*95-885 Data Science and Big Data, Fall 2019*

# A2 Introduction to Pandas

**Due 11:59pm Thursday, September 12**

## Learning Objectives

Explore some of the capabilities of Pandas to extract information from a data set. This assignment is in multiple parts. In each part you will use Pandas to perform a data analysis task we've discussed in class. For full credit, ensure that you are solving these tasks in the *pandorable* way, in particular you do not want to pull data out of a data frame (or series) and iterate with a loop.

### Data Sets

The data sets needed for part I and II have been placed on Github at https://github.com/ProfRaja/67364/tree/master/data/a2 . In your notebooks for this assignment, right after you load the needed libraries (Pandas, Matplotlib etc.). Place code equivalent to the below in a single cell:

```
UseGitHub = False # True
if UseGitHub:
   prefix = 'https://raw.githubusercontent.com/ProfRaja/67364/master/data/a2/'
else:
   prefix = ''  # load locally

expeditions_xls = prefix+'Expeditions.xls'
titles_csv = prefix+'titles.csv'
cast_csv = prefix+'cast_4_1960.csv'
release_dates_csv = prefix+'release_dates.csv'
```

Based on the value of the flag `UseGitHub` your notebooks should load data from your local directory or from github. During development, download the data sets and work with a local copy by setting the variable `UseGitHub` to be `False`. During assessment we will set `UseGitHub` to `True` so that data is pulled directly from GitHub.

## Part I: Pivot Tables

a. Review the contents of `Expeditions.xls` and create the following pivot table with Pandas. Note that the below table was produced with Word. Your Pandas table will cosmetically look slightly different. Each entry in the pivot table is the sum of the revenue.

| | **How** | | | |
|---|---|---|---|---|
| **Where** | Catalog | Store | Web | Total |
| London | | | | |
| New York | | | | |
| Paris | | | | |
| Sydney | | | | |
| Tokyo | | | | |
| Total | | | | |

## Part II: Data Cleansing and Plotting

Benford's law is a fascinating law that describes the frequency of distribution of the leading digits in many real life data sets. Amongst other uses BL has been used to detect fraud in accounting. Wikipedia has a nice article on BL. https://en.wikipedia.org/wiki/Benford%27s_law

This part has the following components:

a. In the Wikipedia article you will see a formula for P(d) which gives the probability of the occurrence of d as the leading digit. Using this formula, plot a line graph depicting the probability distribution. In the second figure of the Wikipedia article you will see a red line. The line you draw should have that form.

Now let us verify Benford's law on a real life data set.

b. The Wikipedia article [https://en.wikipedia.org/wiki/List_of_tallest_buildings_and_structures](https://en.wikipedia.org/wiki/List_of_tallest_buildings_and_structures) has a table of heights of tallest buildings / structures in the world. Using Pandas `read_html`, read the table under the section "Tallest Structure by Category" . Note that `read_html` will read all the tables on a page and return a list. Plot the frequency of occurrence of the leading digit in height (feet) on the same figure of as the bar graph.

c. Repeat (b) but this time using the height in meters.

You will draw a single plot with three quantities: (a) the ideal Benford line graph (b) a bar graph for the frequency distribution of height in feet (c) a bar graph for the frequency distribution of height in meters. Your final graph will be a pictorial representation of the same information in the table in the section labeled "Example" in the Benford's law article [https://en.wikipedia.org/wiki/Benford%27s_law#Example](https://en.wikipedia.org/wiki/Benford%27s_law#Example). Label the X and Y axis appropriately and also create a meaningful legend. [Note that the Wikipedia article on Benford's law uses a dot plot; yours will be a line graph.]

Note that in the final figure you have 3 plots. One of which is a curve which has a x-scale and a y-scale. For bar graphs we only have labels (nominal values) for the x-axis (no scale). Explore on the web how to superimpose a line graph and a bar graph on a single figure.

## Part III: Simple question from the IMDB

Examine the three data files and their constituent fields

| File | Fields |
|---|---|
| `cast_04_1960.csv` | `title, year, name, type, character, n` |
| `release_dates.csv` | `title, year, country, date` |
| `titles.csv` | `title, year` |

Most of the fields should be self-explanatory. The number n in `cast.csv` refers to the importance of the character (which also corresponds to the sequence in the credits). Following are sample lines from cast.csv which illustrate the role of `n`:

```
The Godfather,1972,Marlon Brando,actor,Don Vito Corleone,1
The Godfather,1972,Al Pacino,actor,Michael Corleone,2
The Matrix,1999,Keanu Reeves,actor,Neo,1
```

This data has been extracted from the IMDB data set by a script written by Brandon Rhodes. The full cast.csv file is more than 300M (it has information on all movies and all roles). I've trimmed it to have only the top 4 roles and movies from 1960 onwards.

Using Pandas answer the following questions:

a. What are the ten most common movie names of all time? (list in descending order based on frequency; order of ties don't matter).

b. Produce a bar plot of the number of films that have been released **each decade** over the history of cinema. Use `value_counts()`. Use the data in `release_dates.csv`

c. Produce the same plot of (b), but this time using `groupby()`.

d. In which months are films with Tom Cruise released in the **USA**? Plot a bar graph for your result.

e. Write a query of your interest and answer it with Pandas. Requirements (a) you need to merge at least two of the given data sets and (b) produce some meaningful graph.

I will give a small prize for the most interesting query produced for (e) as determined by the TAs and myself. So do apply your creative juices. ☺

As with assignment 1, enter a short description of the various parts (5.i, 5.ii etc.) in markdown cells followed by your analysis using Pandas.

## What to submit

The deliverable for this assignment is split into multiple files so as to facilitate independent development.

For parts 1-3, prepare your solutions in separate Jupyter notebooks, one for each part. Call them `a2-sraja-part-1.ipynb, a2-sraja-part-2.ipynb, and a2-sraja-part-3.ipynb.` Naturally, replace 'sraja' with your AndrewID ☺. Zip all files and submit the zip bundle to Canvas. As directed at the beginning of this write-up, before making your final submission, ensure that these notebooks are able to load data from the github repository.