An Analysis of NFT Prices

By: Robert Alward

December 11, 2021

In service of the Final Requirement of Math 340 Taught by Professor Kimberly Sellers

Abstract

Non-Fungible Tokens (NFTs) were invented in 2017 and in the past year there has been an exponential increase in their creation, trading, and most notably sale value. This study attempts to determine the best algorithm to predict the future price of NFTs using data on the NFT creation and trading patterns. The models tested include linear regression, robust regression, ridge regression, weighted least squares regression, and random forest regression. Of the different models the ridge regression had the best performance on test data. The analysis also showed that a model containing interaction terms with the NFTs mean transfer value had the best performance of the linear model alterations tested. This analysis shows similar trends to analysis done on other alternative investments, showing that higher trading volume and higher previous prices leads to higher final prices.

1. Introduction

The introduction of Non-Fungible Tokens, NFTs, in 2017 raised little fanfare, yet over the past year there has been an explosion of interest resulting in both excitement and skepticism. The key innovation surrounding NFTs is the ability for an individual to own a piece of digital material. This is possible due to innovations in the technology of blockchains and smart contracts (Ethereum, 2021). NFTs have recorded incredibly high prices with the top sale reaching a value of over $69 million and thousands of trades per month there is undoubtedly a global interest around the technology and potential value creation that NFTs can bring (Cascone, 2021).

The recent activity in regard to NFTs has been primarily performed on the Ethereum blockchain and all of the data for the study is collected from Ethereum based NFT activity. There is some concern that analyzing only a single blockchain could bias the results of this analysis as fees around the purchase and sale of NFT fluctuate at relatively high values, with their recent average being around $40 per NFT. This fluctuation could create an inherent bias within the data as the value of the NFTs could be higher for Ethereum than other blockchains, such as Solana. As Ethereum is the primary blockchain on which the majority of NFT activity happened this year, this is not a major concern for this study but could be investigated in the future as more blockchains compete for usage.

As NFTs have gained popularity, their prices have also increased. In general, NFT prices are denominated in ETH, the cryptocurrency of the Ethereum network. The valuation of NFTs exclusively using ETH provides some concern for analysis of prices as Ethereum's price fluctuates on top of the price changes of the actual underlying NFT. For this analysis I have chosen to look solely at the price of NFTs in ETH or its derivatives such as microeth (one-millionth of ETH), as predicting the underlying price fluctuations of ETH is an unsolved problem. The price of NFTs are influenced by various factors including their exclusivity, visual characteristics, length of existence, and other traits such as network effects or utility. This study looks to analyze the impacts of the characteristics captured by the records of NFTs trading and minting patterns and predict the current price of the NFTs. The two specific research questions addressed in this paper are:

1. What model best predicts current NFT prices with prediction performance assessed based on the Root Mean Squared Error (RMSE) of the model?
2. What are the statistically significant variables in the prediction of NFTs prices?

2. Literature Review

A recent paper by Matthieu Nadini et al. on the NFT boom explores various facets of NFT trading, ownership, and usage. The authors analyze the market trends of NFTs by looking at the value of being in certain collections, the changes in the market over time, and which categories of products are the most traded. The authors split NFTs by type into art, collectible, games, metaverse, other, and utility. This classification helps understand the broader trends and movements in the crypto market. Through plotting the market transactions, the paper shows the dominance of art transactions. Their investigation of trader networks identified some market participants as high-frequency traders, as a large percentage of market transactions pass through their accounts. These traders seem to represent sophisticated market participants and could even represent organizations or large groups of individuals operating trades out of a single wallet address. In the final stage of their analysis, the group used a linear regression model to predict the current price of NFTs using four different variables. These variables included the centrality of the buyer and seller of the NFT within all the NFT networks, the components of the visual features of NFTs, the prior probability of a sale of the NFT, and the past median price of sales of

NFTs within a specific collection. The past median price consistently appears as the most influential variable with an adjusted R-squared of over 50%. In the model of all of the NFTs, all of the above variables were statistically significant at the .01 level. This provides valuable insight into the value of prior prices in predicting the current prices in the market. Additionally, the statistical significance of buyer and seller centrality points to potential future study regarding individual collectors and the impact that they could have on the price of individual NFTs. An additional feature of this paper was the identification of the skew present throughout many of the characteristics of NFT data sales. The majority of NFTs are traded relatively infrequently, the majority of buyers only buy a few NFTs, and the majority of prices are low values (Nadini et al., 2021).

With so few papers on NFTs, it is valuable to look at similar prediction tasks of other alternative assets. An analysis of the alternative asset of rare coins looked into the ability for generalized regression neural networks to predict the price of the Liberty Half Dollar. They identified the variables with the greatest influence as those regarding the state of the conservation of the coin, the age of the coin, and the exclusivity of the coin. The variables of age and exclusivity variables are directly represented in the NFT dataset providing valuable insight into potential significant features. The age variable in regards to NFTs represents a significantly shorter time period of 0-4 years, yet within the time accelerated world of the internet this variable could still have a significant impact on the value of NFTs. Additionally, the exclusivity of NFTs could be represented by the number of copies of a given NFT. With more copies of a given NFT this could result in less exclusivity of a given NFT. This paper provides a valuable basis for variable choice and an interesting point of comparison of the prediction tasks of the value of physical assets versus the value of digital assets (AlAcázar-Blanco et al., 2021).

Another paper on the art market reveals many potential similarities to the NFT market. Both the NFT and art markets command high prices for their most exclusive assets, yet the physical high art market has fewer potential buyers for its work due to higher prices, access difficulties, and limited operating times. The analysis identifies visual features of a given art piece that are not predictive of the actual price of the art. The paper also calls out high correlation between a given works price and its trading volume with high prestige works trading 4.7 times more often than low prestige pieces and claiming values on average 5.2 times above those of low prestige pieces. With the variable number of trades and average trading value it is interesting to see if this relationship holds in the digital art space (Bailey, 2020).

2. Data

One of the key offerings of the blockchain is that all transactions are recorded on a public ledger and can be queried by anyone who wants to investigate activity on any given blockchain. The data for this project comes from this querying of the blockchain done by a project called Moonstream. Moonstream runs analytics for NFTs and other crypto projects. In an effort to raise public interest in the NFT space they released a dataset of all NFT transactions from April 1st, 2021 to September 25th, 2021. In addition to transaction data, Moonstream released a dataset containing all of the minting information[1] of the NFTs. They connected all of the NFT's in their dataset with their current market value in WEI as of September 25th 2021. As the NFT market moves quickly these recorded prices are not representative of the prices as of December 2021, yet this analysis assumes that there has not been a significant change in the dynamics around trading, minting, and pricing in the past months. The data from Moonstream was collected for all copies of any given NFT resulting in multiple copies of the same NFT appearing in the minting and trading datasets provided. Due to this issue the data was summarized by an NFT address

[1] Minting refers to the act of creating and NFT

representing a unique digital item. There is some variability between the prices of copies of NFTs due to changes in popularity and market fluctuations but for the purpose of addressing the question of predicting the future price of unique NFTs the average price of an NFt address was used as the proxy for NFT prices (Moonstream, n.d.).

Due to this quality of the data all features used to predict current NFT price were aggregated across all copies of the same unique NFT address. Another concern regarding this data was that the size of the original dataset was prohibitive to work with. Containing 7.02 million entries of different NFT transactions and over 4 million different NFT mintings the dataset was around 7 gigabytes in size (Moonstream, 2021). To address this a large random sample of over one million different transactions was taken from the original dataset for this analysis. After aggregation the final dataset contained 75 different unique NFTs and data regarding their mining and transfer. Four of these 75 NFTs had never been transferred resulting in missing data. As there was no clear underlying trend to these missing values and they made up less than 10% of the data they were discarded. The final dataset used for analysis consisted of nine unique variables and 71 observations of NFTs. The response variable for this dataset was the average price of the unique NFT on September 25th 2021. The explanatory variables were number of copies of the NFT, number of unique holders of the NFT, average price at minting, maximum price at minting, minimum price at minting, number of transactions of the NFT, average price during all transactions of the NFT, difference in time between first minting and most recent transfer. These were all derived from three datasets, a NFT Transfer dataset, a NFT Minting dataset, and a NFT current price dataset. The final piece of complexity in the data was the denomination of the price of NFTs throughout the dataset. All prices in the data were represented in WEI. WEI is the smallest possible denomination of Ether (ETH), the cryptocurrency associated with the blockchain Ethereum. There are 10^{18} WEI in one ETH and the price of ETH is constantly fluctuating with regard to the US dollar with the price of ETH at September 25th being $3,154.56 to one ETH (Yahoo Finance, n.d.). Additionally due to complications arising from the large integer size of the WEI values, all monetary values were reduced by 10^{12} resulting in final values in the analysis representing 10^{6} to one ETH also known as one microeth. All of the analysis for this project was done using R and R studio with key packages and libraries listed in the appendix (Appendix 10: R Packages).

2. Methodology

A. Exploratory Data Analysis

To begin modeling the data the first step was to perform exploratory data analysis. This analysis includes histograms of the individual variables, correlation matrix, summary statistics, boxplots of individual variables which were used to identify the qualities of the underlying data and the appropriate models to fit to the data. The histograms of the variables showed a consistent trend of right skew in the data. This trend is shown in Appendix 2, with the distribution of current mean NFT prices clustered near zero with a few outliers at much higher values than the mean. This trend can be seen in all variables which include value or count data regarding the NFTs. Specifically, the right skew can be seen in the   number of copies of the NFT, number of unique holders of the NFT, average price at minting, maximum price at minting, minimum price at minting, number of transactions of the NFT, and mean transaction price of the NFT. The one variable that did not follow this trend was the difference between minting price and most recent transfer price. The lack of skew in this variable indicates that there was not a time either recently or at the start of NFT creation that has seen a dominant percentage of NFTs created. The histograms of the data led to concern regarding outliers in the data, thus boxplots for each of the variables were created to identify outliers and the studentized residuals of the response variables were analyzed to check for outliers. In the boxplots shown in Appendix 3, the skewed variables also showed a large number of outliers. The highest number of outliers determined through the IQR method were in the mean transfer value dataset with 10 outliers (Appendix 4). The lowest number of outliers is seen in the mean minting values and the current mean value with both variables only having two outliers. Using studentized residuals, three observations were identified to be outliers (Appendix 15). The high number of outliers throughout the dataset could be a problem for some of the assumptions of a linear regression model, including heteroskedasticity. Overall there are general concerns regarding the normality of this data and the influence of outliers.

The concern regarding this data was the potential for multicollinearity. As the study on predicting physical art prices identified, there is often a correlation between the number of trades of a work and its price. In the NFT market it is also common to see a strong correlation between the price of an NFT at minting and its mean price throughout transfer as the minting price serves as an indicator of the demand of the NFT. To investigate the correlation, a correlation matrix was constructed with all of the explanatory and response variables (Appendix 1). There were four pairs of variables with correlation coefficients over 0.7, confirming the concerns regarding multicollinearity. Some notable correlations were the number of unique owners and the count of the number of NFTs with an extremely high correlation coefficient of 0.919. This makes sense as the more NFTs available the more potential owners there could be. It also reveals that individuals generally do not hold a large number of a single NFT; otherwise, unique owners and number of NFTs would not be as strongly correlated. The count of transfers and count of unique transfers (transfers of different NFTs) were also highly correlated with the count of NFTs with the correlation coefficients being .724 and .7855 respectively. This reveals that more copies of NFTs is correlated with more transfers of those NFTs. This could point to an average range of time that most NFTs are held for. If NFTs were held for drastically different amounts of time these variables could be uncorrelated as certain collections could be rarely traded with other collections being traded often. The variance inflation factor from a linear model of all the explanatory predictors confirms this with a mean VIF of 18.305 for over the desired mean of 1 and a max VIF of 53.038, far over the desired max VIF of 10 further supporting concerns of multicollinearity (Appendix 6).

B. Modeling

Five models were used to predict the current value of NFTs: a linear regression, a robust linear regression, a ridge regression, a weighted least squares regression, and a random forest regression. Each of these regressions used all of the nine predictors to model the current mean NFT price.

The first model attempted was a multiple linear regression shown in Appendix 13. The response variable of mean NFT value was modeled using all of the variables outlined in the equation. To assess the appropriateness of the model, the assumptions for the model were checked. In regards to the linear nature of the variables there were some concerns from the initial histograms. The values generally seemed to follow a more logarithmic pattern in regards to the prices and number of trades. This indicated that there could be some need for variable transformation. Next, looking at the independence of the variables there was also some concern. The multicollinearity of the plots indicated that some explanatory variables were highly correlated and the creation of the variables could lead to some of the explanatory variables not being independent. The variable’s normality was also questioned as many were skewed, not following a centered normal distribution. To investigate concerns of normality further, I conducted a Shapiro Wilks Test. The test resulted in a p-value of $5.219*10^{-6}$ thus we reject the null hypothesis that the data was normal at a 5% significance level (Appendix 11). Finally, to test the homoscedasticity of the variables I used a scale location plot for the residuals of the linear model. The plot showed a clear upward trend indicating increasing variance thus violating the necessary assumption of homoscedasticity (Appendix 5). As the assumptions for a linear model were not met, this indicated the need for other models which could address the issues in this data, the issues with the linear model, and potentially lead to better performance.

To address the issue of heteroscedasticity Weighted Least Squares regression was performed. To address the multicollinearity concerns a Ridge Regression was used. To address the outliers and influential points Robust Regression was used, and to address the potential non-linear relationships in the data a Random Forest Regression was used. These models could not all be compared using traditional measures of performance such as adjusted r-squared, AIC, or BIC, thus I used the root mean squared error (RMSE) of the model's predictions on 25 previously unseen NFTs and their related variables to measure comparable performance (Appendix 12). The models were all trained on 50 NFTs and their related variables. In addition to RMSE where applicable adjusted R-squared, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Mean Squared Error, and F-statistics of the various models were used to add additional insight into model performance.

The Weighted Least Squares Regression model (WLS) was used to address the issue of non-equal variances identified in the residual analysis from the full regression model. This can be seen in the residual plot for the linear regression, the scale location plot and the plot of squared variance against mean NFT values shown in Appendix 5 and Appendix 8. WLS provides advantages over linear regression as it provides weights to emphasize which values in the dataset should contribute more or less to the coefficients. This allows the model to discount the accuracy of values with high variance and emphasize the more accurate values that contain lower variance. While this method helps to create a more accurate equation to predict the mean value of an NFT given the explanatory variables, the theoretical backing of the equation assumes that the appropriate weights of all of the variables are perfectly known. In practice, this is not true and can lead to slightly worse performance of the model. Additionally WLS is sensitive to outliers which are common in this dataset. With these two concerns noted the WLS was used and formulas regarding the weights and regression equation can be seen in Appendix 14. The concerns of the WLS model led to the consideration of a model that specifically addressed the influence of outliers, namely Robust regression (NIST, n.d.).

Upon observing high counts of outliers in the exploratory data analysis as well as multiple values with influential Cook's distance, nine of the forty-six observations in the training dataset had influential cooks distance values as seen in Appendix 15, Robust regression was identified as a model that could more accurately model the data due to the influence of outliers. As the outliers and influential values were a large percentage of the data, simply removing the values would likely lead to a bias in the model, yet their presence reduced the model accuracy. To reduce the influence of these outliers M-estimation with Huber bisquare weighting was used along with the rlm function in the MASS package. The residual plots and table of influential points can be found in Appendix 15 showing the need for robust regression to reduce the influence of outlier values (Fox & Weisberg, 2013).

In addition to issues regarding outliers and variance, the data also showed issues regarding multicollinearity leading to the choice to use Ridge Regression to model the data.

Multicollinearity can occur when there is a strong linear relationship between the explanatory variables of the dataset, such as in this case with four cases of correlation of over .7. The multicollinearity can also be seen in the high Variance Inflation Factors in Appendix 6 which have mean and max VIF of over 10 indicating high multicollinearity. This results in less accurate predictions of the variances of the predictors resulting in a less accurate model. Ridge regression attempts to correct for that. By reducing the weight of the predictors through a penalty term which is changed via a tuning parameter λ, ridge regression is able to more accurately model data with multicollinearity. A lambda of 0 represents the same model as the multiple linear regression whereas a lambda of infinity would move all of the coefficients to 0. Using ridge trace plots shown in Appendix 7 and the R function "select" in the MASS package, the optimal lambda value was determined (NCSS, n.d.).

Finally in another attempt to address the violation of assumption of multicollinearity, normality concerns, and the outliers present in the data a Random Forest regression model was created to predict mean value of NFTs. The random forest model is not based on a linear model but is an ensemble of decision trees (Appendix 16). After training the model on a given set of training data, predictions are then made by taking the average of the two end nodes that the unseen data lands between. The random forest model provides advantages regarding its ability to learn non-linear relationships and its improvement with more data. Unfortunately it is more difficult to address its performance as it does not fit a line of best fit through the data. This resulted in the choice of RMSE as a universal model assessment and comparison metric (Bakshi, 2020).

In addition to using models to fix the assumptions that weren't met in the linear model, changes to the linear model were also made to more closely fit linear assumptions and improve model performance. Interaction terms between mean transfer value and mean minting value were tested due to the previous literature regarding the high impact of mean values on final price prediction. The interaction terms were tested for statistical significance and the overall performance of the model was compared to the basic linear model performance. The performance of a smaller model using only the statistically significant predictors was also tested. This was to see if model assumptions could be met and to see which evaluation metrics improved due to the reduction in size of the model. Also, based on the literature and observations of significance a model with only transfer value as the sole explanatory variable was tested. Finally, to address concerns of normality and skewness, a log transformation of the data was applied to each of the skewed variables to reduce skew and create a more approximate normal version of the data and the performance of this model was compared to the original linear model (Appendix 9).

5. Results

The models had varying performance on the test data with the best performing model based on RMSE being the Ridge Regression model.he best iteration of the linear model was the linear model with mean transfer value interaction terms. The Ridge Regression model had a RMSE of 92850.74. The next best performing model was the Robust Regression with an RMSE of 93429.54. The worst performing model was the Random Forest model with an RMSE of 147013.6. The Linear Regression and Weighted Least Squares had RMSEs of 94900.88 and 113501.8 respectively. In addition to the RMSE, the models which could be assessed with adjusted r-squared, AIC, and BIC showed slightly different trends. The adjusted r-squared of the linear model was .9321 while the adjusted R-squared of the WLS was .9907 indicating that the performance of the WLS model was better than that of the linear model on the training data. The

WLS model also had the lowest AIC of 1214.66 and the lowest BIC of 1234.775 (Appendix 17) . The high adjusted r-squared and low AIC of the WLS model on the training data and lower performance on the test data could represent overfitting in the WLS case and potentially imperfect estimates regarding the weights of the variables.

Over all of the models the most consistently statistically significant variables were difference between minting time and most recents transfer times (diff), mean transfer values (mean_trans_value), and mean minting value (mean_mint_value) (Appendix 18). This list confirms the previous NFT analysis in which the mean value of the NFTs in a group were the most significant predictors of future NFT value. In the weighted least squares model, the count of transfers, count of unique owners, and count of NFTs were all also statistically significant on top of the three variables that were significant across all the models. For ridge regression the optimal λ value was .114 and the coefficients generally did not show a large shift from the coefficients in the linear model. For robust regression the count of transfers and count of NFTs were also statistically significant at the 5% significance level. The coefficients for the statistically significant variables for each of the models are shown in Appendix 18.

Looking into the different iterations of the linear model we can see the combination of explanatory variables which optimizes for adjusted r-squared, the high predictive power of the transfer value model, the statistical significance of interaction terms with mean transfer value, and the results of the variable transformation. The linear model with the fewest predictors and the highest adjusted r-squared was the model containing the three statistically significant variables identified in the full model: difference, mean transfer value, and mean minting transfer value. This model had an adjusted r-squared of .9321 (Appendix 17). The highest adjusted r-squared of any combination of variables was the model consisting of all explanatory variables except maximum minting value. This model had an adjusted r-squared of .9374 which was still below the WLS model. The model which only contained transfer value as the single explanatory predictor had a high adjusted r-squared of .6959 indicating that 69.59% of the variance of current mean NFT prices is explained by the linear model of mean transfer value of NFTs (Appendix 17). Excluding mean transfer value from the predictors resulted in an adjusted r-squared of only .0095. This disparity in model performance between these two models raises concerns as to the impact of transfer value in the prediction tasks. In investigating the interaction terms the explanatory predictor variables of difference, count of transfers, count of unique transfers, count of unique owners, and count of nfts were all statistically significant in their interaction with mean transfer value, yet none were significant in the interaction with mean minting value. The adjusted r-squared of the model which included mean transfer value interaction terms was 0.9826, the second highest after the WLS model. Finally the variable transformation using the natural log did not improve the model performance. While the transformation did reduce the skewness of the data from 3.107 to -1.546 the model with the adjusted variables had an adjusted r-squared of only 64.54%, significantly below the normal linear model. This could be due to the fact that taking the log of zeros created a model whose data was still not normally distributed. Both the various types of models attempted and their performance and the different linear models provided valuable insight into the characteristics of NFT sales including the high impact of outliers, and the value of mean transfer value in the prediction task.

6. Conclusion

This analysis showed a strong ability to predict NFT prices from current data about NFTs. While the high R-squared of the model indicates some level of usefulness, it is important to note that over the time which the data was collected NFTs have been at all time high levels of popularity leading to difficulty in expanding these conclusions beyond the specific time period. The prices of NFTs could be modeled well with linear models due to the fact that the market as a whole has been on an incredible growth spurt. The high performance of the Robust Regression model was indicative of the underlying linear relationship in the data and the influence of outliers on the data. Finally the poor performance of the Random Forest was interesting but not unexpected. Due to the final small size of the data set of 50 training values the Random Forest model was likely not able to create a general enough representation of the data and therefore was overfit to the training data. With more data it is expected that the Random Forest model will improve significantly and potentially surpass the performance of the linear models. This analysis also confirms the literature that as a piece or in this case NFT as more owners this leads to reduced prices. Another key finding of this study, that higher mean trading prices lead to higher overall prices aligns with other research on NFTs. Looking at transfer value specifically, the coefficient for the variable of mean transfer value in the full linear model was .8257 indicating that with every additional increase of 1 microeth in the mean transfer value, the current value of the NFT grows by .8257 microeth. In contrast to the analysis of art, holding all other variables constant, in the full linear model more transfers of NFTs leads to a reduction in the value of the NFTs as with every 1 additional transfer the current price decreases by 8.38 microeth. Overall the impact of transfer value, the high performance of Robust and Ridge Regression, and the poor performance of Random Forest reveal more about the underlying patterns of NFT transactions

Further research into this topic will undoubtedly occur and there are many questions that would be interesting to investigate. Further analysis of the impact of those who trade large volumes of NFTs could provide interesting insights into industry leaders or sophisticated market players. Additionally a further investigation of the impact visual characteristics could lead to interesting prediction and classification ability. Overall the future tracking of NFT prices and an analysis of their inevitable decline would prove interesting to see if the mean transfer value still produced such a significant variable.

References

AlAcázar-Blanco, A. C., Paule-Vianez, J., Prado-Román, M., & Coca-Pérez, J. (2021, September). Generalized regression neuronal networks to predict the value of numismatic assets. Evidence for the walking liberty half dollar. *European Research on Management and Business Economics*, *27*(3). 10.1016

Bailey, J. (2020, April 30). *Can Machine Learning Predict the Price of Art at Auction? · Issue 2.2, Spring 2020*. Harvard Data Science Review. Retrieved December 9, 2021, from https://hdsr.mitpress.mit.edu/pub/1vdc2z91/release/2

Bakshi, C. (2020, June 8). *Random Forest Regression*. gitconnected. https://levelup.gitconnected.com/random-forest-regression-209c0f354c84

Ethereum. (2021, October 30). *Introduction to smart contracts | ethereum.org*. Ethereum.org. Retrieved December 9, 2021, from https://ethereum.org/en/developers/docs/smart-contracts/

Fox, J., & Weisberg, S. (2013, October 8). *Robust Regression*. University of Minnesota.

Logan, M. (n.d.). *Wei Definition - Cryptocurrency*. Investopedia. Retrieved December 9, 2021, from https://www.investopedia.com/terms/w/wei.asp

Moonstream. (2021, October 12). *Ethereum NFTs*. Kaggle. Retrieved December 9, 2021, from https://www.kaggle.com/simiotic/ethereum-nfts

Nadini, M., Alessandretti, L., Di Giacinto, F. et al. Mapping the NFT revolution: market trends, trade networks, and visual features. Sci Rep 11, 20902 (2021). https://doi.org/10.1038/s41598-021-00053-8https://www.investopedia.com/terms/w/wei.asp

NCSS. (n.d.). *Ridge Regression*. NCSS. https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Ridge_Regression.pdf

NIST. (n.d.). *Weighted Least Squares Regression*. Engineering Statistics Handbook. Retrieved December 9, 2021, from https://www.itl.nist.gov/div898/handbook/pmd/section1/pmd143.htm

Yahoo Finance. (n.d.). *Ethereum USD (ETH-USD) Price History & Historical Data*. Yahoo Finance. Retrieved December 9, 2021, from https://finance.yahoo.com/quote/ETH-USD/history/

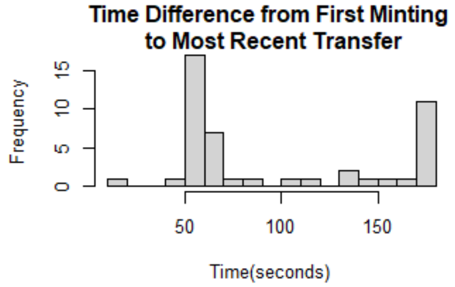
Appendix

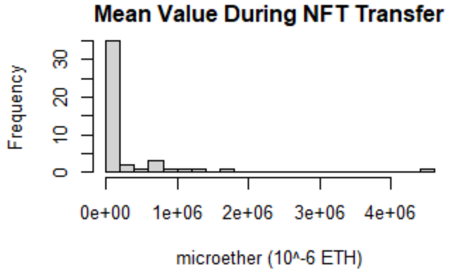
Appendix 1: Correlation Matrix

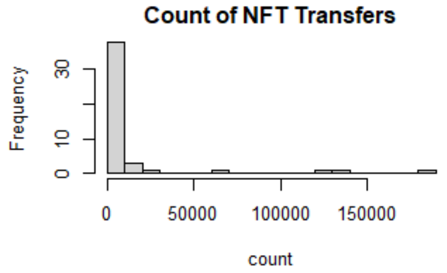
	diff	mean trans. val	count trans.	count unique trans.	count unique owners	count nfts	mint min val	mint max val	mint mean val	current mean val
diff	1.00	-0.06	0.47	0.36	0.44	0.51	-0.01	0.30	-0.12	-0.06
mean trans val	-0.06	1.00	0.00	-0.02	-0.07	-0.06	-0.11	-0.02	-0.07	0.84
count transfers	0.47	0.00	1.00	0.88	0.67	0.79	-0.11	0.43	-0.05	-0.03
count unique trans	0.36	-0.02	0.88	1.00	0.79	0.72	-0.10	0.41	-0.07	-0.05
count unique owners	0.44	-0.07	0.67	0.79	1.00	0.92	-0.12	0.52	0.19	0.07

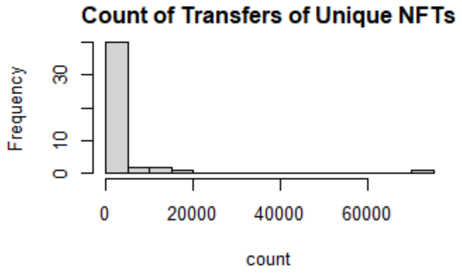
count nfts	0.51	-0.06	0.79	0.72	0.92	1.00	-0.15	0.61	0.29	0.12
mint min val	-0.01	-0.11	-0.11	-0.10	-0.12	-0.15	1.00	0.04	0.10	-0.02
mint max val	0.30	-0.02	0.43	0.41	0.52	0.61	0.04	1.00	0.59	0.28
mint mean val	-0.12	-0.07	-0.05	-0.07	0.19	0.29	0.10	0.59	1.00	0.42
current mean val	-0.06	0.84	-0.03	-0.05	0.07	0.12	-0.02	0.28	0.42	1.00

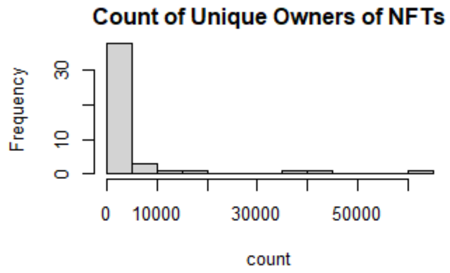
Appendix 2: Histograms of Variables

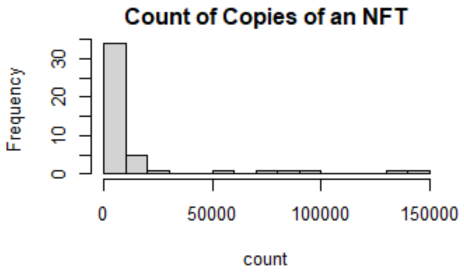

Time Difference from First Minting
to Most Recent Transfer
Frequency
0
5
10
15
50
100
150
Time(seconds)

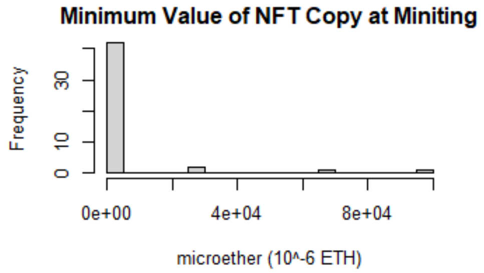

Mean Value During NFT Transfer
Frequency
0
10
20
30
0e+00
1e+06
2e+06
3e+06
4e+06
microether (10^-6 ETH)

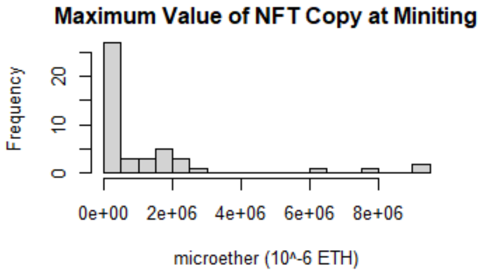

Count of NFT Transfers
Frequency
0
10
30
0
50000
100000
150000
count

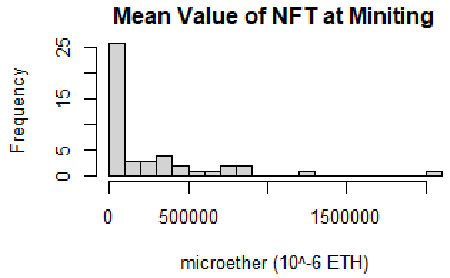

Count of Transfers of Unique NFTs
Frequency
0
10
30
0
20000
40000
60000
count

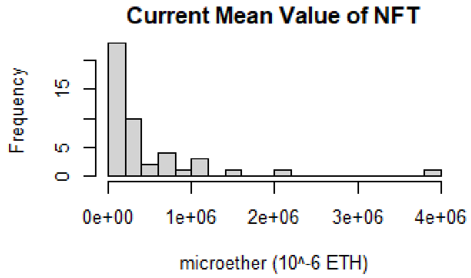

Count of Unique Owners of NFTs
Frequency
0
10
30
0
10000
30000
50000
count


Count of Copies of an NFT
Frequency
0
10
20
30
0
50000
100000
150000
count

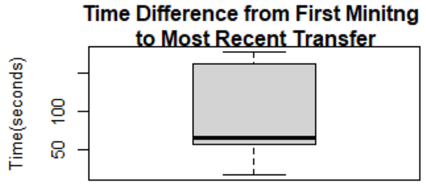

Minimum Value of NFT Copy at Miniting
Frequency
0
10
30
0e+00
4e+04
8e+04
microether (10^-6 ETH)

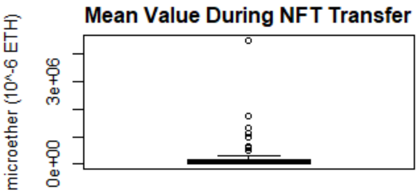

Maximum Value of NFT Copy at Miniting
Frequency
0
10
20
0e+00
2e+06
4e+06
6e+06
8e+06
microether (10^-6 ETH)

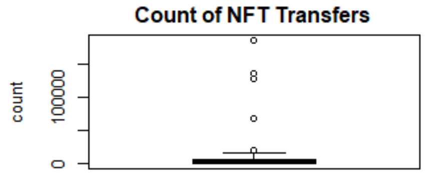

Mean Value of NFT at Miniting
Frequency
0
5
15
25
0
500000
1500000
microether (10^-6 ETH)

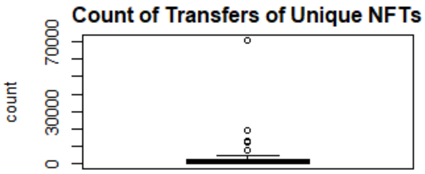

Current Mean Value of NFT
Frequency
0
5
15
0e+00
1e+06
2e+06
3e+06
4e+06
microether (10^-6 ETH)

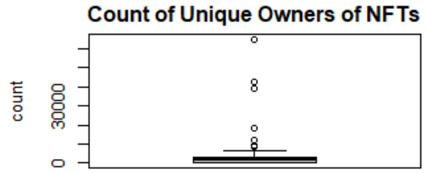
Appendix 3: Boxplots of Variables

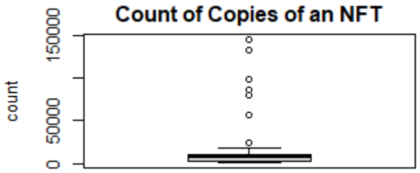

Time Difference from First Minitng to Most Recent Transfer

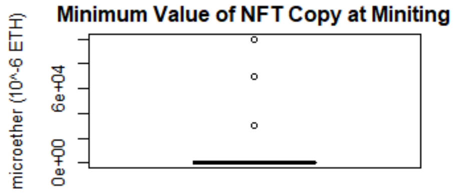

Mean Value During NFT Transfer

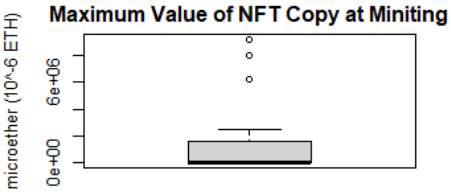

Count of NFT Transfers

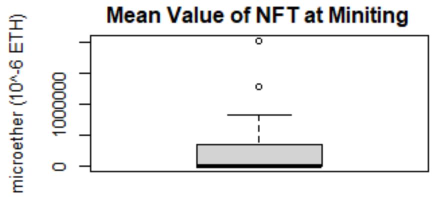

Count of Transfers of Unique NFTs

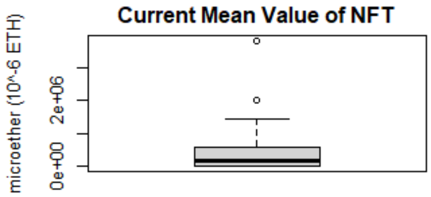

Count of Unique Owners of NFTs


Count of Copies of an NFT


Minimum Value of NFT Copy at Miniting


Maximum Value of NFT Copy at Miniting

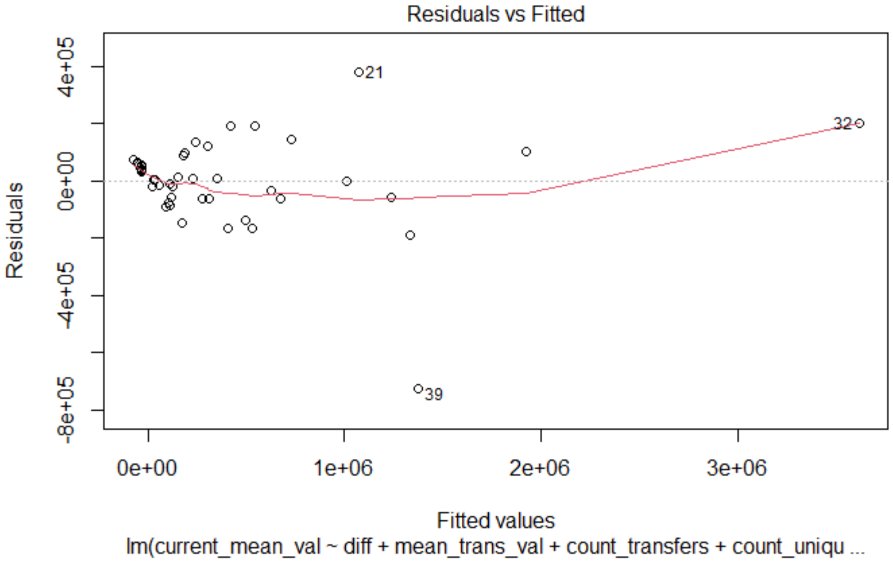

Mean Value of NFT at Miniting

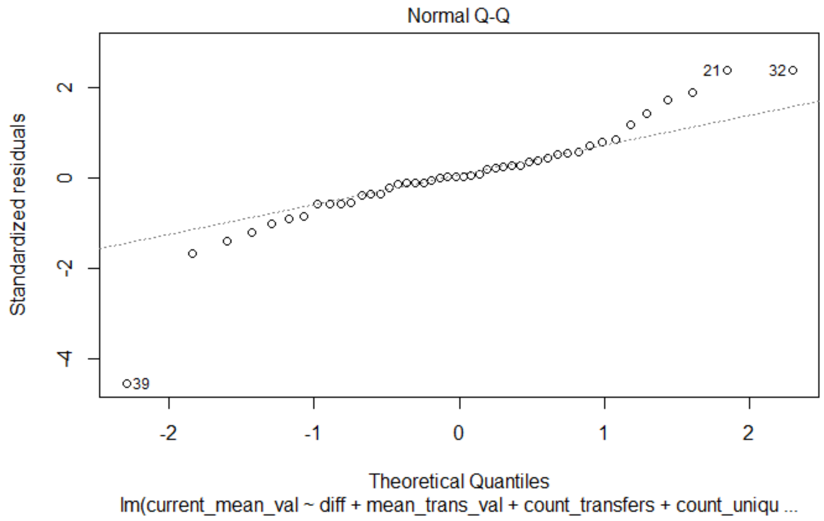

Current Mean Value of NFT

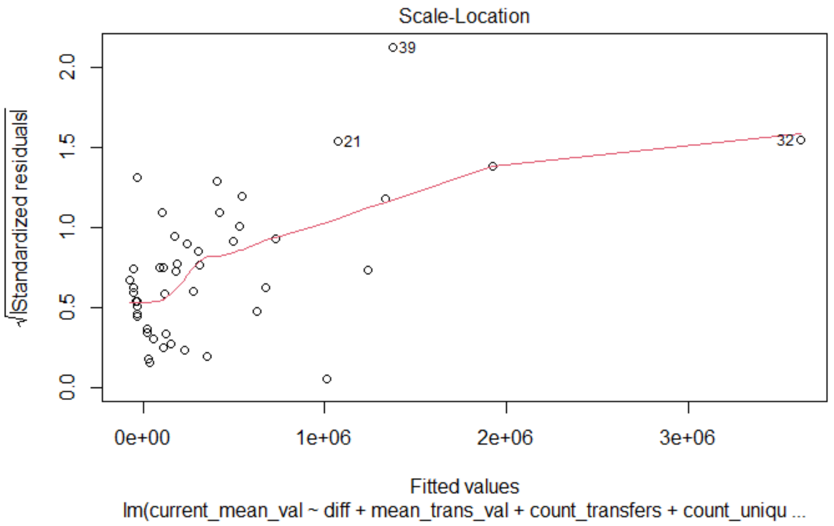
Appendix 4: IQR Outliers of Mean Transfer Value

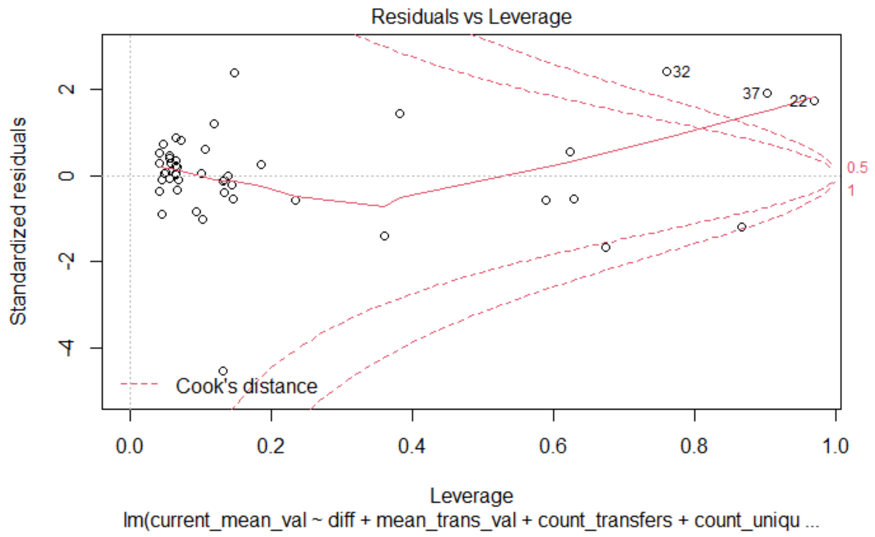
Row #	difference in time	mean trans val	count transfers	count unique trans	count unique owners	count nfts	mint min val	mint max val	mint mean val	current mean val
7	4808460	645201.2	8537	3870	3739	10420	0	2500000	1281046	1140663
21	13983335	1324132	6354	2122	1198	3676	0	0	0	1451704
30	15351288	1102107	135015	19605	18429	98652	0	9223372	537488.6	1181890
32	4939007	4474951	2146	1402	1433	1872	0	0	0	3818318
33	4812422	977221.7	606	432	447	555	0	0	0	874897.9
34	15350949	516191.6	12299	2545	2669	10070	0	0	0	610703
38	13524079	602516.3	17172	5062	4562	17784	0	475000	26.64049	369885.7
39	4889970	1729233	2719	1666	2535	5902	0	0	0	647089
43	5166100	628690.7	11446	4325	2880	10000	0	1600000	808664	1013645

Appendix 5: Residual Charts for Multiple Linear Regression


Residuals vs Fitted
Residuals
4e+05
0e+00
-4e+05
-8e+05
0e+00
1e+06
2e+06
3e+06
21
32
39
Fitted values
lm(current_mean_val ~ diff + mean_trans_val + count_transfers + count_uniqu ...


Normal Q-Q
Standardized residuals
2
0
-2
-4
-2
-1
0
1
2
21
32
39
Theoretical Quantiles
lm(current_mean_val ~ diff + mean_trans_val + count_transfers + count_uniqu ...


Scale-Location
√|Standardized residuals|
2.0
1.5
1.0
0.5
0.0
0e+00
1e+06
2e+06
3e+06
39
21
32
Fitted values
lm(current_mean_val ~ diff + mean_trans_val + count_transfers + count_uniqu ...



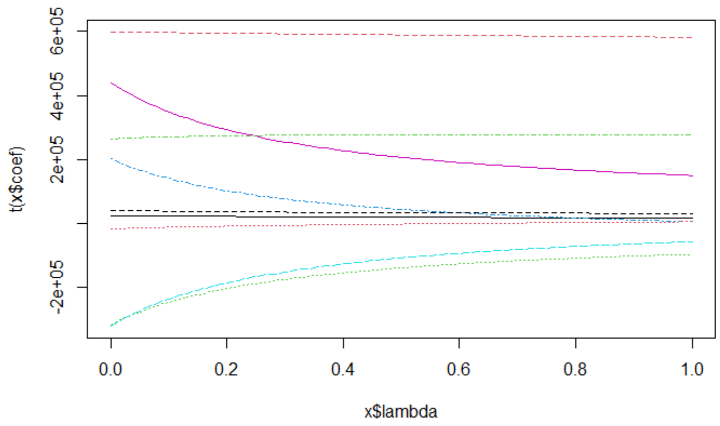
Appendix 6: Variance Inflation Factors

Mean VIF: 18.30504

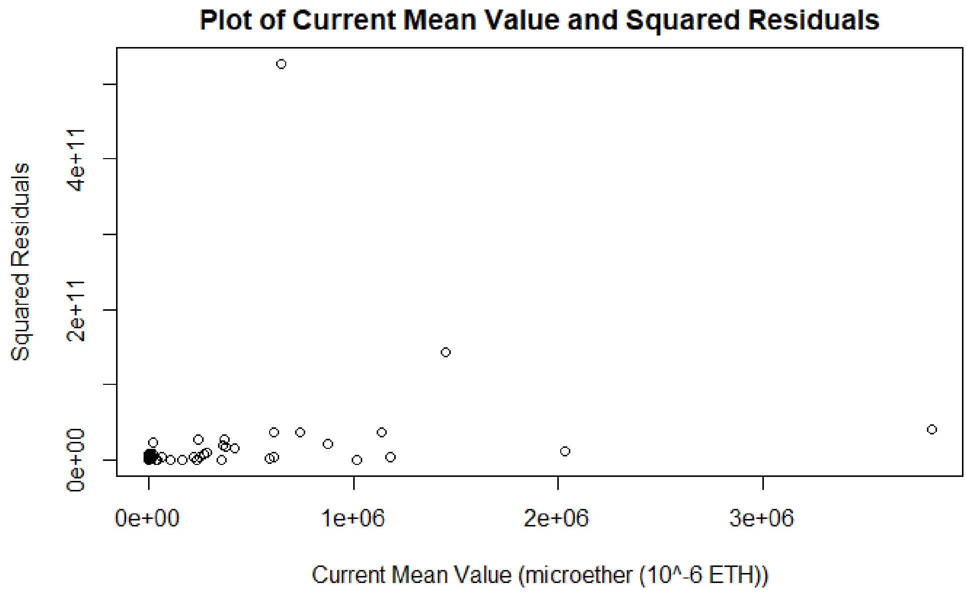
Max VIF: 53.38582

difference in time	mean trans val	count transfers	count unique trans	count unique owners	count nfts	mint min val	mint max val	mint mean val
1.710925	1.03978	31.86507	26.503019	43.60279	53.3858	1.127560	2.639146	2.871243

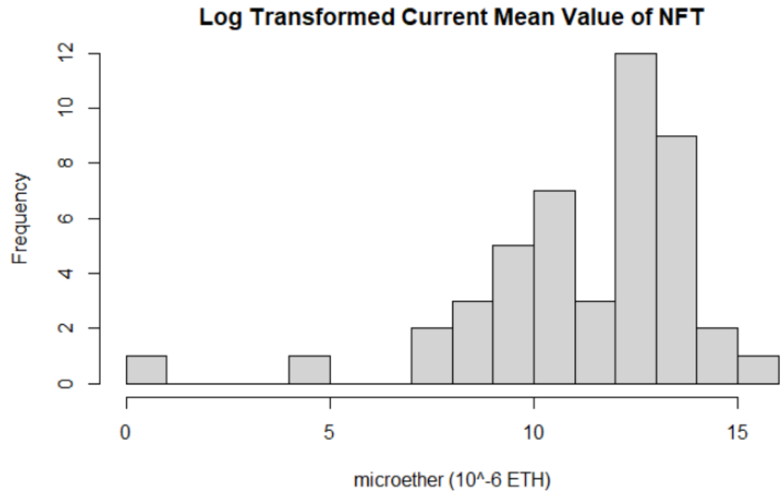
Appendix 7: Ridge Trace Plot



Appendix 8: Plot of Current Mean Value and Squared Residuals



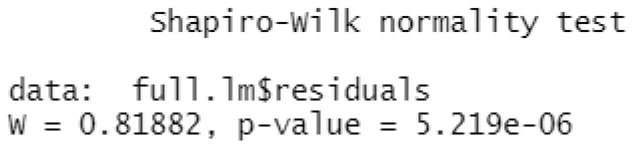
Appendix 9: Log Transformed Histogram of Current Mean NFT Prices



Appendix 10: R Packages

(tidyverse), (foreign), (ggplot2), (MASS), (car), (leaps), (stats), (ISLR), (glmnet), (randomForest), (caTools), (varImp)

Appendix 11: Shapiro Test



```
        Shapiro-Wilk normality test

data:  full.lm$residuals
W = 0.81882, p-value = 5.219e-06
```

Appendix 12: Root Mean Square Error Formula

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} \| y(i) - \hat{y}(i) \|^2}{N}},$$

N: number of observations,

Y(i): i-th measurement

Y(hat)(i): prediction corresponding to the i-th measurement

Appendix 13: Regression Equations

A. Linear Regression

current mean value = B0 + B1*(difference) + B2*(mean transfer value) + B3*(count of transfers) + B4*(count of unique transfers) + B5*(count unique owners) + B5*(count of NFTs) + B6*(mint minimum value) + B7*(mint maximum value) + B8*(mint mean value)

Appendix 14: Weighted Least Squares Regression Equation

wls_practice.lm <- lm(sqr_resid ~ diff + mean_trans_val + count_transfers + count_unique_trans + count_unique_owners + count_nfts + mint_min_val + mint_max_val + mint_mean_val, data=combo_df_wls)

```
full.wls <- lm(current_mean_val ~ diff + mean_trans_val + count_transfers +
count_unique_trans + count_unique_owners + count_nfts + mint_min_val + mint_max_val +
mint_mean_val, data=combo_df_wls, weights=(1/(wls_practice.lm$fit)^2))
```

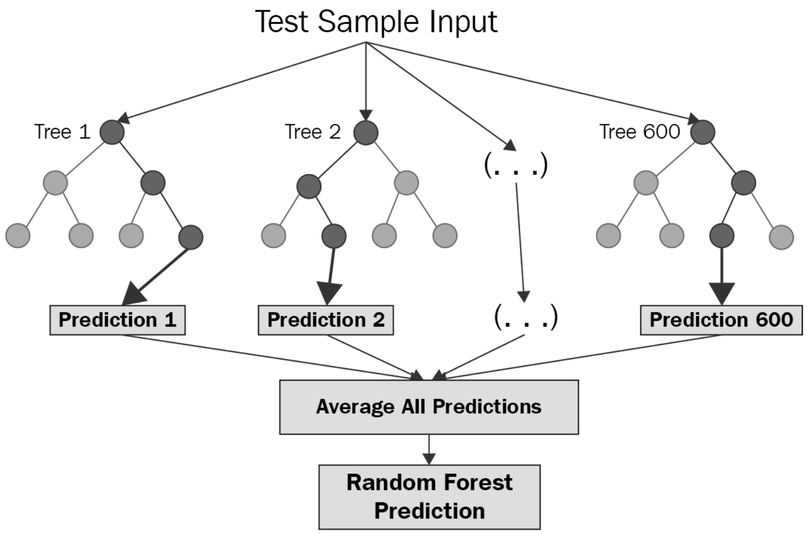
Appendix 15: Outliers and Influential Points

	Cook's Distance	Cook’s Cutoff	Influential	studentized residuals	t-value	outlier
1	0.000838	0.086957	No	-0.338158691	2.019541	Non-outlier
2	0.005297	0.086957	No	-0.552414978	2.019541	Non-outlier
3	0.009694	0.086957	No	-0.557105515	2.019541	Non-outlier
4	7.00E-06	0.086957	No	0.037727331	2.019541	Non-outlier
5	1.10E-05	0.086957	No	0.030902007	2.019541	Non-outlier
6	2.20E-05	0.086957	No	-0.060034747	2.019541	Non-outlier
7	0.109314	0.086957	Influential	-1.410213295	2.019541	Non-outlier
8	0.005092	0.086957	No	0.79821578	2.019541	Non-outlier
9	0.007189	0.086957	No	-0.82956843	2.019541	Non-outlier
10	0.001484	0.086957	No	0.251676291	2.019541	Non-outlier
11	0.002509	0.086957	No	0.713288813	2.019541	Non-outlier
12	0.000207	0.086957	No	-0.114531218	2.019541	Non-outlier
13	0.050559	0.086957	No	0.547248025	2.019541	Non-outlier
14	0.003797	0.086957	No	-0.887077976	2.019541	Non-outlier
15	0.000858	0.086957	No	0.345920947	2.019541	Non-outlier
16	0.000835	0.086957	No	-0.21942891	2.019541	Non-outlier
17	0.000271	0.086957	No	0.19199549	2.019541	Non-outlier
18	4.00E-06	0.086957	No	0.024223608	2.019541	Non-outlier
19	6.20E-05	0.086957	No	-0.09027927	2.019541	Non-outlier
20	0.124387	0.086957	Influential	1.441256328	2.019541	Non-outlier

21	0.098404	0.086957	Influential	2.554028685	2.019541	outlier
22	9.197065	0.086957	Influential	1.780222276	2.019541	Non-outlier
23	0.000564	0.086957	No	-0.354404154	2.019541	Non-outlier
24	0.921632	0.086957	Influential	-1.197468218	2.019541	Non-outlier
25	0.00026	0.086957	No	-0.129426021	2.019541	Non-outlier
26	2.00E-05	0.086957	No	0.052790113	2.019541	Non-outlier
27	6.10E-05	0.086957	No	-0.111355461	2.019541	Non-outlier
28	3.00E-05	0.086957	No	0.074146724	2.019541	Non-outlier
29	0.002267	0.086957	No	-0.380081823	2.019541	Non-outlier
30	0.048988	0.086957	No	-0.532007554	2.019541	Non-outlier
31	0.572068	0.086957	Influential	-1.709557571	2.019541	Non-outlier
32	1.812712	0.086957	Influential	2.572624518	2.019541	outlier
33	0.005197	0.086957	No	0.862373186	2.019541	Non-outlier
34	0.019376	0.086957	No	1.200964669	2.019541	Non-outlier
35	0.004191	0.086957	No	0.585862786	2.019541	Non-outlier
36	0.000316	0.086957	No	0.20875864	2.019541	Non-outlier
37	3.353684	0.086957	Influential	1.980344175	2.019541	Non-outlier
38	0.011688	0.086957	No	-1.007199869	2.019541	Non-outlier
39	0.310646	0.086957	Influential	-6.802407934	2.019541	outlier
40	0.001178	0.086957	No	0.44036498	2.019541	Non-outlier
41	0.000509	0.086957	No	0.285097353	2.019541	Non-outlier
42	0.000361	0.086957	No	0.281007595	2.019541	Non-outlier
43	0	0.086957	No	-0.002994113	2.019541	Non-outlier
44	0.000904	0.086957	No	0.384762015	2.019541	Non-outlier
45	0.001242	0.086957	No	0.523665611	2.019541	Non-outlier
46	0.048862	0.086957	No	-0.577575587	2.019541	Non-outlier

Appendix 16: Random Forest Model

Random Forest Overview: supervised learning algorithm that uses ensemble learning method for regression. This technique combines predictions from multiple decision trees to make a more accurate prediction than a single model.



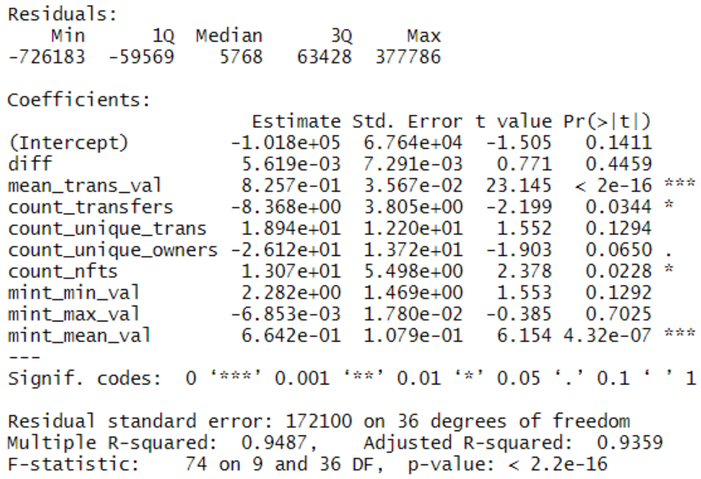
Appendix 17: Model Performance Metrics

	Linear Model	Robust Regression	Ridge Regression	WLS	Random Forest
MSE	90.06	87.2908	86.2126	128.8266	216.13
RMSE	94900.88	93429.54	92850.74	113501.8	147013.6
AIC	1250.417	1252.579	NA	1214.66	NA
BIC	1270.532	1272.694	NA	1234.775	NA
Adj R-Squared	0.9359008	NA	NA	0.9907447	NA

	Three Predictor	Eight Predictor	Transfer Interaction Model	Log Linear Model	Only Transfer Model	No Transfer Linear Model
MSE	86.32751	86.96879	69.33876	1339.609	284.7817	696.6222
RMSE	92912.6	93257.06	83269.9	366006.7	168754.8	263936
AIC	1248.149	1248.606	1194.879	NA	1315.269	1375.612
BIC	1257.292	1266.893	1229.623	NA	1320.755	1393.899
Adj R-Squared	0.932114	0.9373764	0.9825984	NA	0.695883	0.009566915

Appendix 18: Model Outputs

A. Linear Regression

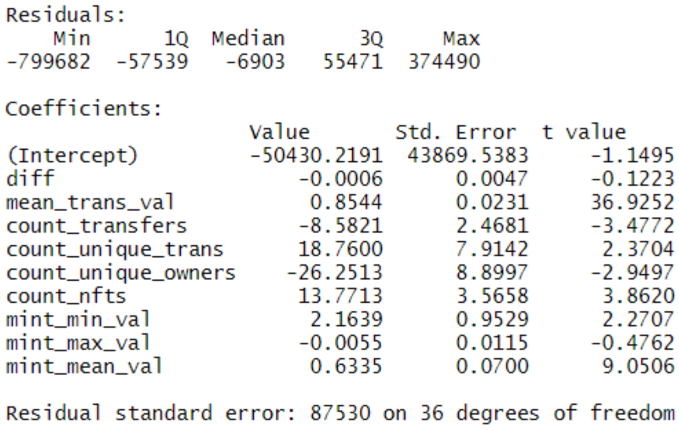


```
Residuals:
    Min      1Q  Median      3Q     Max 
-726183  -59569    5768   63428  377786 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)         -1.018e+05  6.764e+04  -1.505   0.1411    
diff                 5.619e-03  7.291e-03   0.771   0.4459    
mean_trans_val       8.257e-01  3.567e-02  23.145  < 2e-16 ***
count_transfers     -8.368e+00  3.805e+00  -2.199   0.0344 *  
count_unique_trans   1.894e+01  1.220e+01   1.552   0.1294    
count_unique_owners -2.612e+01  1.372e+01  -1.903   0.0650 .  
count_nfts           1.307e+01  5.498e+00   2.378   0.0228 *  
mint_min_val         2.282e+00  1.469e+00   1.553   0.1292    
mint_max_val        -6.853e-03  1.780e-02  -0.385   0.7025    
mint_mean_val        6.642e-01  1.079e-01   6.154 4.32e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 172100 on 36 degrees of freedom
Multiple R-squared:  0.9487,	Adjusted R-squared:  0.9359 
F-statistic:    74 on 9 and 36 DF,  p-value: < 2.2e-16
```

B. Robust Regression

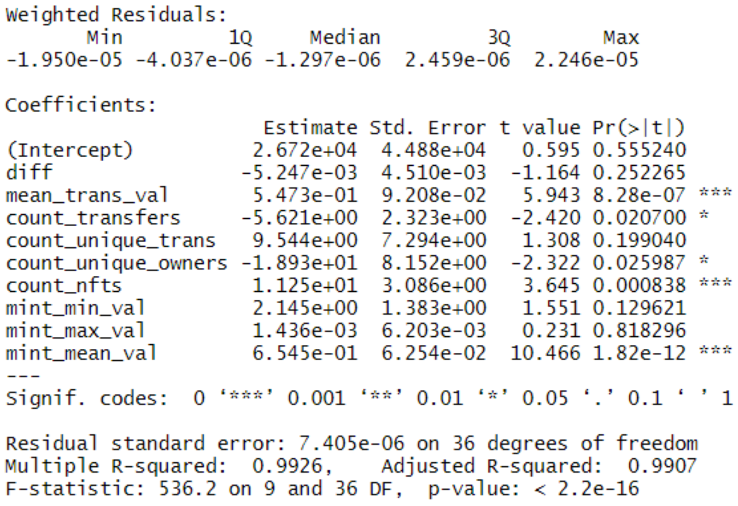


```
Residuals:
    Min      1Q  Median      3Q     Max 
-799682  -57539   -6903   55471  374490 

Coefficients:
                         Value      Std. Error  t value
(Intercept)         -50430.2191  43869.5383     -1.1495
diff                    -0.0006      0.0047     -0.1223
mean_trans_val           0.8544      0.0231     36.9252
count_transfers         -8.5821      2.4681     -3.4772
count_unique_trans      18.7600      7.9142      2.3704
count_unique_owners    -26.2513      8.8997     -2.9497
count_nfts              13.7713      3.5658      3.8620
mint_min_val             2.1639      0.9529      2.2707
mint_max_val            -0.0055      0.0115     -0.4762
mint_mean_val            0.6335      0.0700      9.0506

Residual standard error: 87530 on 36 degrees of freedom
```

C. Weighted Least Squares Regression

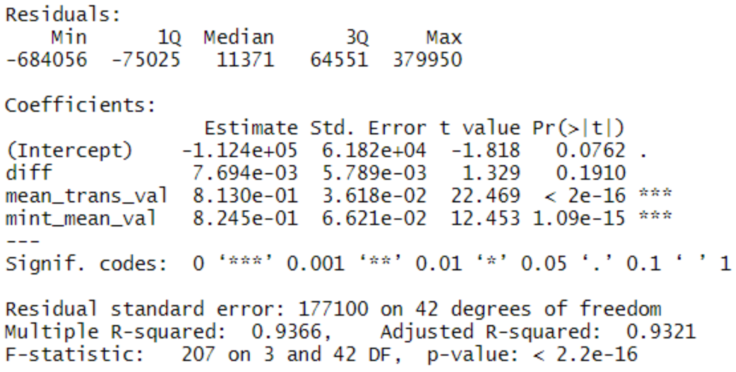


```
Weighted Residuals:
       Min         1Q     Median         3Q        Max 
-1.950e-05 -4.037e-06 -1.297e-06  2.459e-06  2.246e-05 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)         2.672e+04  4.488e+04   0.595 0.555240    
diff               -5.247e-03  4.510e-03  -1.164 0.252265    
mean_trans_val      5.473e-01  9.208e-02   5.943 8.28e-07 ***
count_transfers    -5.621e+00  2.323e+00  -2.420 0.020700 *  
count_unique_trans  9.544e+00  7.294e+00   1.308 0.199040    
count_unique_owners -1.893e+01  8.152e+00  -2.322 0.025987 *  
count_nfts          1.125e+01  3.086e+00   3.645 0.000838 ***
mint_min_val        2.145e+00  1.383e+00   1.551 0.129621    
mint_max_val        1.436e-03  6.203e-03   0.231 0.818296    
mint_mean_val       6.545e-01  6.254e-02  10.466 1.82e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.405e-06 on 36 degrees of freedom
Multiple R-squared:  0.9926,	Adjusted R-squared:  0.9907 
F-statistic: 536.2 on 9 and 36 DF,  p-value: < 2.2e-16
```

D. Three Predictor Model Linear Regression

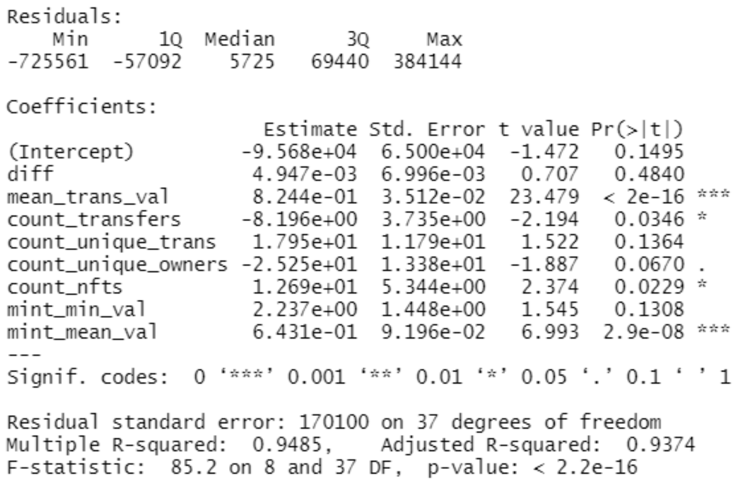


```
Residuals:
    Min      1Q  Median      3Q     Max 
-684056  -75025   11371   64551  379950 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -1.124e+05  6.182e+04  -1.818   0.0762 .  
diff           7.694e-03  5.789e-03   1.329   0.1910    
mean_trans_val 8.130e-01  3.618e-02  22.469  < 2e-16 ***
mint_mean_val  8.245e-01  6.621e-02  12.453 1.09e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 177100 on 42 degrees of freedom
Multiple R-squared:  0.9366,	Adjusted R-squared:  0.9321 
F-statistic:   207 on 3 and 42 DF,  p-value: < 2.2e-16
```

E. Eight Predictor Model Linear Regression

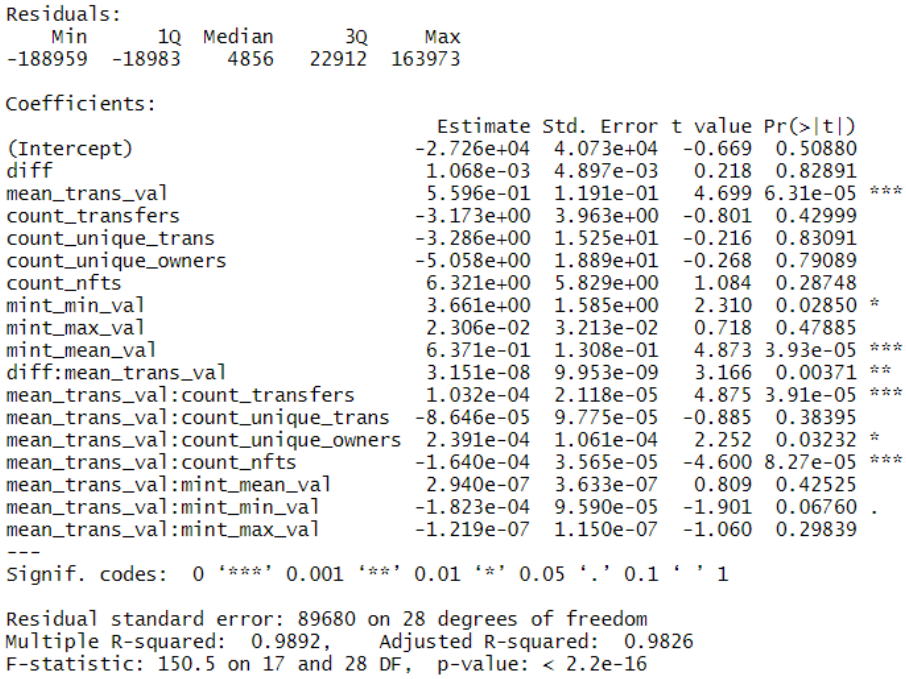


```
Residuals:
    Min      1Q  Median      3Q     Max 
-725561  -57092    5725   69440  384144 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)         -9.568e+04  6.500e+04  -1.472   0.1495    
diff                 4.947e-03  6.996e-03   0.707   0.4840    
mean_trans_val       8.244e-01  3.512e-02  23.479  < 2e-16 ***
count_transfers     -8.196e+00  3.735e+00  -2.194   0.0346 *  
count_unique_trans   1.795e+01  1.179e+01   1.522   0.1364    
count_unique_owners -2.525e+01  1.338e+01  -1.887   0.0670 .  
count_nfts           1.269e+01  5.344e+00   2.374   0.0229 *  
mint_min_val         2.237e+00  1.448e+00   1.545   0.1308    
mint_mean_val        6.431e-01  9.196e-02   6.993  2.9e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 170100 on 37 degrees of freedom
Multiple R-squared:  0.9485,    Adjusted R-squared:  0.9374 
F-statistic:  85.2 on 8 and 37 DF,  p-value: < 2.2e-16
```

F. Mean Transfer Value Interaction Linear Model

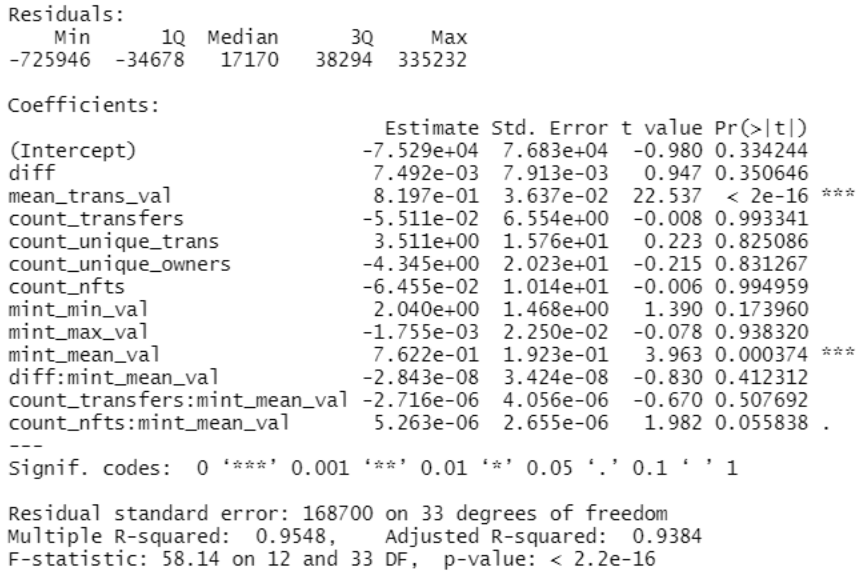


```
Residuals:
    Min      1Q  Median      3Q     Max 
-188959  -18983    4856   22912  163973 

Coefficients:
                                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)                        -2.726e+04  4.073e+04  -0.669  0.50880    
diff                                1.068e-03  4.897e-03   0.218  0.82891    
mean_trans_val                      5.596e-01  1.191e-01   4.699 6.31e-05 ***
count_transfers                    -3.173e+00  3.963e+00  -0.801  0.42999    
count_unique_trans                 -3.286e+00  1.525e+01  -0.216  0.83091    
count_unique_owners                -5.058e+00  1.889e+01  -0.268  0.79089    
count_nfts                          6.321e+00  5.829e+00   1.084  0.28748    
mint_min_val                        3.661e+00  1.585e+00   2.310  0.02850 *  
mint_max_val                        2.306e-02  3.213e-02   0.718  0.47885    
mint_mean_val                       6.371e-01  1.308e-01   4.873 3.93e-05 ***
diff:mean_trans_val                 3.151e-08  9.953e-09   3.166  0.00371 ** 
mean_trans_val:count_transfers      1.032e-04  2.118e-05   4.875 3.91e-05 ***
mean_trans_val:count_unique_trans  -8.646e-05  9.775e-05  -0.885  0.38395    
mean_trans_val:count_unique_owners  2.391e-04  1.061e-04   2.252  0.03232 *  
mean_trans_val:count_nfts          -1.640e-04  3.565e-05  -4.600 8.27e-05 ***
mean_trans_val:mint_mean_val        2.940e-07  3.633e-07   0.809  0.42525    
mean_trans_val:mint_min_val        -1.823e-04  9.590e-05  -1.901  0.06760 .  
mean_trans_val:mint_max_val        -1.219e-07  1.150e-07  -1.060  0.29839    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 89680 on 28 degrees of freedom
Multiple R-squared:  0.9892,    Adjusted R-squared:  0.9826 
F-statistic: 150.5 on 17 and 28 DF,  p-value: < 2.2e-16
```

G. Mean Minting Value Interaction Linear Model



```
Residuals:
    Min      1Q  Median      3Q     Max 
-725946  -34678   17170   38294  335232 

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)    
(Intercept)                  -7.529e+04  7.683e+04  -0.980 0.334244    
diff                          7.492e-03  7.913e-03   0.947 0.350646    
mean_trans_val                8.197e-01  3.637e-02  22.537  < 2e-16 ***
count_transfers              -5.511e-02  6.554e+00  -0.008 0.993341    
count_unique_trans            3.511e+00  1.576e+01   0.223 0.825086    
count_unique_owners          -4.345e+00  2.023e+01  -0.215 0.831267    
count_nfts                   -6.455e-02  1.014e+01  -0.006 0.994959    
mint_min_val                  2.040e+00  1.468e+00   1.390 0.173960    
mint_max_val                 -1.755e-03  2.250e-02  -0.078 0.938320    
mint_mean_val                 7.622e-01  1.923e-01   3.963 0.000374 ***
diff:mint_mean_val           -2.843e-08  3.424e-08  -0.830 0.412312    
count_transfers:mint_mean_val -2.716e-06  4.056e-06  -0.670 0.507692    
count_nfts:mint_mean_val      5.263e-06  2.655e-06   1.982 0.055838 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 168700 on 33 degrees of freedom
Multiple R-squared:  0.9548,	Adjusted R-squared:  0.9384 
F-statistic: 58.14 on 12 and 33 DF,  p-value: < 2.2e-16
```