2A. Shark Tank

January 22, 2019

1 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.

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
        %matplotlib inline
```

1.1 Question 0. Getting and Cleaning the Data

The data is stored in the CSV file `https://raw.githubusercontent.com/dlsun/data-science-book/master/da` Read in the data into a Pandas `DataFrame`.

```
In [2]: st = pd.read_csv("https://raw.githubusercontent.com/dlsun/data-science-book/master/data
        st.head()
```

```
Out[2]:    Season  No. in series                      Company Deal          Industry  \
        0     1.0            1.0             Ava the Elephant  Yes        Healthcare
        1     1.0            1.0        Mr. Tod's Pie Factory  Yes  Food and Beverage
        2     1.0            1.0                      Wispots   No  Business Services
        3     1.0            1.0  College Foxes Packing Boxes   No   Lifestyle / Home
        4     1.0            1.0                   Ionic Ear   No  Uncertain / Other
```

```
  Entrepreneur Gender    Amount Equity  Corcoran  Cuban  Greiner  Herjavec  \
0              Female   $50,000    55%       1.0    NaN      NaN       NaN
1                Male  $460,000    50%       1.0    NaN      NaN       NaN
2                Male       NaN    NaN       NaN    NaN      NaN       NaN
3                Male       NaN    NaN       NaN    NaN      NaN       NaN
4                Male       NaN    NaN       NaN    NaN      NaN       NaN

   John  O'Leary  Harrington  Guest Details / Notes
0   NaN      NaN         NaN    NaN             NaN
1   1.0      NaN         NaN    NaN             NaN
2   NaN      NaN         NaN    NaN             NaN
3   NaN      NaN         NaN    NaN             NaN
4   NaN      NaN         NaN    NaN             NaN
```

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]: st.loc[:,["Corcoran", "Cuban", "Greiner", "Herjavec",
                  "John", "O'Leary","Harrington", "Guest"]]=st.loc[:,
                ["Corcoran", "Cuban",
                 "Greiner", "Herjavec",
                 "John", "O'Leary",
                 "Harrington", "Guest"]].fillna(0.0)
        st.head()
```

```
Out[3]:    Season  No. in series                        Company Deal           Industry  \
        0     1.0            1.0              Ava the Elephant  Yes         Healthcare
        1     1.0            1.0         Mr. Tod's Pie Factory  Yes  Food and Beverage
        2     1.0            1.0                       Wispots   No  Business Services
        3     1.0            1.0  College Foxes Packing Boxes   No   Lifestyle / Home
        4     1.0            1.0                     Ionic Ear   No  Uncertain / Other

          Entrepreneur Gender    Amount Equity  Corcoran  Cuban  Greiner  Herjavec  \
        0              Female   $50,000    55%       1.0    0.0      0.0       0.0
        1                Male  $460,000    50%       1.0    0.0      0.0       0.0
        2                Male       NaN    NaN       0.0    0.0      0.0       0.0
        3                Male       NaN    NaN       0.0    0.0      0.0       0.0
        4                Male       NaN    NaN       0.0    0.0      0.0       0.0

           John  O'Leary  Harrington  Guest Details / Notes
        0   0.0      0.0         0.0    0.0             NaN
        1   1.0      0.0         0.0    0.0             NaN
        2   0.0      0.0         0.0    0.0             NaN
        3   0.0      0.0         0.0    0.0             NaN
        4   0.0      0.0         0.0    0.0             NaN
```

In this problem I decided to change the NaN's to 0 by using the .fillna function using .loc and the names of the investors or the "sharks".

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]: st.Amount = st.Amount.str.replace('$', '')
        st.Amount = st.Amount.str.replace(',', '')
        st.Amount = st.Amount.fillna(0.0)
        st.Amount = pd.to_numeric(st.Amount)
```

```
In [5]: st.Equity = st.Equity.str.replace('%', '')
        st.Equity = st.Equity.fillna(0.0)
        st.Equity = pd.to_numeric(st.Equity)
        st.Equity = st.Equity/100
```

The Amount and Equity variables are being treated as Categorical variables because of the dollar sign and percent sign. In order to clean these variables, I first stripped the entries of their symbols($, %), using str.replace. Next, I filled the NaN's with 0 and changed the data to numeric. For the equity variable, I refactored the variable from a percent to a decimal by dividing by 100 using a technique called broadcasting.

1.2 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 [6]: st["Valuation"] = st.Amount/st.Equity
        st.iloc[st.Valuation.idxmax()]
```

```
Out[6]: Season                                                      4
        No. in series                                              13
        Company                                        Coffee Joulies
        Deal                                                      Yes
        Industry                                   Food and Beverage
        Entrepreneur Gender                                      Male
        Amount                                                 150000
        Equity                                                      0
        Corcoran                                                    0
        Cuban                                                       0
        Greiner                                                     1
        Herjavec                                                    1
        John                                                        1
        O'Leary                                                     1
        Harrington                                                  0
        Guest                                                                  0
        Details / Notes        plus royalty fee of $6 per Coffee Joulies sold...
        Valuation                                                            inf
        Name: 207, dtype: object
```

The most valuable company according to its valuation was Coffee Joulies. However, after further examination of this company's equity, the valuation is infinity because dividing the amount by 0 is undefined. Therefore, this company may not actually have the greatest valuation and we should remove all the infinite values.

```
In [7]: st.replace(np.inf, 0, inplace=True)
        st.iloc[st.Valuation.idxmax()]

Out[7]: Season                                                                 6
        No. in series                                                         11
        Company                                                             Zipz
        Deal                                                                 Yes
        Industry                                               Food and Beverage
        Entrepreneur Gender                                                 Male
        Amount                                                           2.5e+06
        Equity                                                               0.1
        Corcoran                                                               0
        Cuban                                                                  0
        Greiner                                                                0
        Herjavec                                                               0
        John                                                                   0
        O'Leary                                                                1
        Harrington                                                             0
        Guest                                                                  0
        Details / Notes        with an option for another $2.5 Million for an...
        Valuation                                                        2.5e+07
        Name: 421, dtype: object
```

After removing the infinite values from the data frame using "replace", we see that Zipz had the highest Valuation of 25 million dollars. Next, let's check to see if this company had the most money invested in it by the sharks.

```
In [8]: st.Amount.max() == st.iloc[st.Valuation.idxmax()].Amount

Out[8]: False

In [9]: st.iloc[st.Amount.idxmax()]

Out[9]: Season                                                                 6
        No. in series                                                         27
        Company                                                           AirCar
        Deal                                                                 Yes
        Industry                                                 Green/CleanTech
        Entrepreneur Gender                                                 Male
```

```
Amount                                                     5e+06
Equity                                                       0.5
Corcoran                                                       0
Cuban                                                          0
Greiner                                                        0
Herjavec                                                       1
John                                                           0
O'Leary                                                        0
Harrington                                                     0
Guest                                                          0
Details / Notes       Contingent on getting deal to bring to contine...
Valuation                                                  1e+07
Name: 483, dtype: object
```

It turns out that Zipz was not the most invested in company in this dataset. The sharks invested 5 million dollars in AirCar in comparison to the 2.5 million they invested in Zipz, but the valuation of Zipz is 2.5x greater than that of AirCar. This suggests that the most valuable companies may not always be the ones with the most invested money.

1.3 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 [10]: st["Number of Investors"] = (st.Corcoran + st.Cuban + st.Greiner +
                                      st.Herjavec + st.John + st.Harrington +
                                      st.Guest + st["O'Leary"])

         st["Amount Split by Investors"] = st["Amount"] / st["Number of Investors"]
         st["Amount Split by Investors"].fillna(0)

         ("Corcoran paid",
          st[st.Corcoran == 1]["Amount Split by Investors"].sum(),
          '', "Cuban paid",
          st[st.Cuban == 1]["Amount Split by Investors"].sum(),
          '', "Greiner paid",
          st[st.Greiner == 1]["Amount Split by Investors"].sum(),
          '', "Herjavec paid",
          st[st.Herjavec == 1]["Amount Split by Investors"].sum(),
          '', "John paid",
          st[st.John == 1]["Amount Split by Investors"].sum(),
          '', "Harrington paid",
          st[st.Harrington == 1]["Amount Split by Investors"].sum(),
          '', "Guest paid",
          st[st.Guest == 1]["Amount Split by Investors"].sum(),
          '', "O'Leary paid",
          st[st["O'Leary"] == 1]["Amount Split by Investors"].sum())
```

```
Out[10]: ('Corcoran paid',
          4912500.0,
          '',
          'Cuban paid',
          17817500.0,
          '',
          'Greiner paid',
          8170000.0,
          '',
          'Herjavec paid',
          16297500.0,
          '',
          'John paid',
          8154000.0,
          '',
          'Harrington paid',
          800000.0,
          '',
          'Guest paid',
          400000.0,
          '',
          "O'Leary paid",
          7952500.0)
```

Over the 6 seasons of Shark Tank in this dataset, Mark Cuban invested the most amount of money. He invested a total slightly greater than 17 million dollars, however, Herjavec nearly invested the same amount, only $1.5 million less. I obtained this answer by first creating a variable that calculated the total amount of number of investors. I then used this variable to divide the total amount of money invested by the total number of investors. From here, it was filling the NaN's with 0 and summing the amount of money each shark invested.

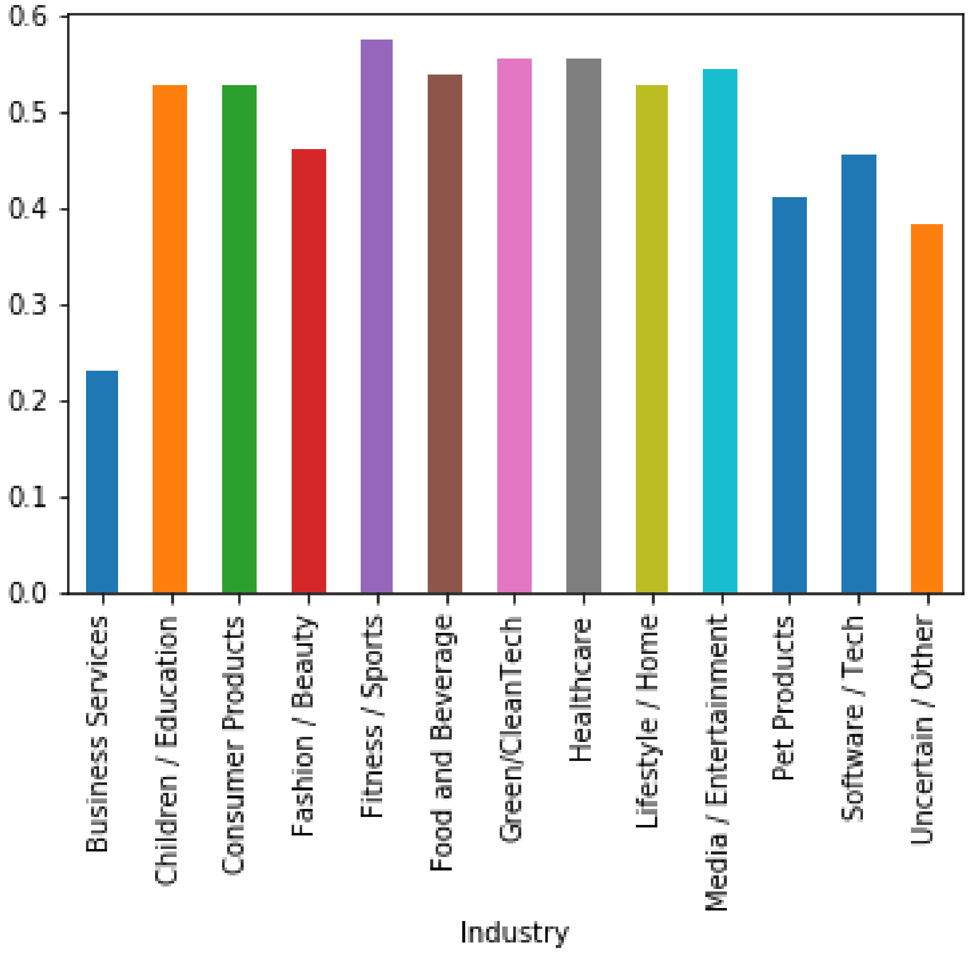
1.4 Question 3. Do the Sharks Prefer Certain Industries?

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

```
In [11]: st["Deal Binary"] = st["Deal"].map({
            "Yes":1,
            "No":0,
         })

         st.groupby("Industry")["Deal Binary"].mean().plot.bar()
```

```
Out[11]: <matplotlib.axes._subplots.AxesSubplot at 0x7f1dd4b9a5f8>
```



In order to answer this quesiton, I wanted to make a new variable in which the variable "Deal" was coded as 1's and 0's indicating either "Yes, a deal was made" or "No, a deal was not made" respectively. Creating this new variable, allowed me to easilly calculate the mean and would have also made calculating the sum or other quesitons easy as well.

The industry with the highest funding rate was "Fitness/Sports" with a rate of nearly 60%. However, from the distribution we can see that the "Food and Beverage", "Green/CleanTech", "Healthcare", "Lifestle/Home", and "Media/Entertainment" industries were invested in at rates similar to "Fitness/Sports". We also see that the sharks least invested in the "Businees Services" industry along with "Pet Products", and "Uncertain/Other".

1.5 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 PDF to PolyLearn.