Shark Tank

Shark Tank is a reality TV show. Contestants pitch their idea for a company to a panel of investors (a.k.a. "sharks"), who then decide whether or not to invest in that company. The investors give a certain amount of money in exchange for a percentage stake in the company ("equity"). If you are not familiar with the show, you may want to watch part of an episode here to get a sense of how it works.

The data that you will examine in this lab contains data about all contestants from the first 6 seasons of the show, including:

- the name and industry of the proposed company
- whether or not it was funded (i.e., the "Deal" column)
- which sharks chose to invest in the venture (N.B. There are 7 regular sharks, not including "Guest". Each shark has a column in the data set, labeled by their last name.)
- if funded, the amount of money the sharks put in and the percentage equity they got in return

To earn full credit on this lab, you should:

- use built-in `pandas` methods (like `.sum()` and `.max()`) instead of writing a `for` loop over a `DataFrame` or `Series`
- use the split-apply-combine pattern wherever possible

Of course, if you can't think of a vectorized solution, a `for` loop is still better than no solution at all!

In [37]:
```
import pandas as pd
```

Question 0. Getting and Cleaning the Data

The data is stored in the CSV file https://dlsun.github.io/pods/data/sharktank.csv. Read in the data into a Pandas `DataFrame`.

In [38]:
```
# YOUR CODE HERE
df = pd.read_csv("https://dlsun.github.io/pods/data/sharktank.csv")
df
```

Out[38]:

	Season	No. in series	Company	Deal	Industry	Entrepreneur Gender	Amount	Equity	Corco
0	1.0	1.0	Ava the Elephant	Yes	Healthcare	Female	$50,000	55%	
1	1.0	1.0	Mr. Tod's Pie Factory	Yes	Food and Beverage	Male	$460,000	50%	
2	1.0	1.0	Wispots	No	Business Services	Male	NaN	NaN	N
3	1.0	1.0	College Foxes Packing Boxes	No	Lifestyle / Home	Male	NaN	NaN	N
4	1.0	1.0	Ionic Ear	No	Uncertain / Other	Male	NaN	NaN	N
...	...	...	...	...	...	...	...	...	
490	6.0	28.0	You Kick Ass	Yes	Children / Education	Female	$100,000	10%	N
491	6.0	29.0	Shark Wheel	Yes	Fitness / Sports	Male	$225,000	8%	N
492	6.0	29.0	Gato Cafe	No	Uncertain / Other	Female	NaN	NaN	N
493	6.0	29.0	Sway Motorsports	Yes	Green/CleanTech	Male	$300,000	20%	N
494	6.0	29.0	Spikeball	Yes	Fitness / Sports	Male	$500,000	20%	N

495 rows × 17 columns

There is one column for each of the sharks. A 1 indicates that they chose to invest in that company, while a missing value indicates that they did not choose to invest in that company. Notice that these missing values show up as NaNs when we read in the data. Fill in these missing values with zeros. Other columns may also contain NaNs; be careful not to fill those columns with zeros, or you may end up with strange results down the line.

In [39]:

```
# YOUR CODE HERE
df[["Corcoran", "Cuban", "Greiner", "Herjavec", "John", "O'Leary", "Harrington"
df
```

Out[39]:

	Season	No. in series	Company	Deal	Industry	Entrepreneur Gender	Amount	Equity	Corco
0	1.0	1.0	Ava the Elephant	Yes	Healthcare	Female	$50,000	55%	
1	1.0	1.0	Mr. Tod's Pie Factory	Yes	Food and Beverage	Male	$460,000	50%	
2	1.0	1.0	Wispots	No	Business Services	Male	NaN	NaN	
3	1.0	1.0	College Foxes Packing Boxes	No	Lifestyle / Home	Male	NaN	NaN	
4	1.0	1.0	Ionic Ear	No	Uncertain / Other	Male	NaN	NaN	
...	...	...	...	...	...	...	...	...	
490	6.0	28.0	You Kick Ass	Yes	Children / Education	Female	$100,000	10%	
491	6.0	29.0	Shark Wheel	Yes	Fitness / Sports	Male	$225,000	8%	
492	6.0	29.0	Gato Cafe	No	Uncertain / Other	Female	NaN	NaN	
493	6.0	29.0	Sway Motorsports	Yes	Green/CleanTech	Male	$300,000	20%	
494	6.0	29.0	Spikeball	Yes	Fitness / Sports	Male	$500,000	20%	

495 rows × 17 columns

Notice that Amount and Equity are currently being treated as categorical variables (`dtype: object`). Can you figure out why this is? Clean up these columns and cast them to numeric types (i.e., a `dtype` of `int` or `float`) because we'll need to perform mathematical operations on these columns.

In [40]:

```
#Cleaning our Amount and Equity columns by removing the $ and %.
df["Amount"]= df["Amount"].astype(str).str.replace("$", "")
df["Amount"] = df["Amount"].astype(str).str.replace(",", "")
df["Amount"] = df["Amount"].astype(float).fillna(0)

df["Equity"] = df["Equity"].astype(str).str.replace("%", "").fillna(0)
df["Equity"] = df["Equity"].astype(float) / 100
```

Question 1. Which Company was Worth the Most?

The valuation of a company is how much it is worth. If someone invests $10,000 for a 40% equity stake in the company, then this means the company must be valued at $25,000, since 40% of $25,000 is $10,000.

Calculate the valuation of each company that was funded. Which company was most valuable? Is it the same as the company that received the largest total investment from the sharks?

In [41]:

```
# Calcuating valuation by dividing the amount by the equity
valuation = df["Amount"] / df["Equity"]
valuation.sort_values(ascending = False).head(5)
```

Out[41]:

```
312           inf
421    25000000.0
464    13000000.0
489    12000000.0
483    10000000.0
dtype: float64
```

In [42]:

```
df.loc[[312]]
```

Out[42]:

	Season	No. in series	Company	Deal	Industry	Entrepreneur Gender	Amount	Equity	Corcoran	Cuba
312	5.0	13.0	The Wall DoctoRX	Yes	Lifestyle / Home	Male	150000.0	0.0	0.0	0.

In [43]:

```
df.loc[[421]]
```

Out[43]:

	Season	No. in series	Company	Deal	Industry	Entrepreneur Gender	Amount	Equity	Corcoran	Cu
421	6.0	11.0	Zipz	Yes	Food and Beverage	Male	2500000.0	0.1	0.0	

In [44]:

```
#Seeing the 5 companies who had the highest amount of investment in dollars
df["Amount"].sort_values(ascending = False).head(5)
```

Out[44]:

```
483    5000000.0
489    3000000.0
421    2500000.0
284    2000000.0
363    1750000.0
Name: Amount, dtype: float64
```

In [45]:

```
df.loc[[483]]
```

Out[45]:

	Season	No. in series	Company	Deal	Industry	Entrepreneur Gender	Amount	Equity	Corcor
483	6.0	27.0	AirCar	Yes	Green/CleanTech	Male	5000000.0	0.5	(

YOUR INTERPRETATION HERE

The company that was the most valuable in funding was a food and beverage venture called Zipz, as they were valued at 25000000-the highest valuation score based on the equity of their product and the amount invested. In this context, equity refers to the percent of the company that the sharks would have a stake in. While The Wall DoctoRX had an infinite valuation according to the sorting of the values, ZipZ was the company that had the largest finite valuation. The company with the highest valuation is not the same as the company that receieved the largest total investment from the sharks because that was AirCar with a total amount of 5000000 invested.

Question 2. Which Shark Invested the Most?

Calculate the total amount of money that each shark invested over the 6 seasons. Which shark invested the most total money over the 6 seasons?

Hint: If n sharks funded a given venture, then the amount that each shark invested is the total amount divided by n.

In [46]:

```
df["Shark Numbers"] = df[["Corcoran", "Cuban", "Greiner", "Herjavec", "John", '
df["Total Per Shark"] = df["Amount"] / df["Shark Numbers"] #Calculating the to
df[["Corcoran", "Cuban", "Greiner", "Herjavec", "John", "O'Leary", "Harrington'
```

Out[46]:

```
Corcoran       4912500.0
Cuban         17817500.0
Greiner        8170000.0
Herjavec      16297500.0
John           8154000.0
O'Leary        7952500.0
Harrington      800000.0
Guest           400000.0
dtype: float64
```

YOUR INTERPRETATION HERE

Based on the calcuations of the total amount invested per shark, it appears that Cuban invested the most over the course of the 6 seasons. Cuban invested a grand total of 17817500 dollars with Herjavec ivesting a close second of 16297500.

Question 3. Do the Sharks Prefer Certain Industries?

Calculate the funding rate for each industry. That is, calculate the conditional distribution $p(funded|industry)$. Make a visualization showing this information.

In [47]:

```
marginal_industry = df['Industry'].value_counts(normalize = True) #Conditionin
marginal_industry
```

Out[47]:

```
Industry
Food and Beverage        0.210101
Fashion / Beauty         0.187879
Lifestyle / Home         0.141414
Children / Education     0.111111
Fitness / Sports         0.080808
Software / Tech          0.066667
Consumer Products        0.038384
Healthcare               0.036364
Pet Products             0.034343
Business Services        0.026263
Uncertain / Other        0.026263
Media / Entertainment    0.022222
Green/CleanTech          0.018182
Name: proportion, dtype: float64
```

In [48]:

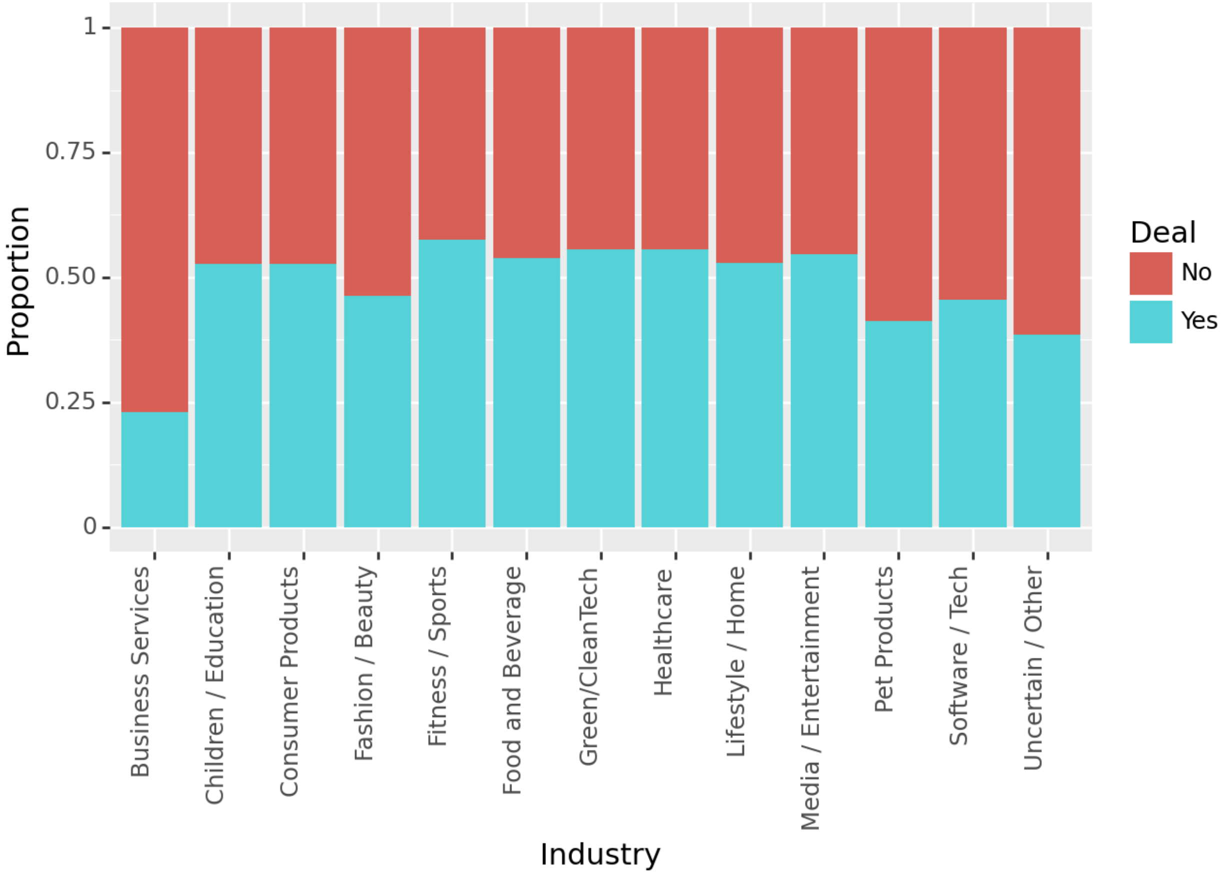
```
#Obtaining a table of the conditional probabilities of funding for each indust
joint_funding_industry = df[["Deal", "Industry"]].value_counts(normalize=True)
joint_funding_industry.divide(marginal_industry)
```

Out[48]:

Industry	Business Services	Children / Education	Consumer Products	Fashion / Beauty	Fitness / Sports	Food and Beverage	Green/CleanTech	Healt
Deal								
No	0.769231	0.472727	0.473684	0.537634	0.425	0.461538	0.444444	0.44
Yes	0.230769	0.527273	0.526316	0.462366	0.575	0.538462	0.555556	0.55

In [ ]:

```
from plotnine import *
#Making a visualization of the funding rates conditioned on industry
(ggplot(df, aes(x = "Industry", fill = "Deal"))
+ geom_bar(position = "fill")
+ ylab("Proportion")
+ theme(axis_text_x=element_text(rotation=90, hjust=1)) #code obtained from Sta
)
```



```
Out[ ]: <Figure Size: (640 x 480)>
```

YOUR INTERPRETATION HERE

The data visualization above demonstrates the funding rate (the frequency with which sharks agreed to make a deal and invest in a certain product) given the industry that product falls under. Based on this information, the fitness and sports industry, those relating to healthcare and sustainable tech, childhood development and education, and consumer products had the highest rates of funding.

Submission Instructions

- Restart this notebook and run the cells from beginning to end.
  - Go to Runtime > Restart and Run All.

```
In [ ]: # @markdown Run this cell to download this notebook as a webpage, `_NOTEBOOK.h

import google, json, nbformat

# Get the current notebook and write it to _NOTEBOOK.ipynb
raw_notebook = google.colab._message.blocking_request("get_ipynb",
                                                      timeout_sec=30)["ipynb"]
with open("_NOTEBOOK.ipynb", "w", encoding="utf-8") as ipynb_file:
  ipynb_file.write(json.dumps(raw_notebook))

# Use nbconvert to convert .ipynb to .html.
```

```
!jupyter nbconvert --to html --log-level WARN _NOTEBOOK.ipynb

# Download the .html file.
google.colab.files.download("_NOTEBOOK.html")
```

- Open `_NOTEBOOK.html` in your browser, and save it as a PDF.
    - Go to File > Print > Save as PDF.
- Double check that all of your code and output is visible in the saved PDF.
- Upload the PDF to Gradescope.
    - Please be sure to select the correct pages corresponding to each question.