Regression on Current Average Selling Price of Pokémon Cards Using Their Characteristics

Kevin Kai Huang

May 7, 2024

Contents

1 Introduction

1.1 Background

Pokémon cards are collectible cards from the Pokémon Trading Card Game originating from the popular animated series "*Pokémon*" that features animal-like creatures (called Pokémon) with supernatural powers. Each card features either a character, item, or location that is reflected through the card's name and illustration. Each card belongs to a set (of cards) representing an overarching theme for that collection of cards. There are different types of Pokémon cards (e.x. Pokémon, Trainer, Energy), each with different functions. A key observation relevant to this study is that Pokémon cards with type "Pokémon" will always feature a Pokémon (character) and this character's name will always appear in the card's name. This observation led to the creation of a significant predictor variable in my study. Thus, this study will focus solely on Pokémon cards with type "Pokémon". For simplicity, assume for the rest of this report that any mention of a Pokémon card refers to a card with type "Pokémon".

Single Pokémon cards cannot be directly bought first-hand from the Pokémon Company, who monopolizes the production of these cards. Instead, Pokémon cards are introduced to consumers through card packs or boxes containing random cards. Each card has a specific (unfortunately unknown) probability of appearing in a pack of cards (related to the card's assigned rarity), introducing an element of luck to obtaining these cards. However, once these cards are opened from packs, they can be sold second-hand by owners/sellers (who decide the price) and bought by bidders (who agree to those prices). Thus, the Pokémon Company has no direct control over the prices of individual cards in this market, which makes it a very interesting market to explore in terms of card prices.

For a more in-depth description of Pokémon cards and their features, including an annotated example of a Pokémon card, please refer to my website's homepage.

1.2 Research Question

With this motivation in mind, it seems interesting to look into the selling prices of Pokémon cards in this second-hand market. However, there are a few things to first consider before defining the population of this study:

First, Pokémon cards each have their own rarity, including "common" and "uncommon" rarities. These cards appear more frequently than cards with other rarities. These common and uncommon cards are generally less desirable and most of them do not get purchased often on the market (or are purchased for very little). Thus, this study will focus on Pokémon cards excluding cards with rarity "common" and "uncommon". This report will refer to these types of cards as rare cards.

Second, due to data availability (discussed further in the next section), this study focuses on the selling prices of Pokémon cards in Euros. Thus, this study is focused on the European market for Pokémon cards in English but may be applicable to global markets for English Pokémon cards.

Finally, after exploring the current average selling prices for cards, it was found that the distribution of average selling prices for Pokémon cards is heavily right-skewed. To combat this, I took the logarithm of the current average selling prices. Finally, I scaled this variable by a multiplicative factor of 10 and used this scaled logarithmic current average selling price as my response variable.

Thus, the research question for this study is: “How do characteristics of a rare Pokémon card affect its current logarithmic average selling price in Euros?”, where a rare Pokémon card refers to Pokémon cards with any rarity excluding common and uncommon. The population that this study applies to is: all Pokémon cards excluding cards with rarity common or uncommon (including future cards to be released).

2 Methods

2.1 Data Acquisition

There are two APIs used to obtain data for this study:

- I used the Pokémon TCG API to retrieve all data on characteristics of both Pokémon cards and sets.
- I used the PokéAPI to retrieve data on Pokémon character names. This is used to help with the construction of one predictor variable (discussed later), but not as a predictor itself.

To expand on the previously mentioned issue on data availability, the Pokémon TCG API contained data on prices for both USD and Euros. However, the structure of the USD price data was inconsistent between different cards, leading to import errors. Thus, this study uses price data in Euros, provided by Cardmarket (and retrieved through the API).

Details on the process for requesting this data can be found in the source code or the instructions in the data folder, both found on the repository hosting this report.

2.2 Raw Data & Variables

First, I extracted just one variable from PokéAPI, the name of every Pokémon (character) that exists as of the date of this paper (May 7, 2024). This contained 1025 rows, representing the 1025 unique Pokémon characters. As a reminder, this variable “Character Name” is not a predictor in this study, but used for the construction of a predictor that is described in later sections. The remaining data was collected from the Pokémon TCG API.

The variables in the raw data collected from the card and set API (Pokémon TCG API) correspond to characteristics of Pokémon cards and sets that can be directly obtained from the cards and sets themselves. I extracted card ID, set ID, card name, current average selling price (in Euros), rarity, and artist of each card in the card dataset. From the set data, I then extracted set ID, total number of cards in the set, and release date of the set.

I merged the card and set data together (by set ID, which is not a predictor variable but only used for joining the data) to create a dataset of 7651 rows (cards) and 6 columns (variables/characteristics of cards). Note that there are more columns in the dataset (card ID and set ID), but these do not correspond to variables of interest (card ID is just used to uniquely identify cards in interactive plots from my website).

I then changed the type of the artist and rarity variables from string to factor in preparation for further data processing. Here is a summary of the raw variables (apart from changing type for artist and rarity) representing characteristics of each Pokémon card in the data:

Table 1: Summary of Variables of Interest in Raw Data

Variables	Type	Description
Average Selling Price (Euros)	Numeric	Related to response variable. Current average market selling price of a card in Euros. Prices set by sellers and purchases made by bidders.
Card Name	String	The card's name. Describes the primary character featured in the card
Rarity	Factor (31 levels)	Rarity of the card (assigned by Pokémon company)
Artist	Factor (278 levels)	Illustrator of card's artwork
Number of Cards in Set	Integer	Total number of cards in the card's set
Release Date	String	The date that the set which contains the card was released. Formatted as 'Year/Month/Day'

2.3 Data Errors

As previously mentioned, the inconsistent structure of price data in USD from the card API led to this study focusing on price data in Euros. However, current average selling price was not the only variable that led to import issues. Any observation that had missing values for a variable did not have a value (or even a cell/slot) for that variable. Thus, importing these observations led to severe import issues with the structure of the dataset. Thus, I had to deal with missing values on the API side by specifying queries to request for observations without any missing values for the variables in the data (instructions for this can be found in the data folder on this project's repository). After successfully loading the data in, I confirmed that there were no missing values for any of the variables.

Despite dealing with missing values through the API, there were still data errors regarding the current average selling price variable. Some observations (272 cards) had current average selling prices of 0 Euros. After checking recent selling prices of some of these cards on Cardmarket as well as other external sources (price tracking sites), I observed that these values were incorrect, and they indicated errors in the price data. This makes sense since some of these cards have high recent selling prices (and it does not make sense to have card sales online for 0 Euros). I removed these 272 observations from the data to be left with a dataset of 7379 rows and 6 columns.

As the price data is right-skewed, there were outliers on the greater side of the distribution. After taking the logarithm of the current average selling price, the distribution became less skewed and I observed very few outliers with larger prices. I decided to keep the outliers because I want to capture what makes those cards so expensive. Besides the previously mentioned data errors, there were no implausible values for any of the variables (Ex. number of cards in a set is always greater than 0).

2.4 Processing the Data

Most of the predictor variables in my study are processed variables, as most of the raw variables/characteristics do not directly provide enough information with respect to regression on current average selling price. Here is a step-by-step process I took to create my processed predictor variables:

First, I transformed the current average selling price variable into my desired response variable by taking its logarithm (to reduce the previously discussed skewness of its distribution) and scaling it by a multiplicative factor of 10 to help interpret data relative to its scale. During this process is where I identified and filtered out the observations with data errors regarding this variable. This scaled logarithmic current average selling price is my response variable.

Moving on to the predictor variables, I first created the variable "Days Old" which originates from the release date variable. This variable counts the number of days since the release date until the current date (May 7, 2024). I did this by converting the string variable, release date, to a Date variable, then calculated the difference in days and stored it as an integer.

Next, I addressed the artist variable, which had too many (278) levels. I collapsed this factor variable into 3 levels based on how many rare Pokémon cards each artist has illustrated. To do this, I first counted how many rare cards each artist illustrated. The cutoff for each level is chosen based on the quartiles of the distribution for the number of illustrations for each artist. I then assigned the values to each of the cards/observations in the dataset according to the card's illustrator. Here is a summary of the levels and cutoffs for this variable named Artist Frequency:

Table 2: Summary of Levels for Predictor Variable "Artist Frequency"

Level	Cutoff (Quartiles)	Cutoff (Number of Illustrations)	Number of Artists	Number of Cards
Infrequent	Less than median	$[1, 7)$	129	294
Frequent	Between median and 3rd quartile	$[7, 24)$	70	998
Abundant	Larger than 3rd quartile	$[2, 1082)$	69	6087

The next predictor variable I created originates from the rarity variable. Originally, I wanted to find a way to collapse or substitute the rarity variable as it has 31 levels that could make models too complicated and risk overfitting. I could not find a way to collapse the rarity variable and keep the relationship with the response variable. However, with the help of some feedback, I observed that some rarity groups had larger standard deviations in price than others, even with a larger number of cards in that rarity. Thus, I created a "log price standard deviation of rarities" variable to represent the standard deviation in logarithmic average selling price of cards for each rarity. Note that this variable is perfectly collinear with the rarity variable in thus study. Thus, **only one** of the predictors can be used in each model. This predictor variable serves as a substitute for rarity, and both predictors will be evaluated, with the **most significant predictor chosen for final models**.

2.5 Constructing the "Character Appearances" Predictor Variable

Moving on to the final processed variable, the "character appearances" variable has a much more complex construction relative to the other processed variables in this study. Additionally, the construction of this variable came with some consequences regarding a small loss of observations, warranting its own section for discussion.

The character appearances variable counts how many times each Pokémon character is featured in rare Pokémon cards. The motivation behind this variable is to attempt to capture how popular each Pokémon character is. This is because I hypothesize that more popular Pokémon characters would generate more demand for their cards compared to less popular characters. The justification behind this variable's relation to popularity is that the Pokémon Company features popular characters in each set to make the set more popular and generate more sales. Since any Pokémon character can have multiple cards, the number of rare cards that feature a Pokémon may be a good indicator as to how popular that Pokémon character is.

This variable is constructed based on the key observation that for every card with type Pokémon, the Pokémon character's name will always be in the name of that card. Thus, this variable is constructed through text mining on each Pokémon card's name. As described in my website's introduction to Pokémon cards, Pokémon card names may sometimes contain words describing the rarity they belong to (this does not apply to every rarity). Since rarity descriptions and words other than character names can appear in card names, I cannot apply regular text mining methods on card names without introducing some correlation between rarity and my new variable.

To solve this, I used the name data requested from the PokéAPI that contains the names of every single Pokémon character to this date. However, some of the names in the data had extra words describing things such as forms the Pokémon characters take (e.x. Eiscue-ice). Luckily, the number of character names that had this issue in the data was small, so I was able to manually add in the regular character names for those observations. Another small issue was that the names in the character name data is in lowercase, while the card names had capitalized first letters (this was easily fixed by setting all tokens/words in card names to lowercase). With these issues fixed, I was able to tokenize the card names, then filter the words to only include Pokémon names.

After extracting the Pokémon names in each card name, I counted how many times each character appeared in cards within my data (all rare cards). I stored this integer in a variable called "character appearances" which is my last predictor variable. Finally, I assigned the values for this variable to each card by merging according to the name of the character featured in the card. This was a technically difficult task, which is described in my source code.

This variable was successfully created, but not without some consequences regarding the data. Since each card name is tokenized by special characters (such as space, hyphen, dot) to differentiate character names from other words, Pokémon characters whose names contain a special character cannot be matched with cards through this method of variable construction. Examples of these Pokémon are "Mr. Mime" and "Ho-oh", who unfortunately have all Pokémon cards featuring them removed from the dataset since they do not have a value for the character appearances variable. Fortunately, the number of observations lost to this reason is relatively small at 200 observations lost. This reduces the total number of observations in my data from 7379 to 7179 cards, which is the final amount of observations after data cleaning and wrangling.

2.6 Summary of Response and Predictor Variables

Through the procedure of data processing described above, we have all our variables to consider in this study, summarized in Table 3 below:

Table 3: Variable Summary of Processed Data

Variables	Type	Description
Logarithmic Average Selling Price (Euros)	Numeric	Response Variable. Logarithmic current average selling price of a card in Euros. Log to combat skewness. Scaled by a Multiplicative Factor of 10
Number of Days Passed Since Release	Integer	The number of days from the current day (May 7, 2024) to card's release date
Rarity	Factor (31 levels)	Rarity of the card (assigned by Pokemon company)
Log Price Standard Deviation of Rarities	Numeric	Standard deviation of the distribution of logarithmic average selling price for the card's assigned rarity
Artist Frequency	Factor (3 levels)	How often this card's illustrator designs cards. Levels: Infrequent (1 to 6 cards), Frequent (7 to 24 cards), Abundant (25 to 1213 cards)
Number of Cards in Set	Integer	Total number of cards in the card's set
Character Appearances	Integer	The number of rare cards that feature the Pokemon character featured in the current card.

Comparing with Table 1, only two of the predictor variables are raw variables (number of cards in set and rarity), while all other variables are processed. After creating the variables, I checked each variable for data errors and missing values, for which I found none. After cleaning and wrangling the data, the final dataset this study uses has 7179 rows and 7 columns/variables of interest. Again, note that there are more columns in the dataset for variables that are not used as predictors, but for annotation purposes in the interactive plots on my website (such as card ID and artist). Thus, this study explores the relationship between the six predictors and the response variable, (scaled) logarithmic current average selling price.

2.7 Data Exploration Methods

After data wrangling and cleaning which required access to the full dataset (e.x. to find data errors), I randomly allocated 70% of my data into a training set and the other 30% into a testing set before performing exploratory data analysis. This is to ensure that I do not allow the testing dataset to affect my analysis during my data exploration through plots and my model analysis.

Throughout my exploratory data analysis, I incorporated multiple predictors into various plots to attempt to find any interesting interactions between variables with respect to the response variable. Here is a summary of tools I used in my data exploration:

- kable and kableExtra to create tables of summary statistics for some variables.
- ggplot2 to:
    - create a correlation matrix heatmap between numeric predictors to evaluate correlations.
    - create scatterplots with regression lines to evaluate numeric variables against the response variable.
    - create barplots to evaluate categorical variables against the response variable.
    - boxplots and histograms for distribution of response variable with respect to predictors.
- Plotly to create interactive plots (with the assistance of ggplot2) on my website's visualizations page

2.8 Modelling Methods

For my regression model analysis, I used the classical linear regression model along with machine learning models to regress (scaled) logarithmic current average selling price using my predictors. With the Linear Regression models, I will build models for both the rarity and rarity PSD predictors and compare the models (even after exploratory data analysis results). For all other models, however, I will select the predictor from the two that has a significant relationship with the response.

- **Multiple Linear Regression** with Backwards Variable Elimination using AIC (built-in lm function)
    - 4 Models
        - Model starting with all predictors excluding rarity PSD.
        - Model starting with all predictors excluding rarity.
        - Model starting with all predictors excluding rarity PSD. With Interactions between correlated variables identified through EDA
        - Model starting with all predictors excluding rarity. With Interactions between correlated variables identified through EDA
    - Variable Elimination performed with Akaike Information Criterion on training data
- **Regression Tree** (rpart)
    - Optimal complexity parameter chosen by cross validation error
- **Bagging** (random forest with mtry = number of predictors [6])
- **Random Forest** (random forest library)
- **Boosting** (gbm)
    - 1000 trees, 10-fold cross validation, 0.1 learning rate, Gaussian distribution for all models
    - *5 models*
        - Each model corresponds to an interaction depth from 1 to 5
- **Extreme Gradient Boosting** (XGBoost)
    - 10-fold grid search cross validation
        - grid search on max depths: (1,3,5,7), number of iterations: (50, 100, …, 500)
        - learning rate = 0.01

To compare between the multiple linear regression models, I will used the adjusted coefficient of determination R^2, since it is less naïve than the non-adjusted counterpart. However, for non-linear models such as extreme gradient boost, the coefficient of determination is a misleading measure. Therefore, I will use the root mean squared error (RMSE) measure to evaluate performance across all models.

After these models are fully trained, here are the steps I will take during my model analysis (with access to test data):

- Compare coefficients, individual t-tests, adjusted R^2 between linear regression models
- Compare RMSEs across the boosting models
- Report the cross validation results for the extreme gradient boosting model
- Analyze Variable Importance of relevant models
- Finally, evaluate model performance across all models using RMSE

2.9 A Note on Reproducibility

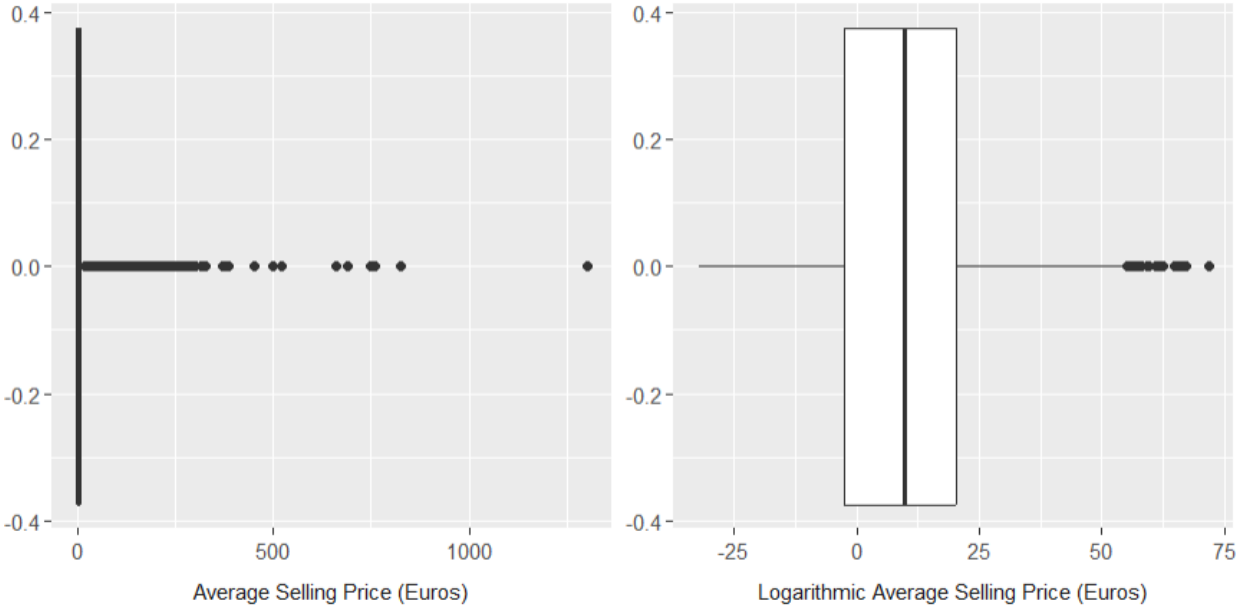
The price data from the Pokémon TCG API updates daily. Thus, it would be impossible to replicate this study when making requests from the API on a later date. To make this study reproducible, I have exported all the raw data I have collected from the API on the date of submission (May 7, 2024) and added a reproducible version of my source code in my repository that loads these datasets rather than requesting data from the API. Additionally, the "days old" predictor variable is constructed using the current date. In the reproducible version, this date is fixed to the date of submission: May 7, 2024. Finaly, all randomized code has a random seed set, ensuring the study 100% reproducible when loading the saved datasets found in the repository's data folder and using the reproducible source code.

3 Results

3.1 Exploratory data analysis

As previously mentioned, all plots are generated on training data so none of my analysis is affected by the testing data. First, I explore the effects of using the scaled logarithmic current average selling price of a card (in Euros) as my response variable (scaled to help with interpretability during model evaluation). The following Figure 1 shows the comparison between the raw variable's distribution and the processed response variable's distribution:

Figure 1:
Distribution Comparison Between Average Selling Price and
Scaled Logarithmic Average Selling Price of Pokemon Cards in Euros



As seen by Figure 1, the original variable, current average selling price of a card in Euros (left boxplot), is heavily right skewed with a median near 0. This is primarily because even among rare cards, most cards are not very desirable (they act as common cards among rare cards) which is reflected by their price. After taking the log of this variable (right boxplot), we observe that distribution of logarithmic average selling prices is much less skewed as seen by the median near the center of the box. As expected, the median price of the logarithmic average selling price is still near 0 and the range of prices is much smaller after taking the log (some of this is counteracted by scaling the variable). By logarithmic rules, observations with a logarithmic average selling price less than 0 have an average selling price less than 1 Euro. Thus, using the logarithmic average selling price of a card in Euros as

the response variable is justified as it successfully combats the skewness of the original price distribution and still reflects the original research question as previously discussed.

Next, the following Figure 2 illustrates Pearson correlation coefficients for all pairs of numerical predictors:

Figure 2: Pearson Correlation Matrix Heatmap for Numerical Predictors

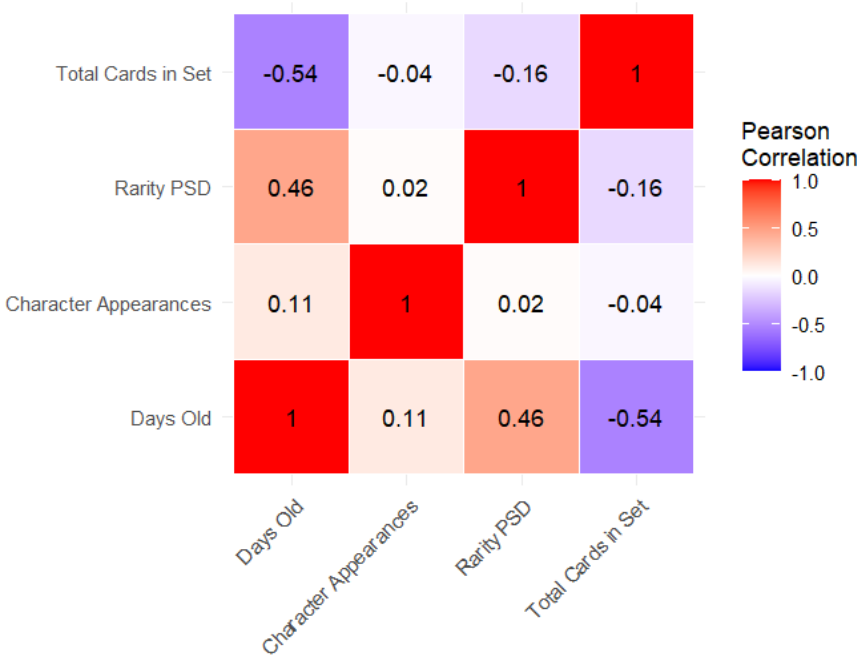


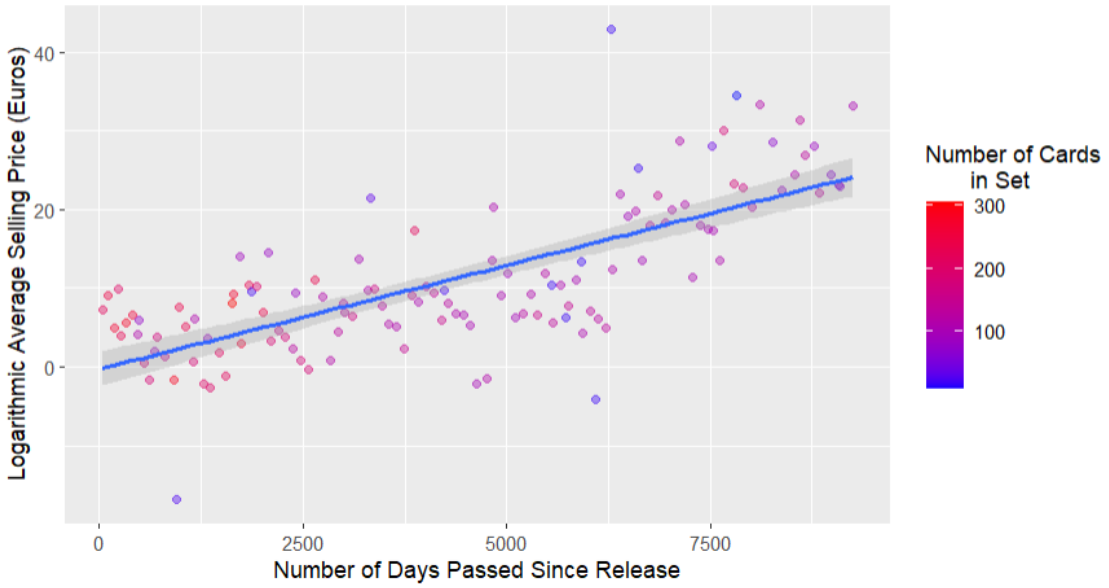
Figure 2 highlights two pairs of variables that appear to be somewhat correlated. These pairs of predictors are:

- Total cards in set and Days old. Pearson Coefficient $= -0.54$
- (Scaled) Log Price Standard Deviation of Rarity (Rarity PSD) and Days Old. Pearson Coefficient $= 0.46$

Thus, these pairs of variables will be considered for the linear regression model with interaction terms. It is not too surprising that the days old variable, which is related to time, is the predictor most correlated with others. I will explore this variable's correlation relationships further in later plots. It is surprising, however, that the character appearances variable demonstrates almost no correlation with other predictors. Its Pearson correlation coefficients with rarity PSD and total cards in set are both very close to 0. Thus, this variable could provide a lot of information on average selling price that would not be captured by the other numerical predictors (if it has a significant relationship with the response).

Next, I will begin to observe the relationships between the response and the predictors, starting with the days old predictor in Figure 3 below:

Figure 3:
Mean of Scaled Logarithmic Average Selling Prices (in Euros)
Against Number of Days Passed Since Release



In Figure 3, we can observe a positive relationship between the number of days passed since release (days old predictor) and the mean of the response variable (grouped by number of days passed). This positive relationship is reflected by the positively sloped regression curve. The points consistently follow this pattern, which is reflected by the relatively small standard error of the regression curve. Overall, the number of days passed since release appears to have a strong positive relationship with the mean of the response variable. Expanding on my findings from Figure 2, I included a gradient colour scheme for each point corresponding to the total number of cards in the card's set. Looking at the red points (high total number of cards in set), we can observe that most of these points are located in more recent times (closer to the left side). Thus, we can see that more modern sets contain more cards in general compared to older sets, which suggests some correlation (specifically, a negative relationship) between the two predictors. However, the correlation between the two predictors does not appear to be extreme as we can observe purple points (middle of gradient) scattered throughout the values of the days old variable. Thus, the Pearson correlation coefficient value between these predictors of -0.54 (from Figure 2) seems plausible as the variables demonstrate a limited amount of correlation.

Next, we can explore the relationship between the rarity variable and the response variable. This analysis can be observed in **Figure B**, found in the visualizations page on the website for this project. Looking at different lengths of bars in Figure B, we can observe that different rarities have different mean values for the response. The differences in lengths between some bars significant, indicating the effectiveness of rarity as a predictor for average selling price. Since the boxplot is ordered and coloured by number of cards in each rarity, we can try to observe the distribution of cards to each rarity. Unfortunately, we cannot see any clear pattern on how the number of cards affects average selling price, so this cannot act as a substitute for rarity (does not carry the same necessary information). However, we can make an interesting observation that the "Rare" rarity has 1489 cards (seen by hovering over the bar with mouse) and a very low (scaled) log average selling price compared to the other rarities. In fact, it is the only rarity with a negative log price, which means its average price is below 1 Euro. Thus, with its abundance of cards and low average selling price, we can induce that the "Rare" rarity generally acts as a the "common" rarity among rare cards. Overall, we have seen through Figure B that the rarity variable has effects on the response variable and that summary statistics of rarity have a hard time capturing the relationship between rarity and the response variable.

Overall, the pairs of variables appear to have negative relationships across states, which is intriguing since in the heatmap, they are positively correlated without grouping by state. This is more intuitive for the first two pairs of variables, since hospitals with a high proportion of vaccinated workers are likely more careful when handling potential sources of health hazards. Thus, states with these hospitals are likely to have fewer postoperative complications overall.

Expanding on results from Figure B, we can continue to explore how accurately summary statistics can capture the relationship between the rarity predictor and the response variable. **Figure C** (in my website) displays boxplots for distributions of the response variable for each rarity with over 25 observations. This series of boxplots (ordered in descending order of number of cards in rarity) allows us to investigate the standard deviation in logarithmic average selling price for each rarity. Unfortunately, there does not seem to be an identifiable relationship between the standard deviation (spread) of each rarity and its median logarithmic average selling price. On the other hand, we can observe that some rarities with larger amounts of observations have larger standard deviations. However, this is not universally true as we can see this pattern broken as travel down the boxplots. Additionally, we have already shown through Figure B that summary statistics (number of observations) of the of the logarithmic price distribution for each rarity do not capture rarity's relationship with the response variable well. Thus, the standard deviation of logarithmic average selling price for each rarity (Rarity PSD predictor) seems to lack a strong relationship with the response variable.

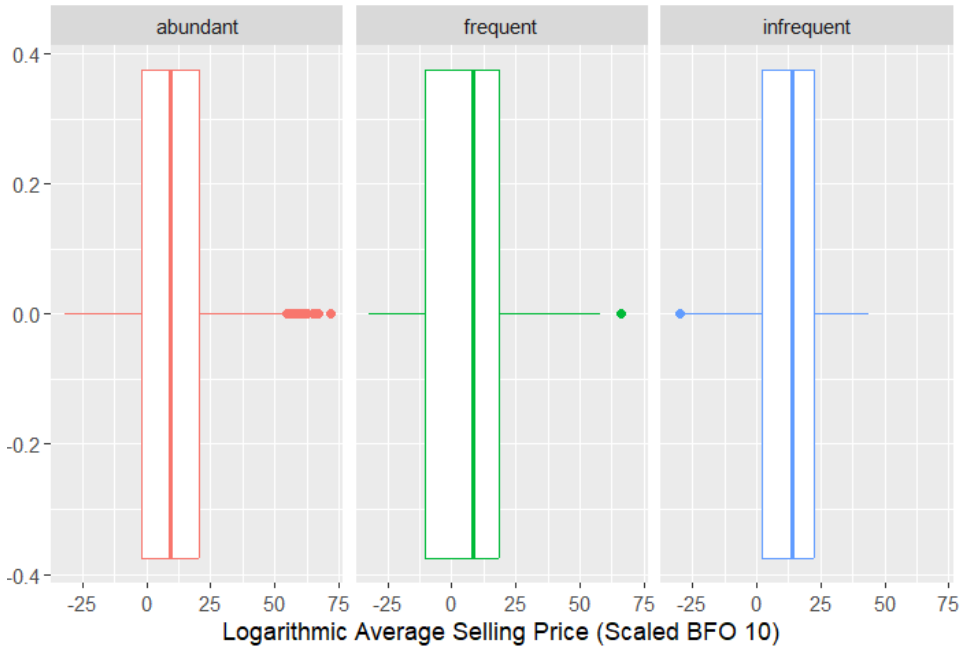
With the log price standard deviation of rarities predictor variable (rarity PSD) defined, we can officially explore its relationship with the response using Figure 4 below:

Figure 4: Mean Scaled Logarithmic Average Selling Prices for Values of Standard Deviation in Log Average Selling Price for Each Rarity (Rarity PSD)

Figure 4 illustrates the same findings identified in Figure C. First, we can see a regression line with a near horizontal slope and a large standard error, indicating a very weak relationship between rarity PSD and the response variable. Second, we can observe some correlation between rarity PSD and number of cards in rarity (as seen by the red points at larger values of rarity PSD), but we have now found that both of these summary statistics have failed to capture the relationship between the rarity predictor variable and the response variable and act as poor predictors.

Exploring the artist frequency predictor, the distributions of the response for each level are shown in Figure 5 below:

Figure 5: Distribution of Scaled Logarithmic Average Selling Prices of Pokémon cards for Each Level of the Artist Frequency Predictor Variable



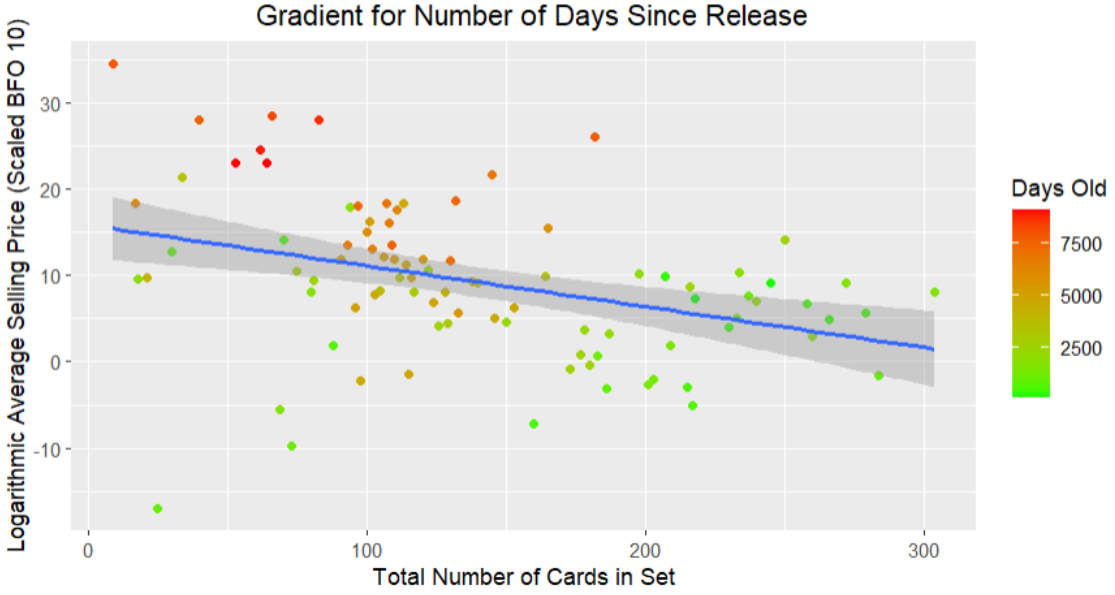
Looking at the boxplots for each level, there are distinct features that can be observed across each level:

- The abundant level and frequent level have similar median logarithmic average selling prices (< 12.5) while the median of the infrequent level is different (> 12.5)
- The abundant level and the infrequent level have similar interquartile ranges, while the frequent level has a larger interquartile range (illustrated by the width of the box)
- The distributions each have similar minimums, but the maximum logarithmic average selling price for the infrequent distribution is comparatively smaller than the maximums for the abundant and frequent levels
- The abundant level contains more outliers than the other levels

These distinct characteristics of each factor level may indicate that the artist frequency variable may have a somewhat significant relationship with the response. However, the magnitude of these differences is relatively low, so this relationship may not be very strong. Thus, the predictive capabilities of this variable will be further evaluated in the modelling analysis.

Next, Figure 6 shows the relationship between the predictor "total cards in set" with the response variable. Like Figure 3, I will use this plot to additionally explore the correlation between this predictor and the days old predictor:

Figure 6:
Coloured Scatterplot of Logarithmic Average Selling Price over Total Number of Cards in Set With Gradient for Number of Days Since Release



Looking at the regression line, we can see that there is a negative relationship between the total number of cards in a set and the response variable. However, some points in the scatterplot are relatively further away from the regression line. This is reflected in the larger standard error of the regression line, indicating this relationship may not be very strong (although I would still consider it significant). Looking at the colour of the points, we can more clearly see the correlation between this predictor and the days old predictor compared to through Figure 3. In Figure 6, we can see clusters of green, orange, and red points in the 2D plot space, showing the correlation between the two variables (first identified in Figure 2) and how it may affect the relationship between the total number of cards in a set and the response variable.

Finally, we can perform exploratory data analysis on the final predictor variable: the number of character appearances. We can first start by evaluating how well this variable captures popularity. Table 4 below provides a summary of the top 8 Pokémon characters (specifically their names) sorted by the character appearances variable, along with some summary statistics of their distributions for the response variable:

Table 3: Top 8 Pokemon Characters with Most Card Appearances Alongside Summary Statistics (rounded to 2 decimals) for Distributions of Scaled Log Average Price

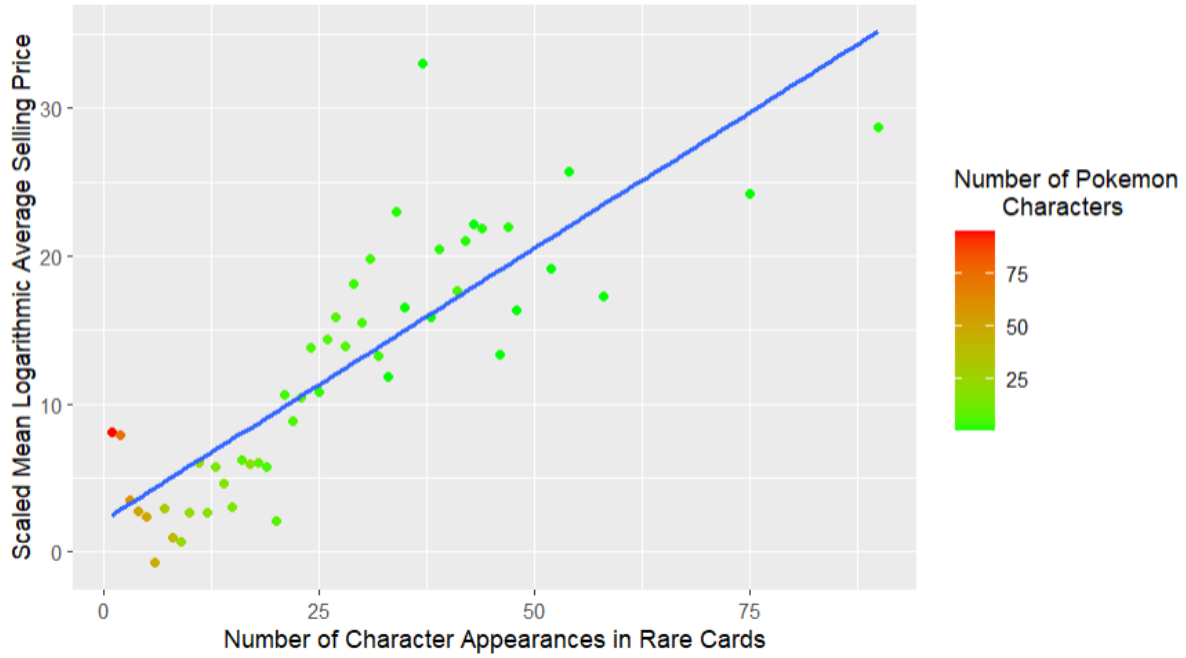
Card Name	Max Log Average Price	Mean Log Average Price	Min Log Average Price	Standard Deviation of Log Average Price	Character Appearances in Rare Cards
charizard	66.30	35.07	11.76	14.95	90
pikachu	59.46	23.02	-13.86	14.75	90
mewtwo	57.04	24.23	-16.61	15.33	75
raichu	50.29	17.28	-28.13	17.08	58
rayquaza	47.87	25.70	-3.42	11.95	54
gyarados	50.22	19.12	-22.07	17.83	52
unown	33.14	16.36	-9.42	10.42	48
gardevoir	50.48	15.37	-20.40	14.24	47

To judge how popular these characters are, I refer to an external source which is a survey conducted by the Pokémon Company in collaboration with Google during 2020. The results of this survey are available on Bulbagarden (they are no longer stored on the official site). Checking with this external source, 6 of the 8 Pokémon characters appear in the survey votes (Gyarados and Unown did not appear in the results). Furthermore, while some of the ordering matched ordering of the survey (e.x. Charizard above all other 7 characters), some of the ordering did not match (e.x. Rayquaza and Gardevoir were more popular than all characters in the table besides Charizard and Raichu was much less popular than characters below it in the table).

Thus, this shows my variable "character appearances" acts as a good approximation for popularity. However, it does not perfectly capture how popular each character is as there are other factors that contribute to popularity that would not be captured by the number of rare cards featuring the character. An example of this is that Pikachu is the mascot of the franchise, so the Pokémon Company would feature this character on more rare cards (along with its closely Pokémon character Raichu). One issue I hypothesized is that newly introduced characters may be popular, but the number of rare cards they would appear in would be limited compared to other characters who have appeared earlier. However, this hypothesis was contradicted by the Pearson correlation coefficient of 0.11 between the character appearances variable and the days old variable. While there is a positive relationship, which agrees with my earlier argument, the magnitude is low. I believe this is because after new characters are introduced, sets in the Pokémon Trading Card Game will focus on this new generation of characters for a while, which counteracts the effect from my proposed argument.

After analyzing how well the variable approximates popularity, we can now analyze its relationship with the response variable through Figure 7 below:

Figure 7: Mean of Scaled Logarithmic Average Selling Prices (in Euros) for Each Number of Pokemon Character Appearances in Rare Cards



The scatterplot in Figure 7 illustrates some remarkable findings about the significance of the predictor variable "character appearance". First, the regression line features the largest slope among all other numerical variables previously analyzed, including "days passed since release". Secondly, many of the points are plotted very close to the regression line and a positive linear pattern can clearly be seen in the points. Thus, the evidence in Figure 7 suggests that the character appearances variable could be the most significant predictor in this study.

In Figure 7, I coloured the points by the number of Pokémon characters with that number of rare card appearances. This helps visualize how only some characters truly stand out from others as popular, as seen through the green points with over 25 appearances, while most of the characters lie in the red points with less appearances in rare cards. This also helps to potentially explain the two red points with the smallest number of character appearances, who do not seem to follow the overall trend. I believe they do have a higher mean response value than the regression line because the number of characters within those points is high, leading to a large number of cards with a large variation in price caused by other factors. Additionally, cards featuring these characters may conversely be desirable, since there are very few rare cards that feature the characters.

To further analyze this variable, I created two interactive plots on my website's visualizations page. Figure D features similar data to Table 3, helping to visualize the distributions of logarithmic average selling price for each of the top 8 most popular Pokémon characters. In Figure D, we can observe that the median prices do not follow the same descending order of number of character appearances. Additionally, since the number of character appearances changes quite drastically throughout the top 8 Pokémon characters, we see the median scaled logarithmic average selling price also change by a relatively large amount across characters.

The second interactive plot is Figure E which plots the individual observations/cards for the top 8 Pokémon characters with most rare card appearances. In this plot, I added additional information on variables that were not used as predictors, such as the card's ID so one can easily identify the card that is being plotted. An example of this is to take the outlier in the Rayquaza boxplot in Figure D (with log average selling price 71.11). Matching this point in Figure D to the point with the same logarithmic average selling price in Figure E provides information on the card when hovering over it with the mouse. We can see this card has rarity "Rare Holo Star" which may explain why its price is so high compared to other Rayquaza cards. This further emphasizes the significance of rarity as a predictor variable, as no other variable can capture this relationship with the response.

3.2 Modelling Analysis

3.2.1 Linear Regression Models

Throughout all my linear regression models, each model minimized AIC with the full model including all predictors (except whichever rarity or rarity PSD is excluded from the model).

I will evaluate each of these models to see which predictor is more significant between rarity and rarity PSD (despite already obtained evidence from exploratory data analysis). Afterwards, I will provide model summaries for the models that include the more significant predictor.

Here is a table to summarize the evaluation of these final linear regression models, using adjusted R^2 as an evaluation metric:

Table 4: Evaluating Linear Regression Models After Variable Elimination

Model	**Final Predictors**	**Minimal AIC**	Adjusted R^2
Start with all variables, except rarity PSD. No interaction	Rarity, Days Old, Artist Frequency, Number of Cards in Set, Character Appearances	4895.76	0.754
Start with all variables except rarity. No interaction	Rarity PSD, Days Old, Artist Frequency, Number of Cards in Set, Character Appearances	4941.68	0.4619
Start with all variables except rarity PSD Interaction	Rarity, Days Old, Artist Frequency, Number of Cards in Set, Character Appearances, Days Old*Rarity, Number of Cards in Set*Days Old	4892.47	0.7722
Start with all variables except rarity. Interaction	Rarity PSD , Days Old, Artist Frequency, Number of Cards in Set, Character Appearances, Days Old*Rarity PSD, Number of Cards in Set*Days Old	4937.88	0.4678

Comparing the models with rarity against the models with rarity PSD, we can see a clear difference in performance that is reflect through both the AIC and the adjusted Pearson coefficient (seen more easily through the latter). This agrees with the evidence from the exploratory data analysis. Thus, for the rest of the modelling process, the rarity PSD variable will be dropped and the rarity variable with 31 levels will be used (as the summary statistics fail to capture the relationship between rarity and the response, despite having perfect collinearity with rarity).

We can also see that in general, the models with interaction terms (derived from Figure 2) perform slightly better, by a very small margin. It should be noted that although these adjusted Pearson coefficients seem promising, they could be misleading if model assumptions are violated (such as a non-linear relationship between the response and predictors). Thus it is better to consider the RMSE, which will be evaluated along with all other models at a later section. Here are model summaries of significant predictors in the two models including the rarity predictor:

Table 5:

Model Summary of Predictors Excluding Rarity_PSD in Linear Regression Without Interaction

Variable	**Coefficients**	**Standard Error**	**p-value**
(Intercept)	-3.776	03.545	0.287
Days Old	4.348×10^{-3}	6.967×10^{-5}	$< 2 \times 10^{-16}$
Character Appearances	0.2259	7.333×10^{-3}	$< 2 \times 10^{-16}$
Artist Frequency (Frequent)	1.582	-3.776	2.28×10^{-5}
Artist Frequency (Infrequent)	0.6598	0.6724	0.326556

Total Cards in Set	7.786×10^{-3}	2.404×10^{-3}	1.21×10^{-3}
Rarity (Rare)	-24.23	3.519	6.52×10^{-12}

Note that table 5 is missing information on many dummy variables for the remaining rarities. These contain both significant and insignificant variables (based off of t-test p-values). To avoid making this section messy with a very long table, I abstained from including all those variables. To see the rest of the rarity terms, please refer to the source code. The "rare" term for rarity is the most significant dummy variable from rarity. This is interesting because the "rare" rarity was the only rarity with a negative value for the mean response (recall Figure B from my website). This is reflected in the negative coefficient with a very large (relative) magnitude. The days old and character appearances variables are both very significant predictors in this model, with the smallest p-values out of all predictors. This accurately reflects my findings during EDA. We can also observe artist frequency and total cards in set having a significant effect (p-value), but both the p-values are not as small as the first two variables. Additionally, the scale for the artist frequency (factor) and total cards in set (integer) coefficients are relatively small (e.x. days old has a range 10 times longer than total cards in set, but their coefficients are similar).

Next, we can observe the same table, but for the linear regression model with interaction:

Table 6:

Model Summary of Predictors Excluding Rarity_PSD in Linear Regression With Interaction

Variable	Coefficients	Standard Error	p-value
(Intercept)	-161.2	161.1	0.1649
Days Old	0.1317	0.09241	0.1542
Character Appearances	0.2255	7.105×10^{-3}	$< 2 \times 10^{-16}$
Artist Frequency (Frequent)	1.444	0.3615	6.56×10^{-5}
Artist Frequency (Infrequent)	0.3748	0.6553	0.5673
Total Cards in Set	-1.582×10^{-2}	3.374×10^{-3}	2.83×10^{-6}
Rarity (Rare)	136.5	116.1	0.2397
Days Old & Total Cards Interaction	$7.904 \times 10\text{^} - 6$	1.067×10^{-6}	1.48×10^{-13}

Looking at Table 6, the linear regression model with interaction effects is completely different from the linear regression model without interaction. Aside from the coefficients (such as the negative intercept), the biggest difference is the significance of predictors. In the linear regression model without interaction, almost every predictor Is significant (with many significant dummy variables for rarity). In this interaction model, we only observe four significant predictors: character appearances (still with a very small p-value), artist frequency (frequent), total cards, and an interaction term between days old and total cards. This means that not a single rarity dummy variable is significant in this model, including any interactions with the rarity variable. Despite this huge change in model, it is reassuring to see that the character appearances variable still performs well across both linear regression models.

3.2.2. Boosting Models

As mentioned in the methods section, I built 5 boosting models where each model correspond to an interaction depth from 1 to 5 (the maximum number of predictors in the model). As previously discussed, this model and future models will exclude the rarity PSD predictor in favour of the rarity model. The following table describes each model's performance, evaluated using RMSE. For a sense of scale, the range of the scaled log average selling price is 127.513 Euros and the interquartile range is 23.526.

Table 7: Evaluating Boosting Models Excluding Predictor Rarity_PSD

Interaction Depth	Train RMSE (rounded to 3 decimals)	Test RMSE (rounded to 3 decimals)
1	7.499	7.951
2	6.429	7.501
3	6.245	7.387
4	6.093	7.383
5	5.984	7.396

Comparing the train MSEs, the boosting model with 5 interaction depth performs the best on the training data. However, when generalizing to the testing data, we observe a different pattern. All three models with interaction depths 3,4, and 5 have very similar test RMSEs. The model with 4 interaction depth has the lowest test RMSE. It is worth noting that the difference between the test RMSE and train RMSE for most boosting models may be an indicator of overfitting. For simplicity, I will report only the boosting model with 4 interaction depth when evaluating all models together.

3.2.3 Extreme Gradient Boost Cross Validation

As a reminder, I performed 10-fold cross validation on the extreme gradient boost model with a 0.4 learning parameter and a grid search on the following hyperparameters:

- Max Depths: {1,3,5,7}
- Number of iterations {50,100,150,…,500}

Through cross validation, the best tuning of hyperparameters was discovered to be:

- Max Depth = 3
- Number of Iterations = 150
- Learning Rate = 0.4
- Gamma = 0
- Colsample bytree = 0.6
- Min child weight = 1
- Subsample = 1

3.2.4 Regression Model Evaluation:

The following table summarizes model evaluations for all types of models constructed in this study. Excluded models from this table include some boosting and linear regression models that were not chosen as previously discussed.

Once again, please note that all models exclude the predictor variable "rarity PSD" in favor of the rarity predictor, after evidence suggests that rarity PSD and other summary statistics of rarity cannot capture its relationship with the

response variable. For reference, the range of the scaled log average selling price is 127.513 Euros and the interquartile range is 23.526.

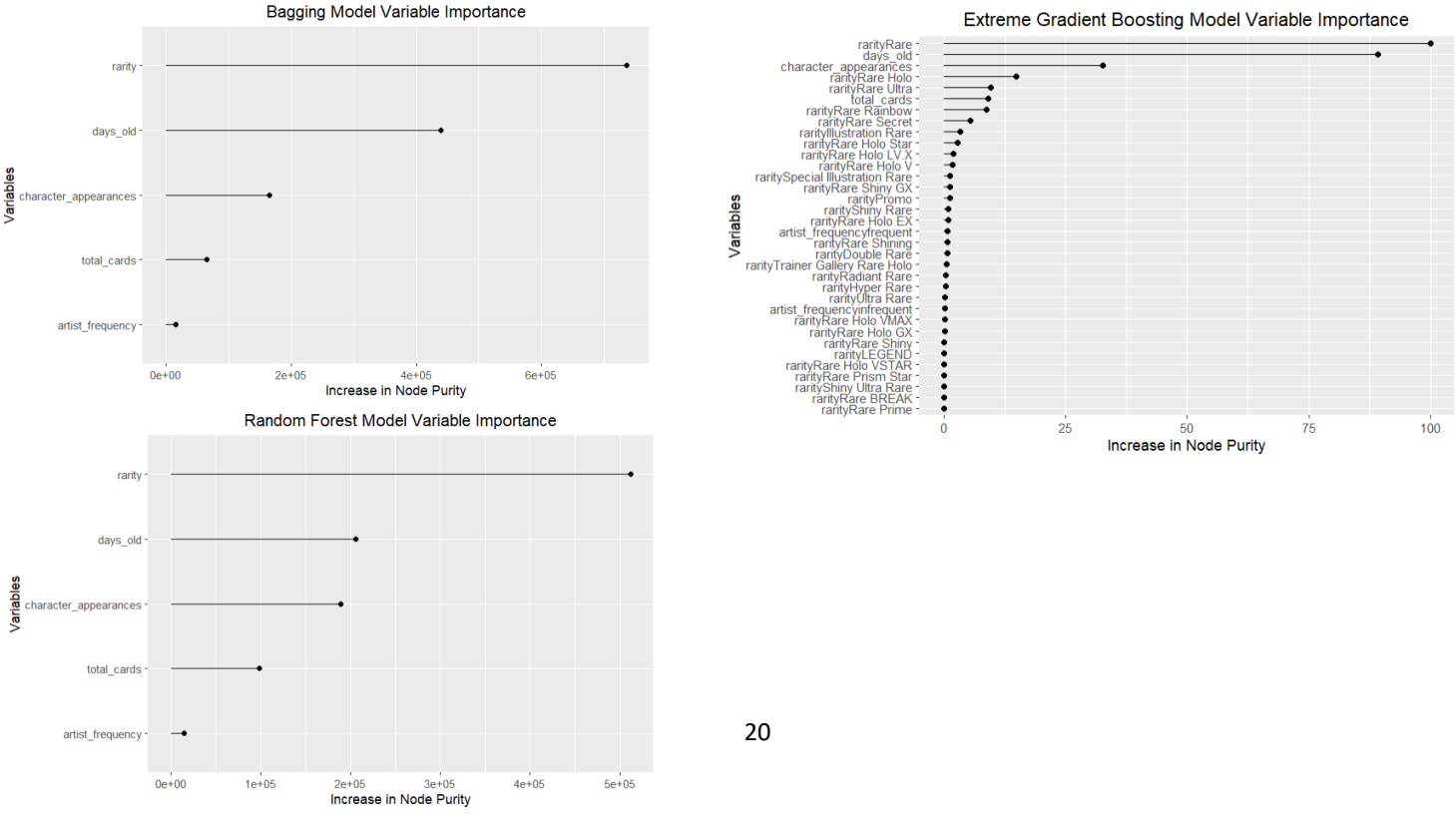
Table 8: Model Evaluation with RMSE (Sorted by Test RMSE) Using All Predictors Except Rarity PSD		
Model	**Train RMSE (rounded to 3 decimals)**	**Test RMSE (rounded to 3 decimals)**
Boosting (4 Interaction Depth)	6.093	7.383
Bagging	4.487	7.744
Extreme Gradient Boost	7.499	7.951
Linear Regression (Interaction)	8.209	8.490
Random Forest	7.600	8.821
Linear Regression (No Interaction)	8.555	8.827
Regression Tree	9.608	9.990

With the ordering provided in Table 8, it can easily be observed which models performed best on the testing data. However, some models demonstrated signs of overfitting (smaller training RMSE) such as Boosting, Bagging and Random Forest. Considering the differences between the two linear regression models, I find it surprising how they both have similar results when it comes to evaluation. Taking into account overfitting, I believe the model that is best for this regression problem is the extreme gradient boosting model, with the tuned hyperparameters previously described.

3.2.5 Variable Importance

Finally, we can evaluate the variable importance in the bagging, random forest, and extreme gradient boosting models. The variable importance plots can be seen in the figure below:

Figure 8: Variance Importance Plots For Various ML Models



As seen through the variable importance plots, rarity consistently appears to be the most significant variable in these machine learning regression models. Considering how rarity has 31 levels, it may be possible that rarity is the source of overfitting in the bagging and random forest models. However, we can also observe the extreme gradient model that has a rarity dummy variable as the most important variable (but closely followed by the days old predictor). The second most important variable, number of days since release, is also consistent across all models. However, how important this variable is depends on the model (as seen through its relative position to rarity). Next, the character appearances variable also appears to be important in each model, but not as much as the first two variables. Given the results from my exploratory data analysis, it is surprising to see character appearances with less importance (especially after observing its significance in linear regression models). The total cards variable also consistently shows up as somewhat significant while the artist frequency variable is only a little significant, both of these agree with the findings from my EDA.

4 Conclusion

4.1 Summary

In summary, this study helped identify some key variables to regress the (scaled) logarithmic current average selling price (in Euros) of rare Pokémon cards. Throughout my exploratory data analysis and modelling analysis, three predictor variables stood out to be significant predictors in most scenarios. These variables are:

- Rarity
    - Demonstrated a strong relationship with the response in EDA
    - Appeared to be the most significant predictor through model analysis and variable importance
    - Has 31 levels, concern for making models complex and overfit
- Number of Days Since Release
    - Showed a strong positive relationship with the response variable in EDA
    - Appeared to be slightly correlated with some predictors (rarity and total cards in set)
    - Consistently acted as a significant predictor throughout various models
- Character Appearances in Rare Cards
    - Displayed a very strong linear and positive relationship with the response variable in EDA
    - Showed almost no correlation with other numerical predictors
    - Was a significant predictor in many models (especially both linear regression models)

The total number of cards in a set variable and the artist frequency variable showed some correlation to the response variable, but it was not very strong and not consistently shown throughout various methods of this study. The exploration of the logarithmic price standard deviation of rarities variable (rarity PSD) was created to explore whether the rarity variable and its relationship with the response could be captured by a simpler variable. Unfortunately, it was discovered through my analysis that summary statistics of the rarity variable carried almost no information about the relationship between rarity and the response. Thus, this variable demonstrated poor potential as a predictor throughout this study and is an insignificant predictor for regressing average Pokémon card selling prices. This variable was also perfectly collinear with the rarity predictor, so this variable was excluded in favor of the significant predictor rarity for final models.

Most of the models displayed similar levels of performance on the test data. However, some models appeared to have issues with overfitting the training data (such as bagging), while other models showed no signs of overfitting (such as extreme gradient boosting). There were some surprising results, especially with the linear regression model

with interaction that had a completely different choice of significant predictors to its other linear regression counterpart. Overall, I would say that extreme gradient boosting is the best model for regressing current average selling prices of Pokémon cards with the predictor variables in this study.

4.2 Expanding on this study

As seen by how most of the predictor variables in this study are processed, useful characteristics of Pokémon cards as predictors are difficult to identify and collect as data. One variable I wanted to work with is the probability of a card appearing in a pack of cards. Unfortunately, this official probability is unknown. However, this variable can be approximated with empirical data. However, the issue with this is that such data may only be available for a limited subset of cards (especially considering how older sets of cards are no longer in production). Thus, considering this variable may severely affect the data available, potentially to the point where no meaningful analysis can be made.

Another alternative I considered is working with "graded" cards which are Pokémon cards sent to third parties to classify based off of their physical qualities. Higher graded cards have a higher selling price as well as an interesting variable representing the population of these graded cards (which is kept track of). Unfortunately, focusing on these variables would reduce this study's scope to graded cards which may have a much smaller data population to consider (same issue as above).

I believe there are confounding variables related to the yet unexplained variance in my response variable. As I am very interested in expanding this study, I am very open to any recommendations on variable selection or modelling methods.