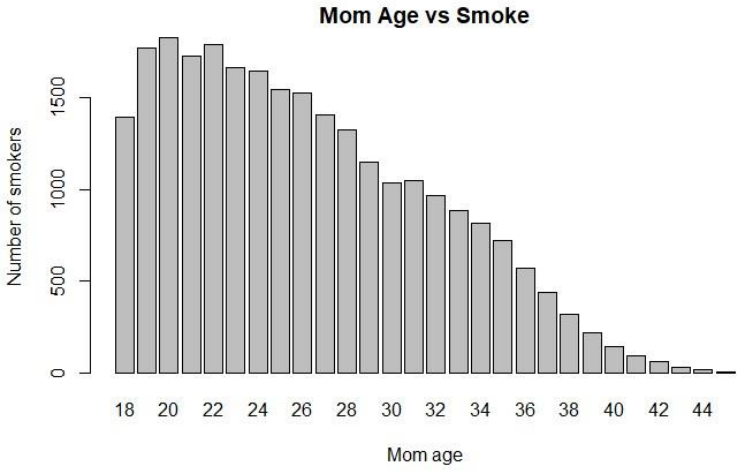
1. **Data Understanding: Two Variables that attempt to show an (interesting) relation via visualization**



One logical question that we asked ourselves was would the number of smokers decrease as the age of mom increased. After computing the aggregate value for total number of smokers per age, we plotted the barplot (above), and we can see that it is indeed the case that the fewer number of mothers smoke as their age increases. This would ideally mean that no-smoking advertisements should ideally be targeted towards younger mothers (Mothers below the age of 40).

After checking the relationship between many variables, we noted that the variables age of the mother vs whether the mother smokes or not have an interesting relationship. As evident from the histogram above, the higher the age of the mother is, the lesser tendency they have to smoke. i.e, the conclusion is that the lesser number of mothers smoke as their age increases. This would ideally mean that no-smoking advertisements should ideally be targeted towards younger mothers (Mothers below the age of 35).

2. **Testing independence among all 45 combinations of binary variables. First using the traditional 0.05 rule for each. Second controlling for multiple testing via Bonferroni correction.**

Testing independence using the traditional rule of p-value > 0.05 : Based on this, the below variable combinations have a p value greater than 0.05.

a. Testing independence using the traditional rule of p-value > 0.05

boy and tri1 boy
and tri2 boy
and tri3 boy
and ed.hs boy
and ed.smcol
boy and ed.col
boy and smoke

b. Testing independence using the Bonferroni rule of p-value>0.05/45: Based on this, the below variable combinations have a p value greater than 0.05/45.

boy and tri1
boy and tri2
boy and tri3
boy and black
boy and married
boy and ed.hs
boy and ed.smcol
boy and ed.col

boy and smoke

Based on the two above tests, we have identified the discrepancies
below: boy and black, boy and married

The traditional testing of independence got it wrong, because we noted that on using the Bonferroni Correction, we were able to identify two more parameters that need to be omitted in order to get a more accurate result.

3. Predicting Birthweight:

Since correlation plot does not indicate any strong correlation between our dependent variable and independent variables, we will try to build different models to try and identify the variables with which we can predict the weight of an infant.

After splitting the cleaned dataset (DATA) into an 80:20 split, we tested out regression models with different combinations of variables. Our conclusion is that the model below yields the lowest residual standard error (which is 0.538) from all the combinations we tested out. Furthermore, the variables below are also statistically significant.

Model =
lm(weight~black+married+boy+tri1+tri2+tri3+ed.hs+ed.smcol+ed.col+mom.age+smoke+cigsper+m.wtgain+mo
m.age2, data = DATA)

This yielded the below summary:

```
Residual standard error: 0.538 on 198362 degrees of freedom
Multiple R-squared:  0.1101,    Adjusted R-squared:   0.11
F-statistic:  1753 on 14 and 198362 DF,  p-value: < 2.2e-16
```

We noted that the Residual standard error is 0.538 which is better than the other models that we tested. We thus used this model to predict the values of weight for the test set (20% of the dataset). On checking the Root Mean Square Error of the forecasted values, we got the value 0.526. This shows that we were able to predict the values considering the most significant variables on the data that our model hasn't seen yet.

Therefore, as per our calculations, the variables are black,married,boy,tri1,tri2,tri3,ed.hs,ed.smcol,ed,col,mom.age,smoke,cigsper,mwtgain,mom.age2 that help predict the birthweight.

4. Multiple Regression To Determine Statistical Significance:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.771e+00  1.434e-02 193.224  < 2e-16 ***
mom.age2     7.822e-05  4.206e-06  18.597  < 2e-16 ***
boy          1.090e-01  2.420e-03  45.021  < 2e-16 ***
m.wtgain     8.840e-03  9.444e-05  93.604  < 2e-16 ***
cigsper     -3.525e-03  4.470e-04  -7.885 3.17e-15 ***
black       -1.991e-01  3.576e-03 -55.671  < 2e-16 ***
married      7.280e-02  3.183e-03  22.871  < 2e-16 ***
tri1         1.851e-01  1.360e-02  13.618  < 2e-16 ***
tri2         1.941e-01  1.389e-02  13.972  < 2e-16 ***
tri3         2.129e-01  1.571e-02  13.549  < 2e-16 ***
ed.hs        2.240e-02  3.740e-03   5.991 2.09e-09 ***
ed.smcol     4.425e-02  4.132e-03  10.710  < 2e-16 ***
ed.col       5.220e-02  4.434e-03  11.774  < 2e-16 ***
smoke       -1.682e-01  6.293e-03 -26.732  < 2e-16 ***
```

All 14 variables are statistically significant applying the 0.05 cut-off value.
The result is the same when the Bonferroni correction is applied with no discrepancy. The Bonferroni threshold we have used is 0.05/14= 0.00357.

5. To increase screening capabilities of detecting risky pregnancies early (end of first trimester), the model can be used to forecast birthweight. However, that comes with issues in implementation:

The adjusted r-squared reading at 0.1086 is not high enough, which means that a consequential component of the birth weights is *not* explained by the input variables (close to 90%). Furthermore, certain variable combinations (such as the ones listed in question 2) are not independent of each other which means the problem of collinearity may arise.

The above two factors will be problematic in terms of implementation. To avoid the issue of collinearity, variables that are not independent of each other (such as the ones listed in question 2) should be avoided.