Assignment 5: Data Cleaning

Machine Learning

Fall 2021

Learning Objectives

- Become familiar with some basic data cleaning tools
- Prepare for your projects, which may involved using un-cleaned data

1 Data cleaning

1.1 Introduction

Data cleaning is an important skill. For your final project, you may choose to use a dataset that has not been nicely cleaned and packaged, like many of the datasets we've worked with in this class have been. You could get a Ph.D. in this, it's a broad topic. In this assignment, we hope to give you a handful of tools that will help you in the future.

In Carrie's opinion, you have two choices– you can clean your data before you work with it, or your can jump to analysis, realize your results make no sense, and data clean after you've wasted a lot of time. (Or, worst case, present results that are nonsense because you didn't realize your data needed to be cleaned. Let's avoid that one.)

Please read this introduction to data cleaning. When the author talks about "throwing a random forest at the data" he's talking about applying a machine learning technique, not about throwing a randomly picked group of trees, brush, moss, and wildlife at something.

I have a philosophical disagreement with this author. What he describes might be an okay practice for analyzing a consumer database for a company. But it's bad practice for science. In science you want to track down what exactly is going on in each instance. It's not enough to know why a data point is incorrect (for example, a negative height). You need to know why you got a negative height measurement, and once you understand that why, you understand what you can do about it. In science, you don't want to get rid of surprising results– after all, you don't want to miss out on a Nobel Prize because you discarded a data point you didn't understand!

1. Carefully consider the following table of moon crater data. This is adapted from a larger dataset[1] you'll be working with for your final project. What data values are suspicious? List them and explain why. You are encouraged to look up information about the moon to help you with this.
   CRATER_ID = identifier assigned to crater by scientist doing study
   LAT_CIRC_IMG =Latitude of center of crater, degrees.
   LON_CIRC_IMG = Longitude of center of crater, from 0 to 360 degrees.
   DIAM_CIRC_SD_IMG = Standard deviation of kilometers of the fit residuals. Each manual rim point's distance from the crater center was calculated and subtracted from the best-fit circle's radius, and this value is the standard deviation of those differences.

Followi is from AstroStat Carrie TBD add census data, make other tweaks to follow ML conventions

CRATER_ID	LAT_CIRC_IMG	LON_CIRC_IMG	DIAM_CIRC_IMG	DIAM_CIRC_SD_IMG
00-1-000000	-19.83040	264.7570	940.960	-21.31790
00-1-000001	44.77630	182.0995	249.840	5.99621
00-1-000002	57.08660	378.6020	3474.891	88.94687
00-1-00000i	1.96124	230.6220	55.762	1.18190
00-1-000004	-49.14960	230.6220	-654.332	17.50970

1.2 Larger Data Sets

What if you have too many values in your data to manually inspect each one, as you did in the previous exercise? Another tool you can use is to plot each value and see if those plots make sense to you. Let's take a minute to think what we expect for a few histograms. *If you are having trouble with any of these, please go outside and LOOK AT THE MOON. Or, at least look at a photo of the Earth-facing and not Earth-facing ("dark") sides of the Moon.*

2. Imagine you were to plot a histogram of LAT_CIRC_IMG. How would you expect it to look?
3. Imagine you were to plot a histogram of LON_CIRC_IMG. How would you expect it to look?
4. Imagine you were to plot a histogram of DIAM_CIRC_IMG. How would you expect it to look?

Now, let's test your results. Work in the same jupyter notebook as you did for the previous question. Please format your notebook nicely to make it easy to grade.

We are now going to work with a portion of the Robbins Moon database. The full database has 1,296,796 craters, which is a little too many for your computers to handle easily, so I've made a smaller version by getting rid of any craters smaller than 5 km in diameter. You can get that file here.

You'll be using Pandas. You'll first need to read in the file. If you're using Google Collab, you'll need to tell it where to look for the file, here's some example instructions.

You'll need to import the Pandas library and read in the file, something like this:

```
import pandas as pd
craters=pandas.read_csv(’AstroStats_Robbins_Moon.csv’, sep = ’,’)
```

Anytime I read in a file, I always like to check the output. This quick check, that prints the data column headers and the first 5 values, lets me learn about the data.

```
print(craters.head(5))
```

5. Check how many craters are in the dataset by typing `len(craters)`. Report the answer.
6. Plot a histogram of LAT_CIRC_IMG.
7. How does this compare to what you were expecting? Are there any obvious problems with the data?
8. Plot a histogram of LON_CIRC_IMG.
9. How does this compare to what you were expecting? Are there any obvious problems with the data?
10. Plot a histogram of DIAM_CIRC_IMG. Hint: For this plot to be useful/meaningful, you may want to use the xlim parameter on matplotlib.
11. How does this compare to what you were expecting? Are there any obvious problems with the data?