Data Analytics

Assignment 2 Report

Team : BlockChainers
Karnati Jahnavi
P Shiridi Kumar

Question - 1:

Part - 1:

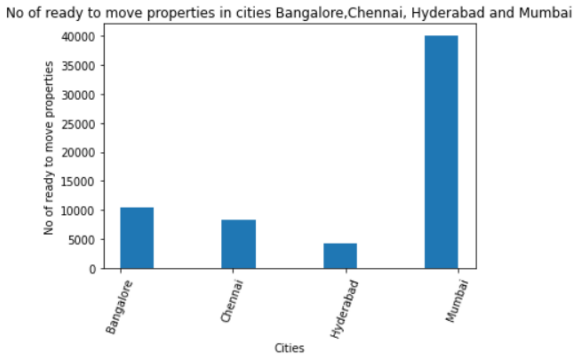
According to Attribute Oriented Induction, data generalization can be performed in either of two ways : 1. Attribute removal 2. Attribute generalization.
First we will perform Attribute removal, we remove the attributes that have a large set of distinct values or its concepts can be expressed by other attributes they can be removed. According to the above rule we remove some attributes from the data such has property_name, property_id, Locality_name, posted_on, buider_name etc (Attribute removal)

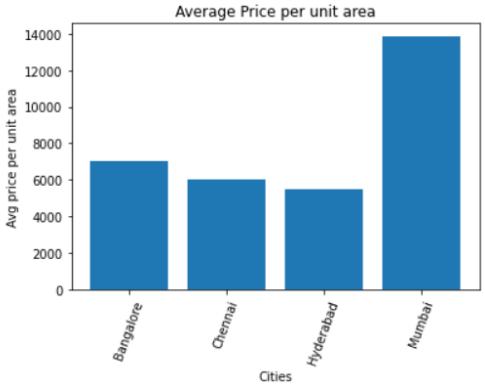
Part - 2:

Now we will be doing Attribute generalization as specified in the question, we will be taking active constructions and generalizing the cities with Technology city and other cities and also we categorize them to budget, luxury and Ultra Luxury based on price per unit area.

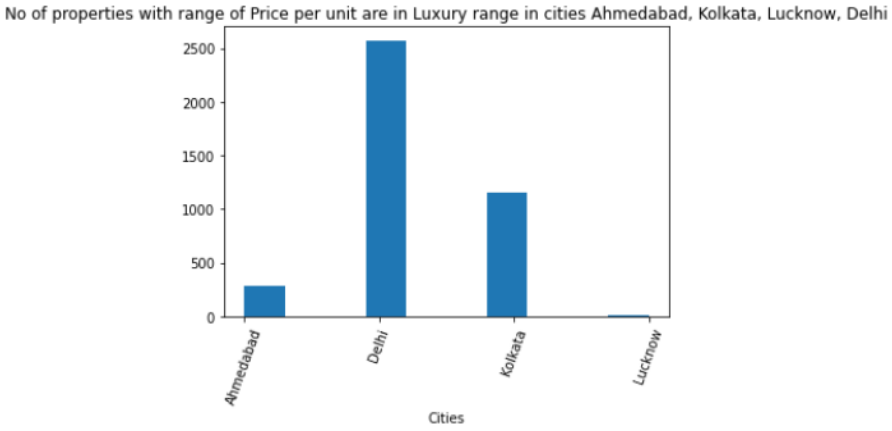
Part - 3:

a. Here we need to summarize properties which are in tech emerging cities that are active. We showed a visualization on Number of properties that are ready to move and the trend between property price and size for each city and also the average price per unit area for each technology city.





b. Here we need to summarize properties which are in non tech emerging cities that are under construction and price per unit area which is in the Luxury range, we showed a visualization of the number of properties that satisfy the given condition.

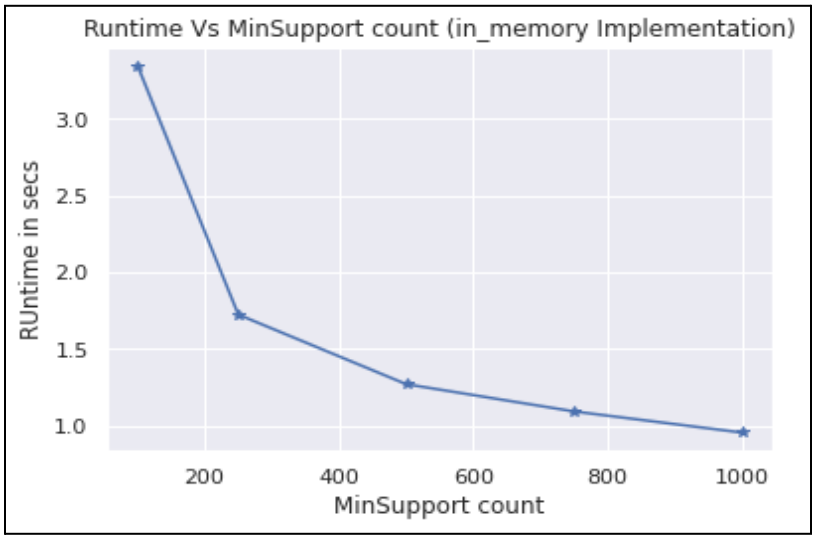


Question - 2:

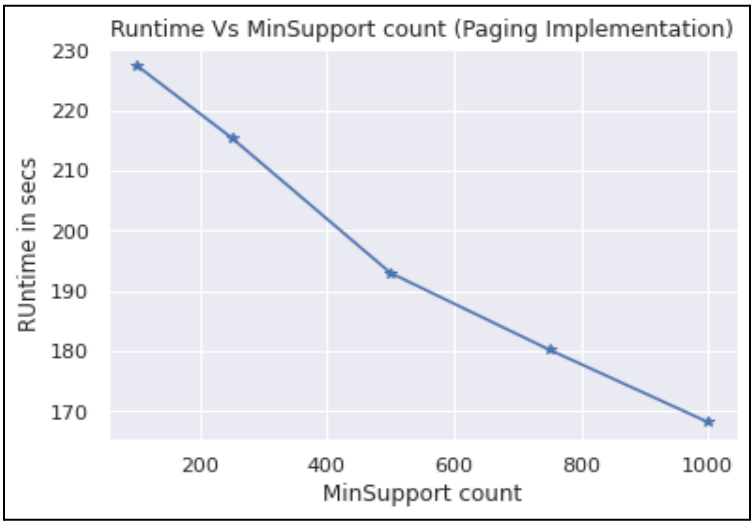
- First 15 rows of result for BUC algorithm with min support =100

	country	year	sex	age	suicides_no	population	gdp_for_year ($)	gdp_per_capita ($)	generation	Count
0	Albania	ALL	female	ALL	ALL	Medium_population	ALL	ALL	ALL	110
1	Albania	ALL	female	ALL	ALL	ALL	Low_income_country	ALL	ALL	102
2	Albania	ALL	female	ALL	ALL	ALL	ALL	ALL	ALL	132
3	Albania	ALL	male	ALL	ALL	Medium_population	ALL	ALL	ALL	110
4	Albania	ALL	male	ALL	ALL	ALL	Low_income_country	ALL	ALL	102
5	Albania	ALL	male	ALL	ALL	ALL	ALL	ALL	ALL	132
6	Albania	ALL	ALL	ALL	1-10	Medium_population	ALL	ALL	ALL	102
7	Albania	ALL	ALL	ALL	1-10	ALL	Low_income_country	ALL	ALL	115
8	Albania	ALL	ALL	ALL	1-10	ALL	ALL	ALL	ALL	135
9	Albania	ALL	ALL	ALL	ALL	Medium_population	Low_income_country	ALL	ALL	170
10	Albania	ALL	ALL	ALL	ALL	Medium_population	ALL	ALL	ALL	220
11	Albania	ALL	ALL	ALL	ALL	ALL	Low_income_country	0-1000	ALL	108
12	Albania	ALL	ALL	ALL	ALL	ALL	Low_income_country	ALL	ALL	204
13	Albania	ALL	ALL	ALL	ALL	ALL	ALL	0-1000	ALL	108
14	Albania	ALL	ALL	ALL	ALL	ALL	ALL	ALL	ALL	264

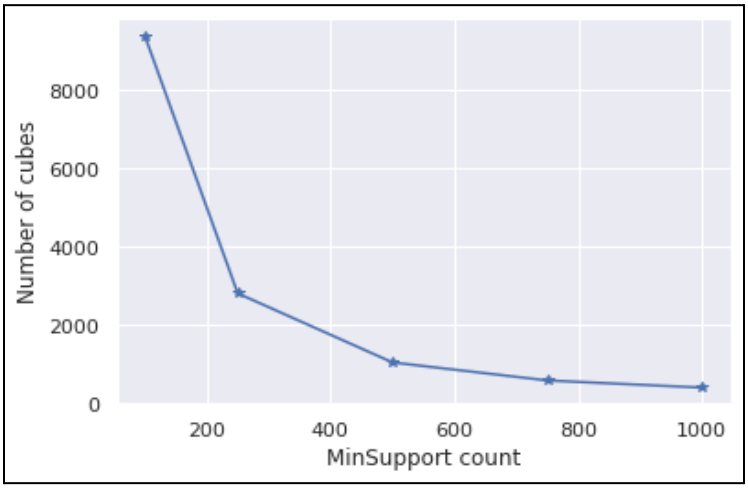
- Min Support vs Runtime without paging



- Min Support vs Runtime paging implementation



- Min Support vs Number of cubes



- Block_size vs Runtime (with support count =2 and block size in k cells)