Rachel Kindangen
Professor Simonoff
Regression and Multivariate Data Analysis

March 2, 2022

**Valuing Vinyl Records**

Music is streamed through various mediums, including digital audio files, streaming platforms, physical CDs and vinyls. The vinyl industry, in particular, has seen a growth in popularity despite the prevalence of digital streaming platforms. The years 2016 - 2017 saw a 18.5% YOY increase in sales of vinyls.

To examine this assumption, data was scraped and analyzed from the website Discogs. Discogs is an online database providing consumers with the ability to appraise and value their vinyl records using pre-existing data. A sample of 278 observations was used to conduct the analyses. The variables collected included:

- Total_albums_artist: Number of albums the artist of the vinyl has released
- Users_have: Number of unique Discogs users who own the record
- Users_want: Number of unique Discogs users who want the record or saved the record to their wishlist
- User_rating: The average user rating for the vinyl
- Total_rating: The number of ratings for the vinyl
- Lowest_price: The lowest reported price of the vinyl on Discogs
- Median_price: The median reported price of the vinyl on Discogs
- Highest_price: The highest reported price of the vinyl on Discogs
- Total_versions: The number of repressings of the vinyl

Vinyl enthusiasts claim that records which are particularly rare or difficult to come by are priced higher in the vinyl market. This assumption can be examined using a simple regression model:

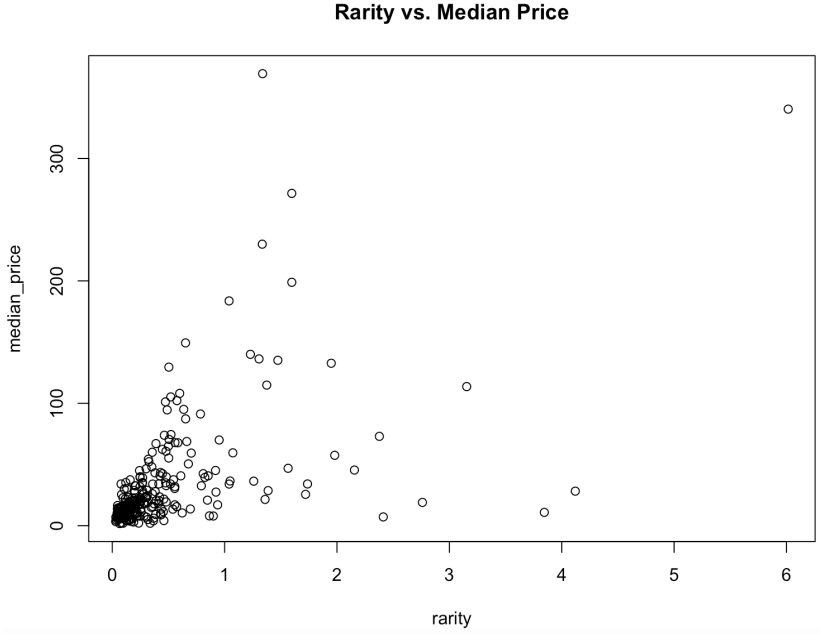
$$Vinyl\ Record\ Price\ =\ Rarity\ *\ \beta_1\ +\ \beta_0\ +\ Random\ error$$

To quantify the vinyl's rarity, a second quantitative predictor was created using the existing predictors. A vinyl's rarity will be quantified as:

$$\frac{\#\ Users\ who\ want\ the\ record}{\#\ Users\ who\ have\ the\ record}$$

A vinyl with a higher rarity value indicates that the vinyl is more rare. It suggests that more users want the record in proportion to the users who have the record. If the predictor for rarity correctly values a vinyl's rarity, then $\beta_1$ would be positive in the simple regression model.

First, we examine a scatterplot of the two variables, with rarity on the x-axis and median price on the y-axis:




The scatterplot shows a nonlinear relationship with most values concentrated in the lower bounds of the x and y values. This indicates the presence of extreme outliers and leverage points. In particular, there seems to be several outliers for values with median price higher than $250. Moreover, there seems to be leverage points for rarity values higher than 2.

Without modifying the dataset, we build a simple linear regression model using median price as the target variable and rarity as the predictor. We obtain:

```
Call:
lm(formula = median_price ~ rarity)

Residuals:
     Min       1Q   Median       3Q      Max
-151.199  -12.007   -6.356    2.184  303.025

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   15.171      2.737   5.542 6.98e-08 ***
rarity        38.203      3.470  11.010  < 2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 37.86 on 276 degrees of freedom
Multiple R-squared:  0.3052,  Adjusted R-squared:  0.3026
F-statistic: 121.2 on 1 and 276 DF,  p-value: < 2.2e-16
```

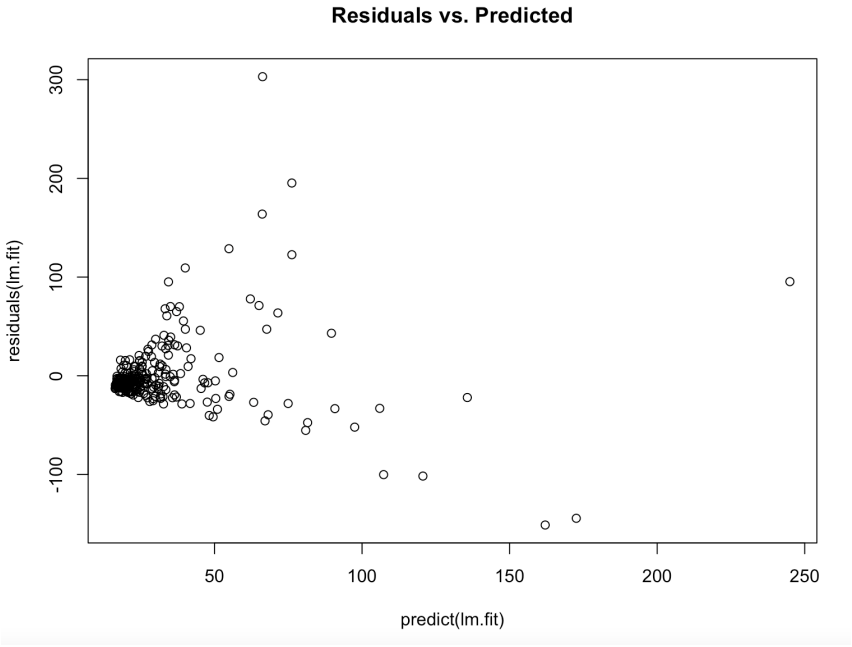
The regression equation is:

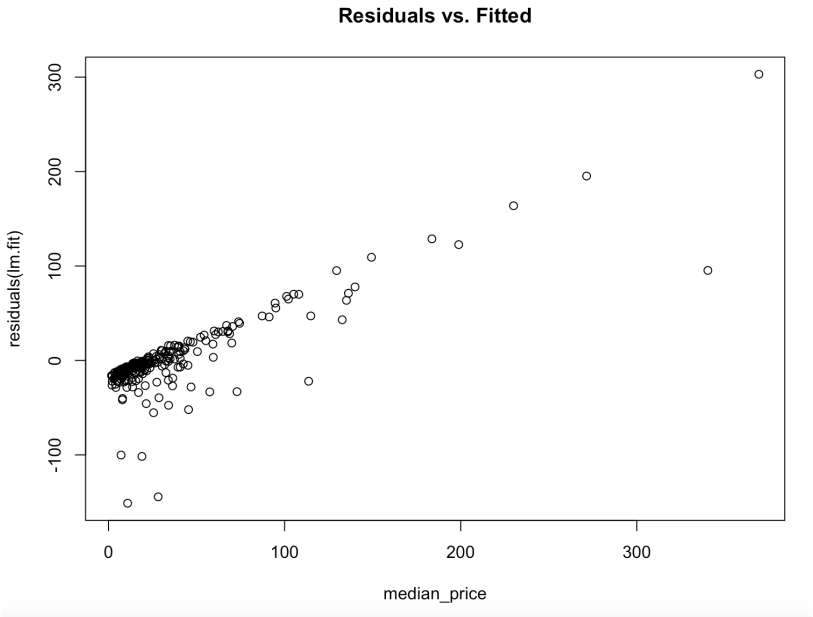
```
Median price = 15.171 + 38.203 Rarity
```

The regression is weak, with an adjusted R-squared of 30.26%. This indicates that only 30.26% of the observed variability in median prices is explained by the rarity value. The regression states that given that the rarity value is zero, or there are zero users who want the record, the median price of the vinyl record will be valued at $15.17. The regression also states that a single unit increase in the rarity value is associated with an increase of $38.20 in the median price of the vinyl record. A single unit increase in rarity value is equivalent to an increase in the number of users who want the record by the current number of users who have the record. As an example, given that 4 users want the record and 10 users have the record, a single unit increase in the rarity value occurs when 10 additional users want the record.

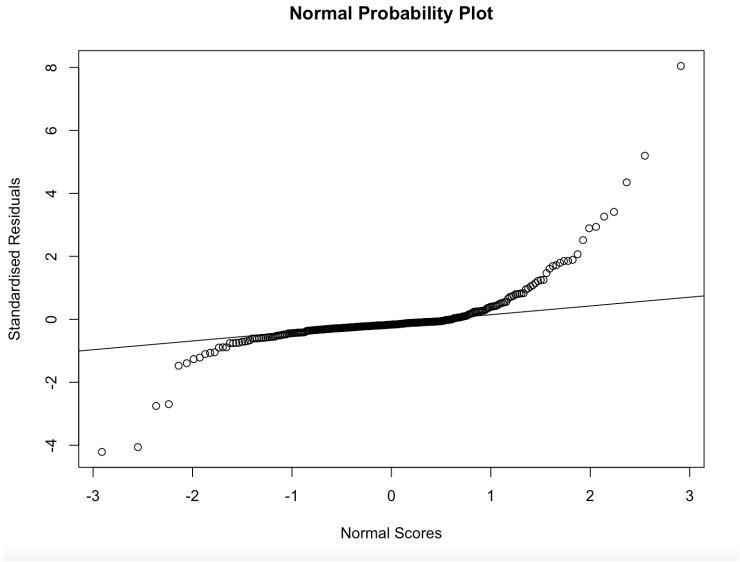
The p-values indicate there is a statistically significant relationship between the predictor and the target variables. At a significance level of $\alpha = 0.05$, we can reject the null hypothesis that $\beta_1 = 0$, since the p-value for rarity is < 2e-16. Similarly, the p-value for the F-test of overall significance is statistically significant.

Upon analyzing the residual plots, however, the observations appear certainly unusual and identify problems with the linear regression. Firstly, the Residuals vs. Predicted and Residuals vs. Fitted graphs indicate the presence of outliers and leverage points.

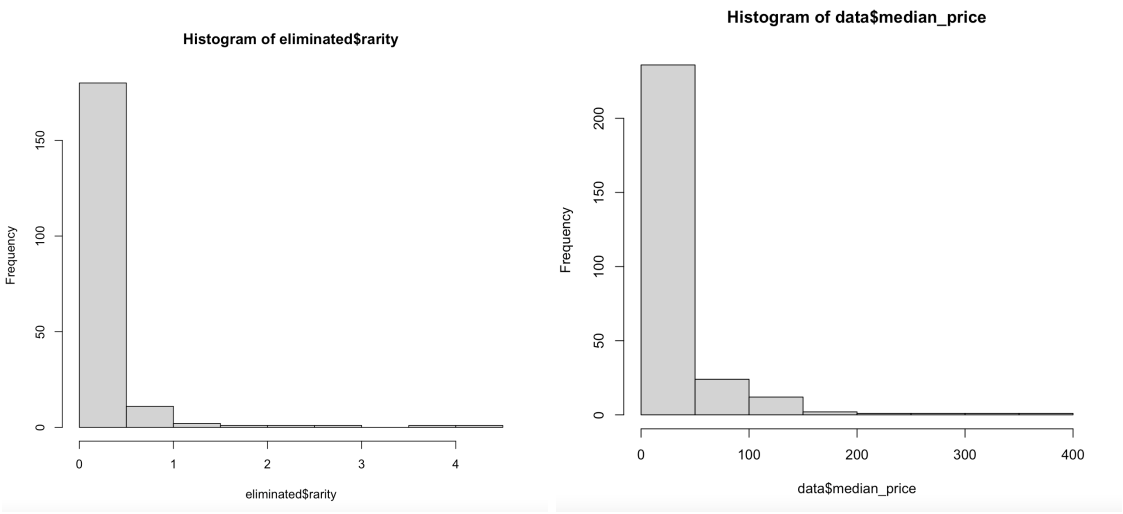





Moreover, the normal probability plot of the residuals indicates that the error terms are not normally distributed. The distribution is heavy-tailed because there are too many extreme positive and negative residuals.




## Removing Outliers and Leverage Points




As observed from the histograms, there seem to be extreme outliers because the histogram for median prices is skewed. Moreover, there seem to be extreme leverage points because the histogram for the rarity values is skewed. To fix the skewed distribution of both variables, we remove the outliers and leverage points from the dataset. Specifically, data points which fall below Q1 - 1.5 * IQR and above Q3 + 1.5 * IQR are removed for both predictor and response variables. Q1 represents the 25th percentile and Q3 represents the 75th percentile.

The removed leverage points are shown in the following table:

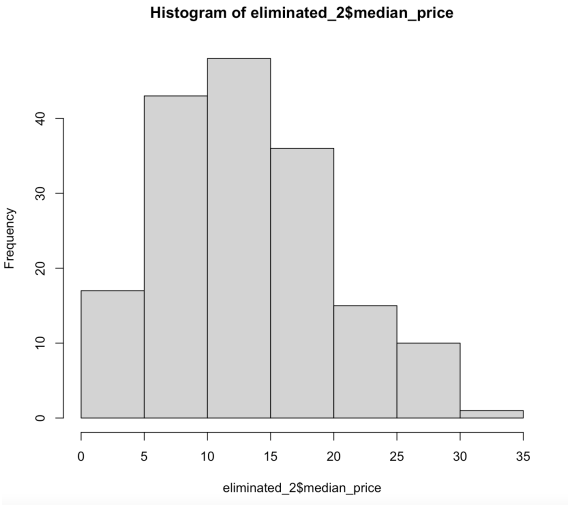
| | years_since_release | total_versions | album_name | rarity | users_have | users_want |
|---|---|---|---|---|---|---|
| **16** | 52 | 6.0 | Led Zeppelin III | 2.155039 | 387.0 | 834.0 |
| **19** | 45 | 2.0 | Animals | 3.844961 | 129.0 | 496.0 |
| **131** | 54 | 3.0 | Beggars Banquet | 2.377193 | 228.0 | 542.0 |
| **217** | 53 | 2.0 | David Bowie | 6.015385 | 130.0 | 782.0 |
| **223** | 31 | 2.0 | Achtung Baby | 3.154122 | 279.0 | 880.0 |
| **224** | 33 | 2.0 | Doolittle | 2.760274 | 146.0 | 403.0 |
| **283** | 35 | 1.0 | Kiss Me Kiss Me Kiss Me | 2.411765 | 85.0 | 205.0 |
| **295** | 51 | 2.0 | The Allman Brothers Band At Fillmore East | 4.120690 | 58.0 | 239.0 |

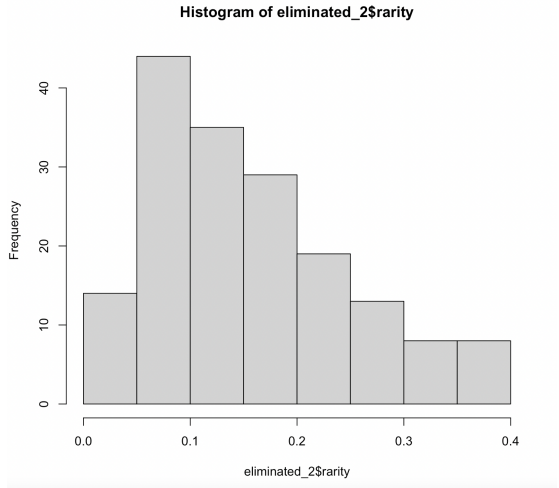
The outliers have high rarity values, indicating there is a higher proportion of users who want the vinyl to the users who have the vinyl. All these vinyls are

particularly rare because there have been fewer repressings or releases of the vinyls, as indicated by the total versions column. Vinyls with rarity values lower than 2, or vinyls which are less rare than vinyls in the preceding table, have an average of 47 different total versions, which is significantly higher than the range of 1 - 6 total versions in the table.

Notably, the vinyl titled “David Bowie " has the highest rarity value, with 782 users who want the album and 130 users who have the album. As reported by Pitchfork, this vinyl is an original repressing of David Bowie’s 1969 self-titled second album. It sold on Discogs for $6,826, which makes it one of the most expensive vinyls sold on the site. The album is particularly valuable and rare because very few of the vinyls were released under the original title, David Bowie. The vinyl was later re-released underneath a different title, “Space Oddity.”

After removing the outliers and leverage points, the same histograms are generated, indicating a more normal distribution for both variables. The dataset was reduced from 278 observations to 170 observations.

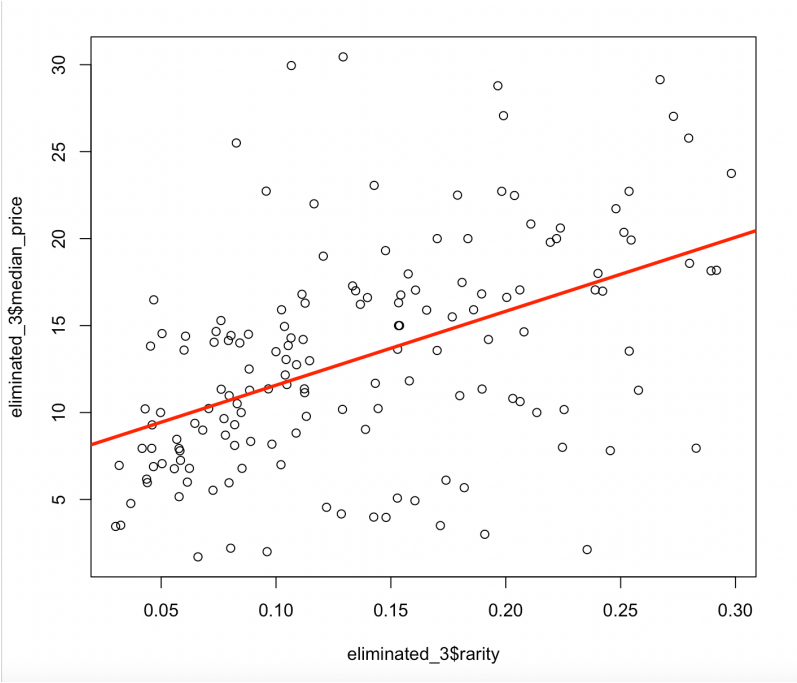



Using the subset of data with leverage points and outliers removed, a second simple linear regression was built. Here is the output:

```
Call:
```

```
lm(formula = eliminated_3$median_price ~ eliminated_3$rarity)

Residuals:
     Min       1Q   Median       3Q      Max
-15.2121  -3.1643  -0.0826   3.4476  18.0949

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)           7.3172     0.9948   7.355  1.1e-11 ***
eliminated_3$rarity  42.5420     6.5167   6.528  9.4e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.631 on 152 degrees of freedom
Multiple R-squared:  0.219,    Adjusted R-squared:  0.2138
F-statistic: 42.62 on 1 and 152 DF,  p-value: 9.395e-10
```




The adjusted R-Squared is lower at 21.38%, which indicates the regression does a poorer job of explainity the variability in median price. This is confirmed by an observation of the scatter plot and the regression line above. There appears to be high variability among the points around the regression line, suggesting a non-linear relationship. The coefficient for the rarity variable is slightly higher at 42.54. This

model predicts a higher increase in median price for a single unit increase in the rarity value.

The regression can be used to build prediction and confidence intervals. Given a vinyl with the rarity value of 0.15, the following 95% prediction and confidence intervals can be obtained:

```
Prediction Interval
fit        lwr      upr
1 13.58546 2.754999 24.41591

Confidence Interval
fit        lwr      upr
1 13.58546 12.69521 14.4757
```

As expected, the prediction interval is wider than the confidence interval. Using the regression, our estimate for median prices of all vinyls with a rarity value of 0.15 is between the range (12.67, 14.48). Our estimate for the median price of a single vinyl with a rarity value of 0.15 falls between the range (2.75, 24.42). In other words, we can be 95% confident that the next vinyl with a rarity value of 0.15 will have a median price between the range (2.75, 24.42).

As observed in the scatterplot, there seems to be stronger correlation in the variables for vinyls with a lower rarity value. Recall that rarity value is calculated by:

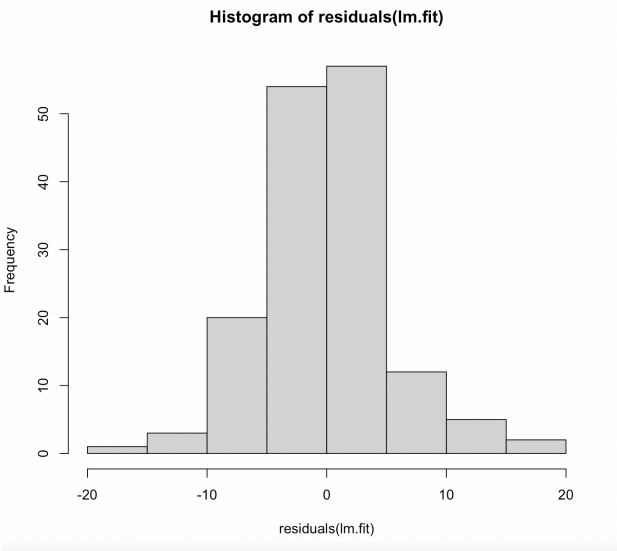
$$\frac{\#\ Users\ who\ want\ the\ record}{\#\ Users\ who\ have\ the\ record}$$

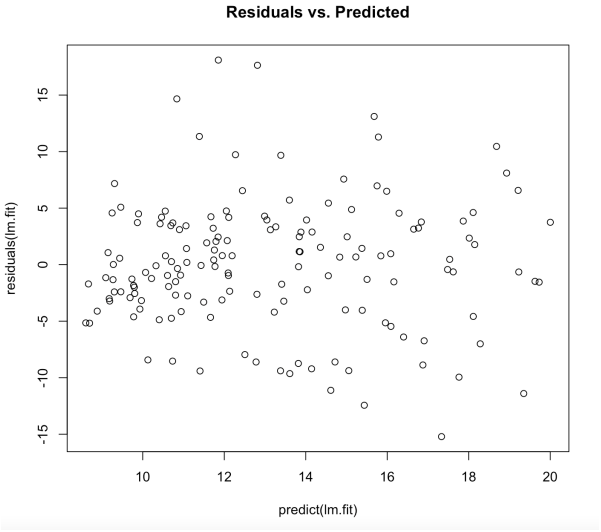
This may indicate that there is a stronger relationship between prices and rarity for vinyls where there is a lower proportion of users who want the vinyl as compared to users who have the vinyl. Let us examine that assumption by building a regression model which omits vinyls with rarity values higher than 0.15:

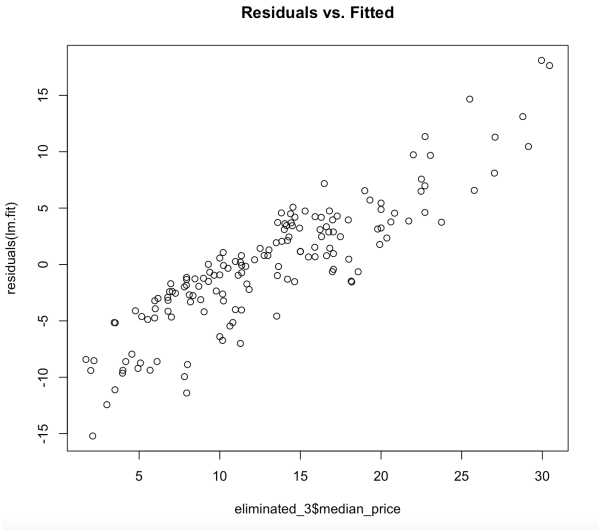
```
Call:
```

```
lm(formula = eliminated_4$median_price ~ eliminated_4$rarity)

Residuals:
     Min       1Q   Median       3Q      Max
-11.4672  -3.0488  -0.4836   2.8409  17.3394

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)           5.296      1.591   3.329 0.001261 **
eliminated_4$rarity   68.570     17.107   4.008 0.000125 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.202 on 91 degrees of freedom
Multiple R-squared:  0.1501,  Adjusted R-squared:  0.1407
F-statistic: 16.07 on 1 and 91 DF,  p-value: 0.000125
```

The adjusted R-Squared is lower at 14.07%. This regression is weaker than the prior model. Moreover, the coefficient for rarity has increased significantly. Therefore, we will disregard this model and examine the histogram of residuals and residual plots for the prior linear regression model:



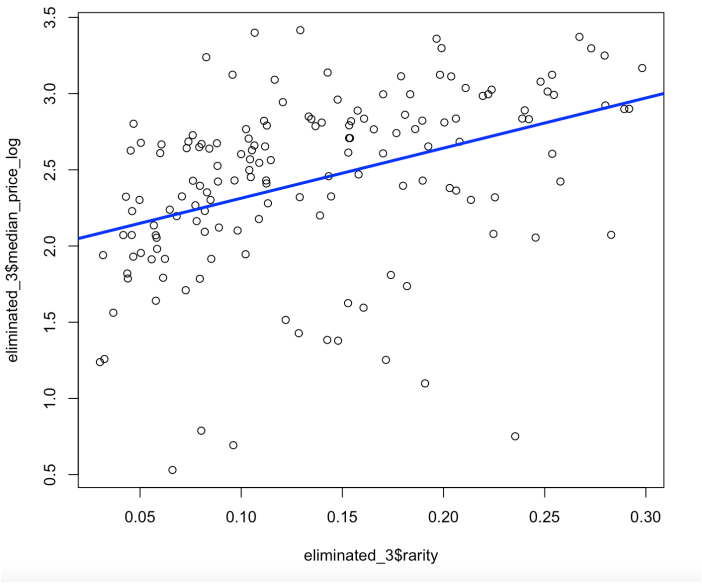


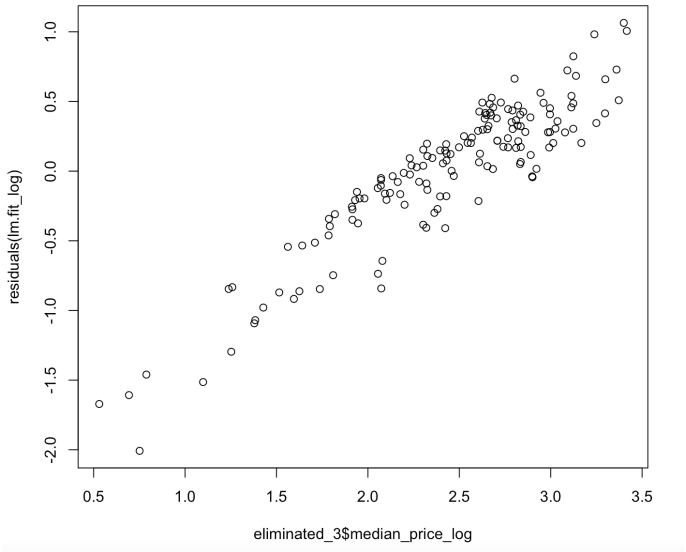




The histogram of the residuals indicates that the error terms are normally distributed. The Residuals vs. Fitted model, however, indicates that the model violates homoscedasticity, and the error terms are not constant along the values of the dependent variable. In particular, the model predictions will be too high for vinyls with a higher median price. Because the model violates homoscedasticity, another regression will be built by transforming the dependent variable and modeling its logarithmic value:

```
Call:
lm(formula = eliminated_3$median_price_log ~ eliminated_3$rarity)

Residuals:
    Min      1Q  Median      3Q     Max
-2.0077 -0.2036  0.1192  0.3501  1.0639

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.98478    0.09375  21.172  < 2e-16 ***
eliminated_3$rarity  3.28916    0.61411   5.356 3.09e-07 ***
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5306 on 152 degrees of freedom
Multiple R-squared:  0.1588,  Adjusted R-squared:  0.1532
F-statistic: 28.69 on 1 and 152 DF,  p-value: 3.089e-07
```



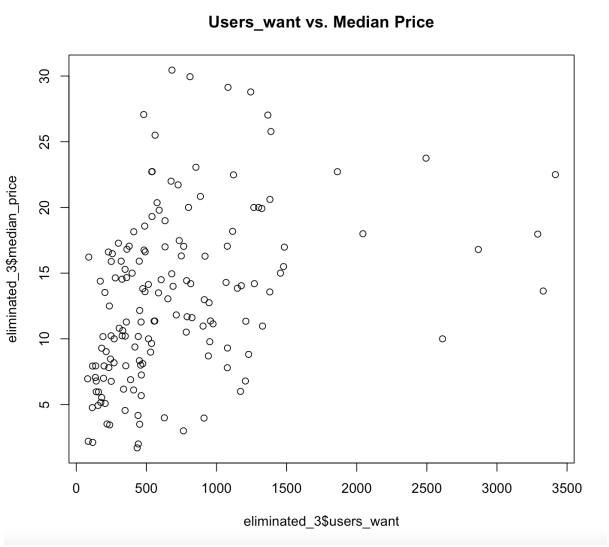



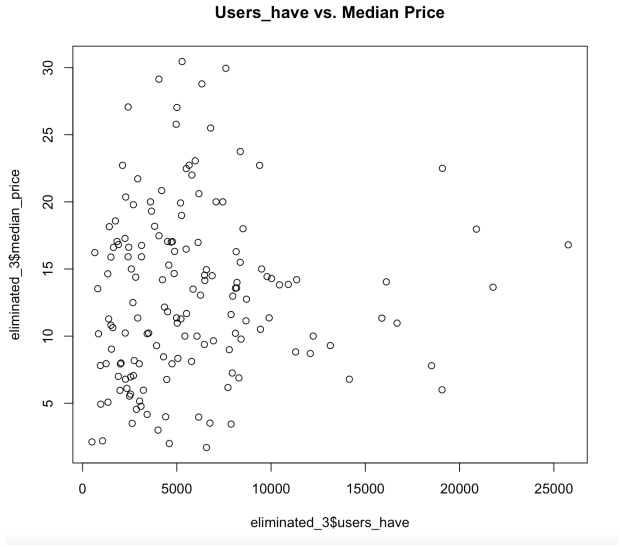

Unfortunately, the model still has not improved with a weaker adjusted r-squared value. Moreover, the model still violates homoscedasticity after observing

the Residuals vs. Fitted plot. To explore why the error in the model increases as the median price increases, we examine the relationship between the determinants of the rarity value and the median price. Recall that rarity value was calculated using users_want and users_have variables:

$$\frac{\#\ Users\ who\ want\ the\ record}{\#\ Users\ who\ have\ the\ record}$$

A scatter plot of those two variables against median price can be observed here:




Note that in both scatterplots, as median price increases there are less observations for both variables, which are determinants of the rarity value. Since there are fewer observations for vinyls with a median price above ~20, the model may be less accurate and overpredict median prices when the true median price is higher than ~20.

In conclusion, there does not appear to be a linear relationship between the predictor, rarity value, and the response variable, median prices. Despite transforming the response variable, and removing the outliers and leverage points,

the linear regression model has a weak r-squared value, indicating that the rarity value does not explain the variation in the median prices of vinyls.