**Assessment - Python for Data Science Level 1**

Note: Work on at least three from the four given

1. **Title:** **Merging csv files from movie lens dataset**

**Background information:** The movies data set provides a solid foundation to Data Science learners for doing data analysis. This data file can be downloaded from the famous movie lens webpage (given in the link below). The three files are movies.csv, users.csv and ratings.csv. The users.csv contains information about the user’s gender, age and preference. The movies data set contains information about the movie name and genre. The ratings.csv file contains information on the ratings given by the users

**Purpose of the case study:** The purpose of this exercise is to understand how to do a meaningful merge twice – first, merge two data files and then merge the resulting file with a third data file

**Problem description:** Do a merge of all the three files movies.csv, users.csv and ratings.csv in an appropriate way as one csv file namely movies\_integrated.csv by using a common column id present in each data file

**Dataset:** <https://grouplens.org/datasets/movielens/>

**Download and unzip your data files:** Use the following files, movies.dat, users.dat and ratings.dat, copy the .dat files as .csv files

1. **Title:** **Comparison of movie ratings given by Ladies and Gentlemen**

**Background information:** The movies data set is downloaded in the previous case study. If you consider the ratings given for a product or food ordered in a restaurant or for movies the ratings vary depending upon the independent variables like gender, age and place. The ratings also vary for the same product in different stores or web sites. In this case study we consider the ratings given by users for movies and check if there are any rating differences

**Purpose:** The purpose of this exercise is to find out if there is any difference in movie ratings between gentlemen and ladies after looking through all the nearly 1 million ratings available in the dataset

**Problem description:** We have created the new merged movies\_integrated.csv from the previous case study. Use that merged file to find if there is any difference in the way ladies and gentlemen rate the movies. If there is a difference in rating between genders, then show it in your data analysis and also through data visualization

**Dataset:** <https://grouplens.org/datasets/movielens/>

1. **Title:** **Data analysis of onion retail prices from Govt of India data home page**

**Background information:** Onion retail prices dataset is released by Govt of India in its data home page <https://data.gov.in/> It has near 1,27,500 samples of data with each day’s onion retail price. The prices range from Rs 2 to Rs 100 and prices are provided from 1997 to 2015. It has date, city name, commodity name and price as the columns. We need to download the dataset and do basic analysis on data, and find if there is any correlation in the prices

**Purpose:** The purpose of this exercise is to perform a first level of data analysis using correlation function in pandas and also find the places maximum and minimum onion prices occurred.

**Problem statement:** The onion\_prices.csv file is attached here. Find the unique names of cities where maximum prices have occurred. Also find the if there is any correlation between change in the onion price of one city to the onion price in any other city, use the retail onion price given in the link below

**Onion retail dataset link :**

<https://data.gov.in/resources/daily-retail-price-onion-upto-april-2015>

1. **Title:** **Data analysis and classification done on the USA university admission data**

**Background information:** Every year many thousands of international students get admitted to universities in the United States. We would like to analyse the GRE and GPA scores needed to get admitted to such universities. We have a very popular dataset (which is attached at the end of this document) called university admission dataset. This dataset has GRE score, GPA score, University Prestige and Selection – YES – NO as columns. It will be useful to build a model to predict if a given student will get admission in the university of his choice

**Purpose:** The purpose of this exercise is to perform a first level of data analysis for the university admission of international students in the United States and to classify selection of those students as “YES” or “NO”

**Problem statement:** Please find the attached university\_admission.csv file here. Perform simple data analysis and visualization using pandas and sea born libraries. Display using a histogram, bar chart and box plot - details of students selected based on GPA, GRE and prestige of the institution. Show the analysis and visualizations using pandas.groupby and pandas.crosstab functions. Build a prediction model using logistic regression algorithm to predict if the applicant will get admission or not

**Dataset:** Attached here in this case study document