# Shark Tank

[Shark Tank Clip](https://www.youtube.com/watch?v=5iKitGJeAZ4)

Let’s start :)

Question 0 - Getting and cleaning the data

1. Reading the “sharktank.csv”

*Hint:* Refer to Lab1 - Q1 for reading the file

1. Display the head of the file
2. Substitute NA fields for company columns with 0

*Hints:*

Select specific columns for substitution (only investors)

* [Select specific rows and columns from a data frame](https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.iloc.html) (iloc)
* [Select specific rows and columns from a data frame](https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.loc.html) (loc)

Fill NA fields with 0

* [Fill a dataframe NA fields with 0](https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.fillna.html)

1. Clean up “Amount” and “Equity” columns and cast them to numerical values

*Hints:*

Convert “Amount” and “Equity” to string and remove “$” and “%” from them respectively

Remove “,” from “Amount” values. Take a look at [this link](https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.replace.html), the last example helps!

Convert values in both columns to float

* [Convert string to float in pandas data frame](https://datatofish.com/convert-string-to-float-dataframe/)

Question 1 - Which company was worth the most?

1. Create a new data frame which includes all the companies with investment (Regarding “Amount” column)

* [How to filter a Pandas Dataframe based on Null values of a column?](https://cmdlinetips.com/2018/03/how-to-filter-a-pandas-dataframe-based-on-null-values-of-a-column/)
* [Reset\_index](https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.reset_index.html) of dataframe after removing null values

1. Find and drop the company with equity = 0 (why?)

* [Select rows from a pandas dataframe based on column value](https://www.interviewqs.com/ddi-code-snippets/rows-cols-python)
* [How to drop rows in Pandas Dataframe by index labels?](https://www.geeksforgeeks.org/how-to-drop-rows-in-pandas-dataframe-by-index-labels/)

1. Calculate the one with greatest value

* [Get the greatest value of a column in a Dataframe](https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.idxmax.html)
* Take a look at the description of the question and explain how we can calculate the total value of a company, given the invested *amount* and the *equity*.

Question 2- Which shark invested the most?

1. Create a new column with number of sharks who invested in each company (You may use [sum](https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.sum.html) method)
2. Remove companies with 0 investors and update the table (You may use [reset\_index](https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.reset_index.html) just like question 1)
3. Find the amount each shark invested for each company, given the total amount and the number of investors.
4. Use a bar plot to find which shark invested more.

Question 3- Do the sharks prefer certain industries?

1. Compute the amount sharks invested on each industry (group of companies)
2. Compute the total amount invested on each industry (group of companies)
3. Explain the preference of sharks based on the outcome of the first two items. You may use barplot to visualize the invested amounts.

# Evidence of Discrimination?

Question 1-

1. Open “ca\_dds\_expenditures.csv”
2. Use Pandas *pivot\_table* to summarize the data into two hispanic and white categories.

* [More information about pivot\_table](https://pandas.pydata.org/docs/reference/api/pandas.pivot_table.html)
* In this example you need two use three of pivot\_table’s input arguments, “index”, “values” and “aggfunc”.
* Think which column can be the *index* (Key’s to group by) and which can be the *value* (columns to aggregate).
* What is aggfunc? You can find examples on the given link.

1. Plot a bar or pie plot for expenditure over two categories, “Hispanic” and “White not Hispanic”.

Question 2-

* Same as the previous section plot expenditure over Hispanic and White for different age cohorts.
* You should add another input argument “column” in case you’re using pivot\_table

Question 3-

* Think and work on the followings:

1. Plot the distribution of age groups among White not Hispanic
2. Plot the distribution of age groups among Hispanic
3. Plot average expenditure over different age groups regardless of the race