Card Price Prediction for Magic The Gathering

Victoria Dorn, Victor Hugo Germano

Shiley-Marcos School of Engineering, University of San Diego

AAI-500: Probability and Statistics

Msc. Leon Shpaner Teacher

October 15, 2024

Abstract

Magic The Gathering Card Game is one of the most famous card games in the world, with a thriving community of players, championships, physical and online games, and a big market for buying and selling cards worldwide. This article takes a deep dive approach using Exploration Data Analysis and Machine Learning techniques to present models of predicting prices based on various information about the game. We begin by analyzing card characteristics and their relationship with the pricing information, generating a basis of predictors like rarity, availability, power, artist (to name a few), and our selected target of price information relative to a single day in the market. Linear Regression, Ordinary Least Square Regressions, Self Organizing maps, and Random Forest Regression were used and compared on their predictive capabilities using our model of numerical and categorical data, and using the Mean Squared Error and R-squared, concluding that EDH Ranking and EDH Saltiness play a significant role as price predictors, beyond external market factors. Future research could try incorporating more information beyond Card attributes to improve the model.

Keywords: regression, machine learning, price prediction, card games

Introduction

Magic the Gathering is one of the most famous card games in the world. First published in 1993 by Wizards of the Coast, it has become a benchmark of modern fantasy games, getting more than 40 million players worldwide (Schmidt., 2023). With more than 23,000 unique cards, the game is a phenomenon, having worldwide championships with cash prizes up to $10 million. For example, The Las Vegas Grand Prix championship receives close to 11,000 participants every edition (Grand Prix, 2024). The game also has a thriving primary and secondary market for cards. The rarest card in the game, the Black Lotus, has been sold for $600,000 in an auction, with some cards receiving even higher expected value.

We believe that analyzing card assets in relation to their general prices in the secondary market can offer valuable insights to collectors. The primary object of this project is to statistically analyze the price variable against various features using methods learned in this course, and employ regression modeling to predict prices. Given the maturity of the Magic: The Gathering card market, we have identified an open-source dataset that includes current (or relatively recent) market prices from a single market day.

This dataset allows us to focus only on key card features, resulting in a simplified solution that does not account for other variables, such as cards used in recent tournaments, potential card bans, or the influence of newly created cards on prices. Secondary markets are influenced by many factors when pricing a card, many of which are external to the data used in the article, and that could be used in future research. Synergy between cards during gameplay, deck strategy in Tournaments, exclusive art, and celebrity endorsement can affect the price and

will not be used in the analysis. We encourage follow-up work to incorporate these and other factors to enhance the fidelity and usability of these models.

The open-source dataset will serve as our ground truth, rather than aggregating data from multiple marketplaces. In this technical report, we will outline and examine our approach to classification and regression techniques for predicting card prices. Specifically, we will use a Self-Organizing Map (SOM) for data visualization and analysis, followed by a comparison of Polynomial Regression, Ordinary Least Squares (OLS) Regression, and Random Forest Regression techniques to identify the best price prediction model for our dataset.

Overview of Magic: The Gathering, the Game

In Magic: The Gathering (MTG), players take on the role of wizards known as Planeswalkers, summoning creatures, casting spells, and engaging in battles. There are 19 different gameplay types ranging from casual formats to multiplayer and digital formats. In the Standard constructed format each player aims to reduce their opponent's life total from 20 to 0 or to outmaneuver them.

Game Components

At the heart of MTG are the cards themselves. Each player builds a deck containing at least 60 cards, which can include various types of cards that serve different purposes:

- Lands are essential for generating mana, the resource required to cast spells. Players can play one land per turn, which helps fuel their magical abilities.

- Creatures can attack opponents and defend against incoming threats. Each creature has two important stats: power, which indicates how much damage it deals in combat, and toughness, which shows how much damage it can withstand before defeate.
- Sorceries represent one-time effects that can be played during a player's main phase.
- Enchantments and Artifacts provide ongoing effects or abilities that can influence the game in various ways.

Turn Structure

The game happens in a series of turns, each turn consists of several phases. It begins with the Untap Phase, where the active player untaps all tapped cards. This is followed by the Upkeep Phase, during which any triggered abilities are resolved. Next, in the Draw Phase, the player draws a card from their library.

During Main Phase 1, players can play one land, cast spells, and summon creatures. The Combat Phase then occurs, where the active player declares attackers, the defending player chooses blockers, and damage is resolved based on power and toughness. After combat, players enter Main Phase 2, allowing for additional spells and land plays. Finally, the turn ends with the End Phase, where any end-of-turn effects are resolved, and players discard down to seven cards if necessary. Each phase provides opportunities for strategic decision-making that shapes the game.

Victory in MTG can come through various means. The most common way is by reducing an opponent's life total from 20 to 0 through strategic attacks and effective spellcasting. Alternatively, a player can win if an opponent cannot draw a card when required because their library is empty.

Card Information

Each card possesses several key attributes that define its role and functionality in the game, Figure A1 and A2 (see Appendix) for further details. Firstly, card type categorizes the card into groups such as creatures, instants, sorceries, enchantments, artifacts, and lands. Each type has specific rules governing its use. For instance, creatures can attack and block, while instants can be played at almost any time. Each card also has a mana cost, indicated in the upper right corner, which specifies the amount and type of mana required to cast it. This cost is crucial for determining when and how a player can use the card.

For creature cards, two important statistics are displayed in the bottom right corner: power and toughness. Power indicates how much damage the creature deals in combat, while toughness shows how much damage it can take before being destroyed. For example, a creature with a power of 5 and toughness of 5 can deal 5 damage and withstand 5 damage before being sent to the graveyard.

Additionally, many cards feature keywords that provide special abilities or effects. Examples include flying (allowing a creature to avoid being blocked except by other flying creatures), haste (enabling a creature to attack immediately after being summoned), and indestructible (preventing a permanent from being destroyed by damage or effects).

Data Cleaning and Preparation

For dataset selection, we found an existing open-source project, MTGJSON, that gears itself towards data aggregation and organization across all MTG play formats (Halpern 2024). They have APIs already set up to pull new information daily along with thorough documentation, which were the main deciding factors in selecting their datasets for use in this project. Specifically, we utilized two main datasets for our analysis, one containing the general Card Attributes for all printed cards coupled with the second that contained card Pricing information on a single market day.

Before the data cleaning process, the Card dataset consisted of 97,145 individual card entries and the Price dataset consisted of 558,079 total entries from 91,302 unique card entries. Further examination of these datasets showed 25 possible variables for each card entry, some of which include Name, Game Attributes, Artist, Collection, EDH Ranking, Card Finish and a mix of other qualitative and quantitative variables, and 8 possible variables for each card entry, some of which include Currency and Price Provider.

Next, we will delineate the steps undertaken to clean and prepare the dataset, to foster transparency in our analytical process as a whole. To streamline our evaluation and reduce complexity of our multidimensional dataset, we reviewed both datasets individually prior to merging them. First, we decided not to incorporate exchange rates and instead selected a single currency. There were 432,688 USD entries and 125,391 EUR entries, so we selected USD for our price comparisons. Subsequently, we removed all constants from both datasets. This meant eliminating any column that contained only a single unique value throughout the data. The columns discarded included Currency (USD) and Data (2024-09-20) from the prices dataset.

An initial comparison of our two datasets led us to determine that there were duplicate variables. Since our end goal with data cleaning was to create a comprehensive primary dataset we removed the ***Finishes***, ***HasFoil***, ***HasNonFoil***, and ***SourceProducts*** column from the Card dataset since this information would be reflected more accurately in the Price dataset ***CardFinish*** variable. Additionally, we filled in the ***isReprint*** variable missing values with False since there were only True or NaN values on ingest. The Colors variables also had repeat permutations of string values, so we cleaned that up from 41 to 32 unique values.

To create our primary dataset, we performed an inner join utilizing the ***uuid*** column, which contained a unique identification number for each individual card. Doing this ensured that only cards present in both datasets would be retained then we dropped missing values. This approach was chosen for our primary dataset to maintain the integrity of the data, especially so missing values did not skew results. The primary dataset enables a more comprehensive analysis, allowing us to examine more features and their corresponding correlations to the price of the card. It is composed of 30 variables and 17,628 total card entries.

Upon further evaluation of the columns we dropped, we decided to create a second, higher-fidelity dataset. Many variables that capture essential game mechanics were removed when we dropped missing values. To preserve card types beyond just Creature, it became necessary to retain the ***Power*** and ***Toughness*** attributes, which were among the missing data. Similarly, most Instant and Sorceries do not have ***Supertypes***, so we opted to retain those missing values as well. Additionally, for the ***Color*** and ***ColorIdentity*** attributes, which had missing values, we choose to keep these as unique identifiers, such as “C” for colorless.

This second dataset allows us to maintain the integrity of the dataset while ensuring that we capture the full range of card types and their characteristics. The second dataset contains 269,807 total card entries and 27 variables.

Given that we wanted to explore categorical variables in our analysis, the variables in both datasets were converted into numerical formats using label encoding. This transformation was applied to all columns, excluding ***uuid***. By encoding these variables, we ensured that they could be effectively included in subsequent modeling, allowing us to explore their impact on card prices.

To further enhance the reliability of the experiment results, ***price*** outliers were identified and removed using the Interquartile Range method (Agresti & Kateri, 2021) . This method calculates the first and third quartiles to determine an acceptable range for data points. Any prices falling outside this range were considered outliers and excluded from further analysis. This step was critical in ensuring that extreme values did not distort the regression results, especially given that there were 2,414 and 41,170 ***price*** outliers in our primary and secondary datasets respectively.

In the end of our preprocessing the secondary dataset contained the 26 variables: *price, artist, cardFinish, colorIdentity, colors, edhrecRank, ehrecSaltiness, gameAvailability, isReprint, language, layout, manaCost, manaValue, name, number, originalType, power, priceProvider, providerListing, rarity, setCode, supertypes, toughness, type, types,* and *uuid.* The primary dataset contained one less variable (*language*) and consisted of 15,214 entries across 3,123 unique cards. The secondary dataset consisted of 228,637 entries across 50,208 unique cards.

In card gaming resell markets, it is common to use and interpret *Rarity* and *Collection*(setCode) as empirical drivers for price. We want to use the relationship between the

variables and clustering algorithms to understand if there are more correlations available prior to predicting prices.

Exploratory Data Analysis

This section aims to uncover patterns and insights within our preprocessed datasets. By thoroughly examining attributes, we will explore the relationships between these characteristics and their impacts on price. Through this analysis, we seek to provide valuable insights for our downstream models and assess the null hypothesis for each variable regarding its influence on price. The same analysis approach was taken for both the primary and secondary datasets.

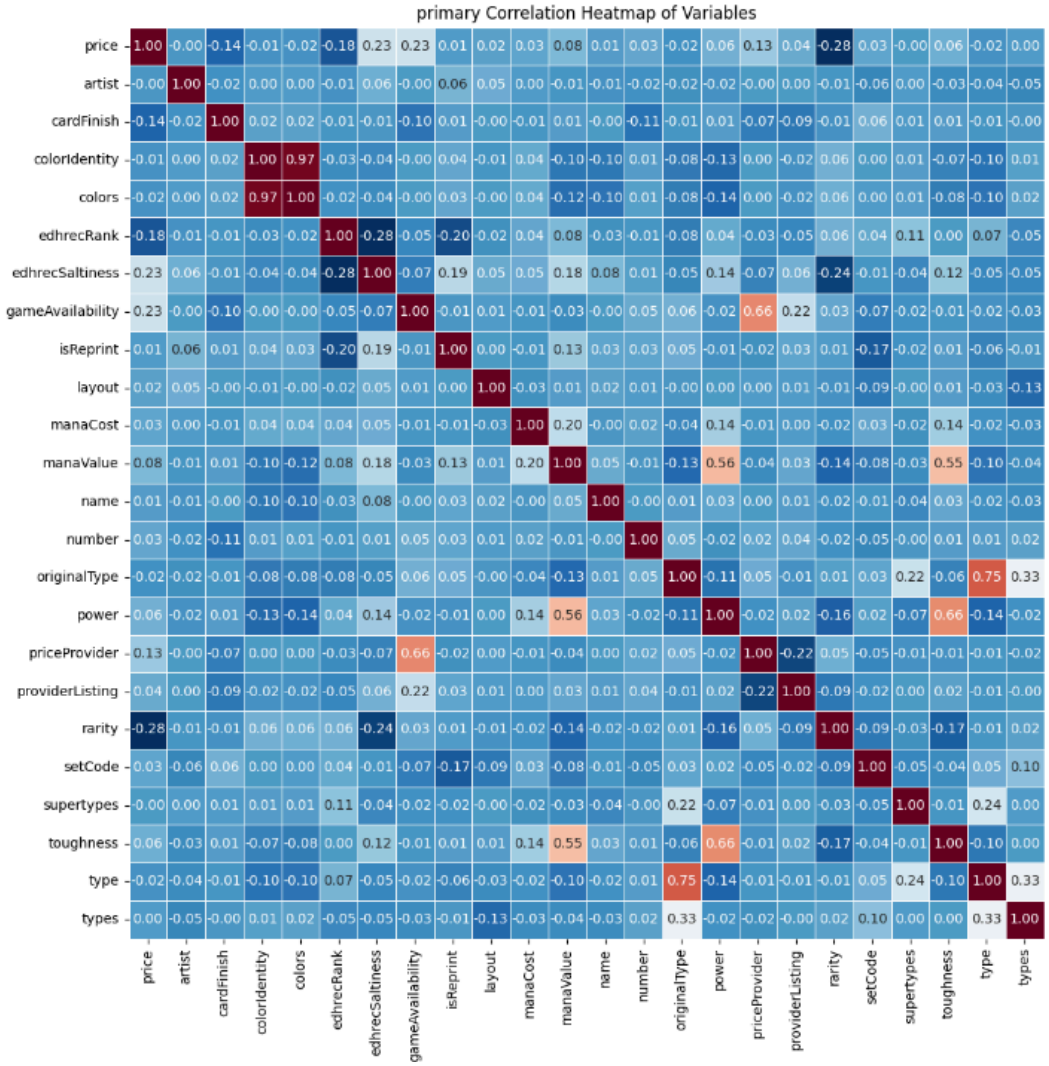
Primary Dataset Exploration

To start evaluation we explored the correlations between variables to get a better understanding if our dataset contained multicollinearity and the impacts of that on regression models. After seeing the strong positive correlation between *colorIdentity* and *colors* at 0.97 and *originalType* and *type* at 0.75, we calculated a Variance Inflation Factor (VIF) to quantify and identify how redundant our variables are (Penn State Eberly College of Science 2018). The VIF values greater than 10 were *colorIdentity* at 20.31 and colors at 20.38, so we dropped the *colorIdentity* column from our primary dataset.

Next, we utilized the Test statistic to measure how far each point estimate deviates from the parameter value of ***price*** in relation to each variable. Additionally, we calculated the p-value to determine whether to reject or fail to reject the null hypothesis regarding the influence of each variable on ***price***. Some key negative relationships to price are isReprint and edhrecSaltiness. Another note is that the priceProvider has some influence on price but not as much as the other relationships with price.

Figure 1

Correlation heatmap of the Primary dataset after feature mapping, but before dropping and transforming features.



Secondary Dataset Exploration

For the secondary dataset we did a similar approach, beginning with coefficient correlations and VIF calculations. The VIF calculations resulted in a 26.92 for *colorIdentity* and a 26.93 for colors, so again we dropped *colorIdentity* from the dataset. The secondary results

from statistical testing indicate that nearly all variables have extremely high T-scores and p-values of 0, suggesting a very strong statistical significance in their relations with price.

Figure 2a and 2b

Shows T-score and P-values for each feature for the primary (Figure 2a, on the left) and secondary (Figure 2b, on the right) datasets.

Primary Dataset Results:

Variable	T-score	P-value
artist	216.822599	0.000000e+00
cardFinish	27.089676	1.704672e-158
colorIdentity	211.185150	0.000000e+00
colors	214.102136	0.000000e+00
edhrecRank	161.394553	0.000000e+00
edhrecSaltiness	-63.338320	0.000000e+00
gameAvailability	-32.594882	4.997116e-226
isReprint	-63.939938	0.000000e+00
layout	196.797219	0.000000e+00
manaCost	172.520311	0.000000e+00
manaValue	160.126938	0.000000e+00
name	216.281241	0.000000e+00
number	157.891695	0.000000e+00
originalType	261.491856	0.000000e+00
power	222.045356	0.000000e+00
priceProvider	14.721119	7.443080e-49
providerListing	-78.676648	0.000000e+00
rarity	46.393962	0.000000e+00
setCode	191.189442	0.000000e+00
supertypes	-90.274842	0.000000e+00
toughness	241.450764	0.000000e+00
type	224.161196	0.000000e+00
types	52.260535	0.000000e+00

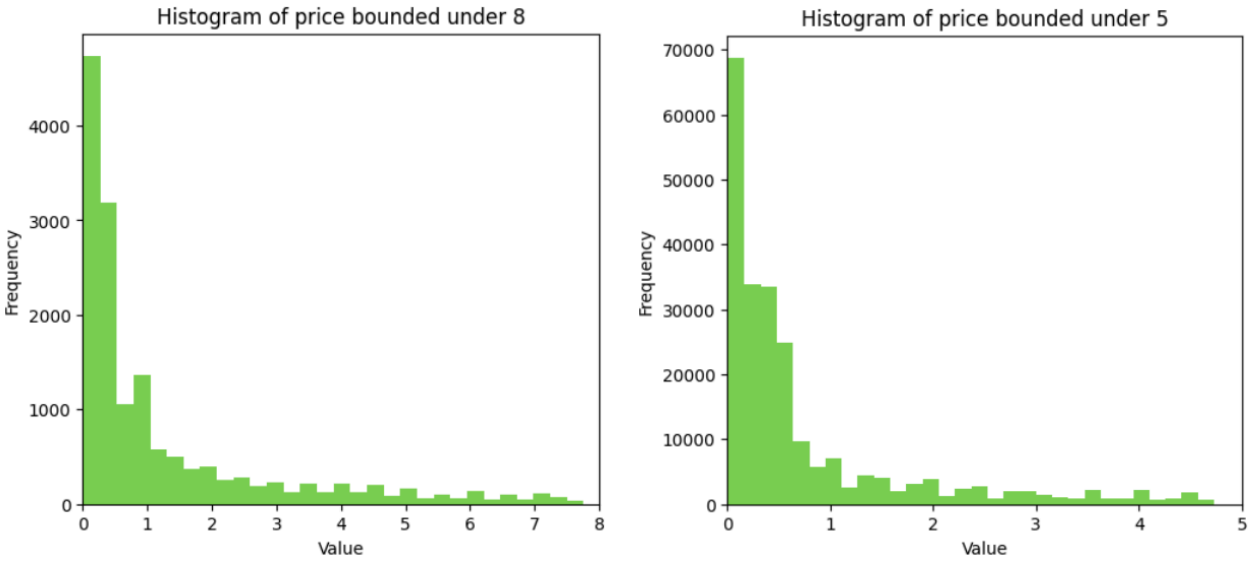
Secondary Dataset Results:

Variable	T-score	P-value
artist	800.471980	0.000000e+00
cardFinish	409.419464	0.000000e+00
colorIdentity	841.600268	0.000000e+00
colors	848.585308	0.000000e+00
edhrecRank	622.054171	0.000000e+00
edhrecSaltiness	-205.092953	0.000000e+00
gameAvailability	29.658054	5.119910e-193
isReprint	-106.482666	0.000000e+00
language	590.618660	0.000000e+00
layout	2345.247354	0.000000e+00
manaCost	668.759183	0.000000e+00
manaValue	650.367138	0.000000e+00
name	823.969217	0.000000e+00
number	656.399831	0.000000e+00
originalType	985.763186	0.000000e+00
power	1131.055230	0.000000e+00
priceProvider	248.340464	0.000000e+00
providerListing	-268.418148	0.000000e+00
rarity	584.873268	0.000000e+00
setCode	654.309904	0.000000e+00
supertypes	401.622635	0.000000e+00
toughness	1268.420404	0.000000e+00
type	828.734928	0.000000e+00
types	504.929862	0.000000e+00

To conclude our exploratory analysis, we examined the price distributions for each dataset. On the left, Figure 3a displays the histogram of prices from the primary datasets, where values are capped at $8 due to outlier removal, resulting in a maximum price of $7.76, a mean of $1.27 and a standard deviation of $1.68. On the right, Figure 3b shows the histogram for the second dataset, which was bounded at $5, with a maximum price of $4.73, a mean of $0.73, and

Figure 3a and 3b

Historgrams for price feature in each datasets (a is primary, b is secondary), without outliers.



a standard deviation of $1.00. This highlights the left-skewed nature of our price data, as the majority of MTG cards are priced below $1.

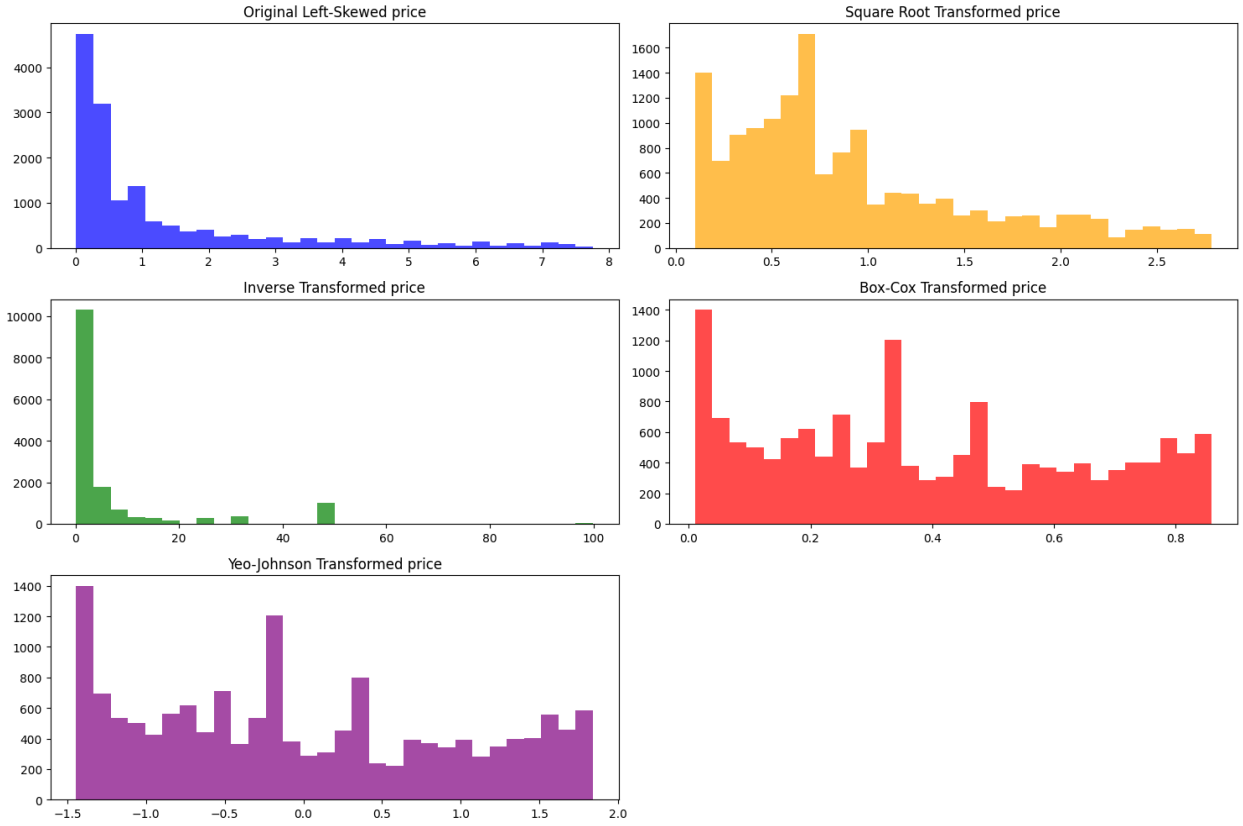
Data Transformation

Our initial analysis revealed the significant positive skew in the price data, with a skewness value of roughly 1.88 and 2.08 for our primary and secondary datasets as seen above. This skewed distribution poses challenges for many statistical techniques that assume normality. We chose to adjust the data distribution through transforms to enable a mode robust statistical analysis and modeling process.To enhance the accuracy and reliability of subsequent insights into card prices and their relationships with other variables, explored multiple transformation methods to normalize the data distribution.

Our results for the primary dataset are explored below and those for the secondary dataset can be found in Figure 4. We implemented four transformation methods: square root, inverse, box-cox, and Yeo-Johnson. Among the applied methods, the Box-Cox and Yeo-Johnson

Figure 4

Historgrams of each transformation performed. (a. Original Price feature, b. Square Root Price transformation, c. Inverse Price transformation, d. Box-Cox Price transformation, e. Yeo-Johnson price transformation)



transformations prove most effective, reducing the skewness from 1.88 to approximately 0.27. This substantial reduction brings the data distribution much closer to normality.

Histograms are provided in Figure 4 for each transformation, allowing for visual comparison of the distribution changes. These visualizations aid in identifying the most effective transformation method for the dataset.

We extended the transformation process to encompass additional left skewed variables such as edhrecRank and edhrecSaltiness, ensuring a comprehensive approach to data

normalization across the entire dataset. By applying the Box-Cox transformation for each variable, we aimed to enhance the predictive accuracy of our model.

However, it's important to acknowledge that this holistic approach introduces certain drawbacks. By adding a layer of complexity we can complicate the interpretation of model results as well as potentially obscure underlying relationships within the data. To mitigate these potential issues, we retained our non-transformed dataset for model evaluation, allowing for us to better understand the impacts of our transformations, which will be discussed in more detail with model results.

Model Selection

We needed to select a model for price prediction, by taking into account the characteristics of our dataset. To gain additional insights beyond our statistical analysis, we decided to evaluate the dataset through clustering. Clustering is a fundamental technique in data analysis where similar data points are grouped together. It facilitates deeper insights into the underlying structure of the dataset.

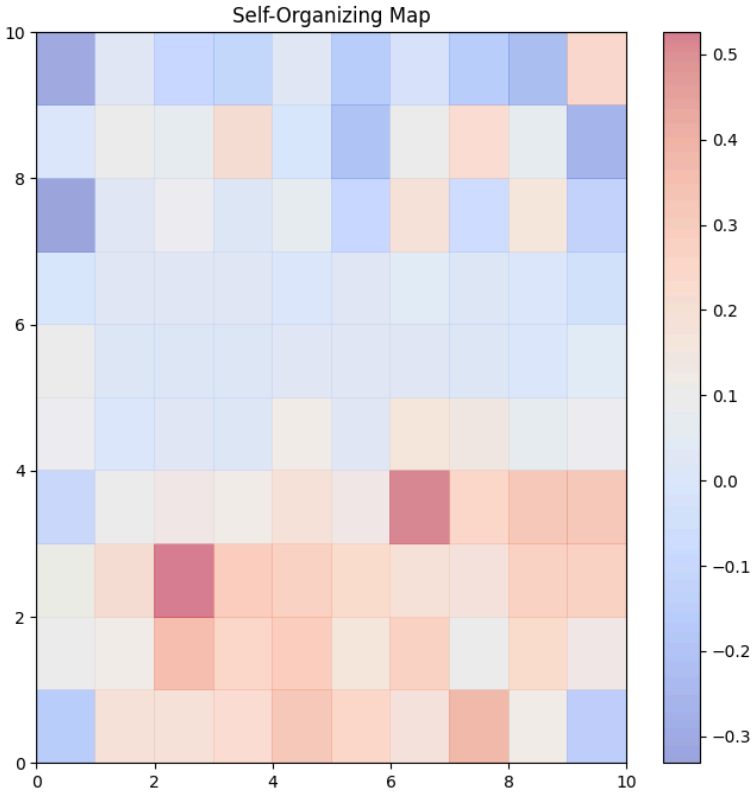
Self-Organizing Maps

In this section we will compare our prior manual insights and clustering analysis to determine our price prediction model. We chose Self-Organizing Maps (SOMs) as our clustering method based on a combination of research and our knowledge of the dataset. SOMs are particularly useful for visualizing high-dimensional data, allowing us to graphically evaluate the entire dataset in an interpretable way. By uncovering patterns and relationships, SOMs can significantly enhance our model predictions.

SOMs are a type of neural network that aid in interpretability through visualizations, often represented as heat maps. In these heat maps, each cell corresponds to a neuron in the SOM, with color–specifically red–indicating a high concentration of data points. Our overall SOM (see Figure 5) reveals two main clusters represented by the red squares on the bottom half, along with several less populated clusters scattered across the graph. This distribution suggests a complex interplay of multiple factors influencing the data.

Figure 5

Self-Organizing Map for the primary (non-transfomred datasets) that shows the clustering in the form of a heat map.



To further analyze these clusters, we overlapped the means of each data feature to illustrate the correlations and differences between neurons in our map. Our findings indicate that price is a significant driving factor within the SOM clusters, suggesting that we can effectively predict prices. These observations reinforce our previous statistical tests, such as p-values and statistical significance, strengthening our findings and allowing us to confidently move forward with the development of our price prediction model.

ANOVA Test

In addition, we aimed to explore one dataset feature in detail, leading us to implement a one-way ANOVA test to show how artists affect price. To do this, we had our null hypothesis (H0) state that artists have no impact on prices, and the alternative hypothesis (HA) posited that they do. The analysis yielded an F-statistic of 5.04 and a p-value of 0.000000 for the primary dataset. Given that the p-value is significantly less than the conventional alpha level of 0.05, we reject the null hypothesis. This result indicates a statistically significant difference in average prices among cards from different artists, suggesting that some artists produce cards that are valued more highly in the market.

We believe that there are important impacts to consider. The results suggest that the artist associated with a card does influence its price, meaning that some artists may produce cards that are valued more highly in the market than others. For collectors and sellers, understanding these differences can inform buying and pricing strategies. We moved on to understand if, beyond the effect of artists to card prices, we would be able to predict card values based on our data.

From our statistical analysis we determined that the relationships in the data might exhibit both linear and non-linear patterns. Therefore we wanted to implement both a random forest and some regression models to capture different aspects of the data and compare their performance. This dual approach allows us to leverage the interpretability of linear regression while also benefiting from the flexibility and robustness of random forests.

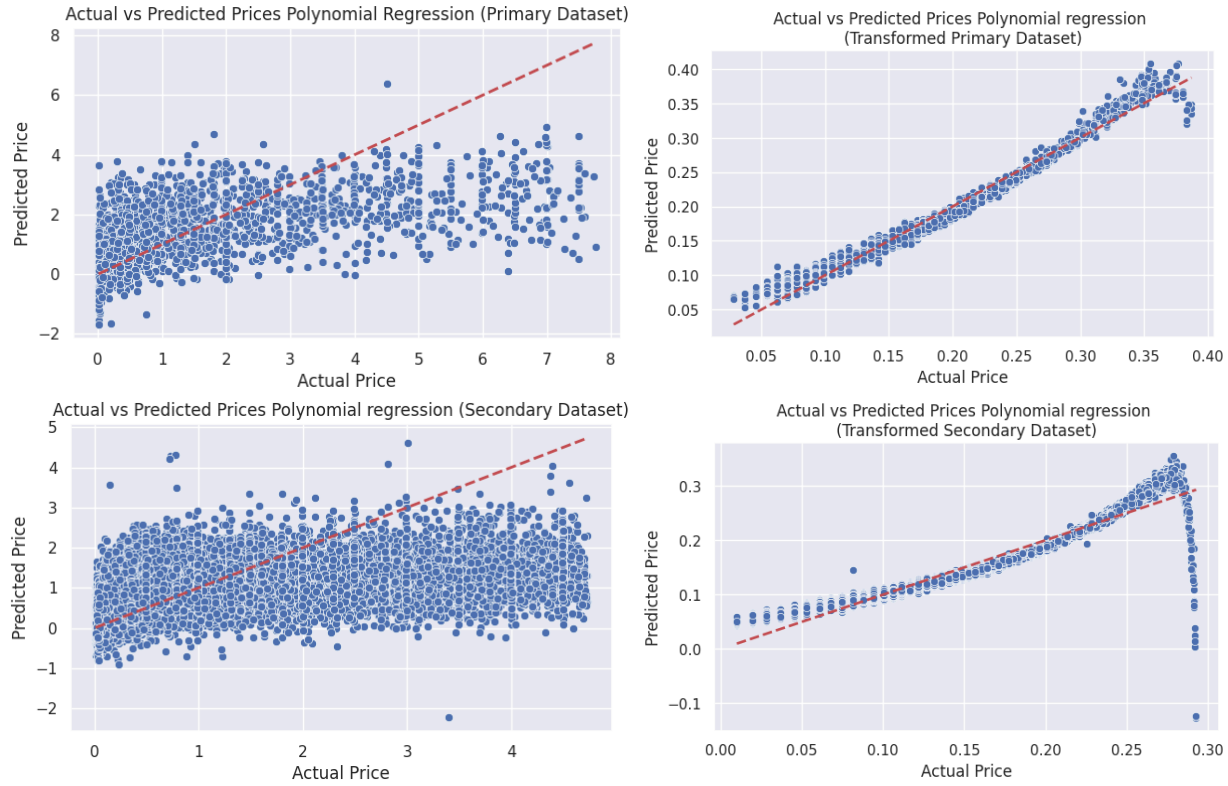
Model Analysis

Following the hypothesis testing, we used both Ordinary Least Squares (OLS) regression and polynomial linear regression to explore the relationships between card features and their market prices. The mapped datasets (both transformed and non-transformed) were divided into training and testing sets using an 80/20 split to validate model performance, and Polynomial features were generated to capture non-linear relationships between predictors and the target variable (price).

Polynomial Regression

Figure 6a-d

Polynomial regression models for both datasets as well as their transformed data.



The polynomial regression model model's performance was evaluated using two key metrics: Mean Squared Error (MSE) and R-squared (R^2) values. The primary dataset model achieved an MSE of approximately 1.91, while the transformed primary dataset achieved approximately 0.00015. This significant reduction in MSE indicates a substantial increase in model accuracy when using the transformed data for predicting actual prices. However on the primary dataset, the R-squared values show a different picture. The non-transformed data had an R-squared value of 0.351, while the transformed data had a lower R-squared value of 0.97. This suggests that the non-transformed dataset explains a greater proportion of the variance in the prices compared to the transformed dataset.

The secondary dataset (recorded in Table 1) showed similar patterns and similar accuracy compared to the respective transformed or non-transformed dataset. One notable difference can be seen in Figure 6d, where the scattered blue points drop drastically around $0.28. We think that the regression model may be overfitting and underfitting in certain regions (particually the higher price values). At the upper end, there is a noticeable deviation where the predictions overestimate the actual values for some points and underestimate other sharply, creating the non-smooth pattern. Ultimately, these models are constrained by our original dataset choices, the feature mapping performed, and the data cleaning steps we outlined, including the removal of outliers so they were not designed to perform well on outliers.

Ordinary Least Squares (OLS) Regression

The OLS model trained on the primary dataset achieved an R-squared value of approximately 0.498, indicating that about 49.8% of the variability in card prices can be explained by the independent variables included in the model. The adjusted R-squared was 0.497, suggesting that the model's explanatory power remains consistent even after accounting

for additional predictors. The F-statistic was 654.9, with a p-value of 0.00, indicating that at least one predictor variable significantly contributes to explaining price variability. This reinforces the overall validity of the model.

When analyzing the results of the individual variables, we can see a clear relationship with multiple attributes of the model, with a few variables being really important:

Significant Variables:

- EDHREC Rank (-4.789e-05): Shows a negative relationship with price, indicating that as a card's rank increases (becomes less popular), its price tends to decrease.
- EDHREC Saltiness (0.8681): Positively correlated with price, suggesting that cards perceived as more powerful or "salty" are valued higher.
- Mana Value (0.0621): Indicates that higher mana value slightly increases card prices.
- Rarity (-0.4133): A slight negative coefficient suggests that as rarity increases, prices tend to decrease, which might be counterintuitive as we would expect to see that rarer cards tend to be more valuable. For this interpretation, we think it is prudent to remind the audience that this is from one days worth of price data. Meaning on this specific day there could have been more market saturation or less demand for rare cards.
- Game Availability (1.389): Strong positive impact on price, indicating that cards available in more game formats are valued higher.

The OLS model trained on the transformed primary dataset achieved an R-squared value of approximately 0.975, indicating that about 97.5% of the variability in card prices can be explained by the independent variables included in the model. The F-statistic was 2.7e4, with a

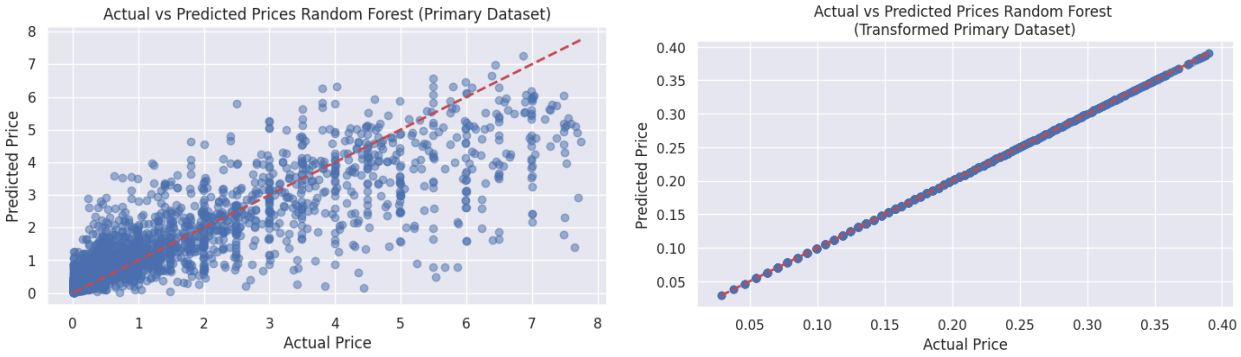
p-value of 0.00, indicating similarly to the previous dataset that at least one predictor variable significantly contributes to explaining price variability, reinforcing the overall validity of our second OLS model. For further numerical representations of the the specific variables and details of the secondary dataset results, consult Appendex 8.

Random Forest Regression

We moved forward to utilize the Random Forest Regression algorithm to understand if we can achieve a better prediction result using the same model definition. For this analysis we will focus in on the primary dataset as well as the transformed primary dataset.

Figure 7a and 7b

Random forest regression model for the primary and transformed primary datasets.



The random forest regression had the best performance on both the transformed and non-transfromed datasets, when analyzing the MSE and R-squared values. For the graphs in Figure 7a, the primary dataset, the MSE was 0.775 and R^2 was 0.741. The transformed dataset showed even better results at an MSE of 8.18e-10 and an R^2 of 0.999. This collaboratesn our previous model results, showing that the data transforms of the significantly left-skewed data results in almost perfect random forest predictions. This suggests that the transformation applied made it easier for the model to predict with higher accuracy.

Another useful feature of random forests is the feature importance values, which we show in Appendix 9. It is important to note that both the primary and secondary datasets had differing feature importance rankings, as well as differences between the transformed and non-transformed data variations. However, across the board, we observed that *edhrecRank* (a card's popularity in EDH/Commander), *edhrecSaltiness* (indicating community sentiment about a card's power level or fairness), and *manaCost* (especially in the secondary dataset) consistently had high feature importance values as strong predictors of price.

The importance of edhrecRank and edhrecSaltiness highlighted the significant impact of community perception and card popularity on pricing, and how this information is important to the perceived value of a card. The high importance of the artist feature, can also be connected to the set published, suggests that certain artists' works command higher prices, which could be valuable information for collectors. Game-related features (manaValue, gameAvailability) have importance, indicating that a card's utility in gameplay does influence its price, but perhaps not as much as its community perception values.

Model Comparisions

Table 1

Shows model Mean squared error and r-squared values for each regression model and each dataset both transformed and non-transformed.

Dataset	Polynomial Regression	OLS Regression	Random Forest Regression
Primary Dataset	MSE ~= 1.91	MSE ~= 2.24	MSE ~= 0.77
	R^2 ~= 0.35	R^2 ~= 0.50	R^2 ~= 0.74
Primary Dataset	MSE ~= 0.00015	MSE ~= 0.0011	MSE ~= 8.18e-10

(Transformed)	R^2 ~= 0.98	R^2 ~= 0.98	R^2 ~= 0.99
Secondary Dataset	MSE ~= 0.67	MSE ~= 0.79	MSE ~= 0.34
	R^2 ~= 0.33	R^2 ~= 0.49	R^2 ~= 0.66
Secondary Dataset (Transformed)	MSE ~= 0.00022	MSE ~= 0.00097	MSE ~= 1.44e-13
	R^2 ~= 0.94	R^2 ~= 0.96	R^2 ~= 0.99

Our radom forest regression consistently outperformed the other models, especially after data transformations, achieving near-perfect R^2 values and minimal MSE. The polynomial regression and OLS regression show similar notable improvements after data transformation but are slightly less accurate. Though our model experimentation process we showed that regression modeling assumes a normal distribution of errors, since our price data was heavily left-skewed the non-transformed models did not perform as well.

The OLS regression provided a stronger explanatory framework for understanding price variability, as evidenced by its higher R-squared value and more significant coefficients.
While polynomial regression allows for capturing nonlinear relationships, it did not outperform OLS, suggesting that the relationships between features and price may be simpler than expected, with a largely linear structure driving the variability in prices. The random forest's ability to capture complex interactions and non-linearities between features proved to be the most effective. The model's flexibility in handling both linear and non-linear relationships likely explains its superior performance, especially after the data transformations. To build on the success of the random forest regression, we should consider further data exploration of additional features not selecting in out datasets.

Conclusion and Recommendations

In this article, we set out to identify attributes influencing price predictions, focusing solely on card attributes and price information from a single market day. Our exploration revealed numerous correlations between price and card attributes, enabling us to develop a regression model capable of predicting prices with an impressive accuracy of approximately 99% for non-outlier values.

The insights gained from our analysis can significantly inform trending strategies for multiple card (on 09/20/2024), providing valuable insight for both buyer and sellers in determining optimal pricing strategies. To improve the fidelity and usability of our model we would first suggest evaluation performance on a new date and possible retraining a model using multiple price date days. We would also recommend for future works to evaluate the datasets without ehrecRank and ehrecSaltiness and/or try to predict these values since they have the strongest overall impact on card price when looking at the results across the board.

By carefully selecting our primary and secondary datasets, we effectively distinguished between creature price predictions and overall card price predictions, enhancing our models specificity for the right user. Moreover, our results highlighted considerable variance in average prices across different artists, suggesting that certain artists produce cards that are consistently more valuable. This knowledge along with other feature insights gained can empower stakeholders in the market to make informed decisions and adjust pricing accordingly, ultimately enhancing trading outcomes.

References

Agresti, A., & Kateri, M. (2021). *Foundations of Statistics for Data Scientists: With R and Python* (1st edition). Chapman and Hall/CRC.

Grand Prix (2024, October 3). MTG Wiki. https://mtg.fandom.com/wiki/Grand_Prix

Halpern, Z. (2024). *MTGJSON* (v5.2.2+20240920) [Source code]. GitHub. https://mtgjson.com/

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*(9), 1464–1480. https://doi.org/10.1109/5.58325

Penn State Eberly College of Science. (2018). Variance inflation factor (VIF). Penn State University. https://online.stat.psu.edu/stat462/node/180/

Schmidt, Gregory (February 16, 2023). "Magic: The Gathering Becomes a Billion-Dollar Brand for Toymaker Hasbro". The New York Times. Archived from the original on January 17, 2024. Retrieved January 17, 2024.

Appendix A

Figure A1

The five mana colors - Wizards of The Coast

Figure A2

Creature card Attributes - Wizards of The Coast

Figure A3

Correlation heatmap of the Primary dataset after feature mapping, but before dropping and transforming features.

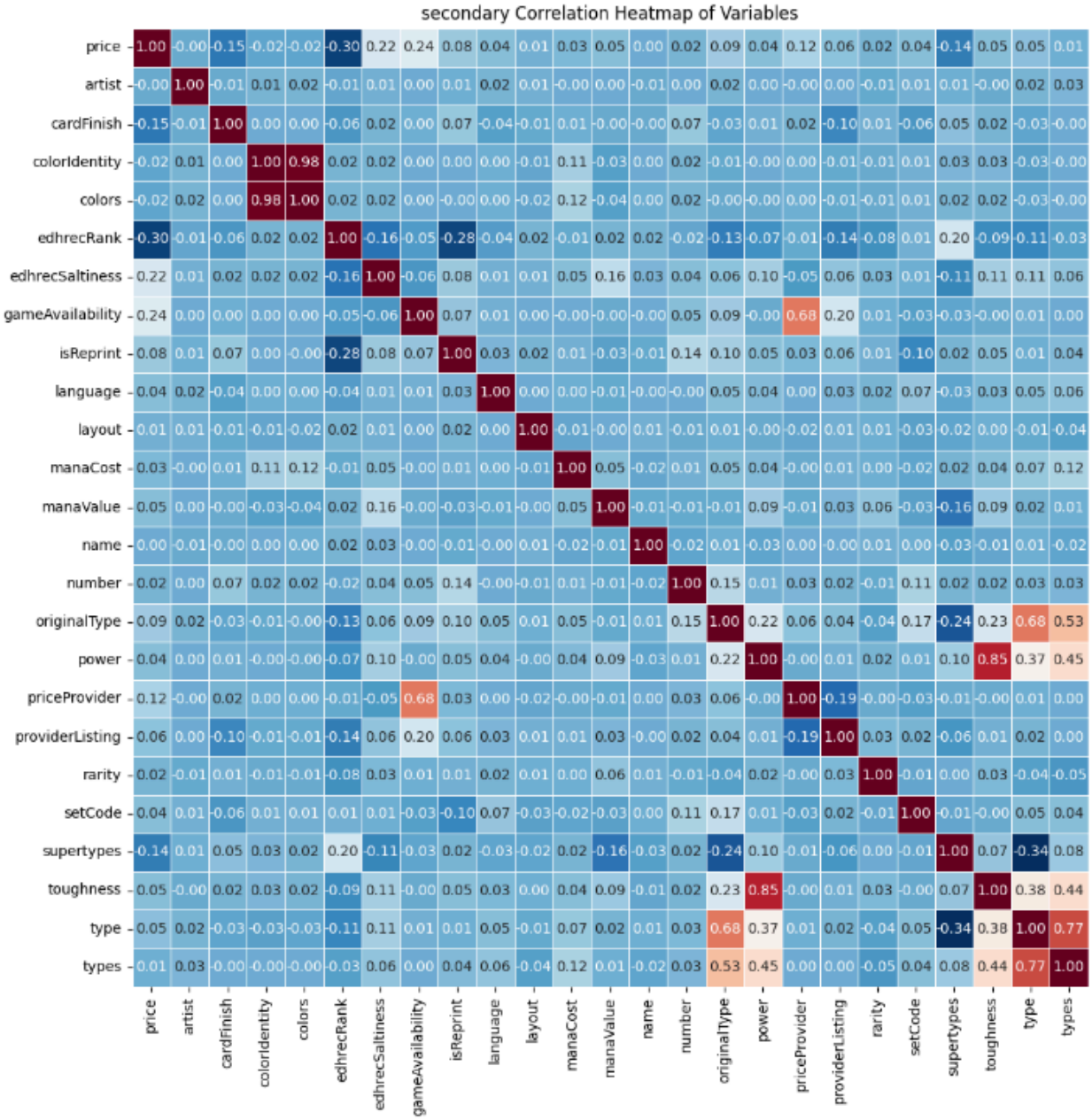


Figure A4

Primary dataset boxplots for price feature.

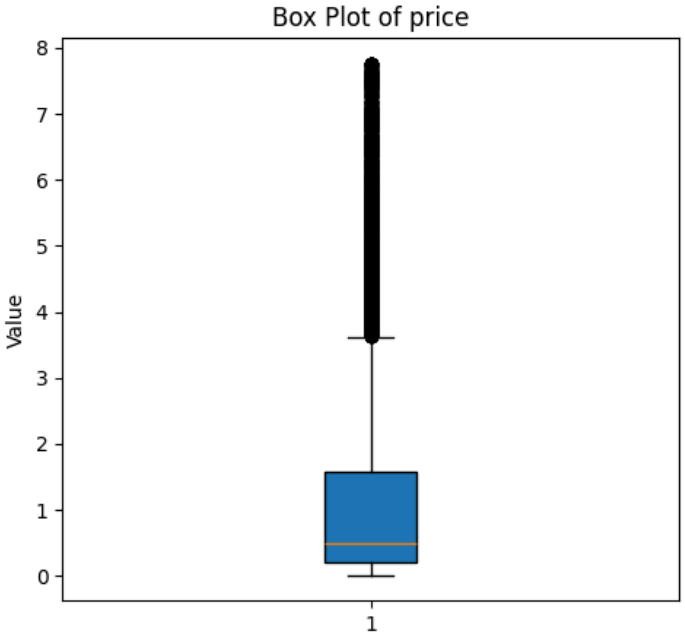


Figure A5

Histograms for each feature from the primary dataset.

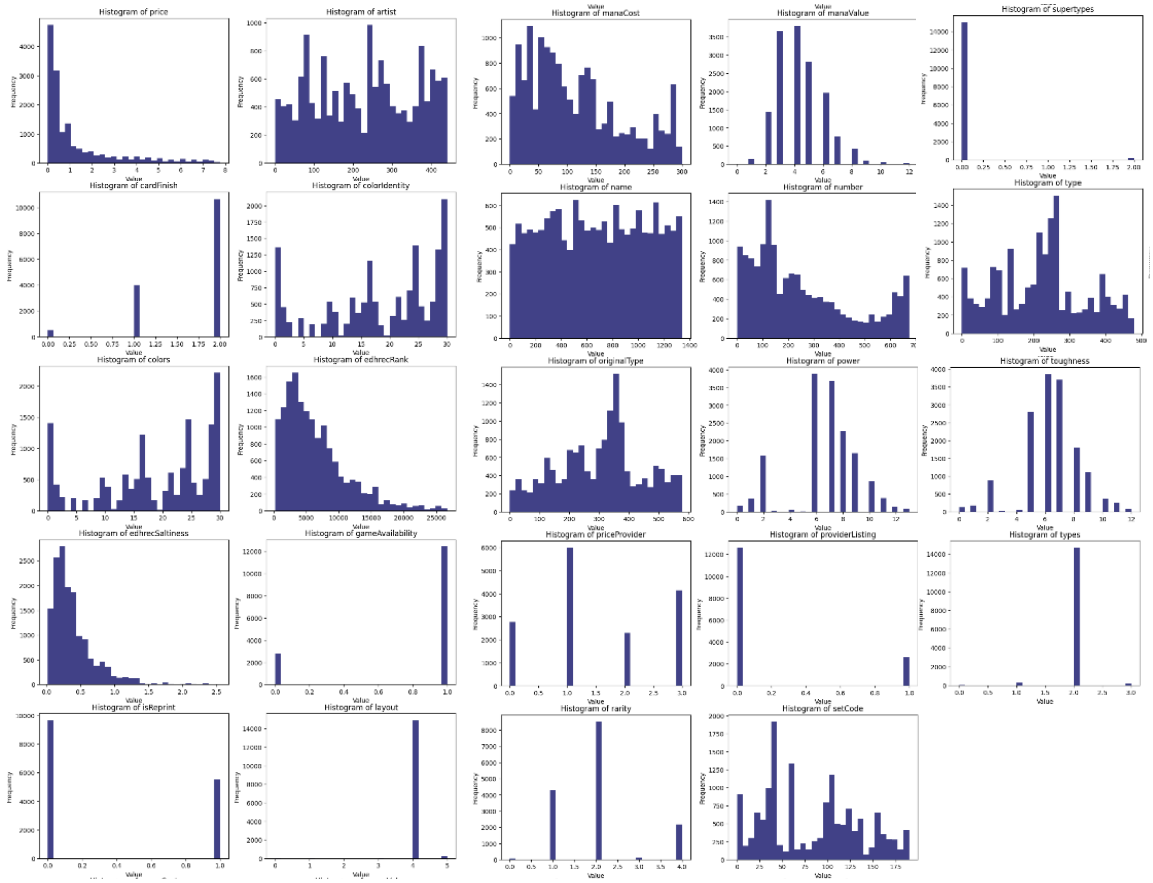


Figure A6

EdhrecRank feature data transformation visualizations.

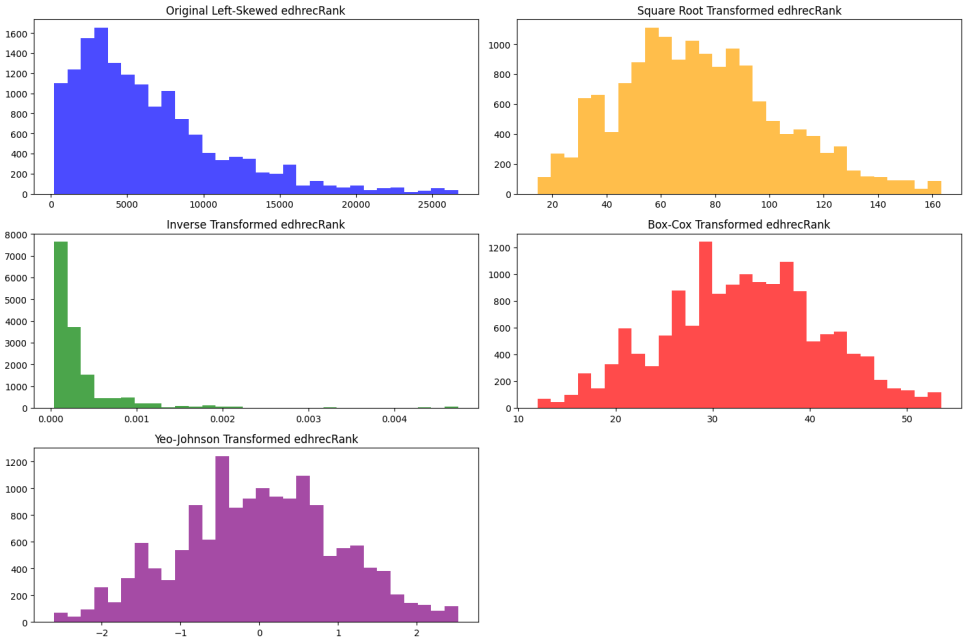


Figure A7

EdhrecSaltiness feature data transformation visualizations.

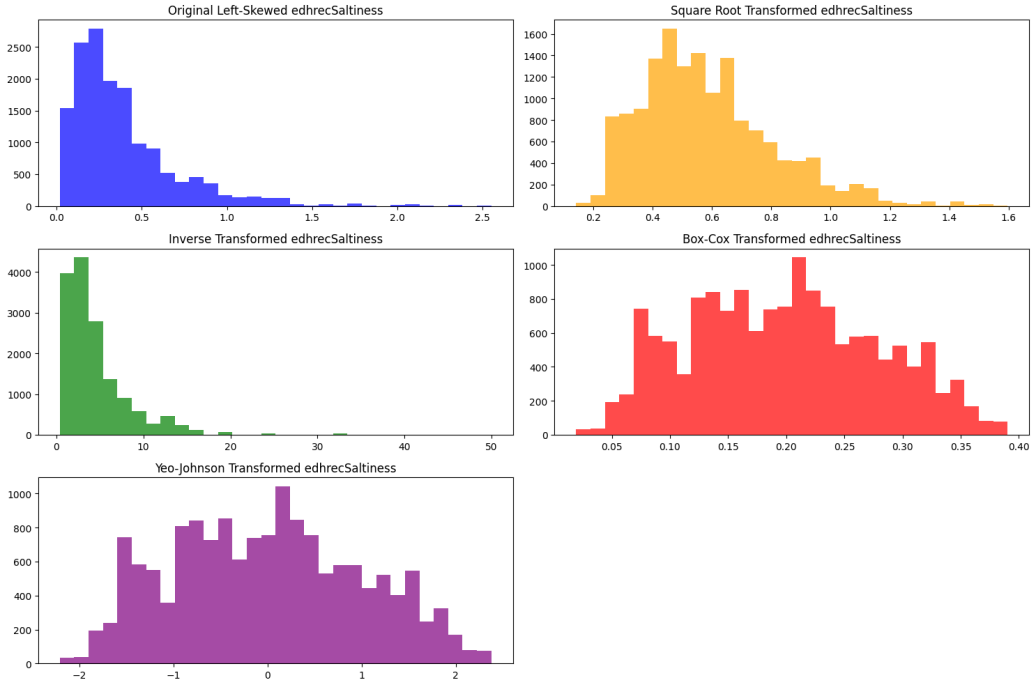


Figure A8a-d

OLS regression results, primary (a), primary transformed (b), secondary (c), secondary transformed (d).

OLS Regression Results

Dep. Variable:	price	R-squared (uncentered):	0.498
Model:	OLS	Adj. R-squared (uncentered):	0.497
Method:	Least Squares	F-statistic:	654.9
Date:	Tue, 22 Oct 2024	Prob (F-statistic):	0.00
Time:	03:17:47	Log-Likelihood:	-27745.
No. Observations:	15214	AIC:	5.554e+04
Df Residuals:	15191	BIC:	5.571e+04
Df Model:	23		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
artist	-0.0001	9.52e-05	-1.390	0.164	-0.000	5.43e-05
cardFinish	-0.3984	0.022	-17.710	0.000	-0.443	-0.354
colorIdentity	0.0150	0.006	2.652	0.008	0.004	0.026
colors	-0.0137	0.006	-2.406	0.016	-0.025	-0.003
edhrecRank	-4.789e-05	2.71e-06	-17.658	0.000	-5.32e-05	-4.26e-05
edhrecSaltiness	0.8681	0.043	20.391	0.000	0.785	0.952
gameAvailability	1.3893	0.049	28.612	0.000	1.294	1.485
isReprint	-0.1448	0.027	-5.333	0.000	-0.198	-0.092
layout	0.2547	0.029	8.863	0.000	0.198	0.311
manaCost	0.0003	0.000	1.657	0.098	-4.67e-05	0.001
manaValue	0.0621	0.010	6.258	0.000	0.043	0.082
name	-1.338e-05	3.18e-05	-0.420	0.674	-7.58e-05	4.9e-05
number	8.447e-05	6.24e-05	1.354	0.176	-3.78e-05	0.000
originalType	-0.0010	0.000	-6.922	0.000	-0.001	-0.001
power	-0.0038	0.007	-0.553	0.580	-0.017	0.010
priceProvider	-0.1241	0.017	-7.138	0.000	-0.158	-0.090

OLS Regression Results

Dep. Variable:	price	R-squared (uncentered):	0.975
Model:	OLS	Adj. R-squared (uncentered):	0.975
Method:	Least Squares	F-statistic:	2.700e+04
Date:	Tue, 22 Oct 2024	Prob (F-statistic):	0.00
Time:	03:17:47	Log-Likelihood:	29964.
No. Observations:	15214	AIC:	-5.988e+04
Df Residuals:	15192	BIC:	-5.972e+04
Df Model:	22		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
artist	-2.501e-07	2.15e-06	-0.117	0.907	-4.46e-06	3.96e-06
cardFinish	0.0012	0.001	2.277	0.023	0.000	0.002
colors	-0.0002	2.91e-05	-6.594	0.000	-0.000	-0.000
edhrecRank	-1.212e-06	6.1e-08	-19.867	0.000	-1.33e-06	-1.09e-06
edhrecSaltiness	0.2169	0.001	226.413	0.000	0.215	0.219
gameAvailability	0.0019	0.001	1.764	0.078	-0.000	0.004
isReprint	0.0062	0.001	10.103	0.000	0.005	0.007
layout	0.0204	0.001	31.549	0.000	0.019	0.022
manaCost	-8.661e-06	3.47e-06	-2.497	0.013	-1.55e-05	-1.86e-06
manaValue	0.0015	0.000	6.520	0.000	0.001	0.002
name	5.836e-06	7.17e-07	8.141	0.000	4.43e-06	7.24e-06

...

OLS Regression Results

Dep. Variable:	price	R-squared (uncentered):	0.491
Model:	OLS	Adj. R-squared (uncentered):	0.491
Method:	Least Squares	F-statistic:	9177.
Date:	Tue, 22 Oct 2024	Prob (F-statistic):	0.00
Time:	03:18:39	Log-Likelihood:	-2.9825e+05
No. Observations:	228637	AIC:	5.966e+05
Df Residuals:	228613	BIC:	5.968e+05
Df Model:	24		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
artist	-1.072e-05	4.62e-06	-2.320	0.020	-1.98e-05	-1.66e-06
cardFinish	-0.3282	0.004	-85.104	0.000	-0.336	-0.321
colorIdentity	-0.0017	0.001	-1.897	0.058	-0.004	5.76e-05
colors	-5.21e-06	0.001	-0.006	0.995	-0.002	0.002
edhrecRank	-3.306e-05	2.58e-07	-128.141	0.000	-3.36e-05	-3.26e-05
edhrecSaltiness	0.7293	0.007	98.973	0.000	0.715	0.744
gameAvailability	0.8596	0.007	118.211	0.000	0.845	0.874
isReprint	-0.0200	0.004	-4.985	0.000	-0.028	-0.012
language	0.2492	0.007	35.434	0.000	0.235	0.263
layout	0.0510	0.002	25.889	0.000	0.047	0.055
manaCost	0.0002	1.12e-05	15.915	0.000	0.000	0.000

...

OLS Regression Results

Dep. Variable:	price	R-squared (uncentered):	0.962
Model:	OLS	Adj. R-squared (uncentered):	0.962
Method:	Least Squares	F-statistic:	2.491e+05
Date:	Tue, 22 Oct 2024	Prob (F-statistic):	0.00
Time:	03:18:41	Log-Likelihood:	4.6828e+05
No. Observations:	228637	AIC:	-9.365e+05
Df Residuals:	228614	BIC:	-9.363e+05
Df Model:	23		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
artist	1.397e-06	1.62e-07	8.644	0.000	1.08e-06	1.71e-06
cardFinish	0.0029	0.000	21.579	0.000	0.003	0.003
colors	3.052e-05	6.22e-06	4.905	0.000	1.83e-05	4.27e-05
edhrecRank	-3.513e-07	9.03e-09	-38.921	0.000	-3.69e-07	-3.34e-07
edhrecSaltiness	0.2070	0.000	802.646	0.000	0.206	0.207
gameAvailability	0.0025	0.000	9.852	0.000	0.002	0.003
isReprint	0.0004	0.000	2.965	0.003	0.000	0.001
language	0.0143	0.000	58.228	0.000	0.014	0.015
layout	0.0043	6.89e-05	61.858	0.000	0.004	0.004
manaCost	8.304e-06	3.91e-07	21.236	0.000	7.54e-06	9.07e-06
manaValue	0.0013	3.98e-05	33.731	0.000	0.001	0.001
...						

Figure A9a and 9b

Random forest regression models for secondary and transformed secondary dataset.

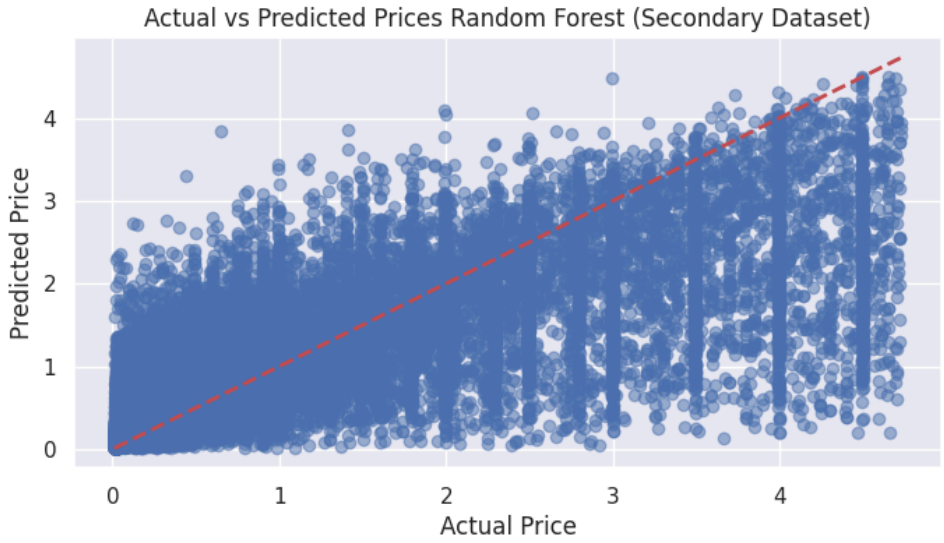
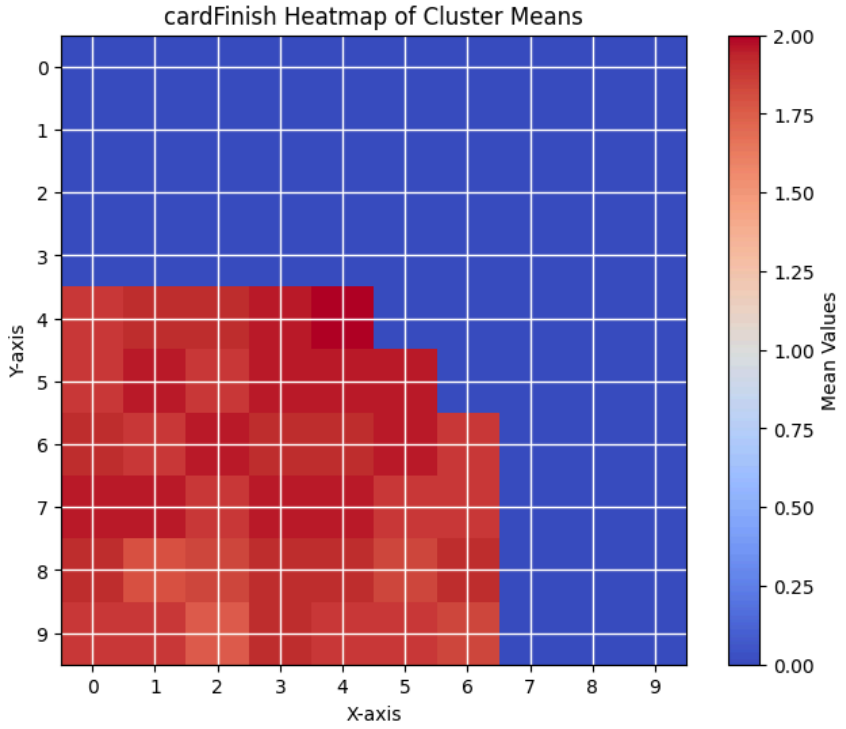
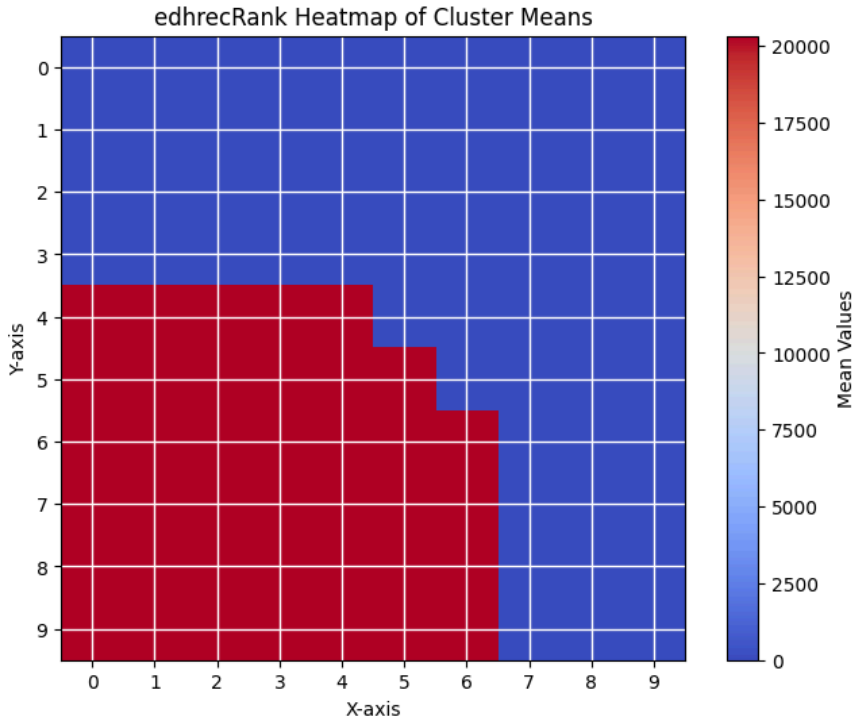
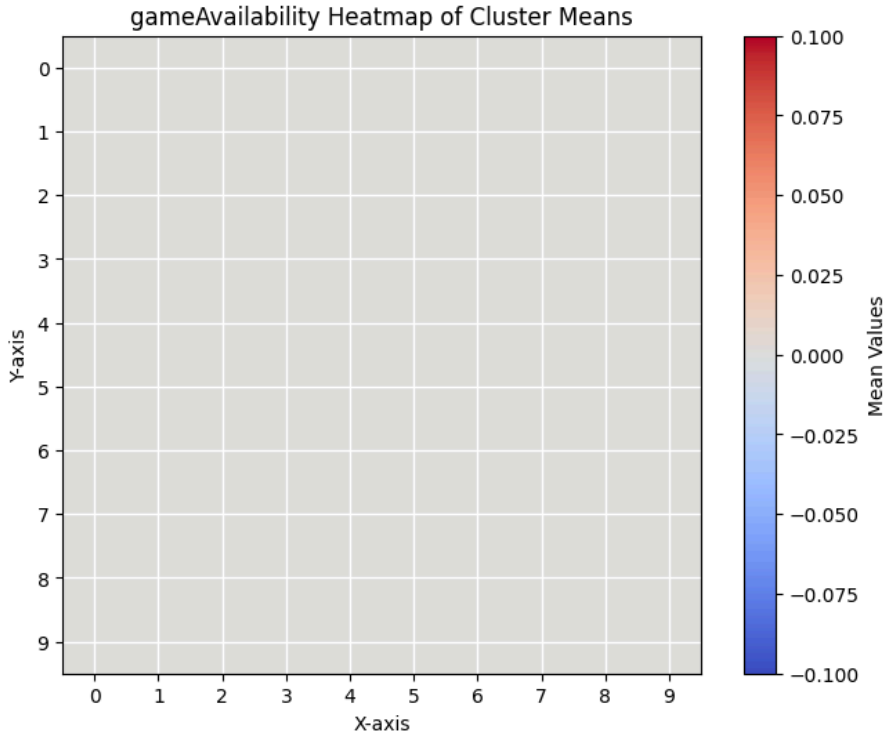


Figure A10a-d

Important SOM feature heatmaps. From an influential feature to a non influential feature.


cardFinish Heatmap of Cluster Means
Y-axis
X-axis
Mean Values
0
1
2
3
4
5
6
7
8
9
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00


edhrecRank Heatmap of Cluster Means
Y-axis
X-axis
Mean Values
0
1
2
3
4
5
6
7
8
9
20000
17500
15000
12500
10000
7500
5000
2500
0



Additional output can be found on our github repository

(https://github.com/victorhg/aai-500-final-project)