Exercises

Module 4

Clustering & Unsupervised Learning

Review Questions

Which is the correct order for the code to run without errors?

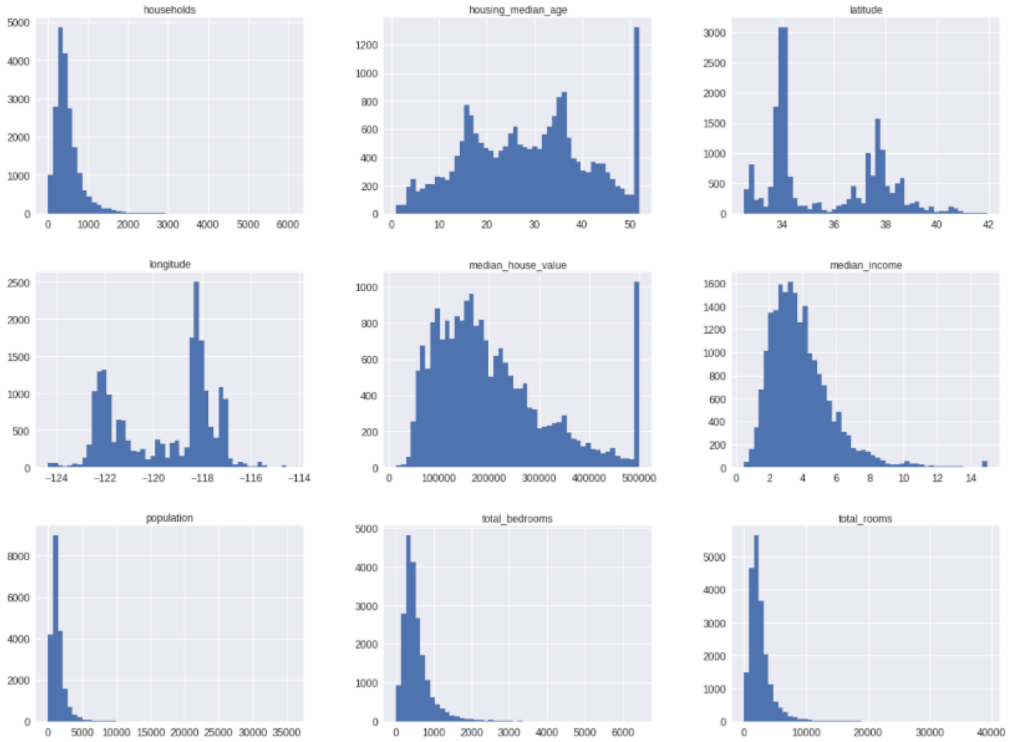
A.
```
lin_reg = LinearRegression()
lin_reg.fit(X_train, y_train)
```

B.
```
lin_reg.fit(X_train, y_train)
lin_reg = LinearRegression()
```

Review Questions

There are 9 histograms plotted below, why is it important to plot these?

A. To gain deeper understanding of features in the dataset prior to building the model
B. To visualize the distribution of values for each feature in the dataset
C. To check if the data is skewed or not
D. To check if a feature has only one value for all observation, so we can drop it
E. All options are correct



Review Questions

Which of following uses sklearn's standard approach to split into train and test sets?

A.
```
from sklearn.model_selection import train_test_split, train_set
test_set = train_test_split(housing, test_size=0.3,
                            random_state=59)
```

B.
```
train_set, test_set = train_test_split(X, y, test_size=0.3,
                                       random_state=59)
```

C.
```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y,
                                                    test_size=0.3,
                                                    random_state=59)
```

D.
```
np.random.seed(59)
train_set, test_set = split_train_test(housing, 0.3)
```

Review Questions

Which statement below is correct?

A. Supervised learning is when there is a supervisor helping with training of the model
B. To group 1000 customers into 3 categories without names, we would use a clustering algorithm
C. Supervised learning is only used for regression
D. Clustering is the most commonly used supervised learning algorithm

Exercise 1

Go to Quercus and download *W04_EX1_clustering.ipynb*

1. Run and review the code in the **Model** section.

   TIP: Don't focus on the creation of the synthetic data, rather focus on how the data it is used to build a clustering model.

2. Try tweaking the hyperparameters:
   - true_k
   - dim_features
   - num_points

Exercise 2

1. Run the code in the **Tune** section and share your inertia results with your neighbour. Are they the same?

2. Complete the **Model Using Real Data** section by training a k means model to the data.

3. Which feature pair best separates the data?

   TIP: Using pandas.plotting.scatter_matrix may be useful here.

4. Does k = 3 get the best performance metrics?