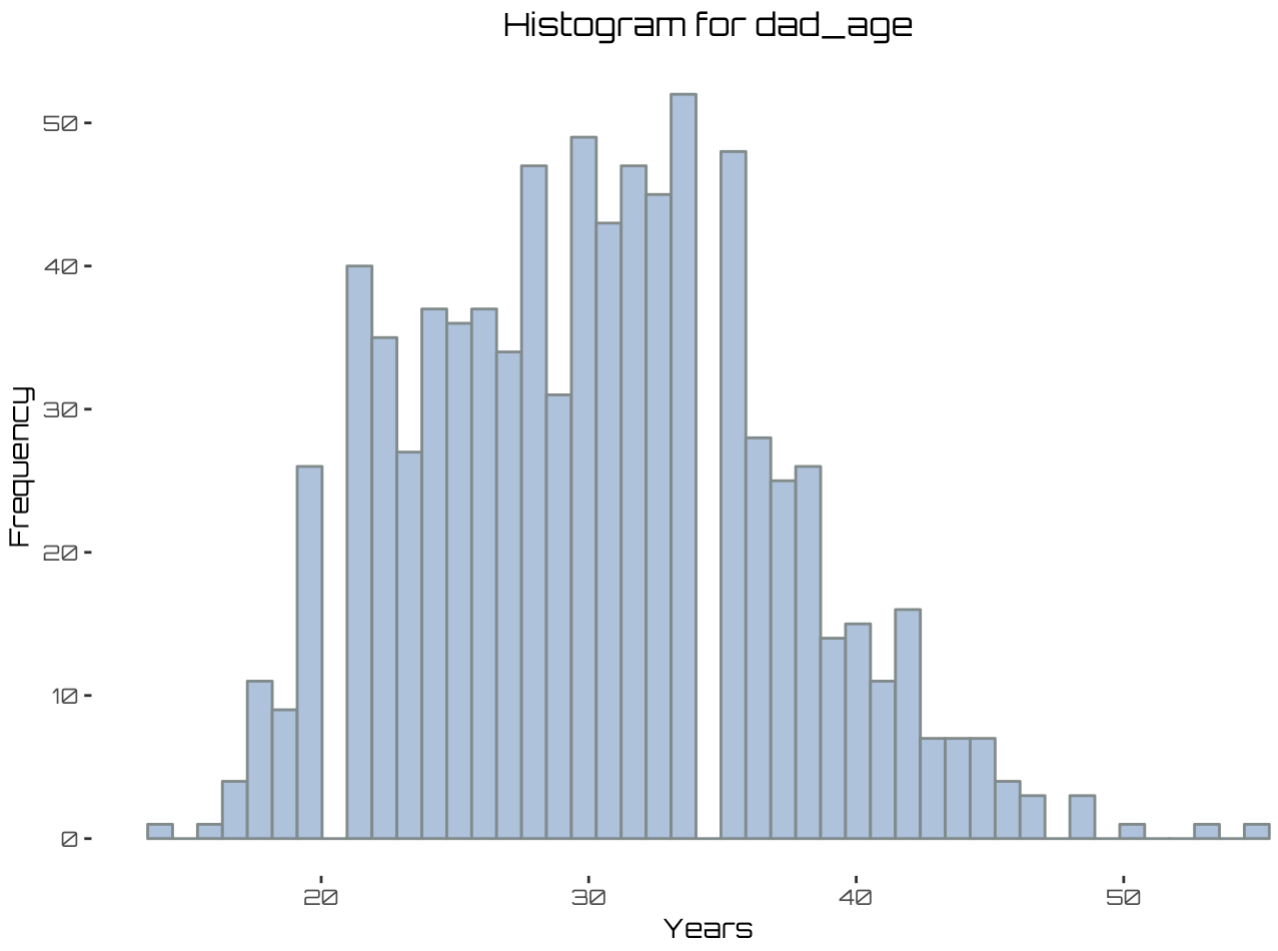
Data Analysis and Visualization Project:
Births Data Frame

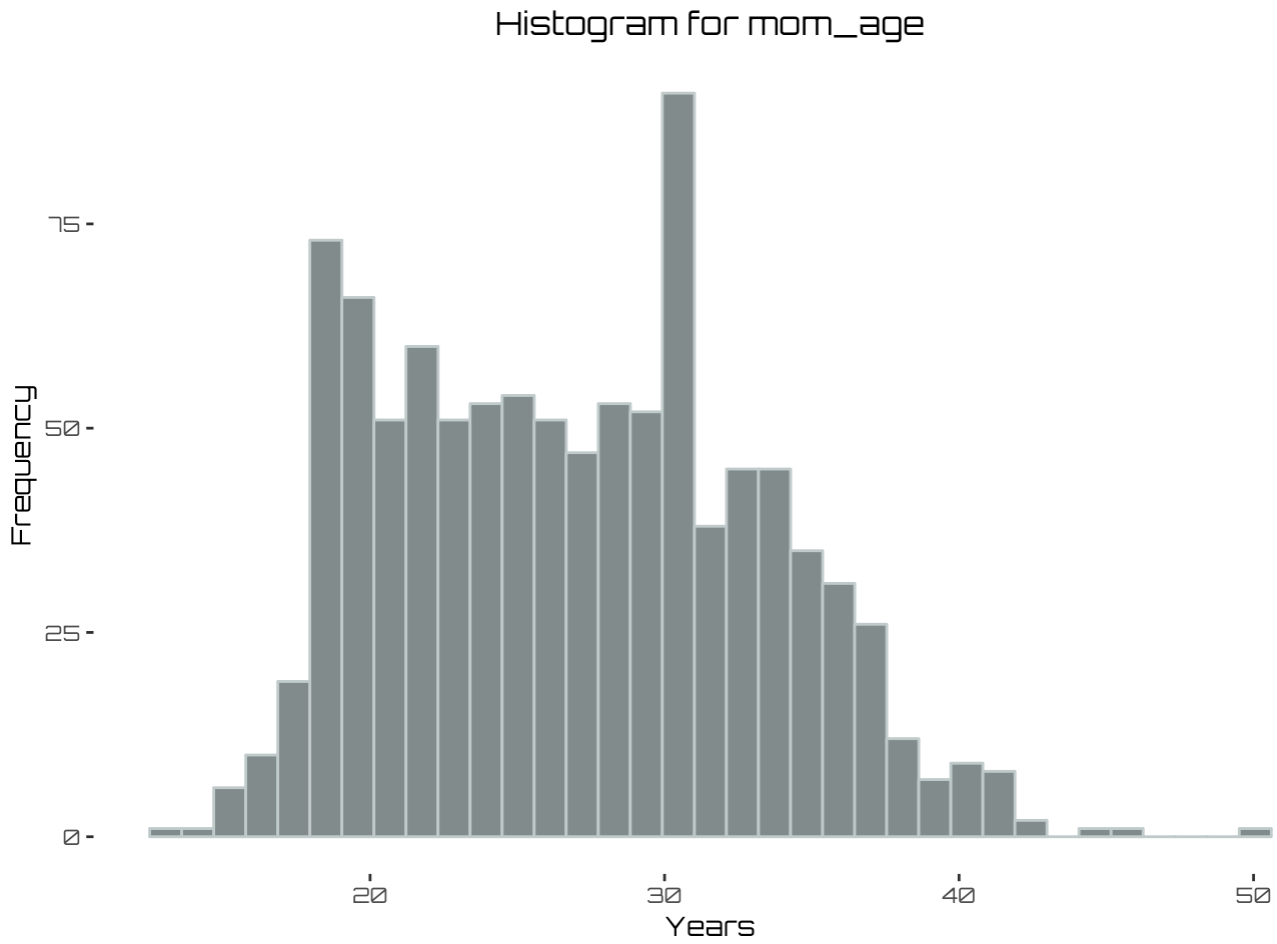
Patrick Boada
9 December 2019

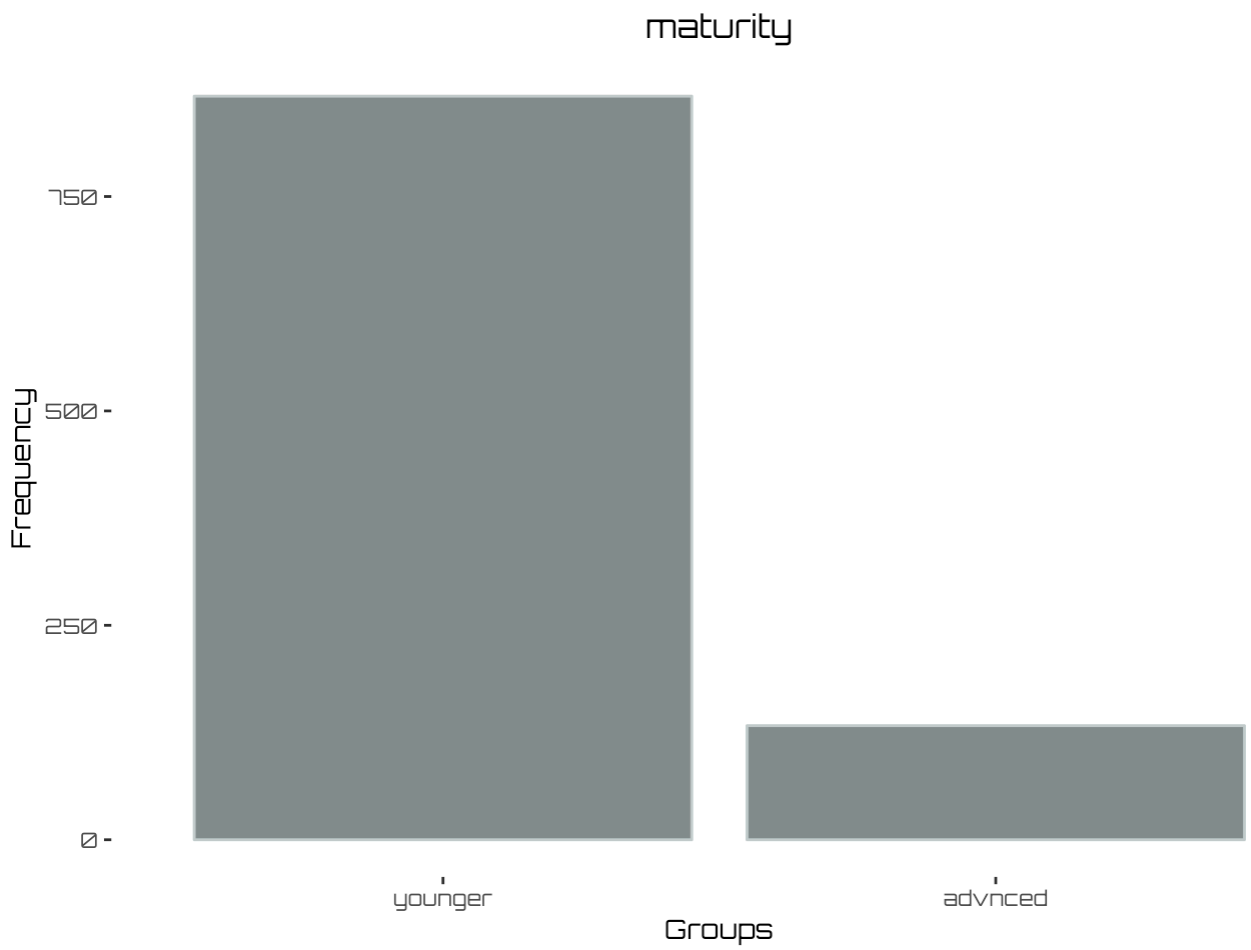
Objective: Data from the births.rdat will be used to create multivariate relations, linear models, and logistic models involving birthweight.

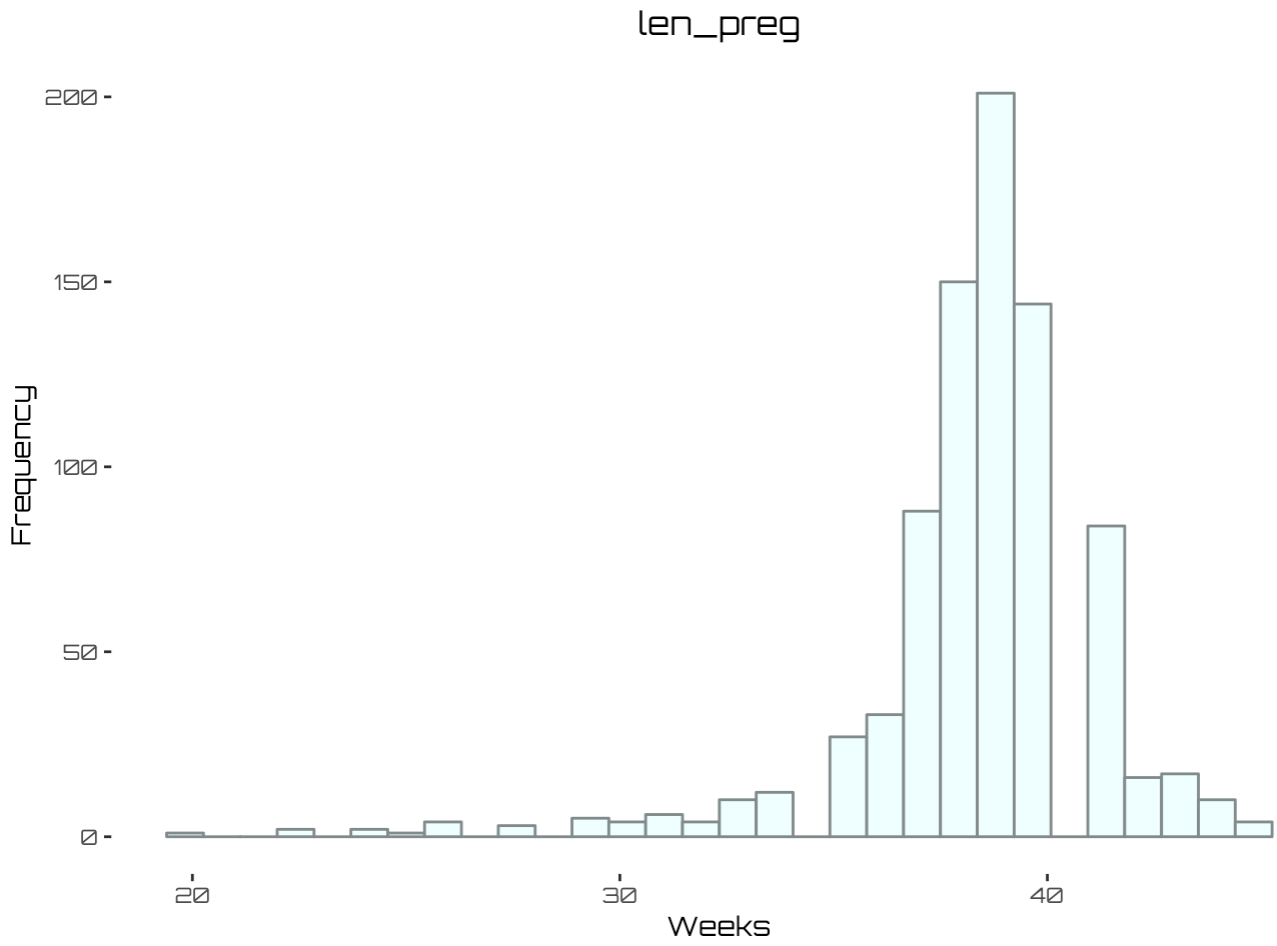
EDA

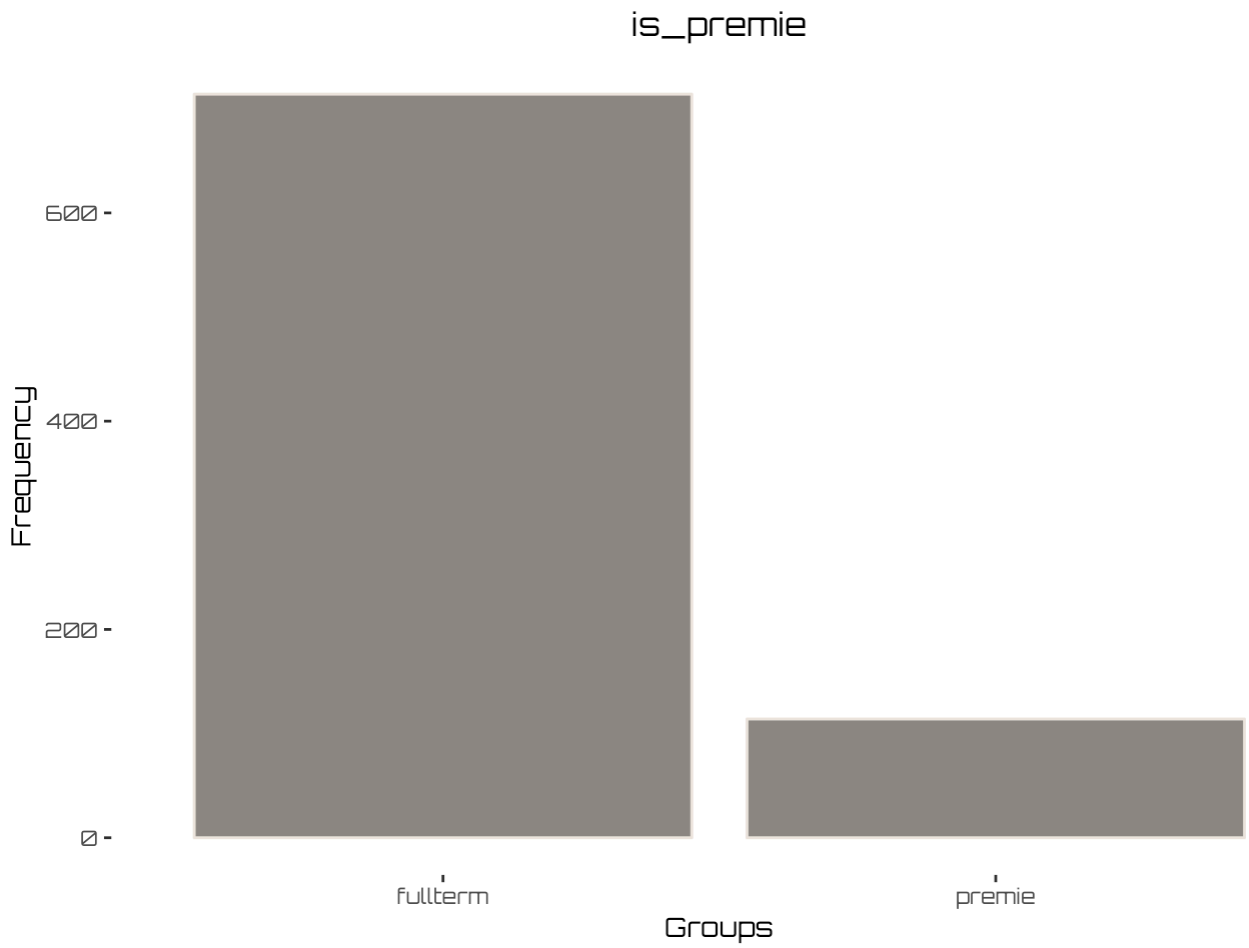
All column variables have been cleaned for outliers. A majority of cleaning involved removing outlier values '999', or releveling categorical variables to restore the dichotomous variable state.

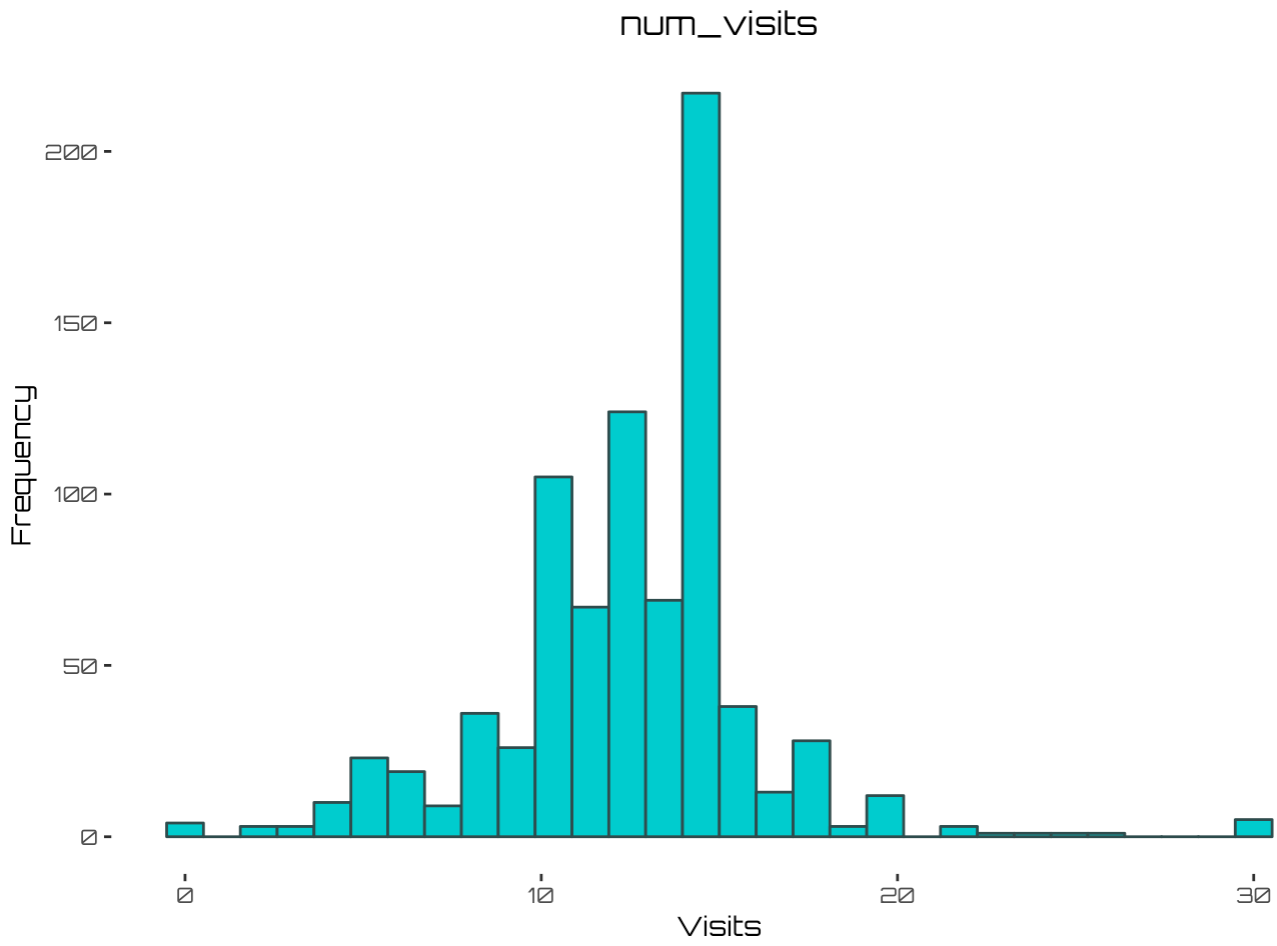


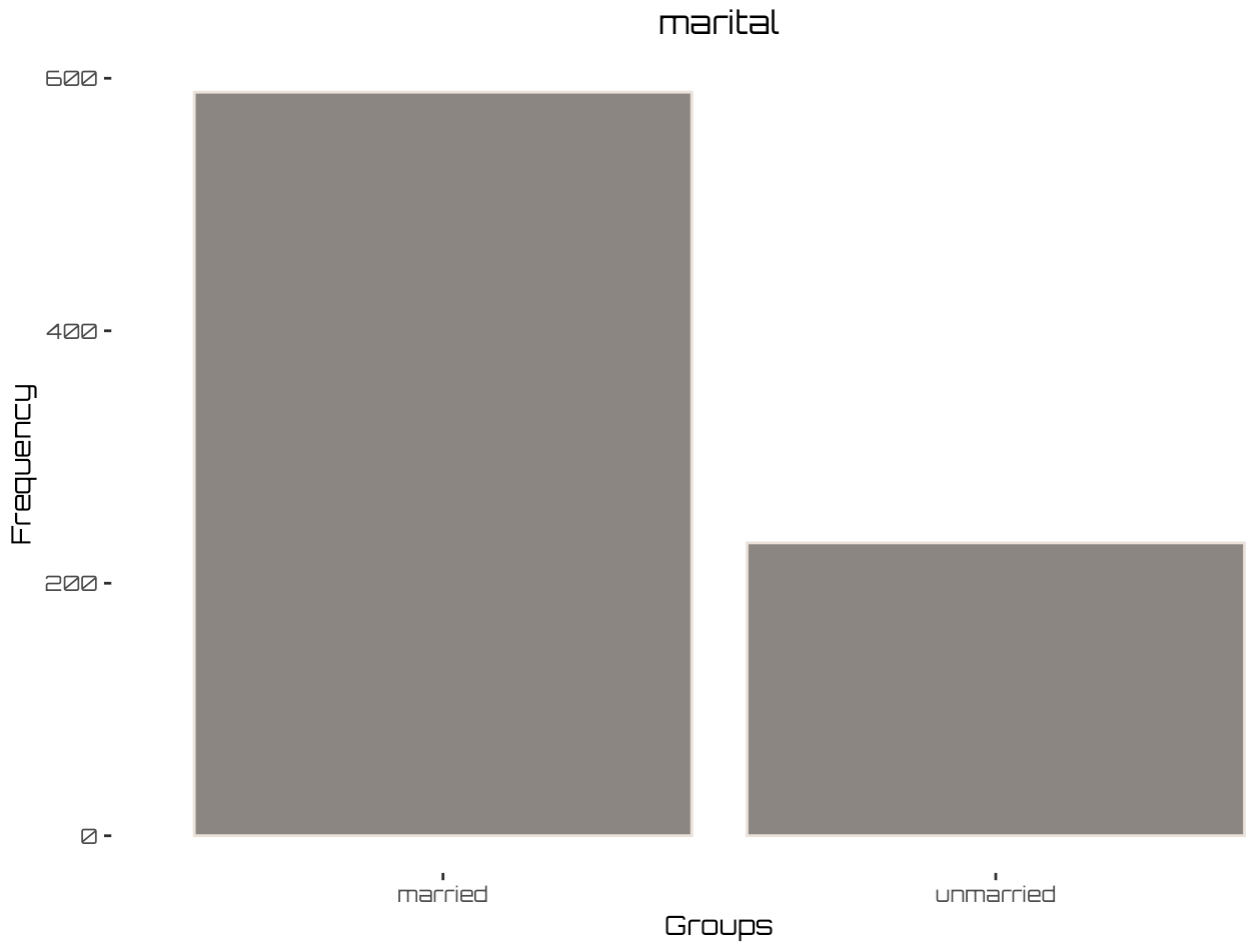

Histogram for mom_age
Frequency
75
50
25
0
20
30
40
50
Years

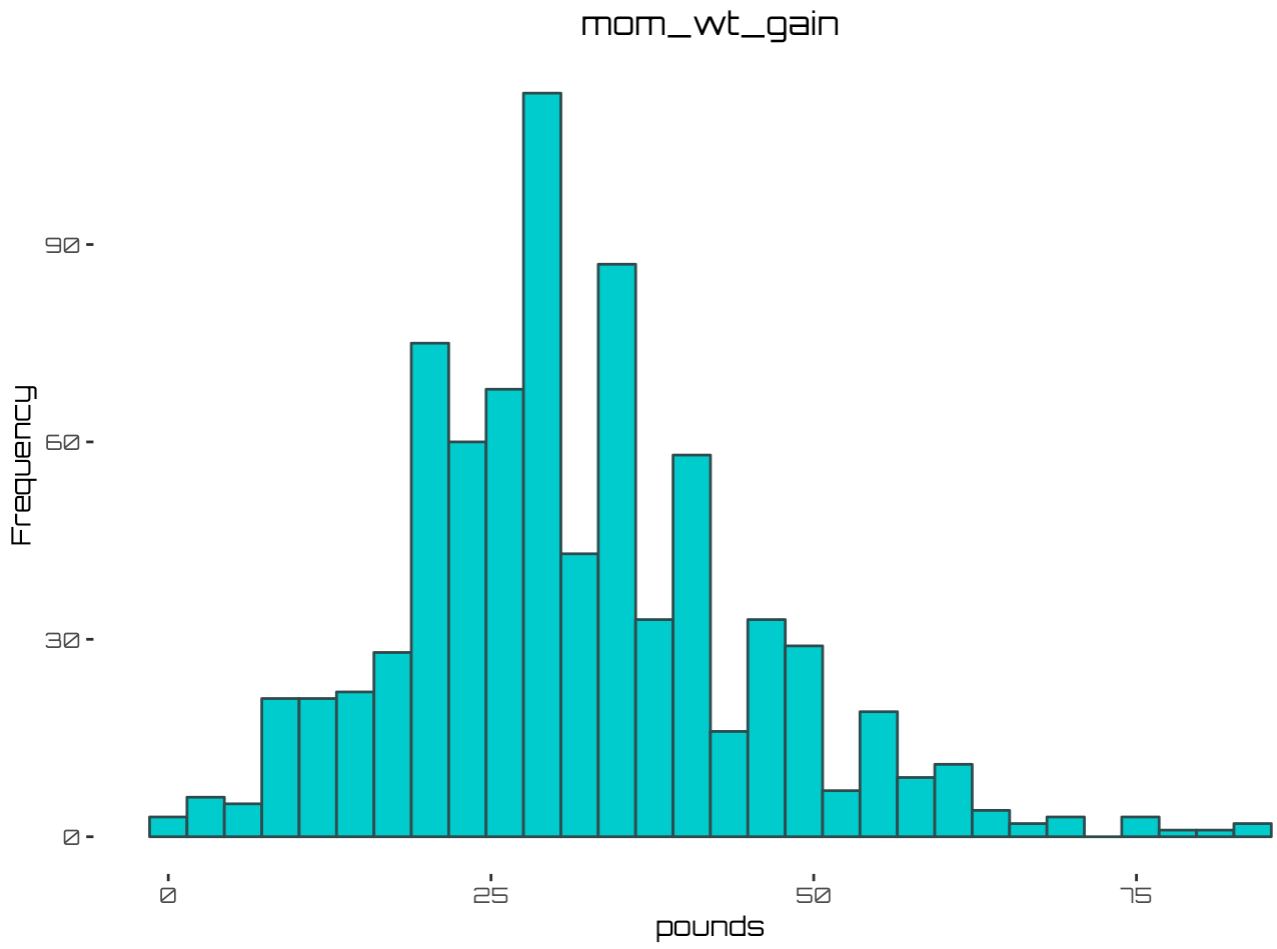

maturity
Frequency
750
500
250
0
younger
advnced
Groups

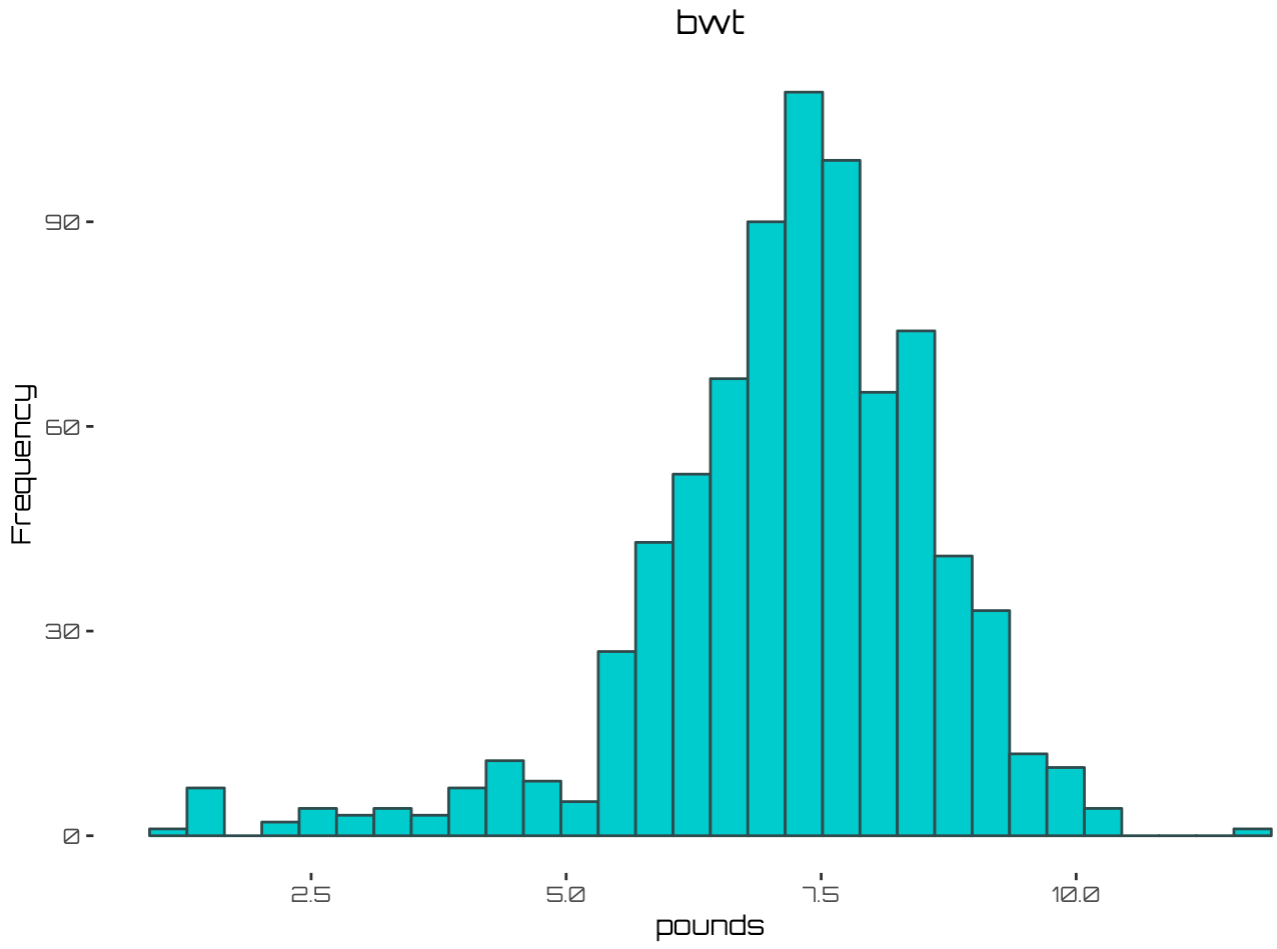

len_preg
Frequency
200
150
100
50
0
20
30
40
Weeks

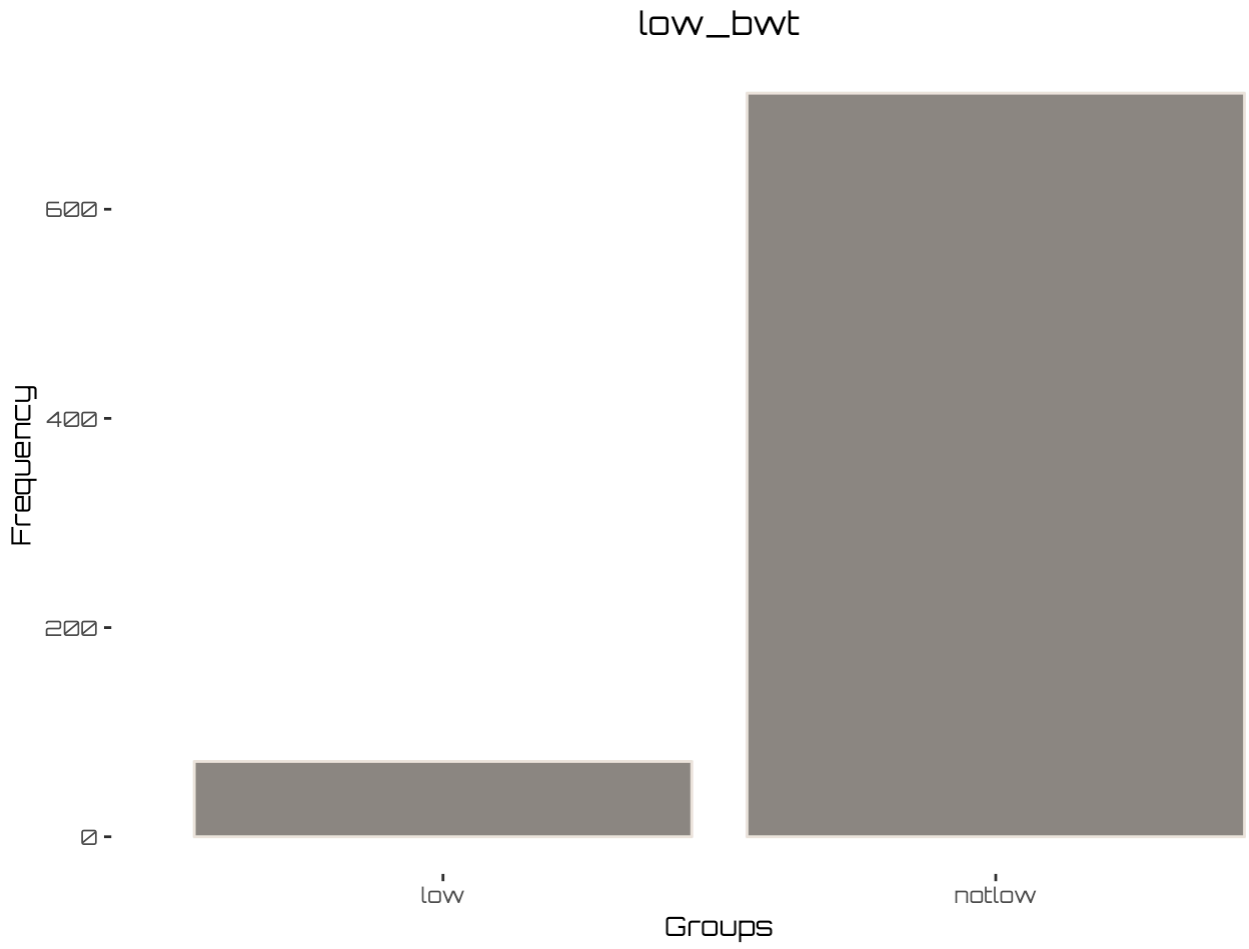

is_premie
Frequency
600
400
200
0
fullterm
premie
Groups

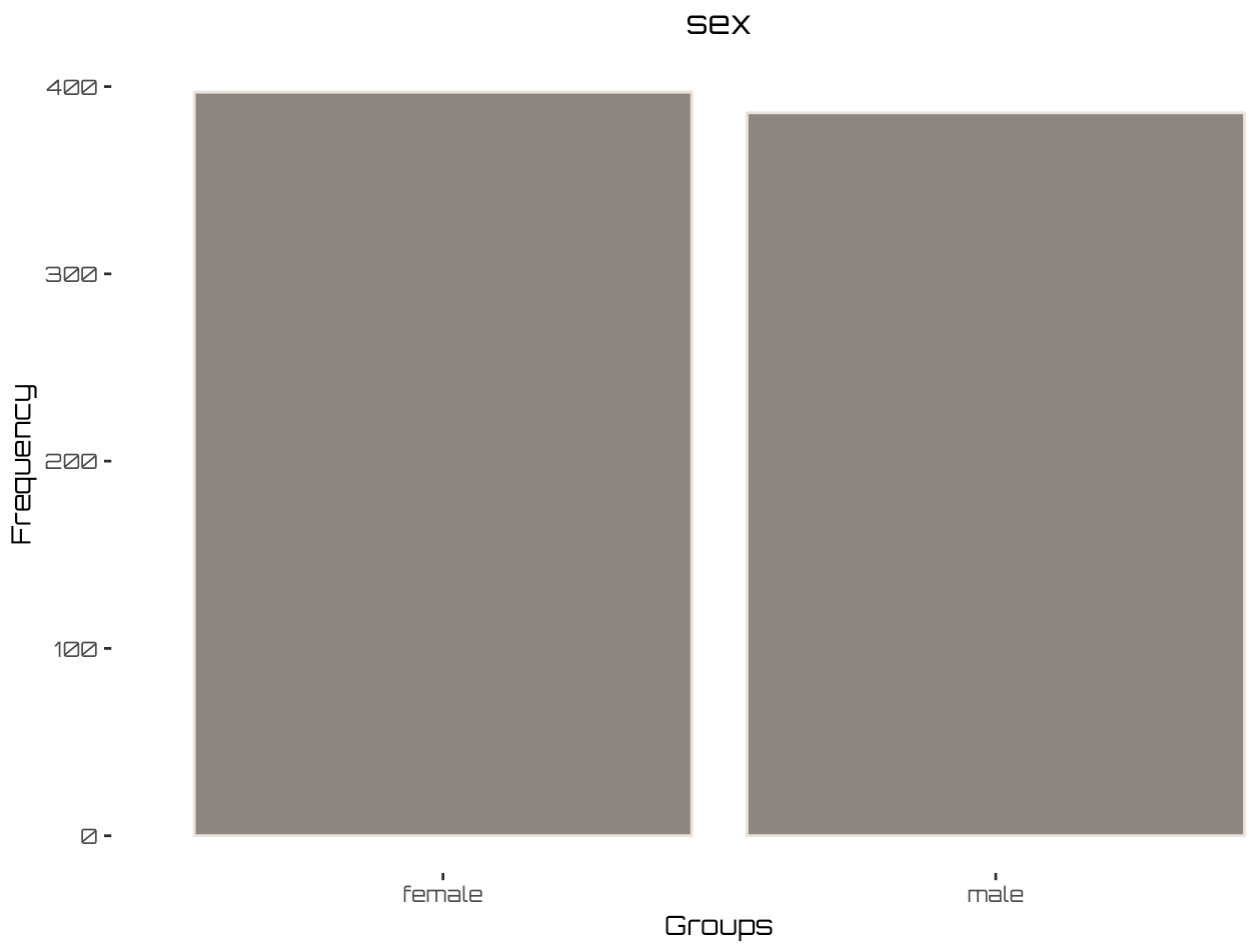

num_visits
Frequency
200
150
100
50
0
0
10
20
30
Visits

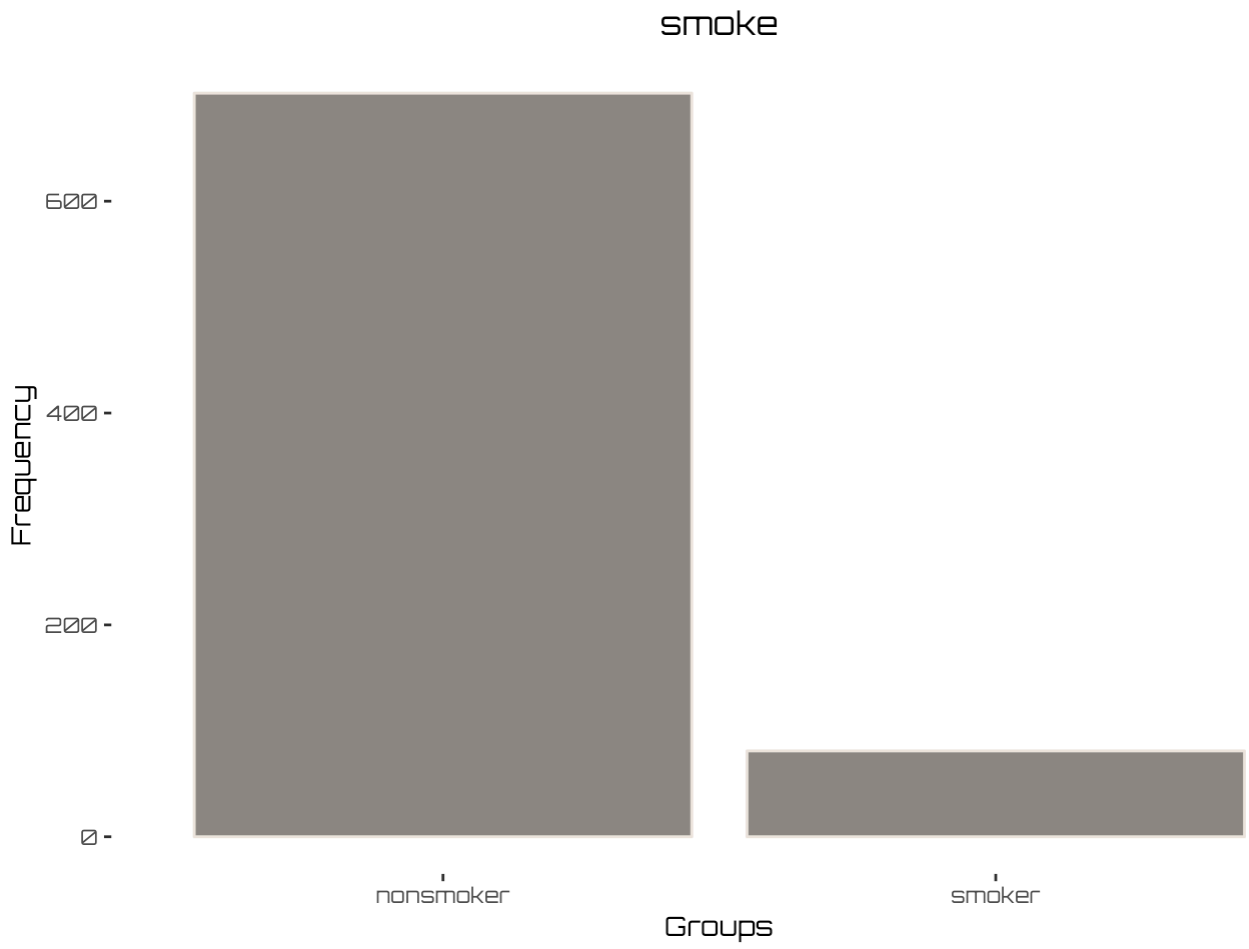

marital
Frequency
600
400
200
0
married
unmarried
Groups

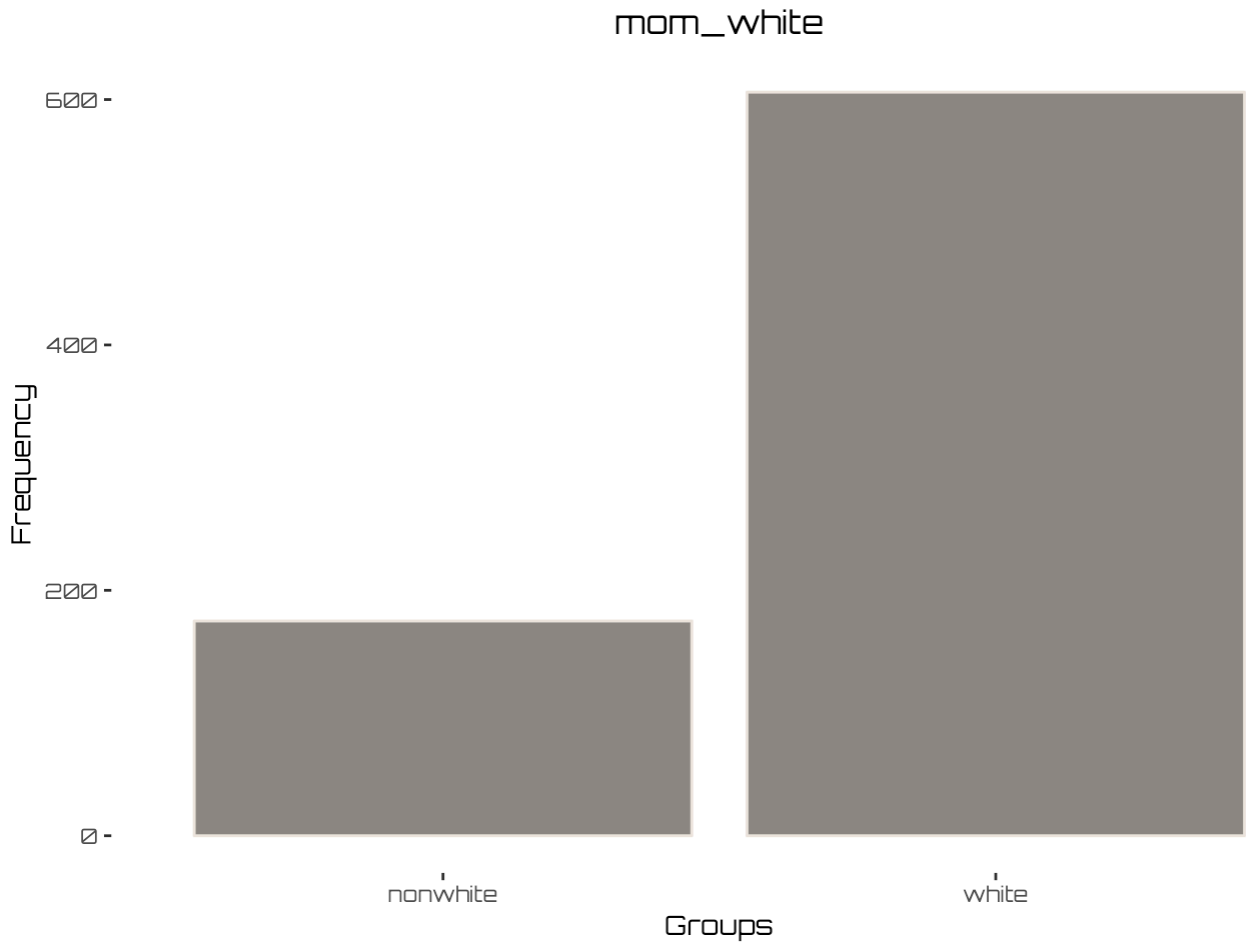

mom_wt_gain
Frequency
90
60
30
0
0
25
50
75
pounds

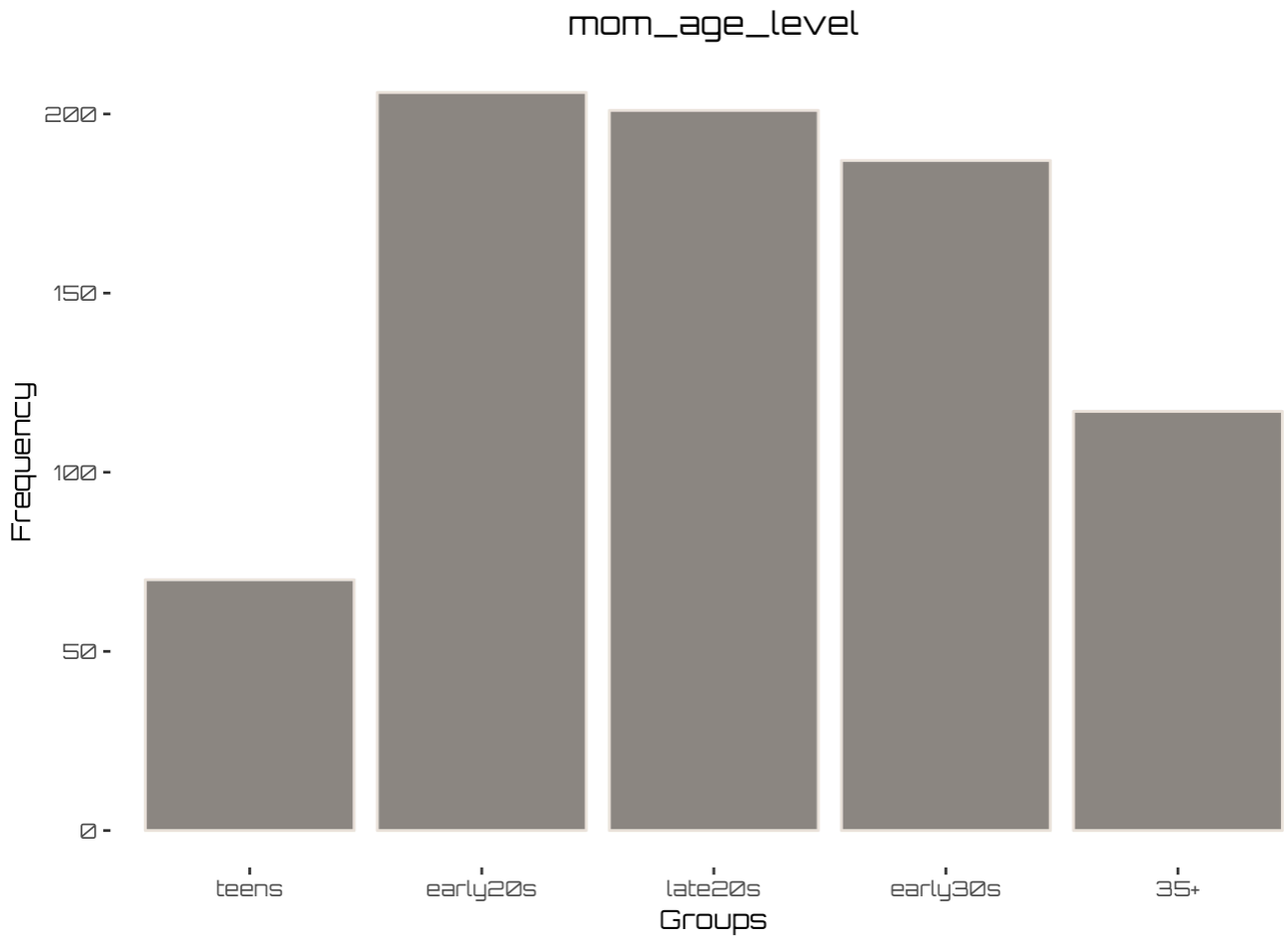

bwt
Frequency
90
60
30
0
2.5
5.0
7.5
10.0
pounds


low_bwt
600
400
200
0
Frequency
low
notlow
Groups


sex
400
300
200
100
0
Frequency
female
male
Groups


smoke
600
400
200
0
Frequency
nonsmoker
smoker
Groups


mom_white
600
400
200
0
Frequency
nonwhite
white
Groups

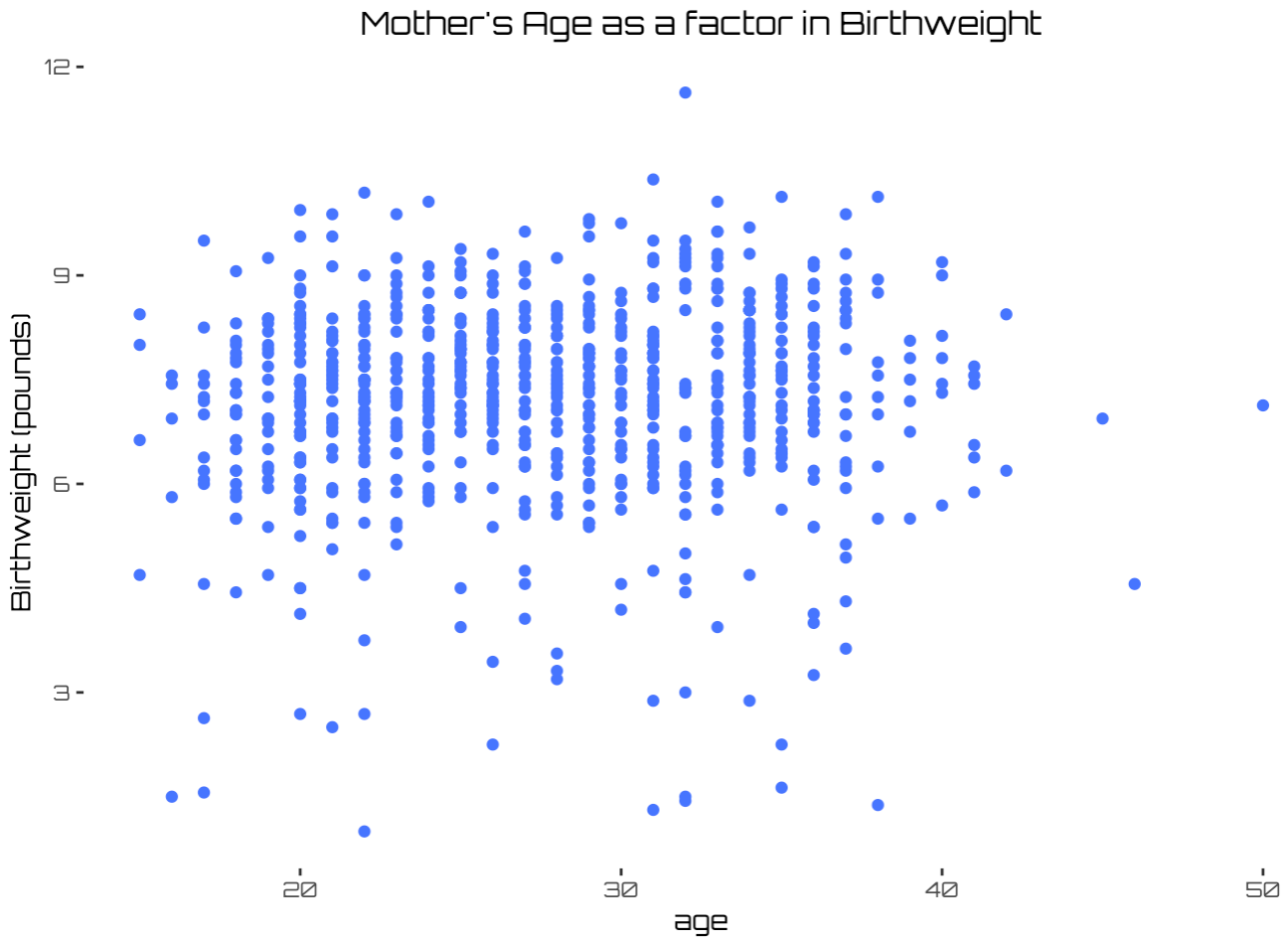


Births Summary

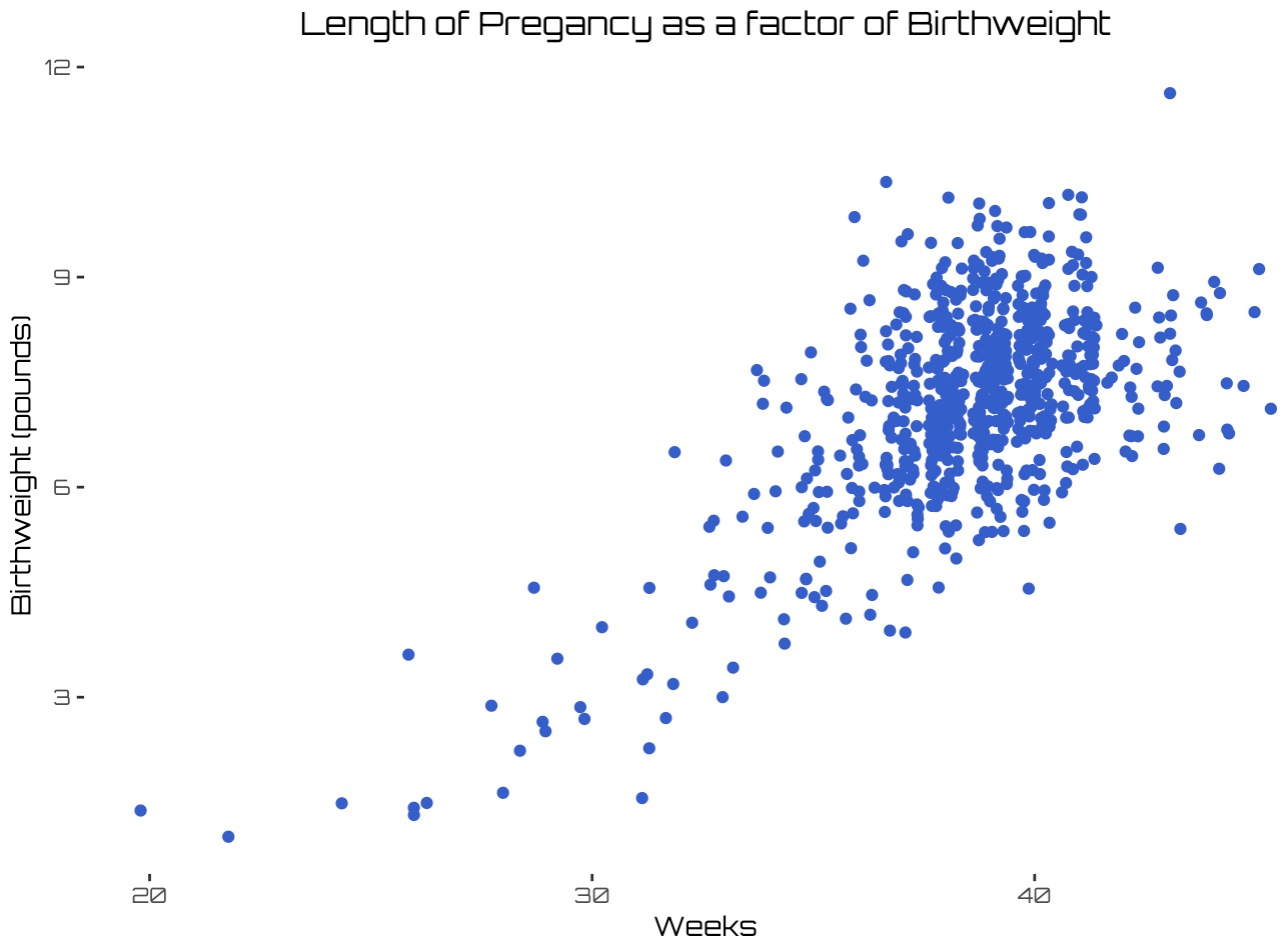
dad_age	mom_age	maturity	len_preg	is_premie
Min. :14.00	Min. :15.00	younger:664	Min. :20.00	fullterm:678
1st Qu.:25.00	1st Qu.:22.00	advnced:117	1st Qu.:38.00	premie :103
Median :30.00	Median :27.00		Median :39.00	
Mean :30.21	Mean :27.56		Mean :38.48	
3rd Qu.:35.00	3rd Qu.:32.00		3rd Qu.:40.00	
Max. :55.00	Max. :50.00		Max. :45.00	

num_visits	marital	mom_wt_gain
Min. : 0.00	married :562	Min. : 1.00
1st Qu.:10.00	unmarried:219	1st Qu.:22.00
Median :12.00		Median :30.00
Mean :12.42		Mean :31.52
3rd Qu.:15.00		3rd Qu.:40.00
Max. :30.00		Max. :85.00

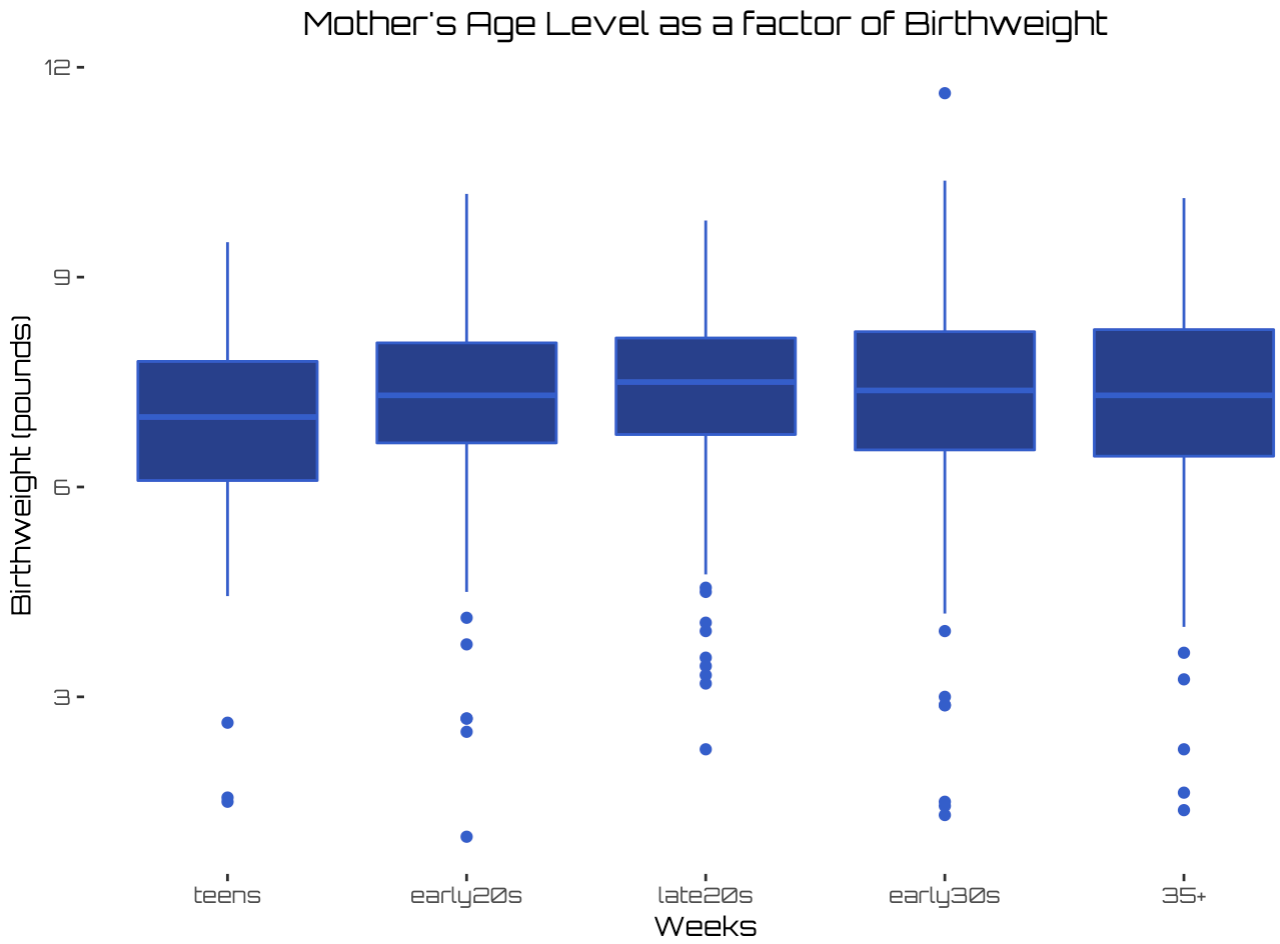
bwt	low_bwt	sex	smoke	mom_white	mom_age_level
Min. : 1.000	low : 72	female:397	nonsmoker:700	nonwhite:175	teens : 70
1st Qu.: 6.500	notlow:709	male :384	smoker : 81	white :606	early20s: 206
Median : 7.380					late20s :201
Mean : 7.208					early30s:187
3rd Qu.: 8.130					35+ :117
Max. :11.630					



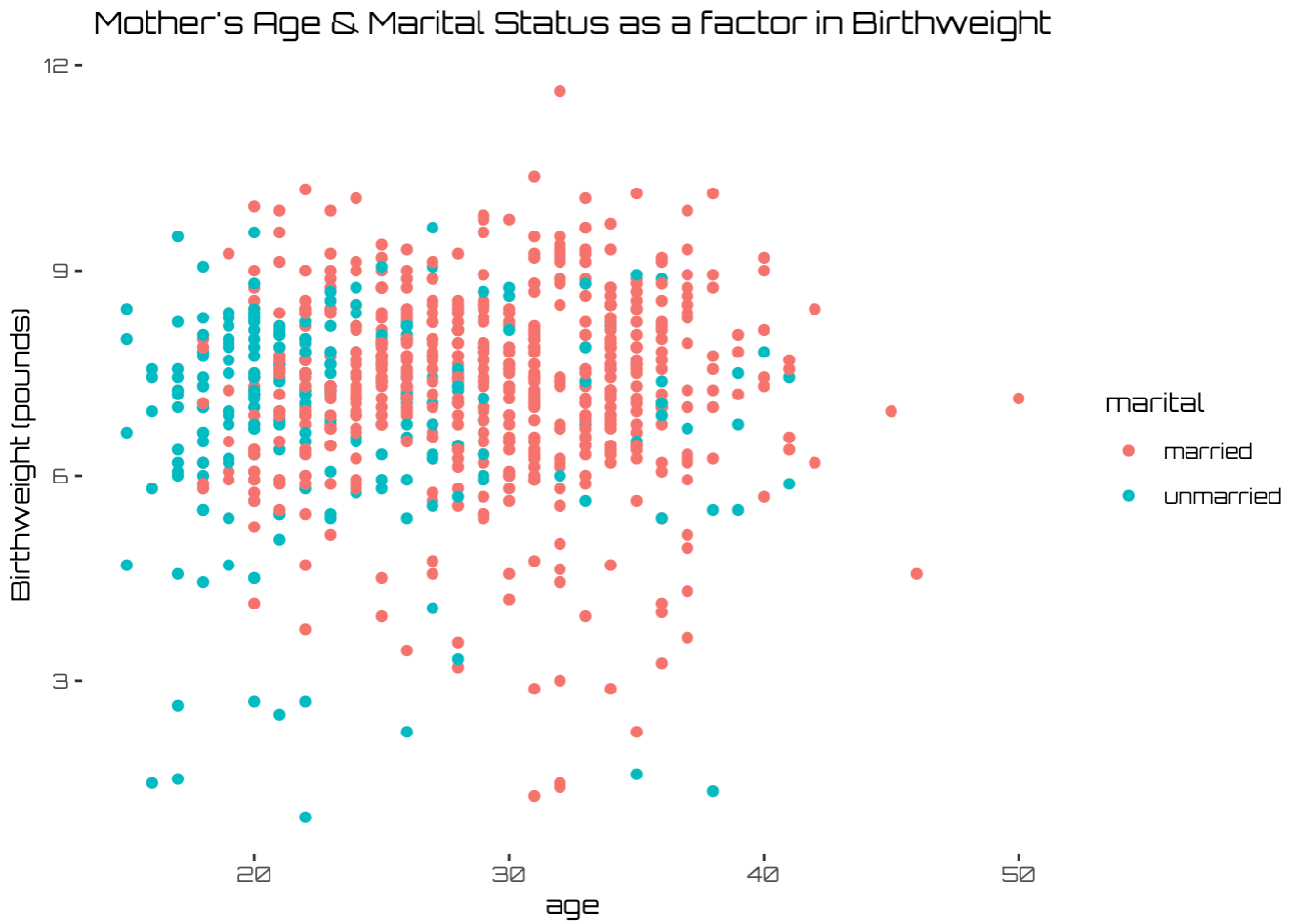
Mother's Age as a factor in Birthweight shows that the highest birthweight achieved was by an individual ~32 years old, and the lowest birthweight ~2 pounds was shared among ages less than 20 to 40 years old. Moreover we see that regardless of age there is a high frequency of babies born with a birthweight between 6 to 9 pounds.



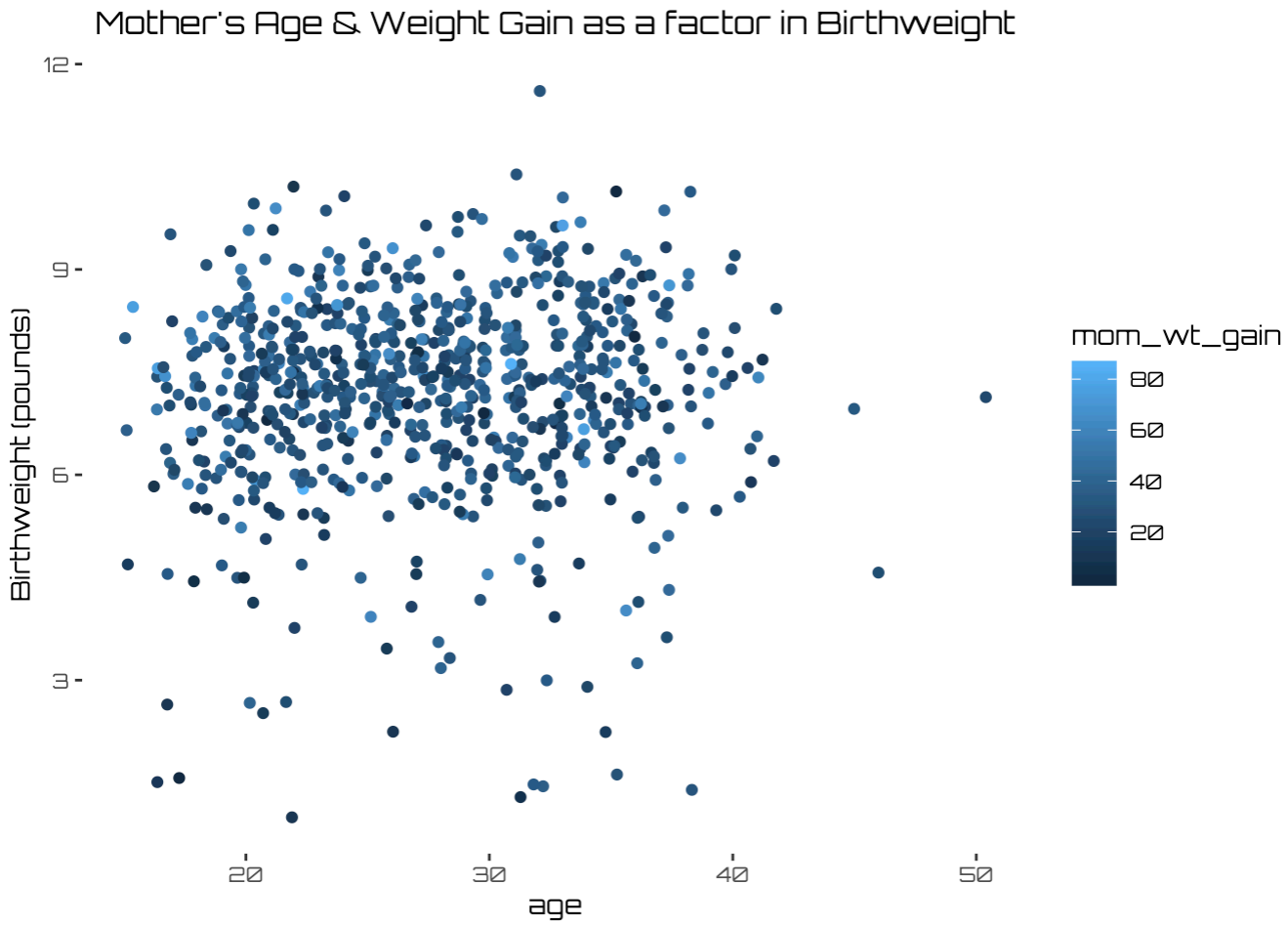
Length of Preganancy as a factor of Birthweight shows that the highest birthweight achieved was from a gestation length above 40 weeks. The lowest birthweights are associated with gestation lengths less than 25 weeks. Moreover we see a high clustering frequency of babies between 35 to 40 weeks with a birthweight ranging from 6 to 10 pounds.



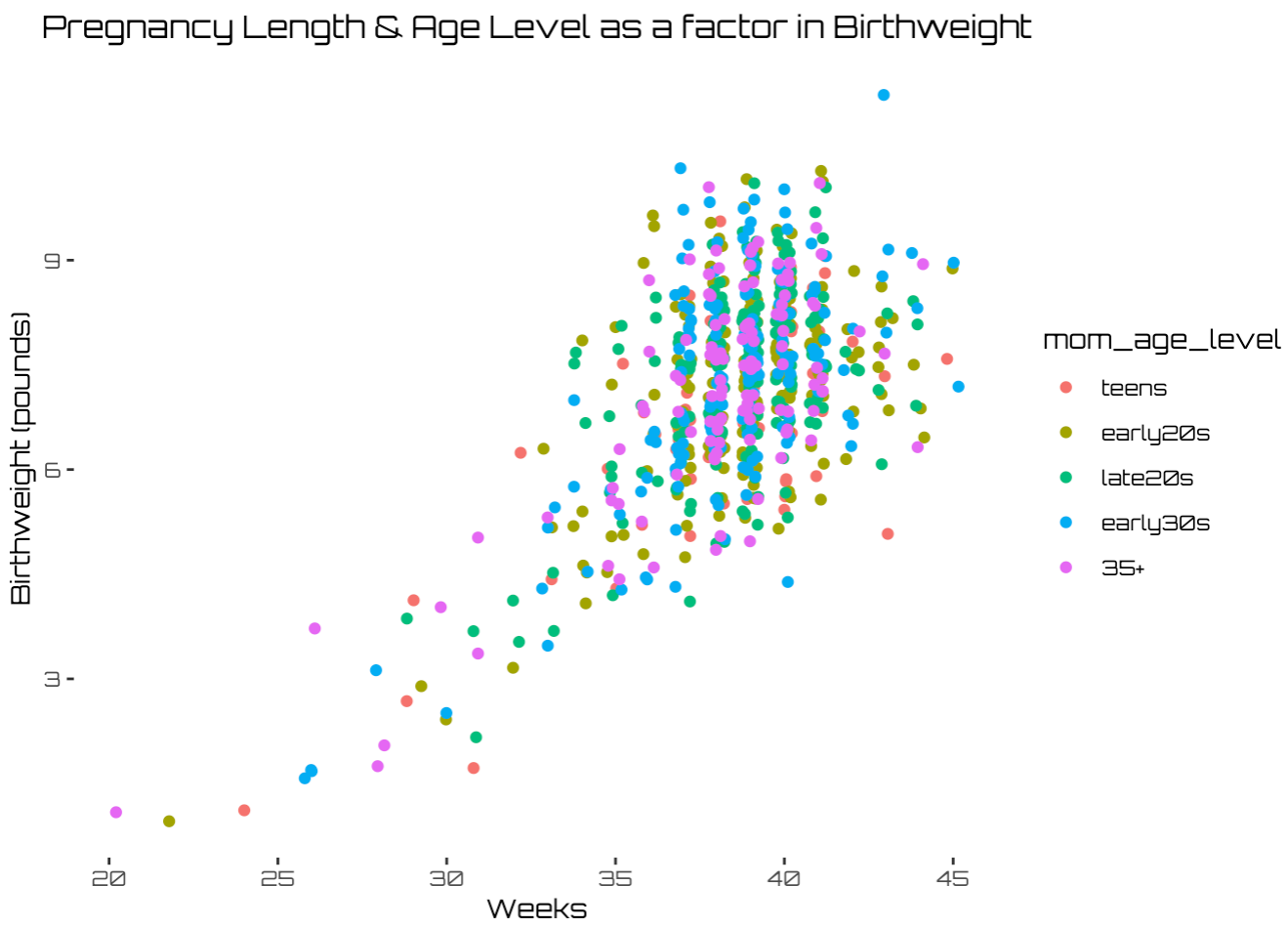
Mother's Age Level as a factor of Birthweight shows that median values are all nearly shared among age groups around 7.5 pounds. All age groups demonstrate an occurrence of low minimum values for birthweight that are less than 3 pounds. However it may be observed that late 20s has the tightest IQR among all groups, with the highest relative median as well. Nevertheless the highest maximum birthweight was rendered from a patient from age group early30s.



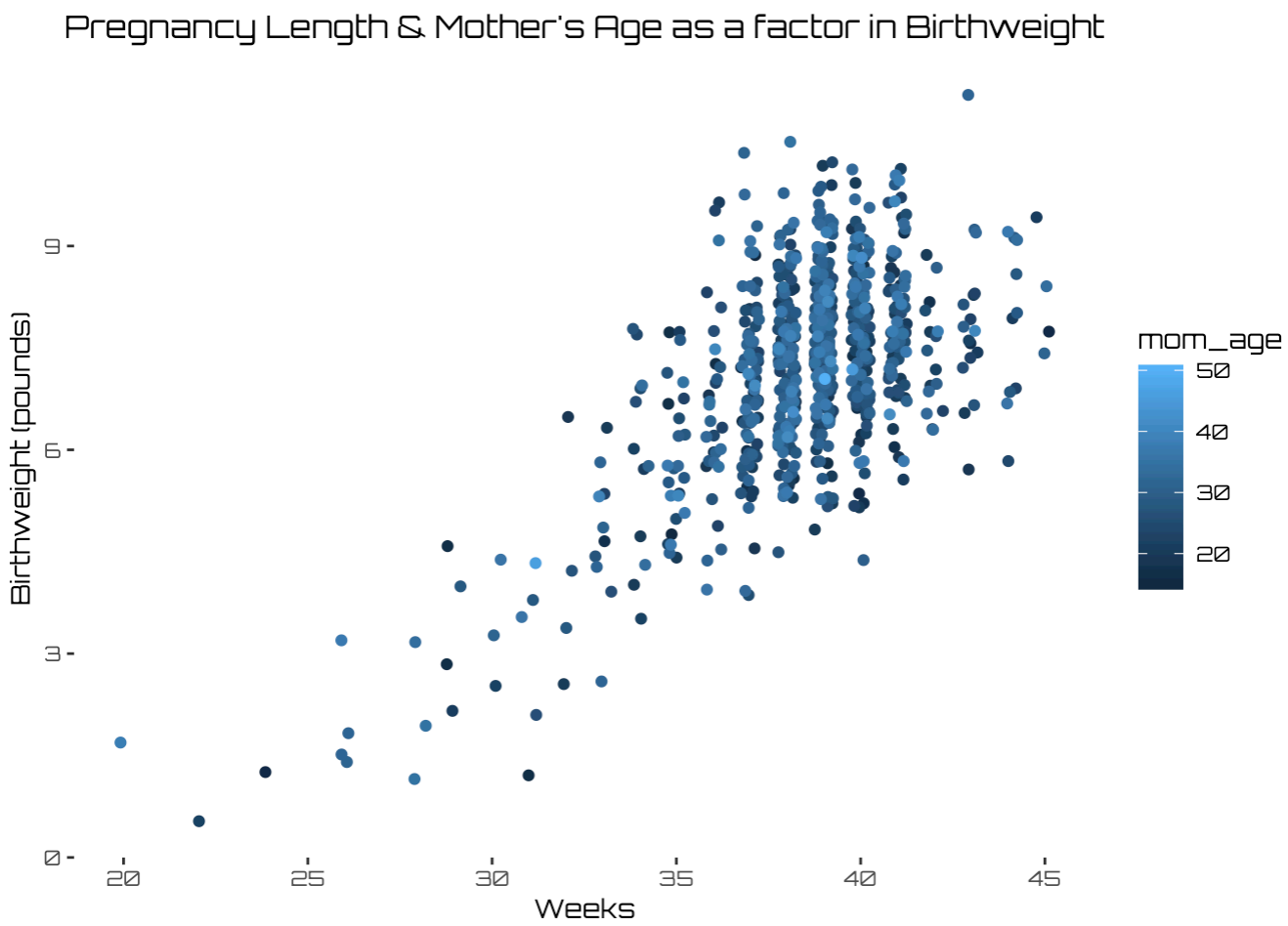
Mother's Age & Marital Status as a factor in Birthweight depicts the amount of weight of women married compared to single. The plot demonstrates that the highest weights were rendered from married woman. Lower birthweights do exists in both groups. Moreover we see a majority of married mothers in the sample.



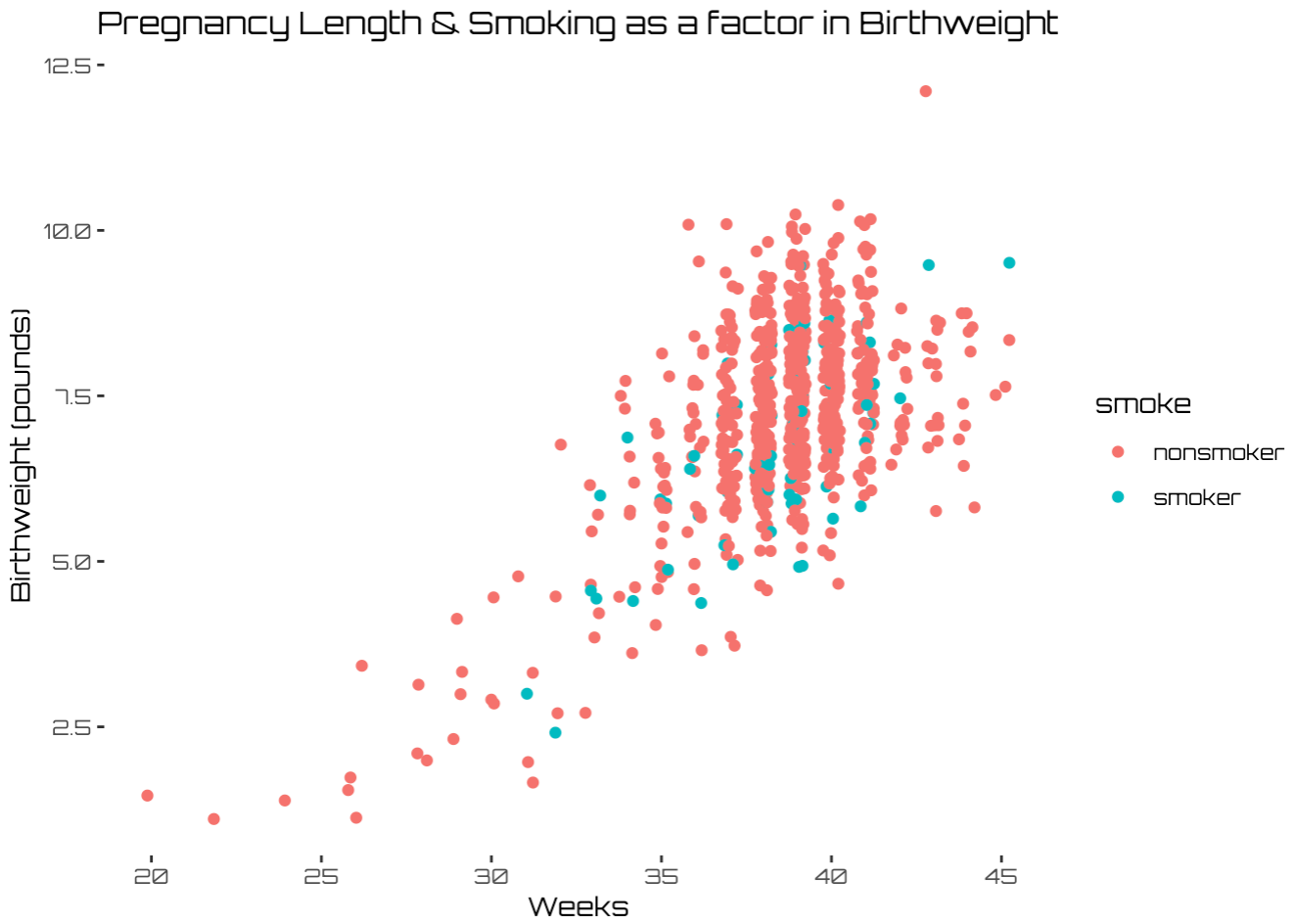
Mother's Age & Weight Gain as a factor in Birthweight depicts the net gained weight of women, with respect to the infant's birthweight and the mother's age. The plot demonstrates that the highest weight gained in mothers resulted in baby most frequently above ~7.5 pounds. Lower weight gain does exists on both sides of the birthweight range. However, a majority of incidents of lower weight gained occur in babies born below 6 pounds and predominantly in babies born below 3 pounds.



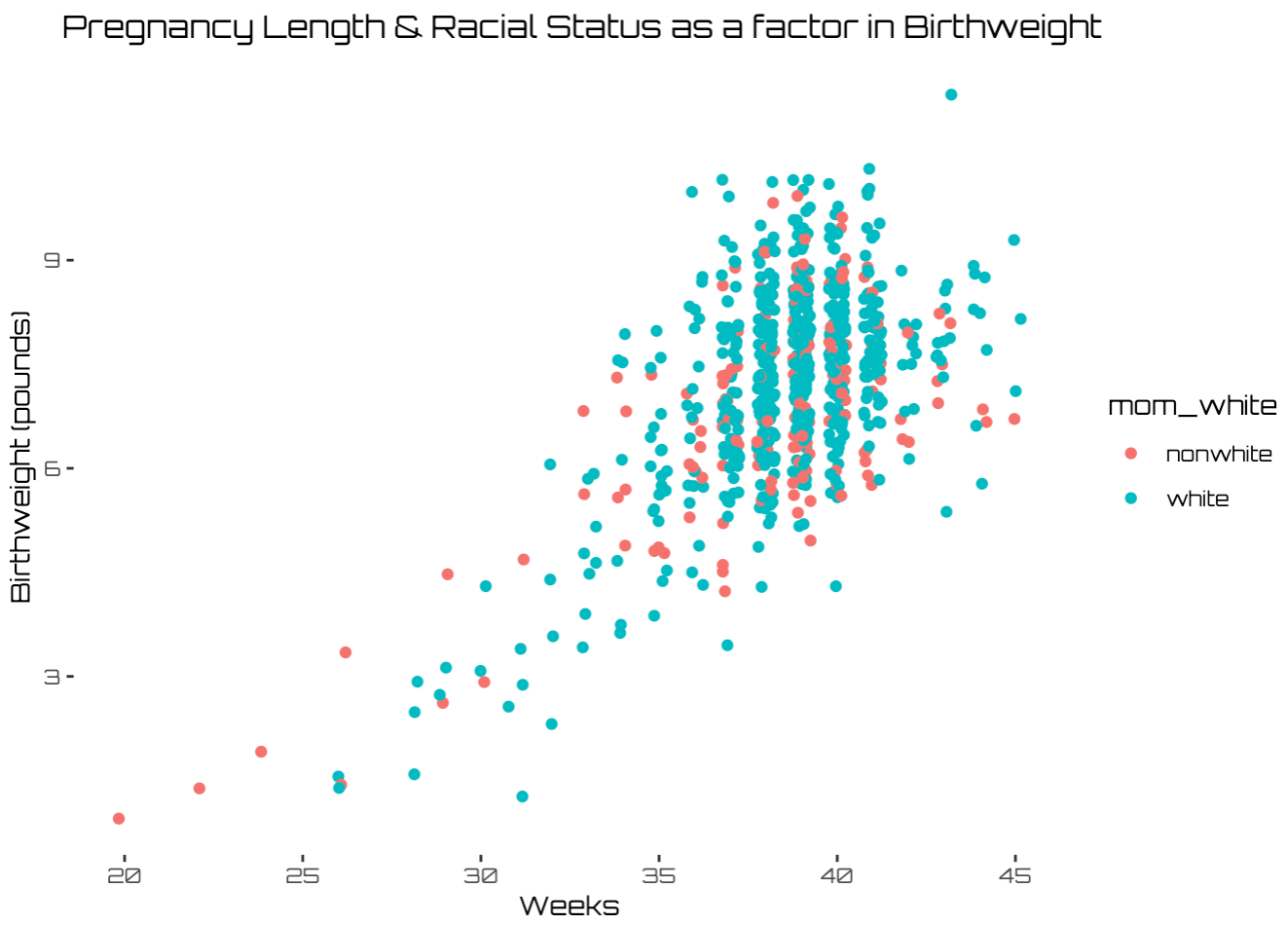
Pregnancy Length & Mother's Age Level as a factor in Birthweight demarcates the frequency of age levels with respect to gestation length and birthweight. One may observe that there is a high frequency of infants born between 6-11 pounds predominantly from 35+. Furthermore there is not one single age level that overwhelming composes lower birthweights. However we do see higher birthweights achieved from mothers in early30s with a gestation length between 35-40 weeks.



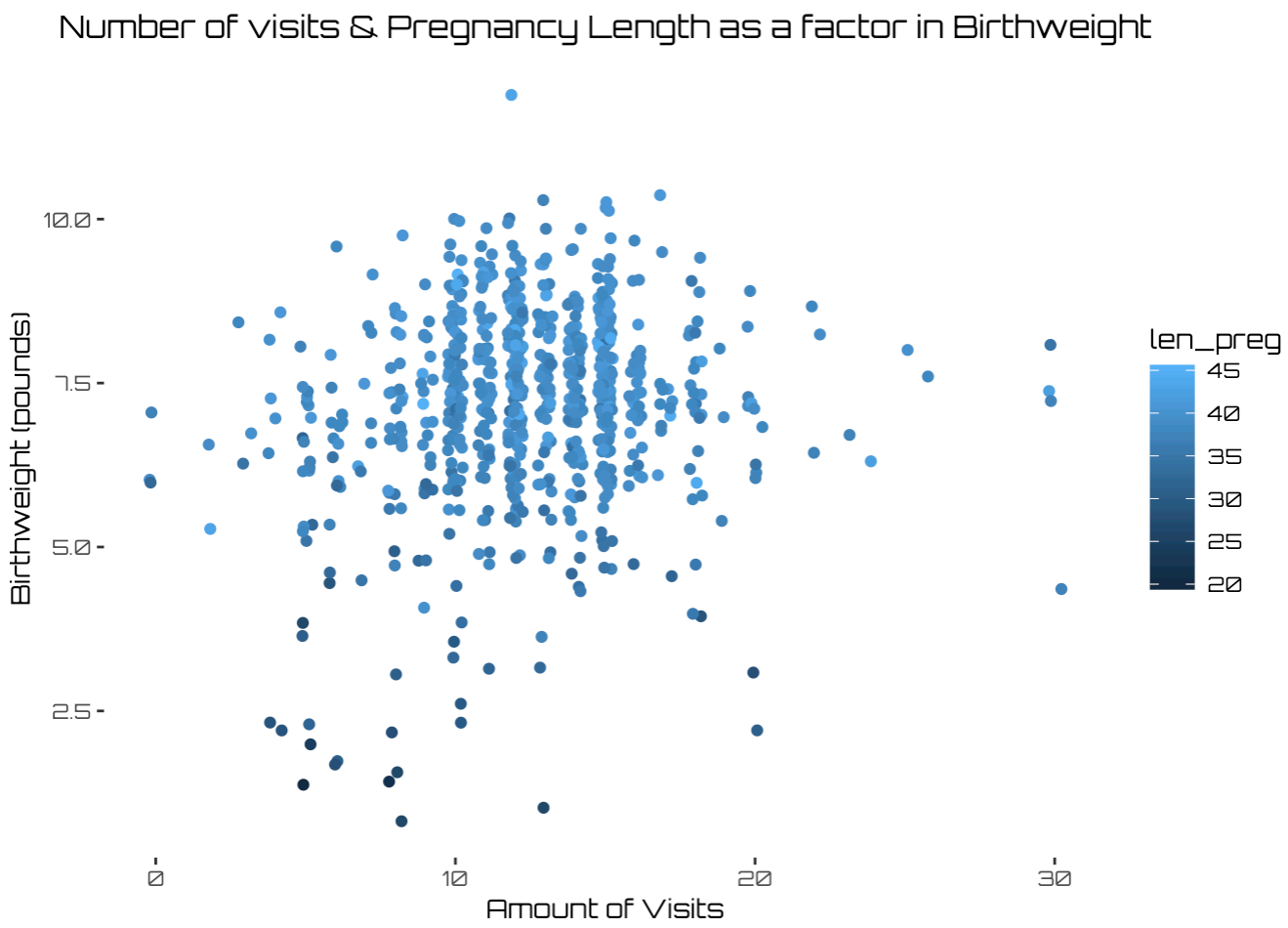
Pregnancy Length & Mother's Age as a factor in Birthweight emphasizes the age of mothers with respect to their gestation length and birthweight of the infant. The plot confirms that babies with a gestation length ranging between 35 - 40 weeks is composed of mothers ~35 years old. Moreover there are not many incidents of mothers above 40 with gestations length shorter than 30 weeks or with babies less than 6 pounds.



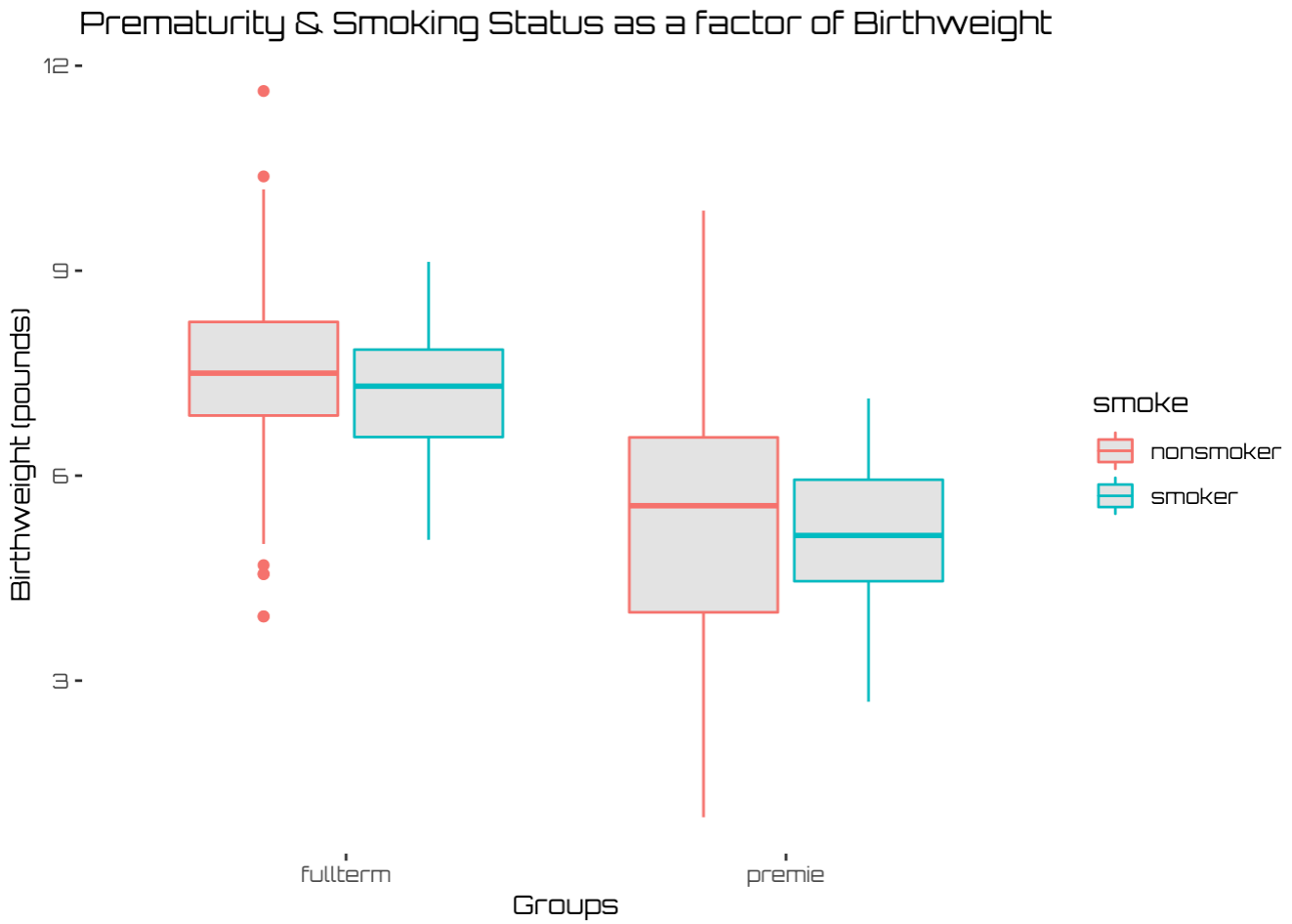
Pregnancy Length & Smoking as a factor in Birthweight separates nonsmoking mothers with respect to gestation length and birthweight. Indisputedly, one may observe that there exist a high frequency of babies born with birthweights equal to or greater than 7.5 pounds with gestation ranging between 35-42 weeks from nonsmoking mothers. Although smoking mothers do exists along the 40-45 week spectrum of gestation, their birtweights do not exceed ~7.5 pounds. Alternatively, the occurence of shorter gestation length and lower birthweight is not only exhibited in smoking mothers. In this sample (that has a larger amount of nonsmoking mothers), we see a high proportion of nonsmoking mothers experience shorter gestation and lower birthweight compared to smoking mothers.



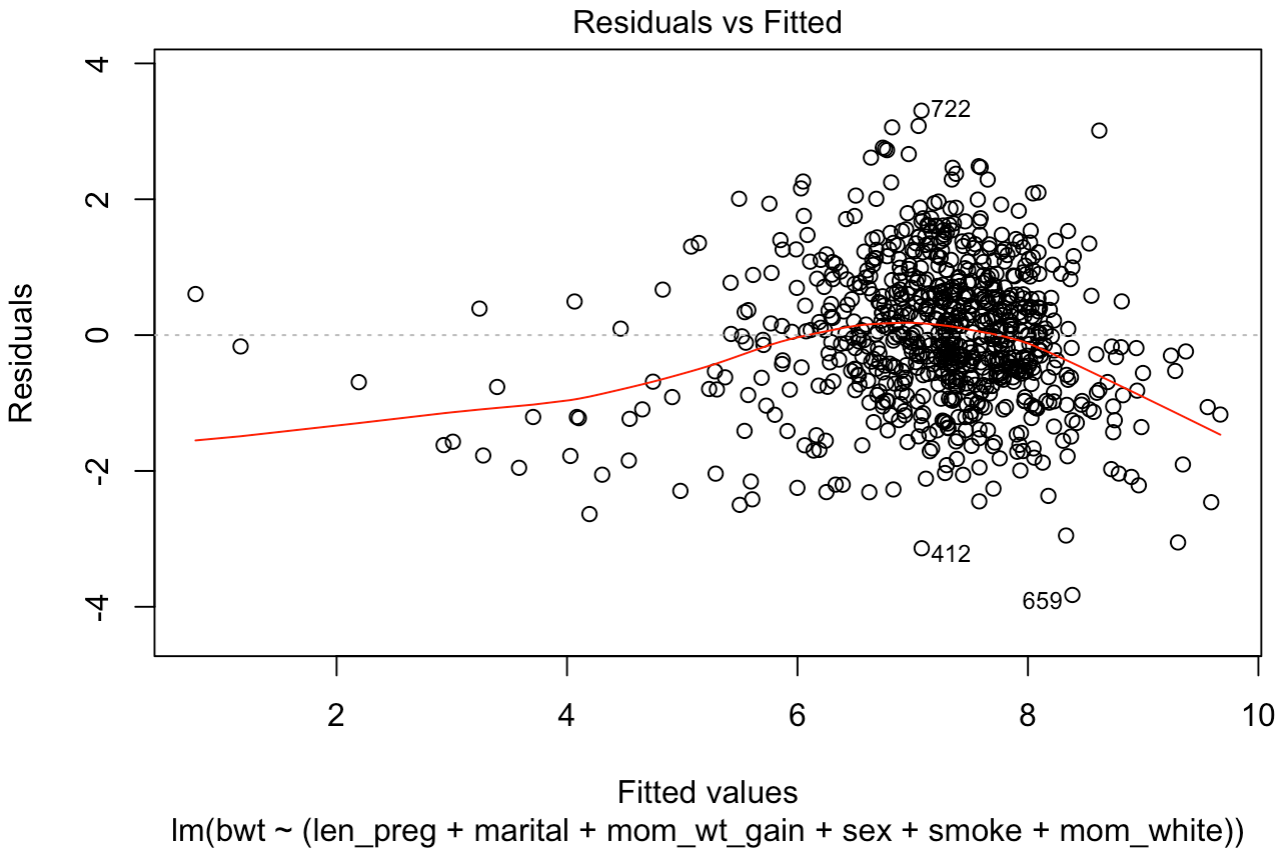
Pregnancy Length & Mother's Racial Status as a factor in Birthweight highlights European descent compared to all other ethnic classes. In the plot one can see that a majority of the sample is derived of white mothers. Although lower birthweights and lower gestation length is shared among ethnic groups one can observe that there exists a higher amount of high birthweights and longer gestation in white mothers compared to nonwhite mothers.

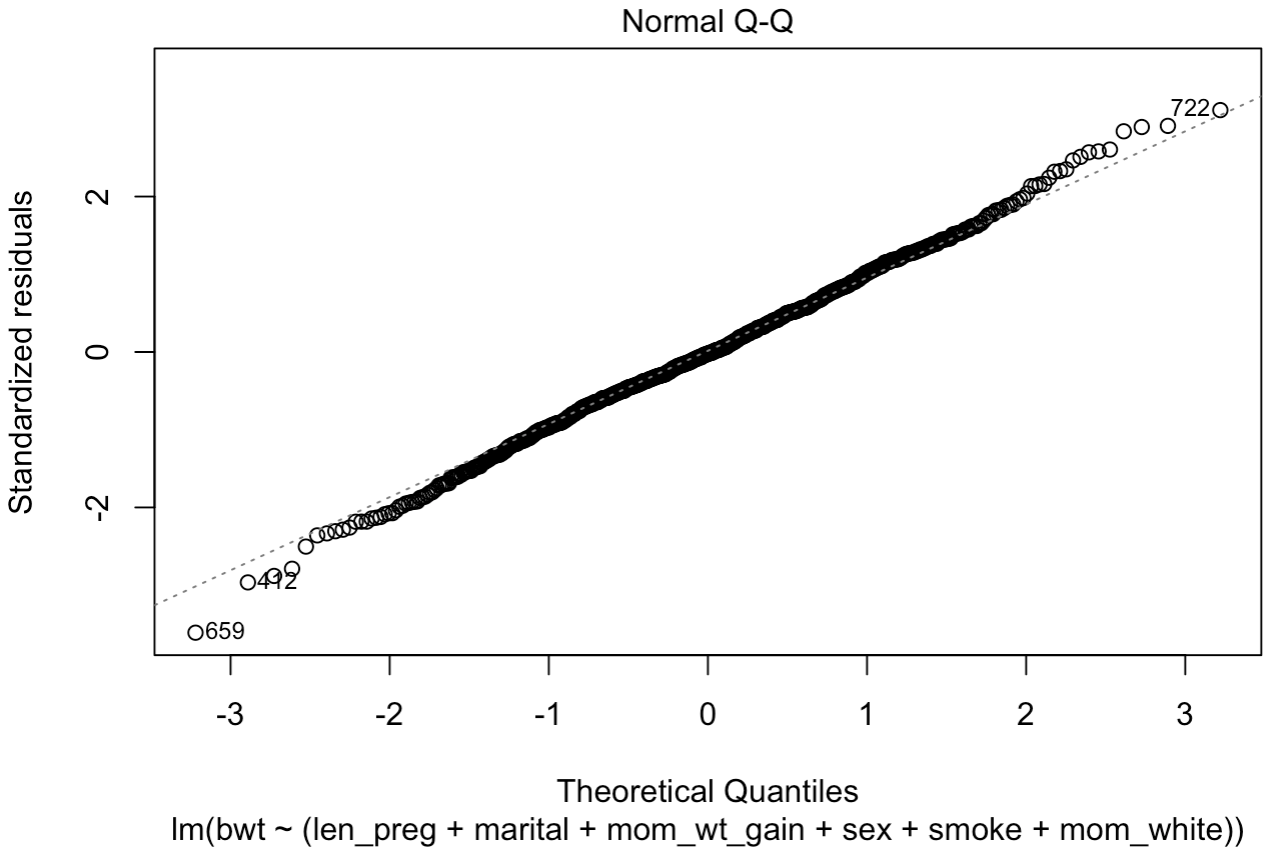


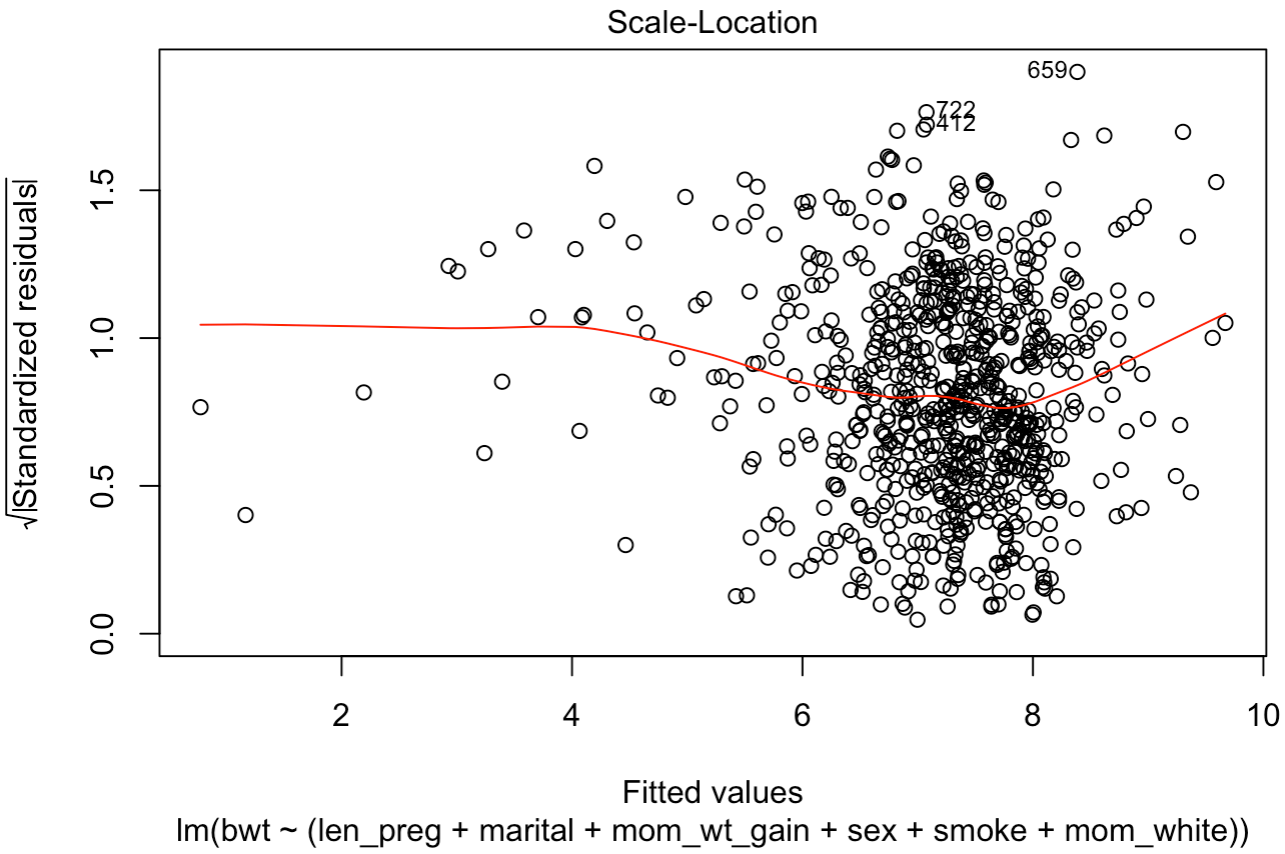
Number of visits & Pregnancy Length as factor in Birthweight demarcates the length of pregancy with respect to amount of visits for resulting birthweight. The highest occurence of visits is 10 -20. Among this range the length of pregnancy is ~40 weeks and above for a birthweight equal or higher than ~7.5 pounds. Naturally, we see that the lowest gestation length and birthweight relates to the lowest amount of visits. Moreover, that the highest number of visits corresponds to the highest gestation length.

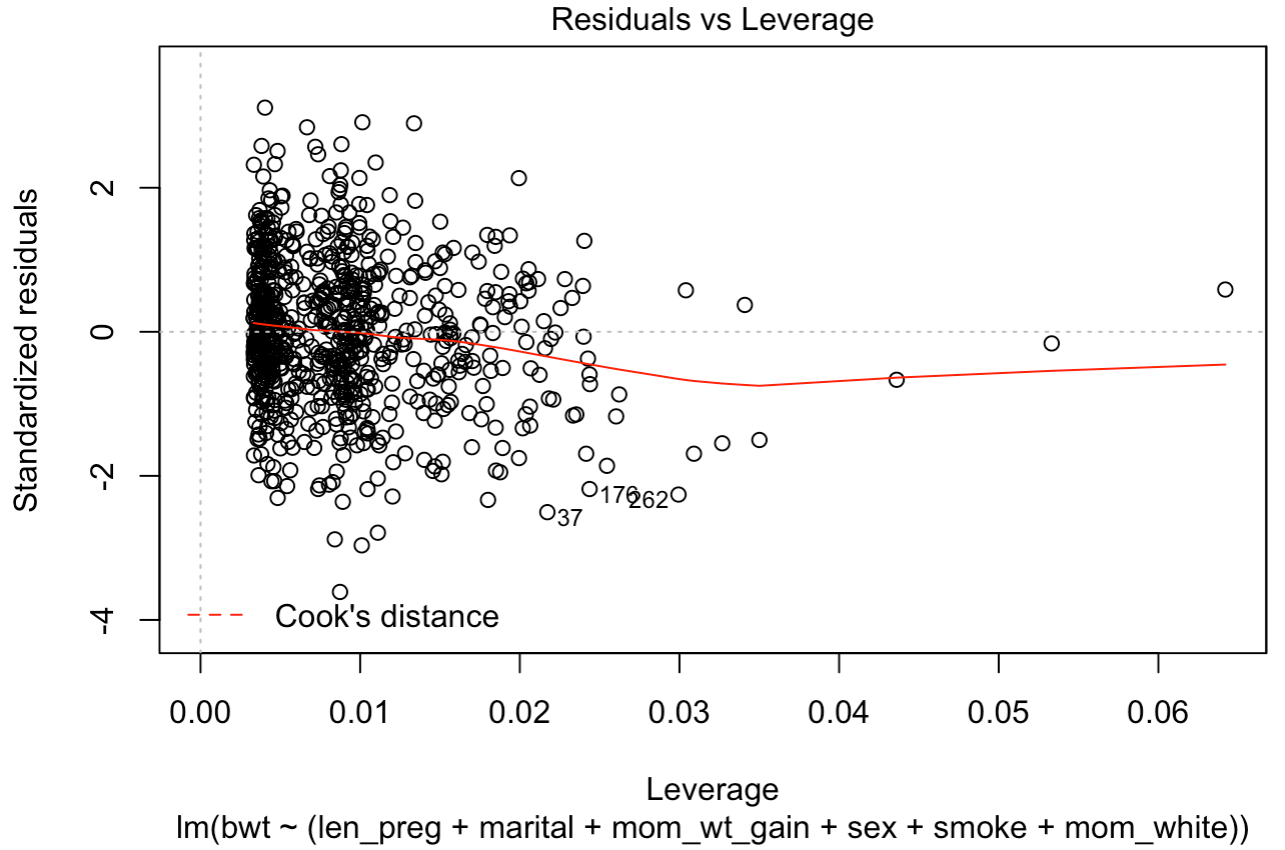


Prematurity & Smoking Status as a factor of Birthweight shows the distinction of premature birthweight with relation to mother's smoking status. In smoking mothers median values of birthweight are both lower than their nonsmoking counterparts. Moreover, fullterm pregnancy of a smoking mother is is ~7.5 pounds, whereas a premature pregnancy of a smoking mother is ~5 pounds. Nonsmoking mothers of premature babies have a wide range when compared to other groups. Fullterm pregancy medians are similar between smoking classes, the range of birthweight of smoking fullterm pregnancies is much lower nonsmoking fullterm pregnancies.


Residuals vs Fitted
Residuals
Fitted values
lm(bwt ~ (len_preg + marital + mom_wt_gain + sex + smoke + mom_white))
722
412
659

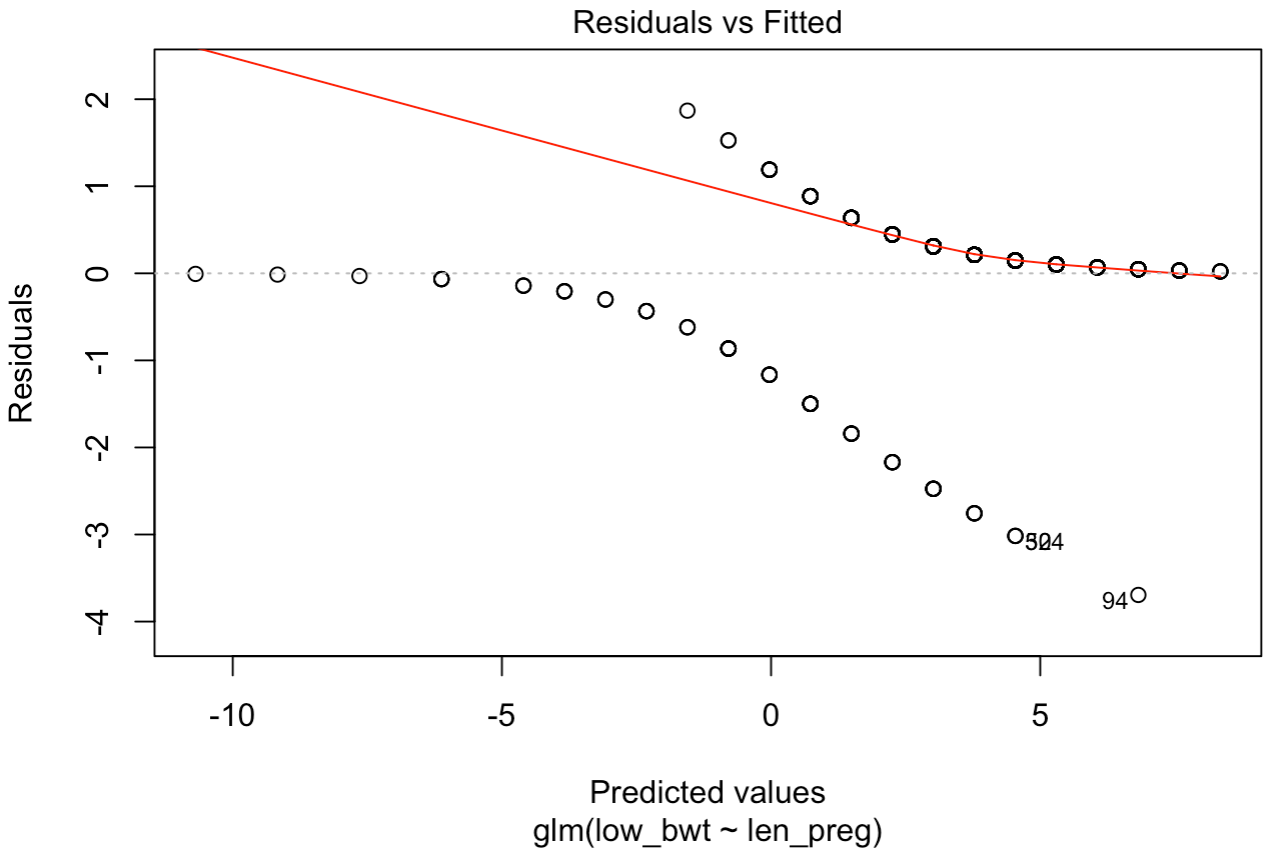

Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(bwt ~ (len_preg + marital + mom_wt_gain + sex + smoke + mom_white))
722
412
659

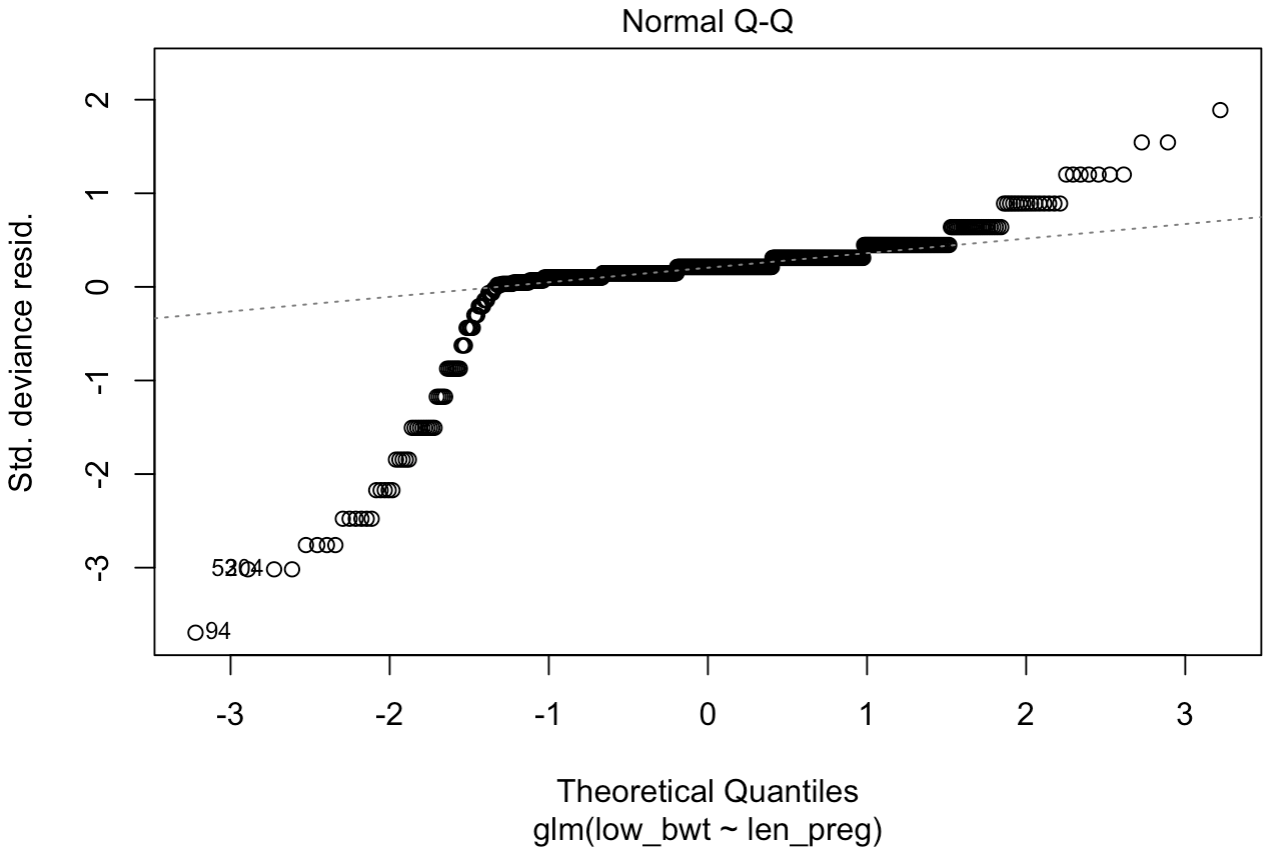

Scale-Location
659
722
412
√|Standardized residuals|
0.0
0.5
1.0
1.5
2
4
6
8
10
Fitted values
lm(bwt ~ (len_preg + marital + mom_wt_gain + sex + smoke + mom_white))

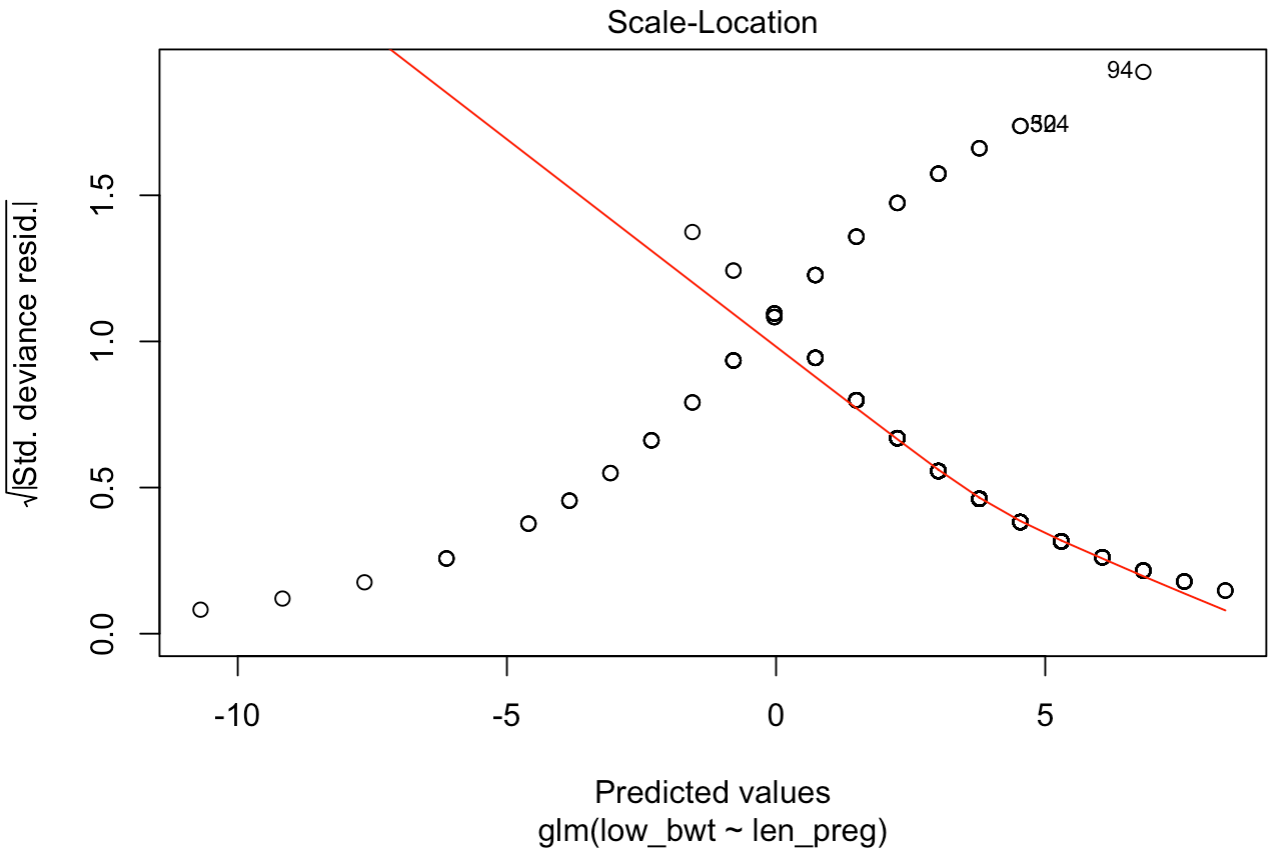

Residuals vs Leverage
Standardized residuals
2
0
-2
-4
176
262
37
Cook's distance
0.00
0.01
0.02
0.03
0.04
0.05
0.06
Leverage
lm(bwt ~ (len_preg + marital + mom_wt_gain + sex + smoke + mom_white))

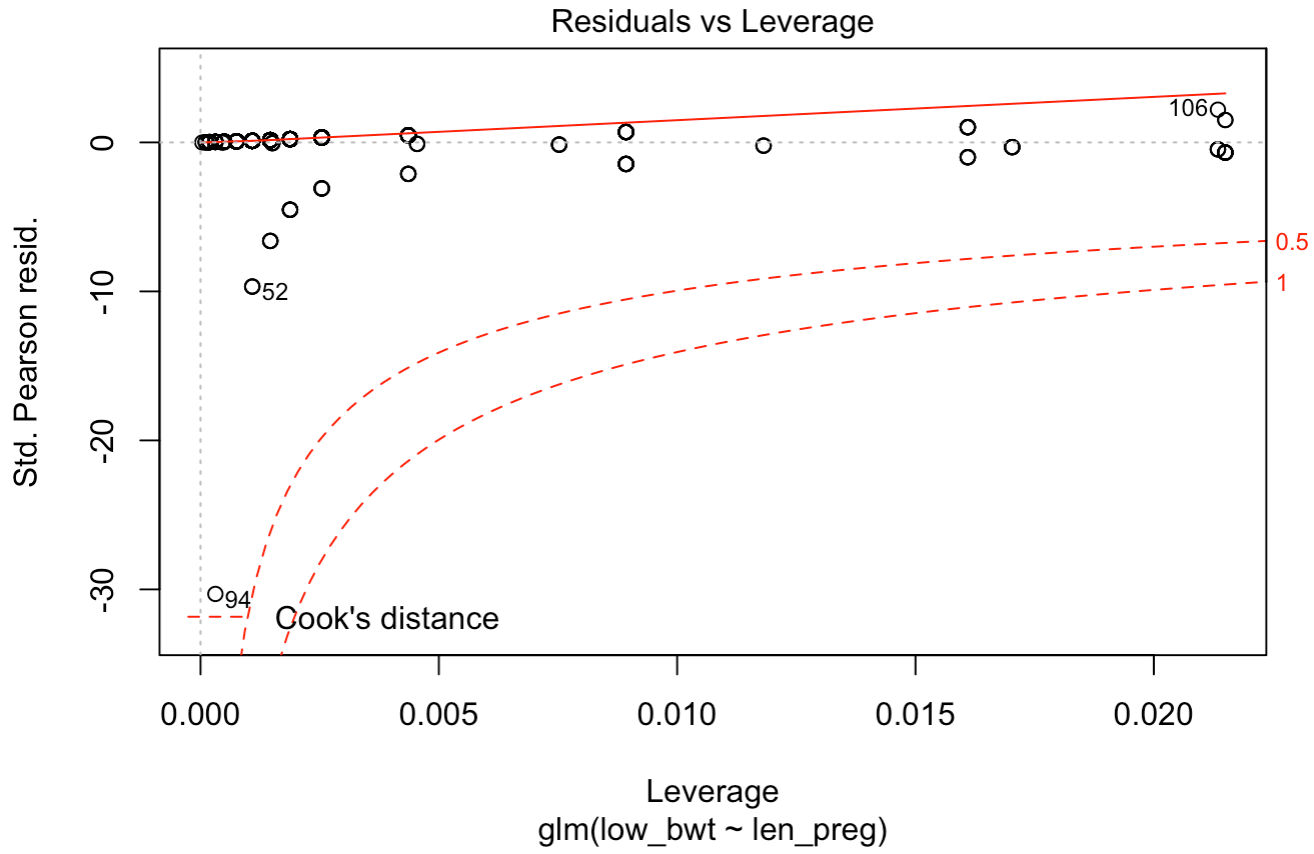
```
##
## Call:
## lm(formula = bwt ~ (len_preg + marital + mom_wt_gain + sex +
##     smoke + mom_white), data = births)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8255 -0.6567 -0.0219  0.6908  3.3043
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -5.702731   0.539536 -10.570  < 2e-16 ***
## len_preg          0.318111   0.013926  22.843  < 2e-16 ***
## maritalunmarried -0.251466   0.087742  -2.866  0.00427 **
## mom_wt_gain       0.012256   0.002867   4.274 2.16e-05 ***
## sexmale           0.365322   0.076352   4.785 2.05e-06 ***
## smokesmoker      -0.372548   0.125480  -2.969  0.00308 **
## mom_whitewhite    0.275294   0.094793   2.904  0.00379 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.064 on 774 degrees of freedom
## Multiple R-squared:  0.4555, Adjusted R-squared:  0.4513
## F-statistic: 107.9 on 6 and 774 DF,  p-value: < 2.2e-16
```

All predictor variables have p values below a cutoff of 0.05. The final linear model accounts for variation in bwt 45.13% of time. Although there is a curve to the fitted line, residuals are more randomly and evenly distributed. he final model shows that length of pregnancy has the highest p value of the predictor variables. Moreover displays that marital status, weight gain, smoking status, and ethnic identity play a role in predicting birth weight.


Residuals vs Fitted
Residuals
Predicted values
glm(low_bwt ~ len_preg)
94


Normal Q-Q
Std. deviance resid.
Theoretical Quantiles
glm(low_bwt ~ len_preg)
94


Scale-Location
√|Std. deviance resid.|
0.0
0.5
1.0
1.5
-10
-5
0
5
94
Predicted values
glm(low_bwt ~ len_preg)


Residuals vs Leverage
Std. Pearson resid.
0
-10
-20
-30
106
52
94
0.5
1
Cook's distance
0.000
0.005
0.010
0.015
0.020
Leverage
glm(low_bwt ~ len_preg)

```
##
## Call:
## glm(formula = low_bwt ~ len_preg, family = binomial(), data = births)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6942   0.0998   0.2128   0.3095   1.8685
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -25.92137    2.84075  -9.125   <2e-16 ***
## len_preg      0.76148    0.07778   9.790   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 480.43  on 780  degrees of freedom
## Residual deviance: 267.43  on 779  degrees of freedom
## AIC: 271.43
##
## Number of Fisher Scoring iterations: 6
```

The final model has one predictor variable (len_preg) with a p value well below the 0.05 cutoff. Moreover the p value for len_preg is very low a 2e-16. Other predictor variables fell out of the model due to the fact that there p values were too high to meet a cutoff of 0.05. Here we can see that the dichotomous variable low_bwt is best predicted by using len_preg. The final logistic model has a McFadden of 0.4433638 and a pseudo R^2 of 0.5195576