EE353 / EE769 Introduction to Data Science and Machine Learning
July-Nov 2024, IIT Bombay
Assignment 1: Exploratory and Descriptive Data Analysis

Instructions:

- Perform all analysis in an ipython notebook environment, such as Google Colab.
- For every line or block of code where you copied the base code from somewhere, cite your source with a source number, such as #[1], #[2], #[3], etc. This should be done even if you modify the base code substantially.
- At the bottom of your assignment, give a reference to the source numbers, e.g. "2. https://stackoverflow.com/questions/17193850/get-column-by-number-in-pandas", or "3. ChatGPT prompt: How to get columns by numbers in pandas", or "1. Discussion with classmate Mukesh Adani Roll No. 213414".
- Make copious use of text cells and inline comments in code cells to explain your intent, observations, and next steps. Without these, your assignment will not be graded, and it will be assumed that you did not understand your own code.
- Record a video shorter than 10 minutes of you demoing your assignment using a screenrecorder and webcam. Host the video in a shared drive (MAKE SURE TO GIVE ANYONE WITH THE LINK ACCESS. WE WILL NOT ASK YOU.) Include the link at the top of your ipynb file in a comment, and submit ONLY the ipynb file.
- Submit your assignment on Moodle before the deadline. After the deadline, there may be a late penalty, but submit it on Moodle only. Do not email or submit on Teams.

Questions:

1. Check out the City of Los Angeles public data sources and test the hypothesis that the statistics of "affordable housing projects" (government housing for low-income people) in a ZIP code has a relation to the health inspection scores of the restaurants in that ZIP code.
   a) Download csv files from:
      i. https://catalog.data.gov/dataset/restaurant-and-market-health-inspections
      ii. https://catalog.data.gov/dataset/hcidla-affordable-housing-projects-list-2003-to-present
   b) Perform EDA on the two files: [4]
      i. Check if the data types are as expected, else convert them
      ii. Check for missing values, then decide to either remove those rows, or fill an imputed value
      iii. Check for unexpected entries in certain columns. Correct them if necessary and feasible.
      iv. Plot some graphs to understand the data
   c) Summarize each file by ZIP code using SQL: [2]
      i. Ensure the right type of summarization (sum, mean, max etc.) for the other variables
   d) Join the files using SQL by ZIP code: [2]
      i. Ensure that the ZIP codes are in compatible formats and lengths
      ii. For each ZIP, get the predictor variable from the housing projects file, and potential predicted variables from the health inspections file
   e) Formulate and test the hypothesis: [2]
      i. Formulate a reasonable alternative hypothesis
      ii. Formulate a null hypothesis
      iii. Select an appropriate test and significance level
      iv. Perform the test and decide if the null hypothesis should be rejected and alternative hypothesis should be accepted
2. Open-ended: Find some interesting data from Indian government data portal https://www.data.gov.in and perform EDA, derive some insights using graphs, and perform a statistical test for an interesting hypothesis. No need to use multiple files for this question, unless you want to do the extra work for your own learning. [4]