# Linear Regression: Ice Cream Sales & Glucose

Alec Brooks

2024-11-09

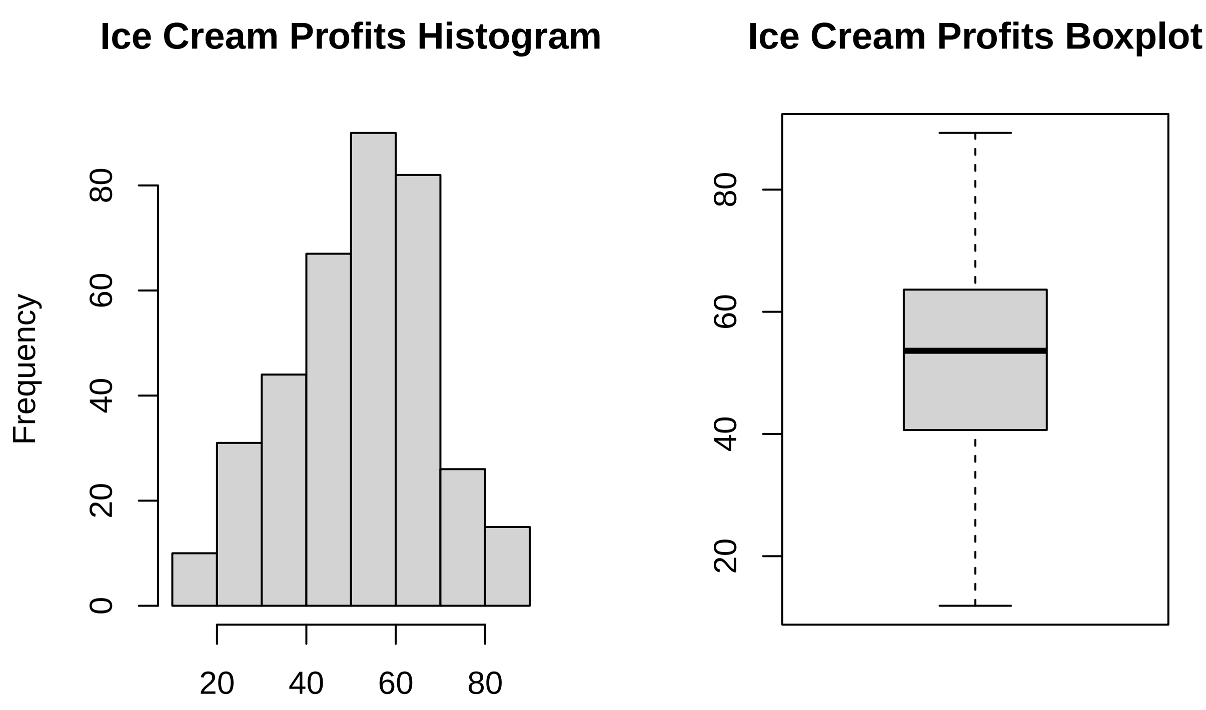
## 1) Ice Cream Sales

*predictor variable:* Temperature

*response variable:* Ice Cream Profits

#### ***Testing Normality:*** Ice Cream Profits

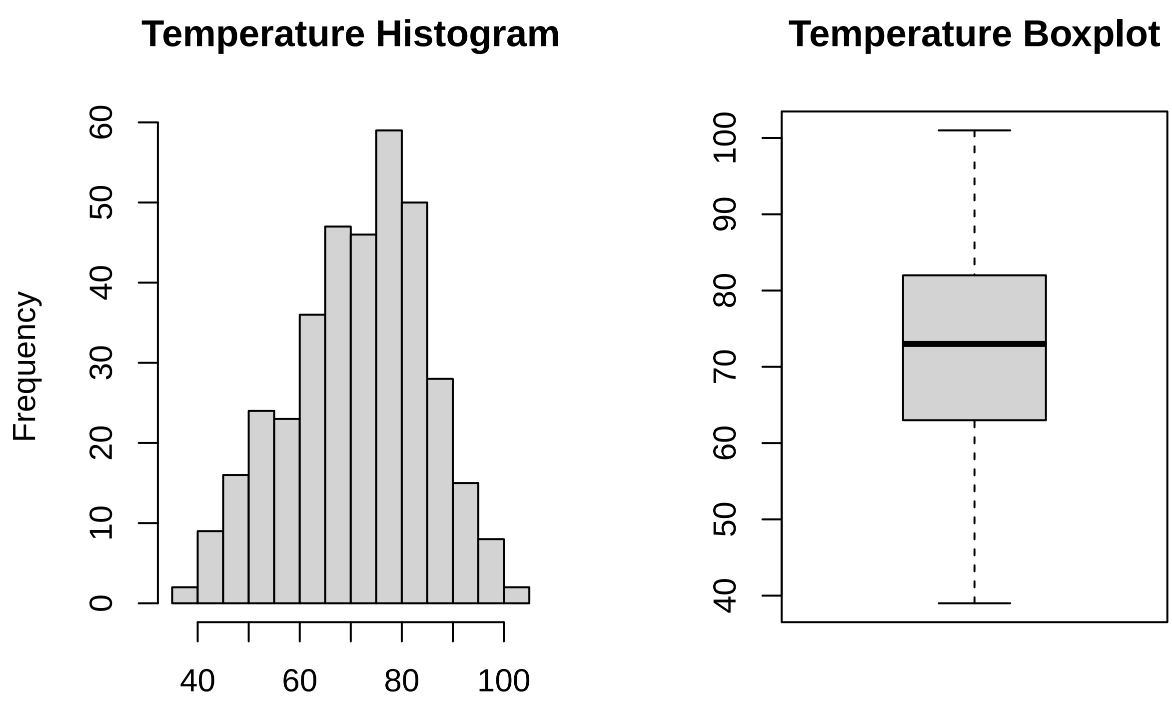
```
par(mfrow = c(1, 2))
hist(icSales$`Ice Cream Profits`,main = "Ice Cream Profits Histogram",xlab = "")
boxplot(icSales$`Ice Cream Profits`,main = "Ice Cream Profits Boxplot")
```



By examining the box plot and histogram, the dependent variable of Ice Cream Profits appears to be somewhat positively skewed, with no apparent outliers. For this reason, it could be considered approximately normal.

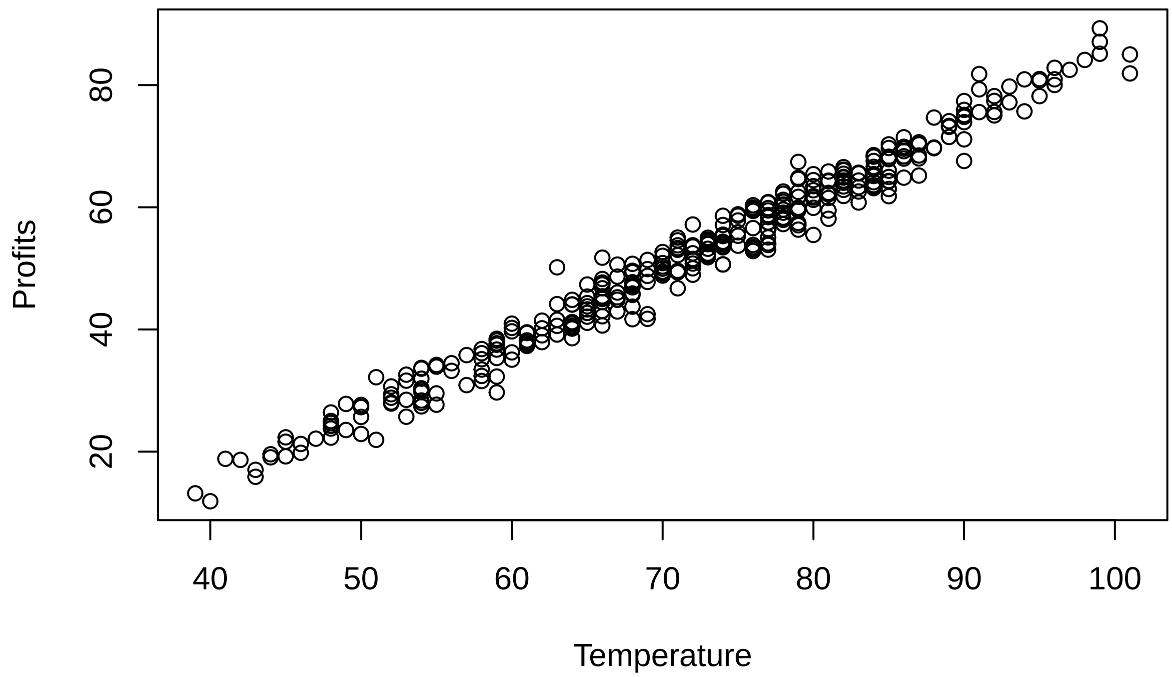
***Testing Normality:* Temperature**

```
par(mfrow = c(1, 2))
hist(icSales$`Temperature`,main = "Temperature Histogram",xlab = "")
boxplot(icSales$`Temperature`,main = "Temperature Boxplot")
```




Similarly, the histogram of temperature also shows a slight positive skew with no outliers. This supports the idea that both the dependent and independent variables are slightly skewed yet are free of significant outliers, making them good candidates for linear regression analysis.

```
plot(icSales$Temperature, icSales$`Ice Cream Profits`,xlab = "Temperature", ylab = "Profits")
abline(0,0)
```




```r
cor.test(icSales$Temperature, icSales$`Ice Cream Profits`)
```

```
##
##  Pearson's product-moment correlation
##
## data:  icSales$Temperature and icSales$`Ice Cream Profits`
## t = 124.24, df = 363, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9858210 0.9905869
## sample estimates:
##       cor
## 0.9884457
```

Visually, we can see that the data appears to be positively correlated, with profits increasing linearly with temperature. This is further supported by the correlation coefficient of approximately 0.99, indicating a very strong positive correlation. Additionally, the p-value of approximately 0.00 from the correlation test confirms that this correlation is statistically significant and reliable.

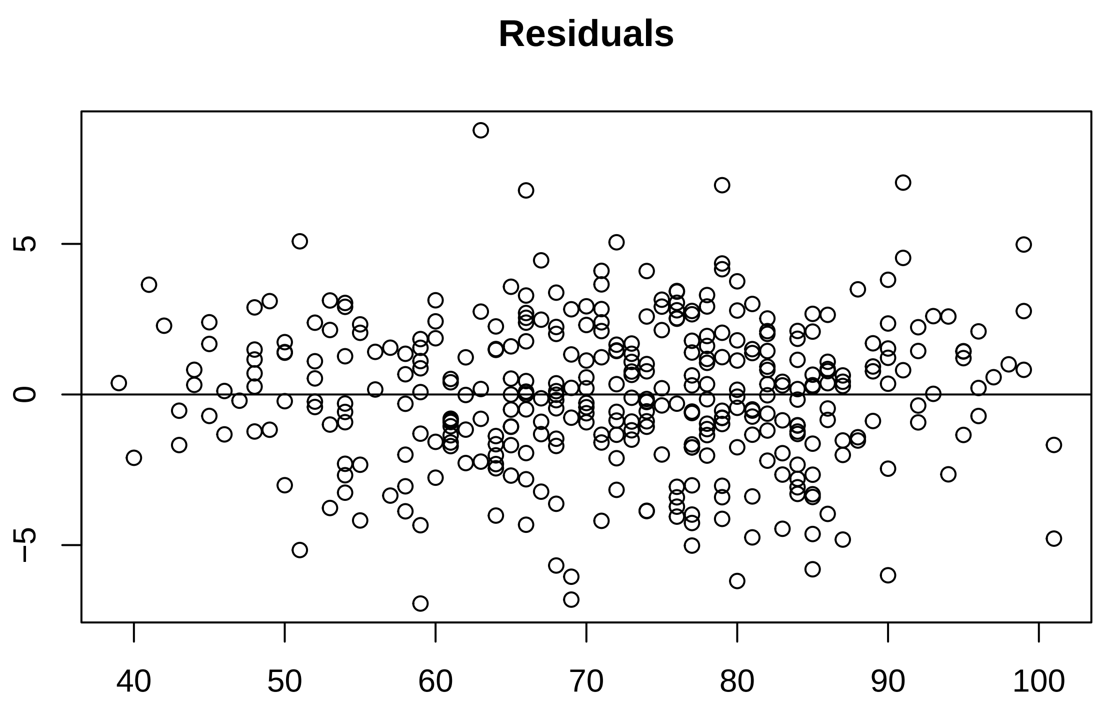
```r
ic_LM <- lm(icSales$`Ice Cream Profits`~icSales$Temperature)
summary(ic_LM)
```

```
##
## Call:
```

```
## lm(formula = icSales$`Ice Cream Profits` ~ icSales$Temperature)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9404 -1.3804  0.0956  1.5976  8.7716
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -33.698166   0.702173  -47.99   <2e-16 ***
## icSales$Temperature   1.192009   0.009594  124.25   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.427 on 363 degrees of freedom
## Multiple R-squared:  0.977,  Adjusted R-squared:  0.977
## F-statistic: 1.544e+04 on 1 and 363 DF,  p-value: < 2.2e-16
```

The summary of the model shows an R-squared value of 0.977, indicating that 97% of the variability in Profits can be explained by the Temperature variable. This suggests a very strong relationship between the two variables.

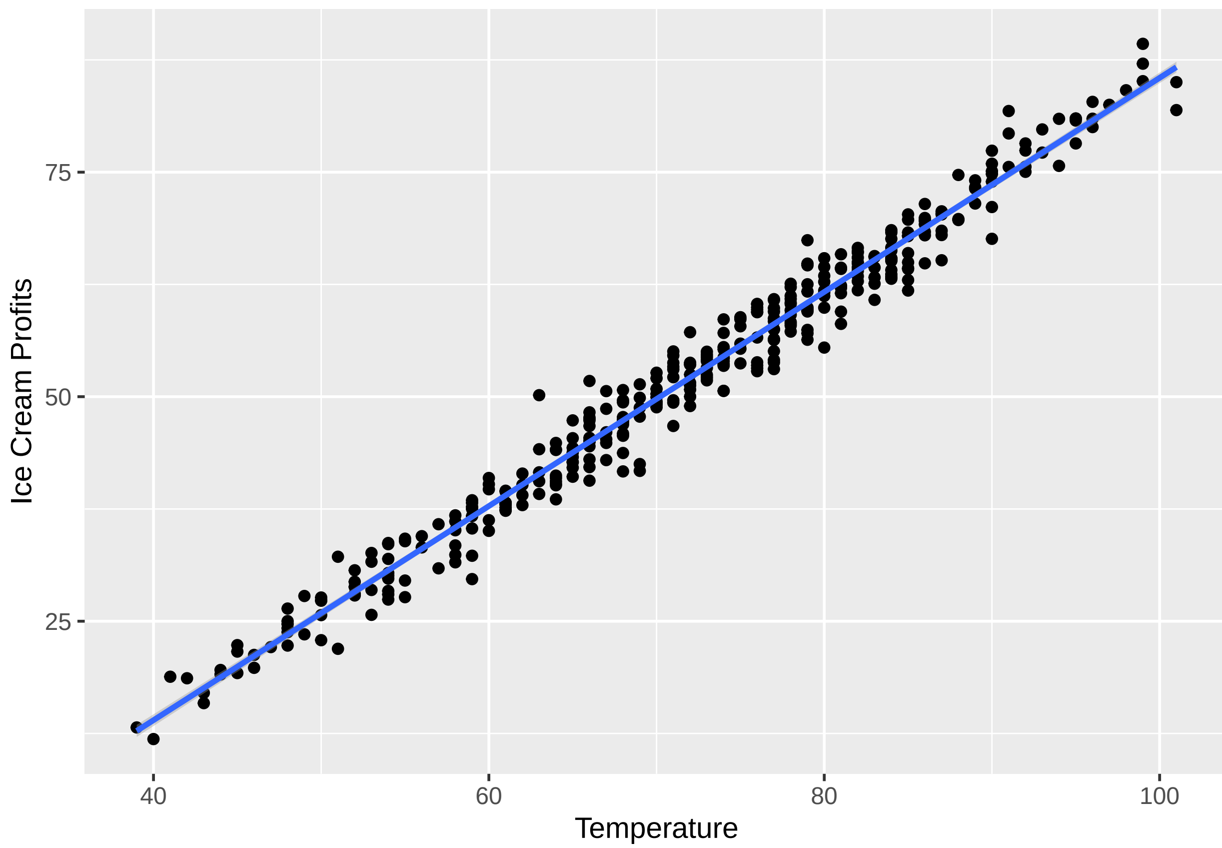
```r
plot(icSales$Temperature, ic_LM$residuals, main = "Residuals", xlab = "", ylab = "")
abline(0,0)
```




The plot of the residuals shows that all the data points cluster around the mean of 0, with relatively equal variability and an even distribution on both sides of the line. This indicates that the model is a good fit for this dataset.

```r
ggplot(icSales, aes(Temperature,`Ice Cream Profits`)) +
  geom_point() +
  stat_smooth(method = lm)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




**Conclusion**

The analysis of the data reveals a strong positive correlation between temperature and ice cream profits, supported by a correlation coefficient of approximately 0.99 and a statistically significant p-value close to 0.00. The linear regression model shows an R-squared value of 0.977, indicating that 97% of the variability in profits can be explained by the temperature variable, demonstrating a strong relationship between the two. The residual plot further validates the model's reliability, as the data points cluster evenly around a mean of 0 with relatively equal variability on both sides, suggesting that the model is a good fit and the data follows a linear trend.

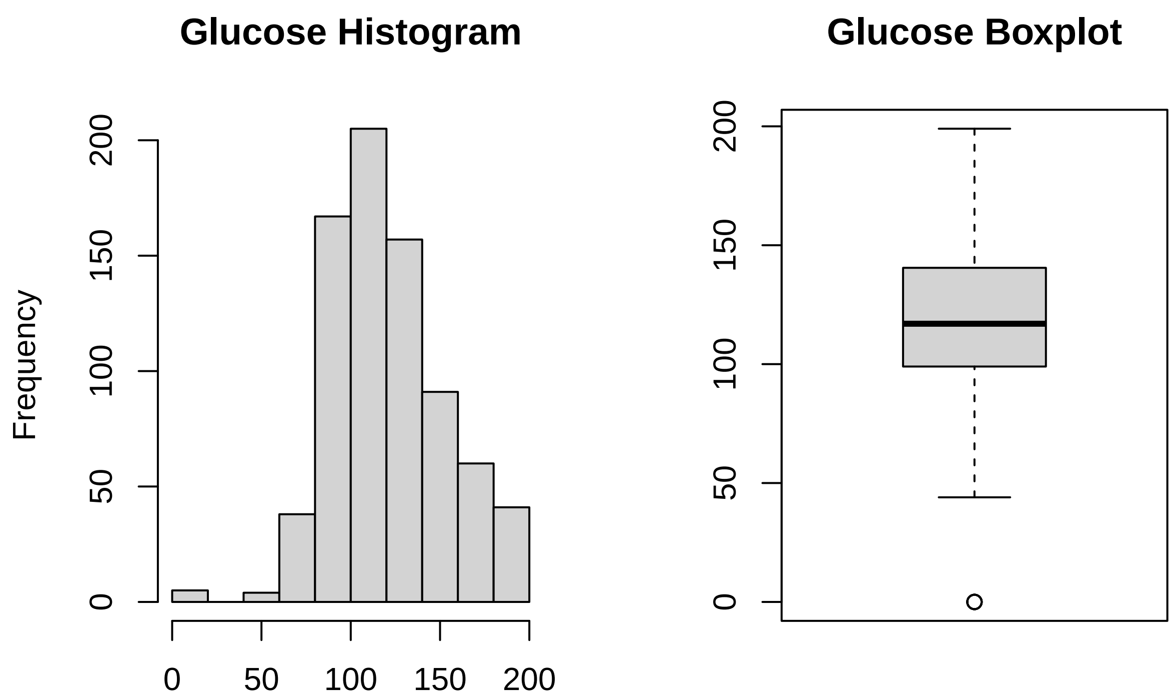
## 1) Diabetes

*predictor variable:* Glucose

*response variable:* Outcome

*Testing Normality:* **Glucose**

```r
par(mfrow = c(1, 2))
hist(d$Glucose,main = "Glucose Histogram",xlab = "")
boxplot(d$Glucose,main = "Glucose Boxplot")
```



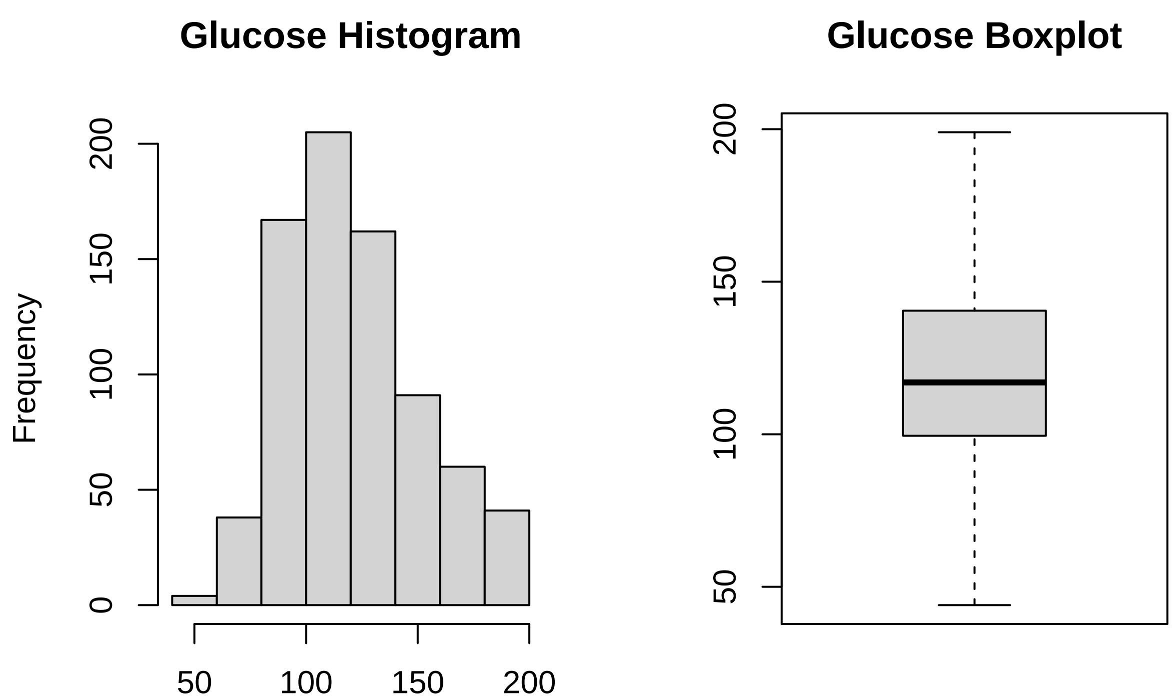

By exsamling the box and histogram for Glucose we can see the data has a heavy positive skew and an outlier at appromitly 0. by looking at a table of the data we can see there are five records where a glucose level of 0 was recorded. this is a recording error becasue a search will tell you that no living human can have a glucose level so low. for this reason, all 0 values were repalced with the data's mean. the proceeding graphs show the data is far more normal with a mean of approxomitly 125. altough the data is still skewed slightly to the right the data is now far more normal and a much better canadate for our analysis.

```r
table(d$Glucose)
```

```
##
##   0  44  56  57  61  62  65  67  68  71  72  73  74  75  76  77  78  79  80  81
##   5   1   1   2   1   1   1   1   3   4   1   3   4   2   2   2   4   3   6   6
##  82  83  84  85  86  87  88  89  90  91  92  93  94  95  96  97  98  99 100 101
##   3   6  10   7   3   7   9   6  11   9   9   7   7  13   8   9   3  17  17   9
## 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121
##  13   9   6  13  14  11  13  12   6  14  13   5  11  10   7  11   6  11  11   6
## 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141
##  12   9  11  14   9   5  11  14   7   5   5   5   6   4   8   8   5   8   5   5
## 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161
```

```
##   5   6   7   5   9   7   4   1   3   6   4   2   6   5   3   2   8   2   1   3
## 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181
##   6   3   3   4   3   3   4   1   2   3   1   6   2   2   2   1   1   5   5   5
## 182 183 184 186 187 188 189 190 191 193 194 195 196 197 198 199
##   1   3   3   1   4   2   4   1   1   2   3   2   3   4   1   1
```

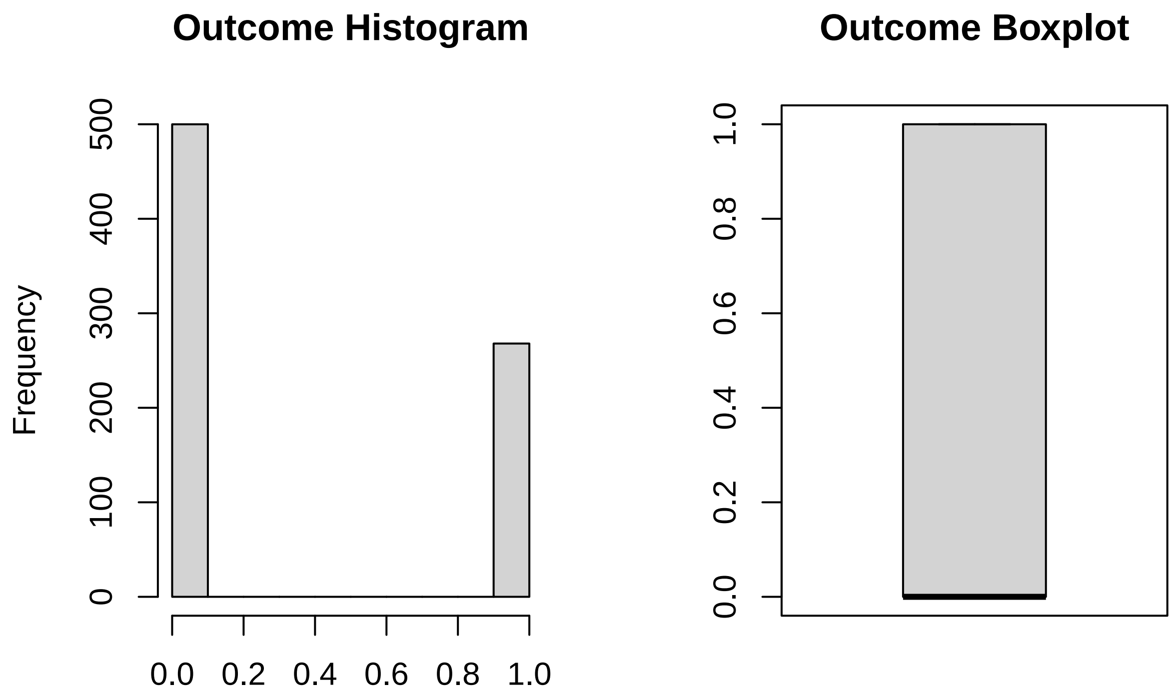
```r
d$Glucose[d$Glucose == 0] <- mean(d$Glucose)
par(mfrow = c(1, 2))
hist(d$Glucose,main = "Glucose Histogram",xlab = "")
boxplot(d$Glucose,main = "Glucose Boxplot")
```




By examining the boxplot and histogram for Glucose, we can see that the data initially has a heavy positive skew with an outlier at approximately 0. A closer look at the data reveals five records with a recorded glucose level of 0. This is a clear recording error, as a search confirms that no living human can have a glucose level that low. For this reason, all 0 values were replaced with the mean of the dataset.

The updated graphs show that the data distribution is now more normal, with a mean of approximately 125. Although the data still has a slight right skew, it is now far more normal and a much better candidate for analysis.

```r
par(mfrow = c(1, 2))
hist(d$Outcome,main = "Outcome Histogram",xlab = "")
boxplot(d$Outcome,main = "Outcome Boxplot")
```




```r
table(d$Outcome)
```

```
##
##   0   1
## 500 268
```

Upon review of the outcome variable, we can see that the data is categorical and denotes a binary outcome: 0 for patients without diabetes and 1 for patients with diabetes. This data is already coded as 0 and 1, so no changes are necessary. The histogram still provides valuable insight by showing that the majority of records represent individuals without diabetes, while a smaller number, approximately 380, are for those with the disease.

```r
D_LM <- lm(d$Outcome~d$Glucose)
summary(D_LM)
```

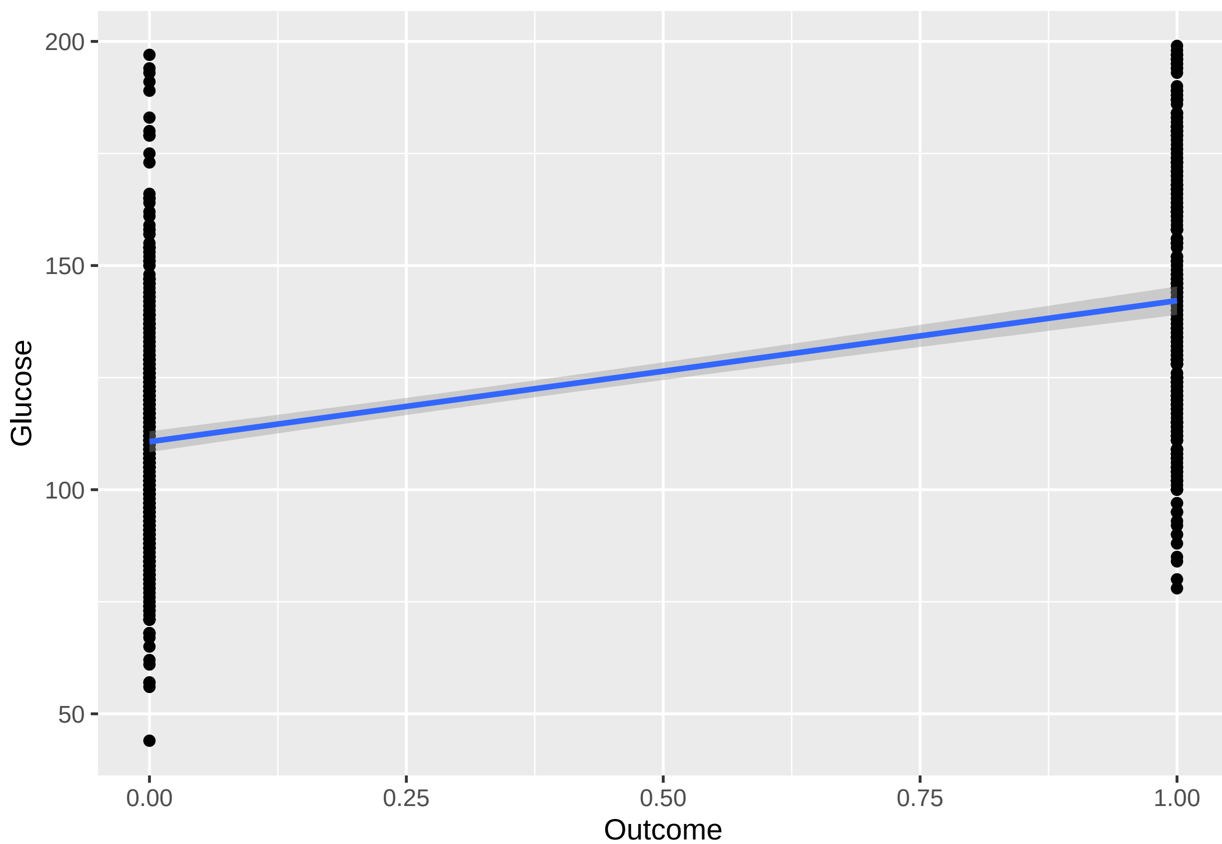
```
##
## Call:
## lm(formula = d$Outcome ~ d$Glucose)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9307 -0.2974 -0.1197  0.3338  0.9885
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.5909326  0.0617899  -9.564   <2e-16 ***
## d$Glucose    0.0077242  0.0004926  15.679   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4153 on 766 degrees of freedom
## Multiple R-squared:  0.243,  Adjusted R-squared:  0.242
## F-statistic: 245.8 on 1 and 766 DF,  p-value: < 2.2e-16
```

The Glucose coefficient of 0.0077 suggests that for every unit increase in Glucose, the predicted value of Outcome increases by 0.0077. This positive coefficient indicates a positive relationship between glucose levels and the likelihood of having diabetes. Additionally, the p-value for Glucose, approximately 0.00, indicates a highly significant relationship between Glucose and Outcome, reinforcing the likelihood that Glucose has a direct impact on diabetes outcomes.

```r
ggplot(d, aes(Outcome, Glucose)) +
  geom_point() +
  stat_smooth(method = lm)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



### Conclusion

The analysis indicates a significant positive relationship between glucose levels and diabetes outcomes. The Glucose coefficient of 0.0077 suggests that as glucose levels increase, the likelihood of a positive diabetes outcome also increases. The p-value of approximately 0.00 supports this relationship, showing that Glucose is a critical predictor of diabetes status.

# Full Code

```r
library(readr)
library(ggplot2)
icSales <- read_csv("Ice Cream Sales - temperatures.csv")
d <- read_csv("diabetes.csv")
par(mfrow = c(1, 2))
hist(icSales$`Ice Cream Profits`,main = "Ice Cream Profits Histogram",xlab = "")
boxplot(icSales$`Ice Cream Profits`,main = "Ice Cream Profits Boxplot")
par(mfrow = c(1, 2))
hist(icSales$`Temperature`,main = "Temperature Histogram",xlab = "")
boxplot(icSales$`Temperature`,main = "Temperature Boxplot")
plot(icSales$Temperature, icSales$`Ice Cream Profits`,xlab = "Temperature", ylab = "Profits")
abline(0,0)
cor.test(icSales$Temperature, icSales$`Ice Cream Profits`)
ic_LM <- lm(icSales$`Ice Cream Profits`~icSales$Temperature)
summary(ic_LM)
plot(icSales$Temperature, ic_LM$residuals, main = "Residuals", xlab = "", ylab = "")
abline(0,0)
ggplot(icSales, aes(Temperature,`Ice Cream Profits`)) +
  geom_point() +
  stat_smooth(method = lm)
par(mfrow = c(1, 2))
hist(d$Glucose,main = "Glucose Histogram",xlab = "")
boxplot(d$Glucose,main = "Glucose Boxplot")
table(d$Glucose)
d$Glucose[d$Glucose == 0] <- mean(d$Glucose)
par(mfrow = c(1, 2))
hist(d$Glucose,main = "Glucose Histogram",xlab = "")
boxplot(d$Glucose,main = "Glucose Boxplot")
par(mfrow = c(1, 2))
hist(d$Outcome,main = "Outcome Histogram",xlab = "")
boxplot(d$Outcome,main = "Outcome Boxplot")
table(d$Outcome)
D_LM <- lm(d$Outcome~d$Glucose)
summary(D_LM)
ggplot(d, aes(Outcome, Glucose)) +
  geom_point() +
  stat_smooth(method = lm)
```