SDS291_FinalProject

Raley Long, Mattea Whitlow, Amelia Tarno, Rachel Lawson

Introduction:

Are Legos sexist? The term "pink tax" is widely used to explain the phenomenon of products aimed at women costing more than similar products aimed at men. In a study published by the Federal Commission Trade, a group of female researchers looked at the verity of the pink tax in the personal care product sector and did not find evidence to support the claim that women do indeed face higher costs than their male counterparts.

In this data analysis, we will be looking at the same question in the toy sector, specifically the brand Lego as this is a large international company with global reach and potential impact on consumers. We will be asking and attempting to answer the question: do feminine Lego sets tend towards higher price tags than their more masculine counterparts? In order to determine the answer, we will create a model including pieces per set, price per set, and gender of set. We hypothesize that feminine lego sets are more expensive on average than masculine or neutral ones. Our null hypothesis is that the gendered models are not necessary and the nested model that only considers price as a function of pieces is enough.

In a modern world on its way to achieving gender equality, it is becoming increasingly important to recognize where inequalities exist in order to work towards eliminating them. By determining if prices differ across a gender spectrum, companies can alter their marketing and pricing tactics to better reflect their audiences.

Methods:

The dataset we will be analyzing uses data from brickset.com, an online resource for Lego fans that compares various prices found online through different retailers for a specific set of Legos. It contains 1,305 observations of fifteen variables. For our analysis, we considered the variables set name, set theme, and total number of pieces in the set, and we also created the variables "masc", "fem", and "neutral", categorical variables with categories "Yes" and "No" based on our interpretation of the set theme as one of the three options. For our explanatory variable we selected pieces per set, a numerical variable measured in integers ranging from 1 to 6020, and filtered by gender which further divided the observations into our three categorical variables.

Our response variable is the price of Lego sets, measured in US dollars, a numeric variable with values ranging from 1.99 to 699.99. In order to address the question of whether Lego set prices depend on the set's targeted gender category, we fit four models. Our first, basic model looks at price as a function of pieces. Our other models consider price as a function of the interaction between pieces and gender category, one for each "masc", "fem", or "neutral" as we defined them in our modified dataset. In evaluating the assumptions of this model, we found that all assumptions are met except normality, which is violated. Because all models violate the normality condition at the extremity points and we are primarily performing a model comparison, this violation carries through and our model comparison should still hold relatively well. This will impact generalizability to all Lego sets, making our conclusions less generalizable.

Data Analysis:

Importing Packages

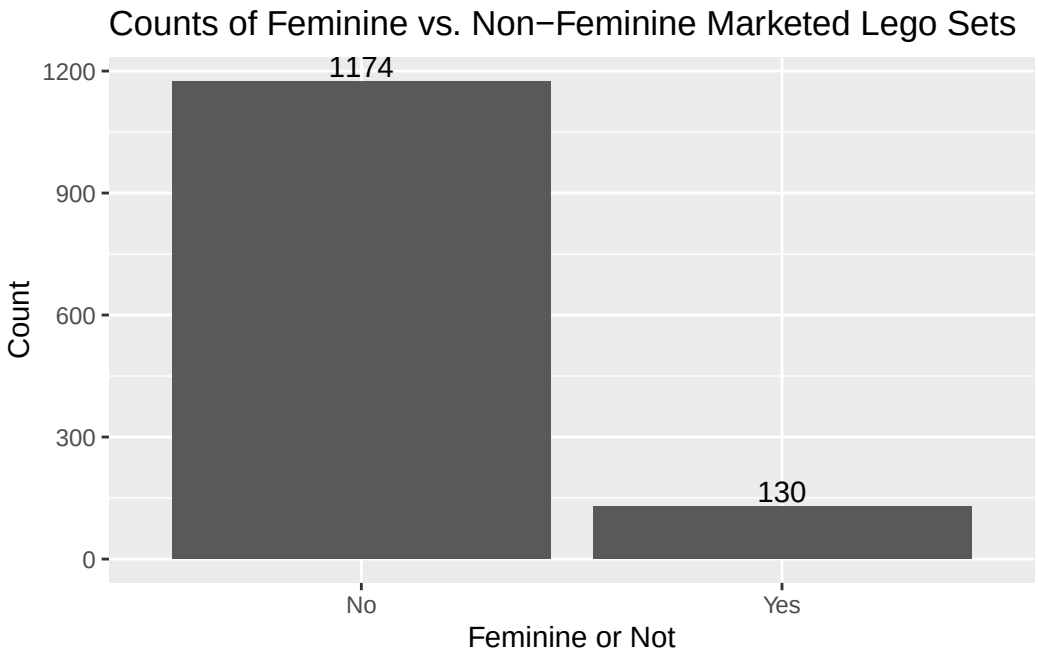
```
library(moderndive)
library(dplyr)
library(tidyverse)
library(ggplot2)
library(Stat2Data)
```

Data Importing

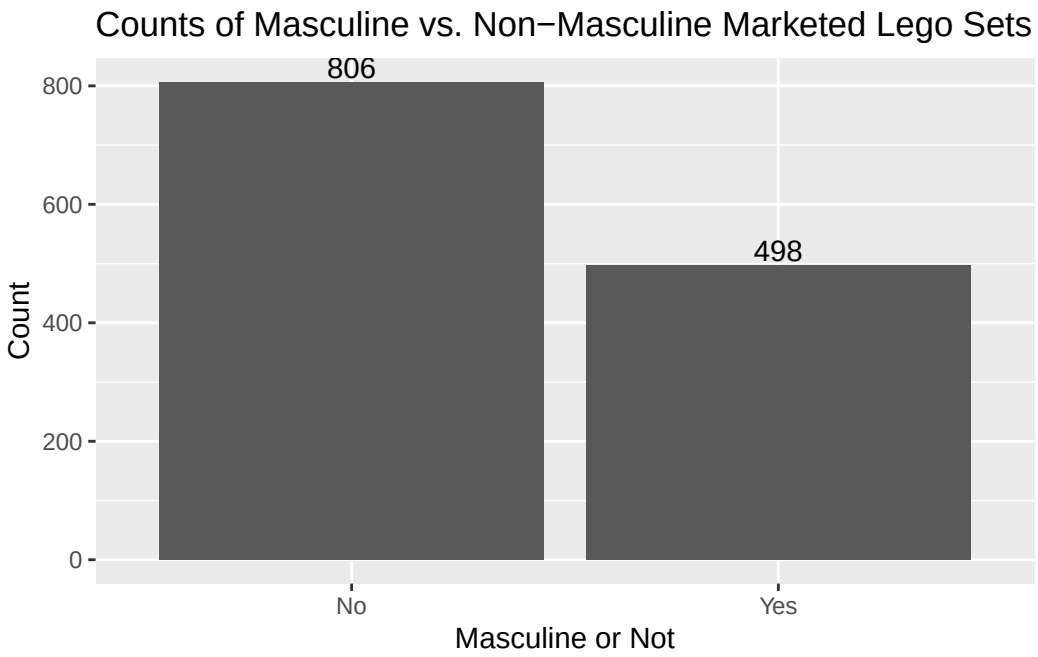
```
lego_clean <- read.csv("lego_clean.csv")
lego_clean$fem <- factor(lego_clean$fem, labels = c("No", "Yes"))
lego_clean$masc <- factor(lego_clean$masc, labels = c("No", "Yes"))
lego_clean$neutral <- factor(lego_clean$neutral, labels = c("No", "Yes"))
```

Exploratory Visualizations

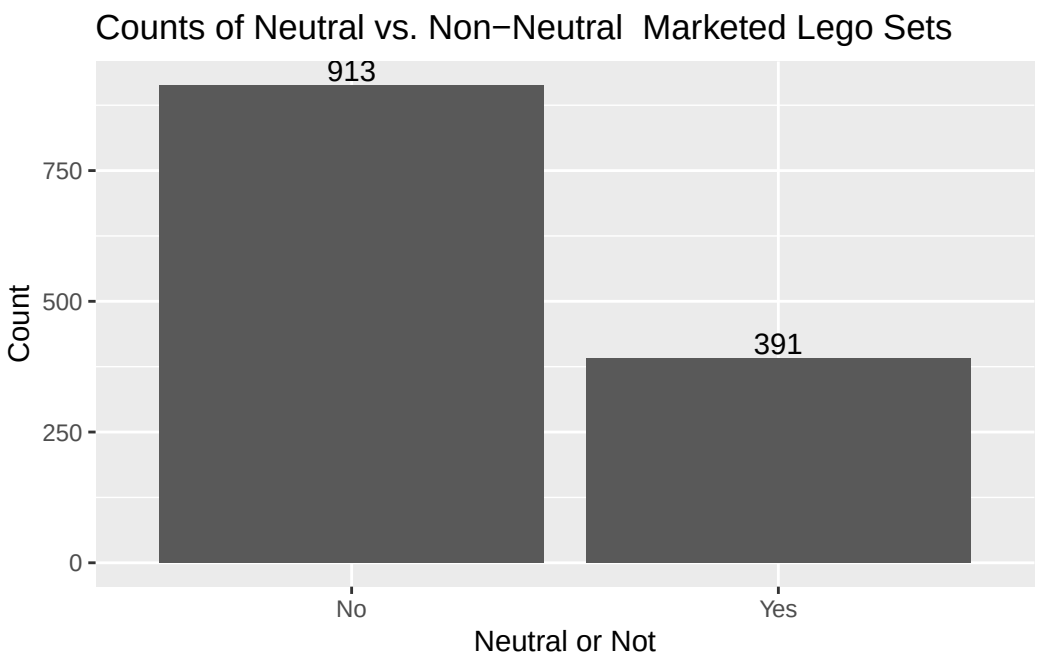
```
ggplot(lego_clean, aes(x = fem)) + geom_bar() + geom_text(stat = 'count', aes(label=after_
```




```
ggplot(lego_clean, aes(x = masc)) + geom_bar() + geom_text(stat = 'count', aes(label=after
```

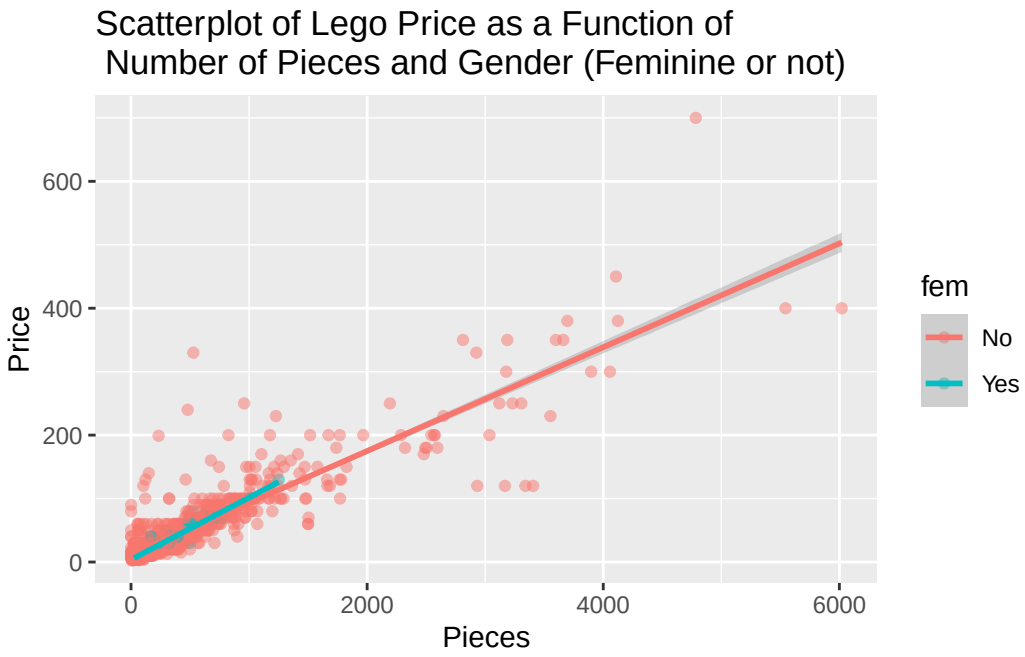


```
ggplot(lego_clean, aes(x = neutral)) + geom_bar() + geom_text(stat = 'count', aes(label=af
```

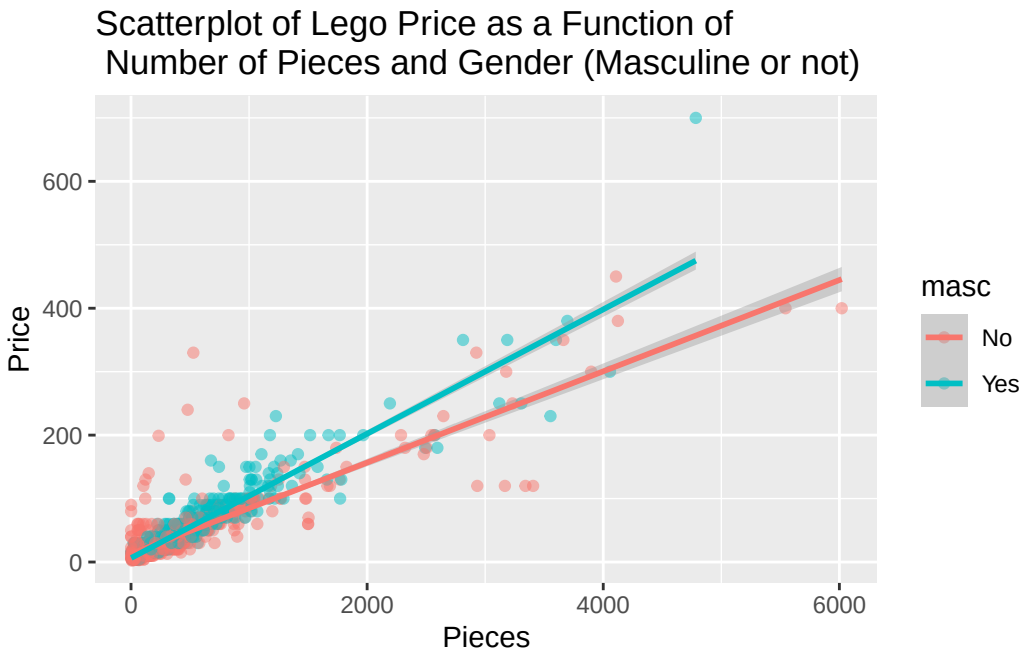


Models:

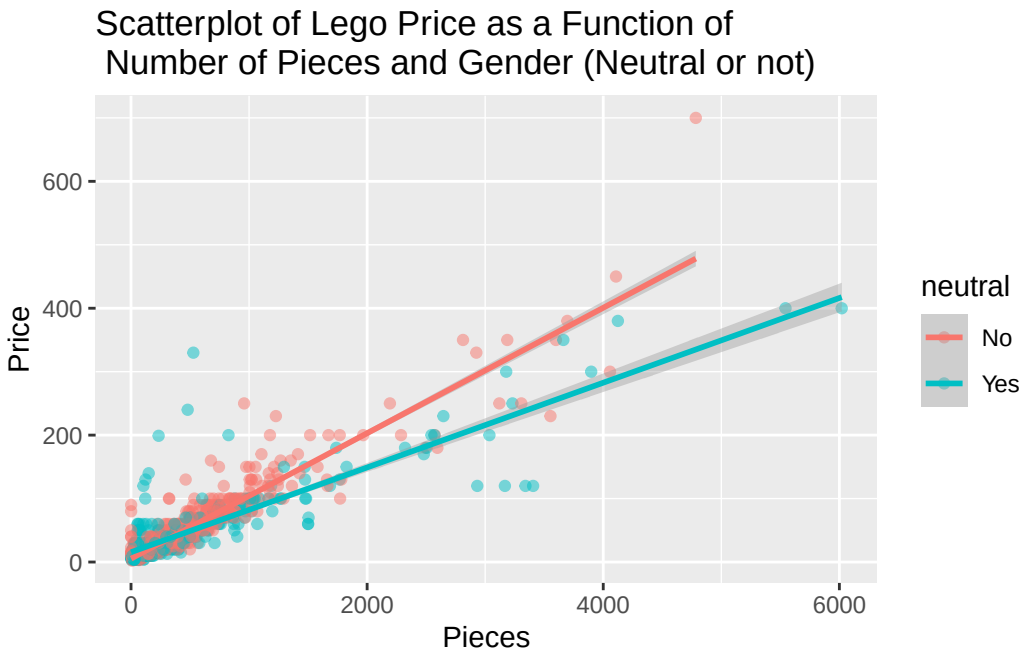
```
price_piece_fem <- lm(Price ~ Pieces * fem, lego_clean)
ggplot(lego_clean, aes( x = Pieces, y = Price, color = fem)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm") +
  labs(title = "Scatterplot of Lego Price as a Function of \n Number of Pieces and Gender
```



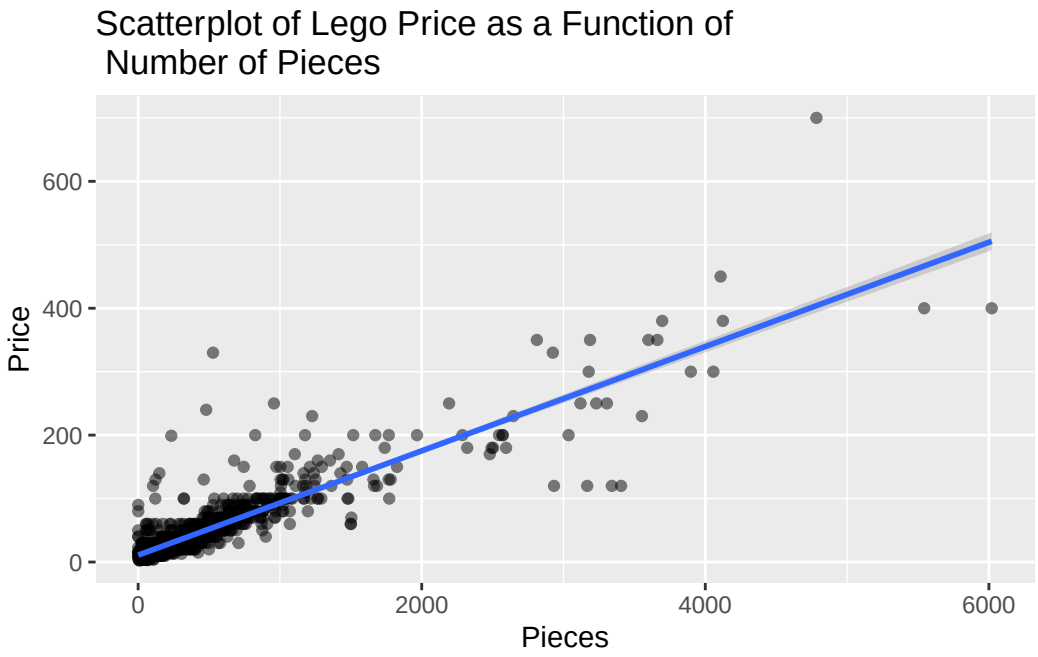

```
price_piece_masc <- lm(Price ~ Pieces * masc, lego_clean)
ggplot(lego_clean, aes( x = Pieces, y = Price, color = masc)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm") +
  labs(title = "Scatterplot of Lego Price as a Function of \n Number of Pieces and Gender
```




```
price_piece_neutral <- lm(Price ~ Pieces * neutral, lego_clean)
ggplot(lego_clean, aes( x = Pieces, y = Price, color = neutral)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm") +
  labs(title = "Scatterplot of Lego Price as a Function of \n Number of Pieces and Gender
```

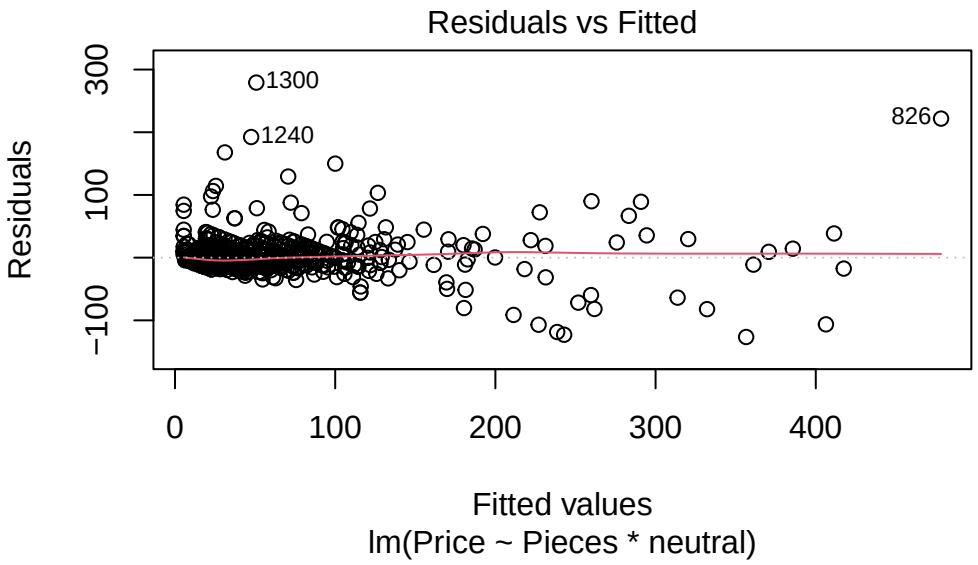


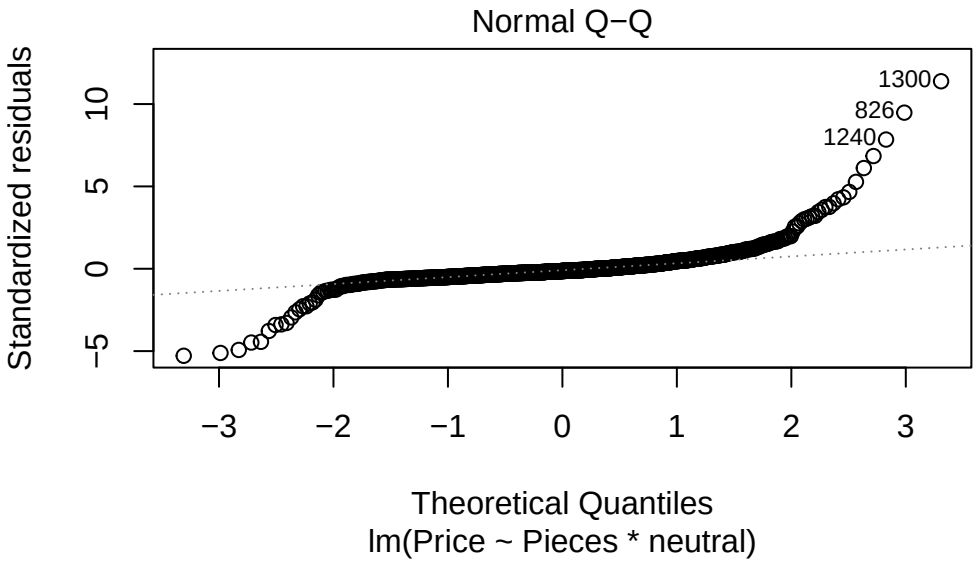

```
price_piece <- lm(Price ~ Pieces, data = lego_clean)
ggplot(lego_clean, aes( x = Pieces, y = Price)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm") +
  labs(title = "Scatterplot of Lego Price as a Function of \n Number of Pieces")
```

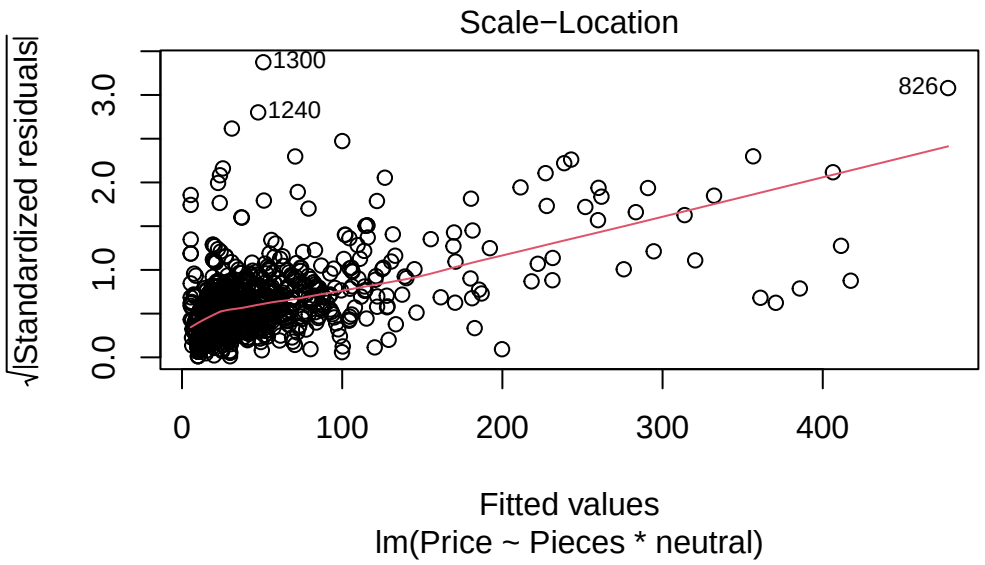


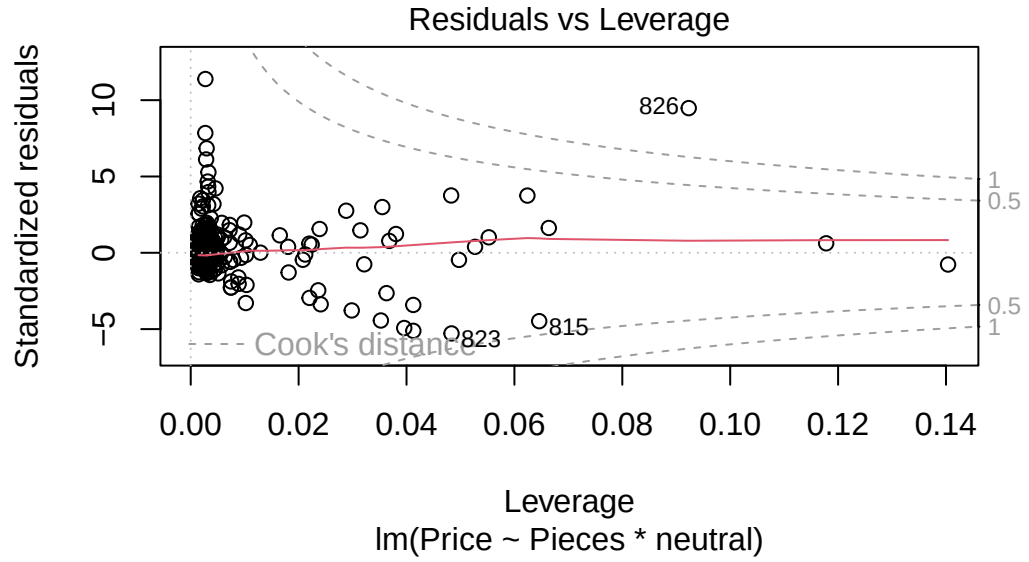
Condition Check

```
plot(price_piece_neutral)
```

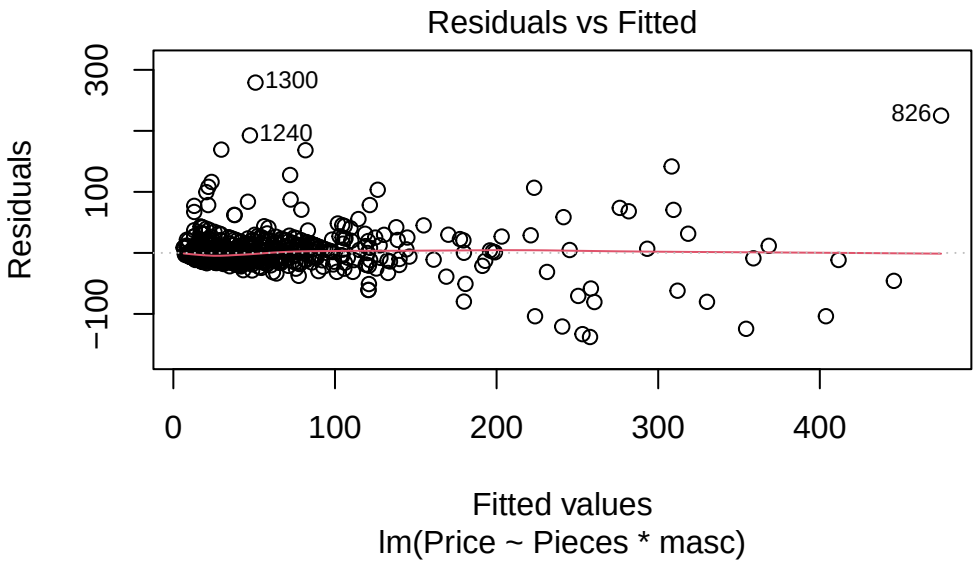

Residuals vs Fitted
Residuals
1300
1240
826
300
100
−100
0
100
200
300
400
Fitted values
lm(Price ~ Pieces * neutral)

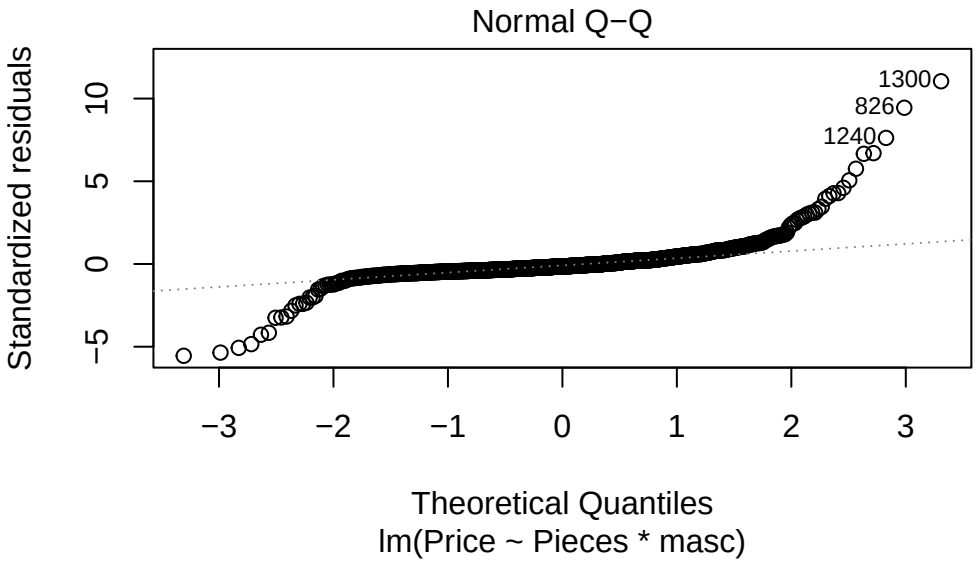

Normal Q−Q
Standardized residuals
1300
826
1240
10
5
0
−5
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(Price ~ Pieces * neutral)

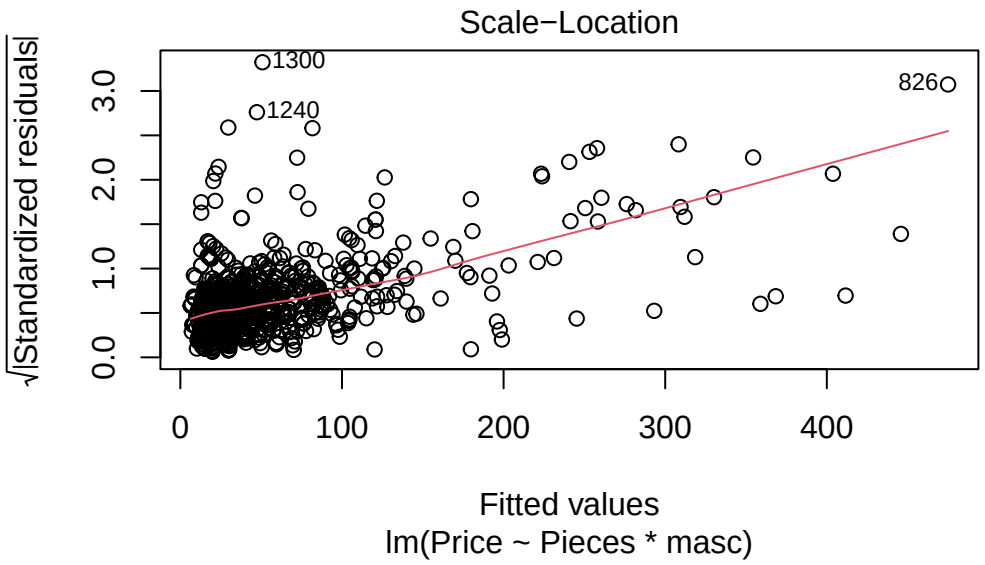


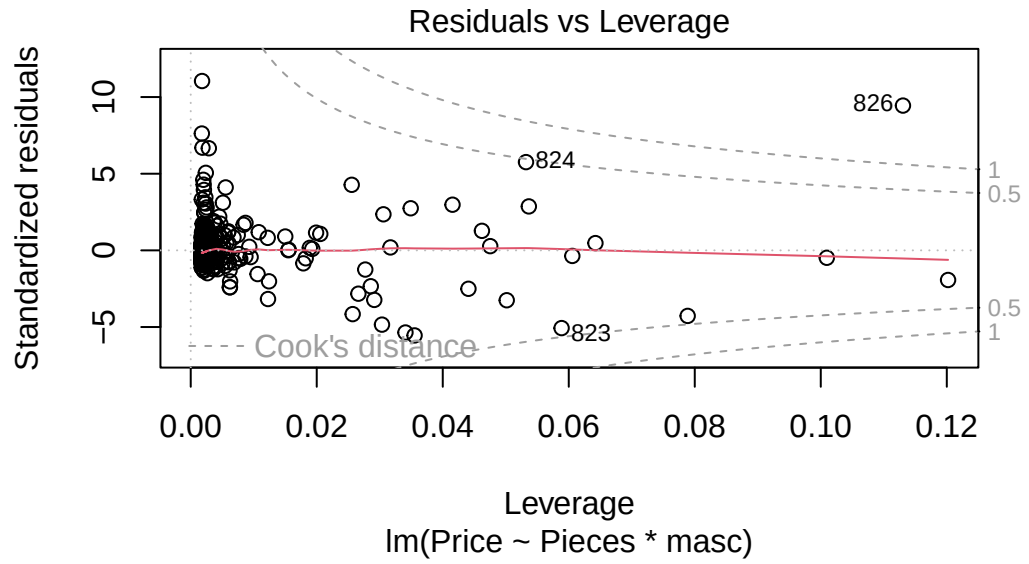



```
plot(price_piece_masc)
```

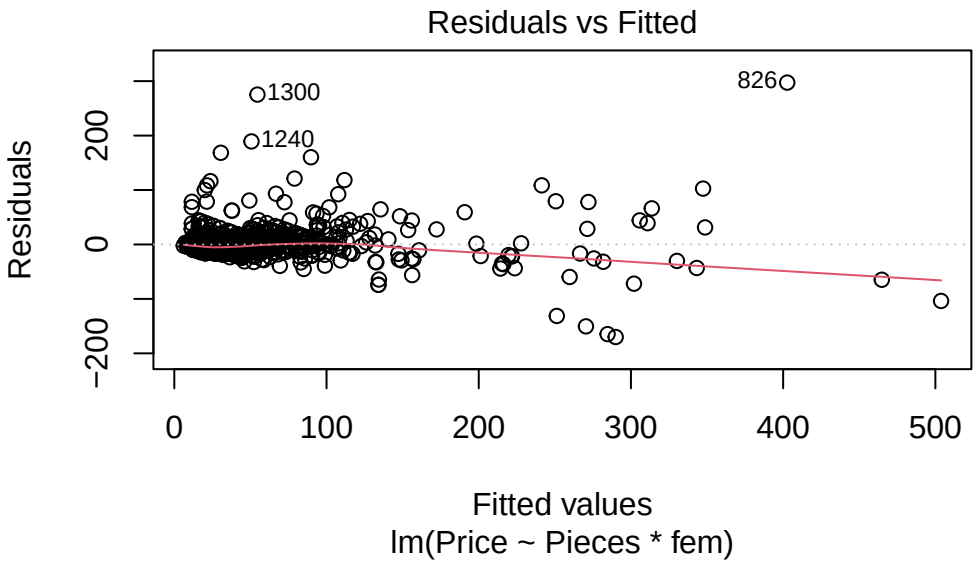

Residuals vs Fitted
Residuals
Fitted values
lm(Price ~ Pieces * masc)
1300
1240
826
300
100
−100
0
100
200
300
400

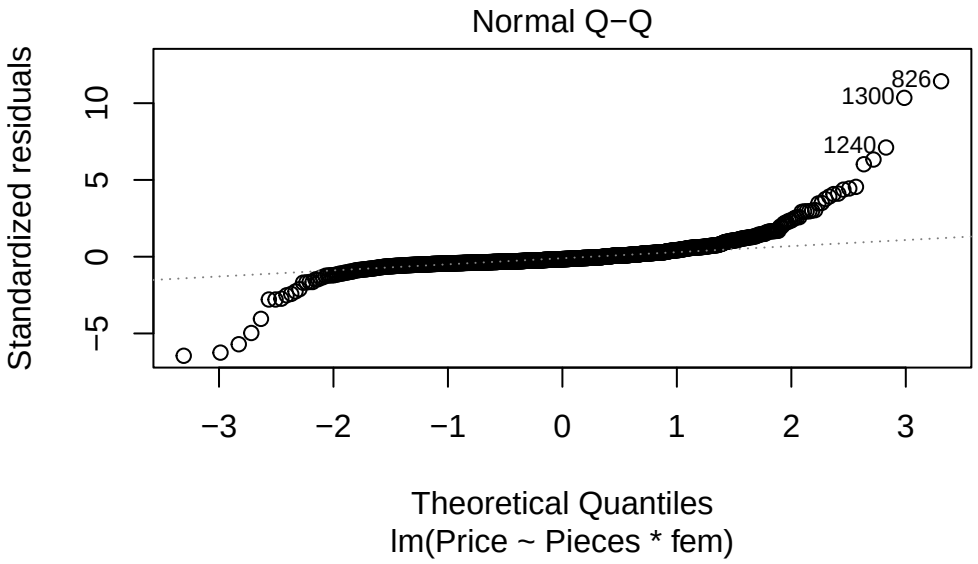

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Price ~ Pieces * masc)
1300
826
1240
10
5
0
−5
−3
−2
−1
0
1
2
3

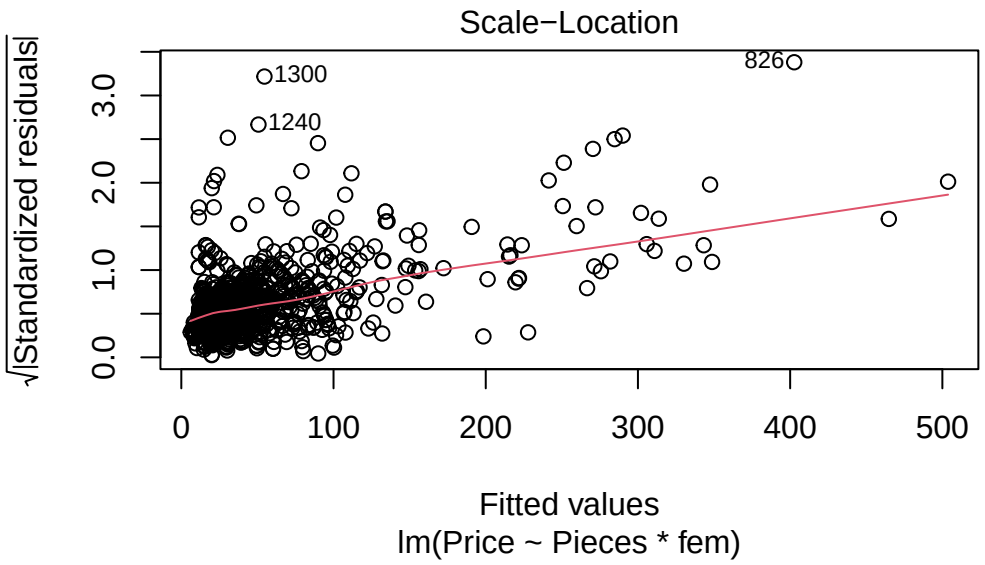


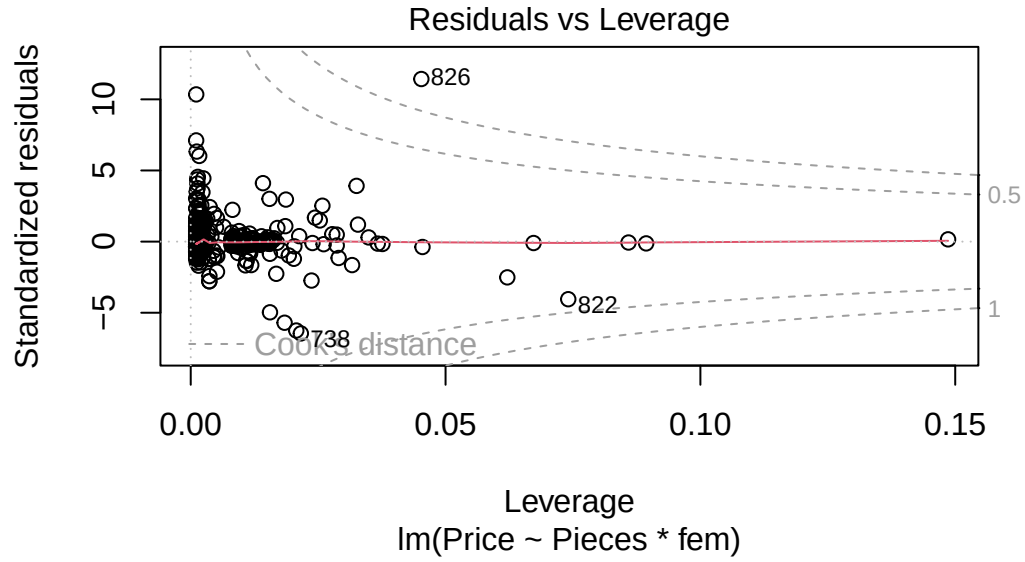



```
plot(price_piece_fem)
```

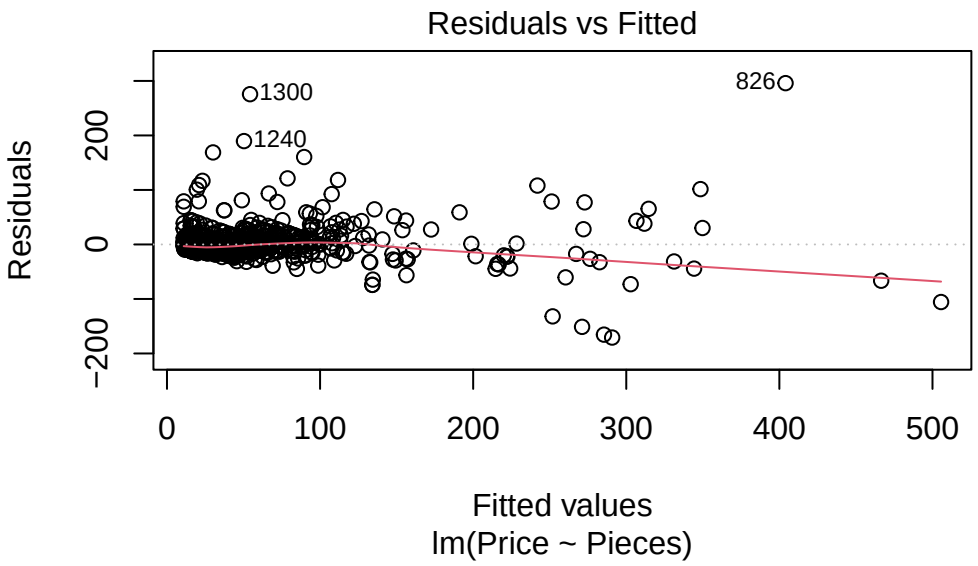

Residuals vs Fitted
Residuals
Fitted values
lm(Price ~ Pieces * fem)
1300
1240
826
−200
0
200
0
100
200
300
400
500

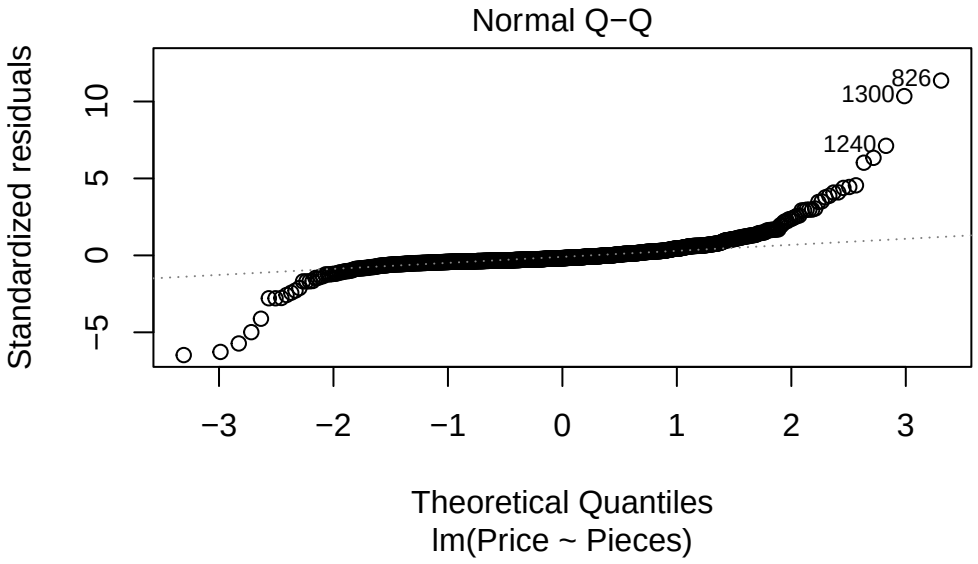

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Price ~ Pieces * fem)
826
1300
1240
−5
0
5
10
−3
−2
−1
0
1
2
3

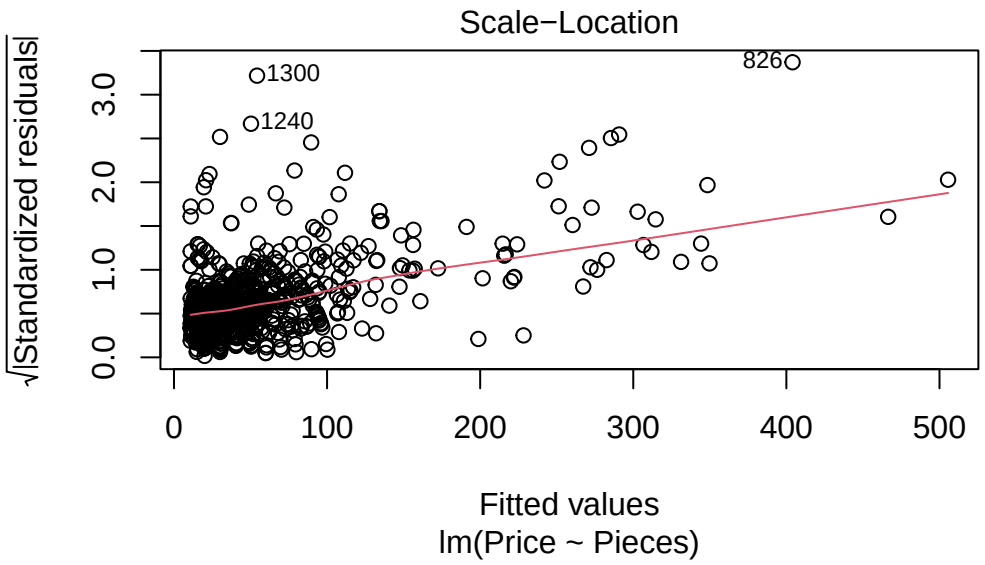

Scale−Location
√|Standardized residuals|
1300
1240
826
0.0
1.0
2.0
3.0
0
100
200
300
400
500
Fitted values
lm(Price ~ Pieces * fem)

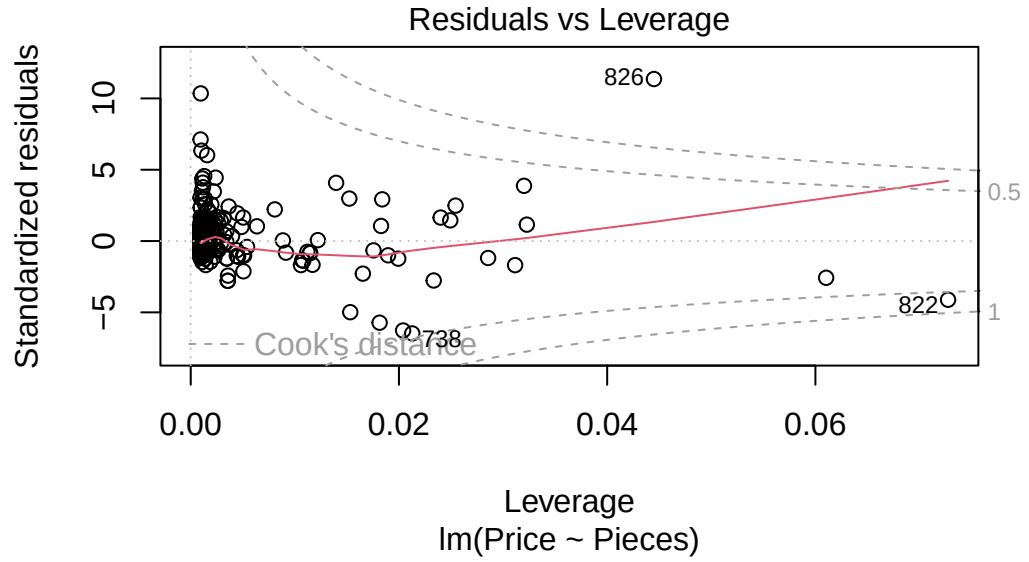

Residuals vs Leverage
Standardized residuals
826
822
738
Cook's distance
0.5
1
−5
0
5
10
0.00
0.05
0.10
0.15
Leverage
lm(Price ~ Pieces * fem)

```r
price_piece <- lm(Price ~ Pieces, data = lego_clean)
plot(price_piece)
```









All models violate the normality condition at the extremity points. Since we are primarily doing a model comparison and this violation carries through, the comparison should still hold

relatively well. This will however make our findings much less generalizable.

Regression Summaries:

```
get_regression_table(price_piece)
```

```
# A tibble: 2 x 7
  term      estimate std_error statistic p_value lower_ci upper_ci
  <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept   10.9       0.986      11.0       0     8.93   12.8
2 Pieces       0.082     0.001      64.3       0     0.08    0.085
```

```
get_regression_table(price_piece_fem)
```

```
# A tibble: 4 x 7
  term          estimate std_error statistic p_value lower_ci upper_ci
  <chr>            <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept       11.4       1.05      10.9    0        9.37    13.5
2 Pieces           0.082     0.001     63.3    0        0.079    0.084
3 fem: Yes        -7.69      3.80      -2.02   0.043  -15.2     -0.228
4 Pieces:femYes    0.016     0.01       1.56   0.12    -0.004    0.036
```

```
get_regression_table(price_piece_masc)
```

```
# A tibble: 4 x 7
  term           estimate std_error statistic p_value lower_ci upper_ci
  <chr>             <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept        12.8       1.23      10.4    0       10.4     15.3
2 Pieces            0.072     0.002     46.4    0        0.069    0.075
3 masc: Yes        -6.62      1.90      -3.48   0.001  -10.4     -2.89
4 Pieces:mascYes    0.026     0.002     10.5    0        0.021    0.031
```

```
get_regression_table(price_piece_neutral)
```

```
# A tibble: 4 x 7
  term              estimate std_error statistic p_value lower_ci upper_ci
  <chr>                <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept            5.35      1.16       4.59       0    3.06     7.63
2 Pieces               0.099     0.002     58.4        0    0.096    0.102
3 neutral: Yes        10.2       1.89       5.37       0    6.45    13.9
4 Pieces:neutralYes   -0.032     0.002    -13.6        0   -0.037   -0.027
```

Model Comparison

```
anova(price_piece, price_piece_neutral)
```

```
Analysis of Variance Table

Model 1: Price ~ Pieces
Model 2: Price ~ Pieces * neutral
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   1063 755013
2   1061 639726  2    115287 95.603 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- This nested f- test concludes that whether a set is marked as gender neutral or not has an effect on price (the full model is better the fit). Gender neutrality is a necessary component of the model

```
anova(price_piece, price_piece_masc)
```

```
Analysis of Variance Table

Model 1: Price ~ Pieces
Model 2: Price ~ Pieces * masc
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   1063 755013
2   1061 678766  2     76248 59.593 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- This nested f-test concludes that whether a set is marked as masculine or not has an effect on price (the full model is a better fit). Masculinity is a necessary component of the model.

```
anova(price_piece, price_piece_fem)
```

```
Analysis of Variance Table

Model 1: Price ~ Pieces
Model 2: Price ~ Pieces * fem
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1   1063 755013
2   1061 752110  2    2903.1 2.0477 0.1295
```

- Since the p-value is above the threshold of 0.05, we fail to reject the null hypothesis and conclude the reduced model (price_piece) explains more variability in the data than a model that also accounts for feminine marketing.

Comparing The Interaction Models to Each other

```
get_regression_summaries(price_piece_neutral)
```

```
# A tibble: 1 x 9
  r_squared adj_r_squared   mse  rmse sigma statistic p_value    df  nobs
      <dbl>         <dbl> <dbl> <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
1     0.827         0.826  601.  24.5  24.6     1688.       0     3  1065
```

```
get_regression_summaries(price_piece_masc)
```

```
# A tibble: 1 x 9
  r_squared adj_r_squared   mse  rmse sigma statistic p_value    df  nobs
      <dbl>         <dbl> <dbl> <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
1     0.816         0.816  637.  25.2  25.3     1570.       0     3  1065
```

Results:

Through this analysis, we concluded that the intended gender demographic of Lego sets does not have a significant impact on the price of the set. We fail to reject the null hypothesis and conclude that the reduced model explains more variability in the dataset than a model accounting for feminine marketing. When comparing the results of our models by using get_regression_summaries(price_piece_neutral) and get_regression_summaries(price_piece_masc), the adjusted r squared values indicate that

the neutral model is a better predictor of price when controlling for pieces. However, the difference in the adjusted r squared values is very low (0.01). Since the number of lego sets marketed to a gender neutral audience is much higher than ones marked to a masculine audience, it is likely that both of these models do a similarly good job of predicting price when controlling for number of pieces. We compared our nested model to the full models, which look at both number of pieces and gender category, in order to determine which had more influence on price using a nested f-test. We found that whether a set is marked as gender neutral or not has an effect on price, and thus gender neutrality is a necessary component of the model. Additionally, whether a set is marked as masculine or not has an effect on price, and thus masculinity is a necessary component of the model. When considering the "feminine" interaction model, we fail to reject the null hypothesis and conclude the reduced model (price_piece) explains more variability in the dataset than a model accounting for feminine marketing. As for coefficients, our model's intercept tells us that for a non-masculine model with 0 pieces, the predicted price is $12.84. "Pieces" tells us that for each additional piece in a non-masculine set, the predicted price increase is $0.07. "Masc: yes" tells us that for a masculine model with 0 pieces, the predicted price is $-6.62. "Pieces: mascYes" tells us that for each additional piece in a masculine set, the predicted price increase is $0.03. These values are quite small and therefore have no significant impact on Lego set prices. Thus, we conclude that the gender of Lego sets did not provide any additional explanatory power beyond what was already conveyed by the number of pieces.

Discussion:

Overall, when considering if feminine Lego sets had higher average prices than masculine ones, we found that gendered marketing had little impact on Lego set price. Despite our conclusions, however, our findings are limited to the dataset we analyzed, which does not account for Lego sets outside of the time period from 2018-2020. Additionally, we used our personal judgment in order to categorize each Lego set as being masculine, feminine, or neutral, categorizations which may be disputed by those with differing opinions about which set should be assigned to each gendered group. Given the scope of the project, our analyses took into account important factors such as gender in order to draw conclusions about price disparities in children's toy marketing. We also accurately differentiated explanatory variables in order to determine which ones most significantly influenced the outcome variables, fitting interaction models for not just one but all of our gender categories.