**AD450 Data Science Development (Winter 2021)**

**Final Exam**

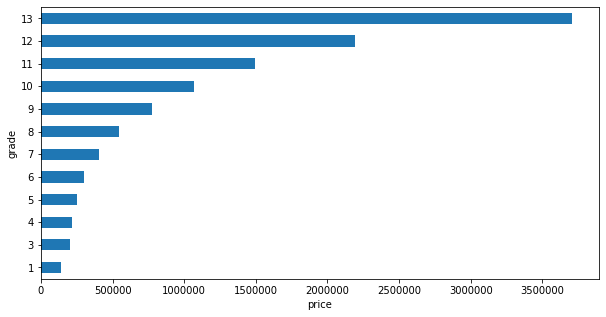
Instructions:

* Answer all the questions.
* Include all the group members in the first markdown cell of the jupyter notebook file.
* Use jupyter notebook to write your codes. Write the question number and the part in the comments. Answers for each part should be in separate blocks of codes.
* Include output of your codes in the file.
* Save your work in a pdf file.
* Submit your file on Canvas.

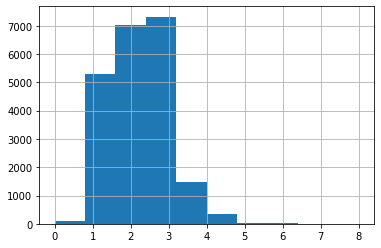
Q.1

(a) Read in the dataset “kc\_house\_data.csv”

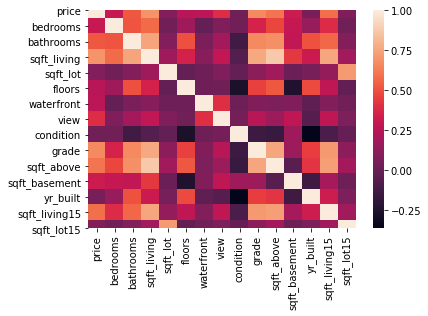
(b) First, use groupby to find the average price for each grade of houses. Then, recreate the following bar plot.



(c) Plot a histogram on “bathrooms”



(d) Create a correlation matrix and then plot the following heatmap.



(e) Check for any missing values in the dataframe.

(f) Set y = price and all other features to be X. Split your data into training set and testing set.

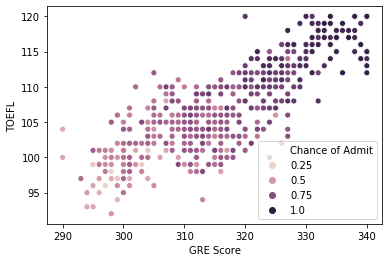
(g) Set up a LinearRegression Model. Train your model with training set.

(h) Show the coefficients on all features using model.coef\_

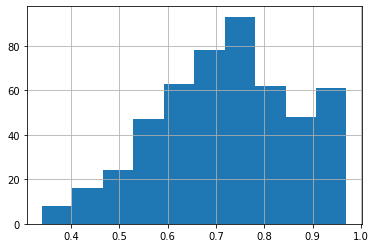
(i) Predict housing price using testing data. Compute the root mean squared error of the prediction.

Q.2 Read in the dataset “Admission\_Predict.csv”, and name it as df.

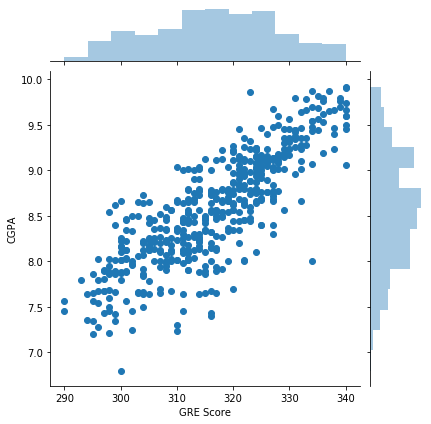
(a) Recreate the following scatter plot where x = “GRE score”, y = “TOEFL score”, hue = “Chance of Admit”

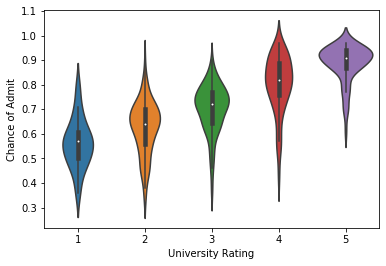


(b) Recreate the following histogram for “Chance of Admit”



(c) Recreate the following jointplot between x = “GRE Score” and y = “CGPA”



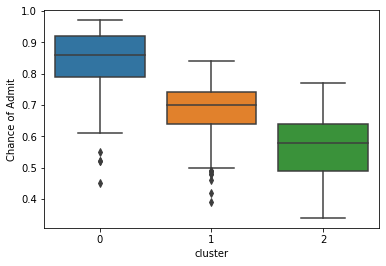
(d) Create a violin plot on “Chance of Admit” for each group of “University Rating”

(e) Create a new dataframe, df2, by dropping the column “Chance of Admit”

(f) In the new dataframe, df2, group the data into 3 clusters. Use K-Means clustering.

(g) In the new dataframe, df2, create a new column “cluster” by predicting the cluster for each row.

(h) Put the column “cluster” in the original dataframe, df. Create the boxplot on “Chance of Admit” for each cluster.



Q.3 In this question, you will train a random forest classifier to recognize a digit in an image. I posted the data file already. You can also download it using the following link:

<http://ufldl.stanford.edu/housenumbers/train_32x32.mat>

(a) Save the file 'train\_32x32.mat' in your working directory. Then, use the following codes to create two arrays, X (the data) and y (the target).

train\_data = scipy.io.loadmat('train\_32x32.mat')

X = train\_data['X']

y = train\_data['y']

(b) Show the first 10 images. Also, put the corresponding target as the xlabel. It should look as follows



Hint:

To display image, use

ax[i].imshow(X[:,:,:,i])

To set the xlabel, use

ax[i].set(xlabel=y[i])

(c) Reshape X and y using the following codes

X = X.reshape(X.shape[0]\*X.shape[1]\*X.shape[2], X.shape[3]).T

y = y.reshape(y.shape[0], )

(d) Split the data into training set and testing set.

(e) Set up a RandomForestClassifier model and train the model with the training data.

(f) Making predictions with the testing data. Find the accuracy score.