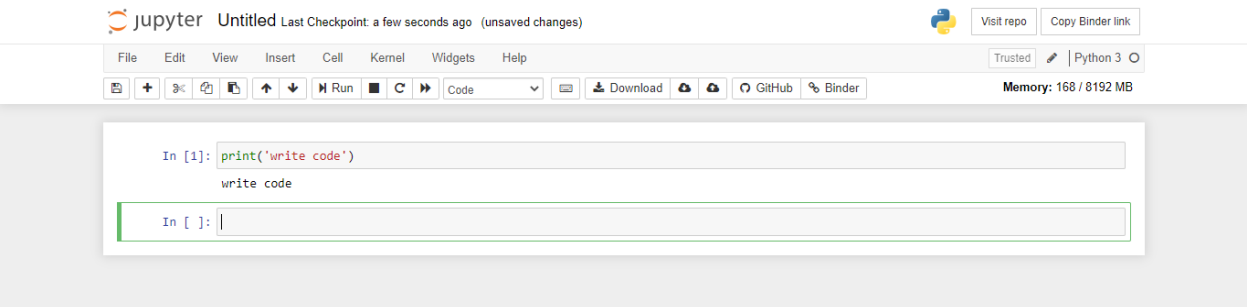
Udacity Data wrangle and analyze

Wrangle report

The time comes to finish this important capstone, an important demanding project in the Data analysis track made by Udacity.

We Rate Dogs is one of the famous Twitter accounts worldwide that has almost 9 million followers, launched in late 2015, it had a goal of rating dogs. It will be a source of the data that will be used in the project.



This project had to be done on Jupyter, an online notebook that processes python code in a neat way and can divides into pieces where one can see the output of last piece of code just written.

The project goal was to learn how to gather data from multiple resources, assess its technical validity as how much it follows the correct coding techniques, then improve any mistake found. Also two reports to explain the efforts produced had to be made.

It's great to be finally writing this document after loads of working hours and thousands of SHIFT+ENTERs -the command needed to load a cell in the Jupyter notebook-.

The project is setup to train good future data analysts, it requires gathering data from three sources. One was a csv file that I downloaded, second one was a list of tweets that had to be downloaded programmatically, and the third was source was Twitter itself by querying Twitter API. I enjoyed making a Twitter developer account and accessing Twitter using my mini-app to access my own secret token.

The three sources were read into 3 dataframes -some sort of list in programming- and copies of them were saved to in order to run any modified code on them again. Dataframes were called to assess their content and I unified the datatype of some of the columns and fixed some of the content in the rows. Data was later assessed and cleaned.

```
memory usage: 313.0+ KB
```

```
In [371]: #change columns to string
          df.name = df.name.astype(str)
          df.doggo = df.doggo.astype(str)
          df.floofer = df.floofer.astype(str)
          df.pupper = df.pupper.astype(str)
```

```
In [372]: type(df['tweet_id'].iloc[0])
```

```
Out[372]: numpy.int64
```

```
In [367]: df['name'].replace('the', 'None', inplace=True)
          df['name'].replace("an",'None', inplace=True)
          df['name'].replace("a",'None', inplace=True)
          df['name'].replace("none",'None', inplace=True)
```

```
In [368]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2356 entries, 0 to 2355
```