# Impact of Maternal Smoking on Infant Birthweight

Author 1 and Author 2

2024-10-08

# Header

## Author Contributions

Brief description of the respective contribution of each team member.

Author 1: Worked on questions #1, #3, and #5, and created the data analysis template for the homework. Additionally, Author 1 worked on the advanced analysis question, creating the visualization for the problem and describing it.

Author 2:

## Use of GPT

ChatGPT was used as a substitute for documentation for R. Since we were unfamiliar with R, we asked ChatGPT how to use R in certain methods in order to find and filter out conditions in the dataset. We additionally used GPT to analyze reasoning and to confirm what we thought was correct about the dataset, as well as to identify extra questions that could be answered for our advanced analysis.

# Introduction

The data provided is a Child Health and Development Studies dataset, which consisted of all pregnancies that occurred from 1960-1967 among women with the **Kaiser Health Plan** in Oakland, CA. Some important things to note are that all 1236 babies in the dataset are boys, there are no twins, and all lived at least 28 days. It's important to keep in mind that this is not classified as a simple random sample of all pregnancies, because the conditions just posed cannot be proven to be a totally random sample of all babies born to mothers. However, we are still studying this data beause it still should be a decent representation of differences in weight between babies born to mothers who smoked during pregnancy and those who didn't, even if it is not totally representative of all babies.

## Main Research Questions

1. What are the numerical distributions of the birth weight for babies born to women who smoked versus those who didn't smoke?
2. Is there a significant difference in these two distributions? If so, what type of conclusion can be reached?
3. What percentage of babies born between these two groups (non-smoking mothers and smoking mothers) are considered low-birth-weight babies? Is there a difference?
4. How does the reliability of the three types of comparisons - numerical, graphical, and incidence - change based on our data, and which was the best?

## Outline

The remainder of the report will go through a basic analysis of the data, including our cleaning methods, basic analysis on various variables in the study, and more. Additionally, we will analysis the questions posed above, along with the conclusions that we came up with in our data. We will also pose an advanced analysis question based on the relationship between the gestation period of a pregnancy and see how the relationship between a child's birthweight and their mother's smoking-status are intertwined.

# Basic Analysis

## Data Processing and Summaries

### Methods

In order to analyze our data, we first have to understand it. To do that, we read the data and got a basic summary of each variable to see what we were working with. Through this, we noticed that the variables 'parity' and 'smoke' were both binary variables, representing a True/False statement.

```
data <- read.table("babies.txt", header = TRUE)

bwt_description <- summary(data$bwt)
bwt_description
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    55.0   108.8   120.0   119.6   131.0   176.0
```

```
gestation_description <- summary(data$gestation)
gestation_description
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   148.0   272.0   280.0   286.9   288.0   999.0
```

```
parity_description <- summary(data$parity)
parity_description
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.2549  1.0000  1.0000
```

```
age_description <- summary(data$age)
age_description
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.00   23.00   26.00   27.37   31.00   99.00
```

```
height_description <- summary(data$height)
height_description
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   53.00   62.00   64.00   64.67   66.00   99.00
```

```
weight_description <- summary(data$weight)
weight_description
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      87     115     126     154     140     999
```

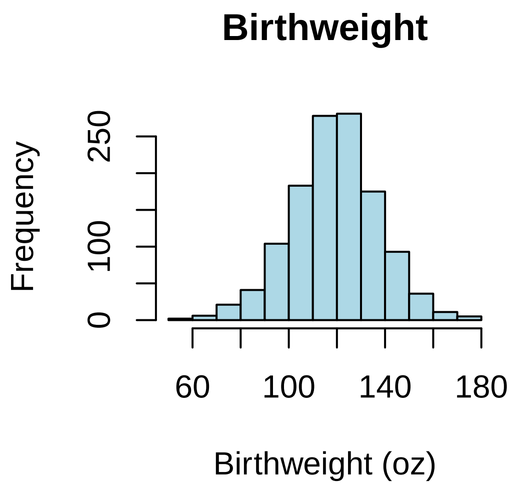
```
smoke_description <- summary(data$smoke)
smoke_description
```

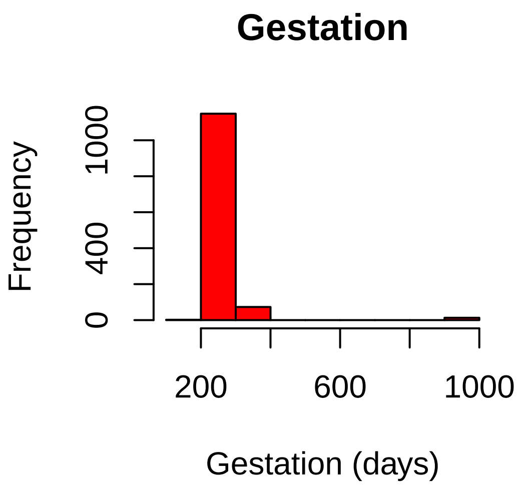
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.4644  1.0000  9.0000
```

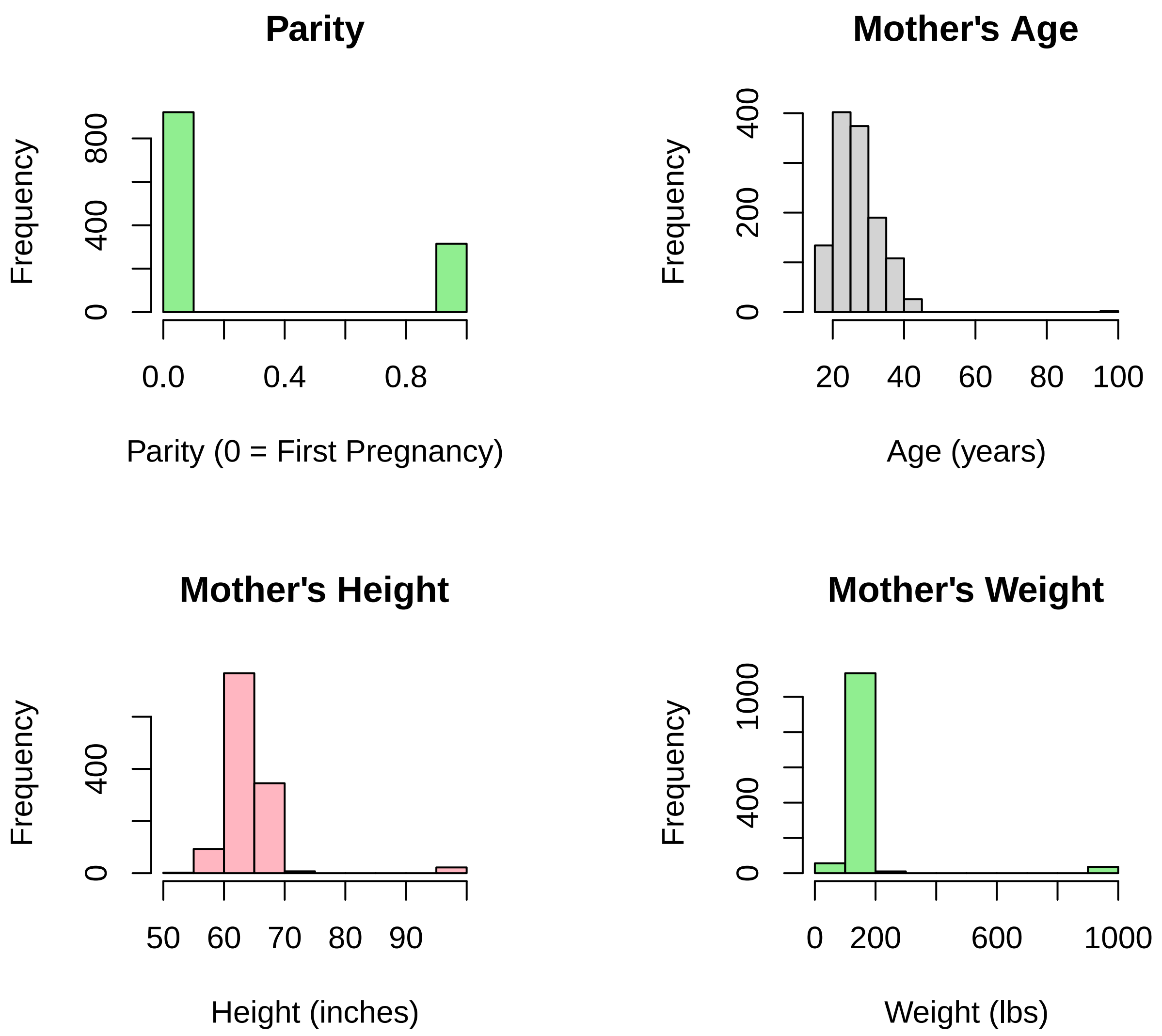
We also looked through the type of each variable, to see if we were mostly working with numerical, categorical, or a mix between the two. As seen below, all of the variables are of type 'int', meaning most are numerical variables. However, as we pointed out above, since 'smoke' and 'parity' were binary variables, these are categorical.

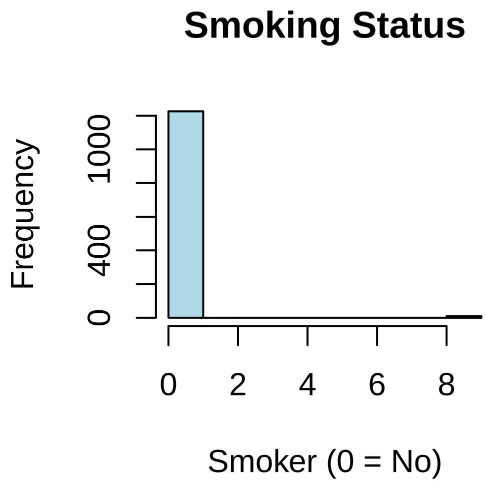
```
## 'data.frame':    1236 obs. of  7 variables:
##  $ bwt      : int  120 113 128 123 108 136 138 132 120 143 ...
##  $ gestation: int  284 282 279 999 282 286 244 245 289 299 ...
##  $ parity   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ age      : int  27 33 28 36 23 25 33 23 25 30 ...
##  $ height   : int  62 64 64 69 67 62 62 65 62 66 ...
##  $ weight   : int  100 135 115 190 125 93 178 140 125 136 ...
##  $ smoke    : int  0 0 1 0 1 0 0 0 0 1 ...
```

We also plotted histograms of our data to observe what type of distributions we were working with. Through this, we were able to notice that some of the histograms that we graphed had what we interpreted as outliers, which were 'gestation', 'age', 'height', 'weight', and 'smoke'. Otherwise, for a variable like 'birthweight', we noticed it was roughly normal, while parity was bimodal (since it is a binary variable).






Parity
Frequency
0
400
800
0.0
0.4
0.8
Parity (0 = First Pregnancy)
Mother's Age
Frequency
0
200
400
20
40
60
80
100
Age (years)
Mother's Height
Frequency
0
400
50
60
70
80
90
Height (inches)
Mother's Weight
Frequency
0
400
1000
0
200
600
1000
Weight (lbs)



## TODO ANALYSIS ON OUTLIERS AND WHY WE CHOSE THIS

```
cleaned_df <- data[data$gestation < 500 & data$age < 50 & data$height < 80 & data$weight < 500 & data$sm
head(cleaned_df, 5)
```

```
##   bwt gestation parity age height weight smoke
## 1 120       284      0  27     62    100     0
## 2 113       282      0  33     64    135     0
## 3 128       279      0  28     64    115     1
## 5 108       282      0  23     67    125     1
## 6 136       286      0  25     62     93     0
```

### Analysis

```
# Your R code for data summary
```

### Conclusions

Your conclusions about data processing and summaries.

## Question 1 (rename)

### Methods

```
# Your R code for methods related to Question 1
```

### Analysis

```r
# Your R code for analysis related to Question 1
```

### Conclusions

Your conclusions for Question 1.

## Question 2 (rename)

### Methods

```r
# Your R code for methods related to Question 2
```

### Analysis

```r
# Your R code for analysis related to Question 2
```

### Conclusions

Your conclusions for Question 2.

## Question 3 (rename)

### Methods

```r
# Your R code for methods related to Question 3
```

### Analysis

```r
# Your R code for analysis related to Question 3
```

### Conclusions

Your conclusions for Question 3.

## Question 4 (rename)

### Methods

```r
# Your R code for methods related to Question 4
```

### Analysis

```r
# Your R code for analysis related to Question 4
```

### Conclusions

Your conclusions for Question 4.

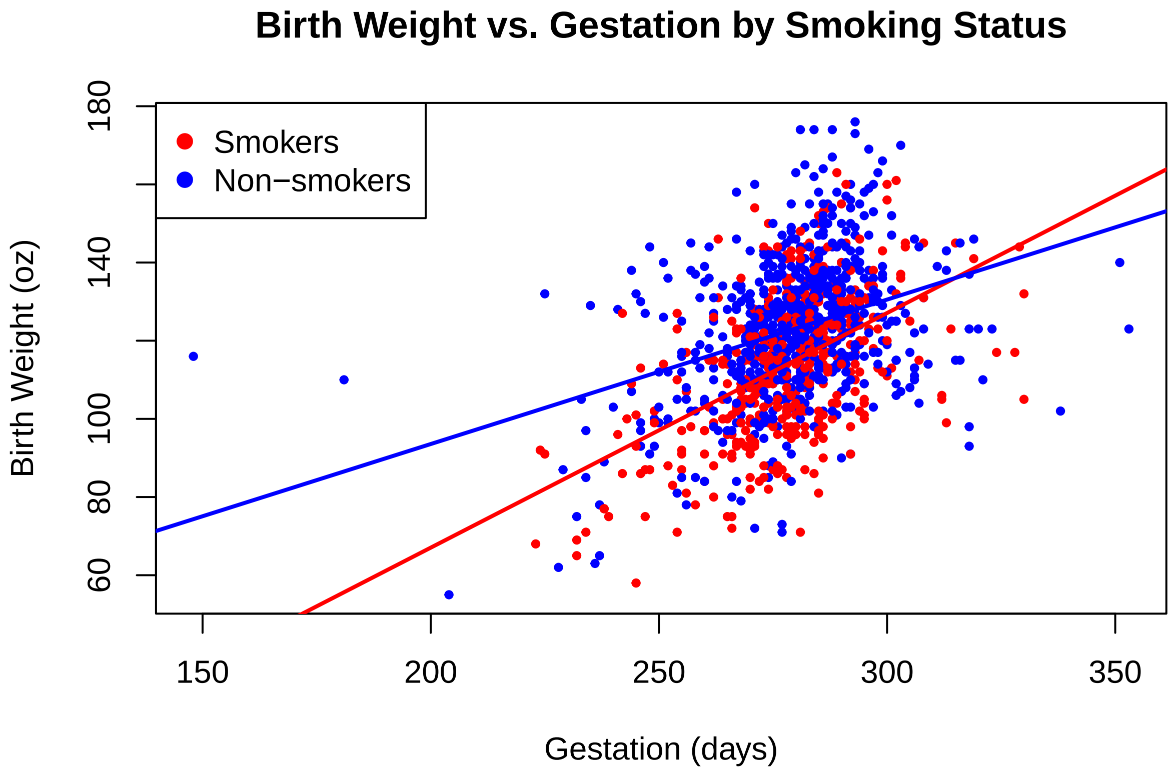
# Advanced Analysis

## Additional Research Question

How does smoking's impact on a baby's birth weight vary depending on how long the pregnancy last?

### Methods

The method we used to analyze the relationship between one of the other variables in the dataset, "gestation", and the babies' weights and parents' smoking status was by using a scatterplot. For starters, we wanted to visually see if there was a relationship between the gestation period and the birth weight, so we used these as our x and y-axes, respectively. In order to identify the difference smoking mothers and non-smoking mothers, we realized that since it was a categorical variable, we could color the dots to indicate it's status. Additionally, we also created a regression line for each subset of mothers, so we could see if a mother's smoking status could visually be represented differently on the plot itself.



**Analysis**

As a general trend, both smokers and non-smokers show a positive correlation between gestation and birthweight, indicating that the longer a gestation period is, the baby's birth weight typically increases.

We also see that the two regression lines (for smoking moms and non-smoking moms) have different slopes. It's shown that the line associated with the smoking moms has a steeper slope than than the line associated with non-smoking mothers. This suggests that for smokers, birth weight increases more rapidly with longer gestation periods for smokers.

We can also notice that over the same gestation periods, babies born to smokers tend to have lower birth weights than their respective counterparts, as observed by the fact that the red line is consistently lower than the blue line for most of the gestation period.

Another thing that can be noticed is the variability of the gestation periods, specifically for non-smoking mothers. Looking through the data, gestation periods for mothers ranged from 150-350 days, while smoking mothers had a much smaller range, typically falling from a range of 220-330 days.

Finally, the last thing to notice from this graph is that there are significantly less mothers that smoke than those that did not.

**Conclusions**

Through this graph, a few conclusions can be reached:

- Smoking is correlated with lower birth weight. The data shows a clear distinction between smokers and non-smokers in terms of their child's birth weight. Since the red line consistenly falls below that of the blue line for most gestation periods, it implies that on average, smoking is associated with lower birth weights.

- Gestation length is positively correlated with birth weight. This means that the longer a gestation period is, the higher the birth weight of a baby usually is.
- Non-smokers show a slower increase in birth weight compared to smokers. There can be many speculations about this, but looking at this graph, it seems that smokers typically have lower birth weights in the same gestation period as non-smokers, leading to their line needing to start at a lower area and climb back up to catch up to other babies that . This could also potentially mean that non-smokers typically have as much of a deviation of a child's birth weight as smokers do, since their slope is smaller, with many of their birth weights being concentrated, while children of smoking mother's have more variability in terms of their birth weight.

# Conclusions and Discussion

## Summary of Findings

The data reveals distinct trends in the birth weights of babies in relation to their mothers' smoking habits. Both non-smoking and smoking mothers show roughly normal distributions for birth weights, but babies born to smoking mothers tend to weigh less on average compared to those born to non-smoking mothers. Additionally, the variability in birth weights among non-smoking mothers is greater, with their distribution closely resembling a Gaussian curve. Furthermore, smoking mothers have a significantly higher likelihood—over 14% more—of having a baby with a low birth weight (under 100 ounces) compared to non-smoking mothers, highlighting a notable disparity in infant health outcomes between the two groups. Finally, our numerical analysis reveals that at every quantile, babies born to non-smoking mothers weigh more than those born to smoking mothers.

This analysis underscores the impact of smoking on birth weight, with clear evidence of increased risk for low-birth-weight babies among smoking mothers

## Discussion

Although there are many warnings everywhere about the dangers of things like alcohol or smoking during pregnancy, many still doubt that doing these things would harm a child. However, our data and our analysis shows a clear trend that smoking can significantly harm a baby's health, specifically towards their birth weight, which may potentially lead towards other health problems in the future.

At the same time however, our analysis is not conclusive. This data cannot be considered a simple random study of all mothers in the world who have had children, because this data is only from a limited period of time (1960 - 1967), only contains male children, and does not have any twins. This data is also from a very specific part of the world (Oakland, CA) that may not be representative of the entire population that we want to study.

Additionally, there may be other confounding factors, like socioeconomic status, access to healthcare, nutrition, and other variables that could potentially influence this study. One question that could be posed is regarding the future implications of smoking on a child. Although this dataset only contains the birthweight of a child, we'd like to see if there are any long term health implications on children whose mothers smoked.

# Appendix

Include any additional technical details, tables, or figures that support your analysis but would disrupt the flow of the main text if included earlier.

```r
# Additional R code or output can be included here
```