CSCI - 5901 - The Process of Data Science - Summer 2019
Assignment 1

Due date: July 2nd, 2019 11:59:59 pm.
The submission must be done through brightspace.
Teams of 2 students

- Cite any and all resources used.
    - books, websites (other than documentation) like stackoverflow.
- I will use plagiarism tools to detect any type of cheating and copying (your code and PDF).
- Write all of your comments and explanations in the code as a text cell.
- Two sides of Data Science (and your mark):
    - Technical // does it work?
        - Quality of code (documentation, naming)
        - Does it work
    - Conceptual // what did you find from the data?
        - Quality of the data insights
        - Quality of the discussion text
    - Questions will be marked individually. Their weights are shown in parentheses after the question number.
- Your submission is a single Jupyter notebook and a PDF (With the compiled results generated by your Jupyter notebook). Filename should be **A1-<your_name1>-<your_name2>.jpynb** and **A1-<your_name1>-<your_name2>.pdf**. Please upload to Brightspace. Please include your B# in your Jupyter notebook and PDF.
- **Forgetting to submit both files results in 0 markings for both students.**

Link for the dataset.
https://www.kaggle.com/himanshupoddar/zomato-bangalore-restaurants

1. **Introduction. (2 marks total)**

    a. Explain the dataset with your own words. Focus on the attributes description. (2 marks)

2. **Data pre-processing and understanding. (28 marks total)**

    a. Load the data. (3 marks)
    b. Explore the data. Plot the distribution of the attributes (frequency). What trends can you find in your data? Are there attributes that are useless at this point? (10 marks)
    c. Are there restaurant duplicates in the data? Detect and if there is, clean it. (5 marks)

d. What is the neighborhood with the highest average rating? What are the major characteristics of this neighborhood (e.g., type of restaurant, type of food they offer, etc). (10 marks)

3. **Build the best model you can that forecasts the approximate cost of a meal for two people using the attributes location, rating, restaurant type, and cuisine. (70 marks total)**
   a. Explain what is the task you're solving (e.g., supervised x unsupervised, classification x regression x clustering or similarity matching x etc). (5 marks)
   b. What models will you choose? Why? (5 marks)
   c. Which metrics will you use to evaluate your model? (5 marks)
   d. How do you make sure not to overfit? (5 marks)
   e. Build your model and verify how it performs (using the metrics you have chosen in Section 3(c)) in your training data. Justify which evaluation approach you are using?(Out of sample validation or Cross-validation). Use a plot to justify your findings. How good is your model? (10 marks).
   f. Test your model in your testing set and evaluate its performance. Use a plot to justify your findings. How is it performing compared to your training data? (15 marks)
   g. Can you tune your model to perform better? Explain the technique you're using and justify why it is improving your results. (25 marks)
   h. **(Bonus)** Use relief feature selection to improve your model.(10 marks)