Exploratory Data Analysis

- StudentID: 21900121

- Name: 김세희

- 1st Major: ICT 융합

- 2nd Major: Data Science

BT_data

Brief summary of your proposed project idea.

은행에서 고객들의 돈을 받거나 돈을 사용할 때, 어떤 부분에서 많이 받고 나가는지 확인하고 그에 대한 대비를 한다.

1. Data overview

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

- 날짜마다 어떤 유형으로 예금, 인출, 잔액이 이루어졌는지 보여줌.
- Sample size: 5000000
- Number of variales: 5
- Data type

#	Column	Dtype
0	Date	object
1	Description	object
2	Deposits	object
3	Withdrawls	object
4	Balance	object

- 2020-08-21 ~ 2155-11-15의 거래내역
- NA 존재하지 않음
- Deposits
    - mean : 89766.407

    - min : 0
    - median : 0
    - 제3사분위수 (75%) : 2359.000
    - max : 2097145.200
    - 표준편차 : 321677.685

- Withdrawls

    - mean : 89766.395
    - min : 0
    - median : 0
    - 제3사분위수 (75%) : 56424.410
    - max : 10546488.840
    - 표준편차 : 264071.197

- Balance

    - mean : 613605.152
    - min : 0
    - 제1사분위수 (25%) : 30754.450
    - median : 267665.200
    - 제3사분위수 (75%) : 937762.085
    - max : 10670658.670
    - 표준편차 : 794376.242

2. Univariate analysis

2.1 Description

Bill, Commission, NEFT, ATM, Miscellaneous, Cash, Reversal, IMPS, Interest, Purchase, Cheque, Transfer, Tax, RTGS, Debit Card 모두 거의 동일한 퍼센트(0.067)로 유형이 나눠져 있음. 자세하게 살펴보면 Bill이 334,106으로 가장 많은 거래 유형이고 Debit Card가 332,703으로 가장 적은 거래 유형임.

2.2 Withdrawls

- mean : 89766.395
- min : 0
- median : 0
- 제3사분위수 (75%) : 56424.410
- max : 10546488.840
- 표준편차 : 264071.197

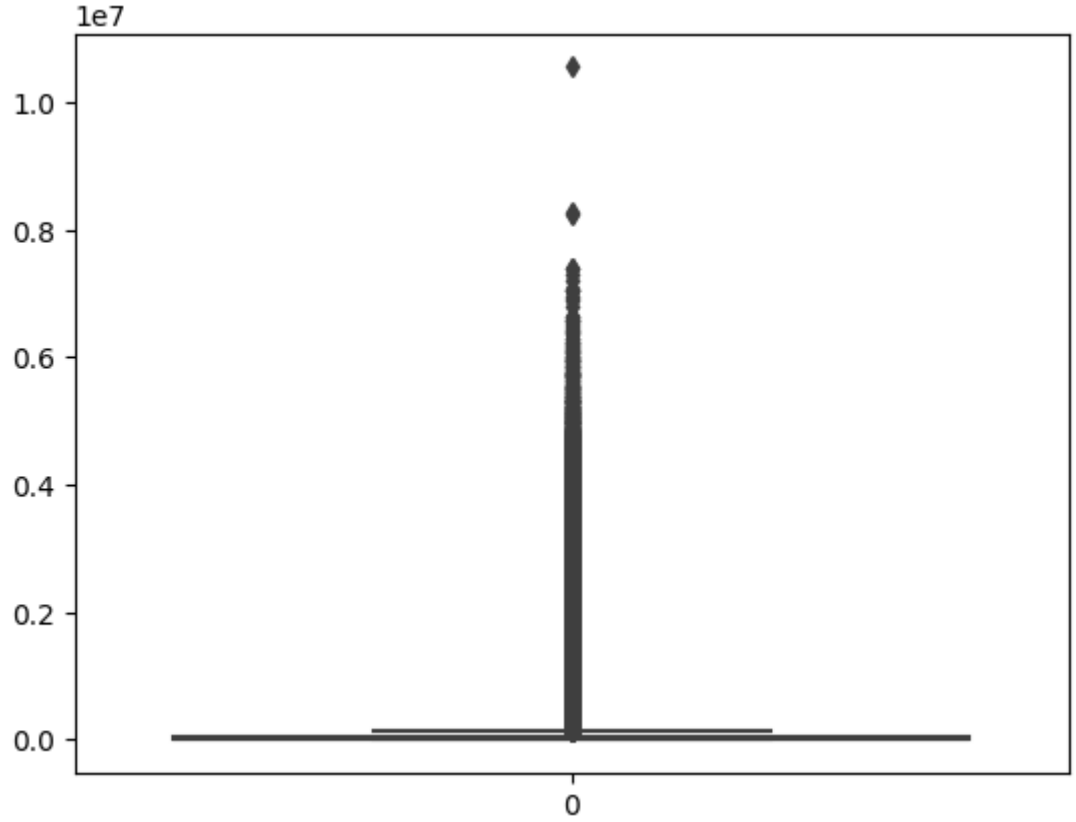

Figure 1. Withdrawls box plot

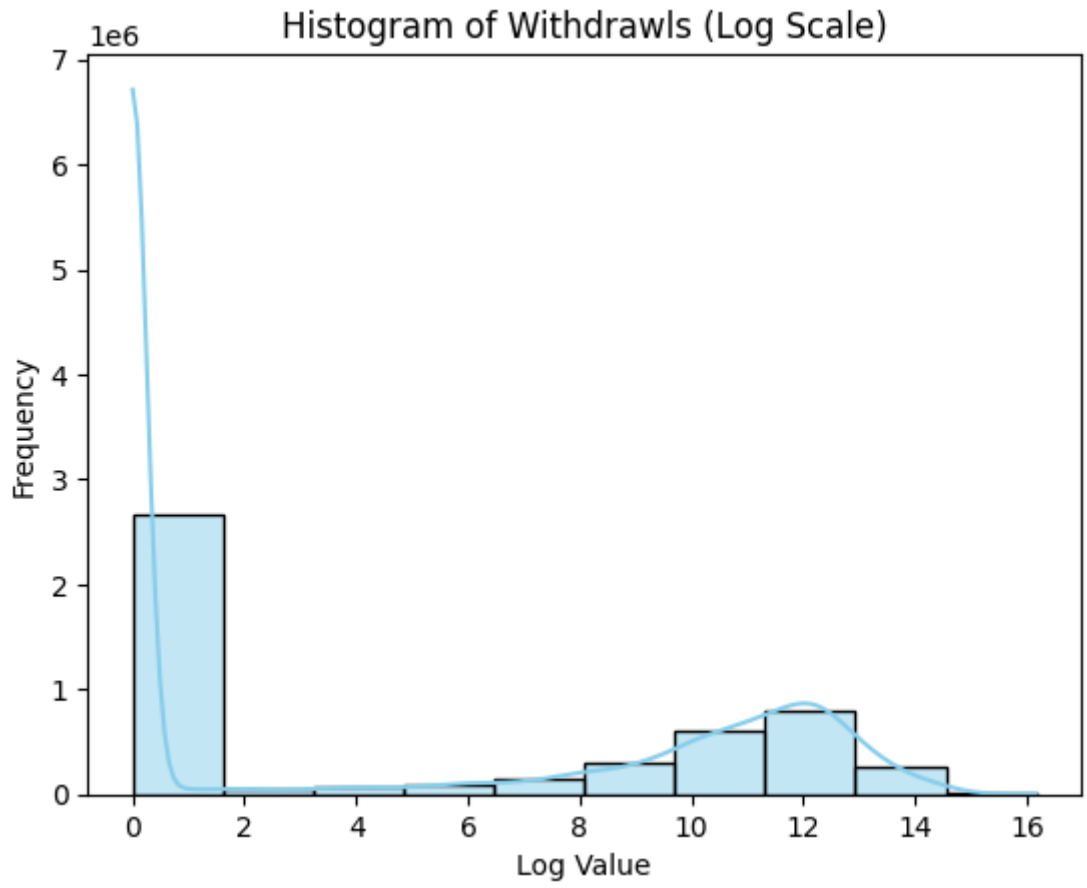

Figure 2. Withdrawls distribution plot

Boxplot과 통계량을 살펴보면 0인 값이 많은 것을 알 수 있음. 최소와 최대가 크게 차이나서 log 변환을 통해서 분포를 살펴봤을 때, 어느 정도 값 이상일 때, 오른쪽으로 치우쳐진 것을 볼 수 있음.

3. Multivariate analysis

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

3.1 Correlation

연속형끼리의 correlation

	Deposits	Withdrawls	Balance
Deposits	1.000	-0.095	0.427
Withdrawls	-0.095	1.000	0.069
Balance	0.427	0.069	1.000

Deposits와 Withdrawls가 음의 상관관계가 높은 이유는 예금을 하지 않고 돈을 인출하거나 결제하는 경우, 그리고 그 반대의 경우가 많기 때문이다.

3.2 Clustering

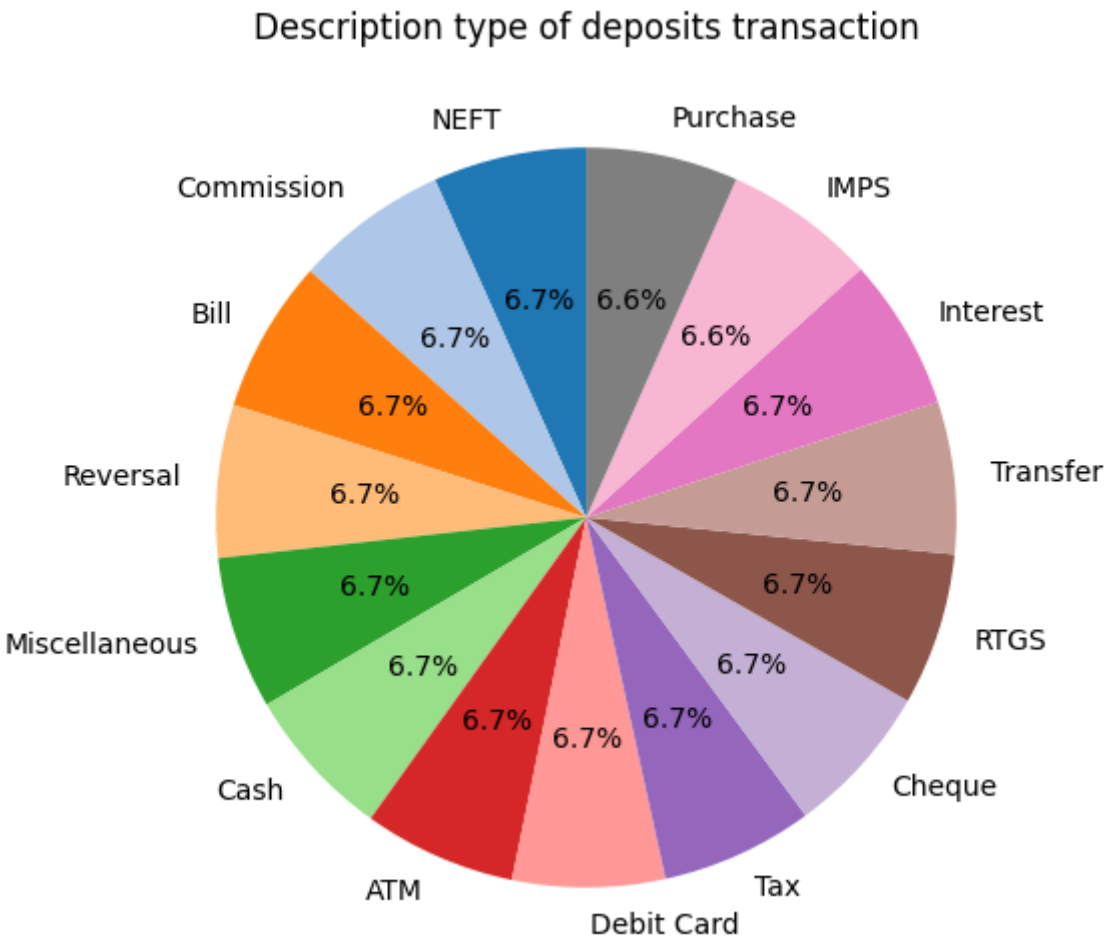

Figure 3. Plot of description type of deposits transaction

예금할 때 가장 많이 사용하는 거래 유형 3개: NEFT, Commission, Bill

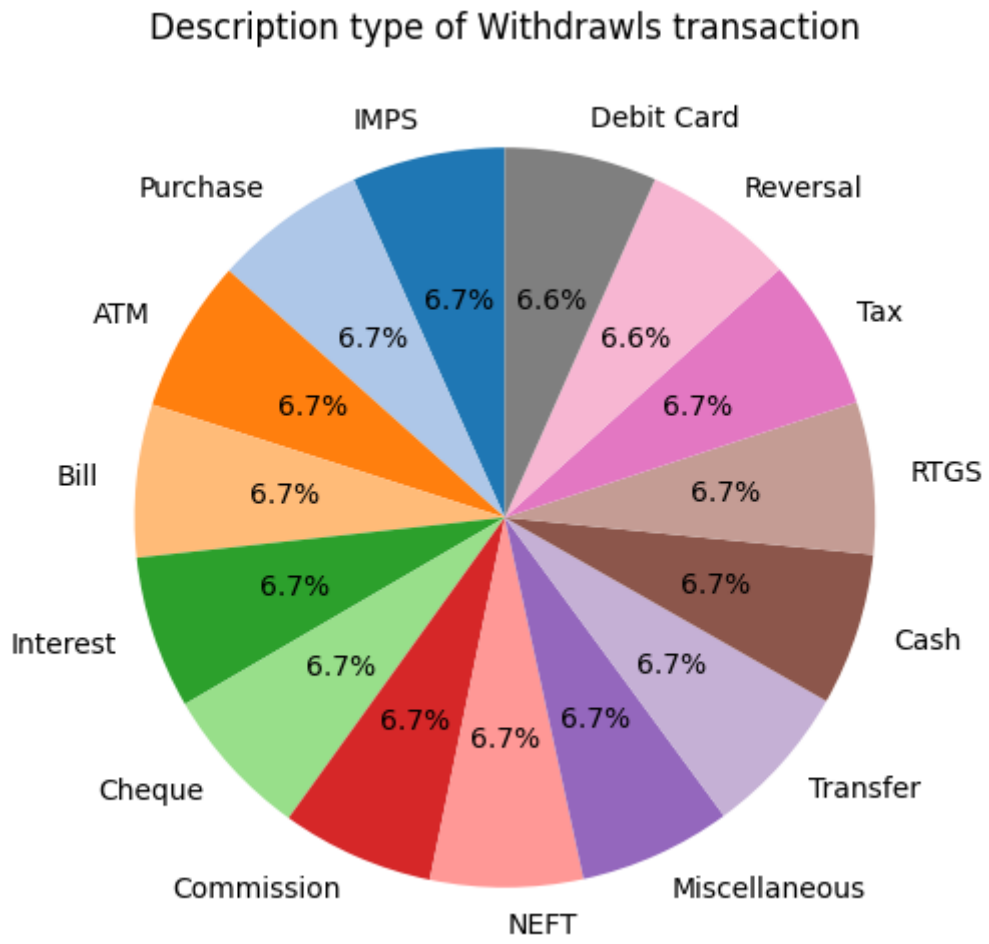

Figure 4. Plot of description type of withdrawls transaction

지출이나 인출할 때 가장 많이 사용하는 거래 유형 3개: IMPS, Purchase, ATM

4. Suggestion

Based on the insights you obtained from the previous stages, propose the potential project idea.

국내 전자 이체을 통해 가장 많이 예금하는 것을 알 수 있고, commision, bill을 통해서 수입이 생긴다. IMPS(실시간 금융 거래), 상품 구매나 ATM기기를 이용해서 출금되거나 인출되기 때문에 ATM기기를 점차 줄이는 방향으로 가서 ATM에서 돈이 나가는 것을 막는다.

CC_data

Brief summary of your proposed project idea.

사람들에게 필요한 카드 타입과 주거래은행에 맞춰서 카드 추천

1. Data overview

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

credic card와 그에 대한 자세한 정보

- Sample size: 5000000
- Number of variales: 11
- Data type

#	Column	Dtype
0	Card Type Code	object
1	Card Type Full Name	object
2	Issuing Bank	object
3	Card Number	int64
4	Card Holder's Name	object
5	CVV/CVV2	int64
6	Issue Date	object
7	Expiry Date	object
8	Billing Date	int64
9	Card PIN	int64
10	Credit Limit	int64

- Range
    - Issue Date (카드발급일)
        - range: 2010-01-01~2020-12-01
    - Expiry Date (만료일)
        - range: 2011-01-01~2040-12-01
    - Billing Date
        - range: 1~28일
        - 평균: 14.5일
        - 표준편차: 8.08일
- NA 존재하지 않음

2. Univariate analysis

Presentation of key variables from various aspects

2.1 Issuing Bank

카드 중에서 Diners Club, JCB, Discover 은행 순으로 많음. 카드 수가 1.7% 이하인 은행들은 Other로 합쳐서 봄. Other 은행은 Wells Fargo, Barclays, GE Capital, PNC, Cabela뭘 WFB, First National임.

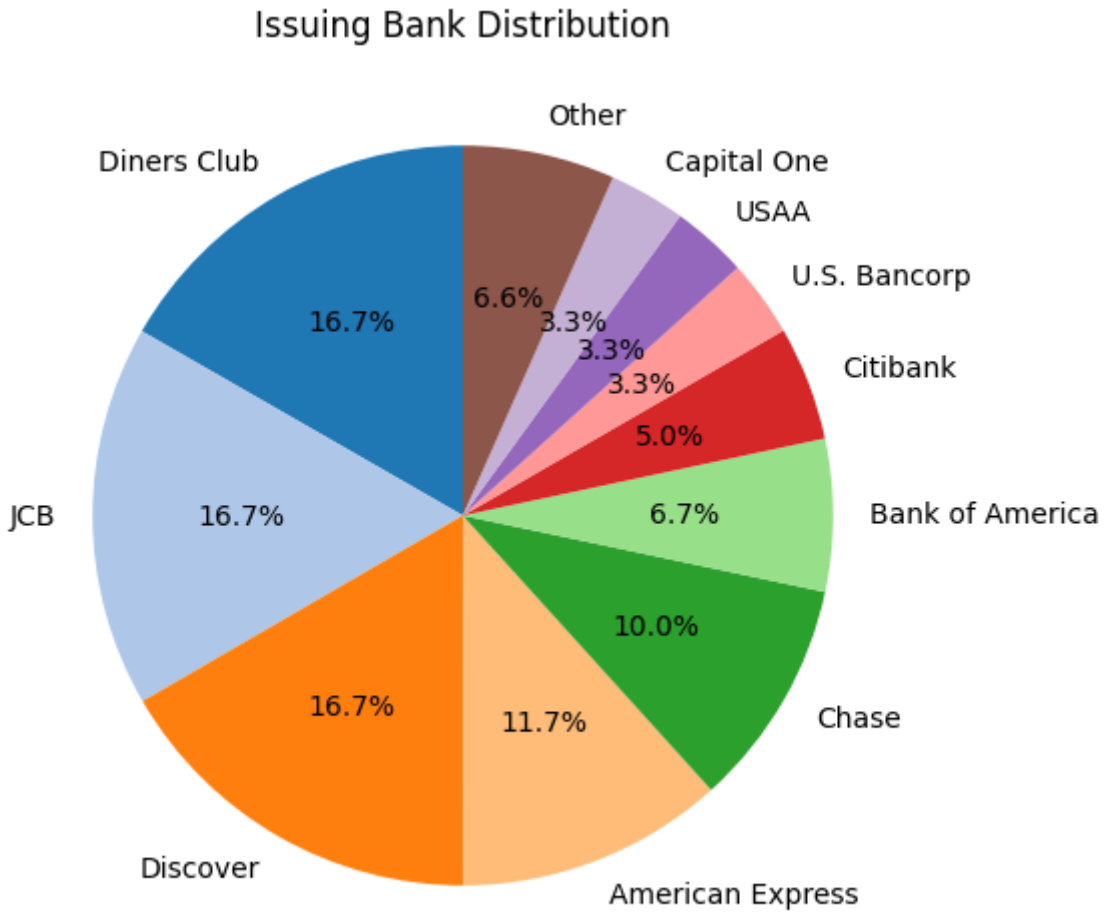

Figure 5. Using percent of issue banks

2.2 Credit Limit

- mean : 104979.497
- min : 10000.000
- 제1사분위수 (25%) : 57400.000
- median : 105000.000
- 제3사분위수 (75%) : 152500.000
- max : 200000.000
- 표준편차 : 54873.808

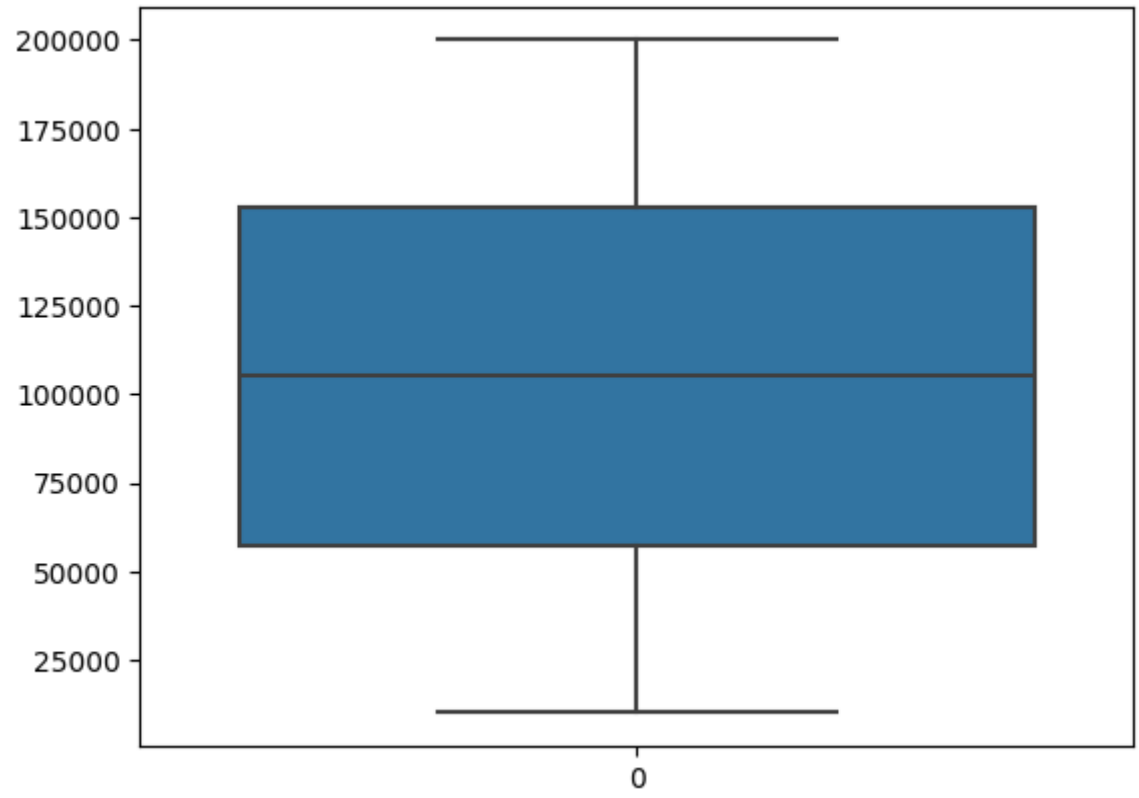

Figure 6. Credit Limit box plot

이상치가 존재하지 않는 것으로 보임. 평균과 중간값이 거의 비슷함.

2.3 Card using duration

Expiry Date에서 Issue Date를 빼 카드를 사용할 수 있는 기간을 새로운 열로 생성함.

- mean : 3834 days
- min : 365 days
- 제1사분위수 (25%) : 1827 days
- median : 3653 days
- 제3사분위수 (75%) : 5479 days
- max : 7305 days
- 표준편차 : 2105 days

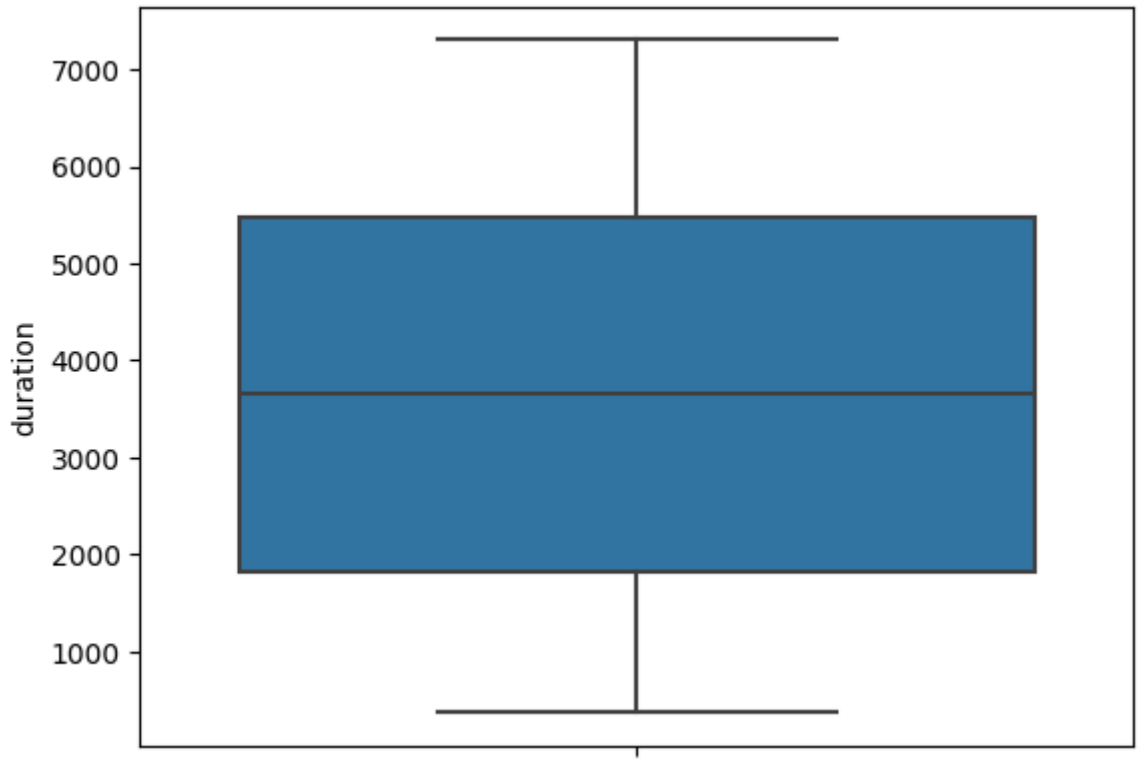

Figure 7. Card using duration box plot

Card using duration Group	Count
15년~20년	1498595
5년~10년	1251032
10년~15년	1250277
5년 이내	1000096

카드 중 사용 기간이 제일 긴 15년~20년이 가장 많은 사람들이 사용하고 있음. 10년~15년보다 5년~10년인 카드를 더 많이 발급함.

3. Multivariate analysis

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

3.1 Correlation

숫자형 변수들의 correlation

	Card Number	CVV/CVV2	Billing Date	Card PIN	Credit Limit	duration
Card Number	1.000	-0.457	0.000	-0.000	0.000	-0.001
CVV/CVV2	-0.457	1.000	0.000	-0.000	-0.000	0.000
Billing Date	0.000	0.000	1.000	0.000	0.000	-0.000

	Card Number	CVV/CVV2	Billing Date	Card PIN	Credit Limit	duration
Card PIN	-0.000	-0.000	0.000	1.000	-0.001	0.000
Credit Limit	0.000	-0.000	0.000	-0.001	1.000	0.001
duration	-0.001	0.000	-0.000	0.000	0.001	1.000

Card Number와 CVV/CVV2와 음의 상관관계가 존재하지만 크게 신경쓸 필요는 없어보임.

3.2 Clustering

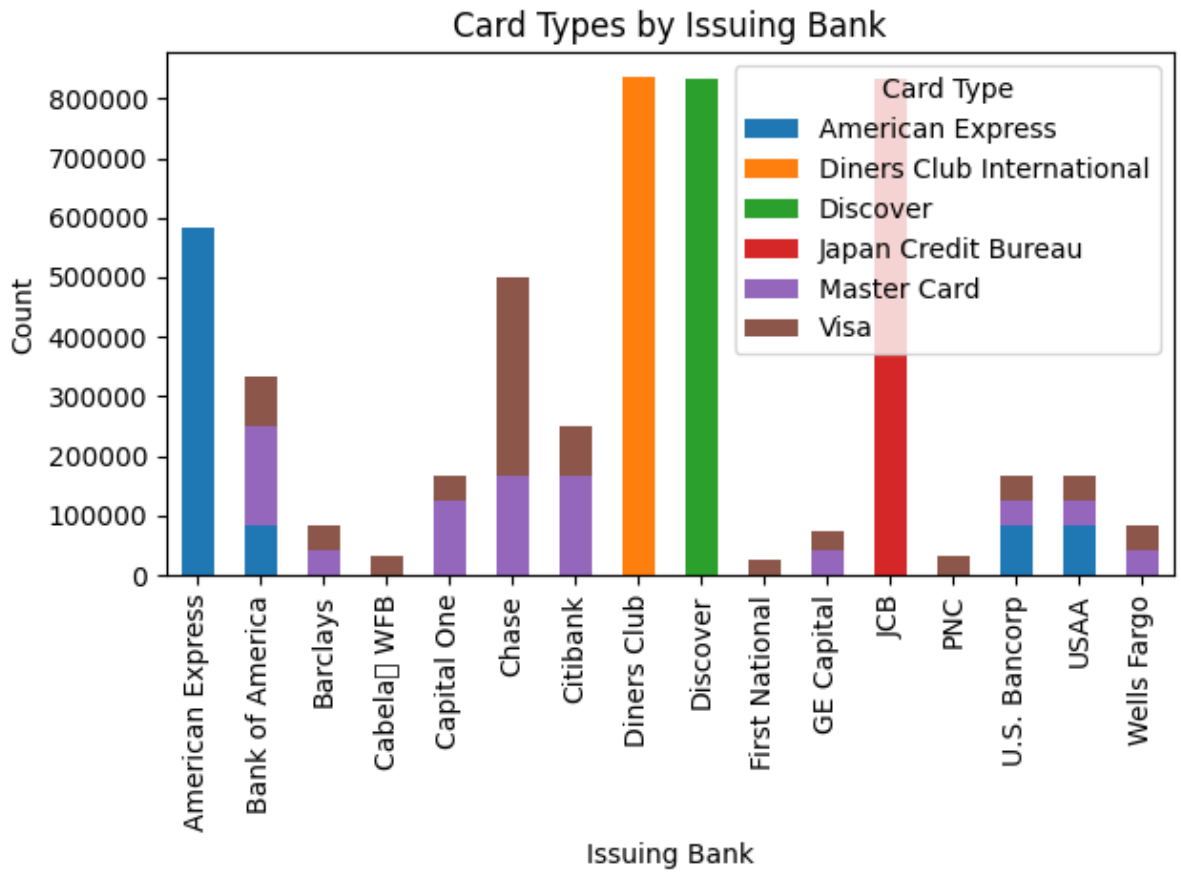

Figure 8. Bar plot of Card Types by Issuing Bank

- Diners Club International, Discover, Japan Credit Bureau 카드는 각 카드에 해당하는 은행밖에 발급할 수 없음.
- Bank of America와 U.S. Bancorp, USAA 은행은 가장 다양한 카드를 발급할 수 있음.

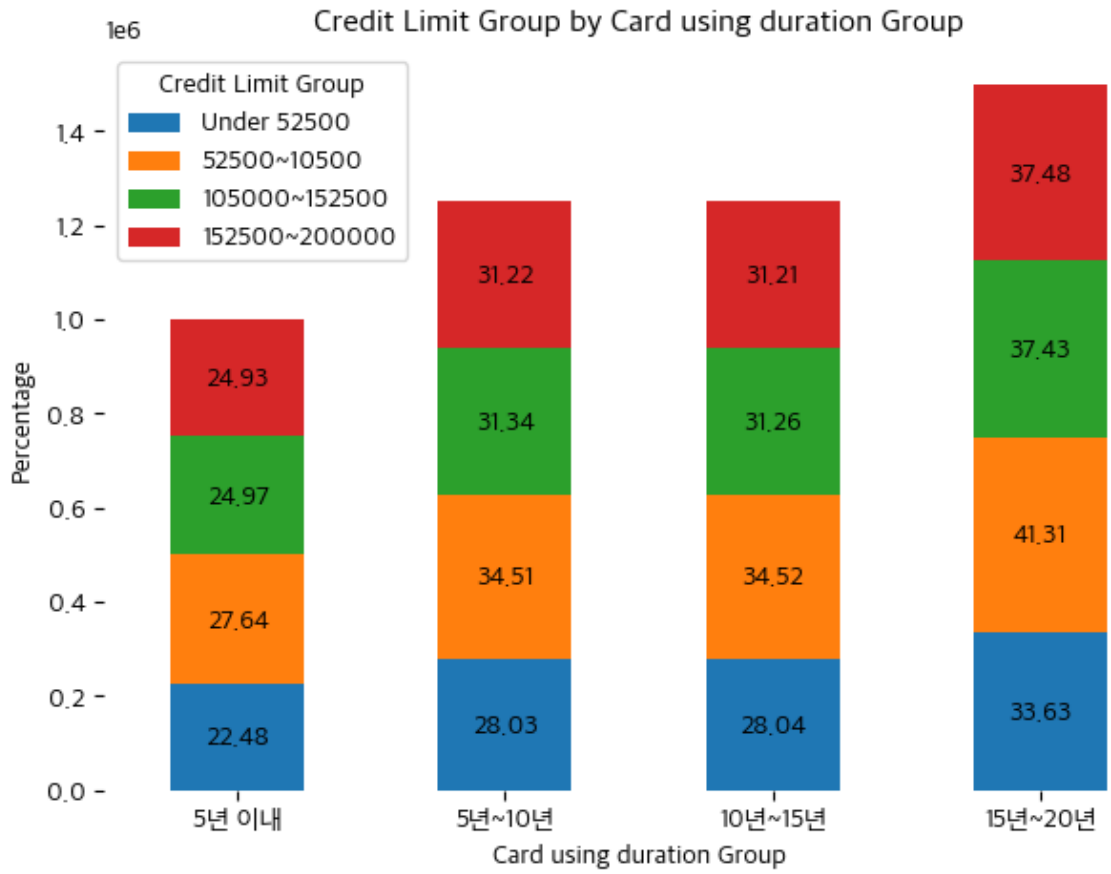

Figure 9. Bar plot of Credit Limit Group by Card using duration Group

- 카드 사용 기간과 상관없이 카드 한도가 $52500 ~ $10500인 것이 제일 많음.
- 카드 사용 기간이 15년 ~ 20년일 때만 2번째로 많은 카드 한도가 $152500 ~ $200000임.

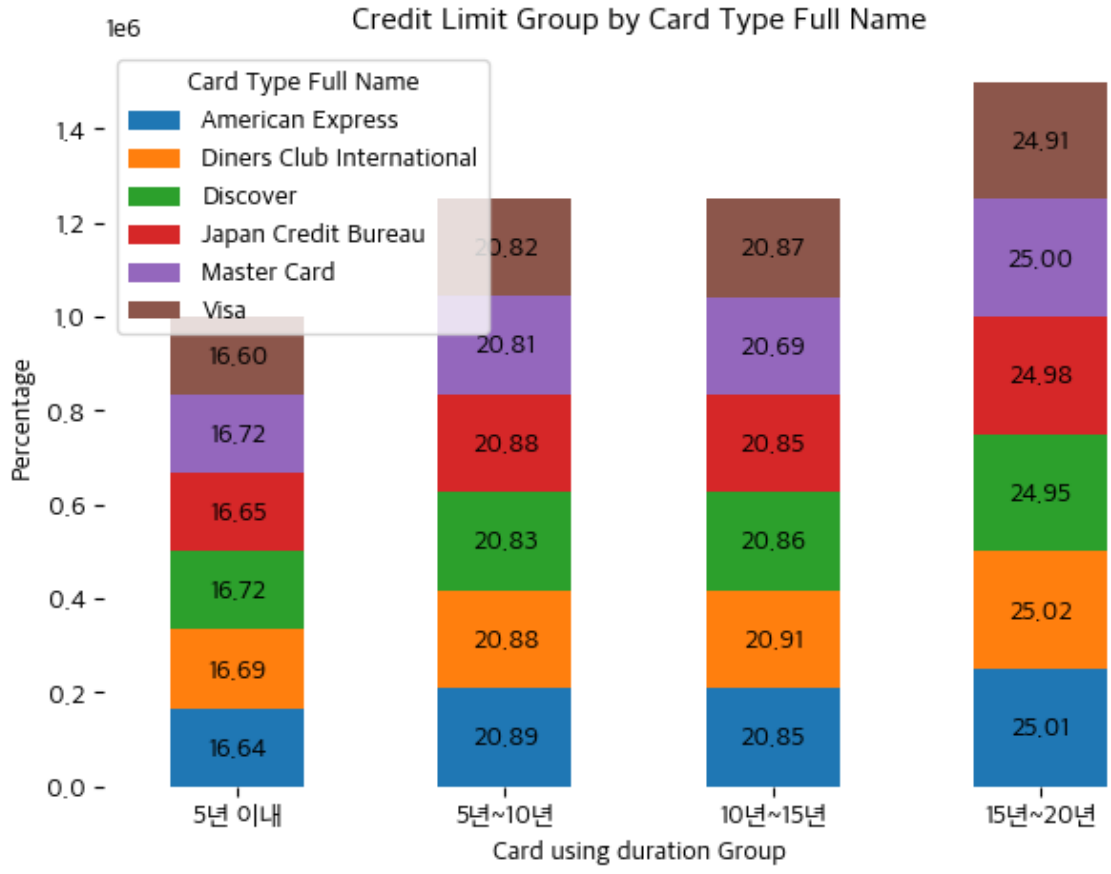

Figure 10. Bar plot of Credit Limit Group by Card Type Full Name

- 카드 기간이 5년 이내일 때는 Master Card 유형이 제일 많고 Visa 카드가 제일 없음
- 카드 기간이 5년 ~ 10년일 때는 American Express 유형이 제일 많고 Master Card 카드가 제일 없음
- 카드 기간이 10년 이상일 때는 Diners Club International 유형이 제일 많음.
- 카드 기간과 상관없이 다 잘 쓰이는 카드 유형은 Diners Club International임.

4. Suggestion

Based on the insights you obtained from the previous stages, propose the potential project idea.

- 카드 추천 가능
  - 카드를 처음 만드는 사람이라면
    - 카드 한도 : 52500~10500
    - 주거래 은행이 Americam Express, Diners Club International, Discover, Japan Credit Bureau 이라면 그에 해당하는 카드 유형으로만 카드 발급 가능
    - 만약 다른 카드 유형을 만들고 싶다면, Bank of America와 U.S. Bancorp, USAA 은행을 추천
    - 카드 사용 기간 : 장기간 사용하길 원한다면 15년 ~ 20년을 추천, 짧은 기간 사용 원하면 5년 이내보단 5년 ~ 10년을 추천
  - 새로운 카드를 만들 사람에게
    - 만약 카드 한도가 큰 카드 발급 원한다면 10년 ~ 15년 사이보다 15년 ~ 20년인 카드, Diners Club International 카드 유형 추천

HR_data

Brief summary of your proposed project idea.

1. Data overview

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

직원들의 성별, 생년월일 같은 기본 정보들과 회사에 언제 들어왔는지 같은 회사와 관련된 정보를 담고 있음.

- Sample size: 5000000
- Number of variales: 37
- Data type

#	Column	Dtype
0	Emp ID	int64
1	Name Prefix	object

#	Column	Dtype
2	First Name	object
3	Middle Initial	object
4	Last Name	object
5	Gender	object
6	E Mail	object
7	Father's Name	object
8	Mother's Name	object
9	Mother's Maiden Name	object
10	Date of Birth	object
11	Time of Birth	object
12	Age in Yrs.	float64
13	Weight in Kgs.	int64
14	Date of Joining	object
15	Quarter of Joining	object
16	Half of Joining	object
17	Year of Joining	int64
18	Month of Joining	int64
19	Month Name of Joining	object
20	Short Month	object
21	Day of Joining	int64
22	DOW of Joining	object
23	Short DOW	object
24	Age in Company (Years)	float64
25	Salary	int64
26	Last % Hike	object
27	SSN	object
28	Phone No.	object
29	Place Name	object
30	County	object
31	City	object

#	Column	Dtype
32	State	object
33	Zip	int64
34	Region	object
35	User Name	object
36	Password	object

- NA 값 존재하지 않음

HRA_data

Brief summary of your proposed project idea.

1. Data overview

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

HR 데이터보다 직원들의 더 사적인 정보를 담고 있음.

- Sample size: 5000000
- Number of variales: 35
- Data type |#| Column | Dtype | |---| ------ | ----- | |0 | Age | int64 | |1 | Attrition | object| |2 | BusinessTravel | object| |3 | DailyRate | int64 | |4 | Department | object| |5 | DistanceFromHome | int64 | |6| Education | int64 | |7 | EducationField | object| |8 | EmployeeCount | int64 | |9 | EmployeeNumber | int64 | |10 | EnvironmentSatisfaction | int64 | |11 | Gender | object| |12 | HourlyRate | int64 | |13 | JobInvolvement | int64 | |14 | JobLevel | int64 | |15 | JobRole | object| |16 |JobSatisfaction | int64 | |17 | MaritalStatus | object| |18 | MonthlyIncome | int64 | |19 | MonthlyRate | int64 | |20 | NumCompaniesWorked | int64 | |21 | Over18 | object| |22 | OverTime | object| |23 | PercentSalaryHike | int64 | |24 | PerformanceRating | int64 | |25 | RelationshipSatisfaction| int64 | |26 | StandardHours | int64 | |27 | StockOptionLevel | int64 | |28 | TotalWorkingYears | int64 | |29 |TrainingTimesLastYear | int64 | |30 | WorkLifeBalance | int64 | |31 |YearsAtCompany | int64 | |32 | YearsInCurrentRole | int64 | |33 | YearsSinceLastPromotion | int64 | |34 |YearsWithCurrManager | int64 |
    - int type : 26
    - object type : 9
- NA 값 존재하지 않음

S_data

Brief summary of your proposed project idea.

나라별로 많이 나가는 상품 유형이나 주문 우선순위 중에서 가장 필요한 상품 유형이 무엇인지 알아보고 상품을 미리 준비하거나 재고관리를 더 잘할 수 있게 만든다.

1. Data overview

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

2010년부터 2020년 10월 30일까지의 배를 타고 이동되는 유통 물품 주문 정보를 담고 있음.

- Sample size: 5000000
- Number of variales: 14
- Data type

#	Column	Dtype
0	Region	object
1	Country	object
2	Item Type	object
3	Sales Channel	object
4	Order Priority	object
5	Order Date	object
6	Order ID	int64
7	Ship Date	object
8	Units Sold	int64
9	Unit Price	float64
10	Unit Cost	float64
11	Total Revenue	float64
12	Total Cost	float64
13	Total Profit	float64

- NA 값 존재하지 않음
- 7개 대륙 185개 국가 거래 정보
- Sales Channel

    - Online, Offline로 나뉨.

- Order Priority
- H : High
- M : Middle
- L : Low
- C : Critical
- Units Sold
    - mean : 4999.991
    - min : 1.000
    - 제1사분위수 (25%) : 2500.000
    - median : 4999.000
    - 제3사분위수 (75%) : 7500.000
    - max : 10000.000
    - 표준편차 : 2886.787
- Unit Price
    - mean : 266.191
    - min : 9.330
    - 제1사분위수 (25%) : 109.280
    - median : 205.700
    - 제3사분위수 (75%) : 437.200
    - max : 668.270
    - 표준편차 : 217.015
- Unit Cost
    - mean : 187.656
    - min : 6.920
    - 제1사분위수 (25%) : 56.670
    - median : 117.110
    - 제3사분위수 (75%) : 364.690
    - max : 524.960
    - 표준편차 : 175.701
- Total Revenue
    - mean : 1331058.049
    - min : 9.330
    - 제1사분위수 (25%) : 277963.730
    - median : 785979.700
    - 제3사분위수 (75%) : 1822443.920
    - max : 6682700.000
    - 표준편차 : 1469901.954

- Total Cost
    - mean : 938378.122
    - min : 6.920
    - 제1사분위수 (25%) : 161925.120
    - median : 467712.000
    - 제3사분위수 (75%) : 1197433.760
    - max : 5249600.000
    - 표준편차 : 1150104.189
- Total Profit
    - mean : 392679.927
    - min : 2.410
    - 제1사분위수 (25%) : 95145.660
    - median : 281655.120
    - 제3사분위수 (75%) : 565962.325
    - max : 1738700.000
    - 표준편차 : 379116.894

2. Univariate analysis

Presentation of key variables from various aspects

2.1 Region

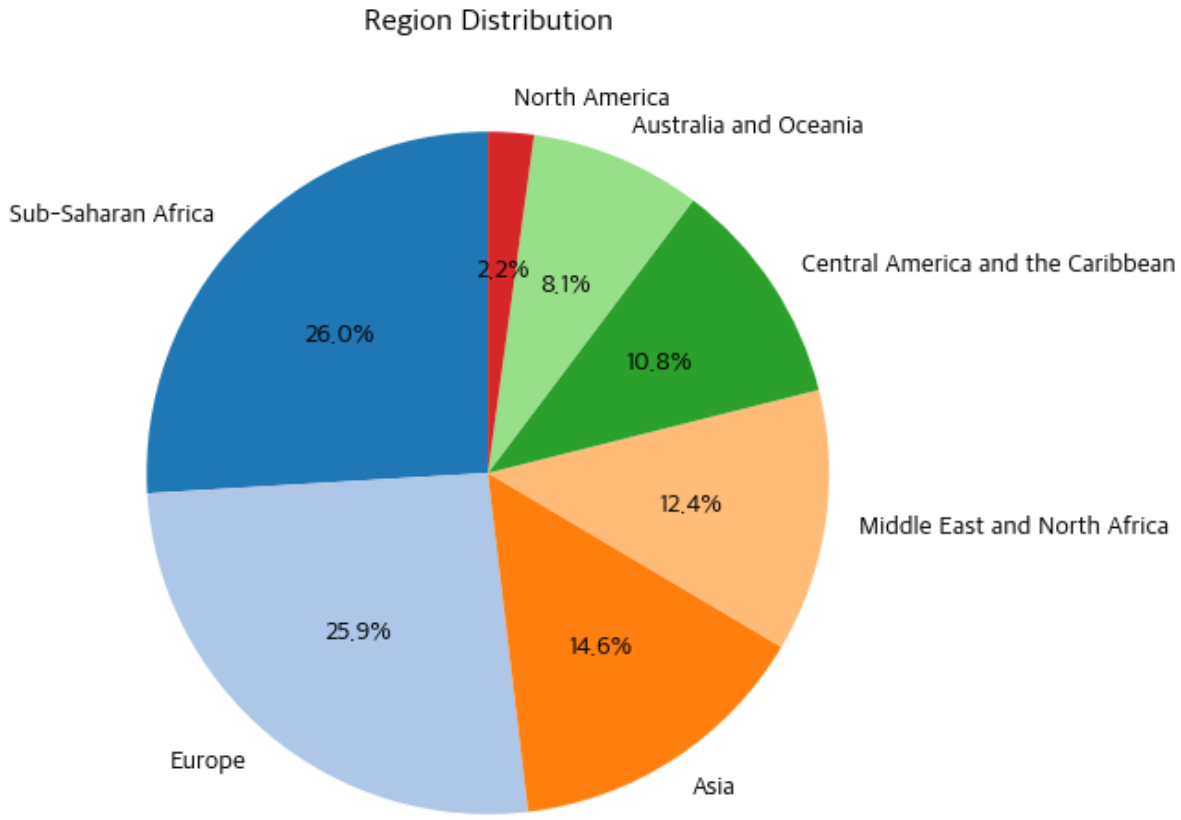

Figure 12. Pie graph of Region Distribution

- 사하라 이남 아프리카 지역이 가장 많이 물건을 구매하고 있음.
- 북아메리카가 가장 적게 물건을 구매함.
- 유럽, 오세아니아 대륙과 달리 아프리카, 아메리카, 아시아는 나눠져있음.
- 중앙아메리카와 캐리비안의 거래는 북아메리카의 거래의 약 5배
- Middle East and North Africa는 중동과 북아프리카를 뜻함.

2.2 Total Profit

총 이익의 통계량

- mean : 392679.927
- min : 2.410
- 제1사분위수 (25%) : 95145.660
- median : 281655.120
- 제3사분위수 (75%) : 565962.325
- max : 1738700.000
- 표준편차 : 379116.894

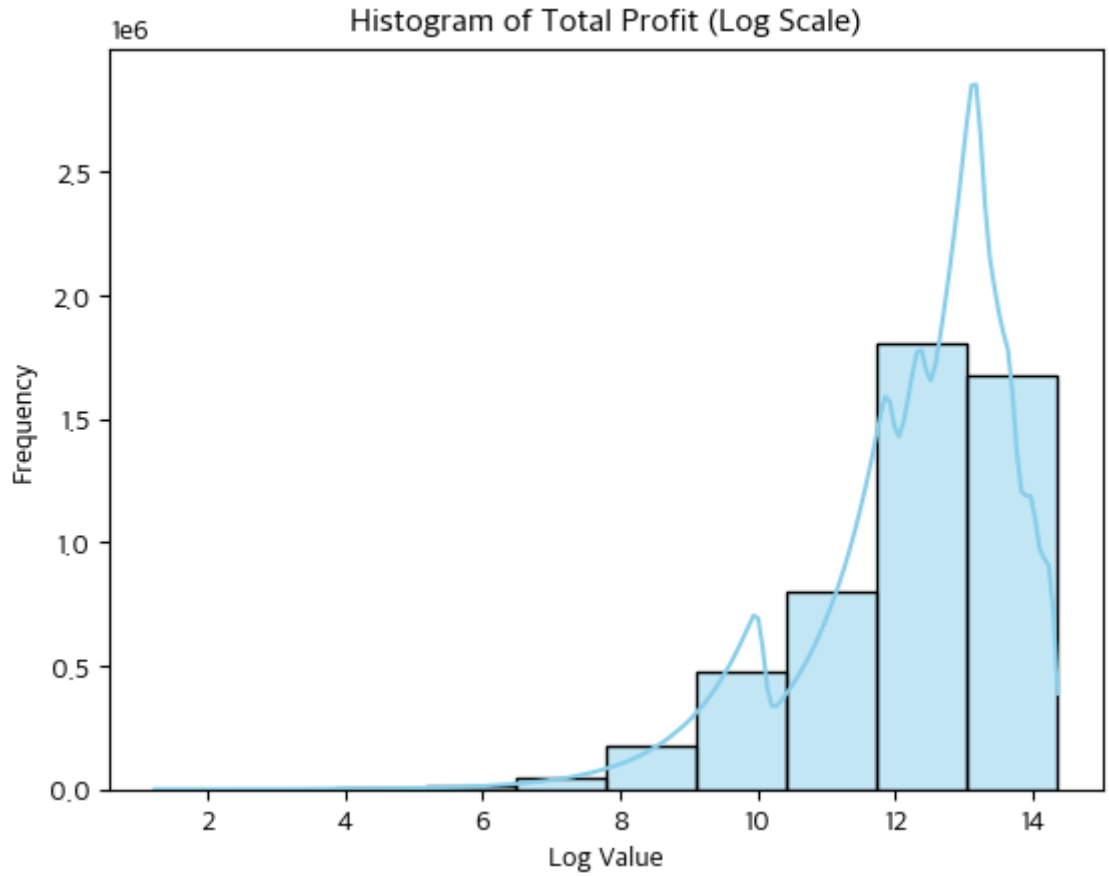

Figure 13. Histogram of Total Profit (Log Scale)

- 하나의 거래로 약 10^{13} 정도의 이익을 가장 많이 얻음.
- 0부터 약 10^7까지의 이익을 얻은 거래는 거의 없음.

3. Multivariate analysis

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

3.1 Correlation

	Order ID	**Units Sold**	**Unit Price**	**Unit Cost**	**Total Revenue**	**Total Cost**	**Total Profit**
Order ID	1.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000
Units Sold	-0.000	1.000	0.000	0.000	0.523	0.471	0.598
Unit Price	-0.000	0.000	1.000	0.986	0.738	0.753	0.577
Unit Cost	-0.000	0.000	0.986	1.000	0.728	0.764	0.505
Total Revenue	-0.000	0.523	0.738	0.728	1.000	0.988	0.881
Total Cost	-0.000	0.471	0.753	0.764	0.988	1.000	0.796

	Order ID	Units Sold	Unit Price	Unit Cost	Total Revenue	Total Cost	Total Profit
Total Profit	-0.000	0.598	0.577	0.505	0.881	0.796	1.000

- 연속형 변수들의 correlation을 나타냄.
- correlation 절대값이 0.5 이상인 관계
    - Units Sold & Total Revenue
    - Units Sold & Total Profit
    - Unit Price, Unit Cost는 Order ID, Units Sold, 자기 자신을 제외한 모든 연속형 열들에 대해 0.5이상의 양의 상관관계를 가짐.
    - Total Revenue가 Total Cost보다 Total Profit에 더 영향이 있음.
    - Total Cost와 Total Revenue는 매우 깊은 양의 상관관계를 가지고 있음.

4. Suggestion

Based on the insights you obtained from the previous stages, propose the potential project idea.

- 사하라 이남 지역이 구매하는 상품들은 재고를 넉넉하게 준비할 것
- 가장 적게 거래하는 북아메리카를 마케팅 대상으로 선정해 광고하기