Descriptive Statistics Analysis of Shark Tank Companies

February 8, 2022

BA222 A1
Professor Carlos Casso Dominguez

Jacob Rose U29999110
Dana Winer U13501316
Gabriel Moncau U83571123
Andres Tamayo U45018174

Section I: Data Overview

Our project is focused on SharkTank, which is an ABC show where entrepreneurs around the United States can pitch their ideas to billionaires. The source of the data is Kaggle, which the data was collected from Shark Analytics. The observation unit is the 6 seasons of Shark Tank, which consists of 122 episodes and 495 companies, to understand which ones performed the best and asked the most. The entities of this dataset are the 495 companies that pitched on the show during this timeframe. The time range is 5 years (from 2009 to 2014) and covers the first 6 seasons of the Shark Tank show. The data was collected from Shark Analytics, which was able to aggregate the information into one relative area. The dataset was created by Chase Willden and contained around 500 samples along with Title, Valuation, technical information, and other features.

Moreover, our group decided to collect this data because we thought it would be interesting to take the 6 seasons of Shark Tank and compare which ones performed the best and asked the most. As well, all members of the group like the show and are interested in learning about entrepreneurship and venture capital. At the moment of comparing the variables, Valuation and Deal are the most important variables. On the one hand, Valuation is crucial because the investors can establish how much the company is worth. On the other hand, Deal is essential because Deal shows the outcome of each pitch (observation). Furthermore, after writing the data dictionary, our group concluded that one variable is not fully defined. The exchangeForStake variable contains numerical values of percentages expressed as whole numbers (percentages multiplied by 100) making it partially undefined. Finally, subsequently analyzing the observations in the data, we concluded that there are no missing observations in the data, as well as 100% of observations are fully complete.

Section II: Summary Statistics

The main focus of this dataset centers around the valuation variable. This variable is related to the askedFor and exchangedForStake variables mathematically. The company's valuation, capital request, and percent offer values all follow the equation $exchangedForStake = \frac{askedFor}{valuation}$.

The valuation variable follows a normal distribution as shown in the left, densely populated side of the histograms below. I included two almost identical histograms to show two different visualizations of the valuation data. Figure A shows the big picture summary of the data, while in Figure B we can observe the distribution enhancing the bulk of the data on the smaller valuation end.

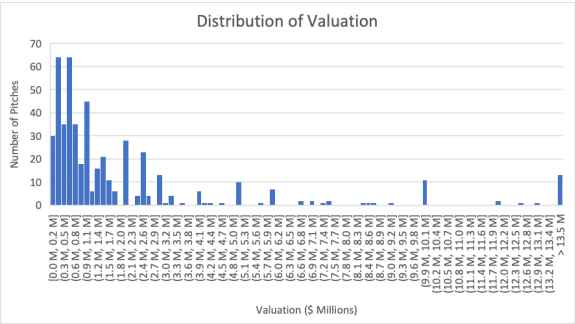


Figure A

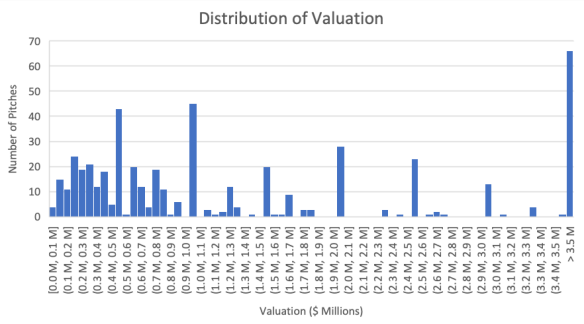


Figure B

We can see a positive skew in the distribution of valuation graphically due to high-value outliers such as the maximum value of $30,000,000, an almost ridiculously high value for a company looking for funding on Shark Tank. Since the median of $1,000,000 is less than the mean of $2,165,614.68, a positive skew is mathematically proven.

With a mode of $1,000,000 appearing 45 times within this set, there are predictable, yet confusing groupings of observations within this variable. There are also notable clumps of data around the $0.5 M, $1 M, $2 M, and $2.5 M marks as shown in Figure A above. These are due to the nature of startup valuations. Valuations are estimates of the value of a startup company, taking into account its present value and potential for future growth. Many startups value their companies themselves, making valuations of nice-looking, rounded numbers such as the clumps in our graphs. We can additionally look at Figure C below and compare it to our valuation distributions in Figures A and B for reference on the type of companies that have their pitches aired on Shark Tank.

Estimated Company Value	Stage of Development
$250,000 - $500,000	Has an exciting business idea or business plan
$500,000 - $1 million	Has a strong management team in place to execute on the plan
$1 million - $2 million	Has a final product or technology prototype
$2 million - $5 million	Has strategic alliances or partners, or signs of a customer base
$5 million and up	Has clear signs of revenue growth and obvious pathway to profitability

Figure C (McClure)

Most companies fall in the first two categories, as Shark Tank deals are usually sought out by extremely fresh startups with minimal current infrastructure. Lastly, this variable passes the common sense test, because it is expected that the distribution of valuations will lean to the left, centering around $2,000,000 as those values encompass the average startup's valuation in their early stages.

The deals variable is a categorical variable classified be true or false. It measures whether one of the sharks accepted or rejected a deal with the startup company. When evaluating the deals variable, we find that there is very close to an even amount of deals made as deals rejected.

Figure D

This variable is very important when considering the relationships between other variables explained in the subsequent section. There is an even amount of deals and rejections due to the nature of the show and the need for viewer retainment. Because there is an element of surprise in the show related to whether a deal is made or not, producers must have an even distribution of the two to keep viewers guessing and interested in the show. Because of this, the variable passes the common sense test and the distribution of deals made is reasonable.

When observing the category variable, we can see that a wide range of types of companies has pitched on Shark Tank. Because this is a categorical variable, there are little to no summary statistics we can include, however, we can see trends in the popularity of different categories of products.

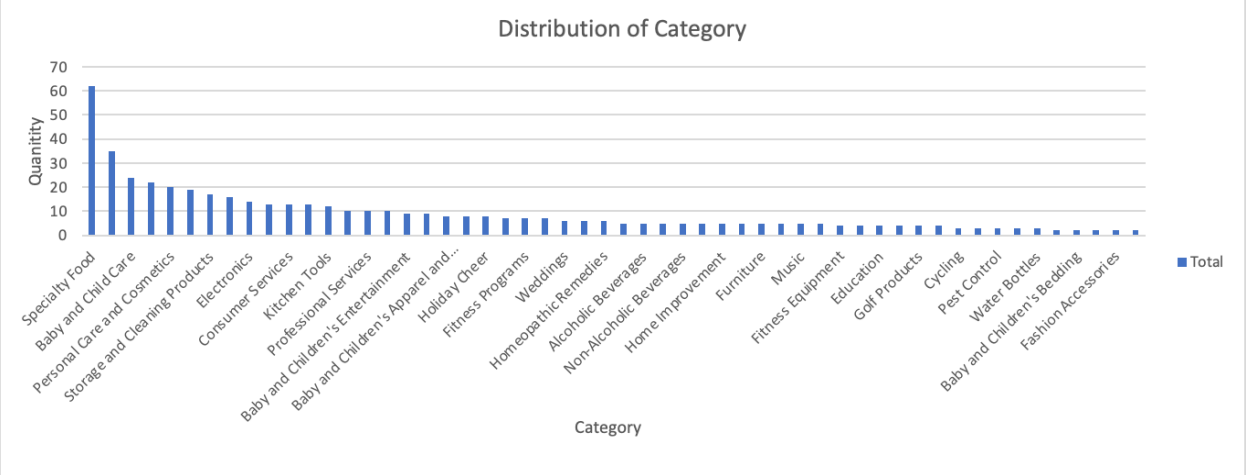


Figure E

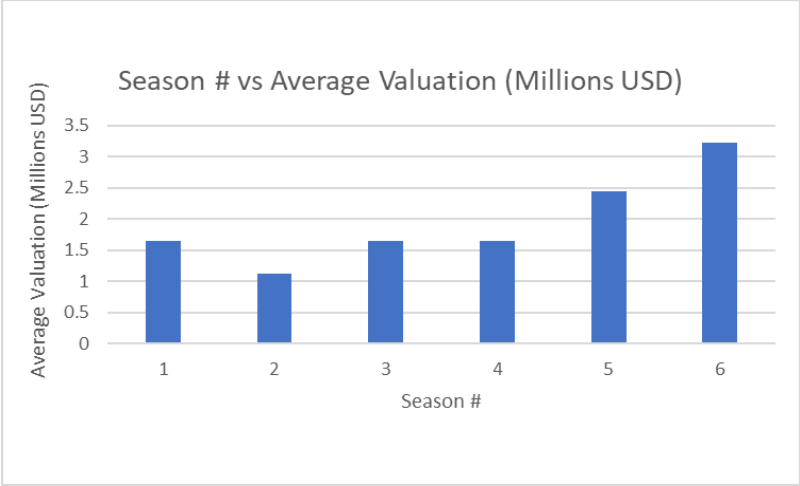
Illustrated in Figure E above, we can see that the top four categories by the number of pitches are Specialty Food, Baby and Child Care, Personal Care and Cosmetics, and Storage and Cleaning Products. All four of these categories can be classified under the broad Food and Home Goods industry. Specialty Food is the maximum category at 62 pitches, while Costumes, Baby and Children's Bedding, Baby and Children's Food, Fashion Accessories, and Maternity are all tied for the minimum category at 2 pitches each. This variable does pass the common sense test for range, however, it only does so if we take into account the usual types of companies that go on

pitch shows similar to Shark Tank. The top categories accurately represent topics that viewers would be interested in, explaining most of this distribution.

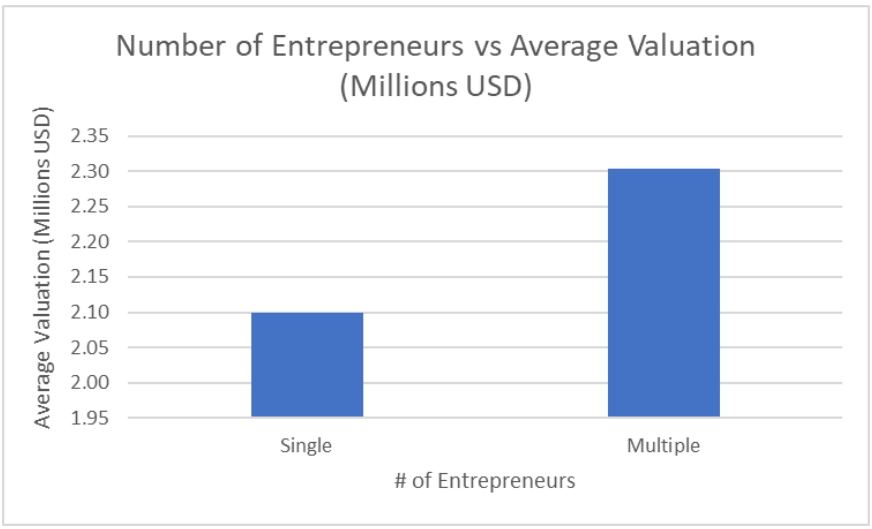
Section III: Relationship between Variables

When it comes to the relationship between some of our variables there were four in particular that caught our attention: season vs valuation, number of entrepreneurs vs valuation, percent of deals made vs location, and percentage of deals made vs count of pitches in location.

The first interesting relationship was season vs valuation (more specifically, the season # vs. the average valuation in all of its episodes in millions of USD). Using a correlation matrix we calculated there to be a weak positive correlation between the two variables of 0.157, visualized on the graph to the right. While the average valuation increases steadily from seasons 4 to 6 the same trend is not observed in the first three seasons with 2 being the lowest of the dataset. As such, the relationship is non-linear, though in the long term it could have straightened out further should the trend have continued. There were no outliers in this data and the correlation holds when adjusted for inflation, ruling that out as a confounding variable. We found this relationship surprising, not because of the weak positive correlation, but rather because of its non-linearity, as we expected the trend to have been steady for all seasons or dip at the end, not at the start.

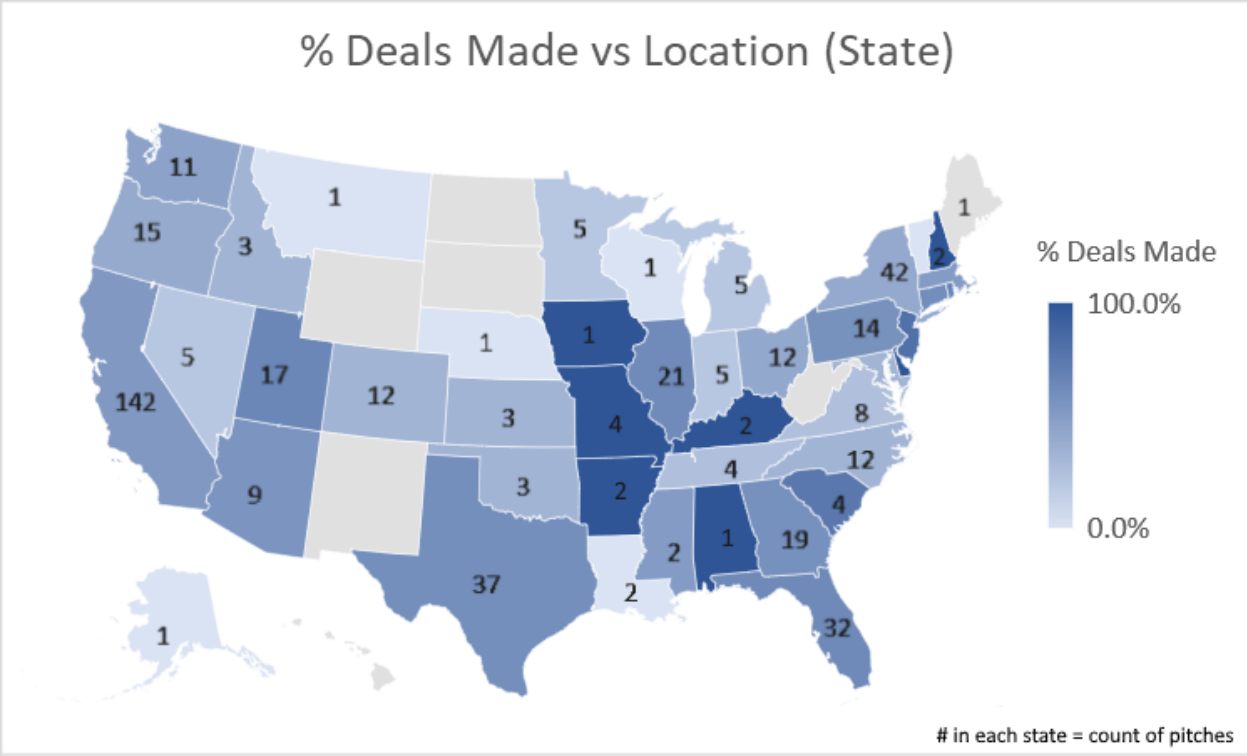


For the number of entrepreneurs vs valuation, we expected to see a similar number for both categories or a weak positive correlation (if any). Instead, pitches with multiple entrepreneurs had a valuation of 2.30 million USD, a 9.7% increase over the 2.09 million average for single entrepreneurs. This moderate positive relation may be indicative of projects with multiple people being more ambitious and therefore being assigned more value. Alternatively, more money may be sought out due to there being multiple people financially



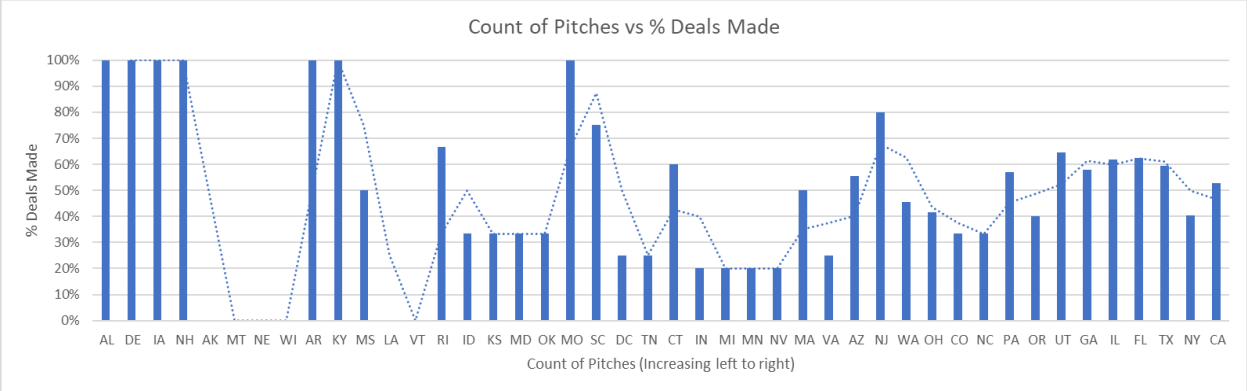
invested. By comparing the mins and maxes we were able to rule out the presence of outliers, as they were virtually identical. This relationship was difficult to analyze much further due to the nature of comparing numerical and categorical variables. However, putting the findings into context revealed further insights as, despite the almost double-digit percentage gains with multiple entrepreneurs, those pitches made up only 32.53% of the dataset.

The third most interesting relationship we observed was deals made (%) vs location (U.S state). To analyze this we first parsed the state abbreviations (ex. MA, CA, etc) from the location column which also includes the city. Then, through multiple pivot table calculations and graphical representations, we generated the map below. This graphic shows every U.S state colored according to their percentage of deals made with darker states being higher and lighter ones being lower. The number of pitches originating from each state is included within them to clarify which are lacking in sample size with empty ones not appearing in the data set at all. From this, we can see that there are 7 states with 100%, New Hampshire, Arkansas, and Missouri among them. However, these states average only 1.70 pitches, much lower than the dataset average of 11.25. Similarly, there were 6 states with 0%, including Louisiana, Vermont, and Wisconsin, averaging only 1.33 pitches each.



From this data a second relationship can be seen, count of pitches vs % deals made, where the states with more pitches trend closer to 50% deals made, while those with less go to the extremes (0 and 100). The relationship is characterized by a very weak positive non-linear correlation of 0.06. This is further supported by the high standard deviation in the number of pitches per state of 22.3, almost double the mean of 11.25, and a range of 141, more than 12 times the mean. It

should also be noted that California leads the number of pitches with 142, exactly 100 more than the state with the next most, New York, at 42, thereby largely inflating the range as an outlier. This is likely due not only to the strong entrepreneurial culture in California but also because the show was shot at Sony Picture Studios in Culver City, CA. These disparities make the data difficult to visualize (and, consequently, the non-linear trend difficult to characterize) as values of 100 and 0 misleadingly average to 50 but are made up of extremes.



To conclude, this relationship (while initially surprising) is reasonable as a limited sample size is likely to produce extremes rather than an expected middle range value and the overwhelming disparity in pitch counts between states is understandable given the external factors mentioned.

Works Cited

McClure, Ben. “Valuing Startup Ventures.” Investopedia. Investopedia, September 13, 2021. https://www.investopedia.com/articles/financial-theory/11/valuing-startup-ventures.asp#:~:text=While%20many%20established%20corporations%20are,investors%20are%20willing%20to%20pay.

Willden, Chase. “Shark Tank Companies.” Boston: Massachusetts, November 28, 2017. https://www.kaggle.com/yamqwe/shark-tank-companiese

Credit to Chase Willden for creating the dataset.

Shark Tank Data Dictionary

Description

A dataset containing the valuations, deal status, and other attributes of 495 pitches in 122 episodes of the ABC TV show Shark Tank over the course of 6 seasons. Data compiled by Shark Analytics (https://sharkanalytics.com/).

Variables

- **deal**: whether the deal went through (TRUE or FALSE)
- **description:** text describing the pitch (ranging from one sentence to a paragraph)
- **category:** characterizes the nature of the pitch into one of 54 categories: Mobile Apps, Education, Alcoholic Beverages, Electronics, Fitness Equipment, Music, Entertainment, Party Supplies, Home Security Solutions, Outdoor Recreation, Automotive, Health and Well-Being, Women's Apparel, Fitness Apparel and Accessories, Men and Women's Apparel, Fashion Accessories, Online Services, Consumer Services, Home Accessories, Men and Women's Shoes, Maternity, Cycling, Specialty Food, Pet Products, Baby and Children's Apparel and Accessories, Men and Women's, Accessories, Toys and Games, Storage and Cleaning Products, Novelties, Water Bottles, Weddings, Baby and Children's Food, Gardening, Fitness Programs, Professional Services, Baby and Child Care, Furniture, Baby and Children's Entertainment, Kitchen Tools, Homeopathic Remedies, Women's Shoes, Wine Accessories, Personal Care and Cosmetics, Women's Accessories, Productivity Tools, Undergarments and Basics, Non-Alcoholic Beverages, Holiday Cheer, Pest Control, Costumes, Home Improvement, Men's Accessories, Baby and Children's Bedding, or Golf Products.
- **entrepreneurs:** the full name of the entrepreneur(s) who pitched (1 or multiple)
- **location:** where the entrepreneurs listed are from (City, STATE).
    - It is unclear if this is where they were born, reside, or identify as being from.
- **website:** full text URL to the company's website (in HTTP:// or HTTPS:// formats)
- **episode:** the numerical episode in which the pitch participated (1 – 29)
- **askedFor:** the value requested as funding (in USD, = **valuation * exchangedForStake**)
- **exchangedForStake**: the % stake in the company the entrepreneur(s) are offering in exchange for the askedFor value from the Sharks (in USD, = **askedFor / valuation)**
- **valuation:** the entrepreneur's self-assigned estimated worth of their company (in USD, = **askedFor / exchangedForStake)**
- **season:** the numerical season in which the pitch participated (1 – 6)
- **shark1-5:** the full names of the judges present during the pitch (always all 5)
- **title:** the name of the entrepreneur's pitch, company, and/or product (<1 sentence)
- **episode-season:** the episode and season numerical values for the pitch (episode-season)
    - Note: automatically formatted to a date in excel
- **Multiple Entrepreneurs:** whether multiple entrepreneurs pitched (TRUE or FALSE)