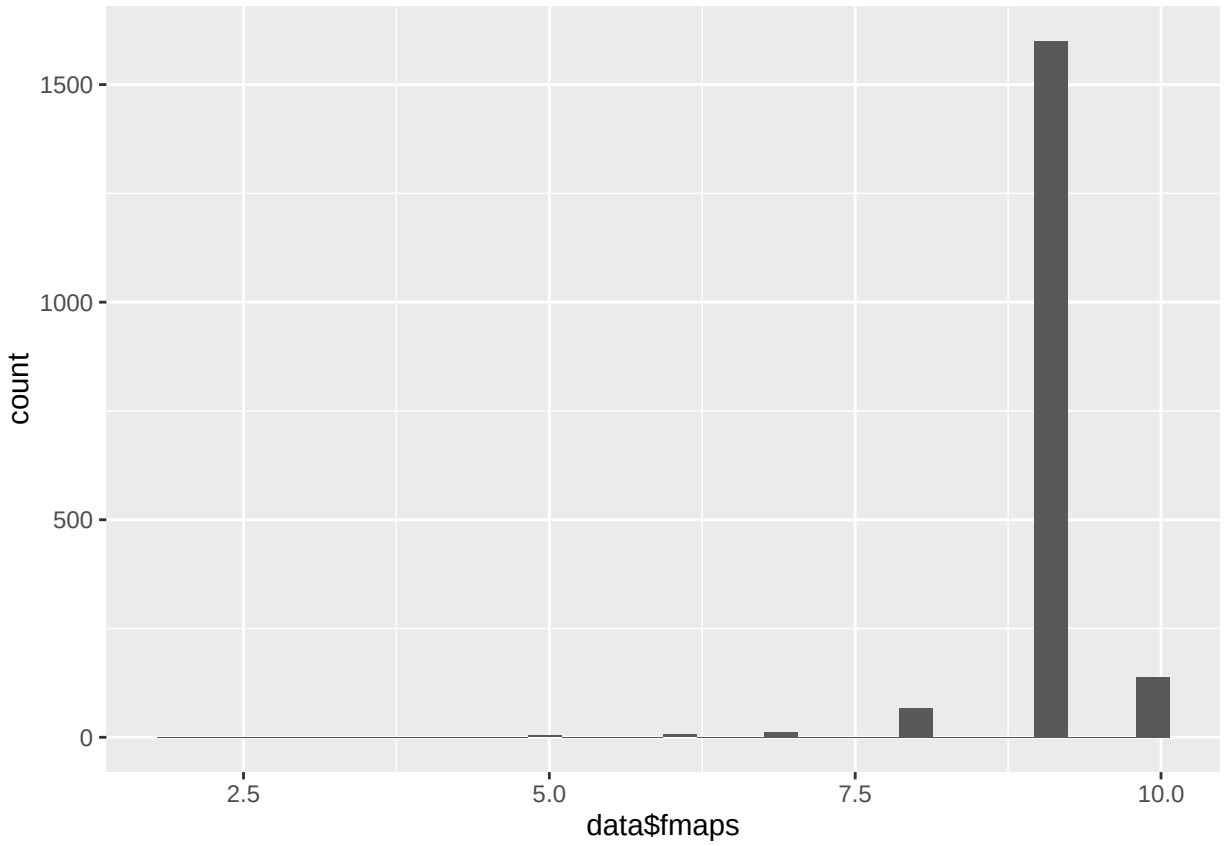
```
## Warning: Removed 3 rows containing non-finite values (stat_bin).
```




```
qplot(data$fmaps)
```

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

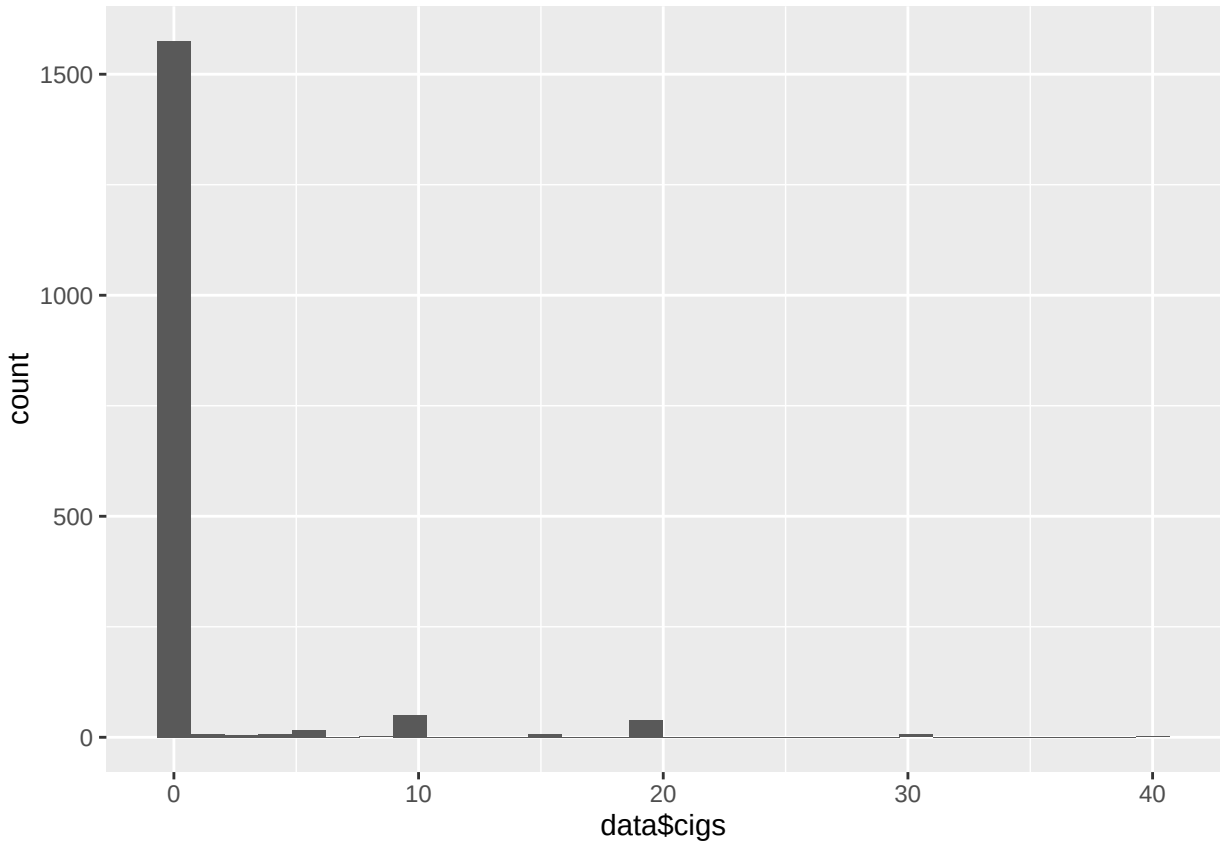
```
## Warning: Removed 3 rows containing non-finite values (stat_bin).
```




```
qplot(data$cigs)
```

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

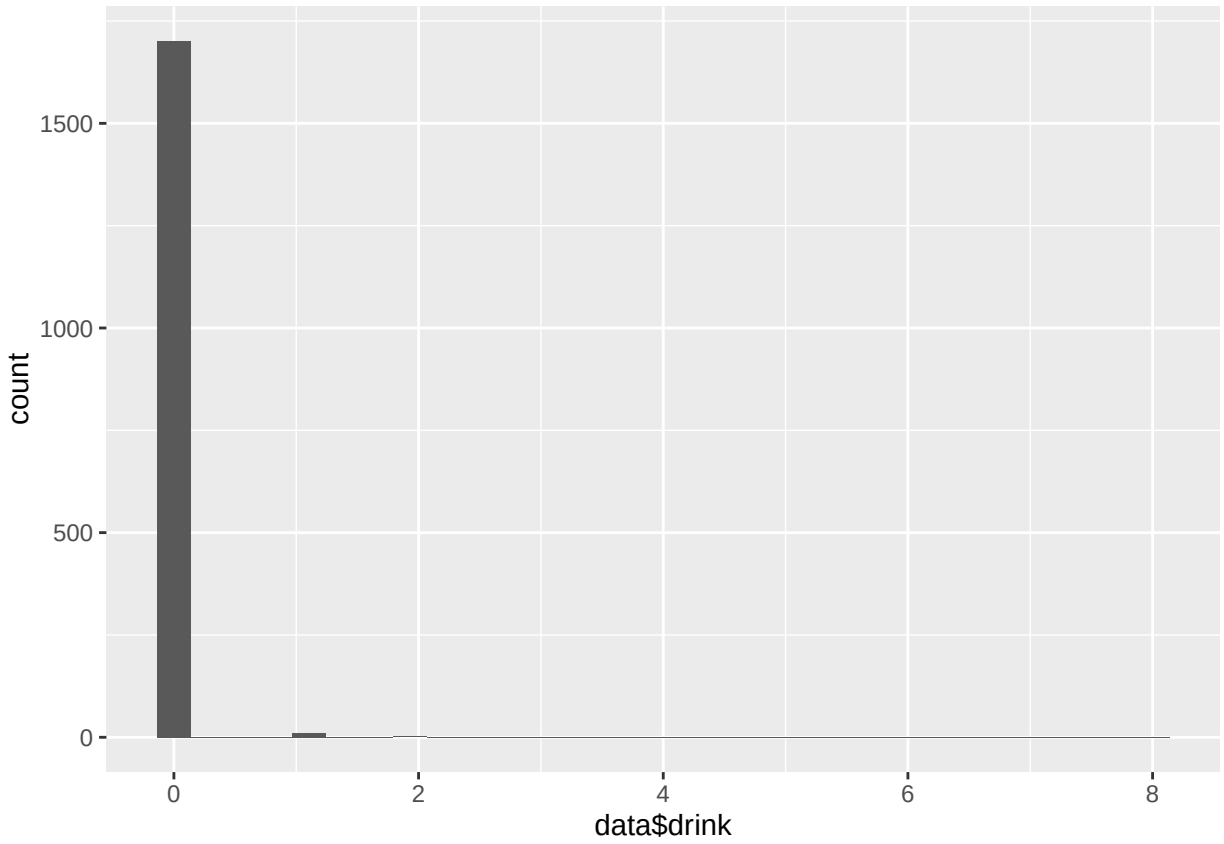
```
## Warning: Removed 110 rows containing non-finite values (stat_bin).
```




```
qplot(data$drink)
```

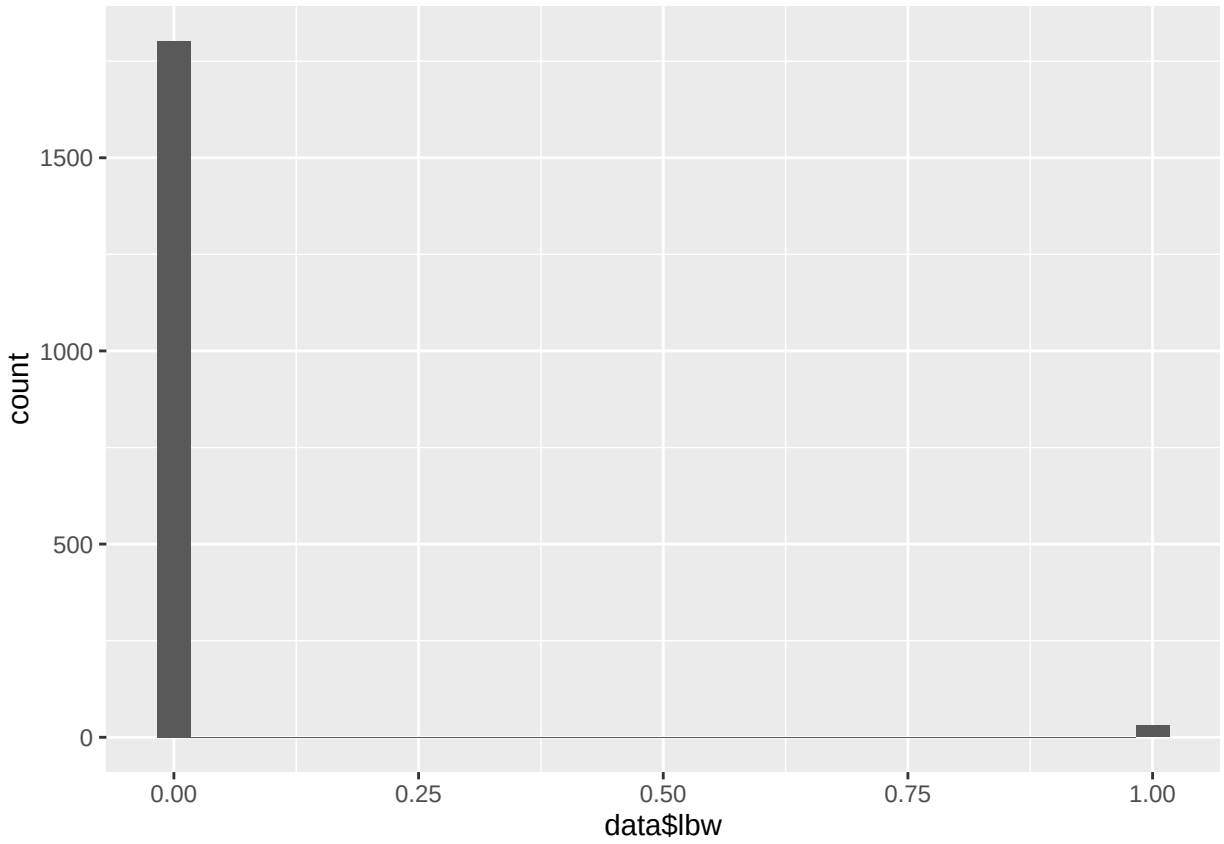
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 115 rows containing non-finite values (stat_bin).
```



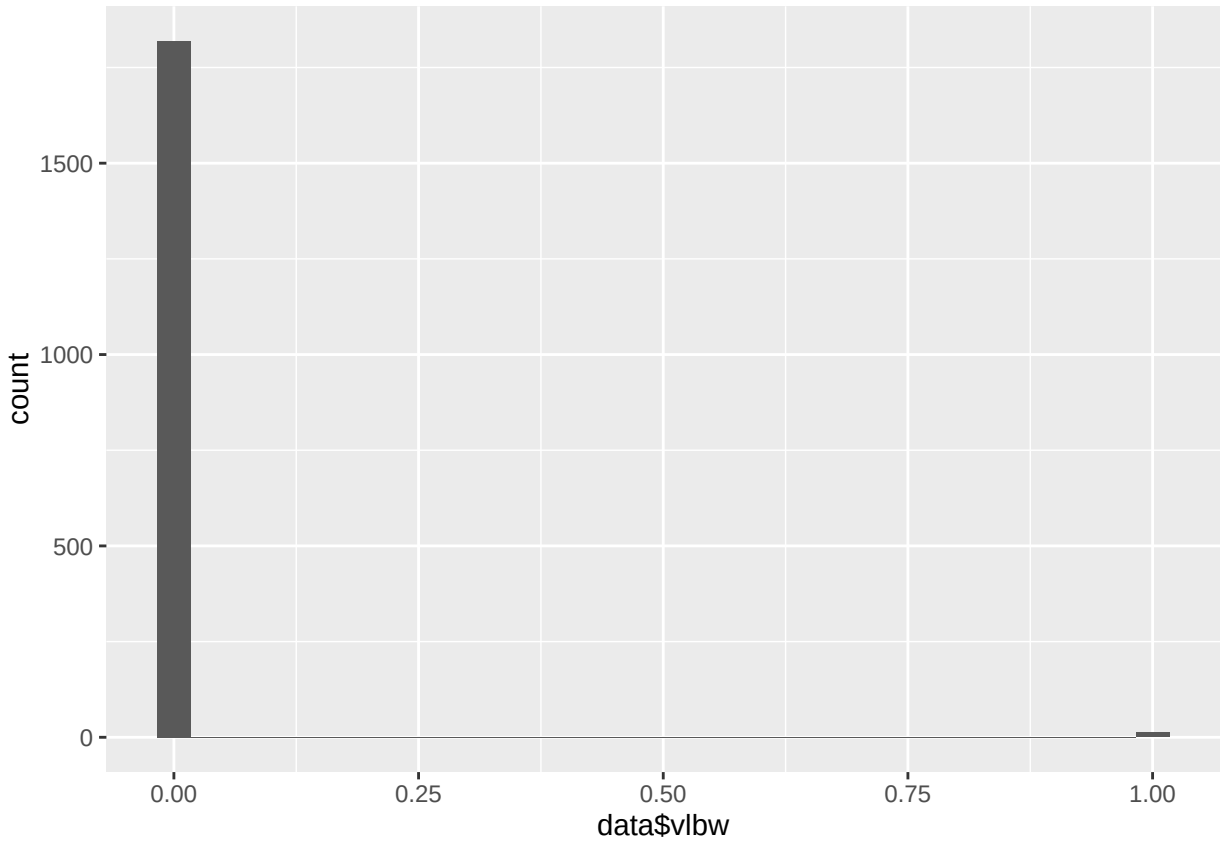

```
qplot(data$lbw)
```

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




```
qplot(data$vlbw)
```

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




```
colMeans(data)
```

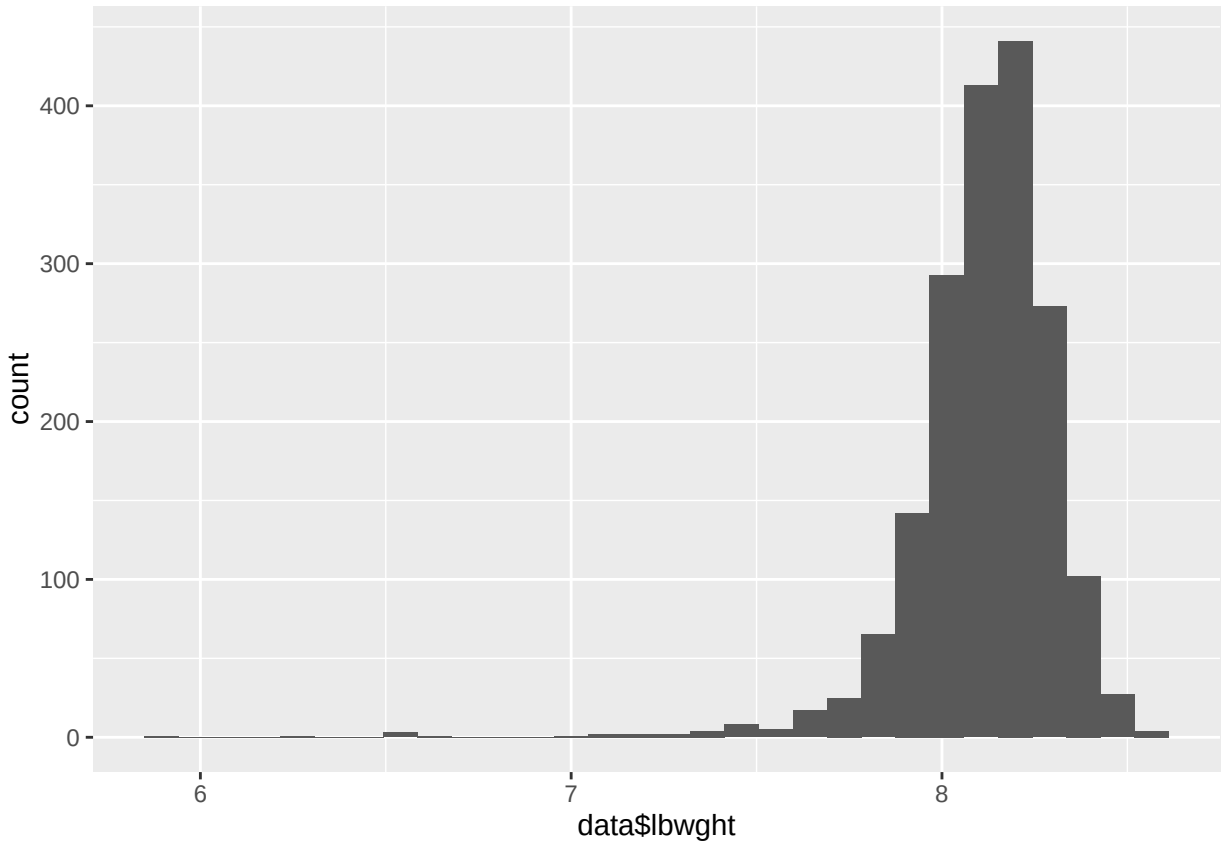
```
##         mage        meduc       monpre        npvis         fage
## 2.955786e+01           NA           NA           NA           NA
##        feduc        bwght        omaps        fmaps         cigs
##           NA 3.401122e+03           NA           NA           NA
##        drink          lbw         vlbw         male        mwhte
##           NA 1.637555e-02 7.096070e-03 5.136463e-01 8.864629e-01
##        mblck         moth        fwhte        fblck         foth
## 5.949782e-02 5.403930e-02 8.897380e-01 5.840611e-02 5.185590e-02
##       lbwght       magesq      npvissq
## 8.114247e+00 8.964170e+02           NA
```

```
mean(data$male)
```

```
## [1] 0.5136463
```

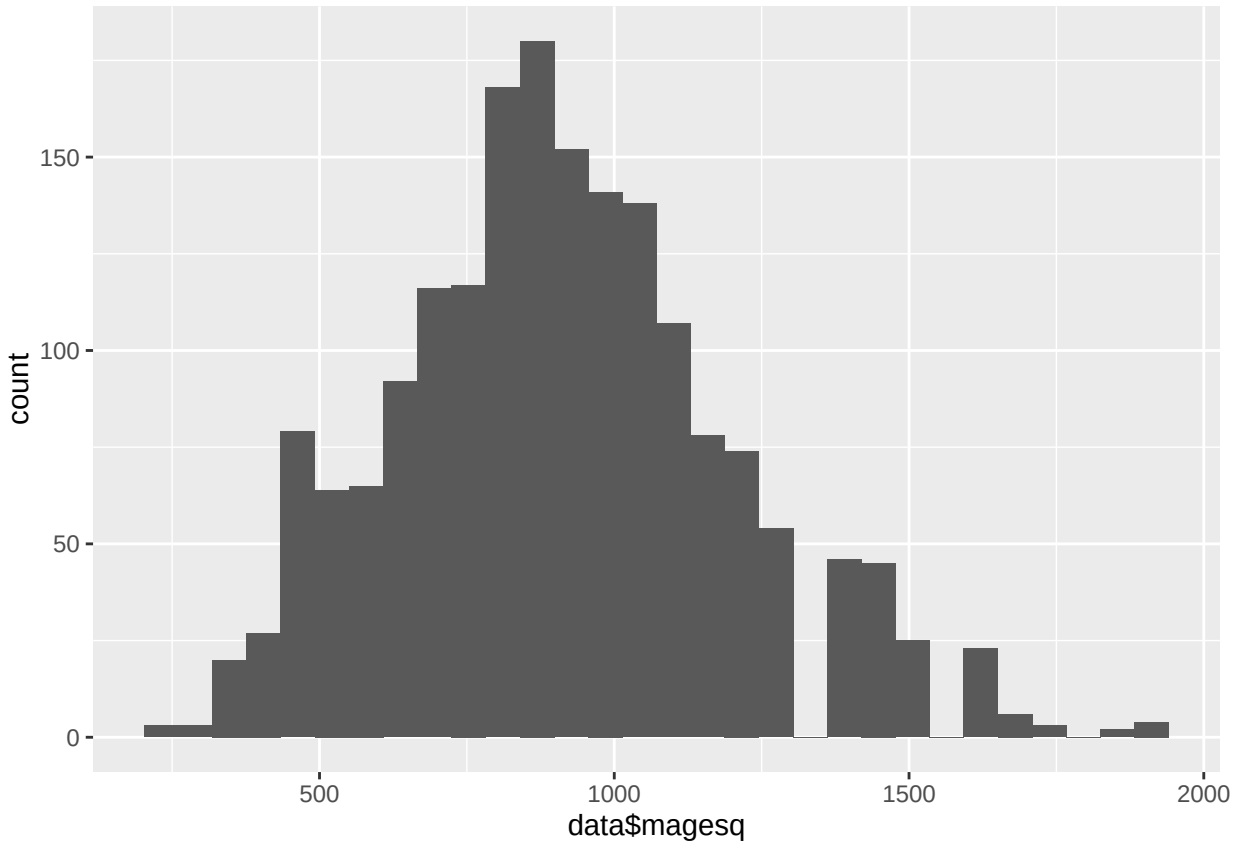
```
qplot(data$lbwght)
```

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




```
qplot(data$magesq)
```

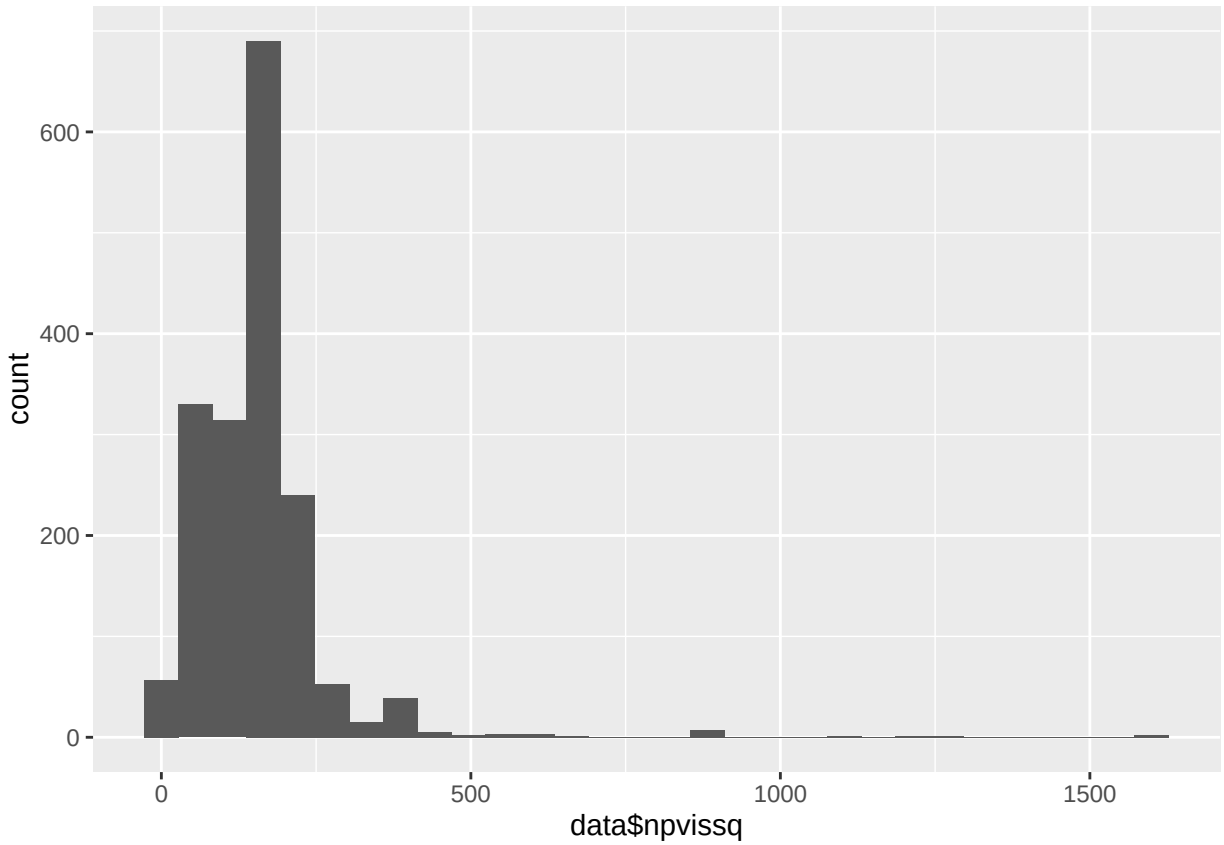
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
qplot(data$npvissq)
```

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

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



Comments on each variable: 1 mage - mother's age, years *Fairly normally distributed. don't think we would want to transform*

2 meduc - mother's educ, years *don't think we would want to transform, possibly do the thing where we bin into hs, college, etc. Spikes at 12 and 16 years as expected*

3 monpre month prenatal care began - *obvious max and min values (0 and ~10 months). Not normally distributed, tail out to the right side. May need to consider bias considering this is survey data*

4 npvis total number of prenatal visits *Somewhat normal looking around 13 visits, however long tail out to the right with very high number of visits*

5 fage father's age, years - *Fairly normal distribution*

6 feduc father's educ, years - *similar to mothers age variable, looks like more college grads than mothers*

7 bwght birth weight, grams - *Very normal looking with the exception of very low values. Will need to explore these low values and how they will impact the model*

8 omaps one minute apgar score - *peak at 9 trailing off to the left. 1 score of 0, 3 scores of 10*

9 fmaps five minute apgar score - *Interesting. Many more high numbers 8-10. Nearly none under 5.*

10 cigs avg cigarettes per day - *We've seen this before in the lab or homework. Possible measurement error due to clustering around 10 and 20 cigs*

11 drink avg drinks per week - *Large number at zero and few a 1 and 2*

12 lbw =1 if bwght <= 2000 - *Can be used to filter out the very low birthweights*

13 vlbw =1 if bwght <= 1500 - *same as lbw but less restrictive*

14 male =1 if baby male - *51% male*

```
mean(data$male)
```

```
## [1] 0.5136463
```

15 mwhte =1 if mother white - *88.6% of mothers white*

```
mean(data$mwhte)
```

```
## [1] 0.8864629
```

16 mblck =1 if mother black *5.9% of mothers black*

```
mean(data$mblck)
```

```
## [1] 0.05949782
```

17 moth =1 if mother is other *5.4% of mothers other*

```
mean(data$moth)
```

```
## [1] 0.0540393
```

18 fwhte =1 if father white *88.9% of fathers white*

```
mean(data$fwhte)
```

```
## [1] 0.889738
```

19 fblck =1 if father black *5.8 of fathers black*

```
mean(data$fblck)
```

```
## [1] 0.05840611
```

20 foth =1 if father is other *5.1% of fathers other*

```
mean(data$foth)
```

```
## [1] 0.0518559
```

21 lbwght log(bwght) 22 magesq mage^2 23 npvissq npvis^2*

Number of NAs for variables

```r
apply(!is.na(data) , MARGIN= 2, mean )
```

```
##      mage     meduc    monpre     npvis      fage     feduc     bwght
## 1.0000000 0.9836245 0.9972707 0.9628821 0.9967249 0.9743450 1.0000000
##     omaps     fmaps      cigs     drink       lbw      vlbw      male
## 0.9983624 0.9983624 0.9399563 0.9372271 1.0000000 1.0000000 1.0000000
##     mwhte     mblck      moth     fwhte     fblck      foth    lbwght
## 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
##    magesq   npvissq
## 1.0000000 0.9628821
```

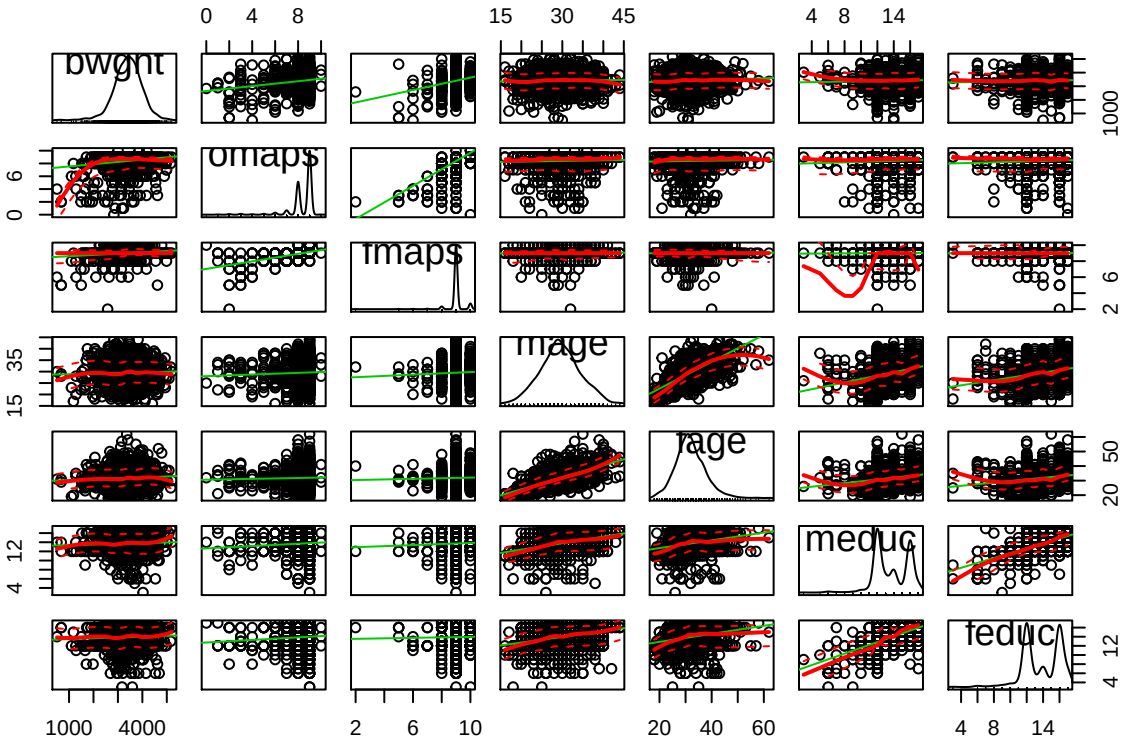
Note the variables with the most NA values (cigs, drink) may introduce bias into our model as the people who did not choose to respond may be to embarrassed to answer

Scatterplot Matrix

```r
scatterplotMatrix(~ bwght + omaps + fmaps + mage + fage + meduc + feduc, data=data)
```

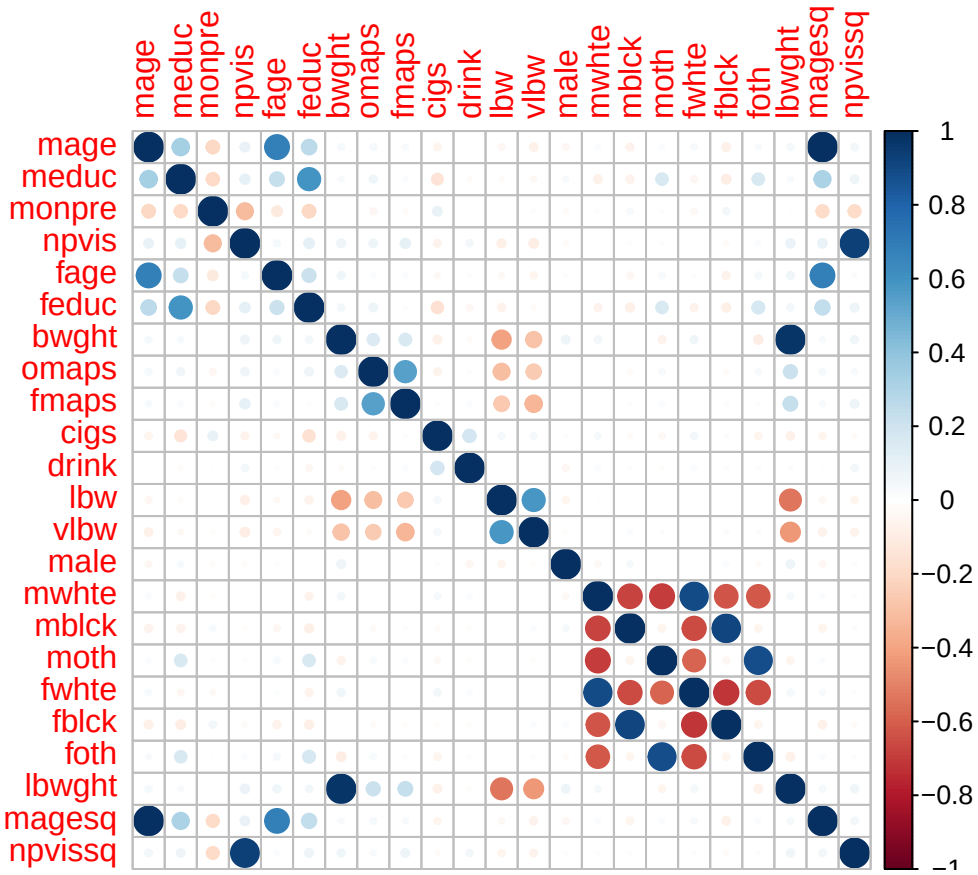
```
## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
```

```
## spread = spread, : could not fit smooth

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth
```



Corr Plot

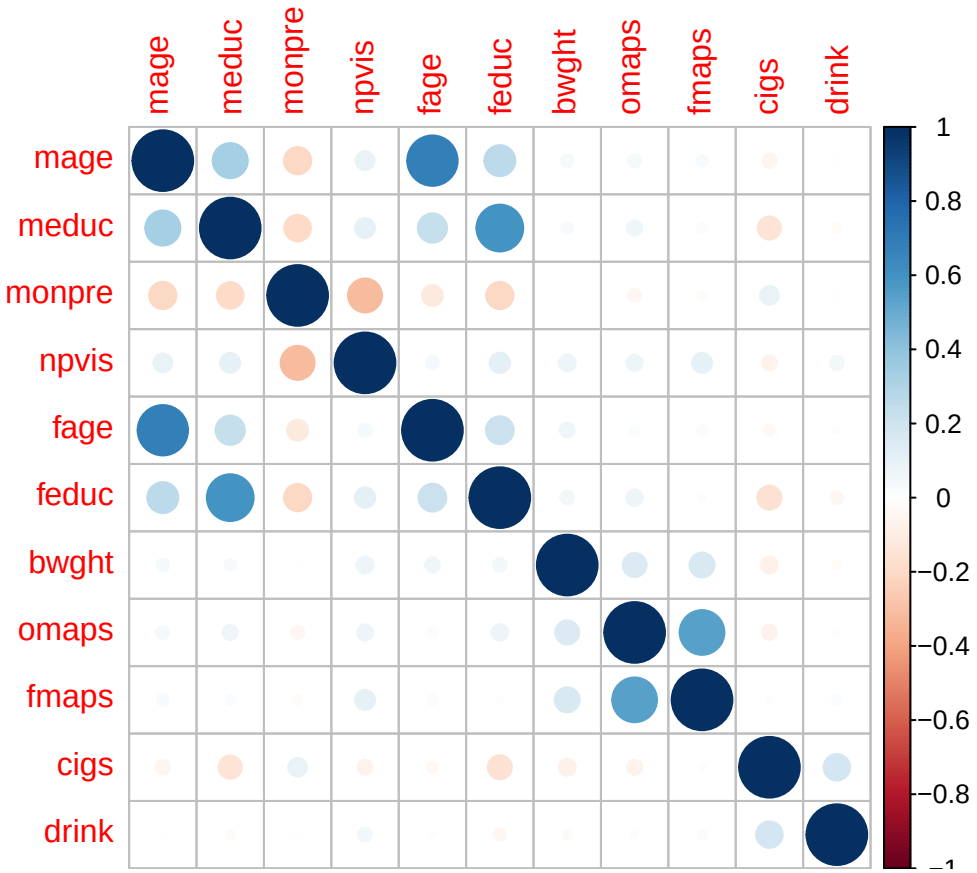
```
data_corr <- cor(data, use="complete.obs")
corrplot(data_corr)
```



variables highly correlated ages and education for mothers and correlated with the age and education of fathers race of mothers is correlated with race of father slight negative correlation between cigs and education of mother apgar scores correlated with eachother as expected

Corr Plot

```
data_corr2 <- cor(data[,(1:11)], use="complete.obs")
corrplot(data_corr2)
```

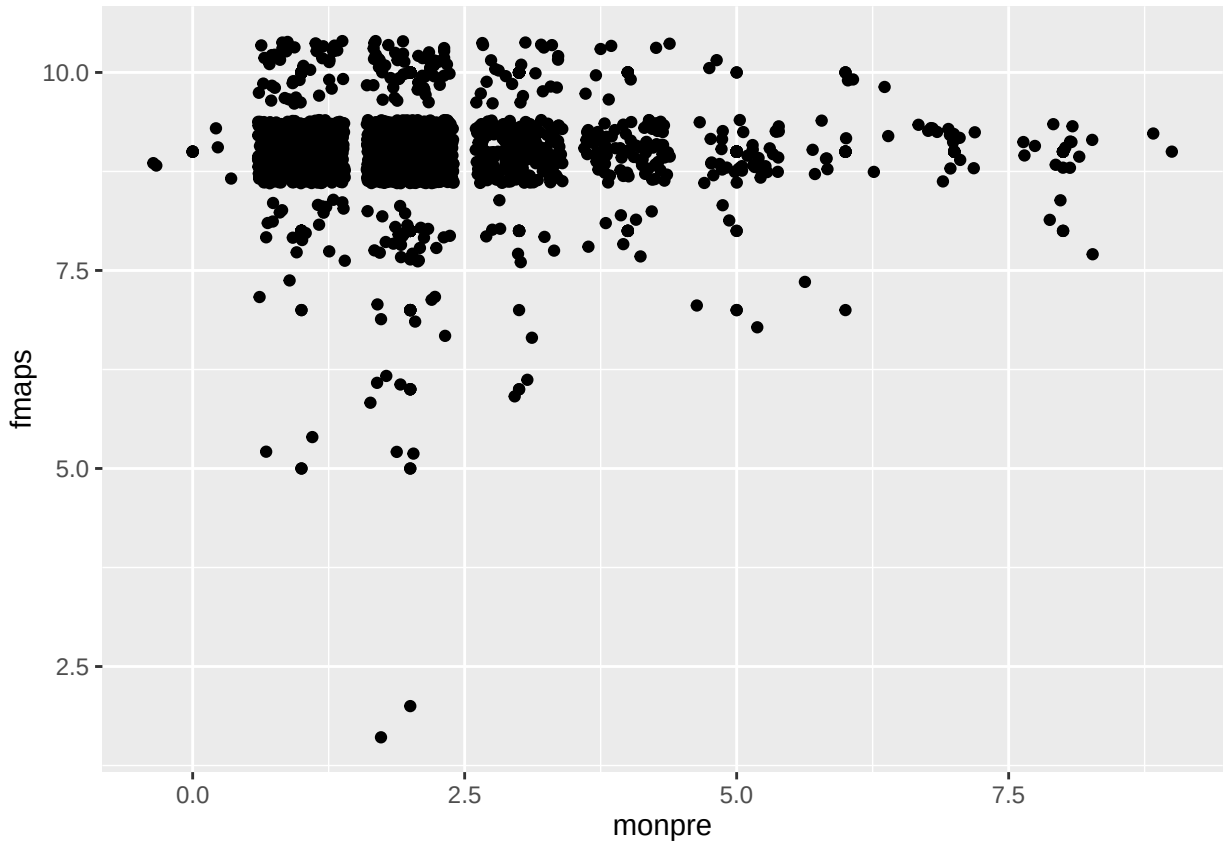


Prenadal care vs apgar

```
qplot(monpre, fmaps, data=data) + geom_jitter()
```

```
## Warning: Removed 8 rows containing missing values (geom_point).
```

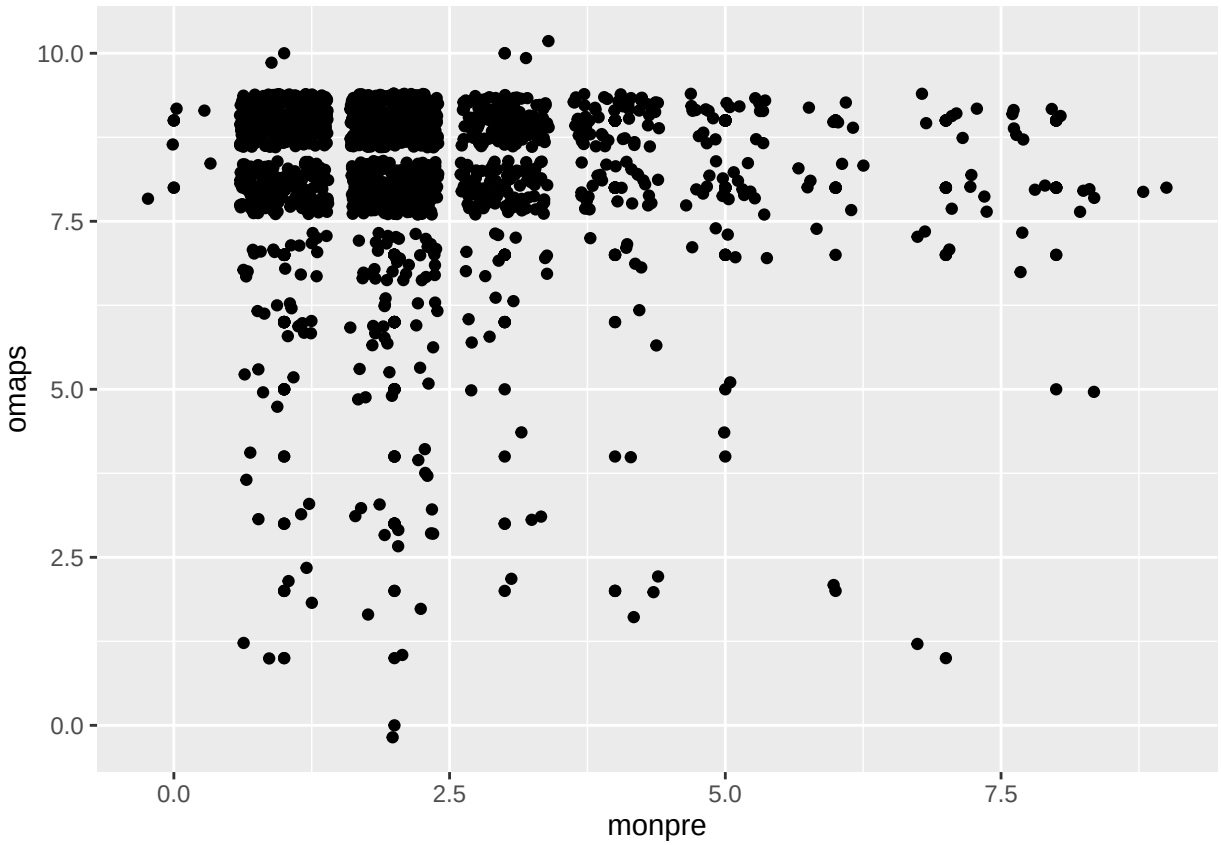
```
## Warning: Removed 8 rows containing missing values (geom_point).
```




```
qplot(monpre, omaps, data=data) + geom_jitter()
```

```
## Warning: Removed 8 rows containing missing values (geom_point).
```

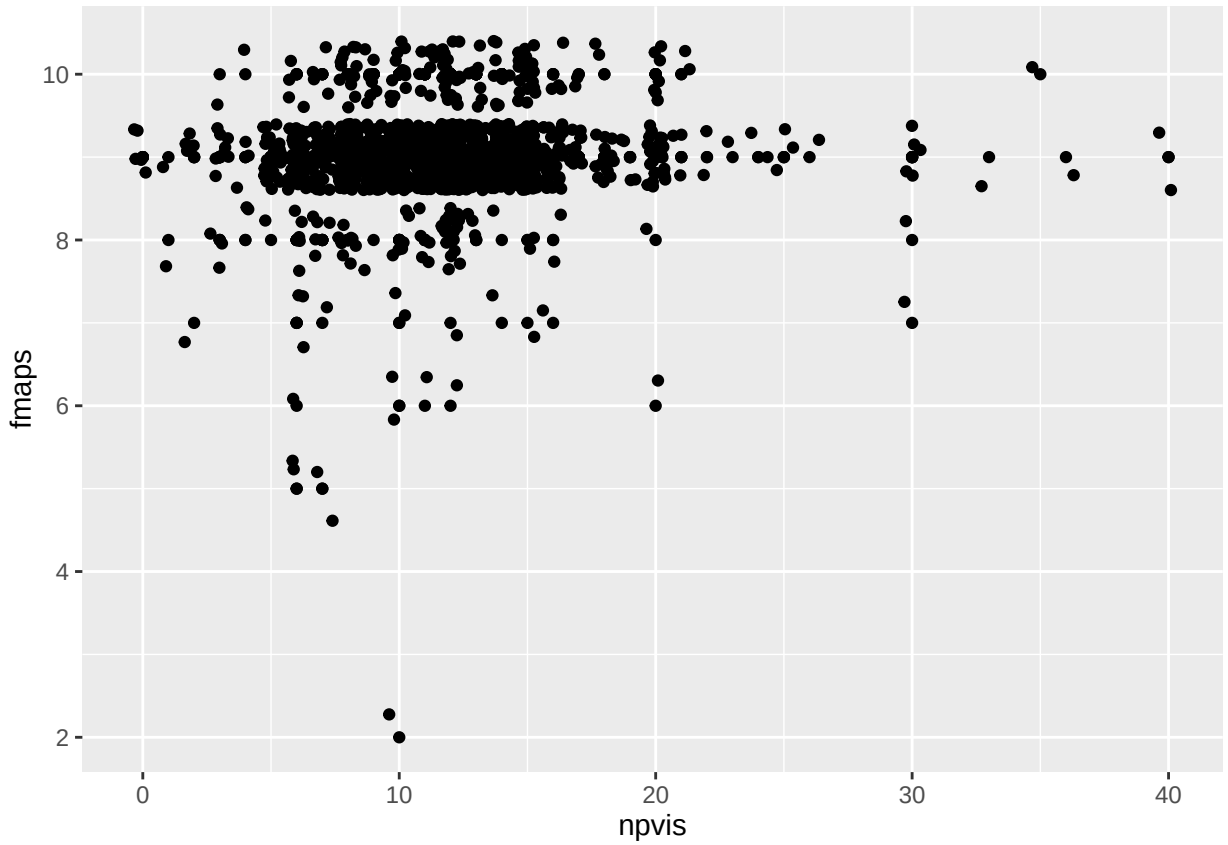
```
## Warning: Removed 8 rows containing missing values (geom_point).
```




```
qplot(npvis, fmaps, data=data) + geom_jitter()
```

```
## Warning: Removed 71 rows containing missing values (geom_point).
```

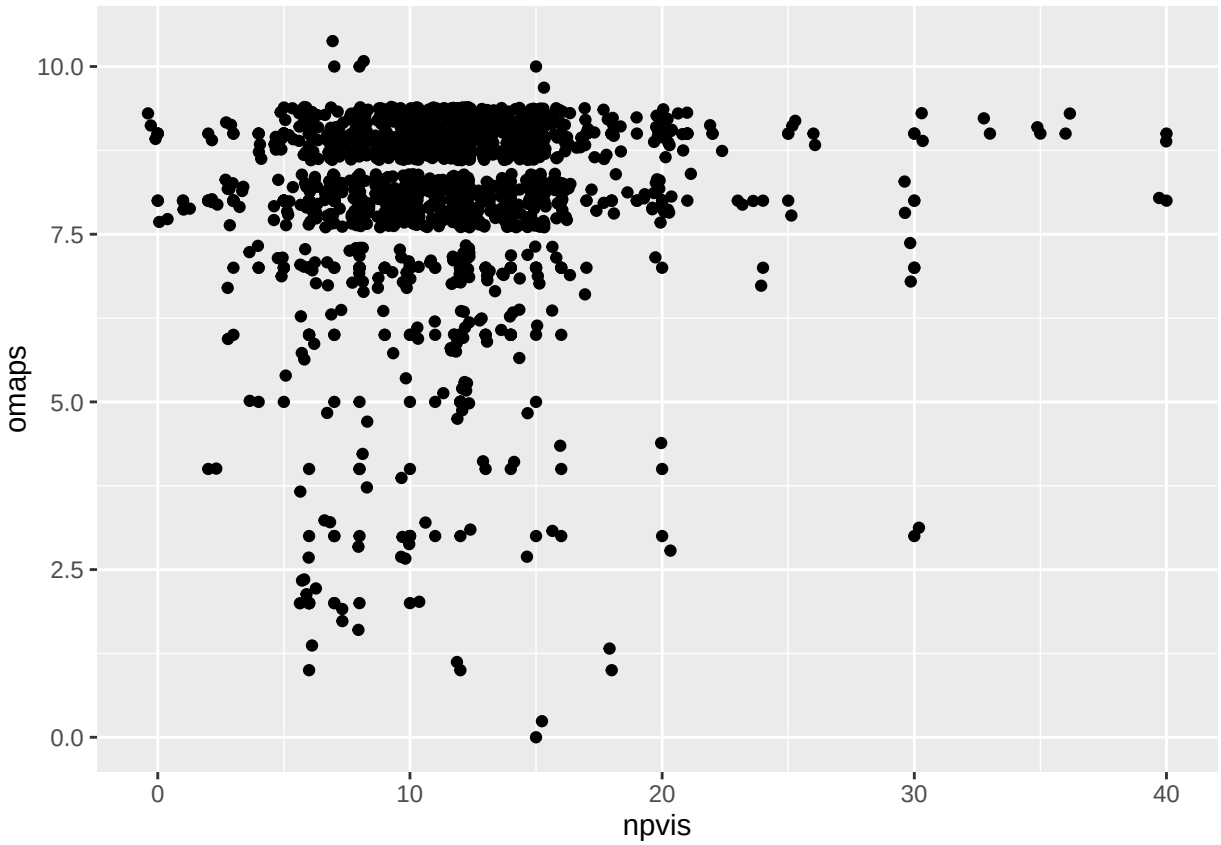
```
## Warning: Removed 71 rows containing missing values (geom_point).
```




```
qplot(npvis, omaps, data=data) + geom_jitter()
```

```
## Warning: Removed 71 rows containing missing values (geom_point).
```

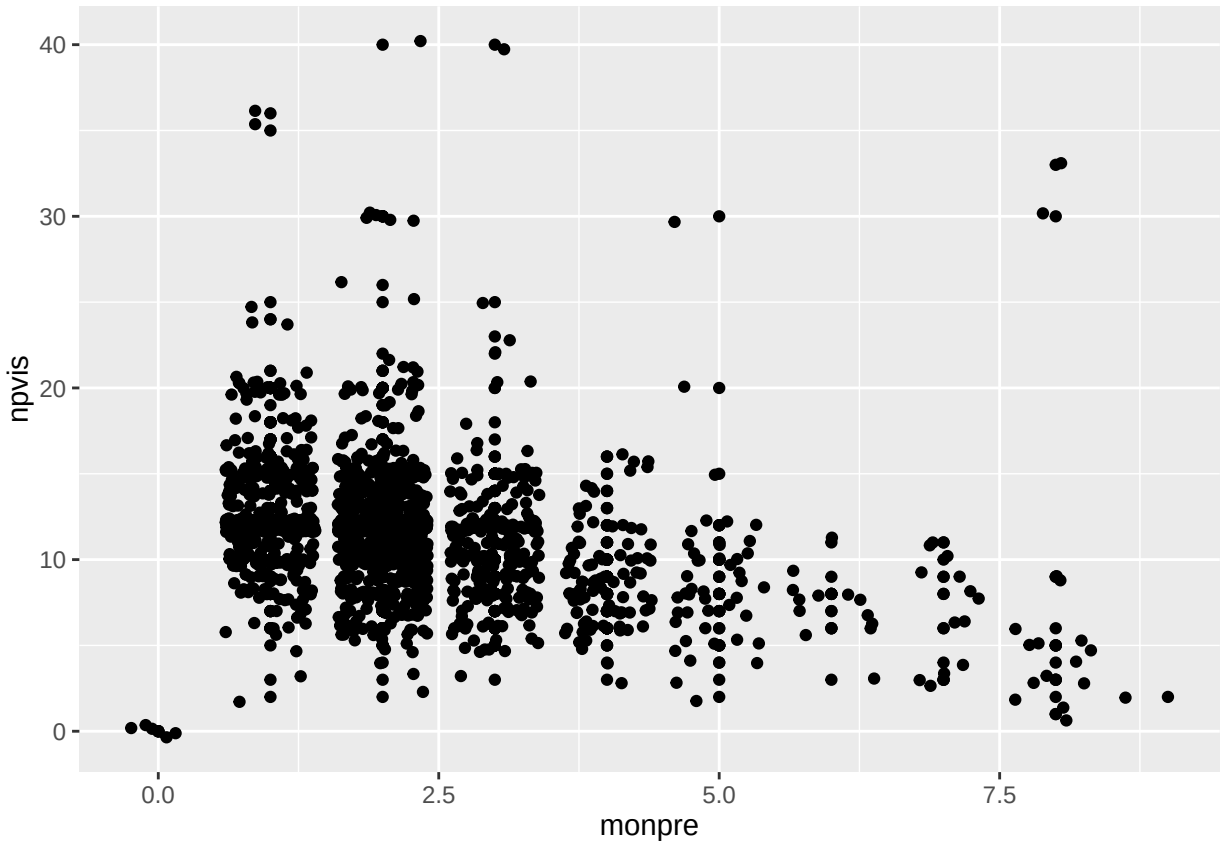
```
## Warning: Removed 71 rows containing missing values (geom_point).
```




```
qplot(monpre, npvis, data=data) + geom_jitter()
```

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

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



Some personal observations about these variables: If a mother has a lot of prenadal visits it could point to an unhealthy pregnancy. If a mother has zero prenadal visits it could point to an unhealthy lifestyle.

- What transformations to apply to variables and what new variables should be created.
- What variables should be included in each model
- Whether model assumptions are met

3. A minimum of three model specifications. In particular, you should include

- One model with only the explanatory variables of key interest.
- One model that includes only covariates that you believe increase the accuracy of your results without introducing bias.
- One model that includes the previous covariates, but also covariates that may be problematic for one reason or another.

4. For your first model, a detailed assessment of the 6 CLM assumptions. For additional models, you should check all assumptions, but only highlight major differences from your first model in your report.
5. A well-formatted regression table summarizing your model results. Make sure that standard errors presented in this table are valid. Also be sure to comment on both statistical and practical significance.
6. A discussion of whether your results can be interpretted causally. In particular, include a discussion of what variables are not included in your analysis and the likely direction of omitted variable bias. Also include a discussion of which included variables may bias your results by absorbing some of the causal effect of prenatal care.

7. A brief conclusion with a few high-level takeaways.

Please limit all submissions to 30 pages. Be sure to turn in both your pdf report and also your source code.