**EXPLORATORY DATA ANALYSIS WITH PYTHON**

Pandas is an important package in python that helps dealing with structured data. Familiarize yourself with some basic tricks in pandas using the following Pandas Cheat Sheet.

1. **Pandas introduction:** <http://nbviewer.jupyter.org/github/pybokeh/ipython_notebooks/blob/master/pandas/PandasCheatSheet.ipynb>
2. **MovieLens Project: This project will help us practice what we learnt in the above exercise and also continue learning the best ways to visualize various aspects of the data given to us.**

Download movielens dataset from here:

<http://grouplens.org/datasets/movielens/1m/>

Create an Ipython Notebook and explore the above data by following these steps:

(*note: there ARE solutions available online for this, but please refrain from copy/pasting. The idea is for you to learn how to explore datasets so you can create a successful project at the end of this workshop)*

1. Get the top 10 genres that have the highest ratings
2. Plot distribution of age among users

<https://www.analyticsvidhya.com/blog/2015/04/comprehensive-guide-data-exploration-sas-using-python-numpy-scipy-matplotlib-pandas/>

1. How is the data distributed by release dates of the movies?
2. How do the ratings differ by location of the users? Are west coast (where Hollywood) users more critical than others in the US? What other dataset/prior knowledge did you have to use to answer this question?
3. Get the average movie rating by age & sex of the users

<http://stackoverflow.com/questions/39420633/how-you-calculate-the-average-rating-per-genre-in-python>

<https://codedump.io/share/JX02sELCVcsl/1/how-you-calculate-the-average-rating-per-genre-in-python>

<https://github.com/justmarkham/DAT3/blob/master/code/05_pandas_class.py>

<http://beyondvalence.blogspot.in/2014/09/python-and-pandas-part-2-movie-ratings.html?m=1>

<http://www.gregreda.com/2013/10/26/using-pandas-on-the-movielens-dataset/>

1. Get the top 20 movies preferred by (a) Men but NOT women (b) Women but NOT men

<http://googleweblight.com/?lite_url=http://www.michaeltsmith.org.uk/ipython/Lesson%25201,%2520(detour)%2520Python,%2520pandas%2520and%2520Movie%2520Ratings%2520(MovieLens).html&ei=9Fi4qZKU&lc=en-IN&s=1&m=933&host=www.google.co.in&ts=1482864347&sig=AF9Nedk2gLbWPWuD7fg9gKMOJyYC7UAfhw>

1. How would you visualize the above disagreement between Men and Women
2. What other interesting questions can you think of asking this dataset.
3. **Top 500 Indian Cities. An important part of being a data scientist is being able to look at our and other’s analysis with a critical eye. How far can we trust what we see in the data?**
4. Download the dataset from Kaggle: <https://www.kaggle.com/zed9941/top-500-indian-cities>
5. Visualize the top 50 cities by female:male literates. What other ‘factors’ in the data could’ve affected this ratio and how would you normalize for it?
6. Take a look at this visualization project on this dataset: <https://www.kaggle.com/anirudhsekar96/d/zed9941/top-500-indian-cities/correlation-between-sex-ratio-and-literacy-rate/code>.
7. What can you say about the hypothesis?

(b) How would you statistically prove that the correlation is ‘significant’?

(c) Think of two other features in this dataset whose correlation might be interesting?