Exploratory Data Analysis

StudentID: 22100128

Name: Kim Seung Hyon

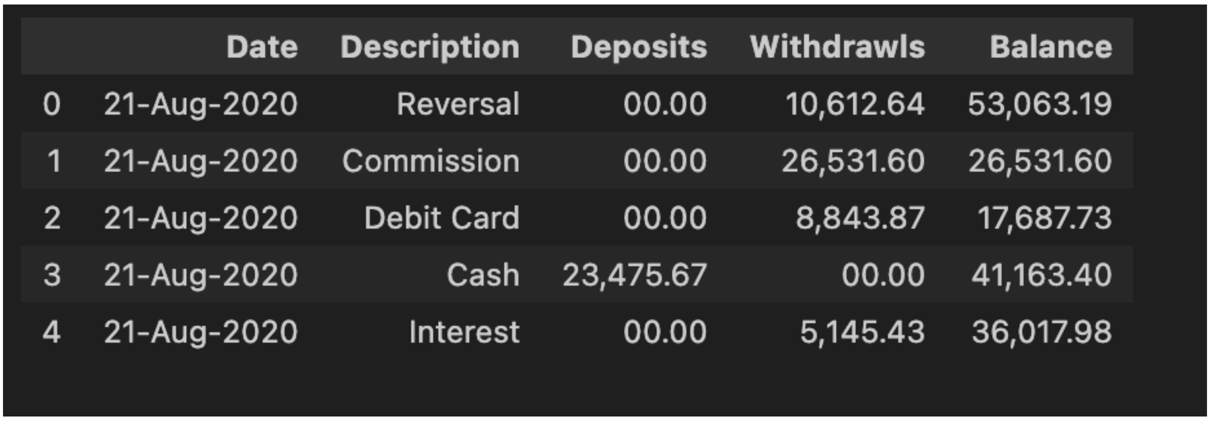
1st Major: ACE

2nd Major: Data Science

Comparing between Date with the change of value of deposits, withdrawls, and balance and check the relationship between that 3 factors.

1. Data overview

Descriptives statistics on overall data (sample size, number of variables, data type, data range, distribution, etc.)



	Date	Description	Deposits	Withdrawls	Balance
0	21-Aug-2020	Reversal	00.00	10,612.64	53,063.19
1	21-Aug-2020	Commission	00.00	26,531.60	26,531.60
2	21-Aug-2020	Debit Card	00.00	8,843.87	17,687.73
3	21-Aug-2020	Cash	23,475.67	00.00	41,163.40
4	21-Aug-2020	Interest	00.00	5,145.43	36,017.98

bt_data head value

sample size: 5,000,000 (1 to 5,000,000)

number of variables: 5(Date, Description, Deposits, Withdrawls, Balance) data type: object

Data range

- Deposits: minimum value: 0.01 / maximum value: 2097145.2
- Withdrawls: minimum value: 0.01 / maximum value: 10546488.84
- Balance: minimum value: 0.01 / maximum value: 10670658.67

2. Univariate analysis

Presentation of key variables from various aspects

Since the data is too big, we read the data until 100000 rows and sample it to 1000. We assume that the data contains the value of time, deposits, withdrawls, and balance which are all linked with bank, so compare the time and other values. We focus on the change of values as time passing.

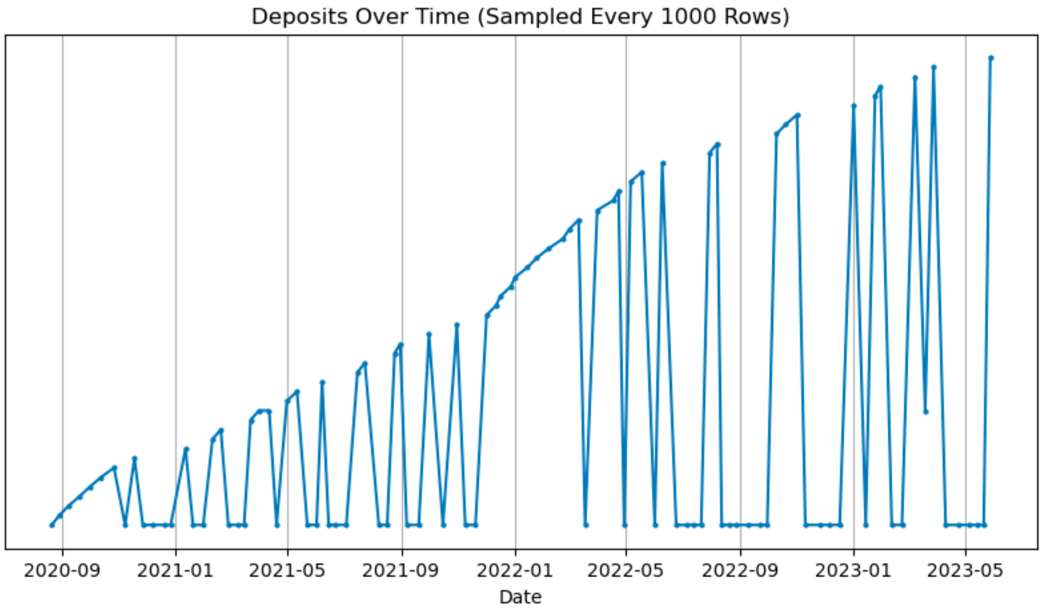


Fig deposits over time (sampled every 1000 rows)

Firstly We restrict the data until 100000 so we get the date value from 2020-09 till 2023-05. This graph represents as deposits increases continuously since time increased. It shows that people spend more as time passed. It can be assumed as high welfare, and currency rate affect to form this structure of graph.

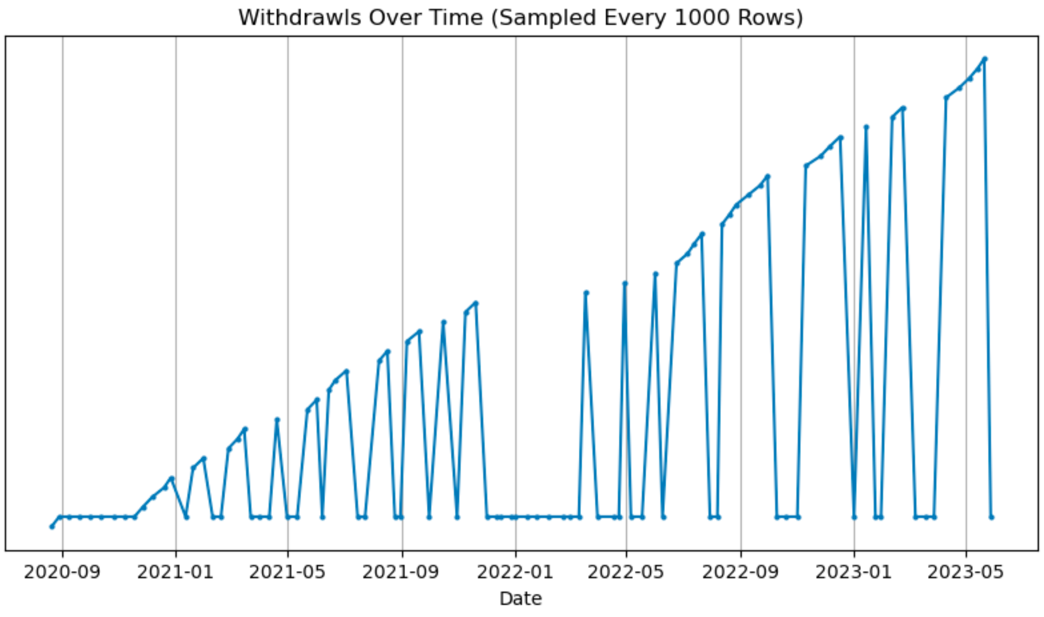


Fig withdrawls over time (sampled every 1000 rows)

This graph represents as withdrawls increases continuously since time increased as well. There are quite many outliers which represent as 0, however overall graph accelerates. It can be assumed as economic growth, change of currency and inflation affect the change of withdrawls as increased.

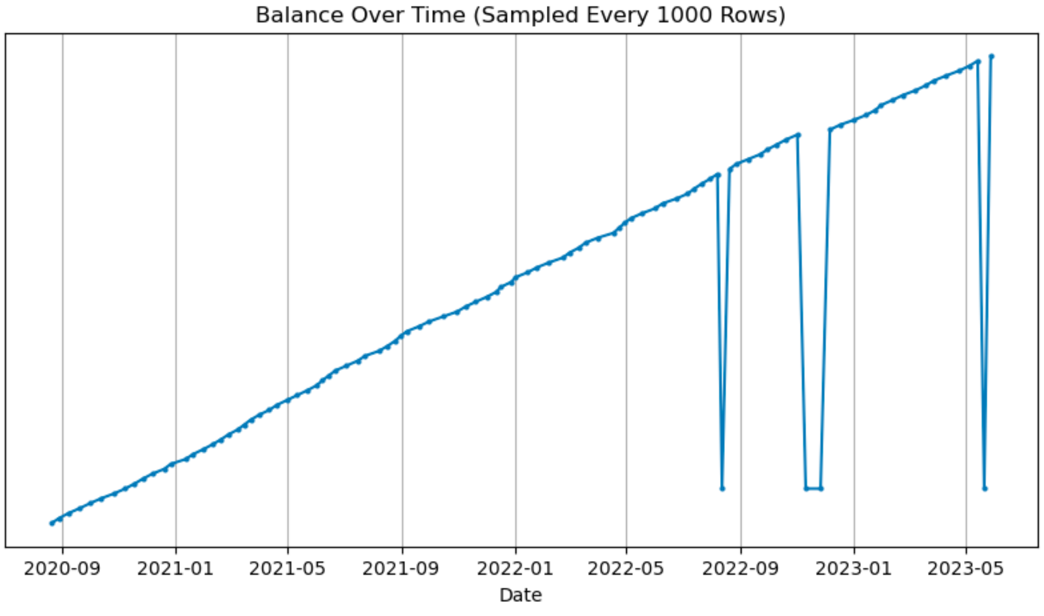


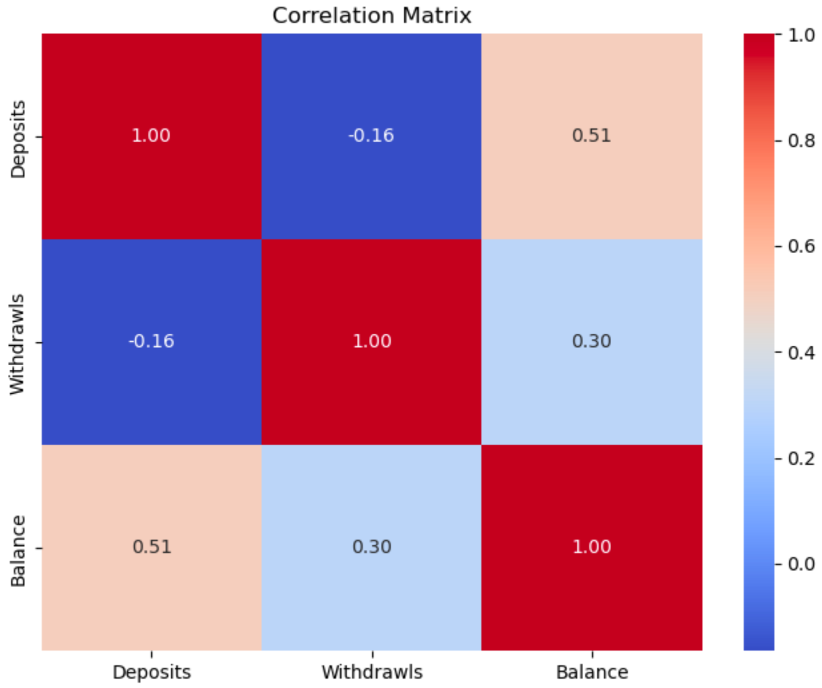
Fig balance over time (sampled every 1000 rows)

It shows balance increases continuously since time increased as well. Although withdrawls increase, deposits increase more than that, so in average the overall balance over time graph accelerates.

3. Multivariate analysis <br>

Presenation of hidden patterns between variables (correlation, clustering, etc.)

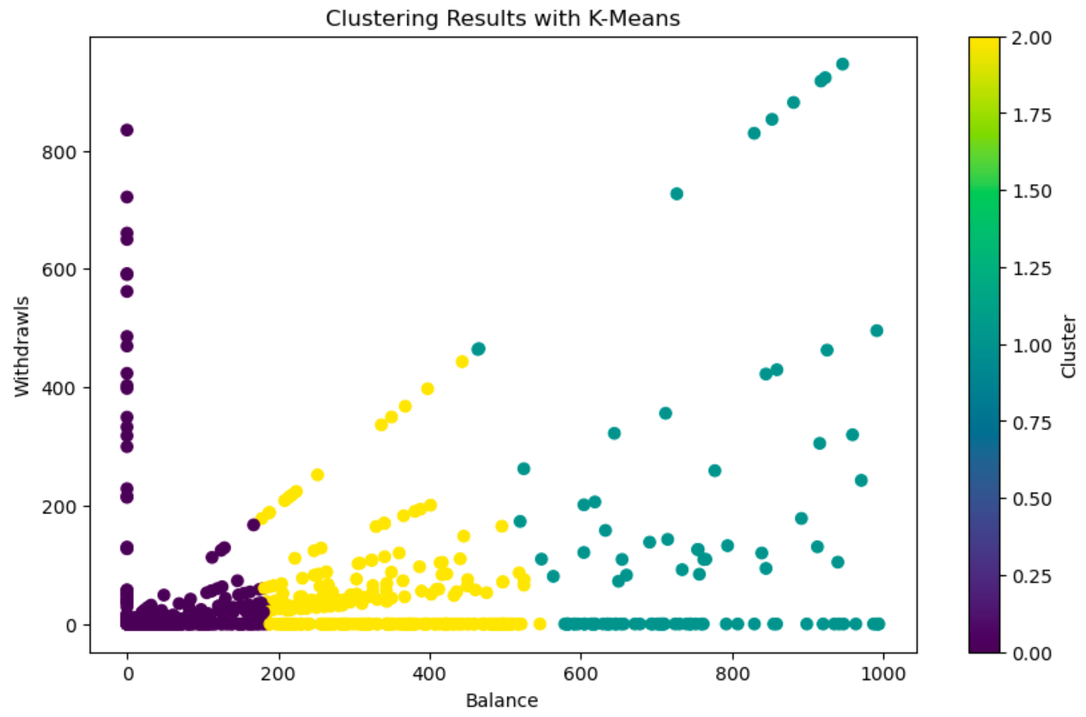
3.1 (ex) Correlation



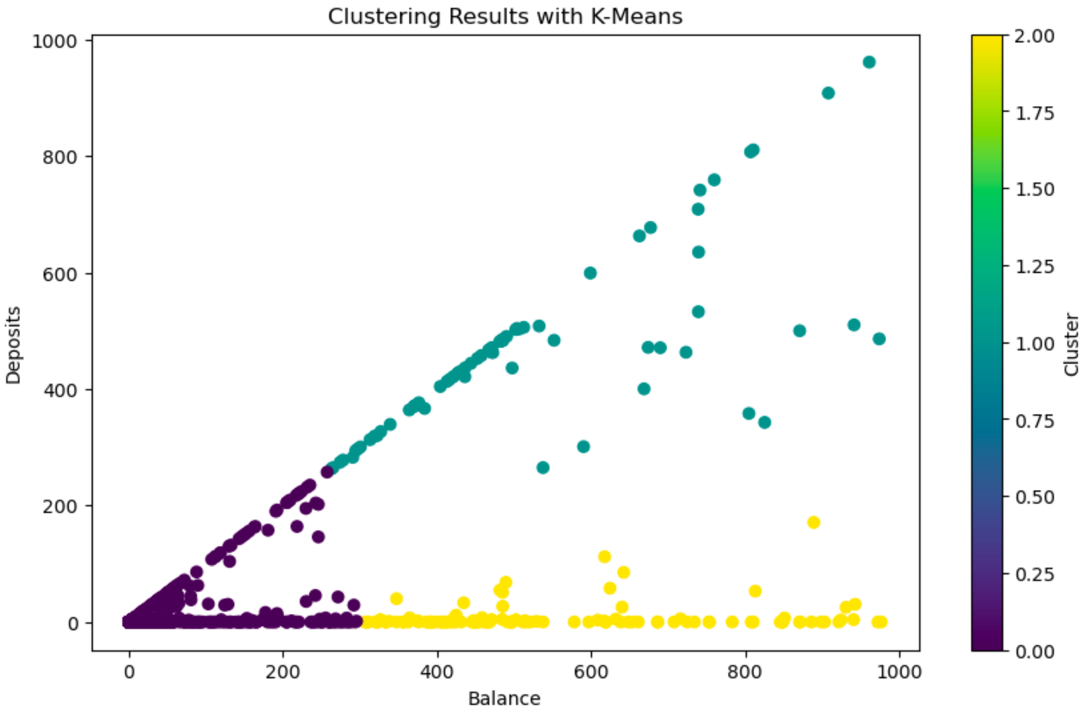
Correlation matrix that compares deposits, withdrawls, and balance.

Despite of deposits and withdrawls have less relationship, balance affects to both deposits and withdrawls rate.

3.2 (ex) Clustering



2D clustering between balance and withdrawls.



2D clustering between deposits and balance.

4. Suggestion

By the result of the graph, as the time passed from past to the future, the overall balance, deposits and withdrawls increase. We are able to check that deposits and withdrawls are not related, however both of the give and get affect from balance. Therefore, for increasing the balance we have to control the deposits and withdrawls. Overall graph shows that more people are wealthy than previously, and also the gap between rich and poor decreases.