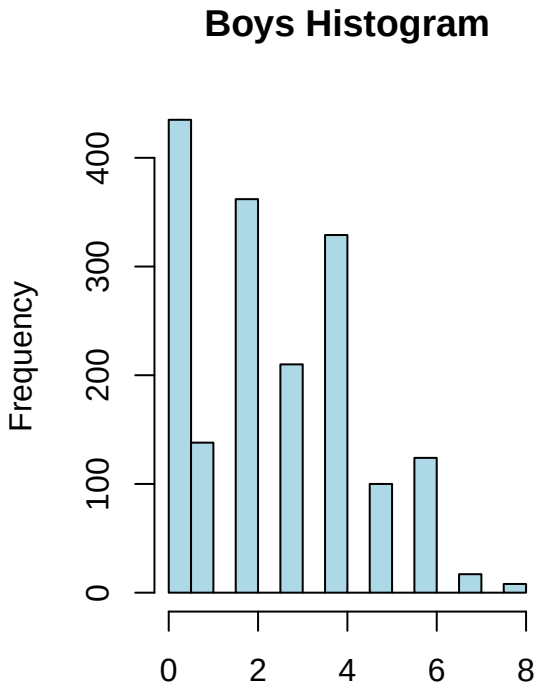
Bad Santa Regression Project

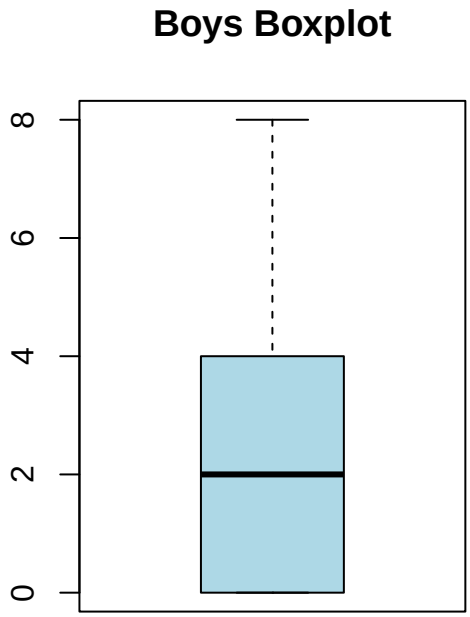
Alec Brooks

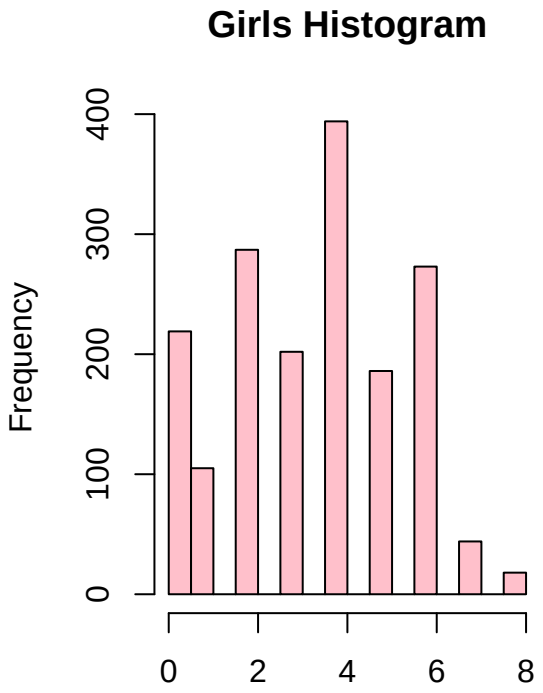
2024-11-23

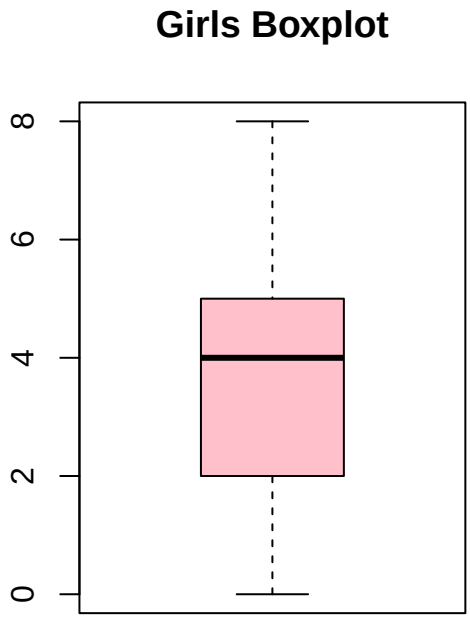
Mission 1

1.1 Is There is a significant difference in presents received for boys and girls?









Both the male and female distributions have no significant outliers. Although the male histogram shows a pronounced positive skew, both distributions will be treated as normal for this test. Visual inspection of the graphs suggests that the mean for girls is significantly higher than that for boys. To confirm whether this difference is statistically significant, we will perform a two-sided t-test.

Null Hypothesis: The mean number of presents gifted to boys is equal to the mean number of presents gifted to girls.

Alternative Hypothesis: The mean number of presents gifted to boys is not equal to the mean number of presents gifted to girls.

- μ_1 Mean number of presents received by boys.
- μ_2 Mean number of presents received by girls.

$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 \neq \mu_2$$

```
##
##  Welch Two Sample t-test
##
## data:  boys and girls
## t = -13.071, df = 24150, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.4284982 -0.3167418
## sample estimates:
## mean of x mean of y
##  1.993947  2.366567
```

With a p-value approximately 0, we can confidently reject the null hypothesis and conclude, at a 95% confidence level, that girls receive more gifts than boys. Based on the test results, girls receive approximately 0.4 more presents on average than boys. Next, we will explore whether this disparity is due to gender bias or influenced by other factors.

1.2 Does this bias appear to arise from gender discrimination or other factors?

```
##
## Call:
## lm(formula = presents ~ ., data = Santa)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3720 -0.7104  0.3289  0.9199  2.8949
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.373730   0.102998 -13.337   <2e-16 ***
## gender       0.075841   0.045713   1.659   0.0972 .
## letter       1.777561   0.045196  39.330   <2e-16 ***
## grades       0.845161   0.021219  39.831   <2e-16 ***
## selfbehav   -0.003708   0.022629  -0.164   0.8699
## SES          0.213727   0.007540  28.347   <2e-16 ***
## coal        -2.652194   0.080877 -32.793   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.271 on 3444 degrees of freedom
## Multiple R-squared:  0.6178, Adjusted R-squared:  0.6171
## F-statistic: 927.7 on 6 and 3444 DF,  p-value: < 2.2e-16
```

After removing the statistically insignificant factors, we are left with the following:

```
##
## Call:
## lm(formula = presents ~ letter + grades + SES + coal, data = Santa)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3461 -0.6986  0.3132  0.9306  2.8904
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.28809    0.07621  -16.90   <2e-16 ***
## letter       1.79804    0.04346   41.37   <2e-16 ***
## grades       0.84909    0.02109   40.26   <2e-16 ***
## SES          0.21379    0.00754   28.35   <2e-16 ***
## coal        -2.66521    0.08048  -33.12   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.271 on 3446 degrees of freedom
## Multiple R-squared:  0.6174, Adjusted R-squared:  0.617
## F-statistic:  1390 on 4 and 3446 DF,  p-value: < 2.2e-16
```

Considering the p-values of all possible factors, we observe that gender is not a statistically significant determinant of the number of presents children receive. Instead, other factors such as writing a letter, grades, socioeconomic status (SES), and receiving coal have a much stronger and more significant influence.

Examining the coefficients, the most impactful factor is whether the child receives coal, with a coefficient of -2.665, followed by whether a letter is written. By isolating the statistically significant factors and ranking them by their coefficients, we can produce the following table:

Table 1: Factors Ranked

Variable	Rank
coal	1
letter	2
grades	3
SES	4

Mission 2 (Child-Friendly Section)

After studying how Santa decides how many presents to give out, we learned girls get a little more presents than boys. But Santa doesn't play favorites just because you're a boy or a girl. Instead, other things matter more when Santa decides how many presents to give you.

For example:

Coal: If you have coal, you get a lot fewer presents.

Letters to Santa: Writing a letter to Santa helps a lot

Grades: Doing well in school helps.

Family Situation: If a family has more money, Santa tends to give the kids more presents too.

So, if you want more presents next year, write a letter to Santa, do well in school, and don't collect coal.

Full Code

```
library(readr)
library(knitr)
library(dplyr)
Santa <- read_csv("~/DataSci 220/Bad Santa: Regression Project/BadSanta.csv")
boys <- subset(Santa,Santa$gender == 1)
girls <- subset(Santa,Santa$gender == 2)
par(mfrow = c(1, 2))
hist(boys$presents,main = "Boys Histogram",xlab = "", col = "#ADD8E6")
boxplot(boys$presents,main = "Boys Boxplot", col = "#ADD8E6")
par(mfrow = c(1, 2))
hist(girls$presents,main = "Girls Histogram",xlab = "",col = "#FFC0CB")
boxplot(girls$presents,main = "Girls Boxplot",col = "#FFC0CB")
boys <- subset(Santa,Santa$gender == 1)
girls <- subset(Santa,Santa$gender == 2)
t.test(boys, girls, var.equal = FALSE)
summary(lm(presents ~., data = Santa))
summary(lm(presents ~letter+grades+SES+coal, data = Santa))
factors <- data.frame(Variable = c("coal", "letter", "grades", "SES"),Rank = c(1, 2, 3, 4))
kable(factors,col.names = c("Variable", "Rank"),caption = "Factors Ranked")
```