Rohit Jyotiba Chougule, 19200240

Assignment No. 3

Question 1:

1. The file '/specs/gpa_question1.csv' dataset contains data related to sample of students, as mentioned in the question the count attribute will not be included in the rule generation, hence we drop it from our data frame.
   We use the below command to drop the count column-

   ```
   df_1 = original_df_1.drop(['count'], axis=1)
   ```

   Here, we will be interested to finding the frequent item sets and we will not be needing the count attribute as it will be unique for each of the data row.

2. We use the apriori algorithm to generate the frequent itemset for the student sample data.
   19 frequent item sets are produced, and the length of the item set lies between 1 and 2.
   When generating item sets, we first take individual item calculate their support and accordingly take itemset with minimum support only.
   First step in finding frequent item sets:

French	6	PhD	5
CS	3	Senior	4
Physics	1	16-20	9
Engineering	3	26-30	8
Philosophy	7	Over 30	4
Chemistry	2	21-25	4
Maths	3	3.2-3.6	12
MA	1	3.6-4.0	7
Junior	11	2.8-.2	6
MS	4		

   From the above frequency table, we take only items with minimum support 15%, which is 3.75 for the data set.
   We then carry forward the same process with each itemset.

3. We then save the generated frequent item set to the csv using df.to_csv function

4. Only one association rule is generated with minimum confidence of 90%. It suggests that most of the students are between the age group of 21 to 25 across the subjects and studying at Junior level.

5. We then save the output of the above association rules to the csv file.

6. Three rules are generated with minimum confidence of 70%. It suggests that majority of the junior students are aged between 21-25 age group, PhD students are aged between 26-30 and most of them have taken Philosophy as their subject.

7. We then save the output of the association rule to the csv file using the df.to_csv function.

Question 2:

1. We import the specs/bank_data_question2.csv dataset and remove the ID attribute is unique for all the rows and would not be useful in generation of frequent item sets.

2. We discretize the numeric columns in our data set- age, income, and children using the cut function in pandas.
   With the help of loop for identifying if the list of numeric attributes in our dataframe we discretize the data into 3 bins.

3. We then use the fpgrowth() function from mlxtend.frequent_patterns and find the frequent items in our data frame with a minimum support of 20% by adding the min_support parameter = 0.2
   We get about 231 frequent itemsets of sizes from 1, 2 and few of 3

4. We then save the output from the frequent item sets to the csv file /output/question_2out_fpgrowth.csv

5. By trying out different values for confidence in using the association_rules from mlxtend.frequent_patterns, we identify that a minimum confidence of atleast 79% is required to generate atleast 10 rules. When using 79% confidence, that is the threshold value in the association_rules=0.79 generates 11 rules.

6. We then save the generated association rules in the csv file using to_csv function

7. - The association rules from the frequent item sets suggest that most of the people having a savings account, have no mortgage and at least 1 child.
   - The age of people having current account lies between roughly 18 years to 34 years and their income is between 4956 to 24386
   - From looking at the association rules, it seems like customers with no mortgage, having current account have signed to sign the Personal Equity plan after the last mailing. The organization can target these group of people who are having current accounts and no mortgage and are not married.