Final_Report

2019

Importing data into R

```
setwd ("/Users/[redacted]")

library(ggplot2)
library(rgl)
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
library(pwr)
library(pscl)
```

```
## Classes and Methods for R developed in the
## Political Science Computational Laboratory
## Department of Political Science
## Stanford University
## Simon Jackman
## hurdle and zeroinfl functions by Achim Zeileis
```

```
library(car)
```

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

```
logisticPseudoR2s <- function(LogModel) {
  dev <- LogModel$deviance
  nullDev <- LogModel$null.deviance
  modelN <-  length(LogModel$fitted.values)
  R.l <-  1 -  dev / nullDev
  R.cs <- 1- exp ( -(nullDev - dev) / modelN)
  R.n <- R.cs / ( 1 - ( exp (-(nullDev / modelN))))
  cat("Pseudo R^2 for logistic regression\n")
  cat("Hosmer and Lemeshow R^2  ", round(R.l, 3), "\n")
  cat("Cox and Snell R^2        ", round(R.cs, 3), "\n")
  cat("Nagelkerke R^2           ", round(R.n, 3),    "\n")
}


load("births.rdat")
summary(births)
```

```
##     dad_age         mom_age      maturity      len_preg
##  Min.   : 14.0   Min.   :13   advnced:133   Min.   : 20.00
##  1st Qu.: 26.0   1st Qu.:22   younger:867   1st Qu.: 37.00
##  Median : 32.0   Median :27                 Median : 39.00
##  Mean   :195.9   Mean   :27                 Mean   : 40.26
##  3rd Qu.: 39.0   3rd Qu.:32                 3rd Qu.: 40.00
##  Max.   :999.0   Max.   :50                 Max.   :999.00
##     is_premie     num_visits          marital     mom_wt_gain
##  fullterm:846   Min.   :  0.00   married  :613   Min.   :  0.00
##  premie  :152   1st Qu.: 10.00   unknown  :  1   1st Qu.: 21.00
##  unknown :  2   Median : 12.00   unmarried:386   Median : 30.00
##                 Mean   : 20.99                   Mean   : 56.48
##                 3rd Qu.: 15.00                   3rd Qu.: 40.00
##                 Max.   :999.00                   Max.   :999.00
##       bwt           low_bwt        sex            smoke        mom_white
##  Min.   : 1.000   low   :111   female:503   nonsmoker:873   nonwhite:284
##  1st Qu.: 6.380   notlow:889   male  :497   smoker   :126   unknown :  2
##  Median : 7.310                            unknown  :  1   white   :714
##  Mean   : 7.101
##  3rd Qu.: 8.060
##  Max.   :11.750
##   mom_age_level
##  35+     :133
##  early20s:281
##  early30s:219
##  late20s :257
##  teens   :110
##
```

Data Cleaning

```
# Cleaning impossible values
births$dad_age[births$dad_age == 999] <- NA
births$len_preg[births$len_preg == 999] <- NA
births$is_premie[births$is_premie == "unknown"] <- NA
births$num_visits[births$num_visits == 999] <- NA
births$marital[births$marital == "unknown"] <- NA
births$mom_wt_gain[births$mom_wt_gain == 999] <- NA
births$smoke[births$smoke == "unknown"] <- NA
births$mom_white[births$mom_white == "unknown"] <- NA

# Drop unused levels
births$is_premie <- droplevels(births$is_premie)
births$marital <- droplevels(births$marital)
births$smoke <- droplevels(births$smoke)
births$mom_white <- droplevels(births$mom_white)

# Reordering levels for correct visualizations

births$maturity <- factor(births$maturity, levels = c("younger","advnced"))
births$mom_age_level <- factor( births$mom_age_level, levels = c("teens","early20
s","late20s","early30s","35+") )
births$is_premie <- factor(births$is_premie, levels = c("premie","fullterm"))
```

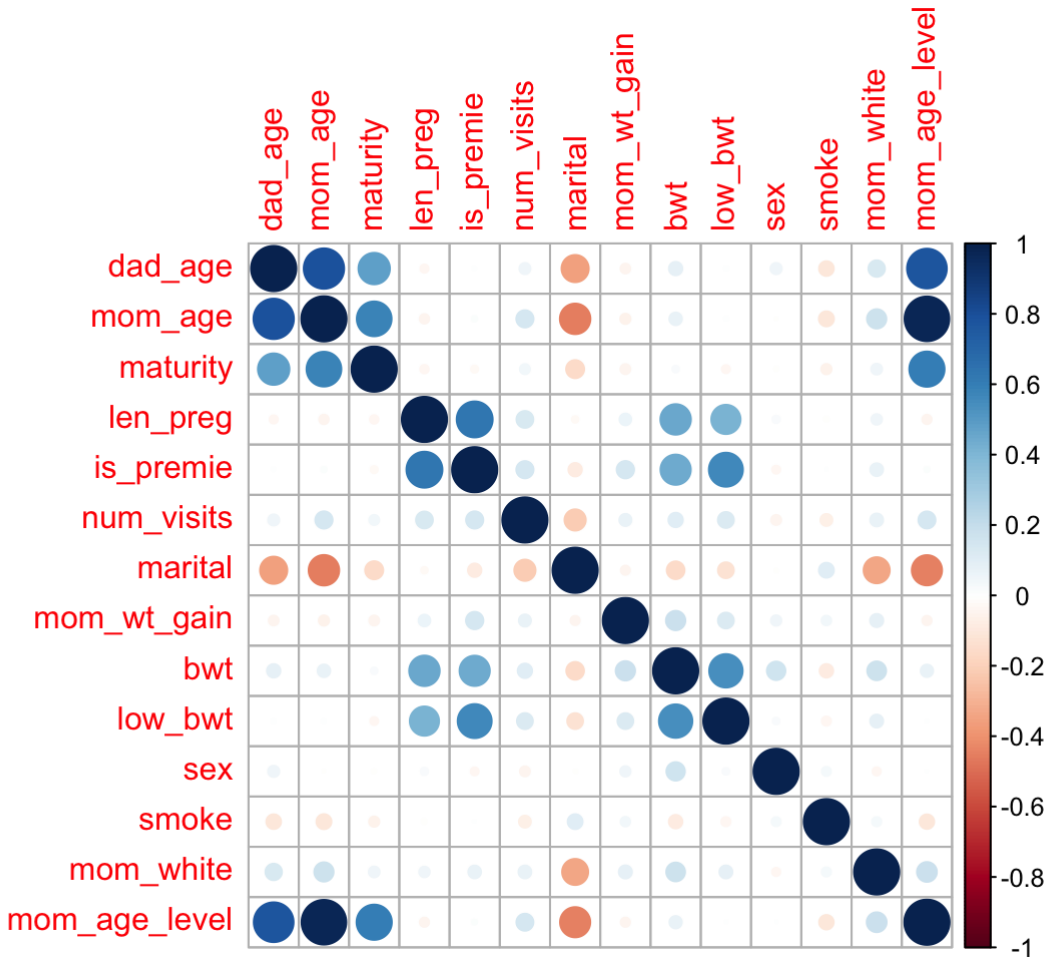
EDA

```
# Summary Statistics : mean, median & quantiles included in summary.for categorica
l variable counts are included too.
summary(births)
```

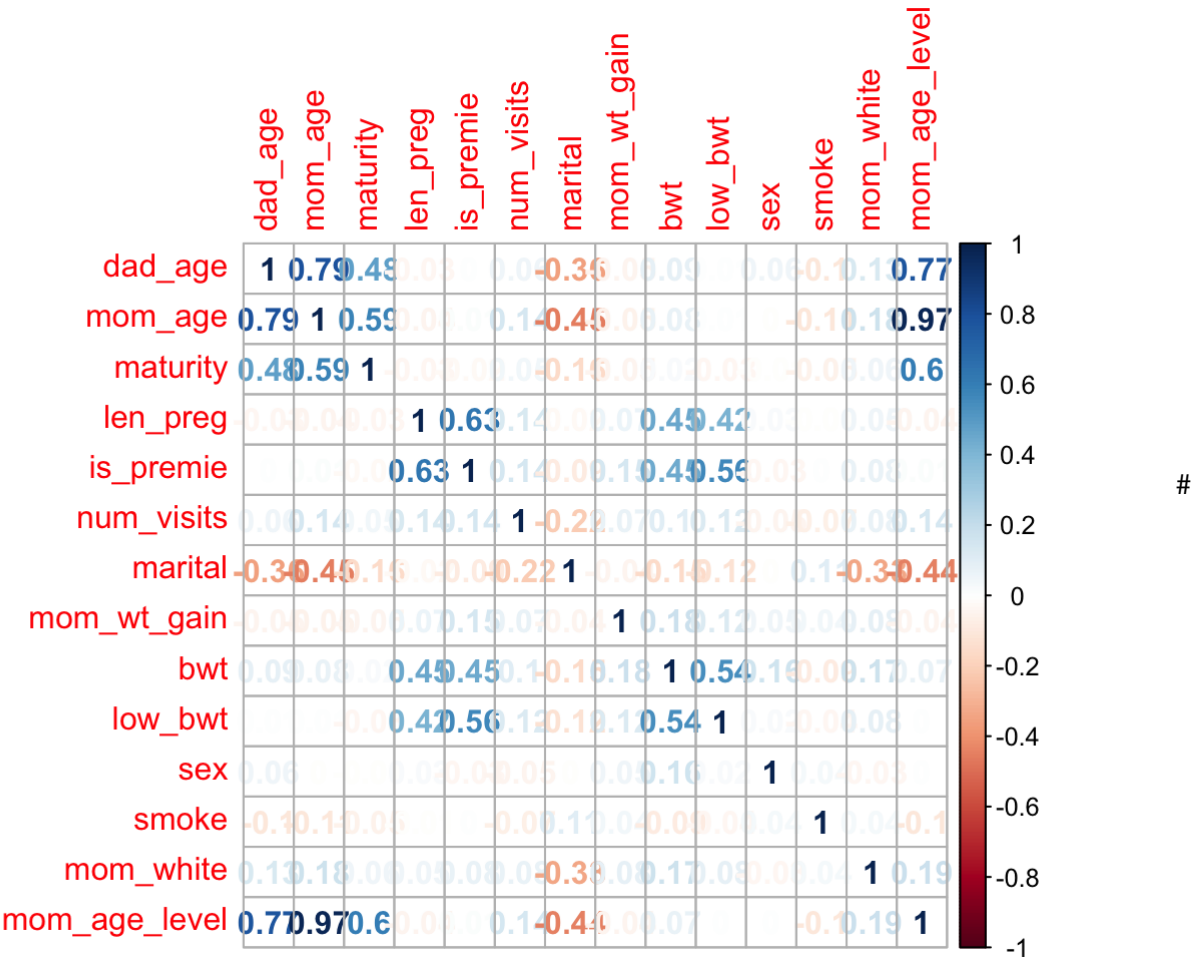
```
##     dad_age         mom_age      maturity      len_preg        is_premie
##  Min.   :14.00   Min.   :13   younger:867   Min.   :20.00   premie  :152
##  1st Qu.:25.00   1st Qu.:22   advnced:133   1st Qu.:37.00   fullterm:846
##  Median :30.00   Median :27                 Median :39.00   NA's    :  2
##  Mean   :30.26   Mean   :27                 Mean   :38.33
##  3rd Qu.:35.00   3rd Qu.:32                 3rd Qu.:40.00
##  Max.   :55.00   Max.   :50                 Max.   :45.00
##  NA's   :171                                NA's   :2
##    num_visits        marital     mom_wt_gain         bwt
##  Min.   : 0.0   married  :613   Min.   : 0.00   Min.   : 1.000
##  1st Qu.:10.0   unmarried:386   1st Qu.:20.00   1st Qu.: 6.380
##  Median :12.0   NA's     :  1   Median :30.00   Median : 7.310
##  Mean   :12.1                   Mean   :30.33   Mean   : 7.101
##  3rd Qu.:15.0                   3rd Qu.:38.00   3rd Qu.: 8.060
##  Max.   :30.0                   Max.   :85.00   Max.   :11.750
##  NA's   :9                      NA's   :27
##    low_bwt        sex            smoke        mom_white    mom_age_level
##  low   :111   female:503   nonsmoker:873   nonwhite:284   teens   :110
##  notlow:889   male  :497   smoker   :126   white   :714   early20s:281
##                            NA's     :  1   NA's    :  2   late20s :257
##                                                          early30s:219
##                                                          35+     :133
##
##
```

Correlation

```
births_numeric <- births[,c(1:14)]
births_numeric$maturity <- as.numeric(births$maturity )
births_numeric$is_premie <- as.numeric(births$is_premie )
births_numeric$marital <- as.numeric(births$marital )
births_numeric$low_bwt <- as.numeric(births$low_bwt )
births_numeric$sex <- as.numeric(births$sex )
births_numeric$smoke <- as.numeric(births$smoke )
births_numeric$mom_white <- as.numeric(births$mom_white)
births_numeric$mom_age_level <- as.numeric(births$mom_age_level)

c <- cor(births_numeric, use = "pairwise.complete.obs", method = "spearman")
corrplot(c)
```




```
corrplot(c, method="number")
```

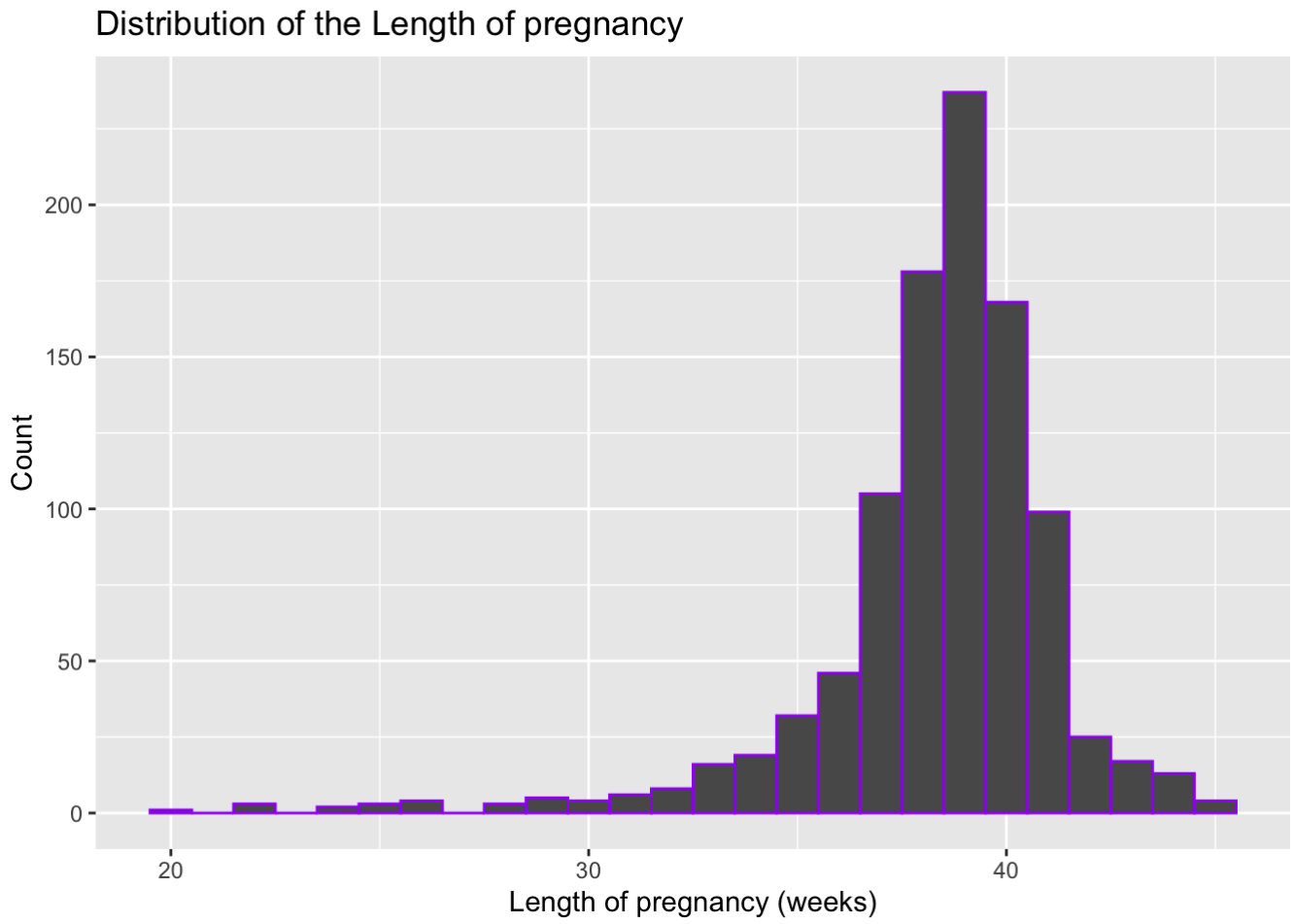


#

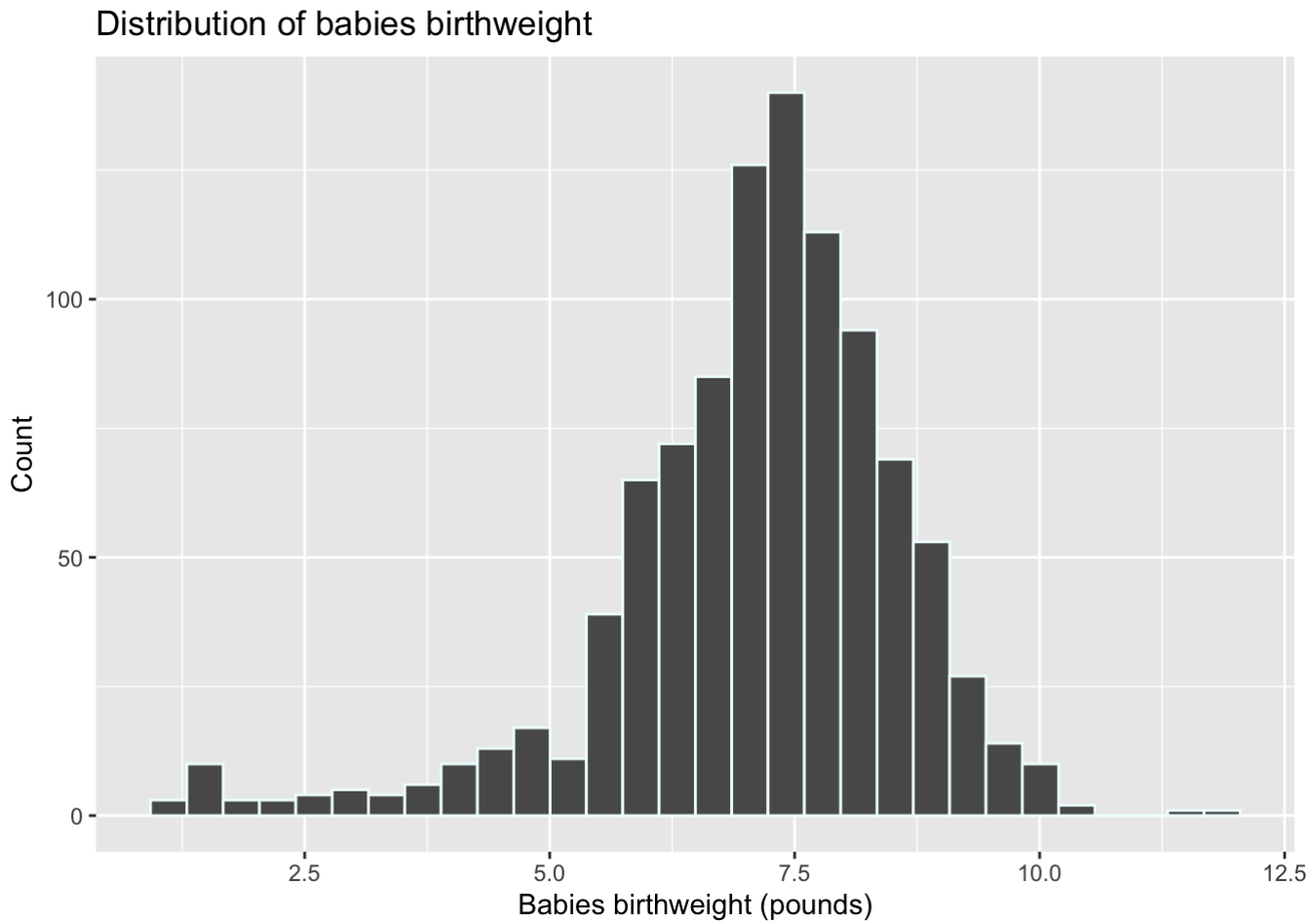
Data Visualizations

```
g4 = ggplot(births, aes(x=len_preg))
g4 + geom_histogram(binwidth = 1,color="purple") + ggtitle("Distribution of the Le
ngth of pregnancy") +
  xlab("Length of pregnancy (weeks)") + ylab("Count")
```

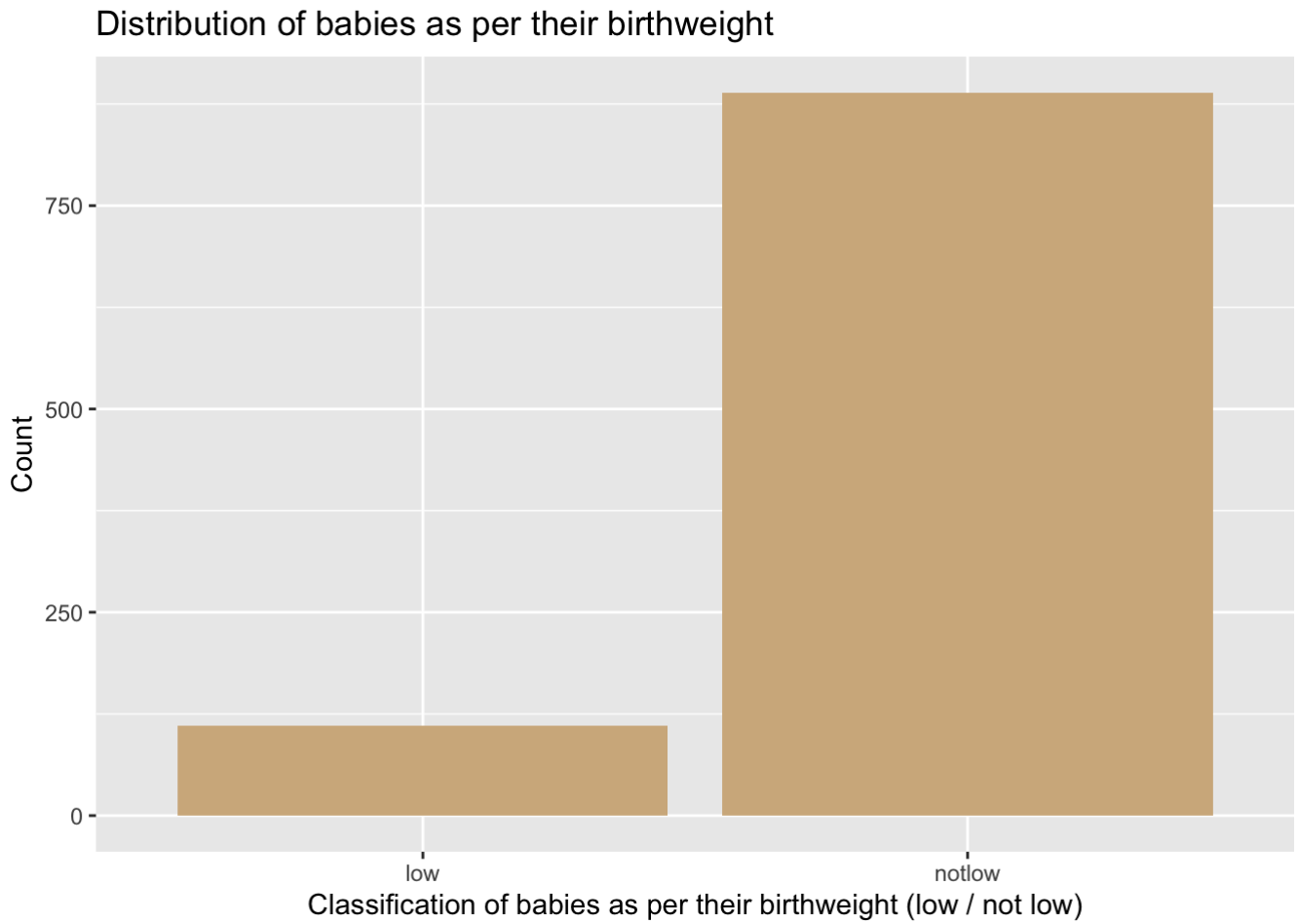
```
## Warning: Removed 2 rows containing non-finite values (stat_bin).
```



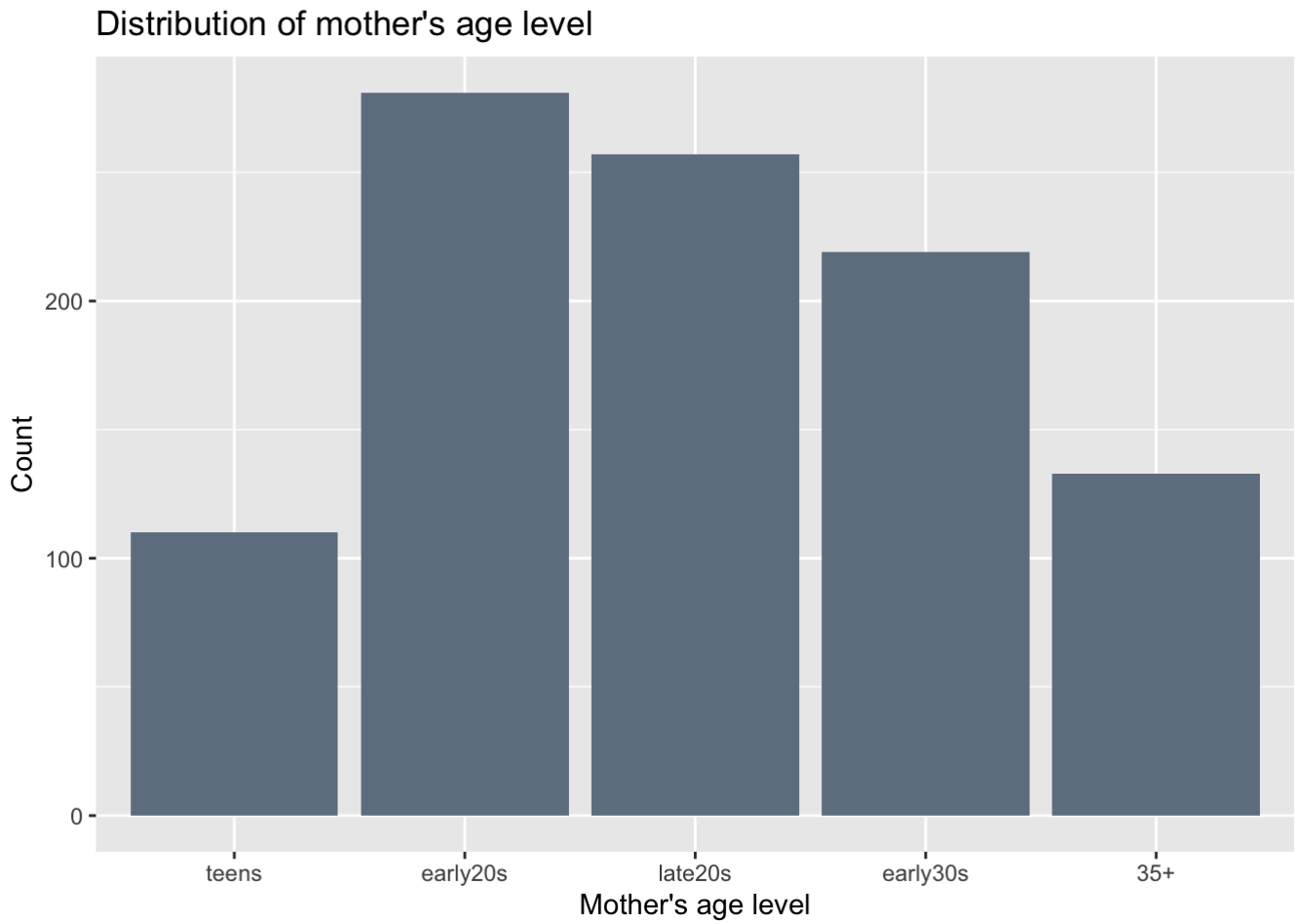

```
g9 = ggplot(births, aes(x=bwt))
g9 + geom_histogram(bins = 30, color = "azure") + ggtitle("Distribution of babies
 birthweight") +
  xlab("Babies birthweight (pounds)") + ylab("Count")
```



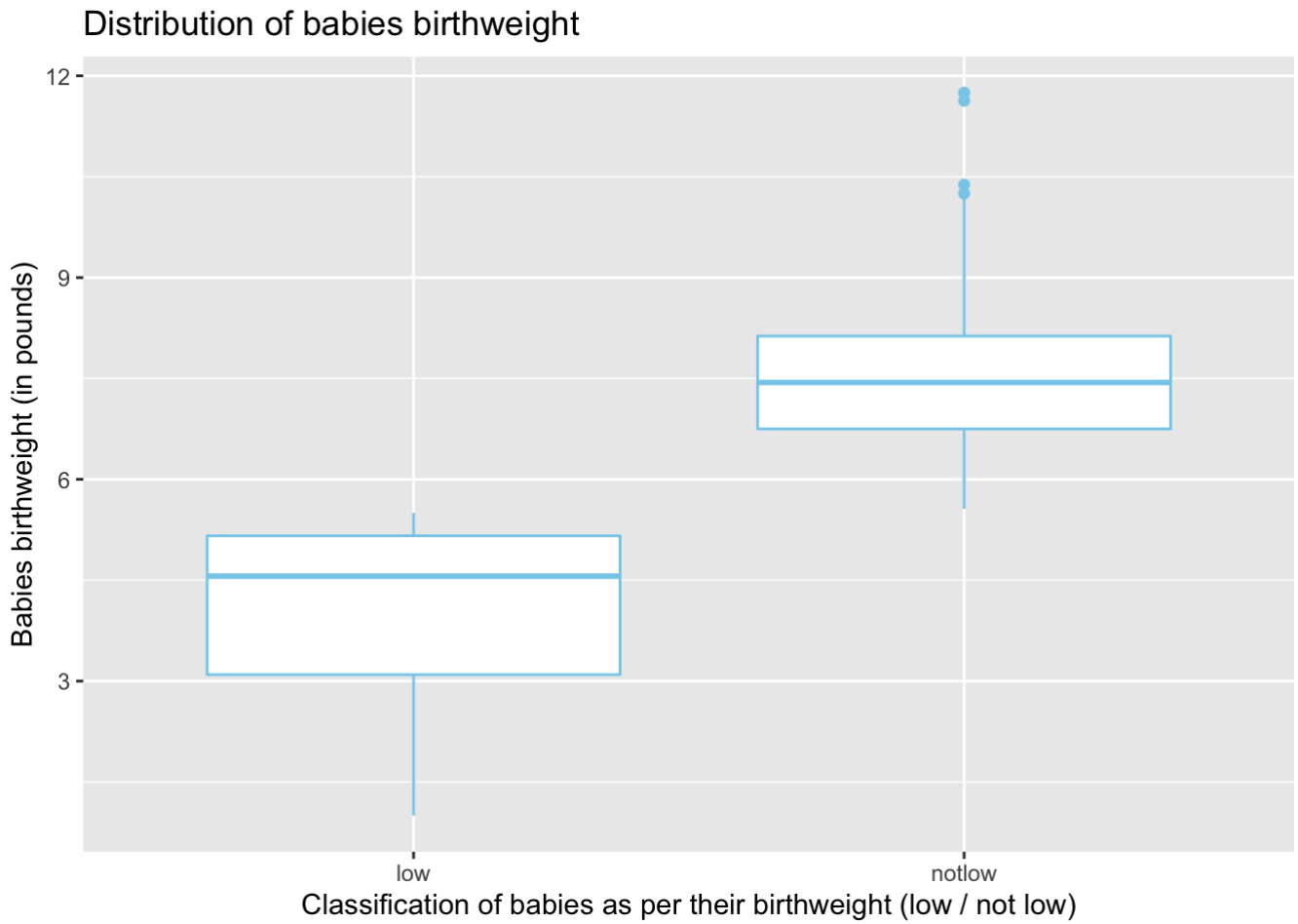

```
g10 = ggplot(births, aes(x=low_bwt))
g10 + geom_bar(fill = "tan")+ ggtitle("Distribution of babies as per their birthwe
ight") +
  xlab("Classification of babies as per their birthweight (low / not low)") + ylab
("Count")
```




```
g14 = ggplot(births, aes(x=mom_age_level))
g14 + geom_bar(fill = "slategrey")+ ggtitle("Distribution of mother's age level")
 +
  xlab("Mother's age level") + ylab("Count")
```



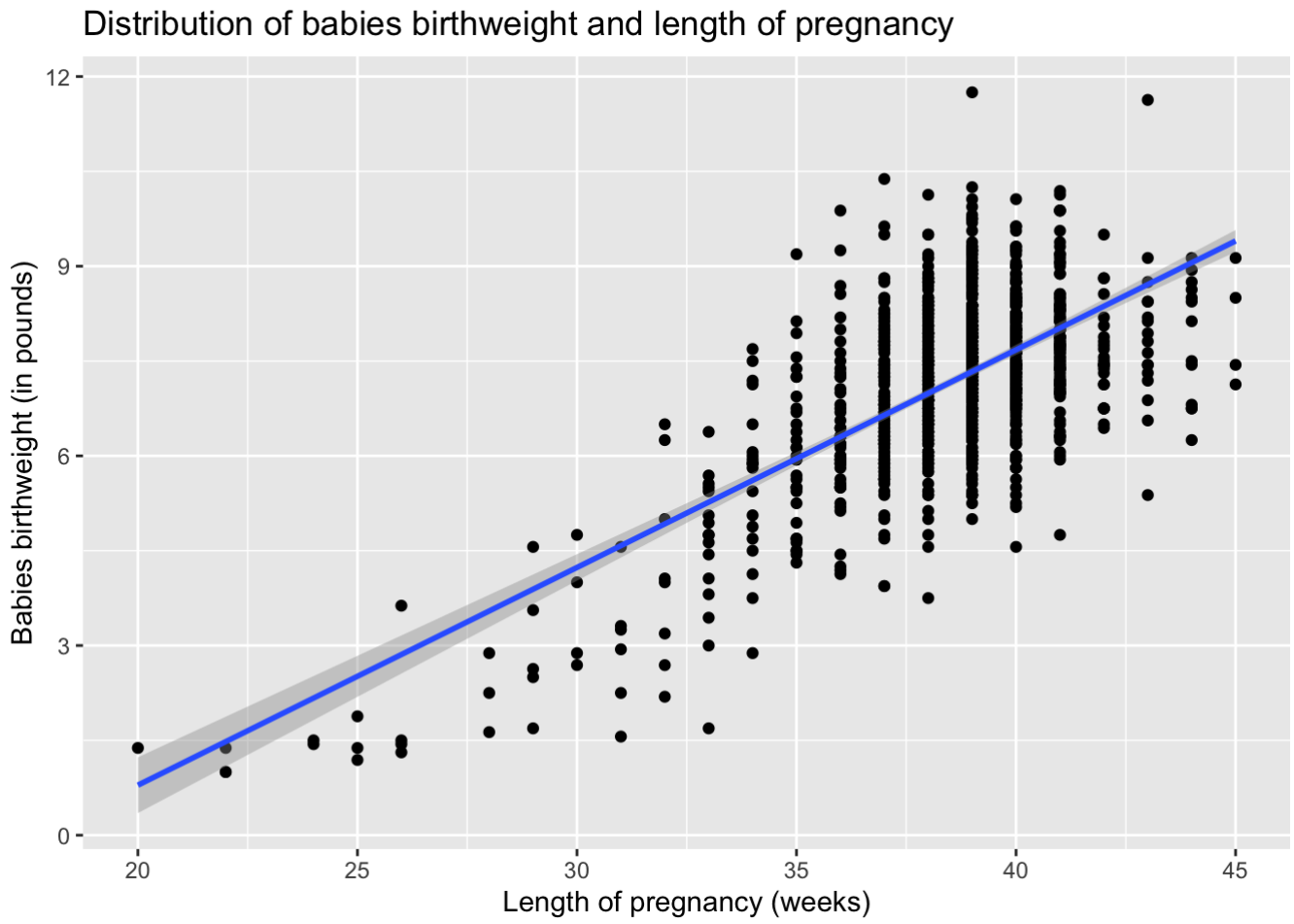

```
# the sample has random selectipn of mother's accross all age levels

g15 = ggplot(births, aes(x=low_bwt, y=bwt))
g15 + geom_boxplot(color="skyblue")+ ggtitle("Distribution of babies birthweight")
+
  xlab("Classification of babies as per their birthweight (low / not low)") + ylab
("Babies birthweight (in pounds)")
```



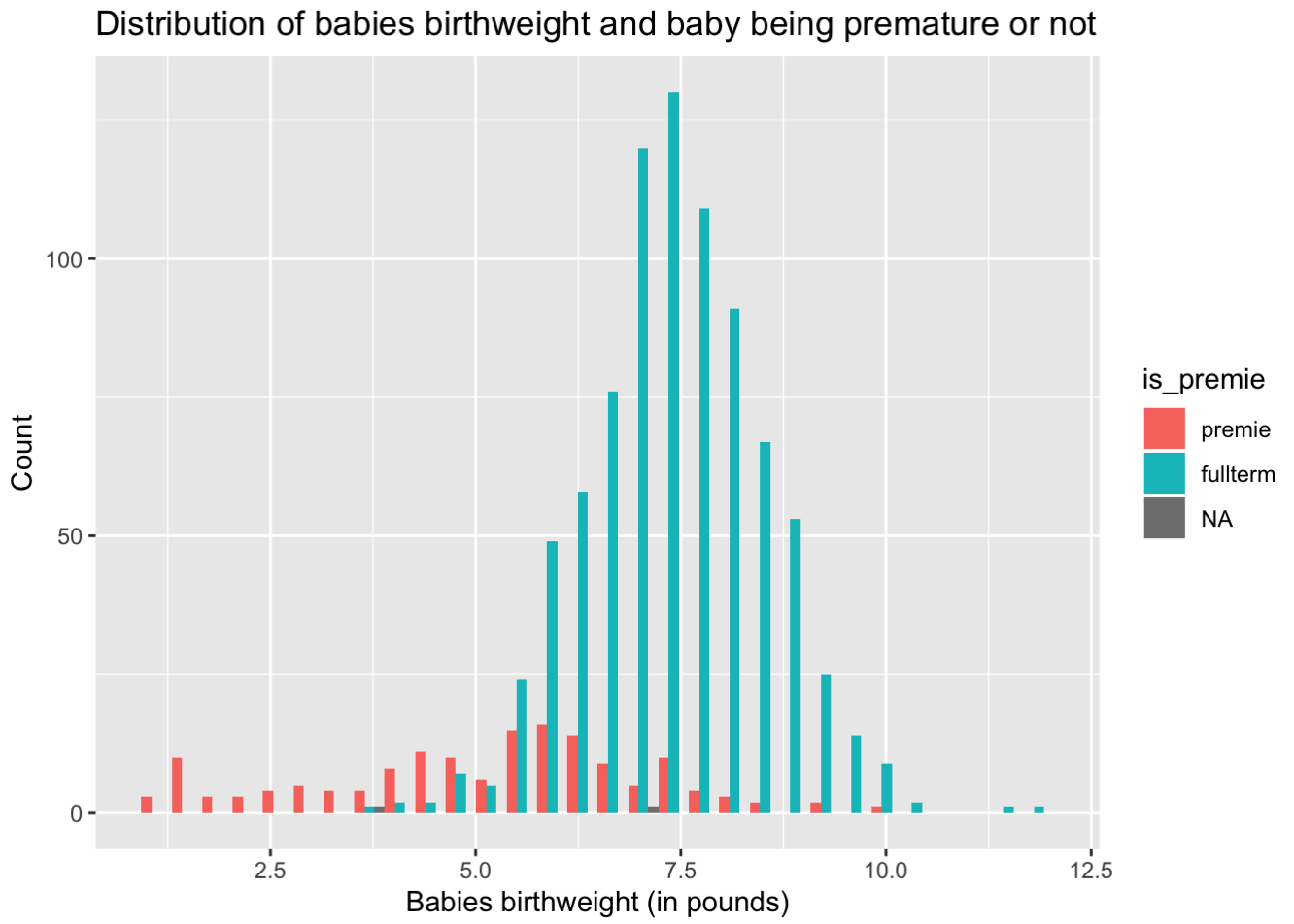

```
# plot implies that the low_bwt variable may be a derived column from the bwt valu
es

g16 = ggplot(births, aes(x=len_preg, y= bwt))
g16 + geom_point()+ geom_smooth(method=lm) + ggtitle("Distribution of babies birth
weight and length of pregnancy") +
  xlab("Length of pregnancy (weeks)") + ylab("Babies birthweight (in pounds)")
```

```
## Warning: Removed 2 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```

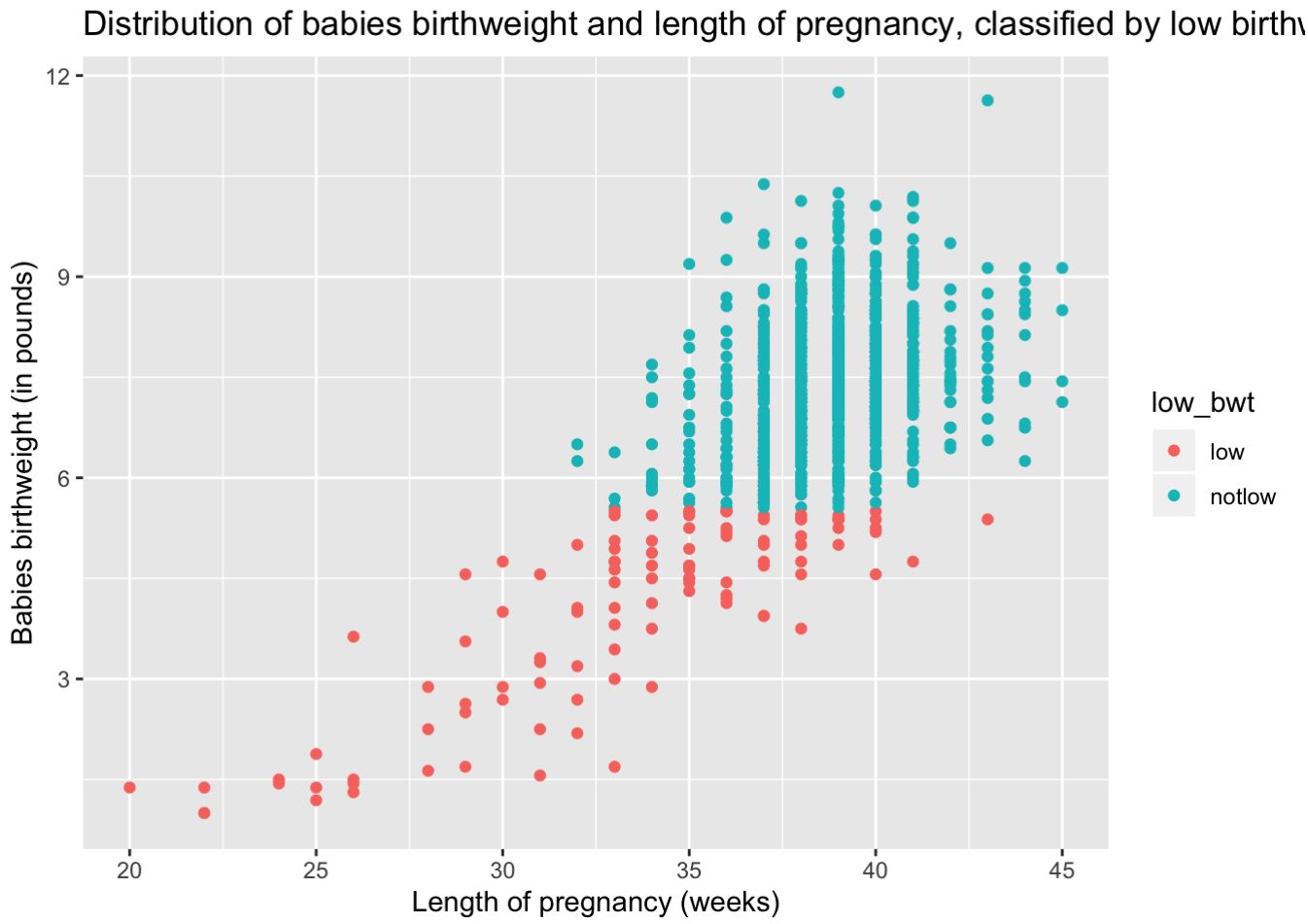



```
# plot implies that there is positive correlation between the length of pregnancy
 and babies birthweight
# the more the length of pregnancy, more will be the babies birthweight

g17 = ggplot(births, aes(x=bwt, fill=is_premie))
g17 + geom_histogram(bins=30,position="dodge")+ ggtitle("Distribution of babies bi
rthweight and baby being premature or not") +
  xlab("Babies birthweight (in pounds)") + ylab("Count")
```



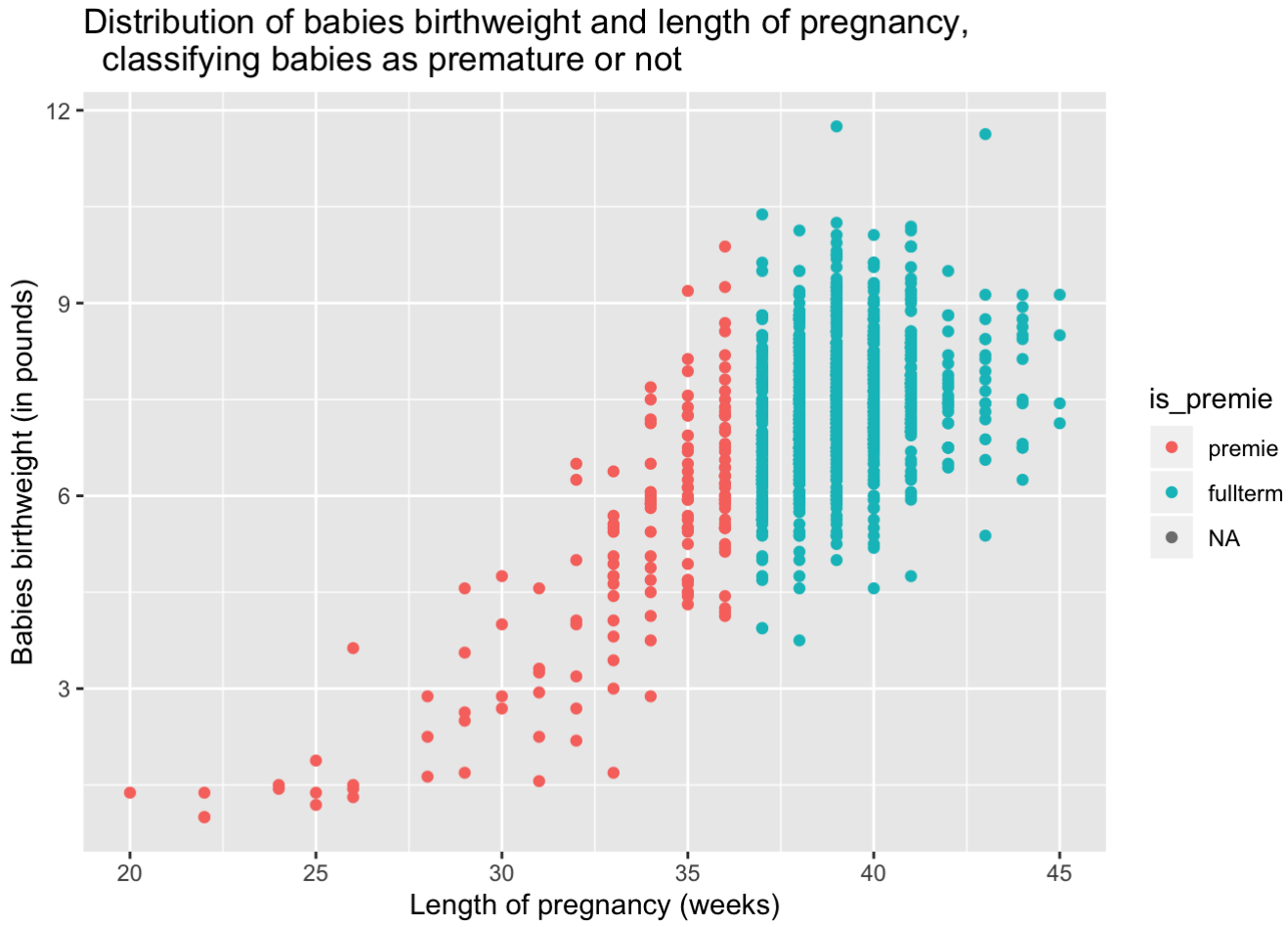

```
# plot implies that there is positive correlation between the babies birthweight a
nd the baby being premature or fullterm
# the fullterm babies will have more birthweight, than the premature babies with l
ower birthweights

g18 = ggplot(births, aes(x=len_preg, y= bwt, col=low_bwt))
g18 + geom_point() + ggtitle("Distribution of babies birthweight and length of pre
gnancy, classified by low birthweight or not") +
  xlab("Length of pregnancy (weeks)") + ylab("Babies birthweight (in pounds)")
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```




```
# plot implies that all the babies delivered after a longer length of pregnancy, h
ave a higher birthweight
# all the babies delivered after a short length of pregnancy, have a low birthweig
ht
# all babies with low birthweights can be classified as low_bwt babies

g19 = ggplot(births, aes(x=len_preg, y= bwt, col=is_premie))
g19 + geom_point() + ggtitle("Distribution of babies birthweight and length of pre
gnancy,
  classifying babies as premature or not") + xlab("Length of pregnancy (weeks)") +
  ylab("Babies birthweight (in pounds)")
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```




```
# plot implies that all the babies delivered after a longer length of pregnancy, h
ave a higher birthweight
# and are fullterm babies
# all babies with shorter length of pregnancy and low birthweights are premature b
abies
```

Statistical Tests

```
# T-test :
t.test(births$bwt~births$is_premie)
```

```
##
##  Welch Two Sample t-test
##
## data:  births$bwt by births$is_premie
## t = -14.216, df = 167.51, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.654709 -2.007267
## sample estimates:
##   mean in group premie mean in group fullterm
##               5.128421               7.459409
```

```
# conclusions from the t test:
# p-value less than significance level=0.05
# and there is a significant difference in means of both groups : birthweight for
 premature and fullterm babies

table(births$is_premie)   # n1=152, n2=846
```

```
##
##   premie fullterm
##      152      846
```

```
pwr.t2n.test(n1=152, n2=846, d=0.2, sig.level = .05 )
```

```
##
##      t test power calculation
##
##              n1 = 152
##              n2 = 846
##               d = 0.2
##       sig.level = 0.05
##           power = 0.6210034
##     alternative = two.sided
```

```
# small effect size can be detected and the power of the study is 80%
# indicates results being statistically significant

# one-way ANOVA:
summary(aov(births$bwt~births$mom_age_level))
```

```
##                       Df Sum Sq Mean Sq F value Pr(>F)
## births$mom_age_level   4   14.8   3.699   1.629  0.165
## Residuals            995 2259.6   2.271
```

```
# null hypothesis is that the mean bwt of baby is same for all groups in mom's age levels
# the p-value = 0.165, is more than significance level of 0.05
# we don't have enough evidence to reject the null hypothesis
# concluding that there isn't much impact of mother's age level on the babies birt hweight

# Chisq Test:
chisq.test(table(births$low_bwt,births$is_premie))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(births$low_bwt, births$is_premie)
## X-squared = 311.57, df = 1, p-value < 2.2e-16
```

```
# null hypothesis: Both low_bwt and is_premie are independent variables
# the p-value is < 2.2e-16, which is less than the significance level of 0.05
# we have enough evidence to reject the null hypothesis
# and conclude that low_bwt and is_premie are not independent variables
```

Linear Regression model

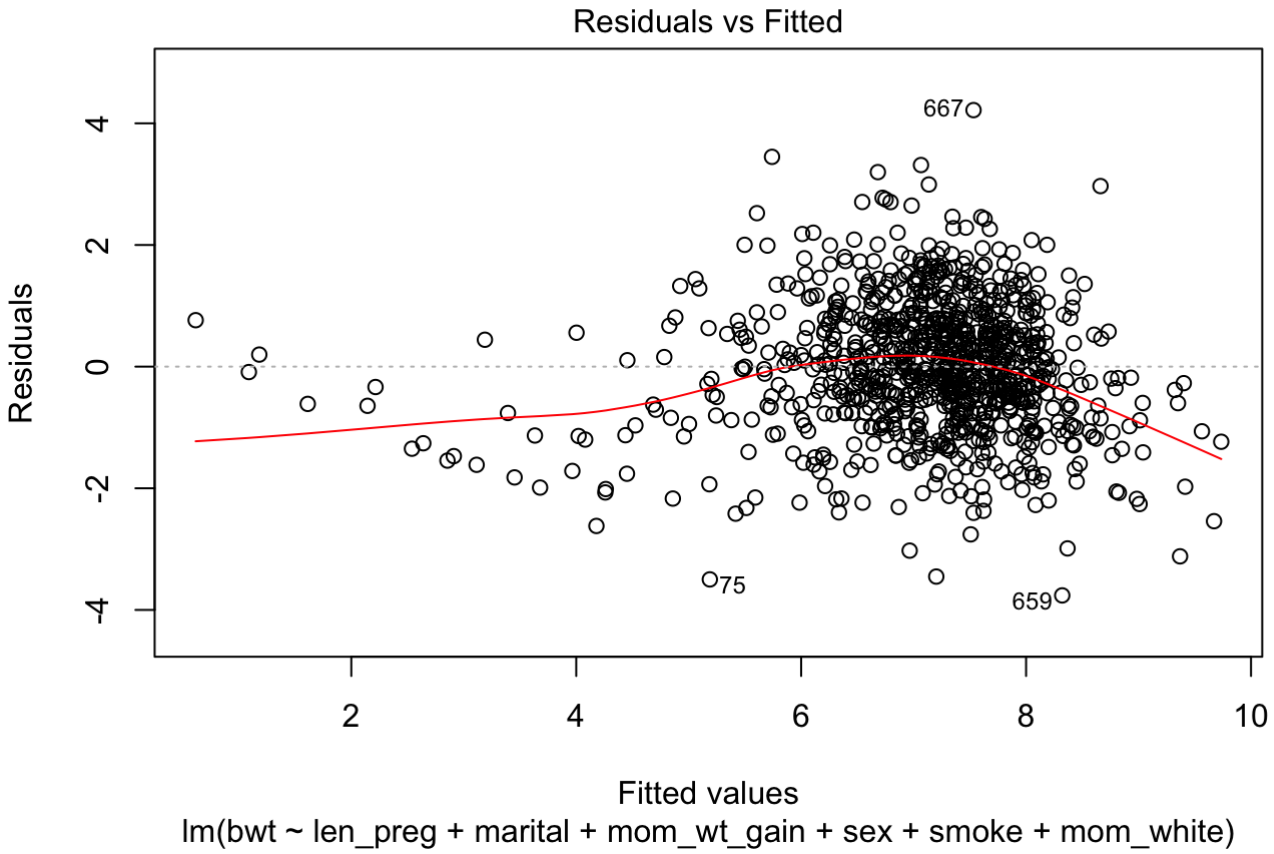
```
fit_all_lm <- lm(bwt ~ dad_age + mom_age + maturity + len_preg + is_premie + num_v isits + marital + mom_wt_gain + low_bwt +
                 sex + smoke + mom_white + mom_age_level, data = births)
summary(fit_all_lm)
```

```
##
## Call:
## lm(formula = bwt ~ dad_age + mom_age + maturity + len_preg +
##     is_premie + num_visits + marital + mom_wt_gain + low_bwt +
##     sex + smoke + mom_white + mom_age_level, data = births)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4770 -0.6012 -0.0021  0.5494  3.5563
##
## Coefficients: (1 not defined because of singularities)
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -2.6503561  0.7170302  -3.696 0.000234 ***
## dad_age                0.0080555  0.0077365   1.041 0.298087
## mom_age                0.0090994  0.0218822   0.416 0.677644
## maturityadvnced       -0.0731887  0.4196130  -0.174 0.861581
## len_preg               0.1718275  0.0181833   9.450  < 2e-16 ***
## is_premiefullterm      0.0442754  0.1440740   0.307 0.758689
## num_visits             0.0003972  0.0089395   0.044 0.964572
## maritalunmarried      -0.1172818  0.0853089  -1.375 0.169590
## mom_wt_gain            0.0071629  0.0023449   3.055 0.002330 **
## low_bwtnotlow          2.3881336  0.1431235  16.686  < 2e-16 ***
## sexmale                0.3821920  0.0653943   5.844 7.45e-09 ***
## smokesmoker           -0.2143158  0.1068650  -2.005 0.045255 *
## mom_whitewhite         0.2458025  0.0807928   3.042 0.002425 **
## mom_age_levelearly20s  0.0825564  0.1538582   0.537 0.591714
## mom_age_levellate20s   0.0187617  0.2309455   0.081 0.935273
## mom_age_levelearly30s -0.0630545  0.3222707  -0.196 0.844929
## mom_age_level35+              NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9136 on 784 degrees of freedom
##   (200 observations deleted due to missingness)
## Multiple R-squared:  0.6055, Adjusted R-squared:  0.598
## F-statistic: 80.24 on 15 and 784 DF,  p-value: < 2.2e-16
```

```
# there are variables that have high correlation with other variables, which might cause multicollinearity
# variables highly correlated (from correlation test above) are :
# maturity with mom_age, is_premie iwth len_preg, low_bwt with bwt and mom_age_levels with mom_age
# removing variables that are derived values from other variables: maturity, is_premie, low_bwt, mom_age_levels

# final linear model derived and is statitically significant is as follows:
final_fit_lm <- lm(bwt ~ len_preg + marital + mom_wt_gain + sex + smoke + mom_white, data = births)
summary(final_fit_lm)
```

```
##
## Call:
## lm(formula = bwt ~ len_preg + marital + mom_wt_gain + sex + smoke +
##     mom_white, data = births)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7613 -0.6582 -0.0263  0.6872  4.2173
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -5.977147   0.462464 -12.925  < 2e-16 ***
## len_preg          0.329042   0.011961  27.509  < 2e-16 ***
## maritalunmarried -0.265136   0.075729  -3.501 0.000485 ***
## mom_wt_gain       0.009258   0.002431   3.809 0.000149 ***
## sexmale           0.378405   0.068971   5.486 5.24e-08 ***
## smokesmoker      -0.388843   0.104758  -3.712 0.000218 ***
## mom_whitewhite    0.212131   0.081324   2.608 0.009236 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.071 on 963 degrees of freedom
##   (30 observations deleted due to missingness)
## Multiple R-squared:  0.4869, Adjusted R-squared:  0.4837
## F-statistic: 152.3 on 6 and 963 DF,  p-value: < 2.2e-16
```

```
# all variables have p-value less than 0.05 and their adjusted R squared=0.48
# indicating a good fit model, that will explain 48% variation of bwt (response variable) from the total explainable variation,
# when all other variables are held constant

plot(final_fit_lm, 1)
```




```
# the residual values are randomly distributed below & above the fitted line.
# the fit line is close to the regression line


###### Brief Interpretation of the model #####
# Birthweight of the baby is dependent on the variables : length of pregnancy(len_
preg), marital status of the mother,
# Mother's weight gain before delivery, sex of the baby, mother's smoking status &
mother being white or not.
# linear model: final model's fitted equation:
# bwt =  0.329042 * len_preg - 0.265136 * marital + 0.009258 * mom_wt_gain + 0.378
405 * sex - 0.388843 * smoke + 0.212131 * mom_white - 5.977147
## For every 1 week increase in the length of pregnancy, the birthweight of the ba
by increases by 0.329042 pounds on average, when all other variables in the model
 are held constant
## For married mothers as compared with unmarried mother's, the birthweight of the
baby decreases by 0.265136 pounds on average, when all other variables in the mode
l are held constant
```

Logistic Regression model

```
fit_glm_all <- glm(low_bwt ~ ., family = binomial(), data = births)
```

```
## Warning: glm.fit: algorithm did not converge
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(fit_glm_all)
```

```
##
## Call:
## glm(formula = low_bwt ~ ., family = binomial(), data = births)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -2.778e-04   2.100e-08   2.100e-08   2.100e-08   2.431e-04
##
## Coefficients: (1 not defined because of singularities)
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -9.264e+02  1.375e+05  -0.007    0.995
## dad_age               -3.359e+00  9.207e+02  -0.004    0.997
## mom_age               -3.718e+00  2.668e+03  -0.001    0.999
## maturityadvnced        1.473e+02  6.438e+04   0.002    0.998
## len_preg               3.282e+00  2.690e+03   0.001    0.999
## is_premiefullterm     -2.840e+01  1.373e+04  -0.002    0.998
## num_visits            -1.576e+00  1.079e+03  -0.001    0.999
## maritalunmarried      -1.859e+01  6.864e+03  -0.003    0.998
## mom_wt_gain           -5.582e-02  2.668e+02   0.000    1.000
## bwt                    1.758e+02  1.299e+04   0.014    0.989
## sexmale                5.662e+00  5.273e+03   0.001    0.999
## smokesmoker           -6.726e+00  6.180e+03  -0.001    0.999
## mom_whitewhite        -1.349e+01  7.363e+03  -0.002    0.999
## mom_age_levelearly20s  3.415e+01  1.116e+04   0.003    0.998
## mom_age_levellate20s   9.080e+01  2.875e+04   0.003    0.997
## mom_age_levelearly30s  1.363e+02  4.292e+04   0.003    0.997
## mom_age_level35+              NA         NA      NA       NA
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4.9325e+02  on 799  degrees of freedom
## Residual deviance: 4.5427e-07  on 784  degrees of freedom
##   (200 observations deleted due to missingness)
## AIC: 32
##
## Number of Fisher Scoring iterations: 25
```

```
# summary shows that there are variables in related to the model indicating a perf
ect linear separation
# the fitted coefficients are not statistically significant to explain the model.
pR2(fit_glm_all)
```

```
##           llh       llhNull            G2      McFadden          r2ML
## -2.271326e-07 -3.486010e+02  6.972020e+02  1.000000e+00  5.816774e-01
##          r2CU
##  1.000000e+00
```

```
# Both McFadden &  r2CU values are 1, indicating a model with perfect fit
# which is not possible

# there are variables that have high correlation with other variables, which might
cause multicollinearity
# variables highly correlated (from correlation test above) are :
# maturity with mom_age, is_premie iwth len_preg, low_bwt with bwt and mom_age_lev
els with mom_age
# removing variables that are derived values from other variables: maturity, is_pr
emie, low_bwt, mom_age_levels
# we arrive at our final interpretable logistic regression model:

fit_final_glm <- glm(low_bwt ~ len_preg + marital,family = binomial(), data = birt
hs)
summary(fit_final_glm)
```

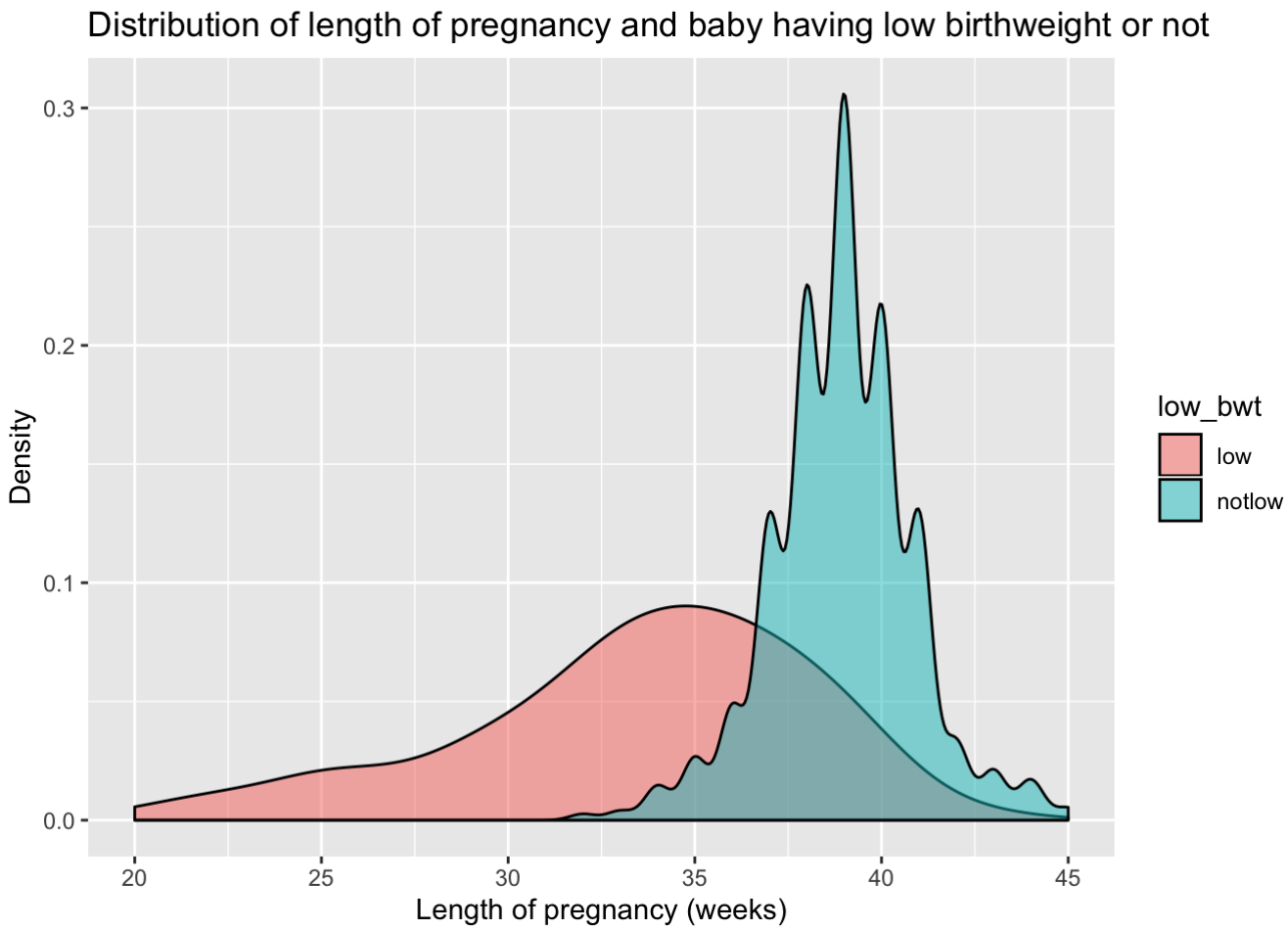
```
##
## Call:
## glm(formula = low_bwt ~ len_preg + marital, family = binomial(),
##     data = births)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4403   0.1479   0.2132   0.3012   1.9834
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -23.63838    2.24101 -10.548   <2e-16 ***
## len_preg           0.70286    0.06131  11.464   <2e-16 ***
## maritalunmarried  -0.66923    0.27075  -2.472   0.0134 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 692.56  on 997  degrees of freedom
## Residual deviance: 396.08  on 995  degrees of freedom
##   (2 observations deleted due to missingness)
## AIC: 402.08
##
## Number of Fisher Scoring iterations: 6
```

```
# both variables have a p-value lesss than significance level 0.05,
# the difference between the null and residual deviance is good
# the degrees of freedom for hte model is very high for our sample size of 1000 ob
servations
# AIC = 402.08, hence indicating that the model is statistically significant
pR2(fit_final_glm)
```

```
##          llh      llhNull           G2     McFadden         r2ML
## -198.0396885 -348.6009843  301.1225915    0.4319015    0.2604594
##         r2CU
##    0.5181017
```

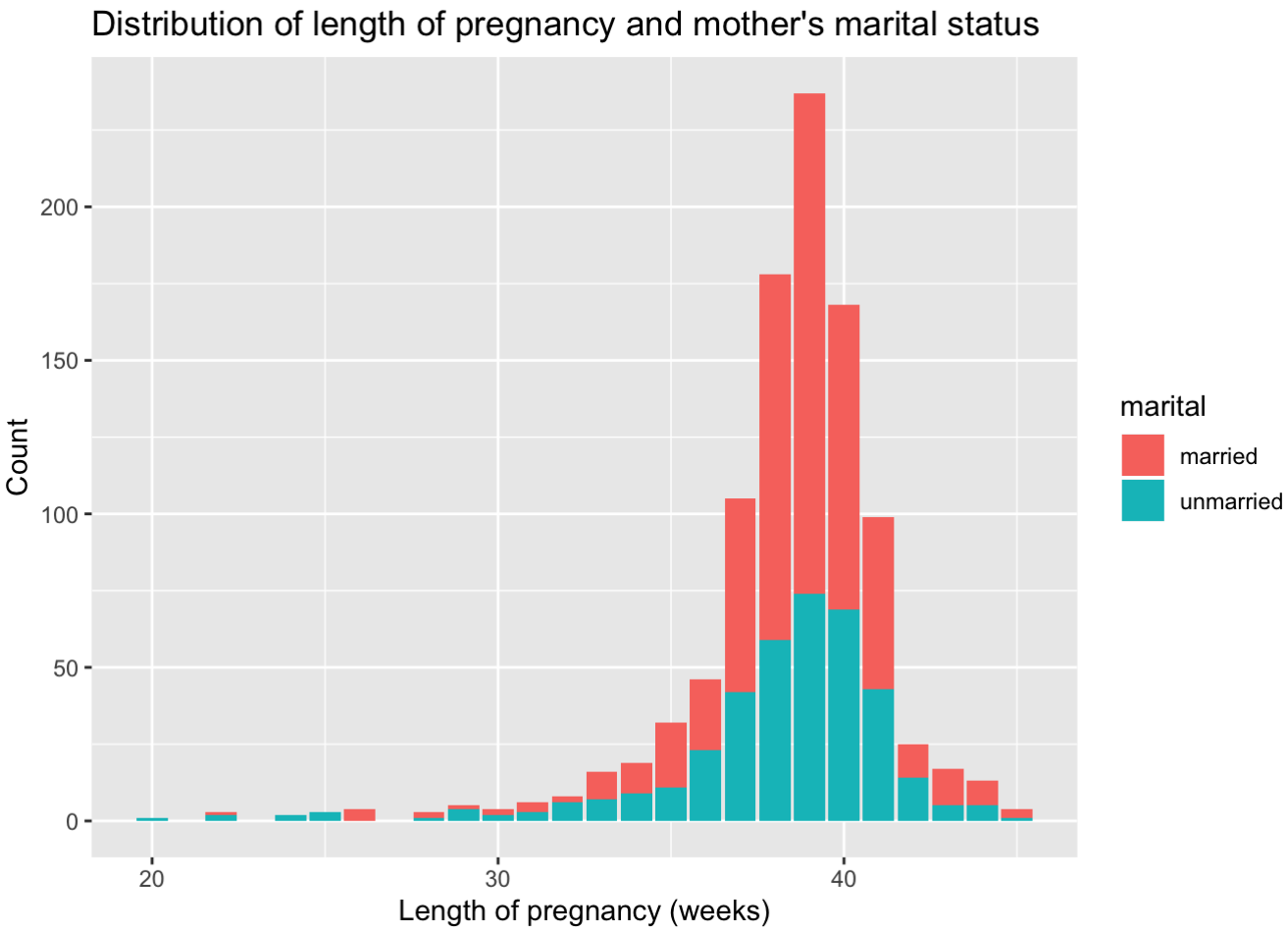
```
# the McFadden value  = 0.4319, also indicates a good fit

g20 = ggplot(births, aes(x=len_preg, fill=low_bwt))
g20 + geom_density(alpha=0.5) + ggtitle("Distribution of length of pregnancy and b
aby having low birthweight or not")+
  xlab ("Length of pregnancy (weeks)") + ylab("Density")
```

```
## Warning: Removed 2 rows containing non-finite values (stat_density).
```




```
# plot shows that higher the length of pregancy, more is the chance that the baby
 will not have a low birthweight

g21 = ggplot(births, aes(x=len_preg, fill=marital))
g21 + geom_bar() + ggtitle("Distribution of length of pregnancy and mother's marit
al status") +
  xlab("Length of pregnancy (weeks)")  + ylab("Count")
```

```
## Warning: Removed 2 rows containing non-finite values (stat_count).
```


Distribution of length of pregnancy and mother's marital status
Count
200
150
100
50
0
20
30
40
Length of pregnancy (weeks)
marital
married
unmarried

```
# plot reveals that married mother's tend to have longer length of pregnancy as co
mpared to unmarried mothers.


##### Brief Interpretation #####
# the probability of the baby having low birthweight or not is highly dependent on
the variables:
# length of pregnancy and nother's marital status
# fitted equation of the logistic model:
# low_bwt = 0.70286 * len_preg - 0.66923 * marital - 23.63838
# for len_preg:
# For every 1 week increase in length of pregnancy, the log odds of having baby wi
th low birthweight
# increases by 0.091152 on average, when all other variables in the model are held
constant
# In the fit_final_glm: exp(coef(len_preg) = 2.019516
# For every 1 week increase in length of pregnancy, the odds of having baby with l
ow birthweight is
# multiplied by 0.091152 on average, when all other variables in the model are hel
d constant
# for marital:
# For married as compared with unmarried mothers, the log odds of having baby with
low birthweight
# decreases by 0.66923 on average, when all other variables in the model are held
 constant
# In the fit_final_glm: exp(coef(marital) = 0.512104
# For married as compared with unmarried mothers, the odds of having baby with low
birthweight
# is multiplied by 0.512104 on average, when all other variables in the model are
 held constant
```