Shark Tank: Analyzing Investment Decisions and Entrepreneurial Outcomes

Jorge Arturo Medina Garduno

06/12/2024

Abstract: This project analyzes investment decisions on Shark Tank, exploring whether business metrics or show dynamics drive outcomes. Using predictive models and clustering techniques, we found that factors like the ratio of amount asked to valuation and project category are the strongest predictors of success, while television-related variables have minimal influence. The findings suggest entrepreneurs should prioritize robust financial valuations and strategic category selection, as rejection on the show does not necessarily reflect business viability.

1 Introduction

Shark Tank is a popular TV series where, in each episode, entrepreneurs pitch their business ideas to a panel of five investors, or "sharks." These sharks evaluate the pitches and, based on their "valuation" of the business proposals, decide whether to invest resources in exchange for equity. The show has gained significant popularity in recent years, to the point where it is now being incorporated into some business school curricula as part of entrepreneurship and business strategy classes (Investopedia, 2024).

However, it's essential to remember that Shark Tank is ultimately a television program, with one of its primary distributors in the U.S. being ABC. Why is this worth noting? Because the valuation methods presented on the show often differ from traditional business valuation practices (Investopedia, n.d.). While the sharks sometimes rely on their instincts or "hunches" to make investment decisions—consistent with the show's entertainment-focused premises—this raises the question: does an innate bias exist to prioritize ratings over well-thought-out investments?

This bias becomes even more relevant for entrepreneurs whose pitches are rejected. Numerous examples exist of rejected pitches achieving remarkable success later. For instance, the founder of Ring, a company that introduced smart doorbells, received an immediate "no" from all five sharks in Season 5—despite his investment of $10,000 for his presentation. However, the company was later acquired by Amazon, generating revenues of approximately $577 million annually (Yahoo Finance, 2023). Such examples highlight not only the emotional toll on entrepreneurs but also the significant investment of time, money, and effort required to participate in the show.

This project aims to analyze the factors influencing the final investment decisions on Shark Tank and challenges the notion of unbiased, well-reasoned investments. Furthermore, it seeks to identify systematic approaches entrepreneurs can use to improve their pitches and increase their chances of success.

1.1 Initial Hypothesis

The primary focus of the sharks may not be a proper valuation of the projects. Instead, their decisions could be influenced by factors such as the season or episode they are situated in, suggesting that external variables might have greater predictive power than the sharks' presence and objective evaluation.

1.2 Objectives

1. **Develop a Tree-Based Predictive Model:** Build a predictive model to analyze the likelihood of a project securing a deal, as well as assess the importance of a variety of factors shaping the sharks' decisions.

2. **Build a Discriminant Analysis Model:** Examine investment patterns by developing a discriminant analysis model to predict the likelihood of investment based on the 10 main sharks and project categories.

3. **Perform K-Means Clustering:** Use clustering techniques to distinguish relevant characteristics between successful and unsuccessful projects, in order to provide actionable insights for entrepreneurs.

2 Data Description

The initial dataset consisted of 15 columns with 496 values (See Appendix Fig 1).

2.1 Data Inconsistencies

The dataset contained 72 missing values in the entrepreneur's column and 38 in the website column. The entrepreneur's column was dropped entirely due to its lack of predictive power, as individual names do not contribute meaningful information to the prediction task.

For the website variable, a binary column (*does_it_have_website*) was created to indicate whether a pitch had a website. However, after analyzing the distribution of this column, it was found that almost 90% of the entries indicated the presence of a website (38 "no" versus 457 "yes"), rendering the feature mostly constant. Consequently, this column was also dropped.

2.2 Testing Interactions Between Variables

Before feature engineering, the relationships between variables were explored using a correlation matrix. Most predictors exhibited low correlations with each other, which later posed a challenge as no numeric variable demonstrated a strong relationship with the target (*deal*). The highest observed correlation was between *askedFor* and *valuation* (0.76).

To address this, a Principal Component Analysis (PCA) was conducted to examine the relationship between these two variables. The first two principal components explained 54.88% of the variance in the data, and both *askedFor* and *valuation* were positioned closely along the first component. This analysis guided subsequent feature engineering to reduce dimensionality (See appendix Fig 2).

2.3 Description of Features

The dataset's features were examined using histograms and boxplots to understand their distributions and characteristics. Among the numeric features, both *valuation* and *askedFor* had noticeable outliers, with values exceeding three standard deviations from the mean. The episode data exhibited an imbalance, with fewer observations in later episodes within each season. The *category* variable was highly dispersed and imbalanced, with most categories accounting for roughly 1% of the dataset, while two categories—*specialty food* (12.53%) and *novelties* (7.07%)—were significantly overrepresented. Similarly, the *location* variable exhibited high cardinality, with 255 unique values. The most represented locations were Los Angeles, CA (8.28%), New York, NY (6.06%), and San Francisco, CA (5.05%) (See appendix Fig 3–Fig 18).

The *shark* variables demonstrated substantial overrepresentation for certain investors. For instance, Kevin O'Leary appeared in 76.5% of entries as *shark3*. Furthermore, several sharks appeared in different roles, such as *shark3* and *shark4*, reflecting repeated values that did not add new information. This redundancy highlighted the need for a more efficient representation of the *shark* data.

2.4 Feature Engineering

To enhance predictive power and reduce dimensionality, various transformations and new features were created. Although natural language processing (NLP) was not applied to the project descriptions, a new variable, *description_length*, was introduced to capture the complexity of the projects by counting the number of characters in each description. This variable, however, also contained two outliers, which were subsequently removed.

For the *askedFor* and *valuation* variables, a new feature, *ratio_asked_valuation*, was created by taking the ratio of the two variables. This transformation addressed the multicollinearity issue while retaining the interaction between them. Upon further inspection, it was discovered that *exchangeForStake* represented precisely this ratio multiplied by 100. Leveraging insights from PCA, the following approach was adopted: the *askedFor* variable was dropped, as it was highly correlated with the primary movement in PC1, while *valuation*, which contributed slightly more to the principal component, was retained. The original *exchangeForStake* was preserved for use in models like decision trees that do not require standardization, while the ratio (*ratio_asked_valuation*) was used for models that benefit from normalized variables.

The *location* variable was initially simplified to state-level information but remained imbalanced, with California having 142 appearances compared to 353 for all other states combined. To address this, states were aggregated into broader geographic regions (North, South, East, and West). Further balancing was achieved by creating artificial groups to distribute California's dominance across multiple categories. However, a new column *CA* was also included to indicate whether the state was California or not. After testing these two variables, the one that produced the best results was *CA*, so the regions were not used.

Shark variables were transformed by consolidating underrepresented sharks (less than 5% occurrence) into an "Others" category. Subsequently, a new column, *investor_combination*, was created to capture unique combinations of sharks appearing in pitches. Rare combinations were also grouped under "Others" to simplify the data. Interestingly, the "Others" category was dominated by combinations involving the main sharks (e.g., Barbara Corcoran, Robert Herjavec, Kevin O'Leary, and Daymond John) and a guest investor.

The *category* variable was restructured into broader groups based on thematic similarities, reducing the original number of categories from 55 to 11. Underrepresented categories were merged into an "Other" category to improve class balance. The *season* variable was grouped into early seasons (Seasons 1–3) and later seasons (Seasons 4–6) to reduce dimensionality. A new variable, *season_part*, was introduced to divide episodes into early, mid, and late parts of each season, capturing the show's chronological progression.

Outliers in numeric features, including *exchangeForStake*, *description_length*, and *ratio_asked_valuation*, were identified and removed based on a three-standard-deviation threshold. This process affected 2.6% of the dataset, a negligible loss of data that resulted in more normalized distributions.

The continuous variables *exchangeForStake*, *description_length*, and *ratio_difference_asked_valuation* were then standardized to ensure compatibility with models sensitive to scale, such as logistic regression and Linear Discriminant Analysis (LDA). (See appendix Fig 19–Fig 21.)

2.5 Removal of predictors

Several variables were removed due to redundancy or lack of predictive power. The *entrepreneur's* variable was dropped as it offered no meaningful contribution to prediction. Similarly, the *title* variable was excluded for the same reason. After feature engineering, other variables were removed, including *episode.season*, which was replaced by more granular variables (*season* and *episode*), and *description*, which was replaced by *description_length*. The individual *shark* columns were replaced with the *investor_combination* variable. The *askedFor* and *valuation* variables were dropped in favor of the newly created *ratio_difference_asked_valuation*, which effectively addressed their redundancy. After this step, the data is ready for the model (See appendix Fig 22–Fig 28).

3 Model Selection

To analyze the factors influencing whether a deal is made, we explored three models: logistic regression, boosted trees, and Quadratic Discriminant Analysis (QDA). While predictive performance was important,

the primary focus was on understanding the contribution of each predictor. Accuracy was chosen as the evaluation metric, ensuring that both "deal" and "no deal" instances were classified correctly.

Given the imbalanced distribution of categorical variables and the relatively small dataset, cross-validation with 10 folds was employed instead of a simple train-test split. This approach provided a more reliable estimate of model performance by using the entire dataset across multiple splits for training and validation. All models were initially trained using the following predictors: *category*, *description_length*, *season_part*, *ratio_difference*, *multiple_entrepreneurs*, *investor_combination*, *season_grouped*, *CA* (California), and *valuation*. The baseline categories for categorical variables were set as follows: the "Baby and Child" category for predictors of type, "early season" for season, "Barbara Corcoran, Lori Greiner, Robert Herjavec, Kevin O'Leary, Mark Cuban" for *investor_combination*, and "first seasons" for the *season_grouped* variable.

3.1 Logistic Regression

Logistic regression was initially applied with all predictors included. To refine the model, predictors with p-values greater than 0.10 were excluded, as these did not contribute significantly to the prediction. For categorical variables, a significant p-value indicated that a category behaved differently from the baseline. *Description_length* emerged as a significant contributor to the model, while the *ratio_difference*, representing the gap between the amount asked and the valuation, proved to be the most important variable. *Season_grouped*, *multiple_entrepreneurs*, *investor_combination*, and *CA* were not significant contributors. *Valuation* was found to be significant.

After removing non-significant predictors and re-running the regression, the model's accuracy improved from 0.58 to 0.60.

Logistic Regression Results	**Estimate**	**Std. Error**	**Pr**
(Intercept)	0.891	0.427	0.0370 *
Category - Entertainment and Recreation	0.351	0.376	0.3448
Category - Fashion and Apparel	-0.862	0.242	0.0003 **
Category - Food and Beverages	-0.234	0.345	0.5048
Category - Home and Living	0.193	0.402	0.6464
Category - Other	-0.241	0.376	0.5194
Category - Personal Care	0.702	0.419	0.0916
Category - Professional Services	-0.827	0.415	0.0476 *
Description Length	0.026	0.012	0.0286 *
Ratio Difference Asked	3.544	1.082	0.0014 **
Valuation	-8.133e-08	3.498e-08	0.0202 *
Observations	469		
Log Likelihood	-311.159		
Akaike Inf. Crit.	644.319		

Table 1: Coefficients in Logistic Regression

3.2 Boosted Trees with Cross-Validation

Boosted trees were chosen over other tree-based models due to their ability to iteratively learn from errors, improving robustness and predictive power. In contrast, traditional tree-based models showed strong performance on the training set but failed to generalize well during validation. The boosted model employed a Bernoulli distribution and was tuned for optimal performance using between 100 and 150 trees. This model achieved the highest accuracy among the three, with a score of 0.70.

Feature importance analysis revealed that *category* was the most influential predictor, followed by *ratio_difference*, *investor_combination*, *CA*, and *season_grouped*.

Feature Importance for Boosted Model	Relative Importance
Description length	30.791
Category	26.799
Valuation	14.272
Exchange for stake	12.187
Season part	6.774
Investor combination	5.800
Multiple entrepreneurs	2.516
CA	0.599
Season grouped	0.262

Table 2: Feature importance for boosted model

3.3 Quadratic Discriminant Analysis (QDA)

QDA was applied to account for non-constant variance among predictors, which was expected due to skewed distributions. This approach also allowed for analysis of prior probabilities. The model achieved an accuracy of 0.66, performing better than logistic regression but slightly below boosted trees.

Most predictors exhibited similar means for "deal" and "no deal" groups. However, some categories such as *description_length*, *fashion and apparel*, *home and living*, and *professional services* showed greater spread and were associated with a lower likelihood of securing a deal. Early seasons and later parts of a season were more likely to result in a deal, reinforcing observations from other models.

QDA Model Results	
Metric	**Value**
Accuracy	0.663
Kappa	0.325
Sensitivity	0.638
Specificity	0.688
Precision	0.661
F1 Score	0.649

Table 3: QDA metrics

4 Clustering Techniques

Before finalizing conclusions, unsupervised clustering techniques were employed to identify patterns in the data.

4.1 Clustering with PCA

After plotting the values on a graph using the two principal components, which together explain 56.99% of the variance in the data, we observed a disparity between the "deal" and "no deal" outcomes. This suggests that no single predictor significantly influences these components. However, there appears to be a slight concentration of "no deal" outcomes associated with the *valuation* and *amount asked* variables.

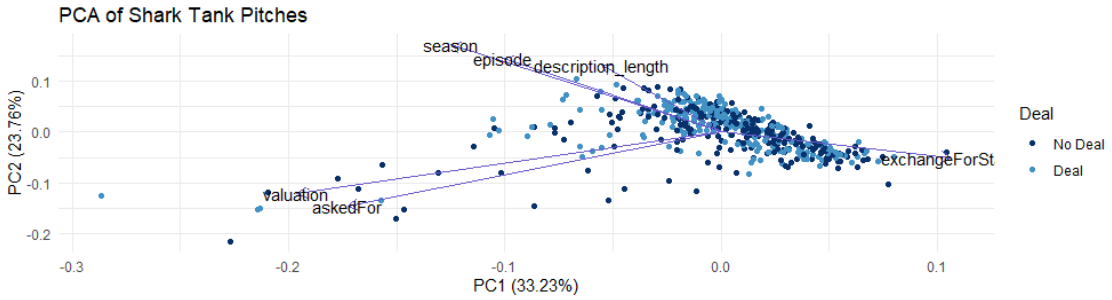


Figure 1: PCA for clusters showing Deal and No Deal pitches.

4.2 K-Means Clustering

K-means clustering was used to explore grouping patterns further, focusing on predictors such as *episode* and *season*. The within-cluster variation declined sharply from 2 to 3 clusters, with diminishing improvements beyond 3 clusters. This observation aligned with the *season_part* variable, which categorized data into early, mid, and late parts of the season.

When plotting "deal" and "no deal" outcomes with distinct symbols, no clear pattern emerged based on *season* or *episode*. This suggests that deals are not strongly influenced by these specific characteristics, emphasizing the need to focus on other predictors.

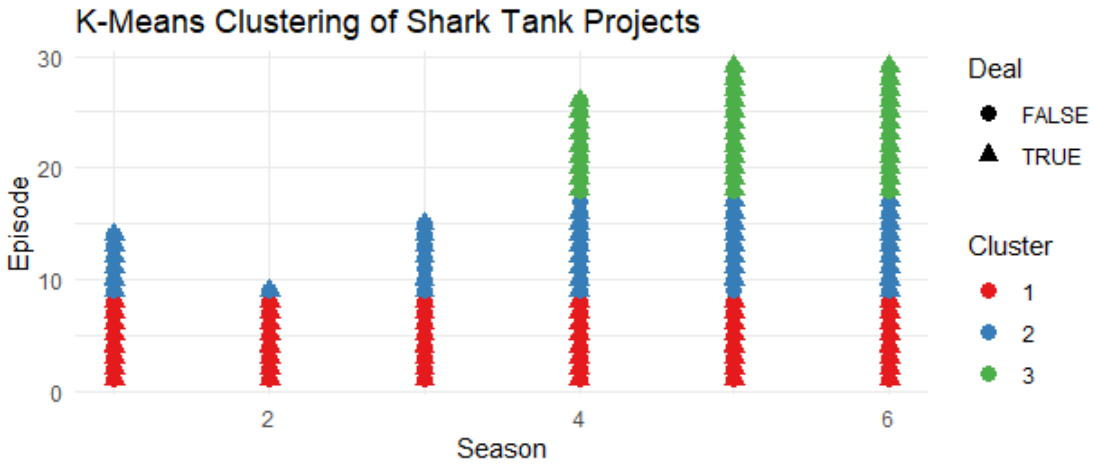


Figure 2: K-means for season and episodes

5 Results and Conclusions

Across all our models, the hypothesis that the television business side strongly influences whether an investment is made appears to be largely rejected. In our highest-accuracy model, the only variable related to the show's television characteristics (season part) had some influence, but its importance was overshadowed by other factors, such as the category of the project. This finding was consistent with logistic regression, where predictors related to the television aspect were dropped due to a lack of statistical significance. On the other hand, *categories* and the *ratio difference* between the amount asked and valuation consistently proved to be significant predictors of whether a deal was made. Similarly, in the QDA model, certain categories were significant; however, season-related data exhibited a wider spread of values, hinting at minor influence.

Our clustering techniques further supported these findings. While there may be some influence from television-specific characteristics, such influence appears to be minimal. However, it is worth noting that variables related to the show were present, albeit to a lesser extent, across multiple models. The primary

drivers of investment decisions were consistently the project's *category* and *financial valuation* (or related metrics). Therefore, entrepreneurs should prioritize strengthening their financial valuations, as these consistently emerged as the most critical predictors of investment success.

Interestingly, the role of investors was consistently insignificant in our predictive models. This finding, coupled with the average values observed in the QDA analysis, suggests a potential bias among investors to balance their decisions—to make deals and reject pitches in roughly equal proportions. Alternatively, it is possible that the application process to appear on the show filters pitches in a way that aligns with what the investors consider good business ideas. Additionally, it is conceivable that some members of specific investor combinations are more likely to make deals than others, an aspect that could not be explored due to the limitations of the available data. Incorporating this information into future models would provide valuable insights.

While the influence of television-related factors was not as strong as initially hypothesized, it was present to some degree. This, combined with the lack of significance of certain predictors, offers important takeaways for entrepreneurs. As mentioned earlier, they should focus on avoiding less promising categories and refining their financial valuations. Since the role of individual investors appears to be limited, entrepreneurs should consider that a rejection on the show does not necessarily reflect the viability of their business. Moreover, they should avoid excessive expenditures or over-preparation for the sake of impressing the investors, as this could jeopardize the financial health of their startup.

6 Appendix

6.1 Data Description

The following table describes the variables for the Shark Tank dataset:

Variable Descriptions for Shark Tank Dataset

Variable	Description	Type
Description	A description of the business idea presented during the pitch.	Text
Category	The industry category of the business idea.	Text
Entrepreneurs	Names of the entrepreneurs.	Text
Location	Neighborhood and state where the episode featuring the pitch took place.	Text
Website	Link to the business website.	Text
Shark 1-5	Names of investors evaluating the pitch. The order has no specific meaning.	Text
Title	Title of the business idea.	Text
Episode.season	A combination of the season and episode number.	Text
Deal	Indicator of whether a deal was made (Yes or No).	Binary
Multiple.Entrepreneurs	Indicates whether there were multiple entrepreneurs (Yes or No).	Binary
Episode	The episode number within the season.	Numerical
AskedFor	The amount of money requested by the entrepreneur.	Numerical
ExchangeForStake	The percentage of equity the entrepreneur asked by the investors.	Numerical
Valuation	The business valuation given by the investors.	Numerical
Season	The season number of the episode.	Numerical

Figure 3: Table for data description.

6.2 Testing interaction between variables

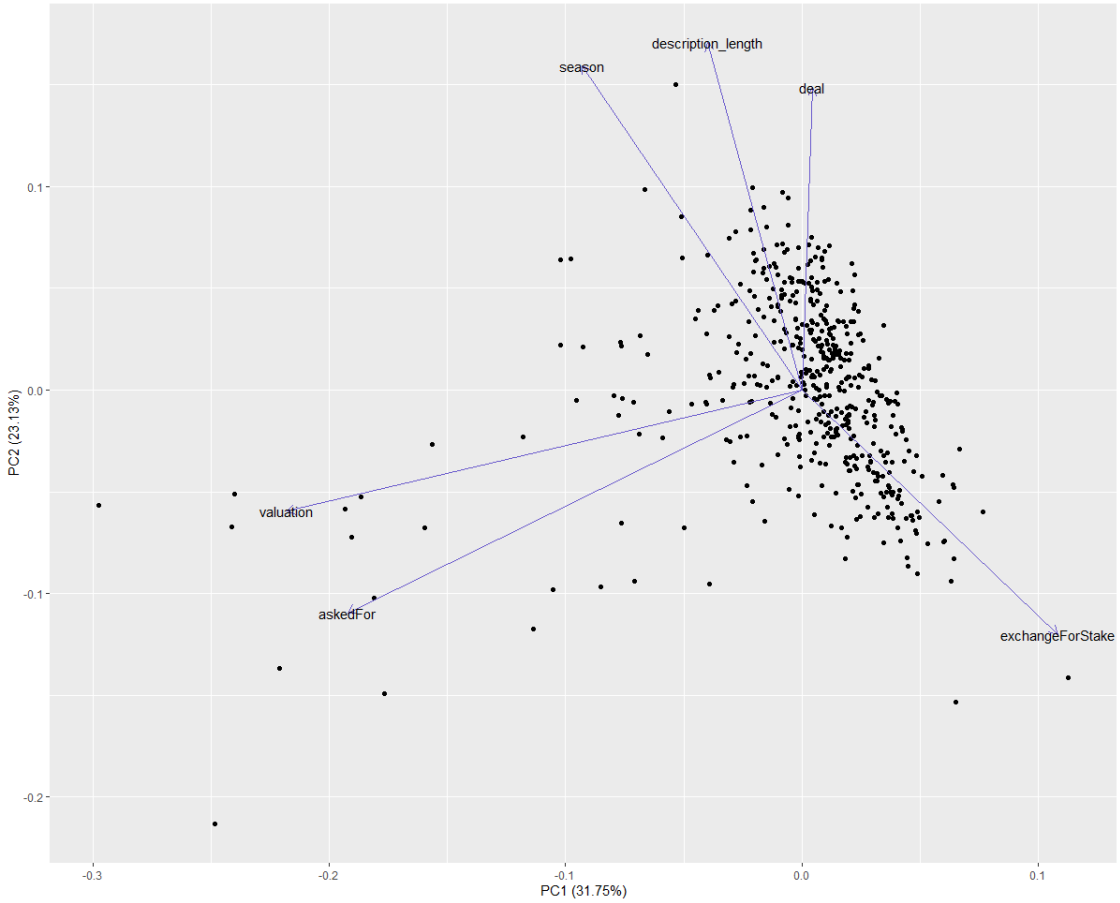
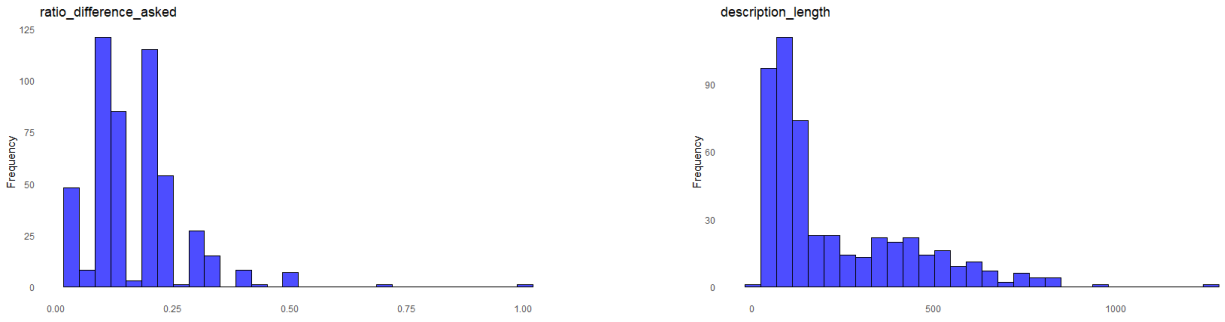


Figure 4: PCA 1 and 2 .

6.3 Histograms for Continuous Variables

The following histograms illustrate the distributions of key continuous variables.

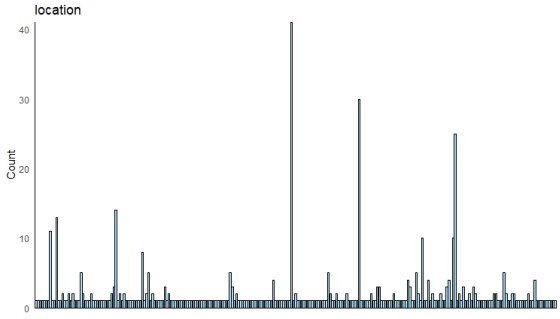


(a) Rate Difference Asked

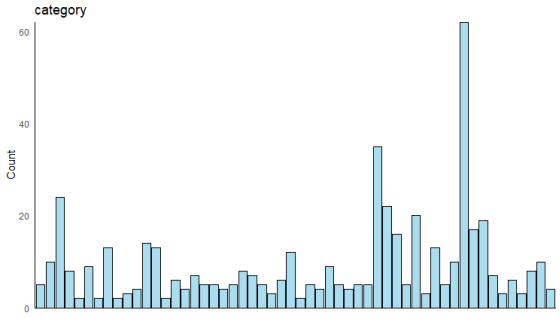
(b) Description Length

Figure 5: Histograms for continuous variables.

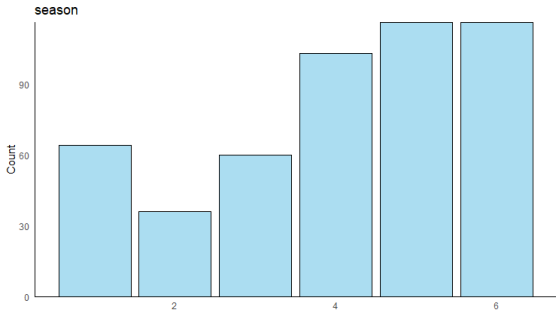
6.4 Bar Charts for Categorical Variables



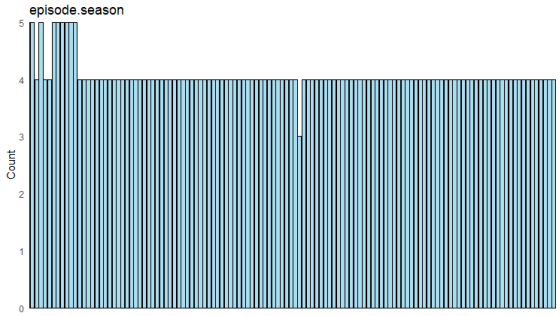
(a) Location



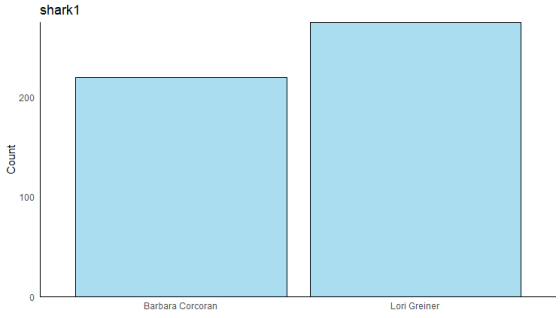
(b) Category



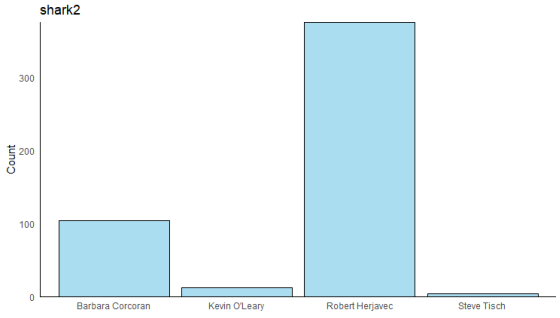
(c) Season



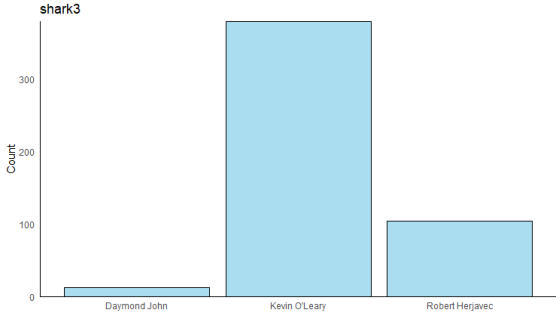
(d) episode.season



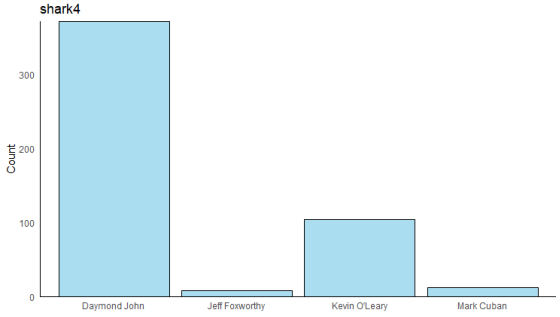
(e) Shark 1



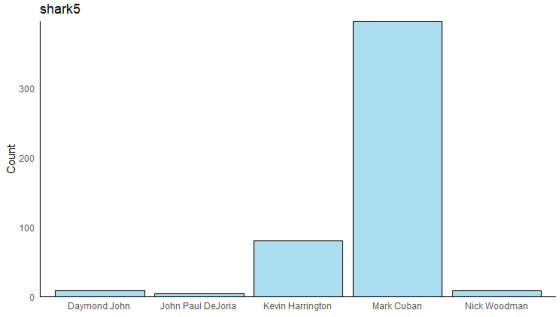
(f) Shark 2



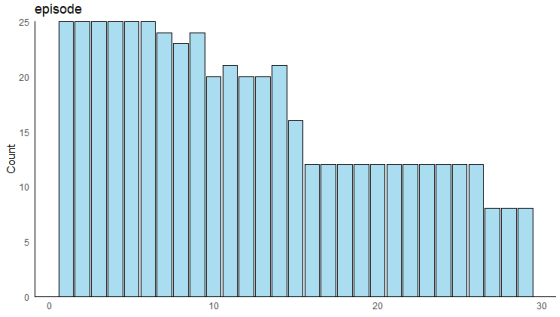
(g) Shark 3



(h) Shark 4



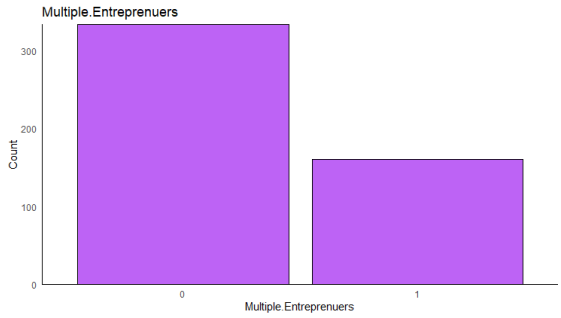
(i) Shark 5



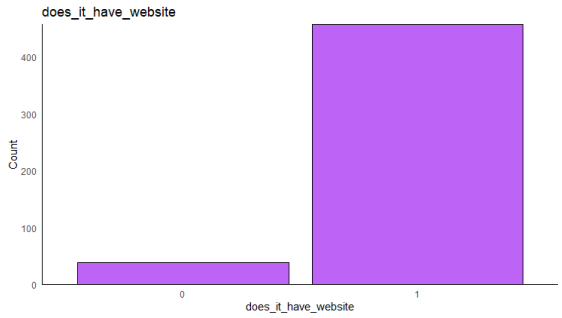
(j) Episode

6.5 Bar Charts for Binary Variables

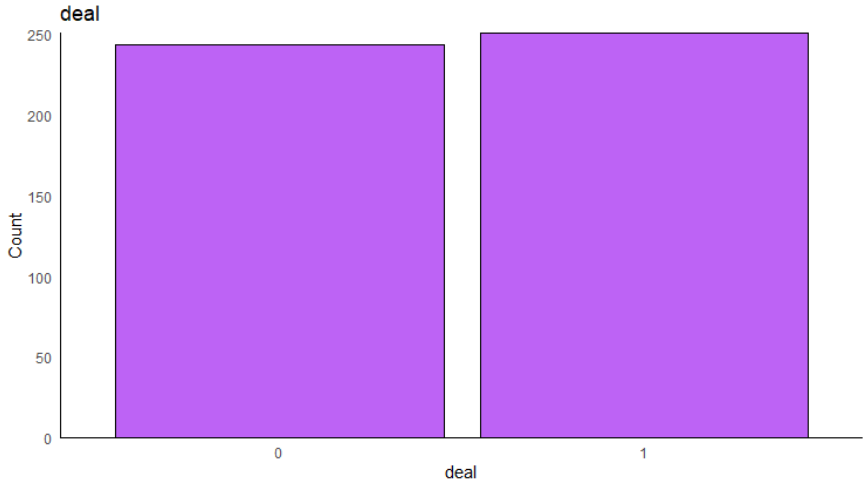
The following bar charts show the distributions for binary variables.



(a) Multiple Entrepreneurs



(b) Does it Have a Website

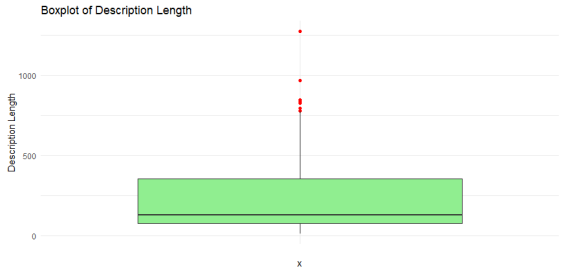


(c) Deal

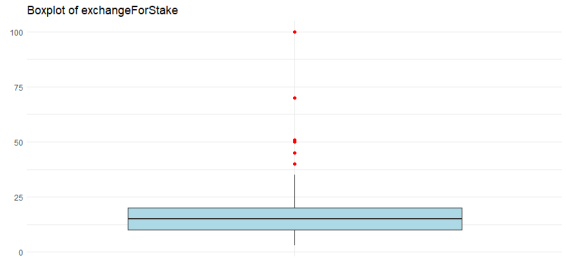
Figure 7: Bar charts for binary variables.

6.6 Box Plot for Outliers Detection

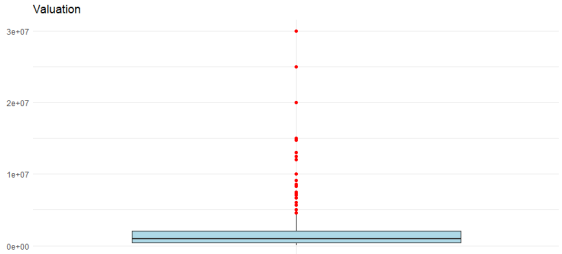
The following box plots illustrate the detection of outliers in numeric variables.



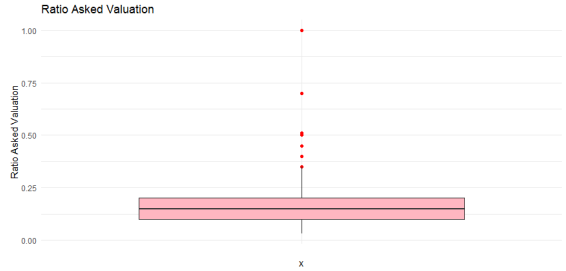
(a) Boxplot of Description Length



(b) Boxplot of Exchange for Stake



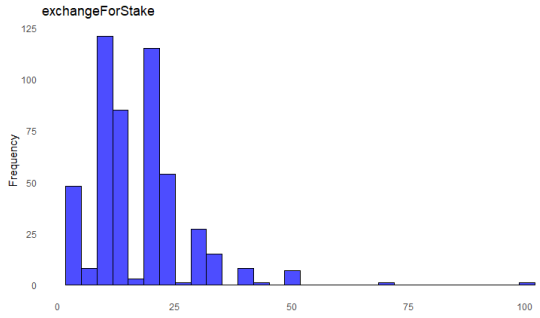
(c) Valuation



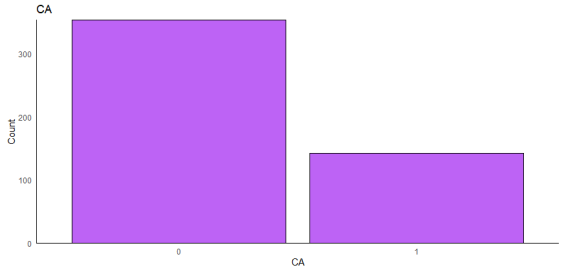
(d) Ratio Asked Valuation

Figure 8: Box plot for outliers detection in numeric variables.

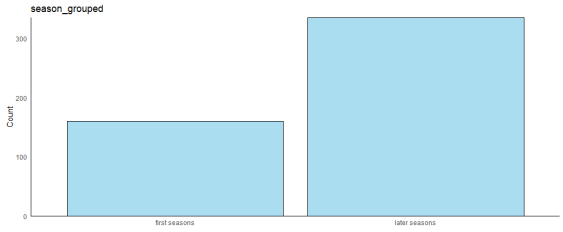
6.7 Numeric, categorical and binary variables created or changed after feature engineering



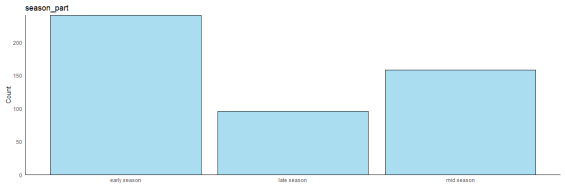
(a) Exchange for Stake



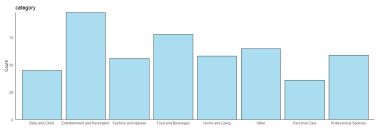
(b) CA



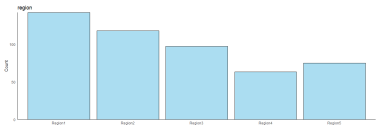
(c) Season Grouped



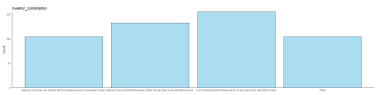
(d) Season Part



(e) Category



(f) Region



(g) Investor Combination

Figure 9: Histograms and bar charts for numeric, categorical, and binary variables after feature engineering. If the variable is not shown, it remained unchanged.

The bar charts display the distributions for key categorical variables. It is worth noting that the charts refer to location category and episode. season are missing the labels in the X axis for readability, but the values can be consulted in the code; they are not disclosed here for extension related purposes.

References

Micah Rosenbloom discusses shark tank versus reality. (2014). In Films On Demand. Films Media Group. https://fod.infobase.com/PortalPlaylists.aspx?wID=103901\allowbreak&xtid=164690

Yu, J. (2024, June 28). How is a business valued on 'Shark Tank'? Investopedia. https://www.investopedia.com/articles/company-insights/092116/how-business-valued\allowbreak-on-shark-tank.asp

Ketchum, D. (2023, April 25). 'Shark Tank' rejects that became super successful. Yahoo Finance. https://finance.yahoo.com/news/shark-tank-rejects-became-super-\allowbreak110008433.html

Walton, J. (2024, May 17). 3 'Shark Tank' failures that made millions. Investopedia. https://www.investopedia.com/articles/personal-finance/101515/3-shark-tank-\allowbreakfailures-made-millions.asp

USTVDB. (2024, November 22). Shark Tank ratings on ABC. https://ustvdb.com/networks/abc/shows/\allowbreakshark-tank/

Investopedia. (n.d.). Business valuation. Retrieved December 1, 2024, from https://www.investopedia.com/terms/b/\allowbreakbusiness-valuation.asp