ECON861 Data Analysis(Spring 2021) Homework 1

Due: 2th March 2021 (23:59)

Name: _____

UID: _____

---

1. (100 points) Train a model which predict the price of the Airbnb listing in New York City.

   - Download the crime dataset from the New York City Airbnb data https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data/?select=AB_NYC_2019.csv. It is a CSV file (approximately 7 MB)
   - This dataset contains data of the Airbnb listings in New York City, which includes the location, room type .....etc.
   - You do not need to use ALL the data. You should use only a subset of it. Some of the observations are obviously problematic, you may need to drop them before you do your analysis.
   - Train a model to predict the price of the Airbnb listing.
   - In your repository, there should be:
     - One (or more) python file, which contains your machine learning code
     - A README.md file, which contains instructions of how to run the python files.
     - A document which contains a three pages (maximum) report of your model. You should write down the results, how you choose the parameters, how you subset your dataset, any limitations of your model.
   - The dataset contains a lot of information you can use to train your model. You may need to use Ridge/Lasso regression to limit the complexity of your model.
   - The latitude and longtitude give you a lot more information than merely a bunch of numbers. Be creative. In this exercise, you do NOT need to use external information (for example map information like amenities around the location of the Airbnb). However, you can use the latitude and longtitude to calculate the distance from a nearby Airbnb listing, which could be useful.
   - You should hand in via Github. Please set your Github repository to a private repository and invite me to be your collaborator. Your repository should include your code and also your write up. However, DO NOT commit your dataset into the repository. I will download the dataset from the link to test run your code.
   - Again, do not print out your answers and hand in a hardcopy to me, you will fail this class if you do that.
   - There is no perfect answer for this task. You need to make simplifying assumptions to make sense of the data. Feel free to simplify the problem and write down your reasons in the write-up document.
   - The deadline is 23:59pm 2th March 2021. If you choose to hand in late (between the 3rd March 2021 and the 9th March 2021), there will be a 20% reduction in homework score.