DS 5230: Unsupervised machine learning and data mining (Fall 2020) Roi Yehoshua
Student name: (Due) October 6, 2020

Data Analysis and Preprocessing

The file housing.csv contains data on median house prices in California districts, derived from the 1990 census data. The data set contains 20,640 rows, one row per district (house block). Each row contains the following features:

- `longitude`: how far west a house is; a higher value is farther west.
- `latitude`: how far north a house is; a higher value is farther north.
- `housing_median_age`: median age of a house within the block; a lower number is a newer building.
- `total_rooms`: total number of rooms within the block.
- `total_bedrooms`: total number of bedrooms within the block.
- `population`: total number of people residing within the block.
- `households`: total number of households (a group of people residing within a home unit) within the block.
- `median_income`: median income for households within the block (measured in tens of thousands of US dollars).
- `median_house_value`: median house value for households within the block (measured in US dollars).
- `ocean_proximity`: location of the house with respect to the ocean. Can have one of the following values: NEAR BAY, NEAR OCEAN, <1H OCEAN, INLAND, ISLAND.

The objective in this data set is to predict the median house value in a given district based on the values of the other features.

Answer the following questions:

1. What is the data type of each feature? (ordinal/nominal/interval/ratio, discrete/continuous)
2. Display summary statistics of the data. What can you learn from it on the data?
3. Compute the correlation between each feature and the target `median_house_value`. Which features have strong correlation with the target?
4. Use data visualization tools to explore the data set. Display at least three different types of graphs.
5. What type of problems can you detect in the data set? Name at least three different problems.
6. Clean the data set using the data preprocessing techniques discussed in class. Show a sample of the data set before and after the cleaning.
7. Extract at least two new features from the data set that have strong correlation with the target feature.