Shark Tank

Shark Tank is a reality TV show. Contestants present 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. You can also search for a clip on YouTube.

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 [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

Question 0. Getting and Cleaning the Data

The data is stored in the CSV file `sharktank.csv`. Read in the data into a Pandas `DataFrame`.

In [2]:
```
# YOUR CODE HERE
df = pd.read_csv('sharktank.csv')
df.head()
```

Out[2]:

	Season	No. in series	Company	Deal	Industry	Entrepreneur Gender	Amount	Equity	Corcoran	Cuba
0	1.0	1.0	Ava the Elephant	Yes	Healthcare	Female	$50,000	55%	1.0	Na
1	1.0	1.0	Mr. Tod's Pie Factory	Yes	Food and Beverage	Male	$460,000	50%	1.0	Na
2	1.0	1.0	Wispots	No	Business Services	Male	NaN	NaN	NaN	Na
3	1.0	1.0	College Foxes Packing Boxes	No	Lifestyle / Home	Male	NaN	NaN	NaN	Na
4	1.0	1.0	Ionic Ear	No	Uncertain / Other	Male	NaN	NaN	NaN	Na

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 [3]:
```
# YOUR CODE HERE
df.loc[:, 'Corcoran':'Guest'] = df.loc[:, 'Corcoran':'Guest'].fillna(0)
df.head()
```

Out[3]:

	Season	No. in series	Company	Deal	Industry	Entrepreneur Gender	Amount	Equity	Corcoran	Cuba
0	1.0	1.0	Ava the Elephant	Yes	Healthcare	Female	$50,000	55%	1.0	0
1	1.0	1.0	Mr. Tod's Pie Factory	Yes	Food and Beverage	Male	$460,000	50%	1.0	0
2	1.0	1.0	Wispots	No	Business Services	Male	NaN	NaN	0.0	0
3	1.0	1.0	College Foxes Packing Boxes	No	Lifestyle / Home	Male	NaN	NaN	0.0	0
4	1.0	1.0	Ionic Ear	No	Uncertain / Other	Male	NaN	NaN	0.0	0

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 [4]:
```
# YOUR CODE HERE
df['Amount'] = df['Amount'].str.replace('$', '').str.replace(',', '').astype(fl
df['Equity'] = df['Equity'].str.replace('%', '').astype(float)

df['Amount'] = df['Amount'].fillna(0)
df['Equity'] = df['Equity'].fillna(0)

# def clean_amount(amount):
#     return float(str(amount).replace('$', '').replace(',', ''))
# df['Amount'].apply(clean_amount)

# df.['Amount'] = df['Amount'].apply(lambda x: float(str(x).replace('$', '').re

df.head()
```

```
/var/folders/8s/b4dwb16174x344bs1ylrswmc0000gr/T/ipykernel_17730/1882617138.p
y:2: FutureWarning: The default value of regex will change from True to False
in a future version. In addition, single character regular expressions will *n
ot* be treated as literal strings when regex=True.
  df['Amount'] = df['Amount'].str.replace('$', '').str.replace(',', '').astype
(float)
```

Out[4]:

	Season	No. in series	Company	Deal	Industry	Entrepreneur Gender	Amount	Equity	Corcoran	Cuba
0	1.0	1.0	Ava the Elephant	Yes	Healthcare	Female	50000.0	55.0	1.0	0
1	1.0	1.0	Mr. Tod's Pie Factory	Yes	Food and Beverage	Male	460000.0	50.0	1.0	0
2	1.0	1.0	Wispots	No	Business Services	Male	0.0	0.0	0.0	0
3	1.0	1.0	College Foxes Packing Boxes	No	Lifestyle / Home	Male	0.0	0.0	0.0	0
4	1.0	1.0	Ionic Ear	No	Uncertain / Other	Male	0.0	0.0	0.0	0

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 [5]:
```
# YOUR CODE HERE
df.drop(df[df['Equity'] == 0].index, inplace=True)
```

In [6]:
```
def calculate_valuation(company):
    return (company.Amount * 100) / company.Equity

df['Valuation'] = df.apply(calculate_valuation, axis=1)
```

In [7]:
```
df['Company'][df['Valuation'].idxmax()]
```

Out[7]:
```
'Zipz'
```

In [8]:
```
df['Company'][df['Amount'].idxmax()]
```

Out[8]:
```
'AirCar'
```

The company that we found to be the most valuable is Zipz. The company that received the largest total investment from the sharks is AirCar.

Question 2. Which Shark Invested the Most?

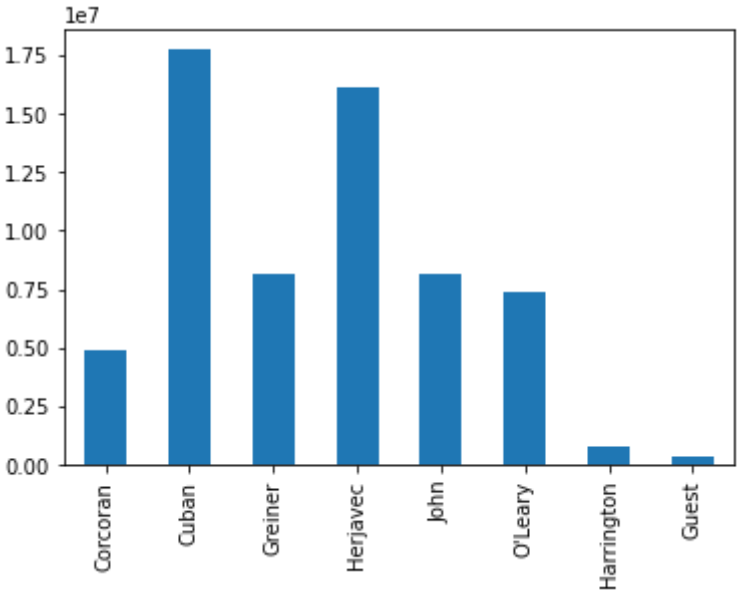
Calculate the total amount of money that each shark invested over the 6 seasons. Which shark invested the most total money over the 6 seasons?

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

In [9]:
```
# ENTER CODE HERE.
num_invested = df.loc[:, 'Corcoran':'Guest'].sum(axis=1)
investments = df['Amount'] / num_invested

df.loc[:, 'Corcoran':'Guest'].multiply(investments, axis=0).sum().plot(kind='ba
```

Out[9]:
```
<AxesSubplot:>
```



The shark that invested the most amount of money over 6 seasons was Mark Cuban.

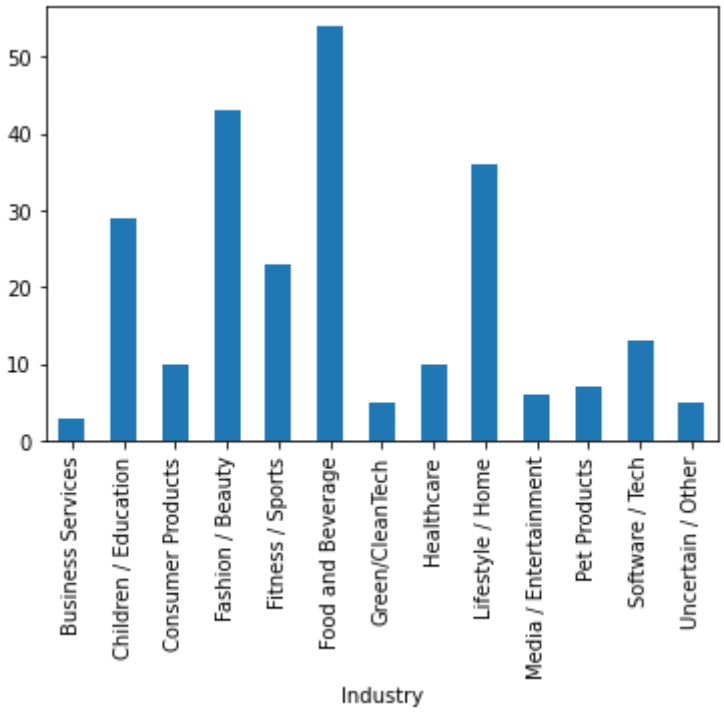
Question 3. Do the Sharks Prefer Certain Industries?

Calculate the funding rate (the proportion of companies that were funded) for each industry. Make a visualization showing this information.

In [10]:
```
# ENTER CODE HERE.
df['Deal'] = pd.get_dummies(df['Deal'])
```

In [11]:
```
df.groupby('Industry')['Deal'].sum().plot(kind='bar')
```

Out[11]:
```
<AxesSubplot:xlabel='Industry'>
```



The top three industries the sharks are most interested in are Food and Beverage, Fashion/Beauty, and Lifestyle/Home.

Submission Instructions

Once you are finished, follow these steps:

1. Restart the kernel and re-run this notebook from beginning to end by going to `Kernel > Restart Kernel and Run All Cells`.
2. If this process stops halfway through, that means there was an error. Correct the error and repeat Step 1 until the notebook runs from beginning to end.
3. Double check that there is a number next to each code cell and that these numbers are in order.

Then, submit your lab as follows:

1. Go to `File > Export Notebook As > PDF`.
2. Double check that the entire notebook, from beginning to end, is in this PDF file. (If the notebook is cut off, try first exporting the notebook to HTML and printing to PDF.)
3. Upload the Notebook (ipynb) to canvas (one submission per group).
4. Demo your lab by next Tuesday for full credit.