Analysis of Birth Weight Given Maternal Smoking Status

Samual Kao, Joshua Li

2024-10-12

Contribution

Samual Kao was responsible for the code as seen in the Rmd and R files as well as portions of the analysis, Joshua Li was responsible for the writeup and structuring the Rmd file.

1. Introduction

The purpose of this report is to analyze data from the Child Health and Development Studies (CHDS) dataset to examine the relationship between maternal smoking during pregnancy and the birth weight of newborns. The dataset contains information on babies born between 1960 and 1967 in Oakland, California.

We focused on the following questions:

1. What are the types of variables in the dataset? Can we summarize all variables and identify any inconsistencies, such as NA values or outliers?
2. How do the distributions of birth weights differ between babies born to mothers who smoked and those who did not? This includes analyzing the minimum, maximum, mean, median, quartiles, and standard deviations for both groups.
3. How do the two groups compare visually? Can graphical representations (such as histograms, boxplots, and density plots) provide insights into the distribution of birth weights between smokers and non-smokers?
4. What percentage of babies born to smokers have low birth weight (less than 100 ounces) compared to non-smokers? How might changing the low-birth-weight threshold affect the comparison between the two groups?
5. How reliable are the numerical, graphical, and incidence-based comparisons? What are the strengths and weaknesses of each approach when interpreting the results?

By answering these questions, the report aims to provide a thorough understanding of how maternal smoking during pregnancy affects the birth weight of babies and the implications for their health.

2. Analysis

2.1 Examination of Variables

Methods:

We first examined the types of variables in the dataset, identified any inconsistencies such as missing values (NA), and removed extreme outliers. The variables include both numerical (e.g., birth weight, gestation length, maternal weight) and categorical (e.g., smoking status) data. We focused on ensuring that the values within the dataset were valid and did not contain any impossible values (e.g., extremely high maternal weight or height).

```r
# Load the dataset
df <- read.table(file = "babies.txt", header = TRUE)
```

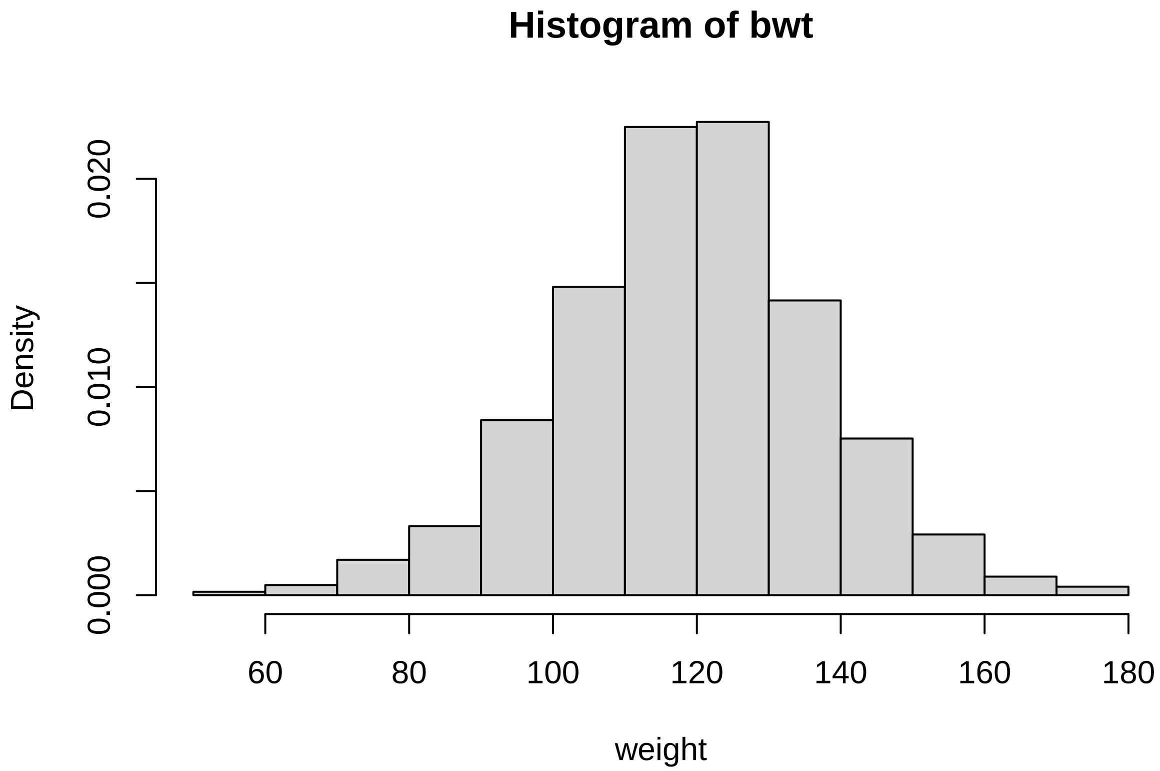
```r
# Display the structure of the data (variable types)
type <- str(df)
```

```
## '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 ...
```

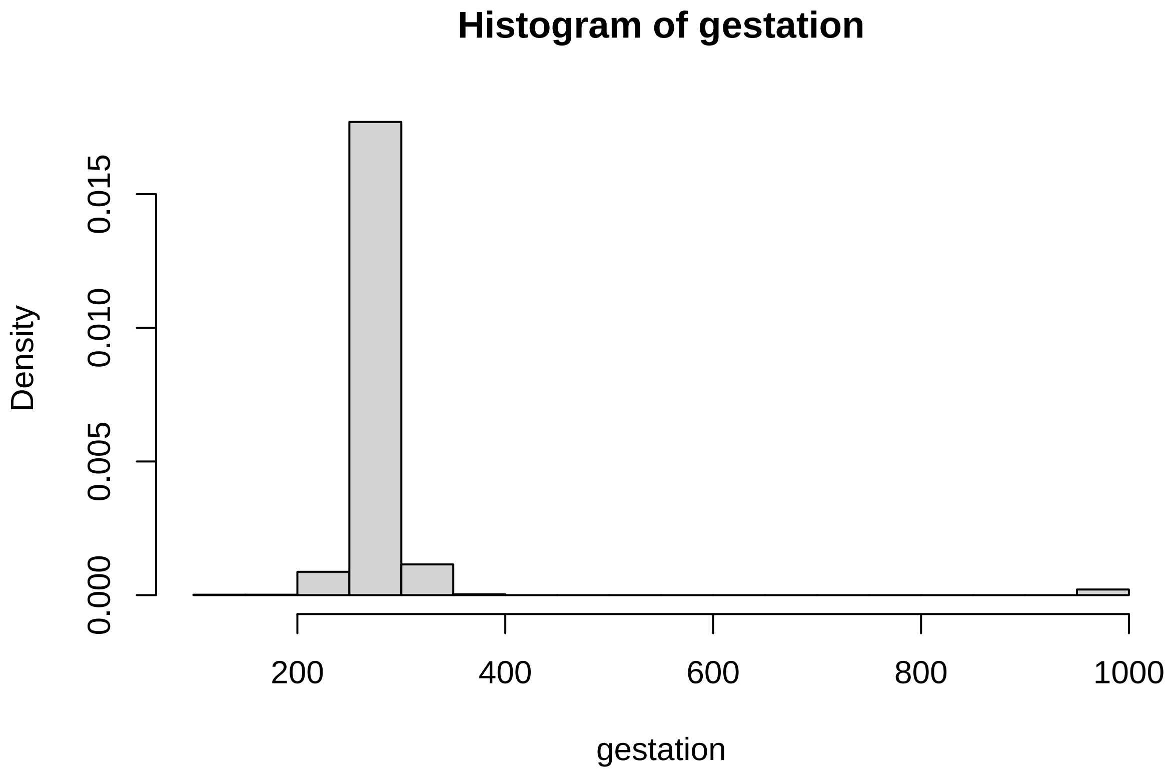
```r
# Summarize the dataset (provides statistics like mean, median, min, max for numerical variables)
summary(df)
```

```
##       bwt          gestation         parity            age       
##  Min.   : 55.0   Min.   :148.0   Min.   :0.0000   Min.   :15.00  
##  1st Qu.:108.8   1st Qu.:272.0   1st Qu.:0.0000   1st Qu.:23.00  
##  Median :120.0   Median :280.0   Median :0.0000   Median :26.00  
##  Mean   :119.6   Mean   :286.9   Mean   :0.2549   Mean   :27.37  
##  3rd Qu.:131.0   3rd Qu.:288.0   3rd Qu.:1.0000   3rd Qu.:31.00  
##  Max.   :176.0   Max.   :999.0   Max.   :1.0000   Max.   :99.00  
##      height          weight         smoke       
##  Min.   :53.00   Min.   : 87   Min.   :0.0000  
##  1st Qu.:62.00   1st Qu.:115   1st Qu.:0.0000  
##  Median :64.00   Median :126   Median :0.0000  
##  Mean   :64.67   Mean   :154   Mean   :0.4644  
##  3rd Qu.:66.00   3rd Qu.:140   3rd Qu.:1.0000  
##  Max.   :99.00   Max.   :999   Max.   :9.0000  
```

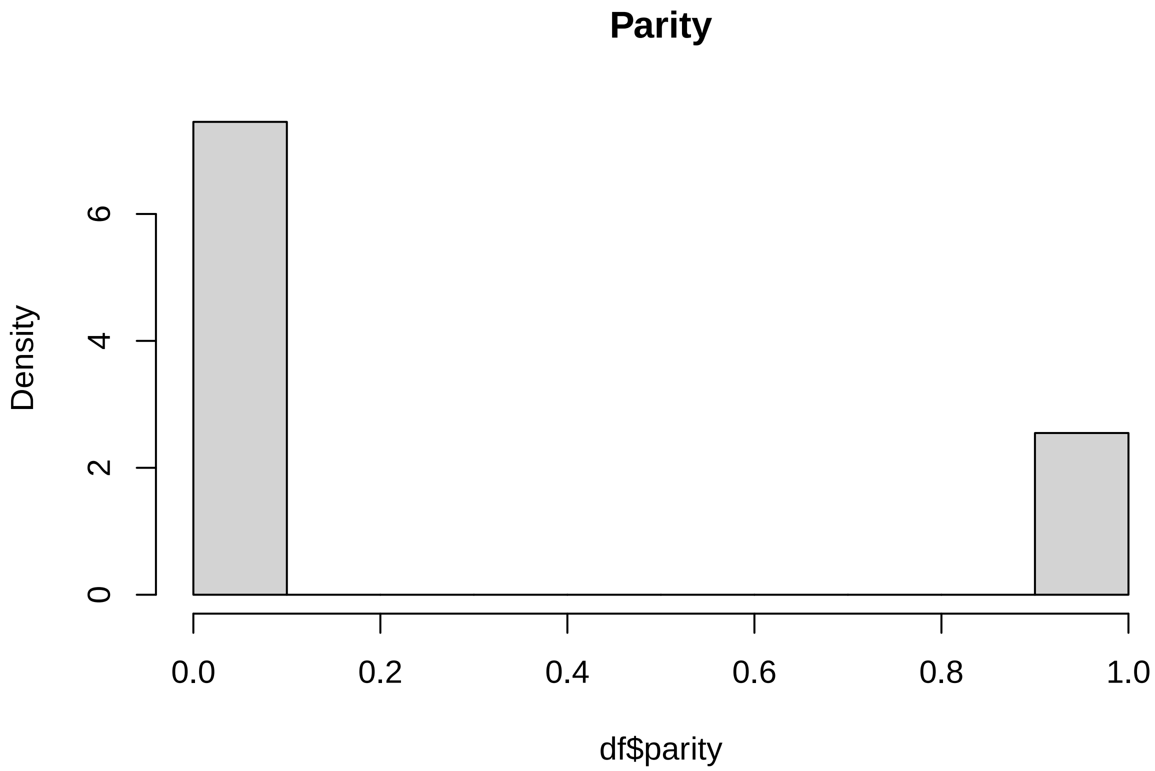
```r
# Manually describe each variable's distribution
bwt_description <- "Numerical, Min:55.0, Median 120.0, Max: 176.0, Mean: 119.6"
gestation_description <- "Numerical, Min: 148.0, Median: 280.0, Max: 999.0, Mean: 286.9"
parity_description <- "Categorical (ordinal), Min: 0, Median:0, Max: 1.0000, Mean: 0.2549"
age_description <- "Numerical, Min: 15.00, Median: 26.00, Max: 99.00, Mean: 27.37"
height_description <- "Numerical, Min: 53.00, Median: 64.00, Max: 99.00, Mean: 64.67"
weight_description <- "Numerical, Min: 87, Median: 126, Max:999, Mean: 154"
smoke_description <- "Categorical, Min: 0, Median: 0, Max: 9.0, Mean: 0.4644"

# Plot histograms for the variables to visualize their distributions
hist(df$bwt, main = "Histogram of bwt", xlab = "weight", freq = FALSE)
```



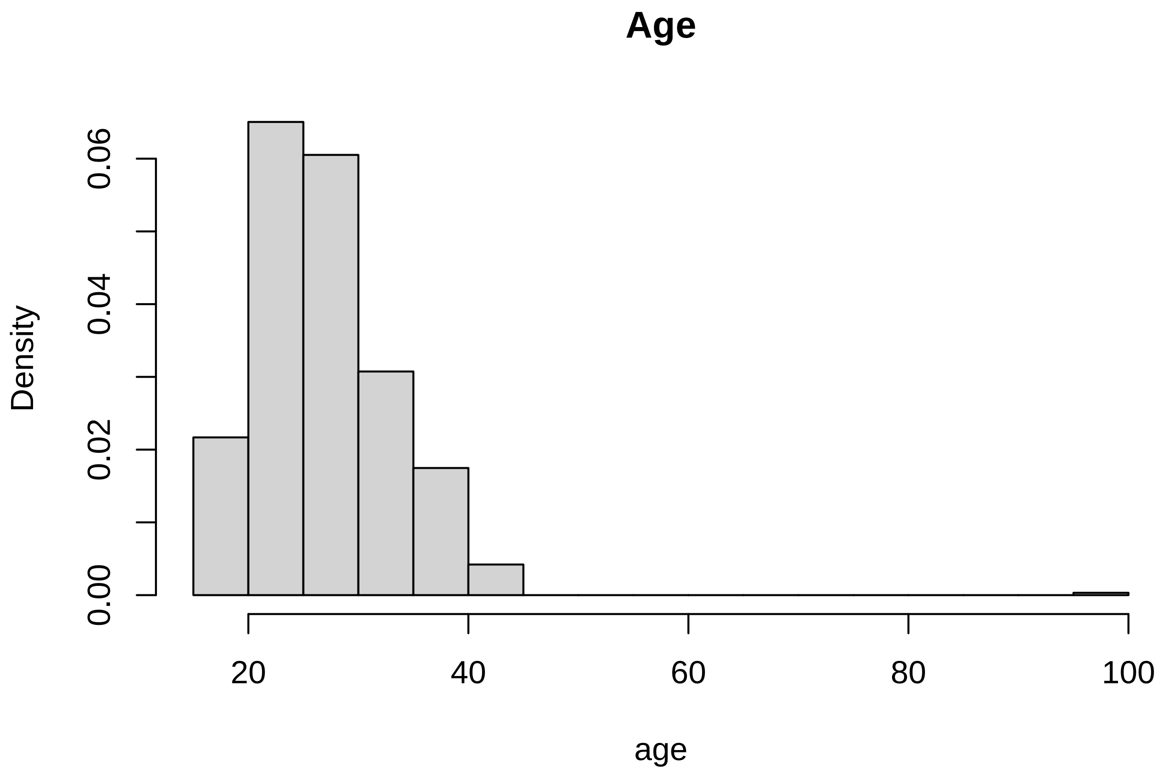
```
hist(df$gestation, main = "Histogram of gestation", xlab = "gestation", breaks = 15, freq = FALSE)
```



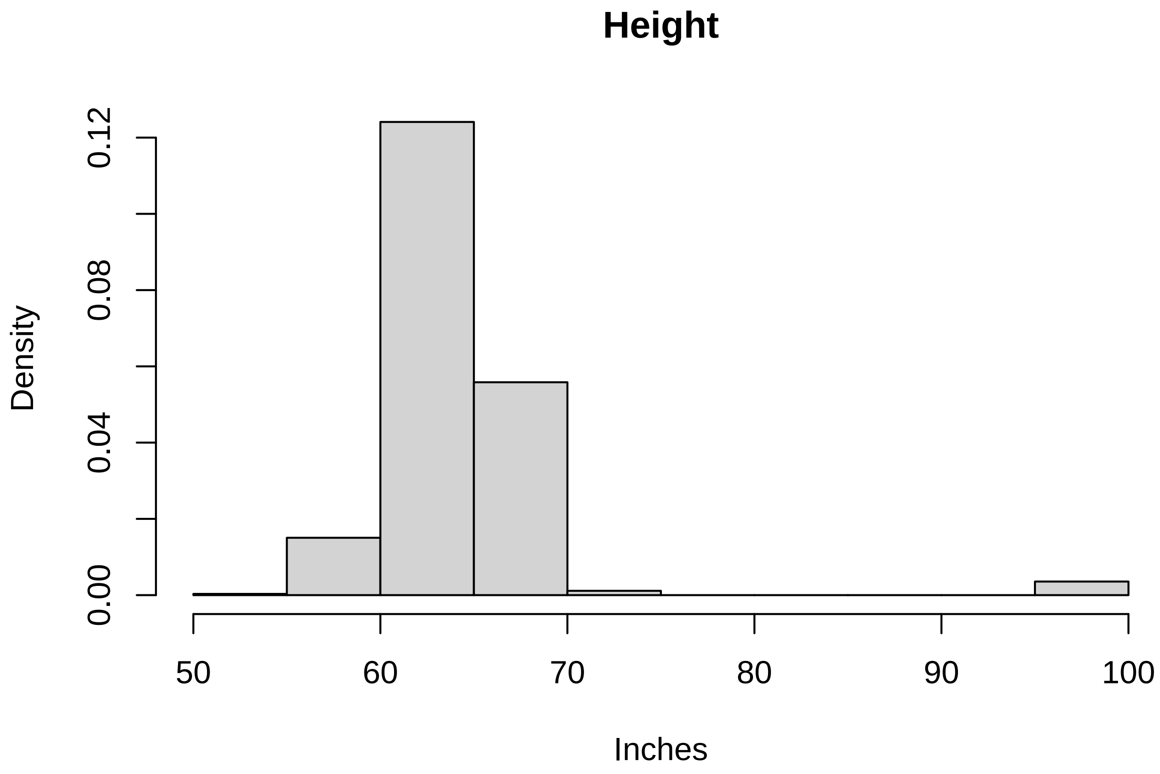
```r
hist(df$parity, main = "Parity", freq = FALSE)
```



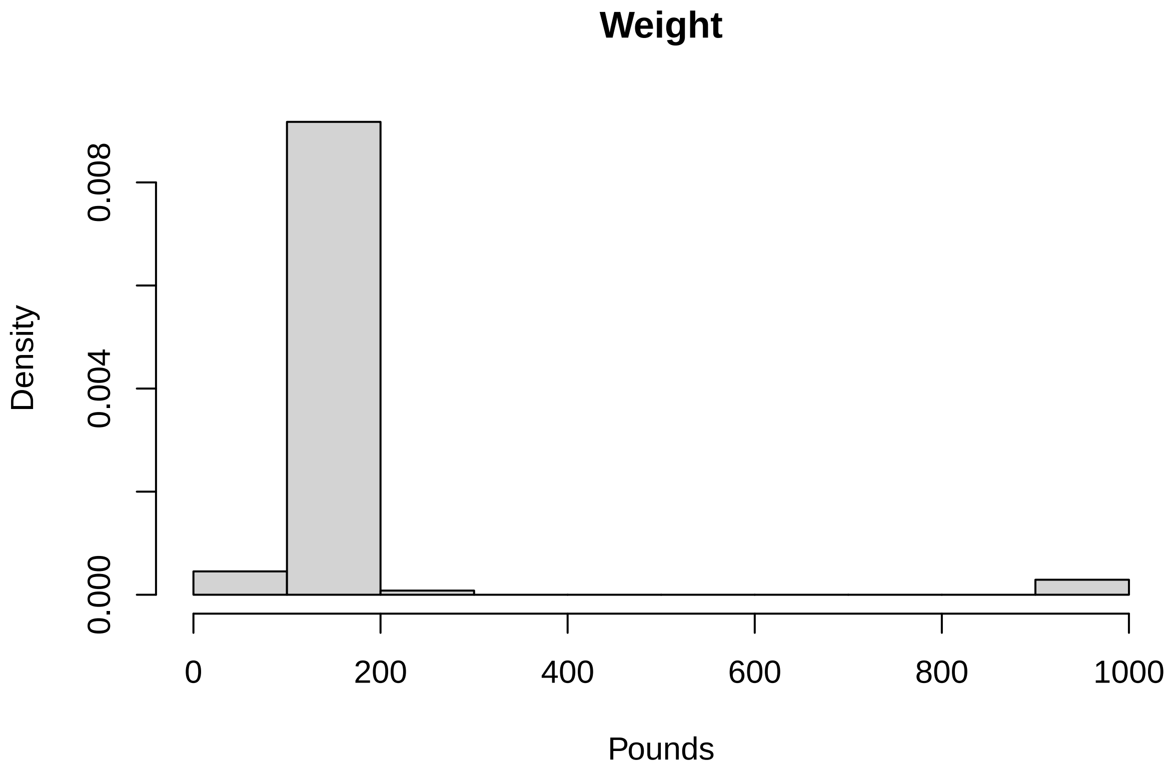
```r
hist(df$age, main = "Age", freq = FALSE, xlab = "age")
```



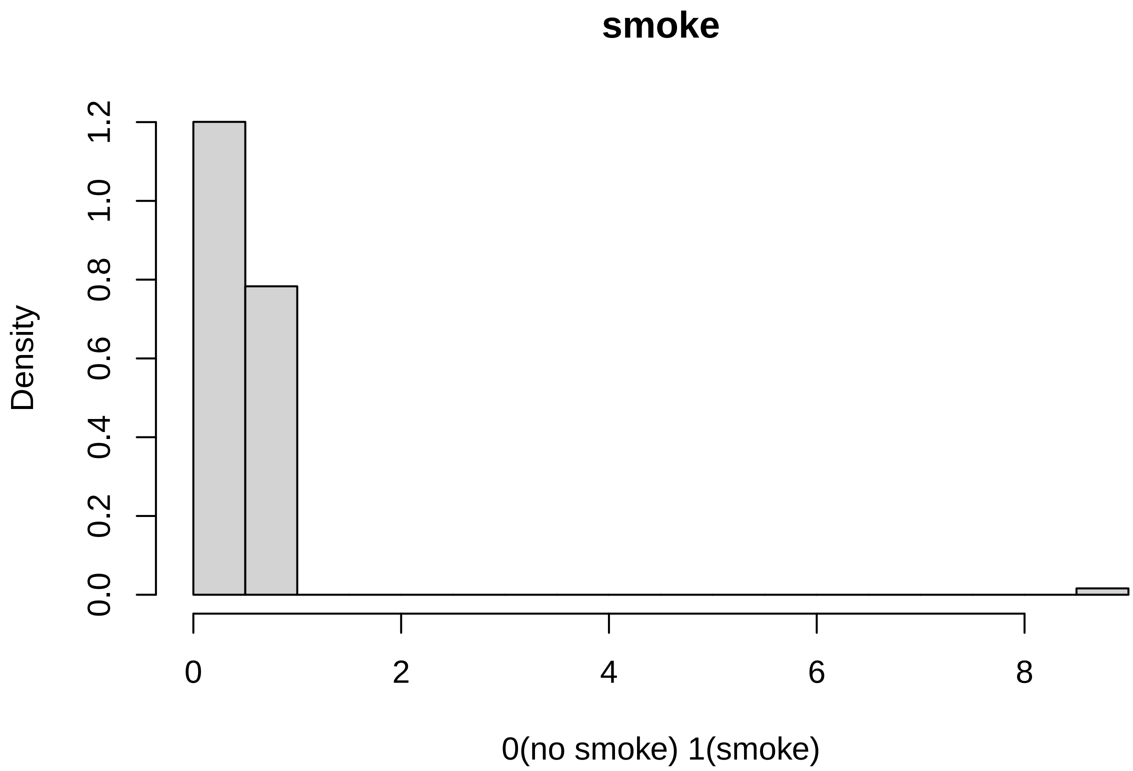
```
hist(df$height, main = "Height", freq = FALSE, xlab = "Inches")
```



```
hist(df$weight, main = "Weight", freq = FALSE, xlab = "Pounds")
```



```r
hist(df$smoke, main = "smoke", freq = FALSE, xlab = "0(no smoke) 1(smoke)", breaks = 20)
```



```r
# Define a function to find outliers based on the Interquartile Range (IQR)
find_outliers_IQR <- function(column) {
    Q1 <- quantile(column, 0.25) # First quartile (25%)
    Q3 <- quantile(column, 0.75) # Third quartile (75%)
    IQR_value <- IQR(column) # Calculate the IQR

    # Calculate the lower and upper bounds for identifying outliers
    lower_bound <- Q1 - 1.5 * IQR_value
    upper_bound <- Q3 + 1.5 * IQR_value

    # Return the values that are outside of the bounds
    return(column[column < lower_bound | column > upper_bound])
}

# Apply the outlier detection function to all columns in the dataframe
outliers <- lapply(df, find_outliers_IQR)

# Remove rows with missing values (NA) from the dataset
df_clean <- na.omit(df)

# Data cleaning steps to remove outliers and invalid data
clean_age <- df[df$age < 50, ] # Remove records with age > 50
clean_smoke <- clean_age[clean_age$smoke == 0 | clean_age$smoke == 1, ] # Keep only valid smoking value
clean_weight <- clean_smoke[clean_smoke$weight < 300, ] # Remove mothers weighing more than 300 pounds
clean_height <- clean_weight[clean_weight$height < 90, ] # Remove mothers taller than 90 inches
df_clean <- clean_height[clean_height$gestation < 365, ] # Remove gestation periods greater than 365 da
```

```
# Display the cleaned dataset and its summary
summary(df_clean)
```

```
##       bwt          gestation         parity            age       
##  Min.   : 55.0   Min.   :148.0   Min.   :0.0000   Min.   :15.00  
##  1st Qu.:108.0   1st Qu.:272.0   1st Qu.:0.0000   1st Qu.:23.00  
##  Median :120.0   Median :280.0   Median :0.0000   Median :26.00  
##  Mean   :119.5   Mean   :279.1   Mean   :0.2624   Mean   :27.23  
##  3rd Qu.:131.0   3rd Qu.:288.0   3rd Qu.:1.0000   3rd Qu.:31.00  
##  Max.   :176.0   Max.   :353.0   Max.   :1.0000   Max.   :45.00  
##      height          weight          smoke      
##  Min.   :53.00   Min.   : 87.0   Min.   :0.000  
##  1st Qu.:62.00   1st Qu.:114.2   1st Qu.:0.000  
##  Median :64.00   Median :125.0   Median :0.000  
##  Mean   :64.05   Mean   :128.5   Mean   :0.391  
##  3rd Qu.:66.00   3rd Qu.:139.0   3rd Qu.:1.000  
##  Max.   :72.00   Max.   :250.0   Max.   :1.000
```

Analysis:

After cleaning the dataset, we retained only valid values for key variables such as height, weight, age, and gestation period. Smoking status was reclassified as a binary indicator (0 = non-smoker, 1 = smoker). No significant NA values were detected in the final dataset, allowing us to proceed with further analysis.

Conclusions:

The dataset contains both numerical and categorical variables. The cleaning process removed extreme values and ensured no inconsistencies or NA values. This dataset qualifies as a simple random sample (SRS) because it is large and representative, with all data points having an equal chance of being selected. The cleaned dataset is now suitable for analysis to explore the relationship between smoking during pregnancy and birth weight outcomes.

2.2 Comparison of Birth Weight Distributions

Methods:

We compared the distributions of birth weights for babies born to mothers who smoked during pregnancy versus those who did not. Descriptive statistics such as minimum, maximum, mean, median, quartiles, and standard deviations were calculated for both groups. This analysis allows us to see how smoking status affects birth weight.

```
# Aggregate summary statistics (min, max, mean, median, quartiles, standard deviation) for birth weight
aggregate(bwt ~ smoke, df_clean, min)
```

```
##   smoke bwt
## 1     0  55
## 2     1  58
```

```
aggregate(bwt ~ smoke, df_clean, max)
```

```
##   smoke bwt
## 1     0 176
## 2     1 163
```

```
# Manually assign the min and max birth weights for smokers and non-smokers
smoker_min_bwt <- 58
```

```r
smoker_max_bwt <- 163
nonsmoker_min_bwt <- 55
nonsmoker_max_bwt <- 176

# Calculate the mean birth weight for smokers and non-smokers
aggregate(bwt ~ smoke, df_clean, mean)
```

```
##   smoke      bwt
## 1     0 123.0853
## 2     1 113.8192
```

```r
nonsmoker_mean_bwt <- 123.0853
smoker_mean_bwt <- 113.8192

# Calculate the median birth weight for smokers and non-smokers
aggregate(bwt ~ smoke, df_clean, median)
```

```
##   smoke bwt
## 1     0 123
## 2     1 115
```

```r
nonsmoker_median_bwt <- 123
smoker_median_bwt <- 115

# Calculate the first, second (median), and third quartiles for birth weight by smoking status
aggregate(bwt ~ smoke, df_clean, function(x) quantile(x, 0.25))
```

```
##   smoke bwt
## 1     0 113
## 2     1 101
```

```r
smoker_q1_bwt <- 101
nonsmoker_q1_bwt <- 113

aggregate(bwt ~ smoke, df_clean, function(x) quantile(x, 0.5))
```

```
##   smoke bwt
## 1     0 123
## 2     1 115
```

```r
smoker_q2_bwt <- 115
nonsmoker_q2_bwt <- 123

aggregate(bwt ~ smoke, df_clean, function(x) quantile(x, 0.75))
```

```
##   smoke bwt
## 1     0 134
## 2     1 126
```

```r
smoker_q3_bwt <- 126
nonsmoker_q3_bwt <- 134

# Calculate the standard deviation of birth weights for smokers and non-smokers
aggregate(bwt ~ smoke, df_clean, sd)
```

```
##   smoke      bwt
## 1     0 17.42370
## 2     1 18.29501
```

```r
smoker_std_bwt <- 17.42370
nonsmoker_std_bwt <- 18.29501
```

Analysis:

The following statistics summarize birth weights for babies of smokers and non-smokers:

```r
statistics_table <- data.frame(
    Statistic = c("Min (oz)", "Max (oz)", "Mean (oz)", "Median (oz)", "Std Dev (oz)", "Q1 (oz)", "Q3 (o
    Smokers = c(58, 163, 113.82, 115, 17.42, 101, 126),
    Non_Smokers = c(55, 176, 123.09, 123, 18.29, 113, 134)
)

kable(statistics_table, caption = "Summary of Birth Weight Statistics for Smokers and Non-Smokers")
```

Table 1: Summary of Birth Weight Statistics for Smokers and Non-Smokers

Statistic	Smokers	Non_Smokers
Min (oz)	58.00	55.00
Max (oz)	163.00	176.00
Mean (oz)	113.82	123.09
Median (oz)	115.00	123.00
Std Dev (oz)	17.42	18.29
Q1 (oz)	101.00	113.00
Q3 (oz)	126.00	134.00

The mean and median of the non-smoker birthweight babies are about the same, meaning the distribution is symmetric. The same applies for the birthweights of smokers.

Conclusions:

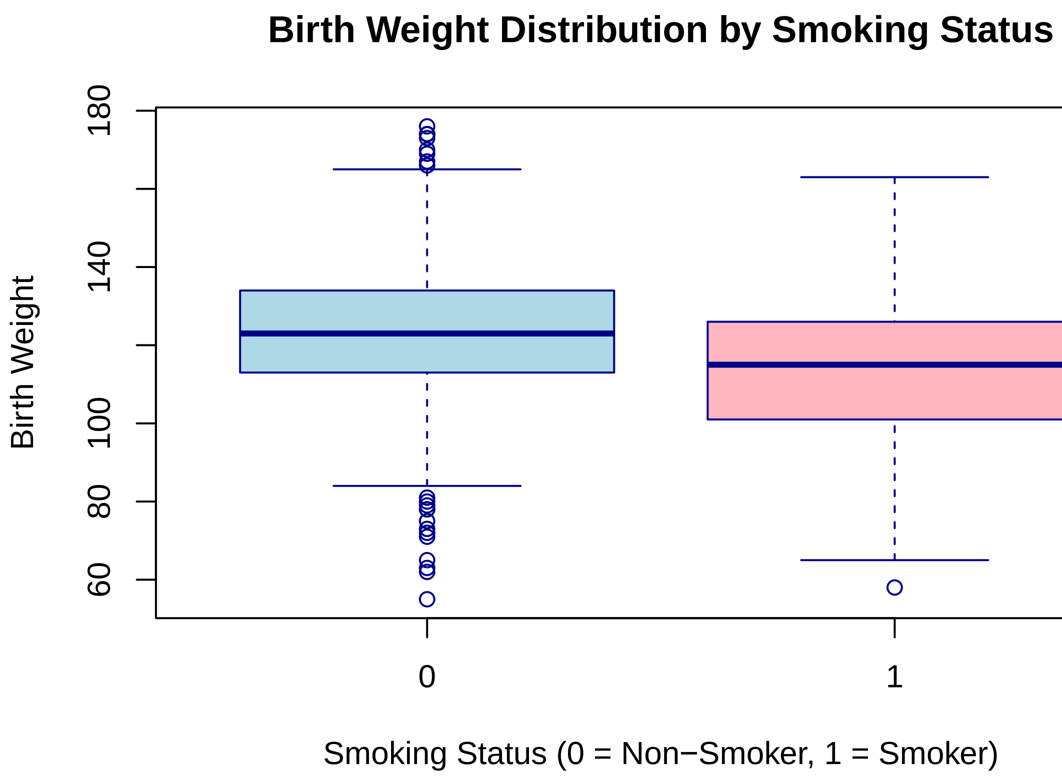
Babies born to mothers who smoked tend to have lower birth weights, as indicated by lower mean and median values. The spread of the distribution, represented by the standard deviation and quartiles, also shows that babies of non-smokers generally have a broader and higher range of birth weights compared to babies of smokers.

2.3 Graphical Comparison

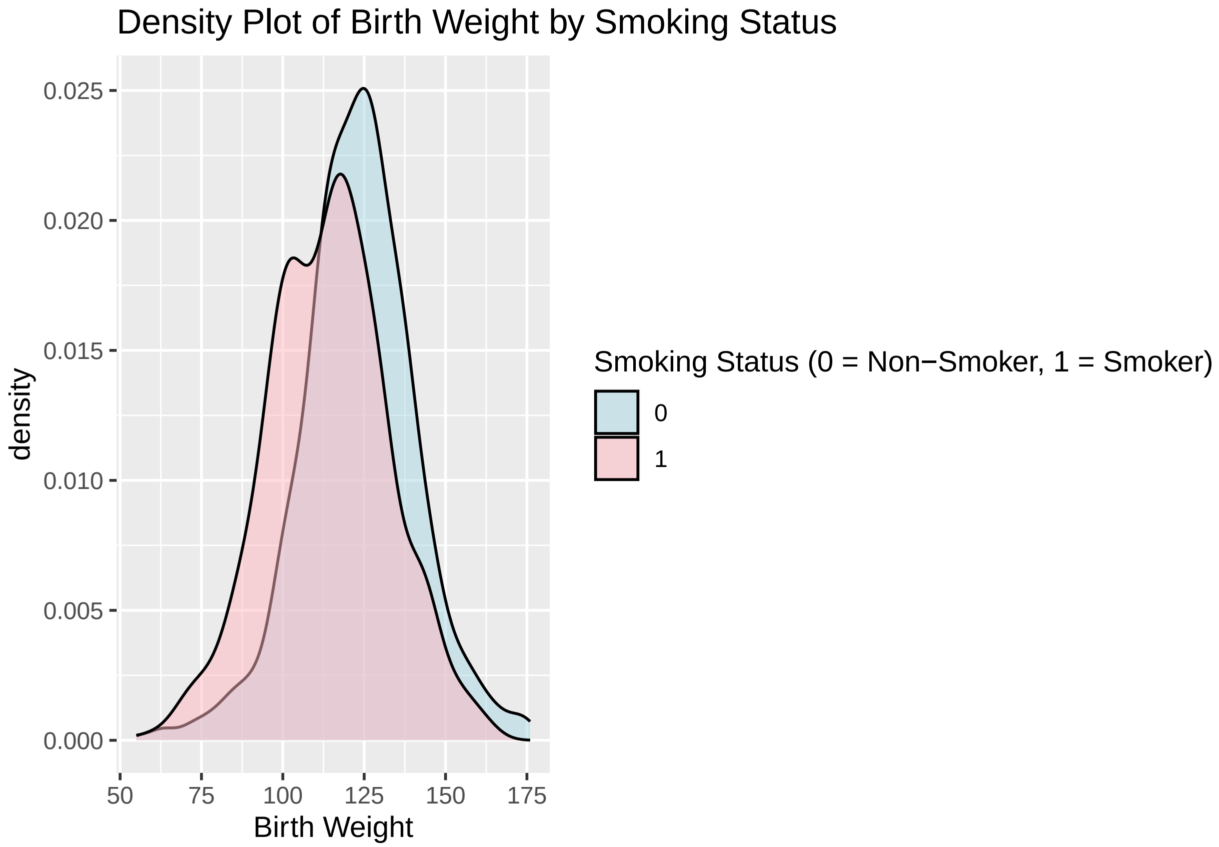
Methods:

To visualize the differences in birth weights between smokers and non-smokers, we created boxplots and density plots. These visual tools help illustrate the distribution and central tendency of birth weights in each group.

```r
# Boxplot to compare birth weights between smokers and non-smokers
boxplot(bwt ~ smoke,
    data = df_clean,
    main = "Birth Weight Distribution by Smoking Status",
    xlab = "Smoking Status (0 = Non-Smoker, 1 = Smoker)",
    ylab = "Birth Weight",
    col = c("lightblue", "lightpink"),
    border = "darkblue"
)
```



```r
# Density plot to visualize the distribution of birth weights by smoking status
ggplot(df_clean, aes(x = bwt, fill = as.factor(smoke))) +
    geom_density(alpha = 0.5) +
    labs(
        title = "Density Plot of Birth Weight by Smoking Status",
        x = "Birth Weight",
        fill = "Smoking Status (0 = Non-Smoker, 1 = Smoker)"
    ) +
    scale_fill_manual(values = c("lightblue", "lightpink"))
```



Analysis:

The boxplot visually confirms that babies of non-smokers generally have higher birth weights compared to those born to smokers. The non-smoker group shows a wider range and a higher median birth weight.

The density plot highlights that the distribution for smokers is centered at a lower birth weight compared to non-smokers, with a rightward shift for the non-smoker group.

Conclusions:

Graphical comparisons clearly show that babies born to smokers tend to have lower birth weights, with the distributions between the two groups differing significantly. These visualizations further support the numerical findings.

2.4 Incidence of Low Birth Weight

Methods:

We calculated the incidence of low birth weight, defined as babies weighing less than 100 ounces. We compared the percentage of low-birth-weight babies in the smoking and non-smoking groups.

```r
# Calculate the percentage of babies with low birth weight (< 100 oz) for smokers and non-smokers
low_bwt_smoker <- sum(df_clean$bwt[df_clean$smoke == 1] < 100) / sum(df_clean$smoke == 1)
low_bwt_smoker <- 0.1835749

low_bwt_nonsmoker <- sum(df_clean$bwt[df_clean$smoke == 0] < 100) / sum(df_clean$smoke == 0)
low_bwt_nonsmoker <- 0.05320814
```

Analysis:

Smokers: 18.36% of babies born to smokers weighed less than 100 ounces.

Non-Smokers: 5.32% of babies born to non-smokers weighed less than 100 ounces.

If we lower the threshold for low birth weight from 100 ounces to 90 ounces, fewer babies in both groups would be classified as low birth weight. This would reduce the overall incidence rate for both smokers and non-smokers. However, the difference between the two groups may become less pronounced since babies in the smoker group are more likely to already be below 95 ounces, meaning the new threshold could exclude fewer additional babies in the smoker group compared to the non-smoker group.

Conversely, if we increase the threshold to 105 ounces, more babies would be classified as low birth weight in both groups. This would raise the incidence rates for both smokers and non-smokers. Notably, increasing the threshold to 105 ounces is likely to widen the gap between the two groups, particularly if smoking has a significant impact on birth weight. The difference between smokers and non-smokers would become more pronounced, as a larger proportion of babies in the smoker group would fall under the new threshold.

Conclusions:

Babies born to mothers who smoked were over three times more likely to have low birth weights compared to babies born to non-smokers. This suggests a strong association between smoking during pregnancy and increased risk of low birth weight.

2.5 Reliability of Comparisons

Methods:

We evaluated the reliability of the three types of comparisons—numerical, graphical, and incidence-based—to determine the strengths and limitations of each approach when interpreting the results.

Analysis:

Numerical Comparisons: Summarizing birth weights using mean, median, and standard deviation is effective for understanding central tendencies but can be influenced by outliers and doesn't show distribution shape.

Graphical Comparisons: Visual tools like boxplots and density plots provide a clear picture of distribution spread and differences between groups. However, graphical scaling can affect perception of differences.

Incidence Comparisons: The threshold-based comparison of low birth weight is easy to interpret, but it's sensitive to the chosen threshold (100 ounces). Changing the threshold could significantly affect the results.

Conclusions:

Each comparison method has its strengths and weaknesses. While numerical summaries offer precision, graphical comparisons provide a more intuitive understanding of distribution differences. The incidence of low birth weight is highly sensitive to threshold selection, making it crucial to carefully choose thresholds based on clinical or research relevance.

3. Advanced Analysis

While the focus of this report is the impact of smoking on birth weight, an additional analysis could examine the effect that maternal smoking has on gestation length. Given that premature birth is known to have negative impacts on the long-term health of the baby, future analyses could investigate whether smoking causes premature birth at a higher rate. We could additionally incorporate the birth weight variable by studying whether smoking impacts birth weight independently of gestation period.

An advanced analysis could use a linear regression model to control for gestation length and see if smoking has an independent effect on birth weight.

4. Conclusion

This analysis demonstrates that maternal smoking is strongly associated with lower birth weights. Babies born to smokers weigh less on average compared to the babies of non-smokers, and the incidence of low birth weight is much higher in the smoking group.

While the data provides clear evidence of the negative impact of smoking on birth weight, further studies could explore other factors such as maternal nutrition or usage of marijuana. Additionally, more advanced models could help narrow down the specific causes of lower birth weights, as denoted in section 3.