Chris Murdter

Springboard Capstone Project 1: Machine Learning with Satellite Imagery

The data for my capstone project is from Kaggle, a website containing many open data sets. This data set contains images extracted from Planet satellite imagery collected over the San Francisco Bay and San Pedro Bay areas of California. Included in the data set are 4000 80x80 RGB images labeled with either a "ship" or "no-ship" classification each given a value of 1 or 0 respectively. After downloading this data set the next appropriate step is to look it over and clean it. First I loaded the data in a jupyter notebook. The image files are .png files. The dataset contains four columns. The data column which is a list of numbers I assume are the color code. The labels column for each image which is a 1 or a 0 depending if the image has a ship in it or not. The locations columns has the coordinates for the image's center point and the last column is the scene ID which contains the individual IDs for the images. The data columns look very useful and the images that I randomly looked through are very clear.

To find any missing values I used the isna() panda function and applied it to every column. I found that there are no missing values from this data set. This is very useful and most likely because the data set has already been cleaned before being posted to Kaggle. I wasn't sure on what to look for with outliers. I wanted to include all images to have a big data set to train my machine learning algorithm. I think maybe if there was an image that had some dead/hot pixels then that should be looked at but I wasn't sure how to go about this. Overall, I think this data set is ready to be implemented for my capstone project.

Below is a picture of my code:

```
In [1]:   1  import pandas as pd

In [3]:   1  data = pd.read_json('shipsnet.json')

In [4]:   1  data.head()
```

Out[4]:

	data	labels	locations	scene_ids
0	[82, 89, 91, 87, 89, 87, 86, 86, 86, 86, 84, 8...	1	[-118.2254694333423, 33.73803725920789]	20180708_180909_0f47
1	[76, 75, 67, 62, 68, 72, 73, 73, 68, 69, 69, 6...	1	[-122.33222866289329, 37.7491755586813]	20170705_180816_103e
2	[125, 127, 129, 130, 126, 125, 129, 133, 132, ...	1	[-118.14283073363218, 33.736016066914175]	20180712_211331_0f06
3	[102, 99, 113, 106, 96, 102, 105, 105, 103, 10...	1	[-122.34784341495181, 37.76648707436548]	20170609_180756_103a
4	[78, 76, 74, 78, 79, 79, 79, 82, 86, 85, 83, 8...	1	[-122.34852408322172, 37.75878462398653]	20170515_180653_1007

```
In [5]:   1  data.tail()
```

Out[5]:

	data	labels	locations	scene_ids
3995	[126, 122, 124, 138, 165, 186, 195, 199, 203, ...	0	[-122.08693255500694, 37.77781408256089]	20170815_180821_102d
3996	[130, 134, 139, 128, 117, 126, 141, 147, 142, ...	0	[-122.10549691828378, 37.76946626247702]	20170730_191230_0f21
3997	[171, 135, 118, 140, 145, 144, 154, 165, 139, ...	0	[-122.48298739296371, 37.684929808845375]	20161116_180804_0e14
3998	[85, 90, 94, 95, 94, 92, 93, 96, 93, 94, 94, 9...	0	[-122.29028216570079, 37.71632091139081]	20170211_181116_0e16

```
In [14]:   1  data.isna()
```

Out[14]:

	data	labels	locations	scene_ids
0	False	False	False	False
1	False	False	False	False
2	False	False	False	False
3	False	False	False	False
4	False	False	False	False
5	False	False	False	False
6	False	False	False	False
7	False	False	False	False