



Forecasting Consumer Spending Amounts Using Machine Learning and Time Series Analysis

Roy

Research Objective

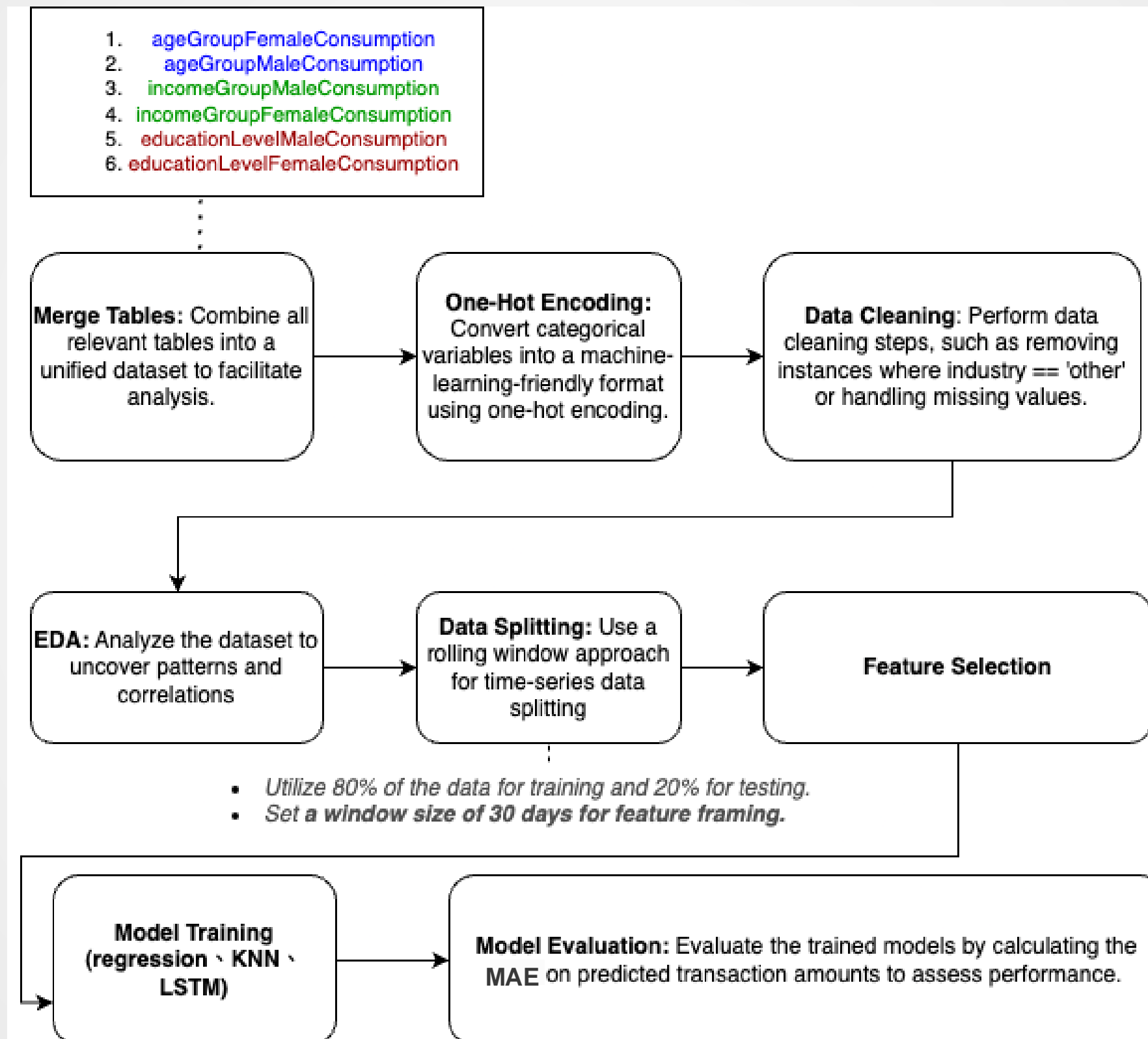
To predict the credit card transaction amounts for the upcoming month across various industries based on different age, income, and education groups, providing precise support for marketing and resource allocation decisions.

Practical Applications

- **Marketing Strategy Optimization**
 - Deliver customized promotional campaigns targeting specific income and education groups to improve market response rates.
- **Credit Card Business Innovation**
 - Design specialized credit card products for specific age and income groups (e.g., education industry cards, entertainment spending cards).
 - Adjust credit limits based on consumption predictions to mitigate financial risks.



Flow chart



Data (data.gov.tw)

- ageGroupFemaleConsumption : **各年齡層女性持卡人於各行業別總簽帳金額及筆數**
- ageGroupMaleConsumption : **各年齡層男性持卡人於各行業別總簽帳金額及筆數**
- incomeGroupMaleConsumption : **各年收入族群男性持卡人於各行業別總簽帳金額及筆數**
- incomeGroupFemaleConsumption : **各年收入族群女性持卡人於各行業別總簽帳金額及筆數**
- educationLevelMaleConsumption : **各教育程度男性持卡人於六都消費樣態**
- educationLevelFemaleConsumption : **各教育程度女性持卡人於六都消費樣態**



Merge Tables -> One-Hot Encoding -> Data Cleaning -> Exploratory Data Analysis (EDA) -> Data Splitting -> Feature Selection -> Model Training -> Model Evaluation





Merged table : Age Group

combined : 年齡層 (25088 instances)

	年月	信用卡產業別	性別	年齡層	信用卡交易筆數	信用卡交易金額 [新臺幣]
0	2014-01-01	食	2	未滿20歲	6367	5630047
12556	2014-01-01	食	1	75(含)-80歲	36983	59655595
12557	2014-01-01	食	1	80(含)歲以上	30221	52358455
12558	2014-01-01	衣	1	未滿20歲	1225	3372107
12559	2014-01-01	衣	1	20(含)-25歲	18667	47403285
...
12514	2024-08-01	文教康樂	2	75(含)-80歲	14103	118381107
12515	2024-08-01	文教康樂	2	80(含)歲以上	7022	52667892
12516	2024-08-01	百貨	2	未滿20歲	242728	161364361
12518	2024-08-01	百貨	2	25(含)-30歲	3178874	3034562989
25087	2024-08-01	其他	1	80(含)歲以上	37245	137102471

25088 rows x 6 columns



Merged table : Income Group

combined : 年收入 (14336 instances)

	年月	信用卡產業別	性別	年收入	信用卡交易筆數	信用卡交易金額 [新臺幣]
0	2014-01-01	食	2	未達50萬	4602444	6589392709
7178	2014-01-01	衣	1	75(含)-100萬	69409	195212721
7177	2014-01-01	衣	1	50(含)-75萬	167294	449058900
7176	2014-01-01	衣	1	未達50萬	241377	643675362
7175	2014-01-01	食	1	200(含)萬以上	261193	912427601
...
7143	2024-08-01	行	2	200(含)萬以上	531201	778191525
7142	2024-08-01	行	2	175(含)-200萬	118659	137239443
7141	2024-08-01	行	2	150(含)-175萬	209030	279408938
7139	2024-08-01	行	2	100(含)-125萬	571821	664353856
14335	2024-08-01	其他	1	200(含)萬以上	715102	10281815851

14336 rows x 6 columns



Merged table : Education Level

combined : 教育程度 (64512 instances)

	年月	信用卡產業別	性別	教育程度類別	信用卡交易筆數	信用卡交易金額 [新臺幣]
0	2014-01-01	食	2	博士	17328	23014654
32332	2014-01-01	百貨	1	高中高職	15146	64613060
32331	2014-01-01	百貨	1	專科	15134	54999974
32330	2014-01-01	百貨	1	大學	37657	147525955
32329	2014-01-01	百貨	1	碩士	15070	55733574
...
32165	2024-08-01	百貨	2	其他	318001	532201566
32164	2024-08-01	百貨	2	高中高職	416935	828824990
32163	2024-08-01	百貨	2	專科	314957	612001223
32176	2024-08-01	食	2	高中高職	78114	89943810
64511	2024-08-01	其他	1	其他	43251	108540541

64512 rows x 6 columns

One-hot encoding : age group

```
Info about ageGroupCombined:
<class 'pandas.core.frame.DataFrame'>
Index: 21504 entries, 0 to 12518
Data columns (total 26 columns):
#   Column                                     Non-Null Count  Dtype
---  -
0   Date                                     21504 non-null  datetime64[ns]
1   Transaction Count                       21504 non-null  int64
2   Transaction Amount (NTD)                21504 non-null  int64
3   Industry_Clothing                       21504 non-null  bool
4   Industry_Department_Store               21504 non-null  bool
5   Industry_Education_Entertainment        21504 non-null  bool
6   Industry_Food                           21504 non-null  bool
7   Industry_Housing                        21504 non-null  bool
8   Industry_Others                         21504 non-null  bool
9   Industry_Transportation                 21504 non-null  bool
10  Gender_Female                           21504 non-null  bool
11  Gender_Male                             21504 non-null  bool
12  AgeGroup_20-25                           21504 non-null  bool
13  AgeGroup_25-30                           21504 non-null  bool
14  AgeGroup_30-35                           21504 non-null  bool
15  AgeGroup_35-40                           21504 non-null  bool
16  AgeGroup_40-45                           21504 non-null  bool
17  AgeGroup_45-50                           21504 non-null  bool
18  AgeGroup_50-55                           21504 non-null  bool
19  AgeGroup_55-60                           21504 non-null  bool
20  AgeGroup_60-65                           21504 non-null  bool
21  AgeGroup_65-70                           21504 non-null  bool
22  AgeGroup_70-75                           21504 non-null  bool
23  AgeGroup_75-80                           21504 non-null  bool
24  AgeGroup_Above 80                         21504 non-null  bool
25  AgeGroup_Under 20                         21504 non-null  bool
dtypes: bool(23), datetime64[ns](1), int64(2)
```



One-hot encoding : income group

```
Info about incomeGroupCombined:
<class 'pandas.core.frame.DataFrame'>
Index: 12288 entries, 0 to 7139
Data columns (total 20 columns):
#   Column                                     Non-Null Count  Dtype
---  -
0   Date                                     12288 non-null  datetime64[ns]
1   Transaction Count                       12288 non-null  int64
2   Transaction Amount (NTD)                12288 non-null  int64
3   Industry_Clothing                       12288 non-null  bool
4   Industry_Department_Store               12288 non-null  bool
5   Industry_Education_Entertainment        12288 non-null  bool
6   Industry_Food                           12288 non-null  bool
7   Industry_Housing                        12288 non-null  bool
8   Industry_Others                         12288 non-null  bool
9   Industry_Transportation                  12288 non-null  bool
10  Gender_Female                           12288 non-null  bool
11  Gender_Male                             12288 non-null  bool
12  IncomeGroup_1.25M-1.5M                   12288 non-null  bool
13  IncomeGroup_1.5M-1.75M                   12288 non-null  bool
14  IncomeGroup_1.75M-2M                     12288 non-null  bool
15  IncomeGroup_1M-1.25M                     12288 non-null  bool
16  IncomeGroup_500k-750k                    12288 non-null  bool
17  IncomeGroup_750k-1M                      12288 non-null  bool
18  IncomeGroup_Above 2M                     12288 non-null  bool
19  IncomeGroup_Below 500k                   12288 non-null  bool
dtypes: bool(17), datetime64[ns](1), int64(2)
memory usage: 588.0 KB
None
```



One-hot encoding : education level

```
Info about educationLevelCombined:
<class 'pandas.core.frame.DataFrame'>
Index: 55296 entries, 0 to 32176
Data columns (total 18 columns):
#   Column                                     Non-Null Count  Dtype
---  -
0   Date                                     55296 non-null  datetime64[ns]
1   Transaction Count                       55296 non-null  int64
2   Transaction Amount (NTD)               55296 non-null  int64
3   Industry_Clothing                      55296 non-null  bool
4   Industry_Department_Store              55296 non-null  bool
5   Industry_Education_Entertainment       55296 non-null  bool
6   Industry_Food                          55296 non-null  bool
7   Industry_Housing                      55296 non-null  bool
8   Industry_Others                       55296 non-null  bool
9   Industry_Transportation                55296 non-null  bool
10  Gender_Female                         55296 non-null  bool
11  Gender_Male                          55296 non-null  bool
12  EducationLevel_Associate               55296 non-null  bool
13  EducationLevel_Bachelor                 55296 non-null  bool
14  EducationLevel_Doctorate                55296 non-null  bool
15  EducationLevel_High School              55296 non-null  bool
16  EducationLevel_Master                   55296 non-null  bool
17  EducationLevel_Other                   55296 non-null  bool
dtypes: bool(15), datetime64[ns](1), int64(2)
memory usage: 2.5 MB
None
```

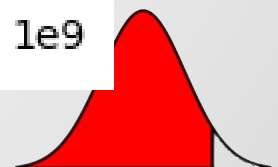
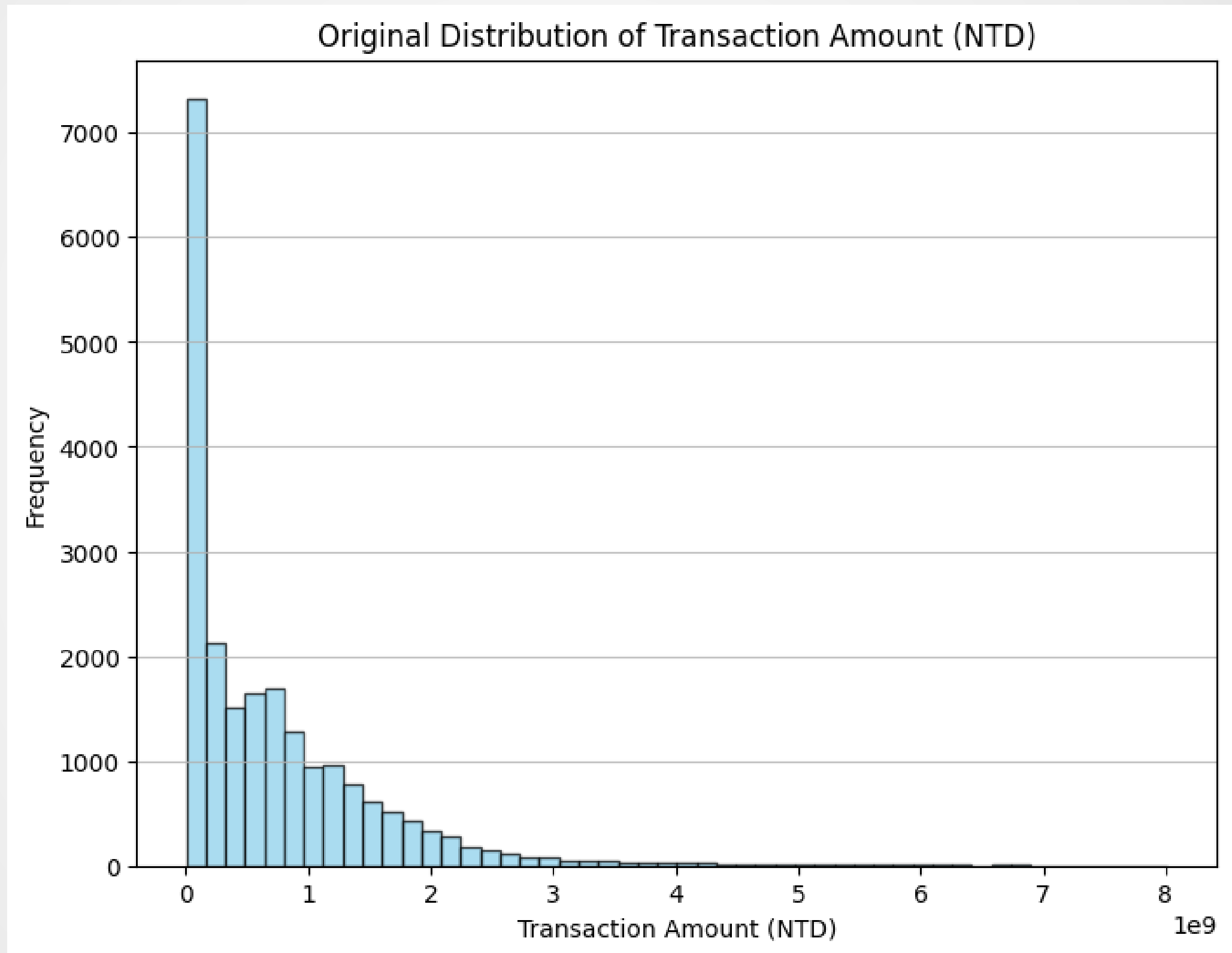


Data cleaning ('other' features)

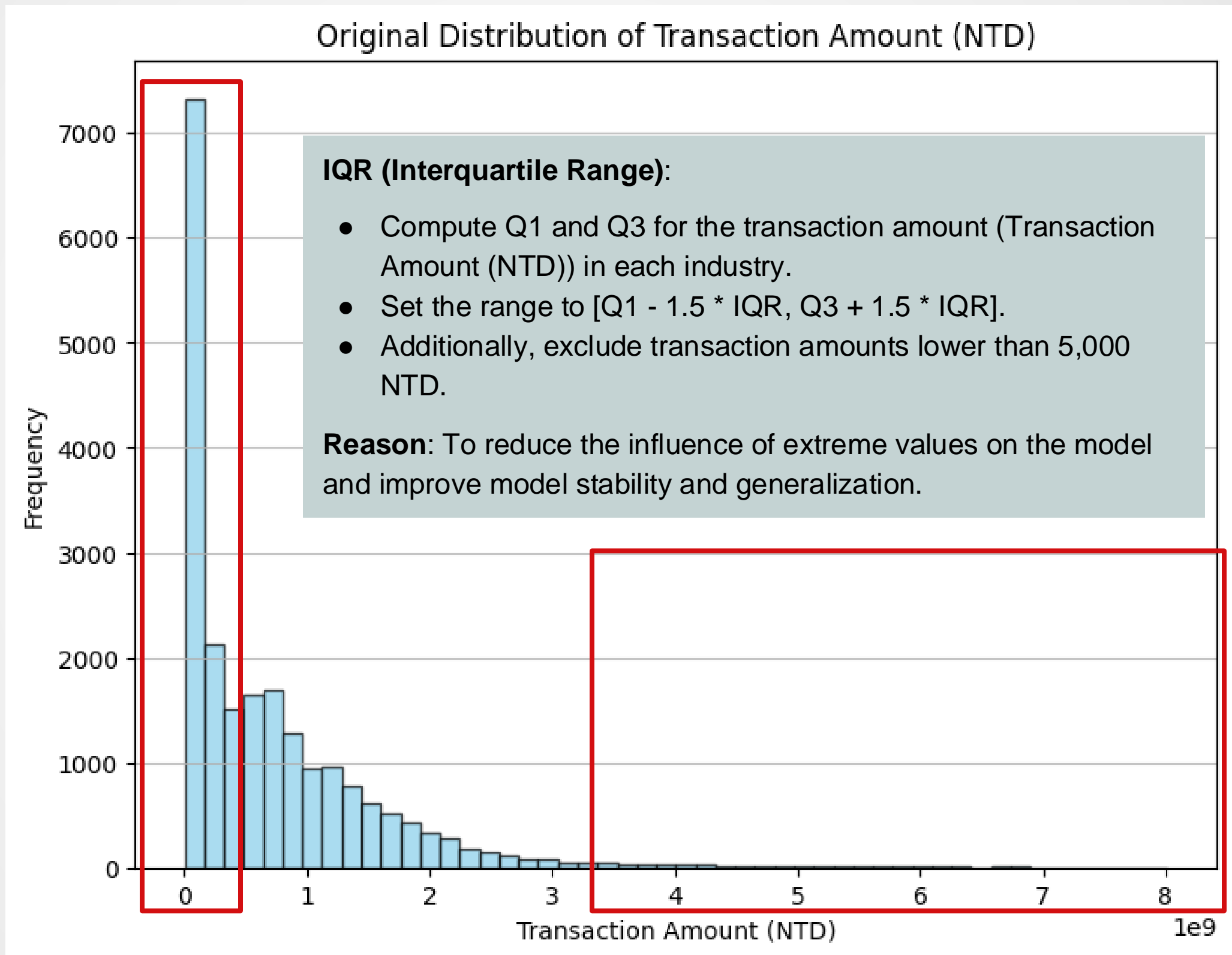
```
Info about incomeGroupCombined:
<class 'pandas.core.frame.DataFrame'>
Index: 12288 entries, 0 to 7139
Data columns (total 20 columns):
#   Column                                     Non-Null Count  Dtype
---  -
0   Date                                     12288 non-null  datetime64[ns]
1   Transaction Count                       12288 non-null  int64
2   Transaction Amount (NTD)                12288 non-null  int64
3   Industry_Clothing                       12288 non-null  bool
4   Industry_Department_Store               12288 non-null  bool
5   Industry_Education_Entertainment        12288 non-null  bool
6   Industry_Food                           12288 non-null  bool
7   Industry_Housing                        12288 non-null  bool
8   Industry_Others                         12288 non-null  bool
9   Industry_Transportation                  12288 non-null  bool
10  Gender_Female                           12288 non-null  bool
11  Gender_Male                             12288 non-null  bool
12  IncomeGroup_1.25M-1.5M                   12288 non-null  bool
13  IncomeGroup_1.5M-1.75M                   12288 non-null  bool
14  IncomeGroup_1.75M-2M                     12288 non-null  bool
15  IncomeGroup_1M-1.25M                     12288 non-null  bool
16  IncomeGroup_500k-750k                    12288 non-null  bool
17  IncomeGroup_750k-1M                      12288 non-null  bool
18  IncomeGroup_Above 2M                     12288 non-null  bool
19  IncomeGroup_Below 500k                   12288 non-null  bool
dtypes: bool(17), datetime64[ns](1), int64(2)
memory usage: 588.0 KB
None
```



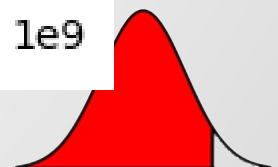
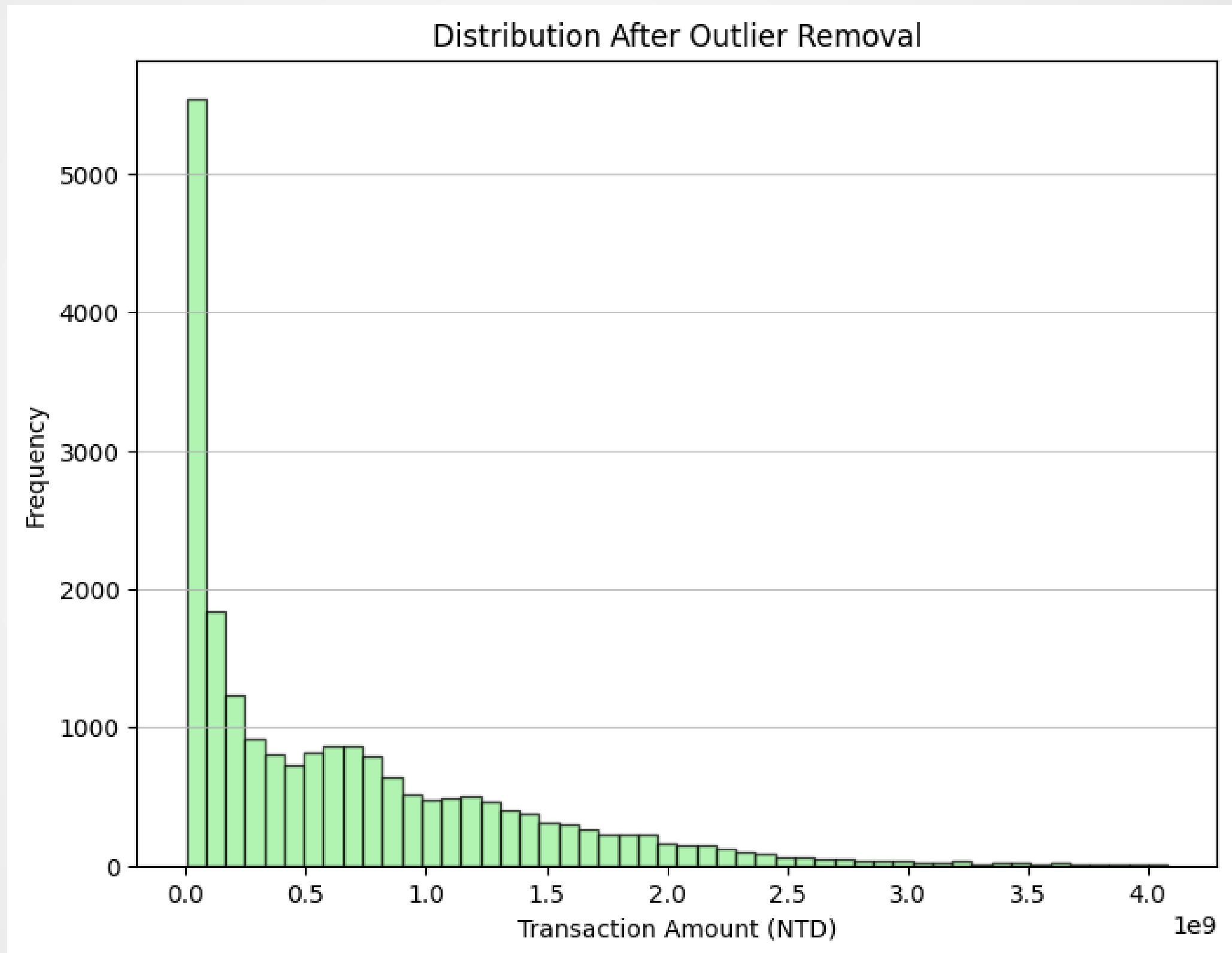
Data cleaning(outlier)



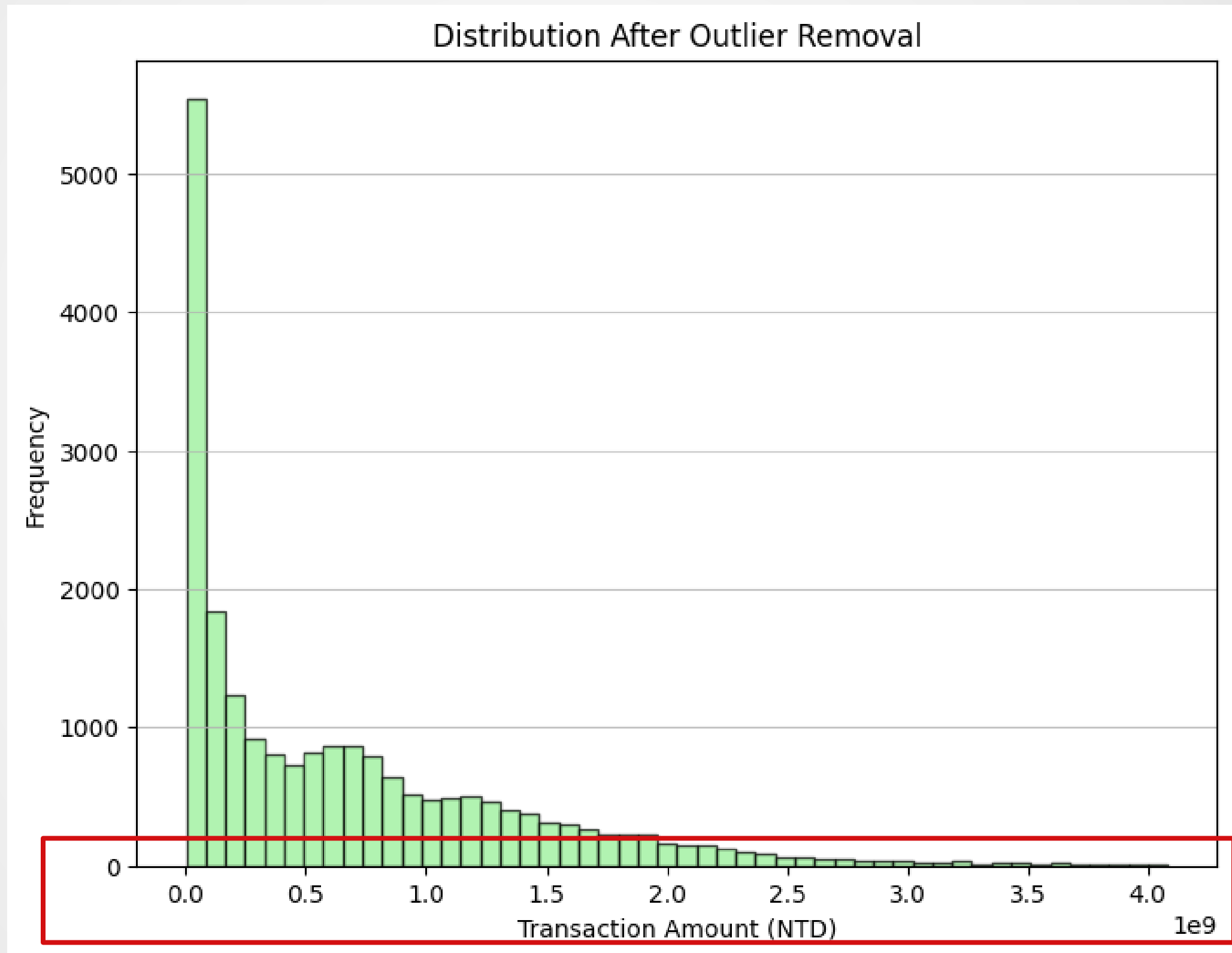
Data cleaning(outlier)



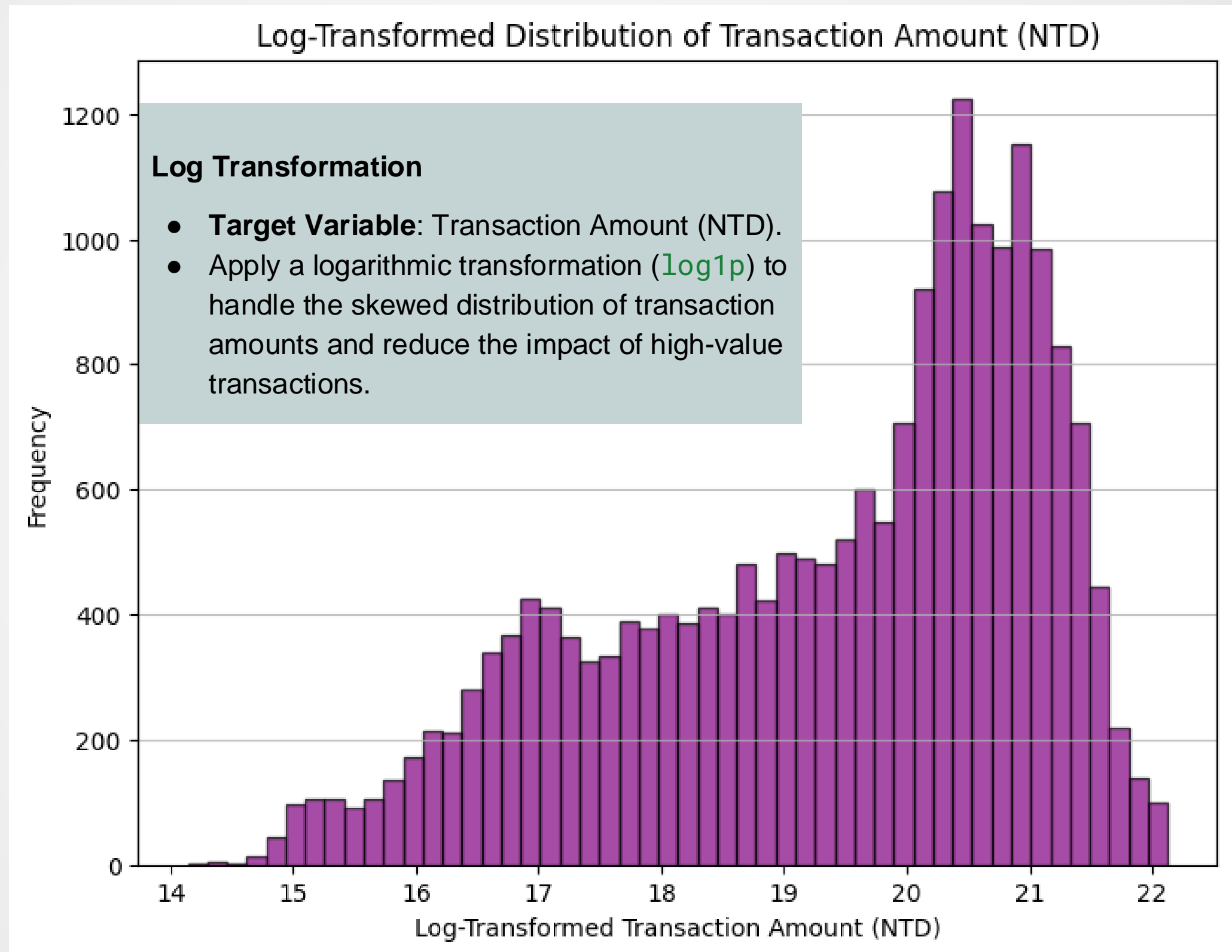
Data cleaning(outlier)



Data cleaning(transform)



Data cleaning(transform)



EDA

1. Do key variables (e.g., Age, Income, Education) significantly impact the target variable (transaction amount)?

1. Do industries influence the relationship between key variables and the target?

1. Does the distribution of industries across different key variables show consistent patterns in their contribution to transaction amounts?

1. Does the target variable exhibit any cyclical patterns over time?



EDA

1. Correlation Analysis of **Key Variables** with **Transaction Amounts**

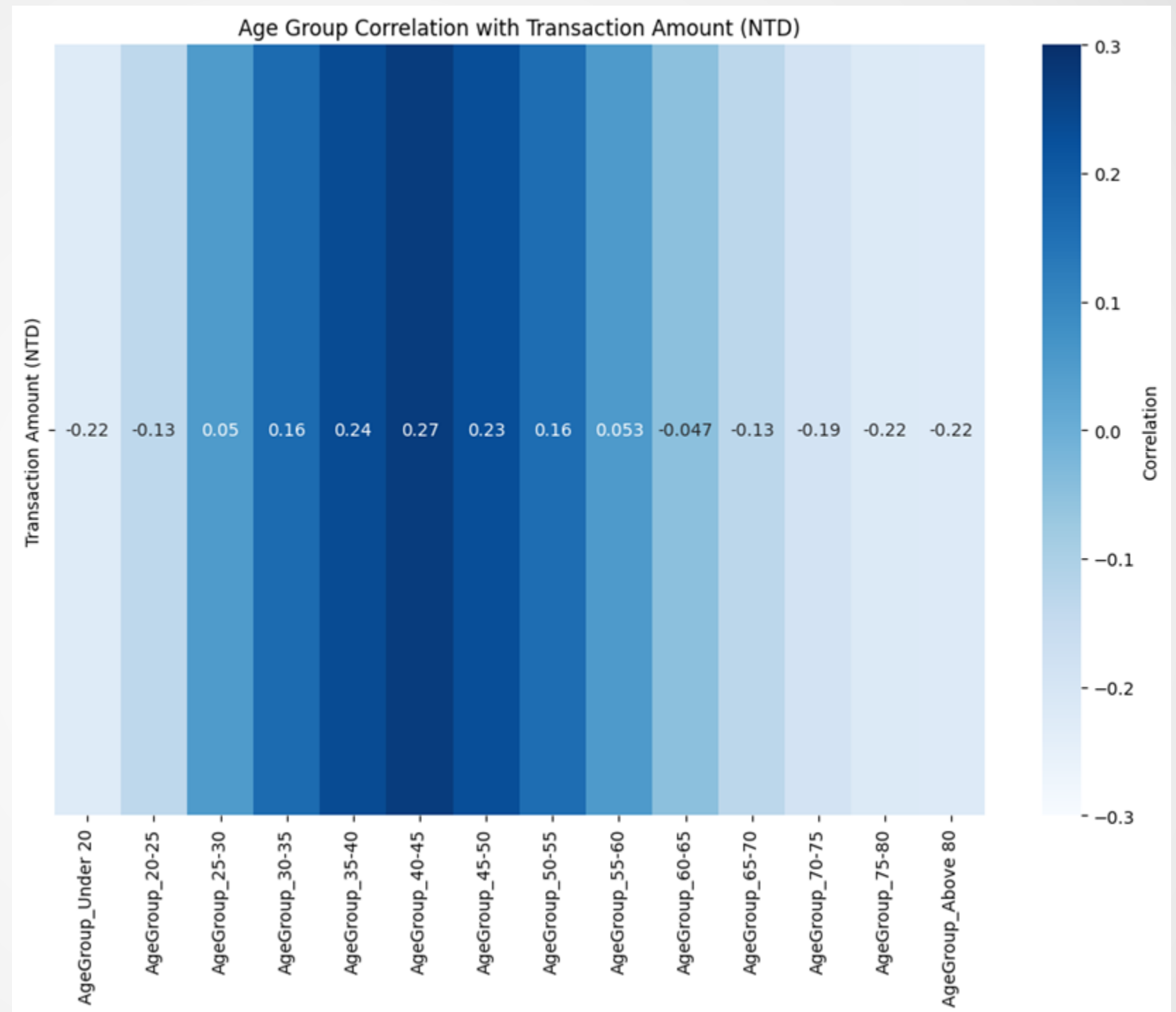
1. Correlation Analysis of **Industry Categories** with **Transaction Amounts**

1. Impact of **Key Variables** (Age, Income, Education) and **Industry Categories** on **Transaction Amounts**

1. **Time Series Analysis** to Identify Cycles and Evaluate the Importance of Dates for Accurate Data Splitting



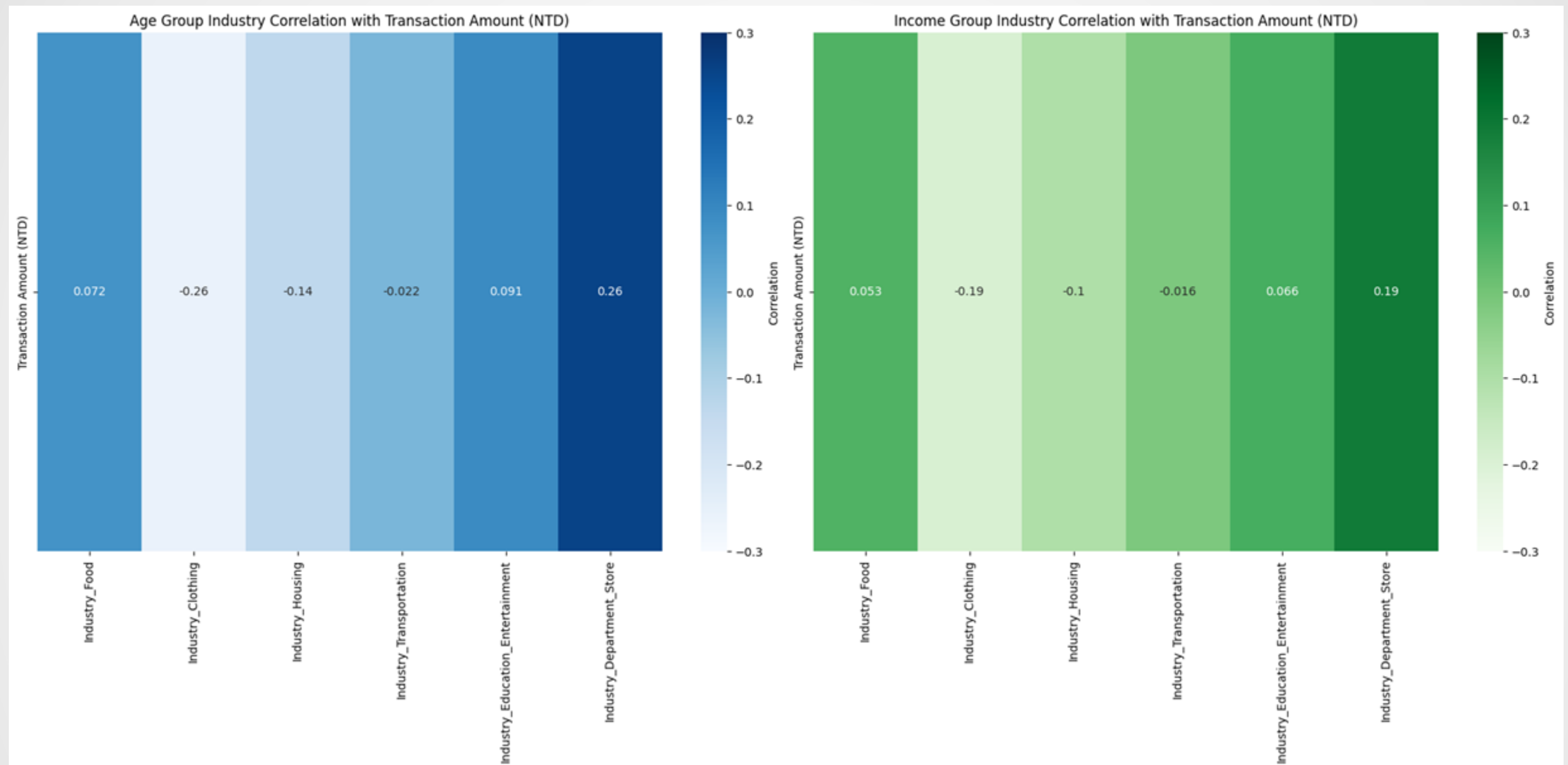
EDA : age group



1. Correlation Analysis of Key Variables with Transaction Amounts



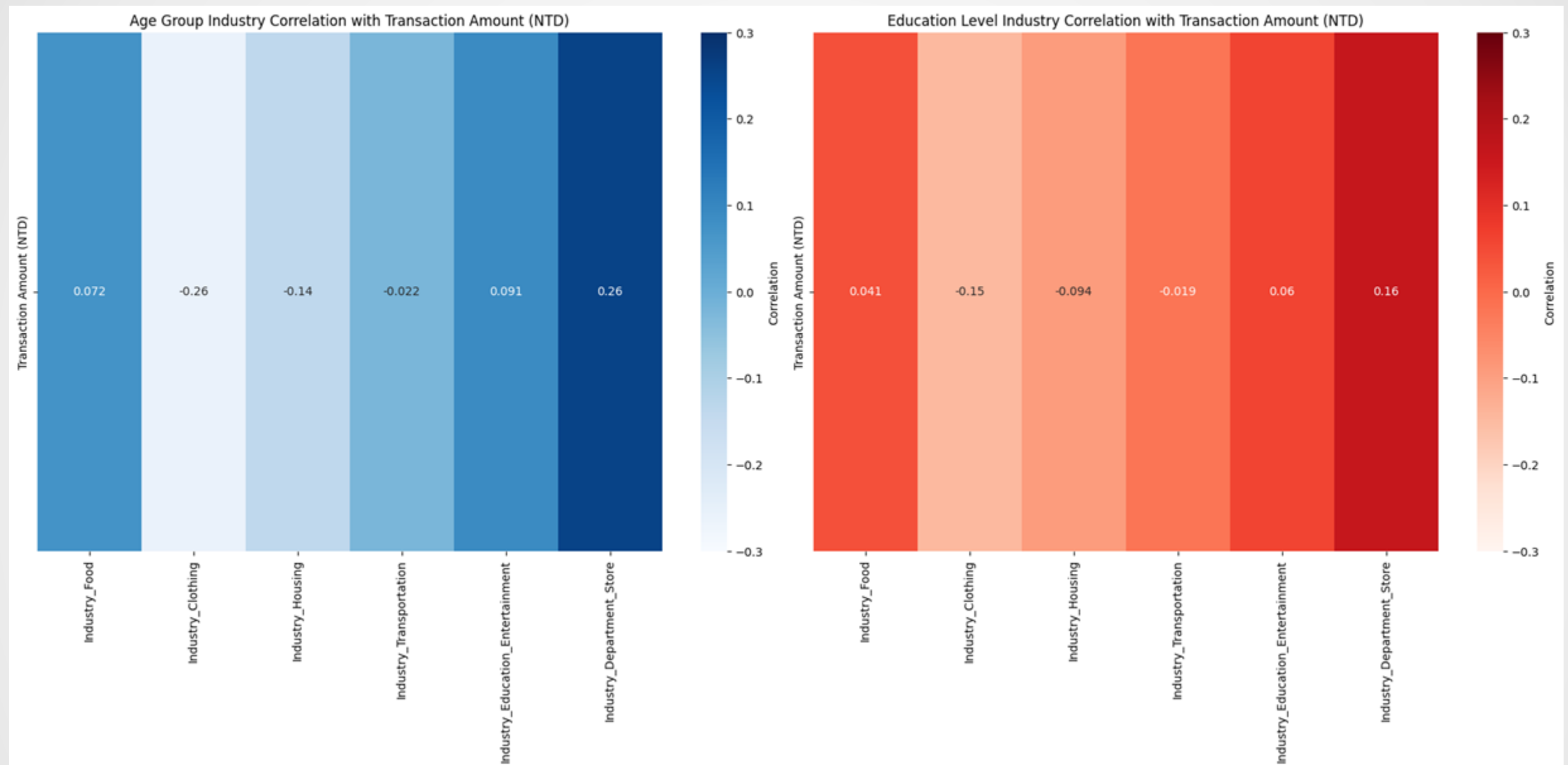
EDA :



2. Correlation of Industry Categories with Transaction Amounts



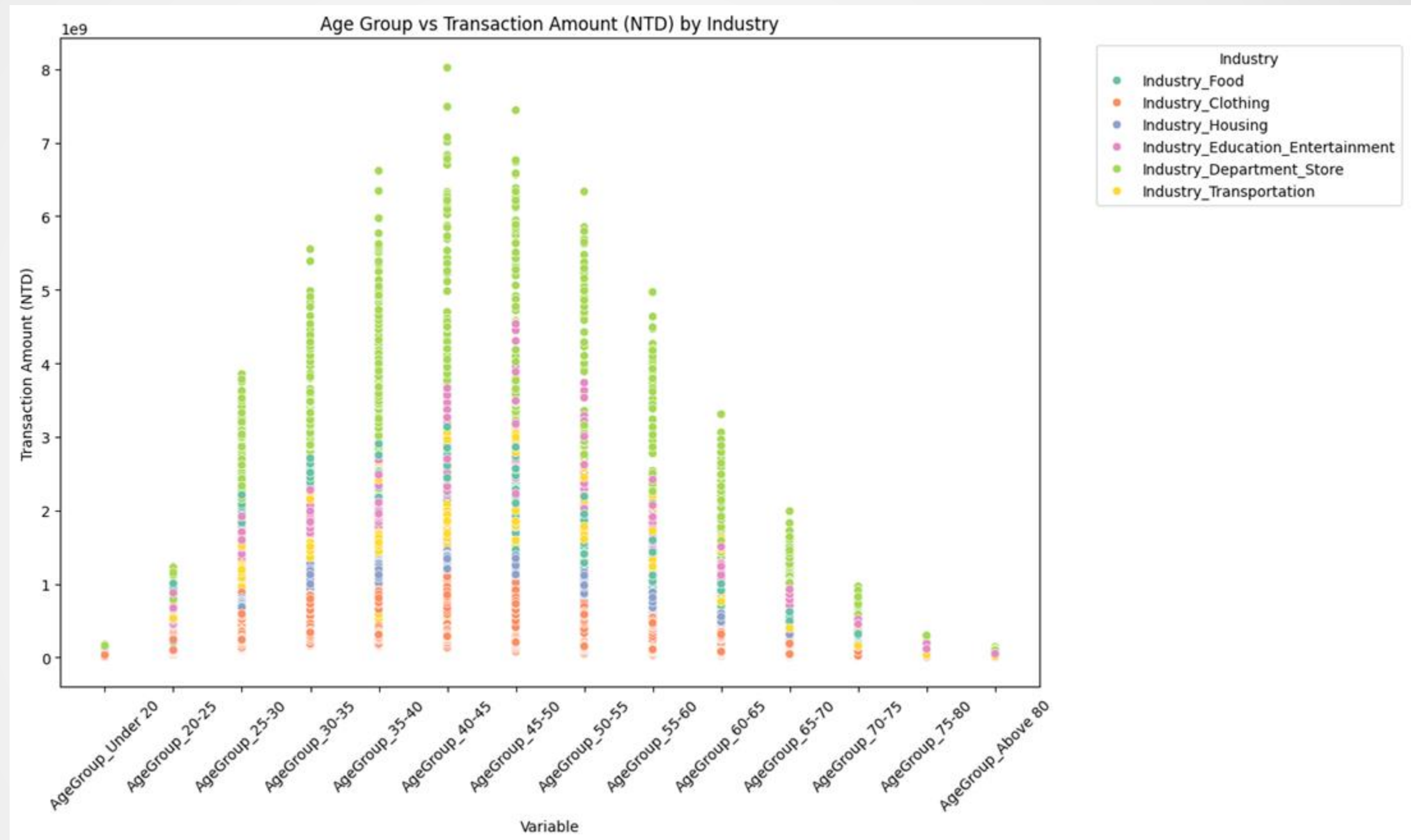
EDA :



2. Correlation of Industry Categories with Transaction Amounts



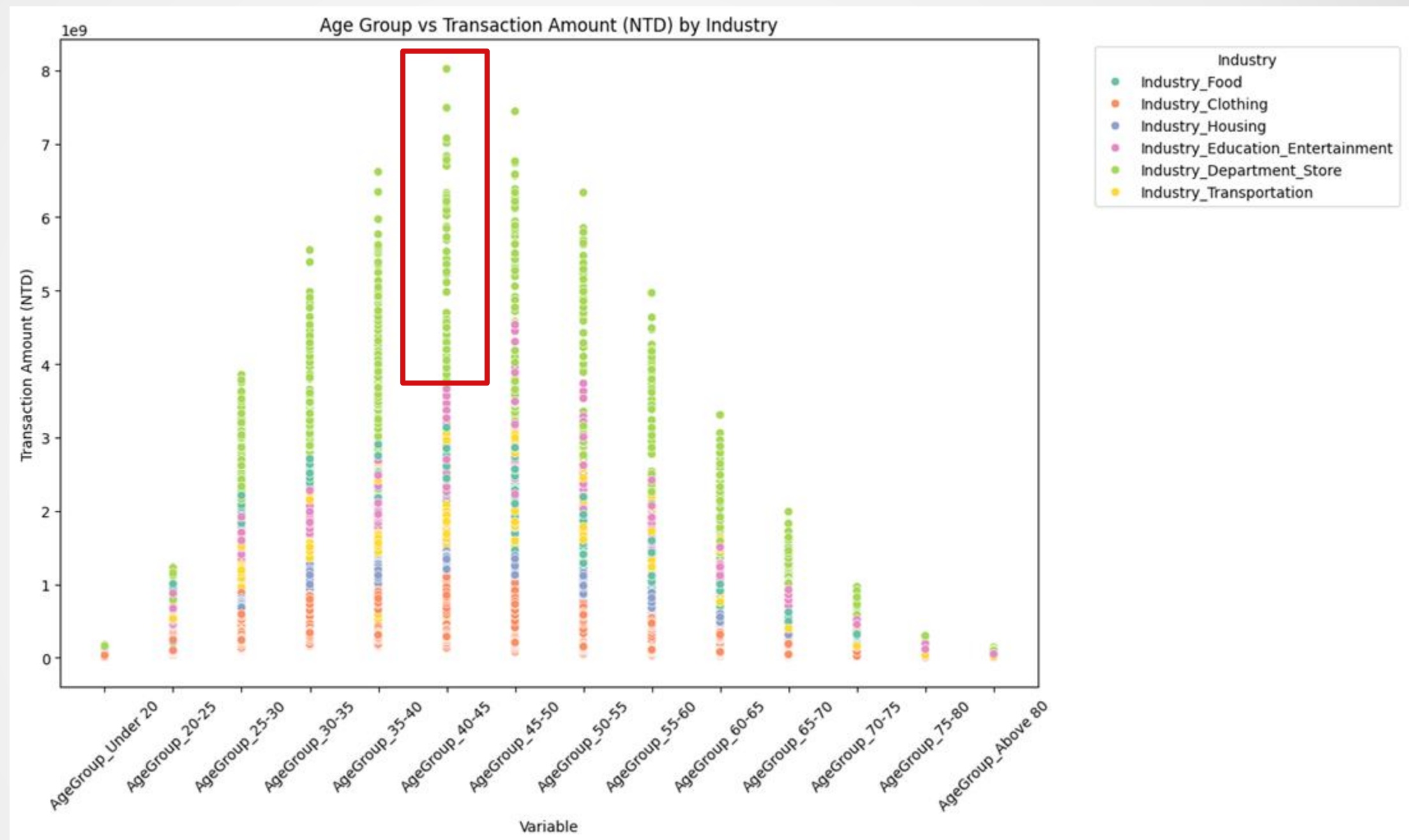
EDA : age group



3. Impact of **Key Variables** (Age, Income, Education) and **Industry Categories** on Transaction Amounts



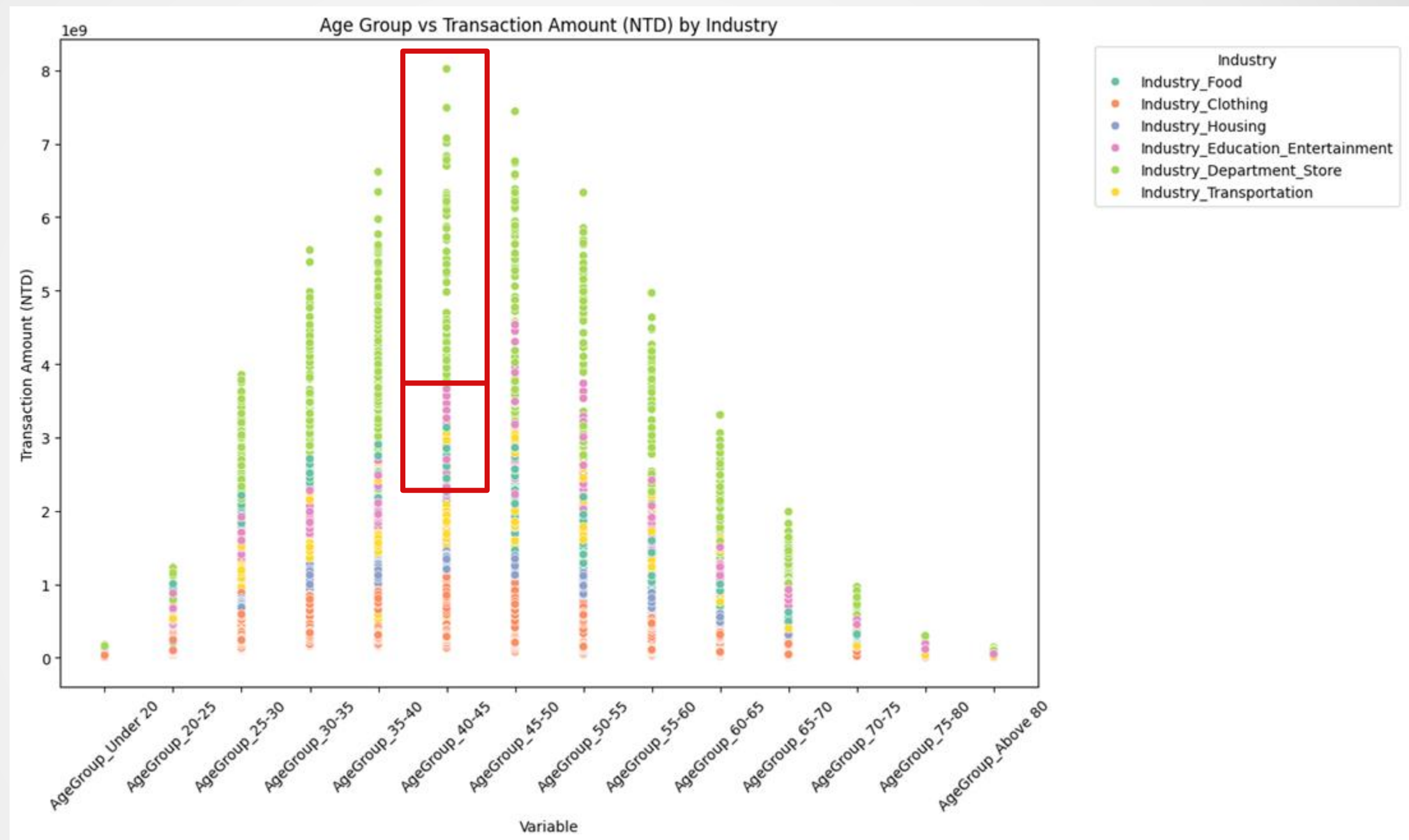
EDA : age group



3. Impact of **Key Variables** (Age, Income, Education) and **Industry Categories** on Transaction Amounts



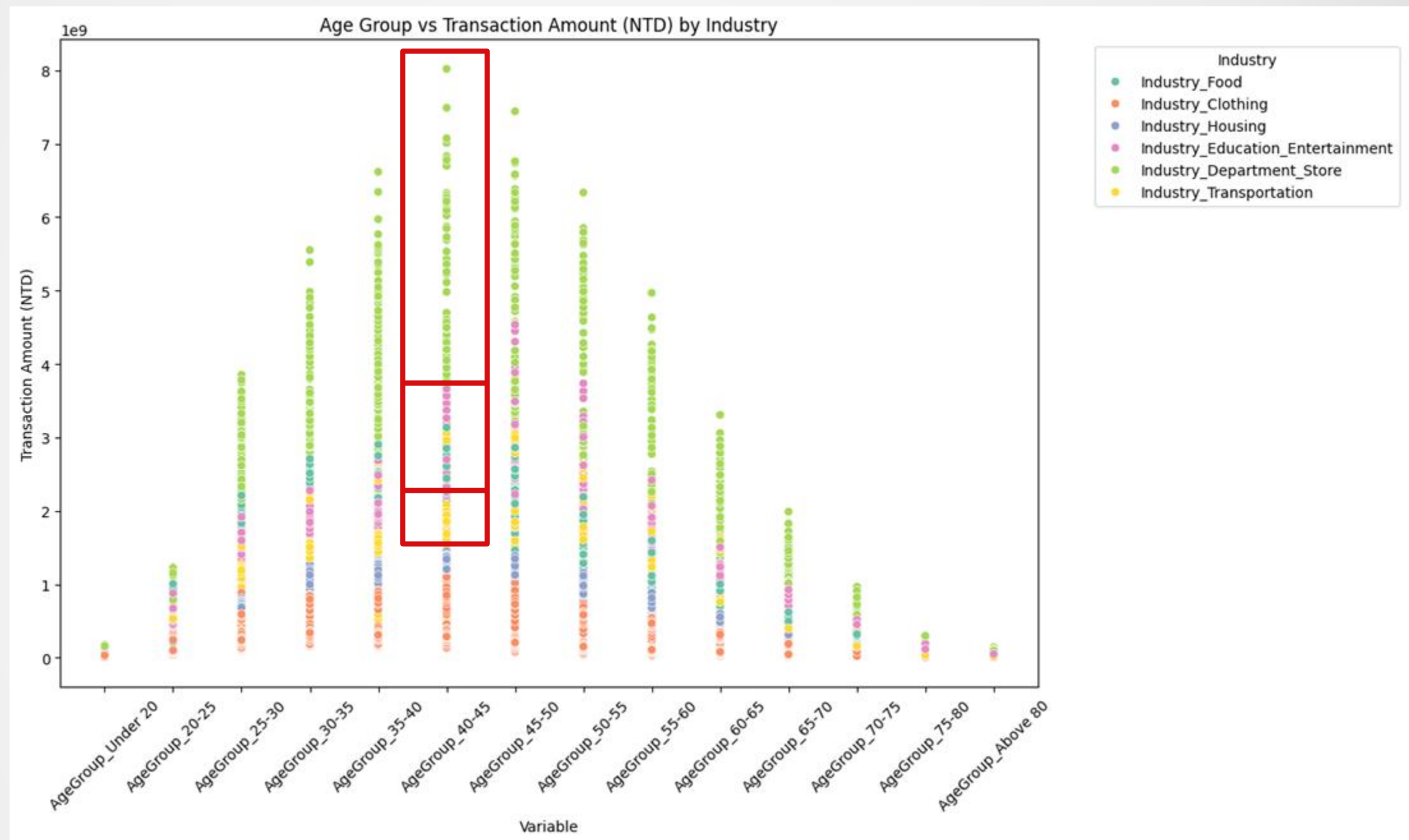
EDA : age group



3. Impact of **Key Variables** (Age, Income, Education) and **Industry Categories** on Transaction Amounts



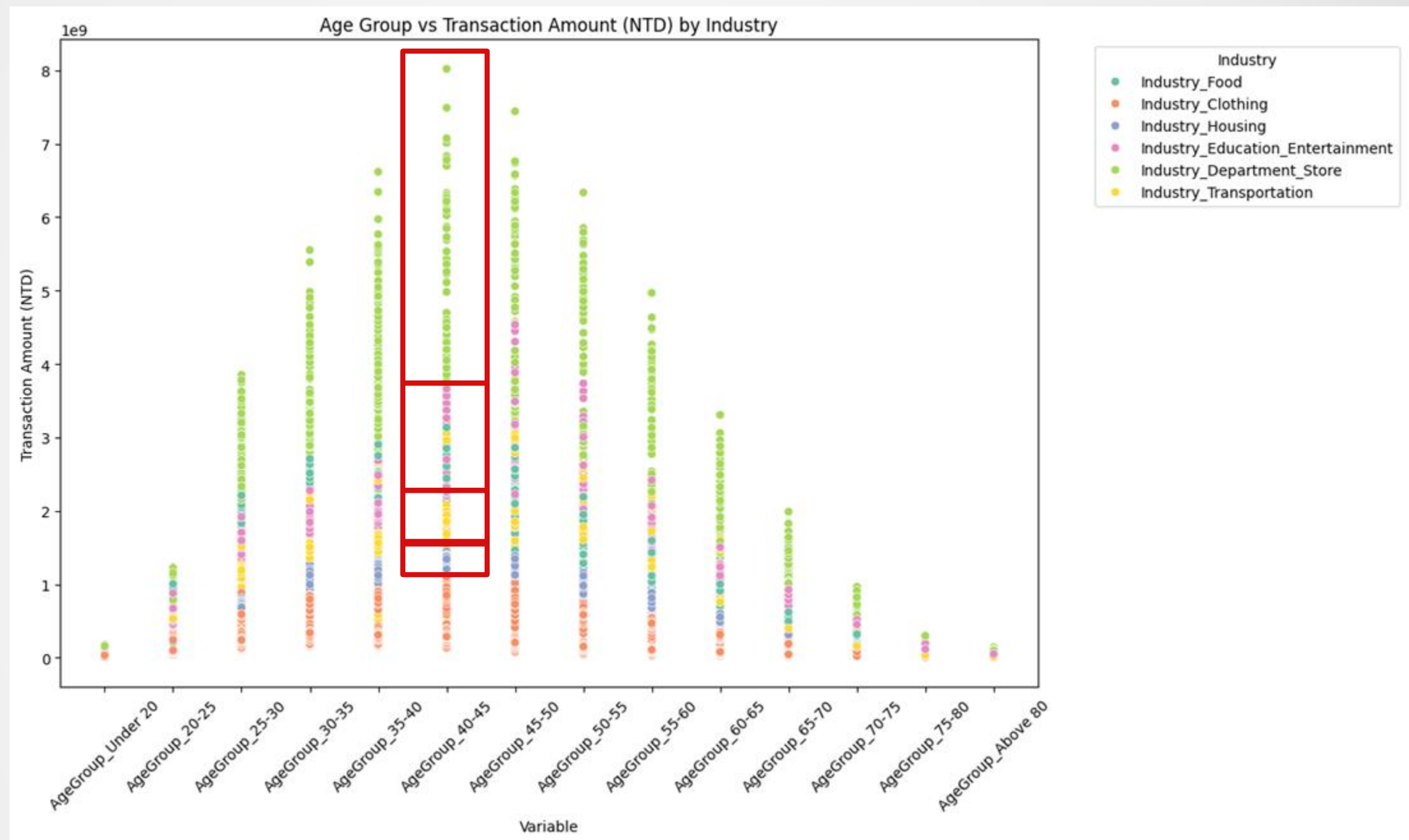
EDA : age group



3. Impact of **Key Variables** (Age, Income, Education) and **Industry Categories** on Transaction Amounts



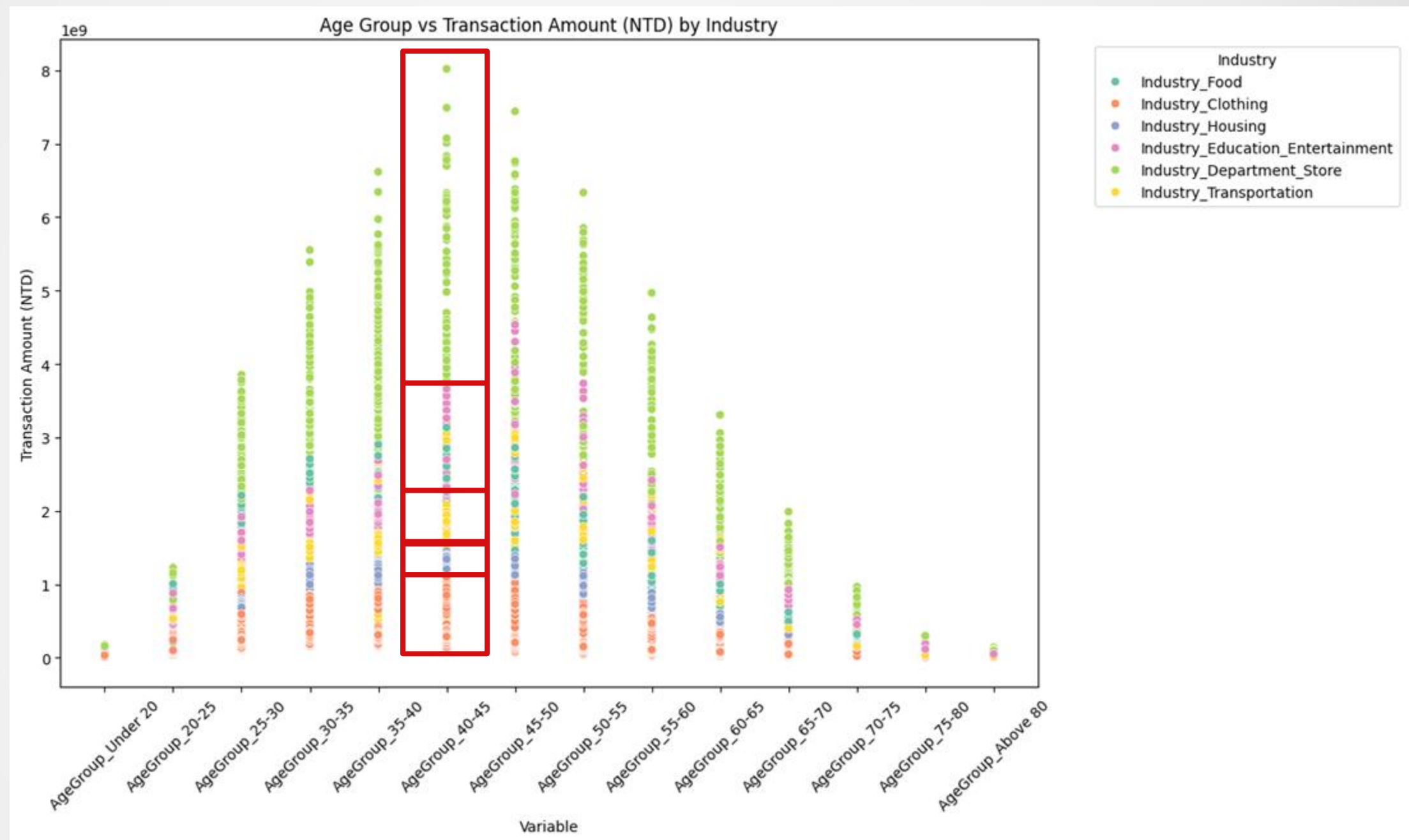
EDA : age group



3. Impact of **Key Variables** (Age, Income, Education) and **Industry Categories** on Transaction Amounts



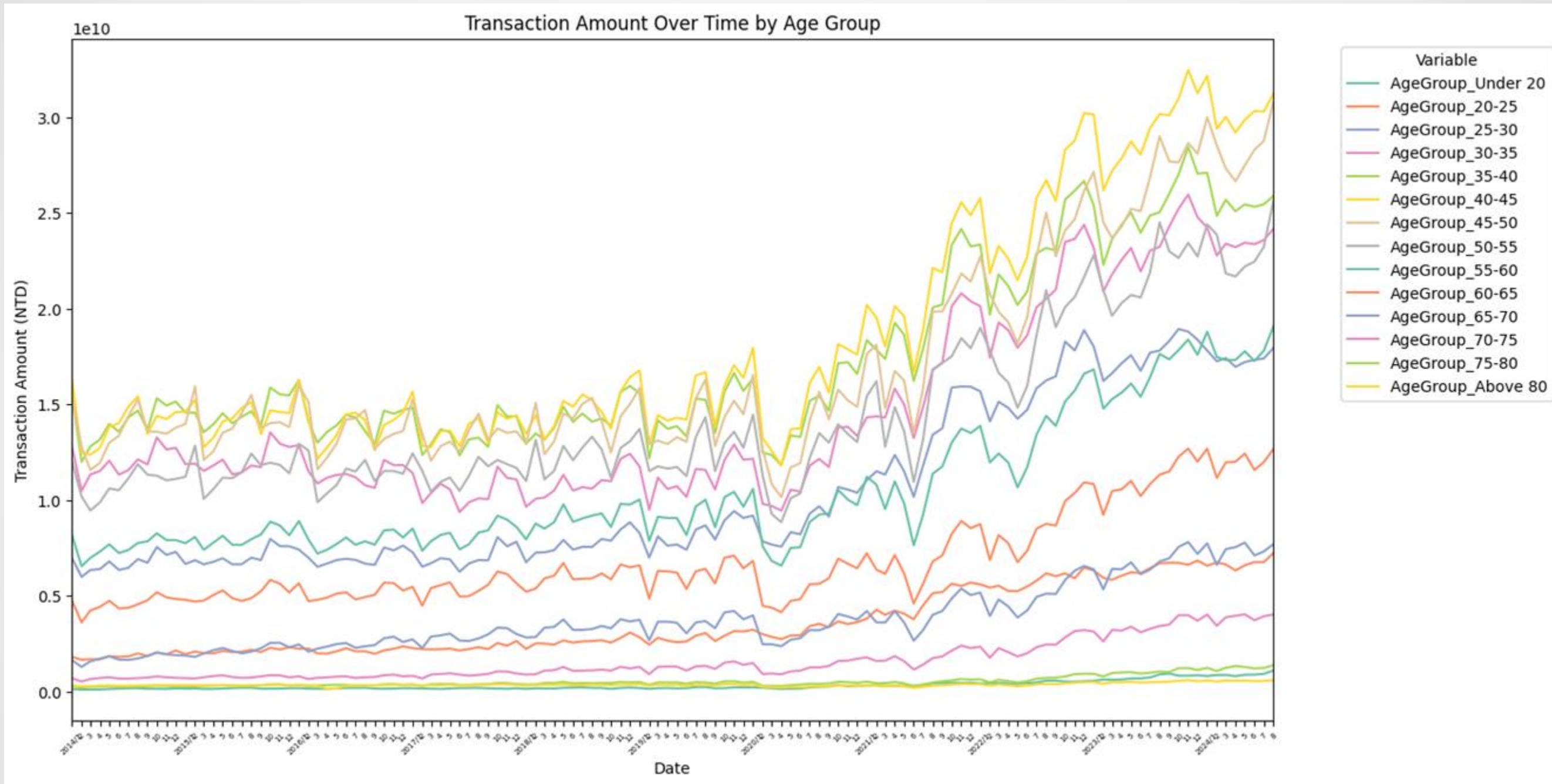
EDA : age group



3. Impact of **Key Variables** (Age, Income, Education) and **Industry Categories** on Transaction Amounts



EDA : age group



4. Time Series Analysis to Identify Cycles and Evaluate the Importance of Dates for Accurate Data Splitting



Data Splitting (industries)

	date	f2	f3	...	fn	target
x1						
x2						
...						
x30						

industry

clothing、department store、education、food
、housing、transportation

	date	f2	f3	...	fn-5	target
x1						
x2						
...						
x30						

*6 data sets



Model training

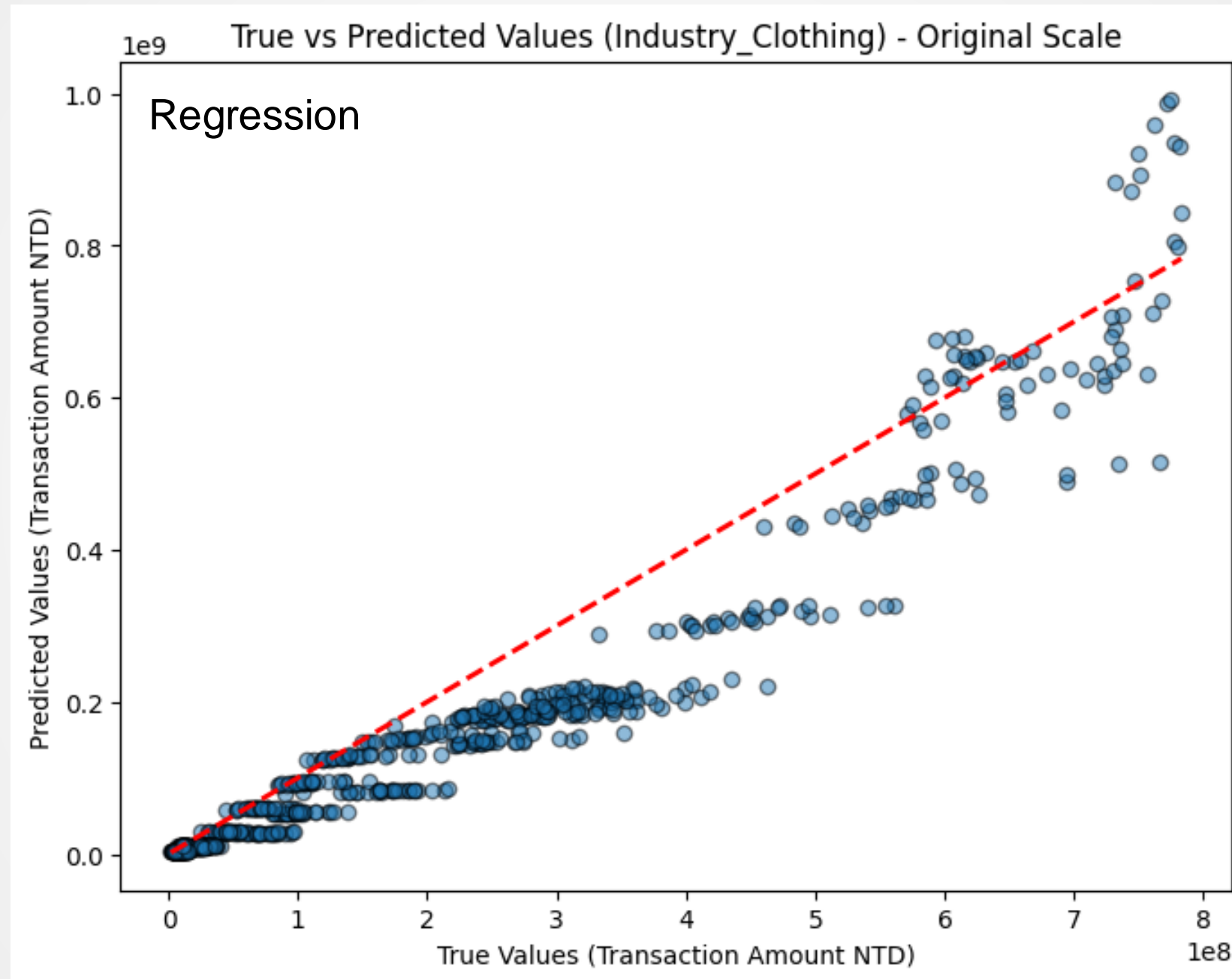
Model Training

- regression
- KNN
- LSTM

Model Evaluation: Evaluate the trained models by calculating the **MAE** on predicted transaction amounts to assess performance.



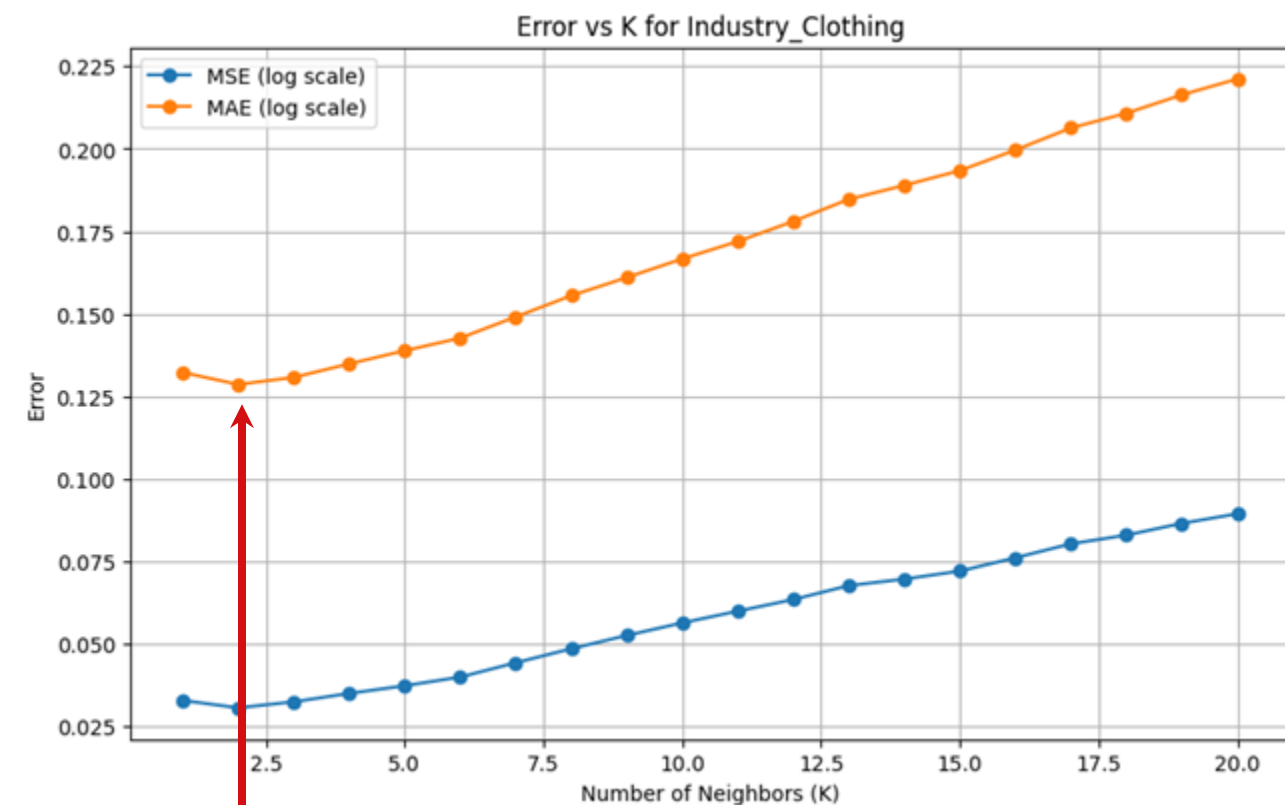
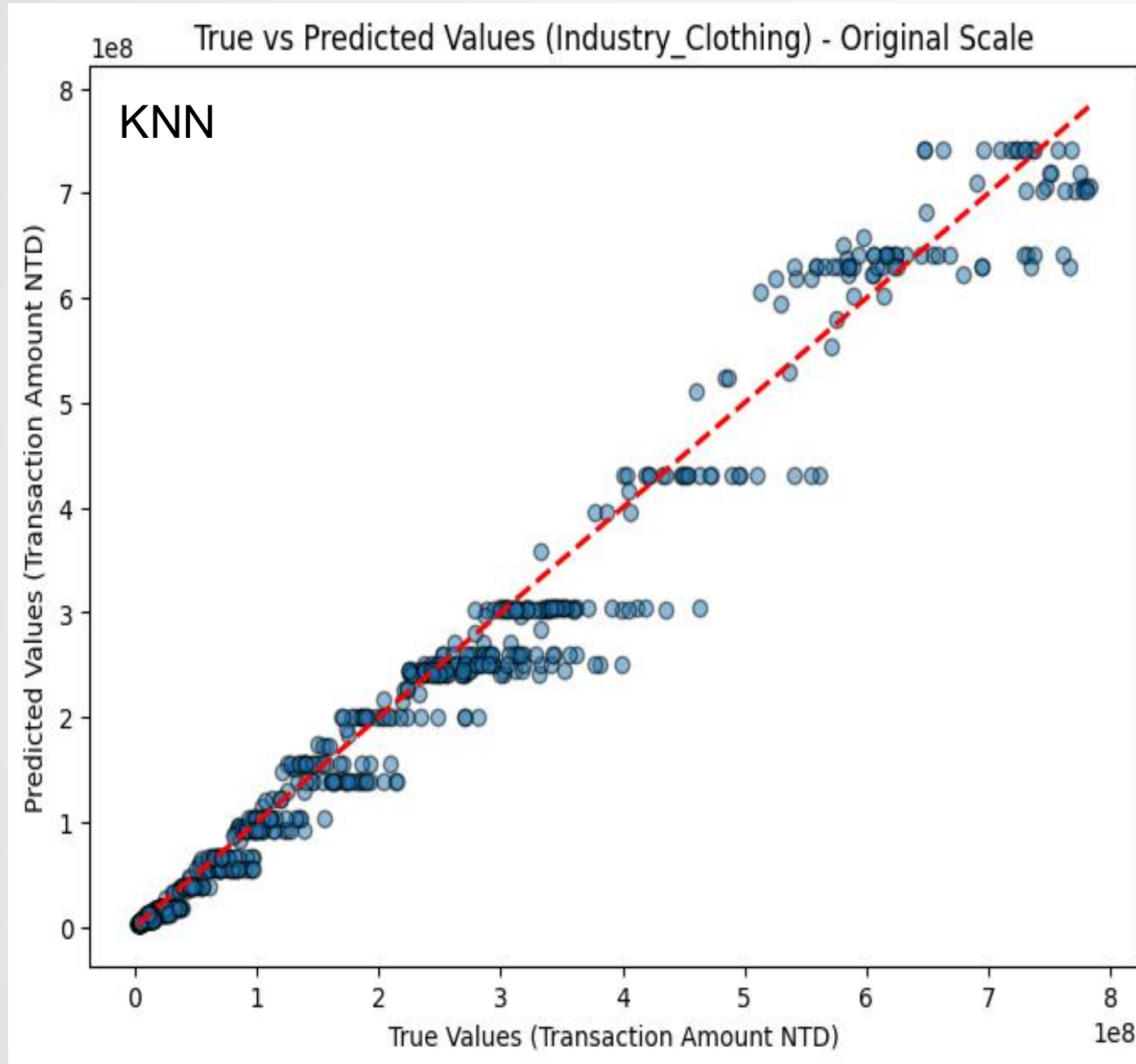
Model Evaluation (linear regression)



MAE Linear regression: 0.727



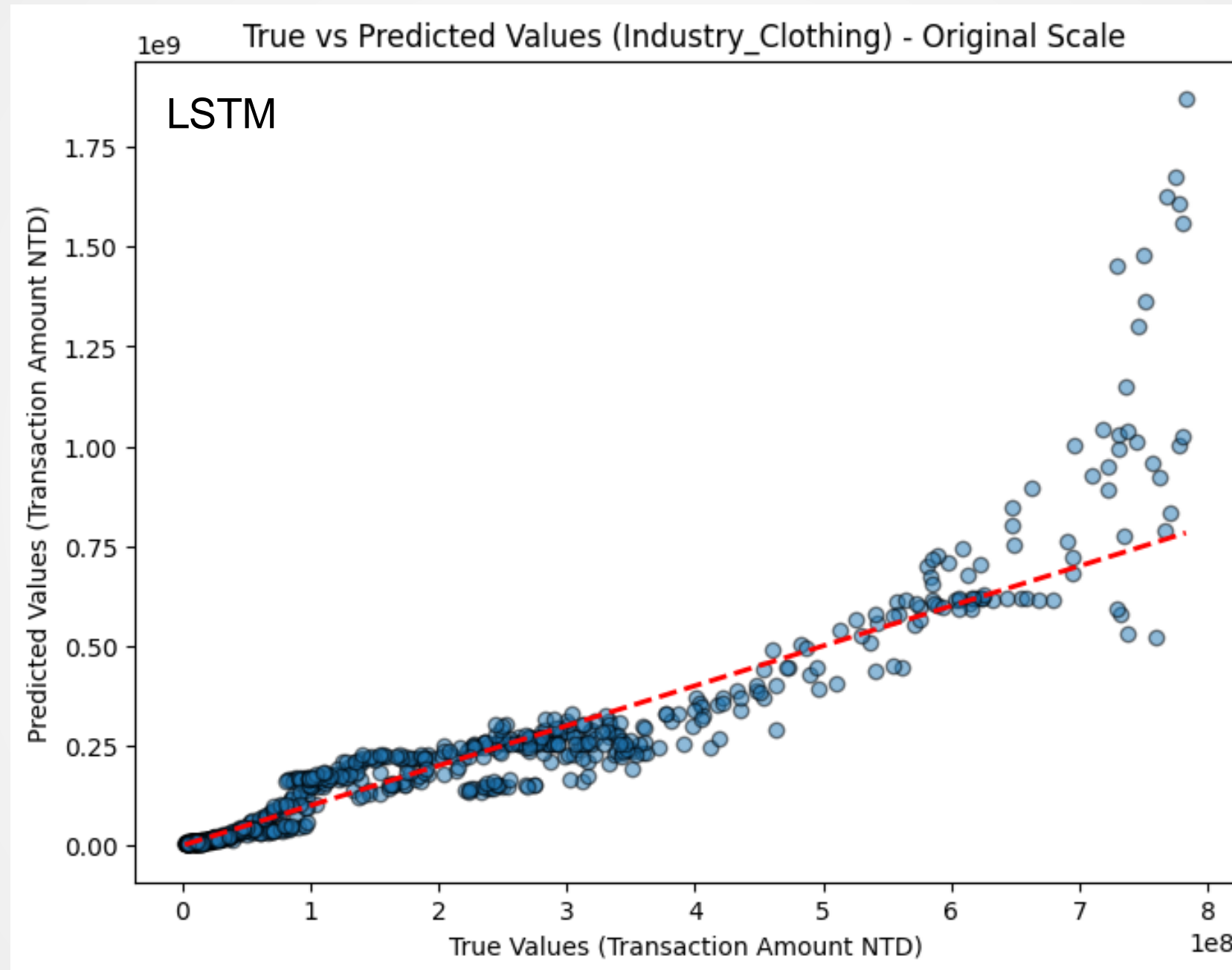
Model Evaluation (KNN)



MAE KNN: 0.167



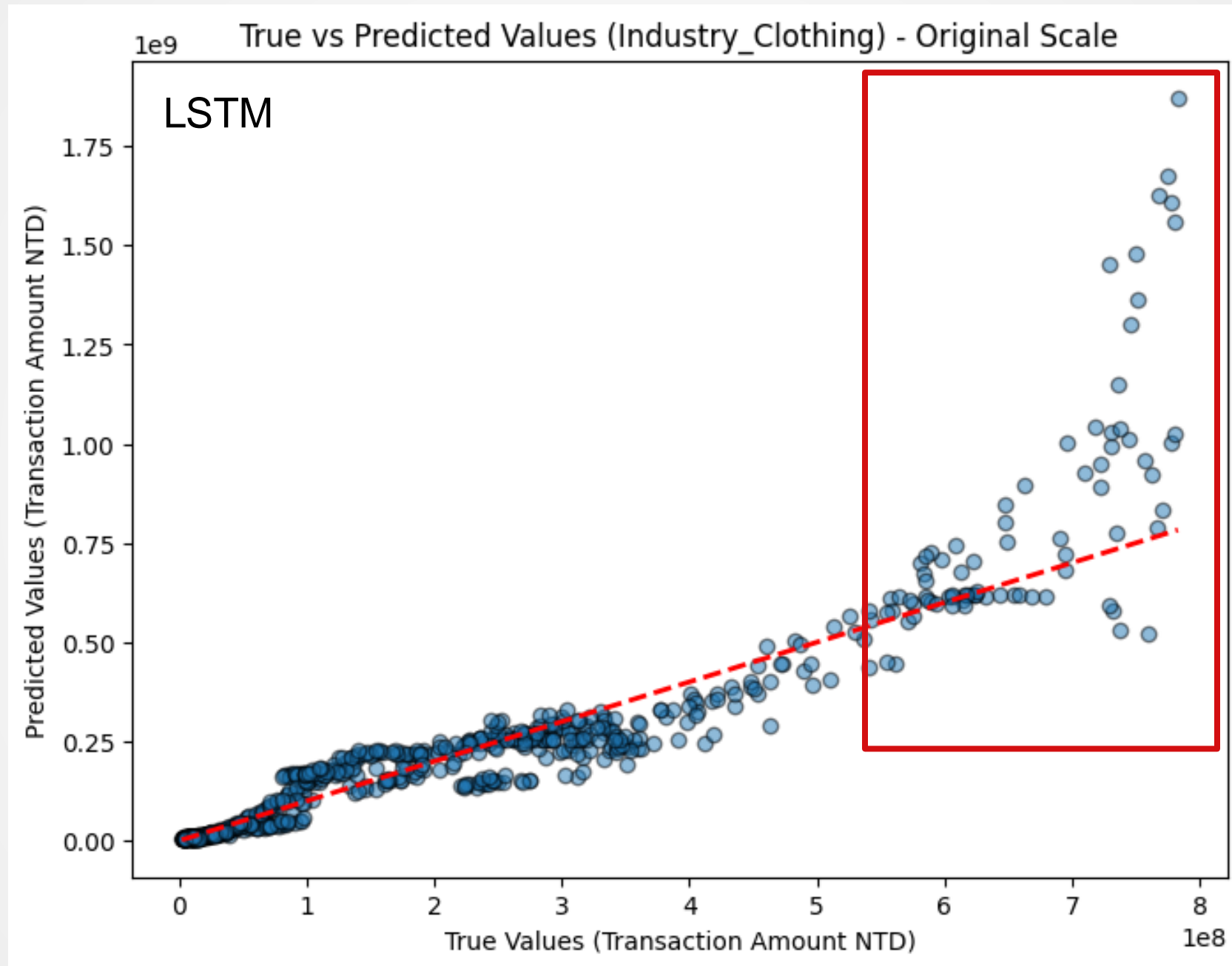
Model Evaluation (LSTM)



MAE LSTM:0.301



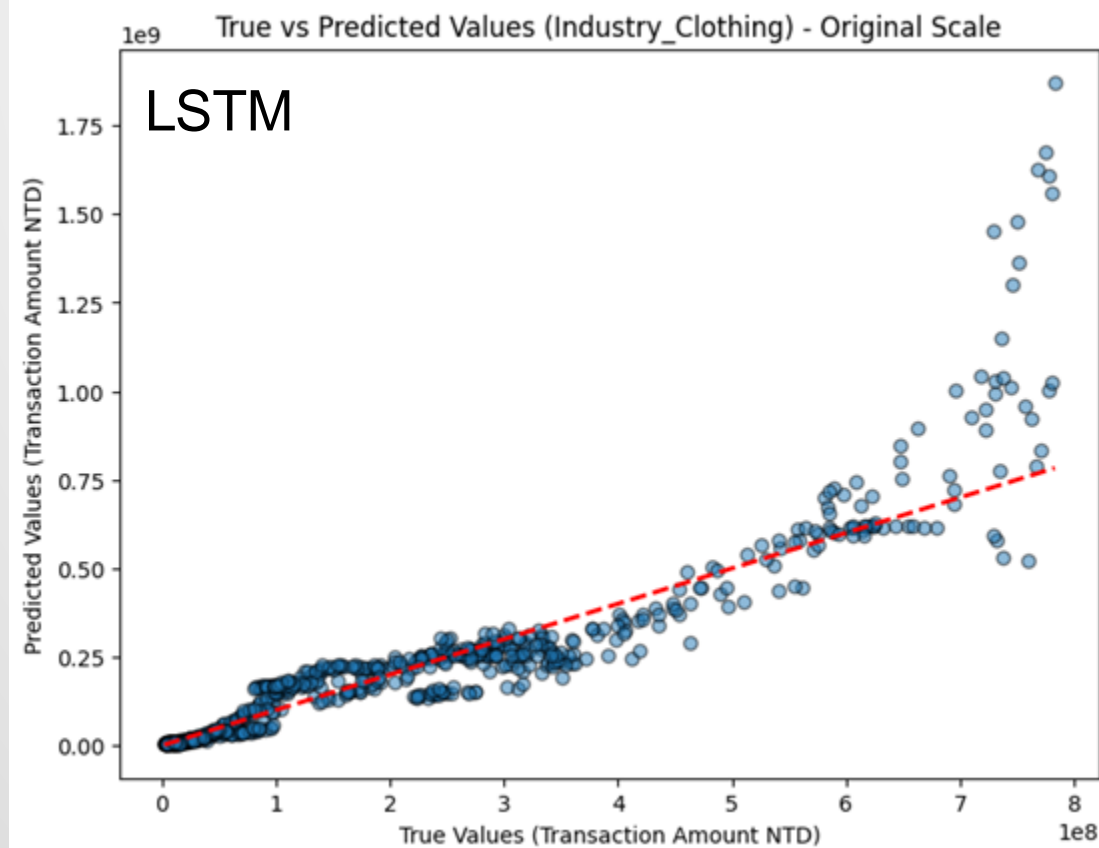
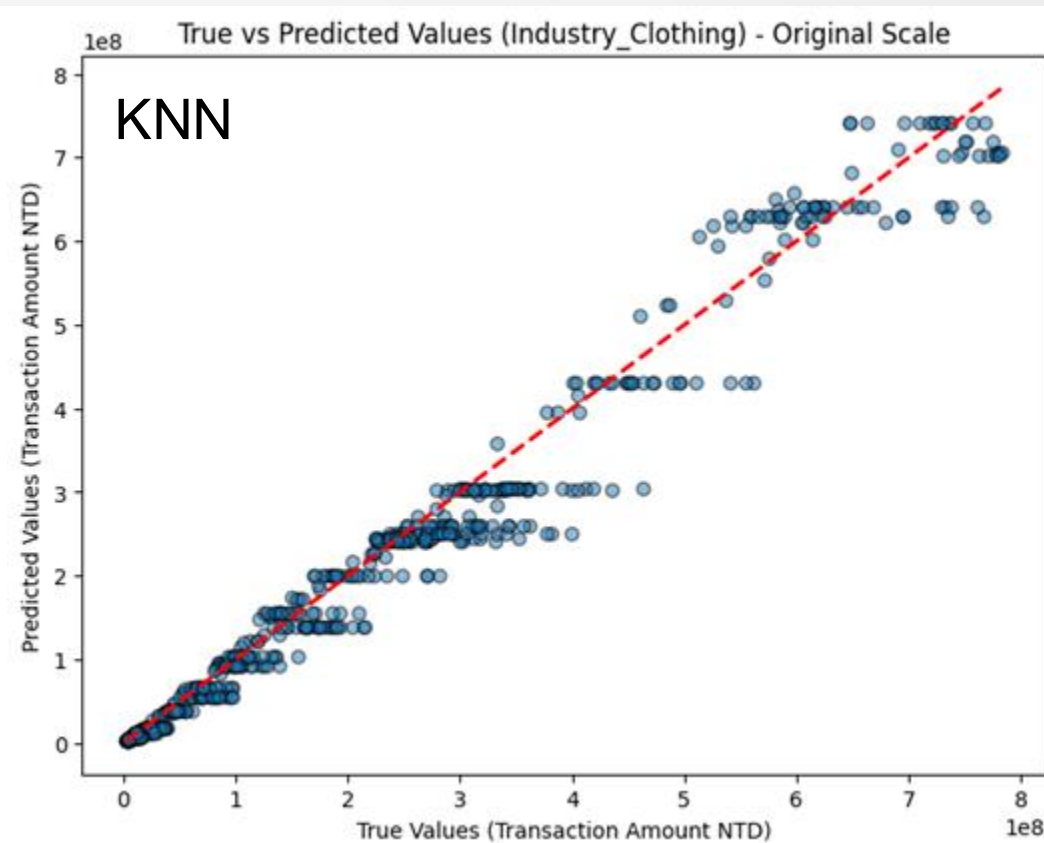
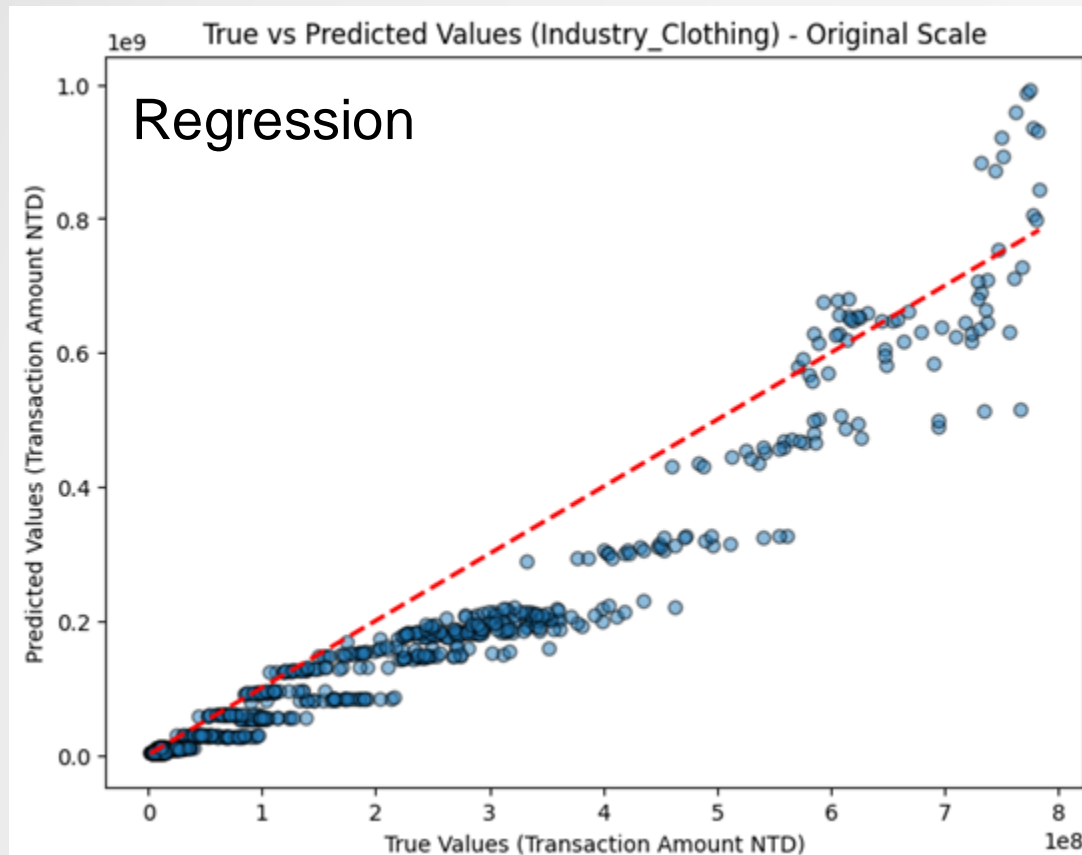
Model Evaluation (LSTM)



MAE LSTM:0.301



Model Evaluation (linear regression, KNN, LSTM)

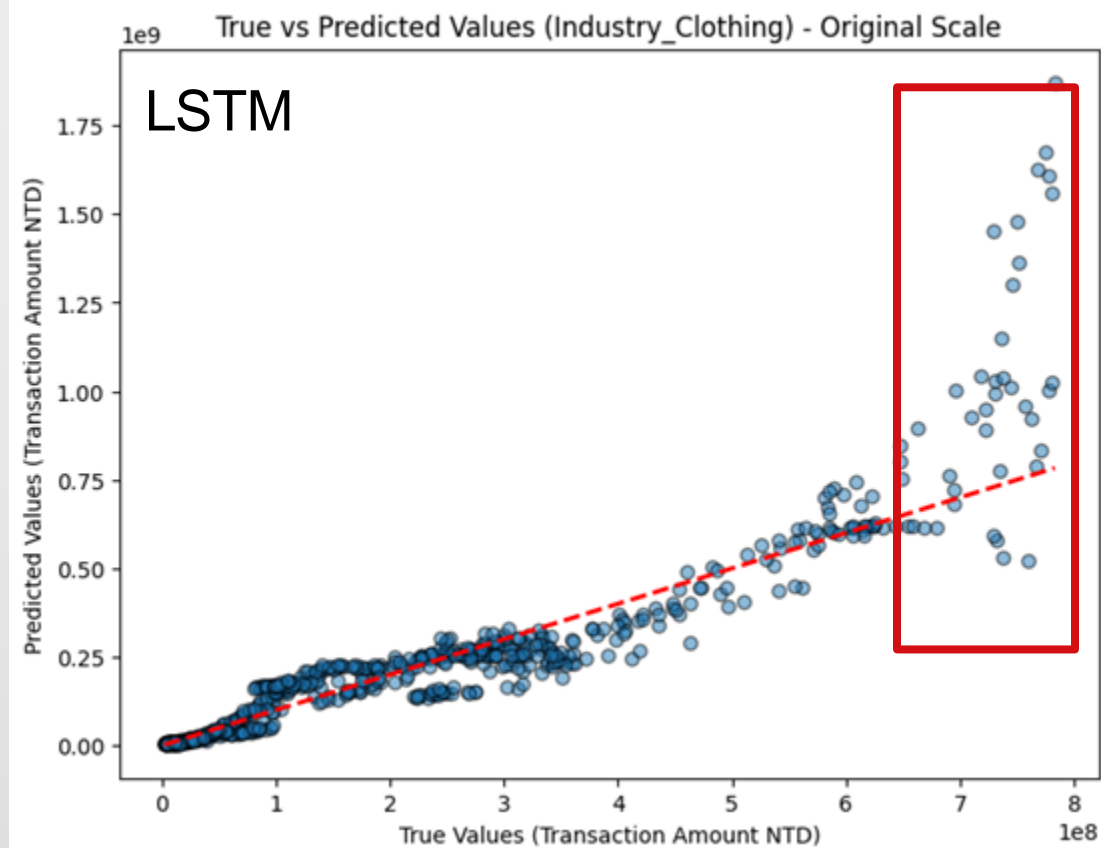
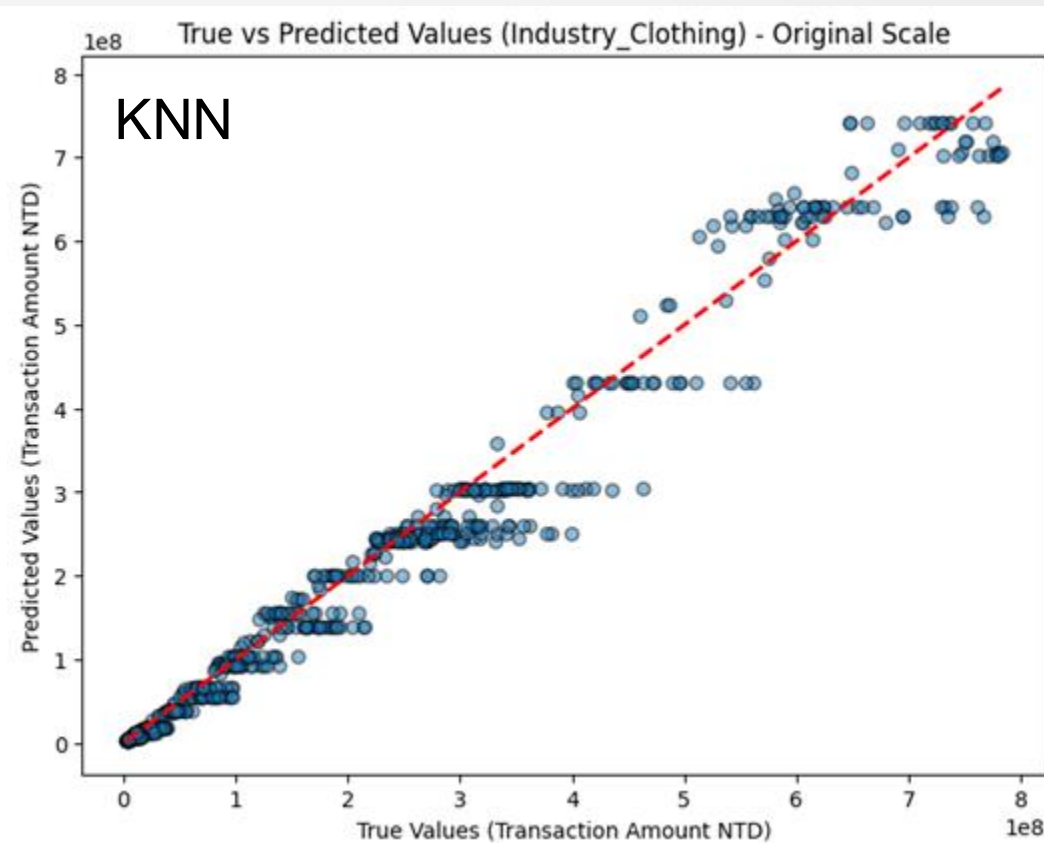
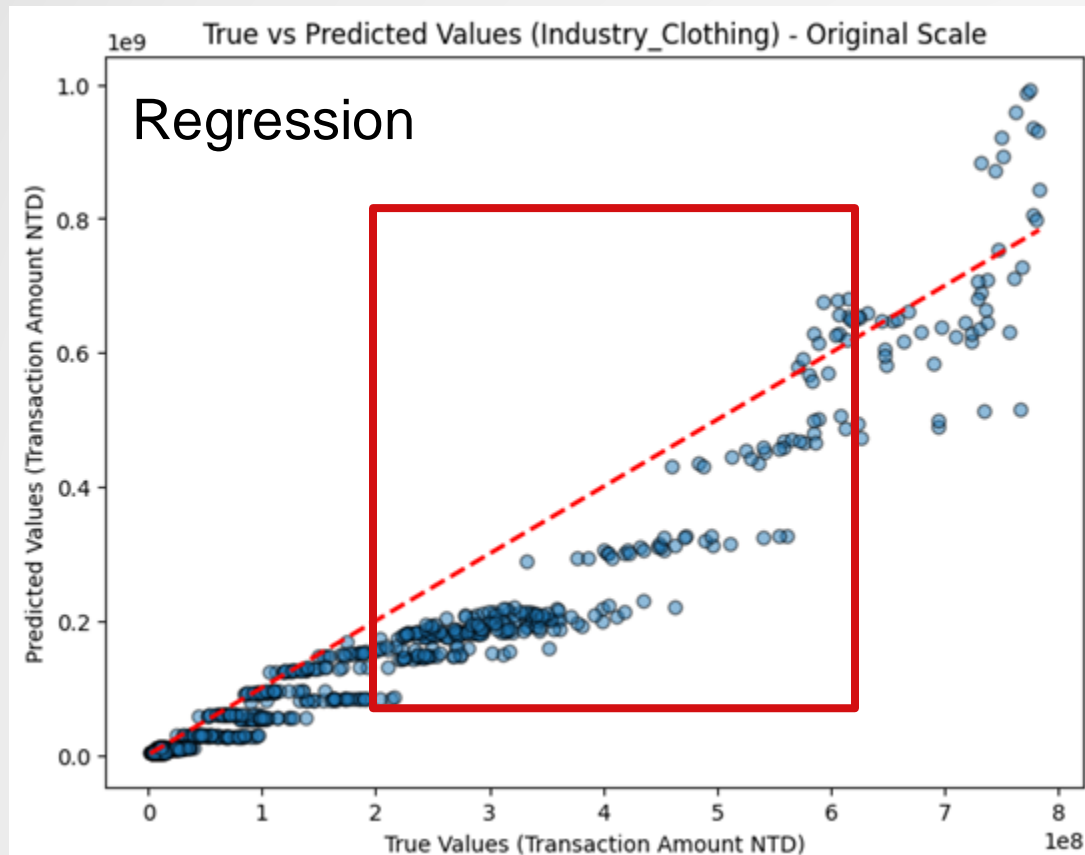


MAE:

1. Linear regression: 0.727
2. KNN: 0.167
3. LSTM: 0.301



Model Evaluation (linear regression, KNN, LSTM)

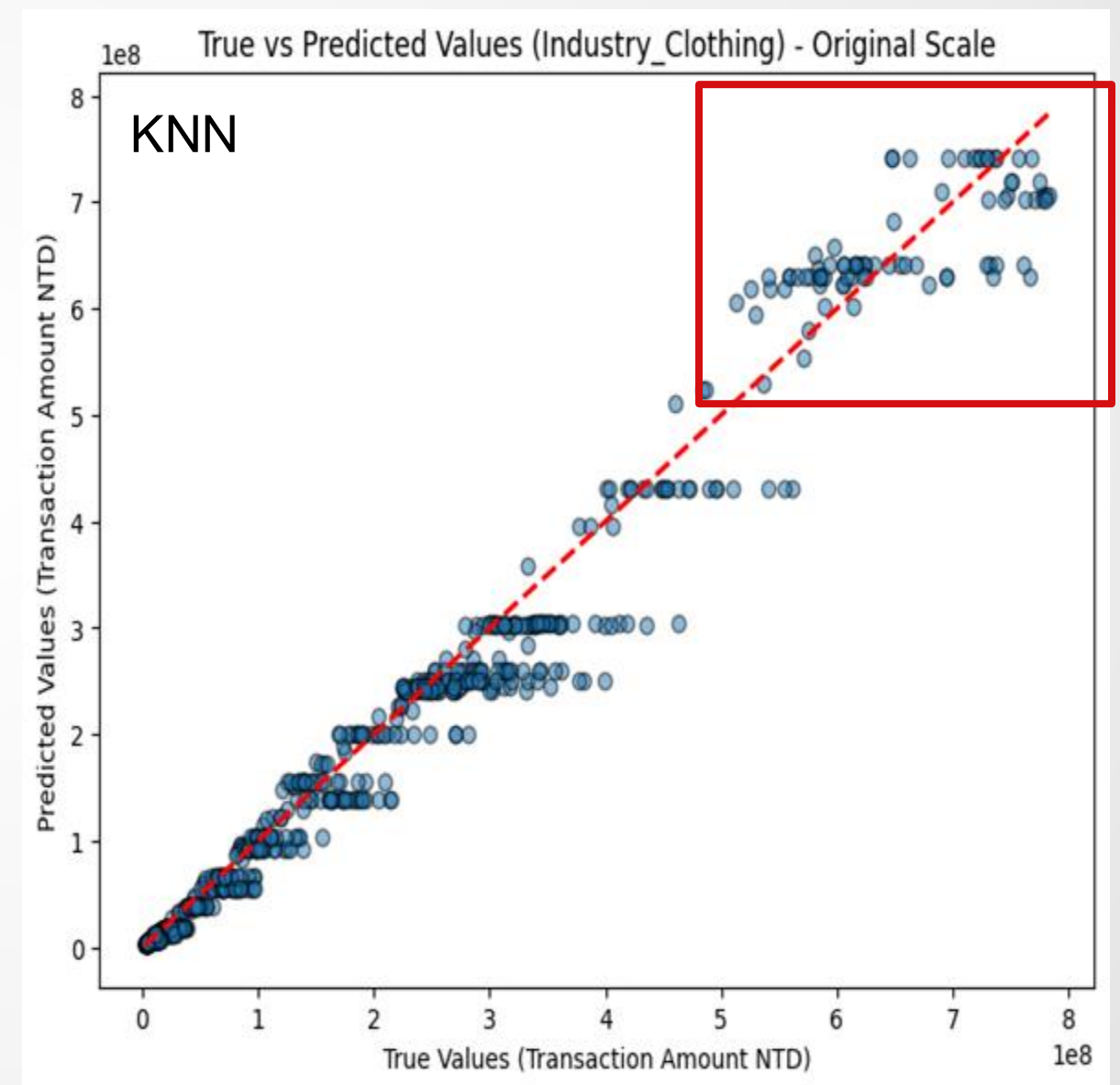
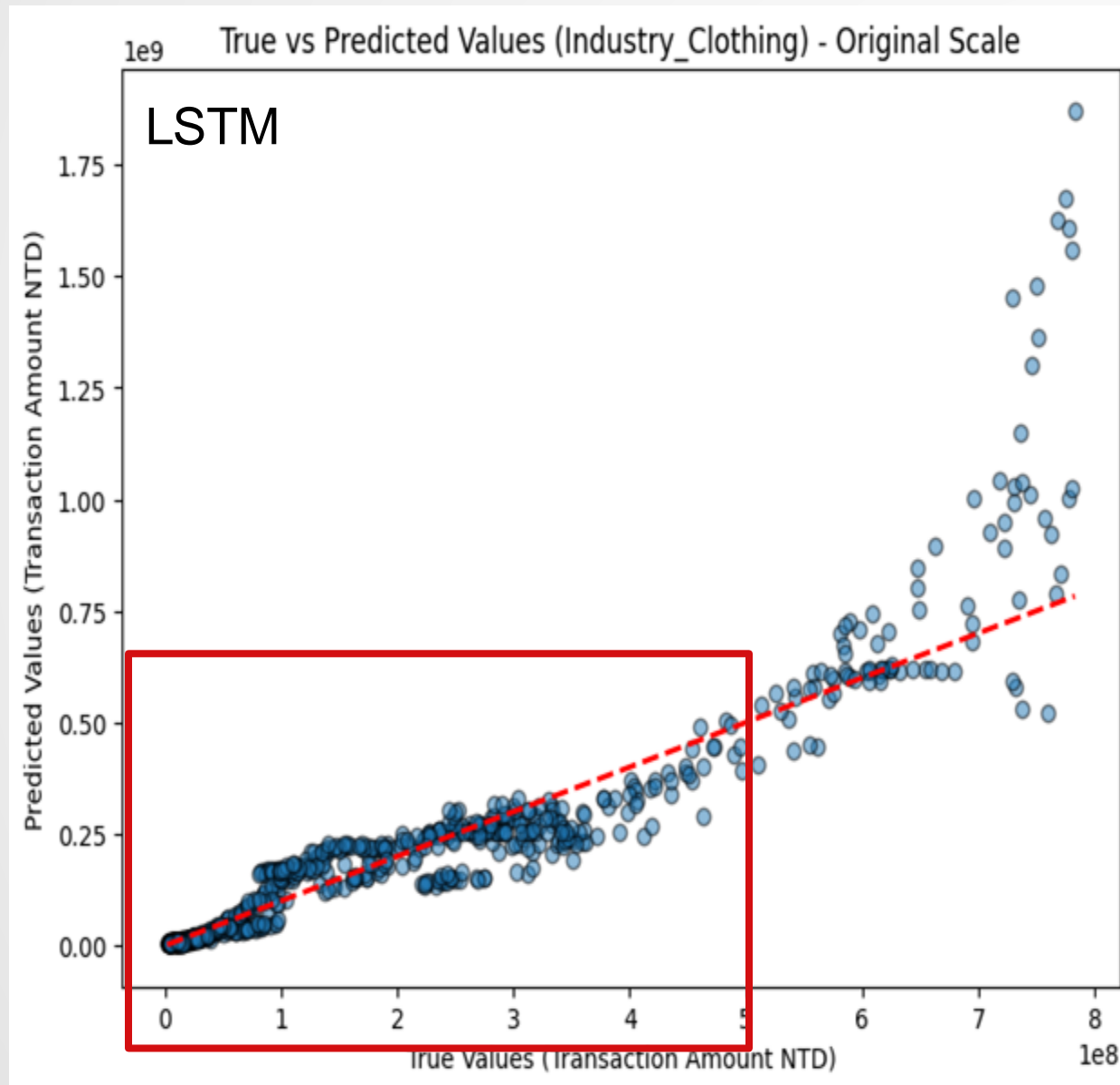


MAE:

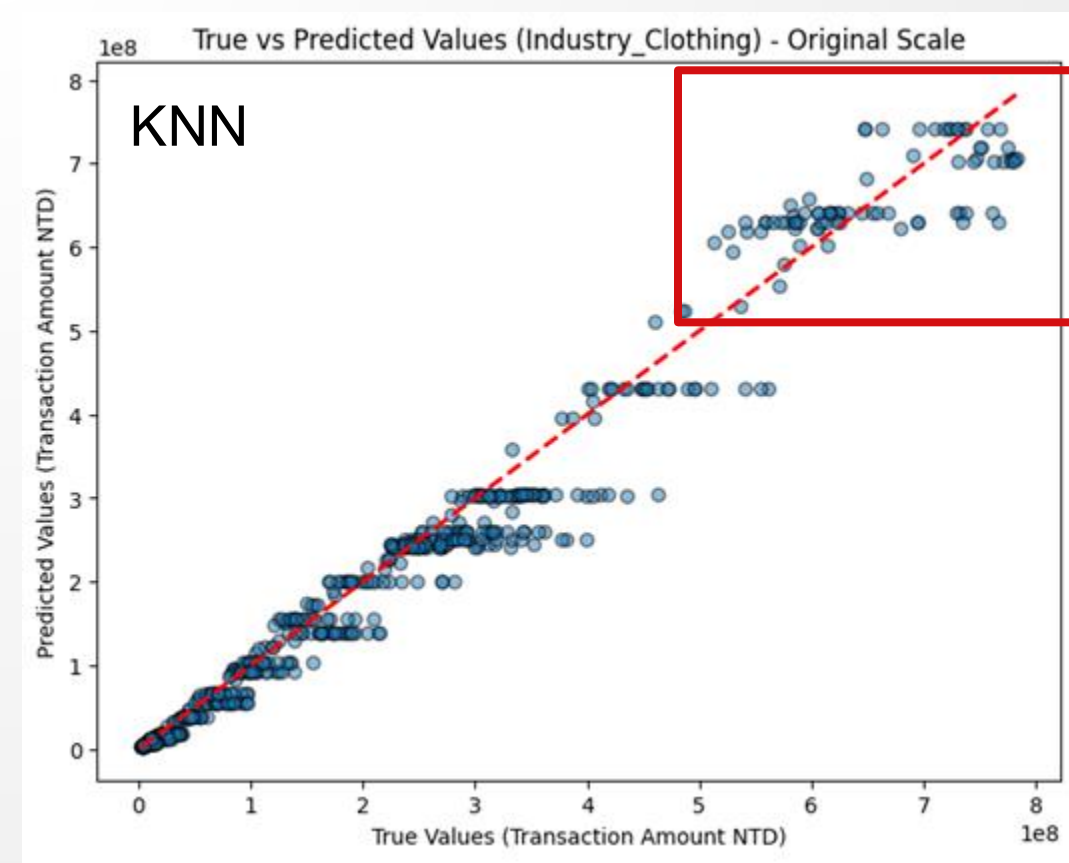
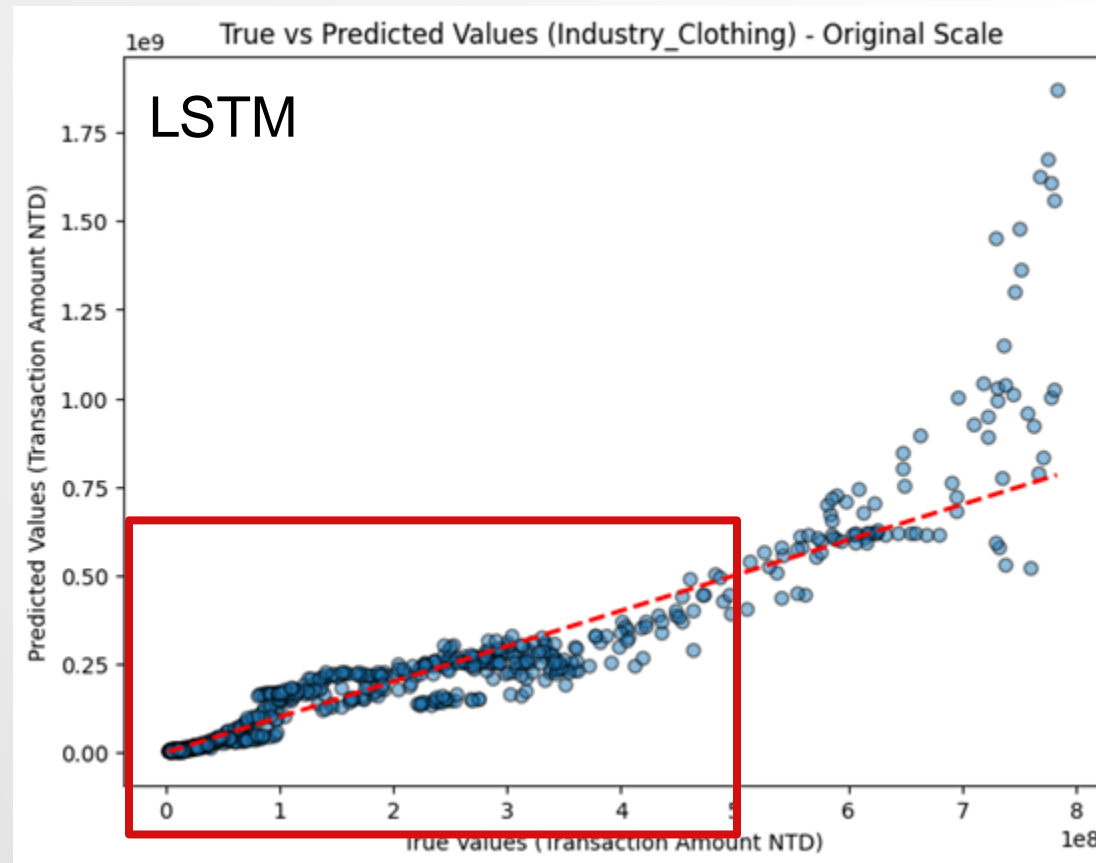
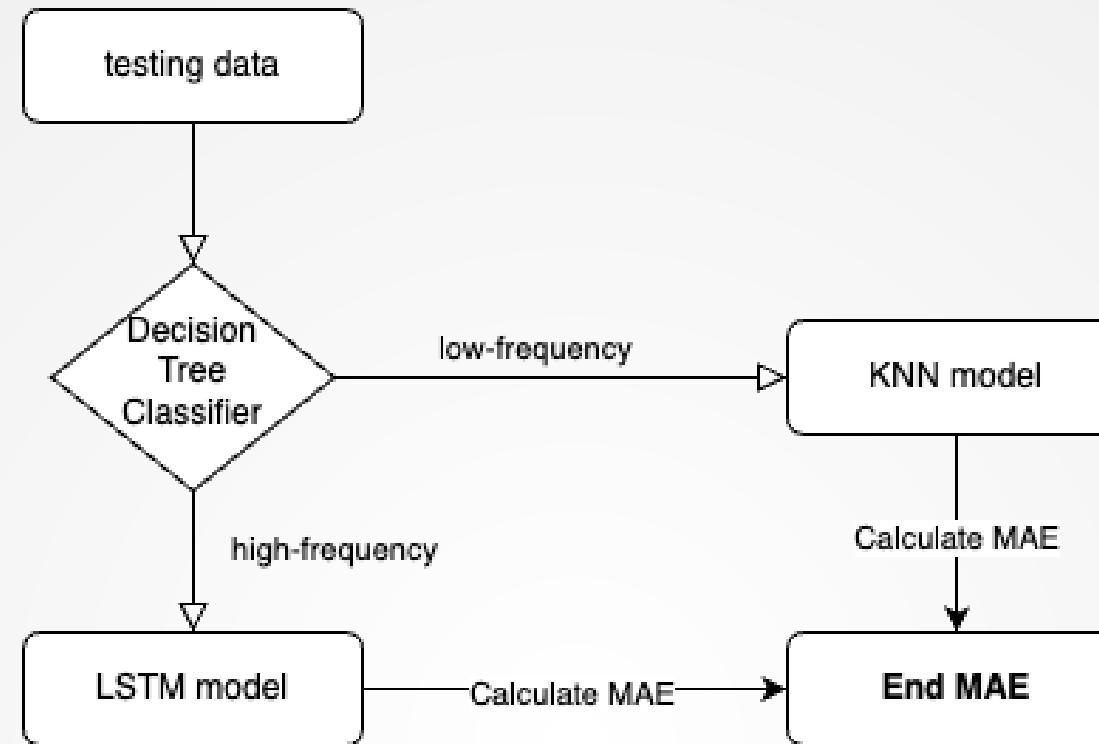
1. Linear regression: 0.727
2. KNN: 0.167
3. LSTM: 0.301



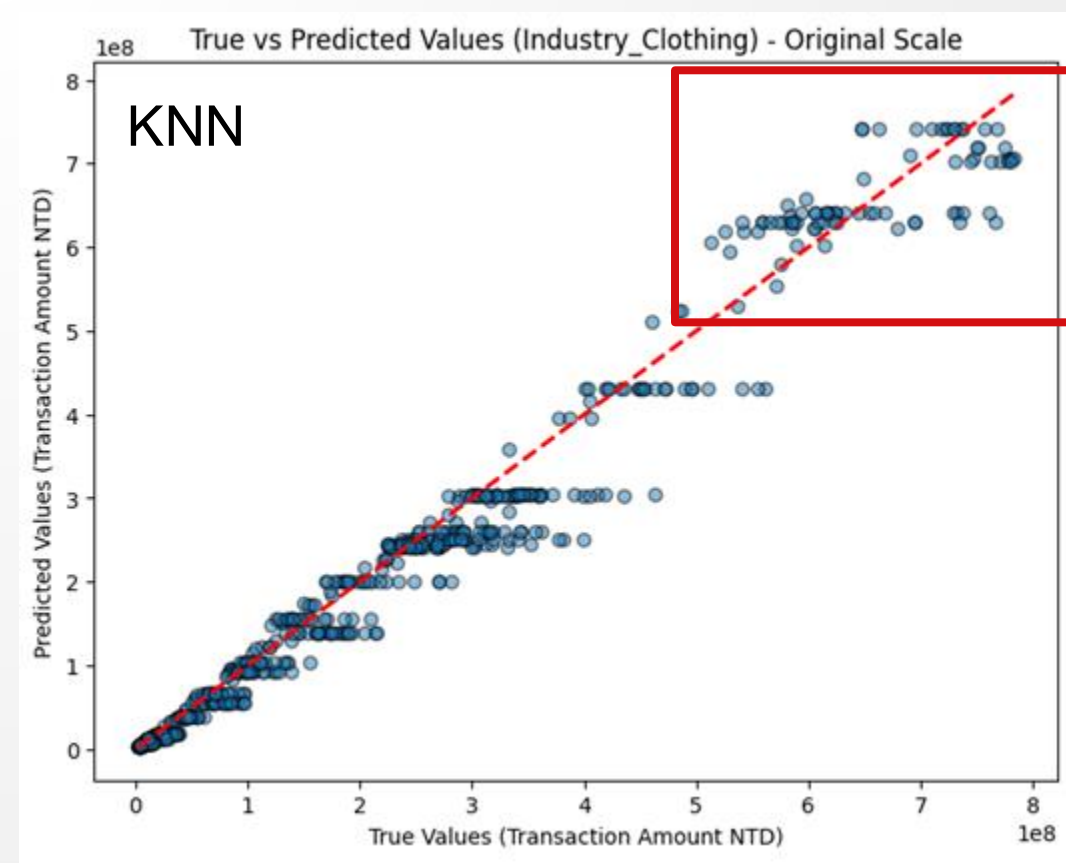
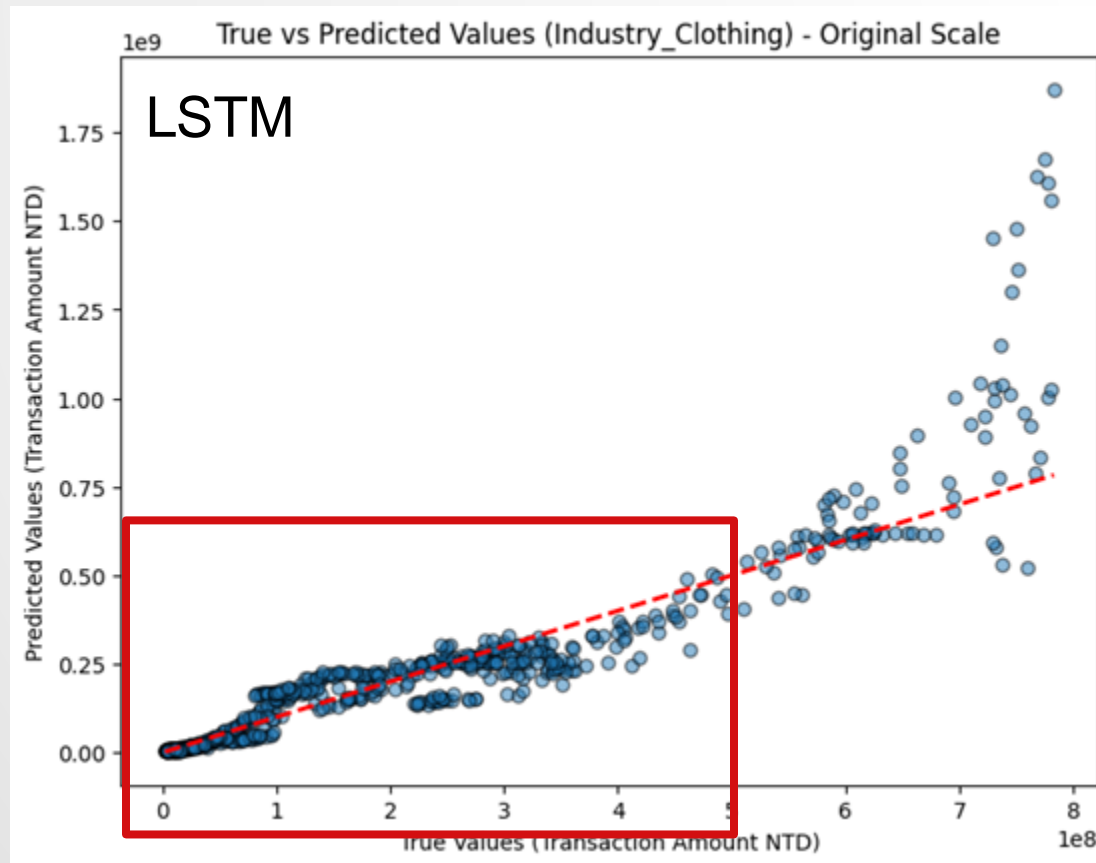
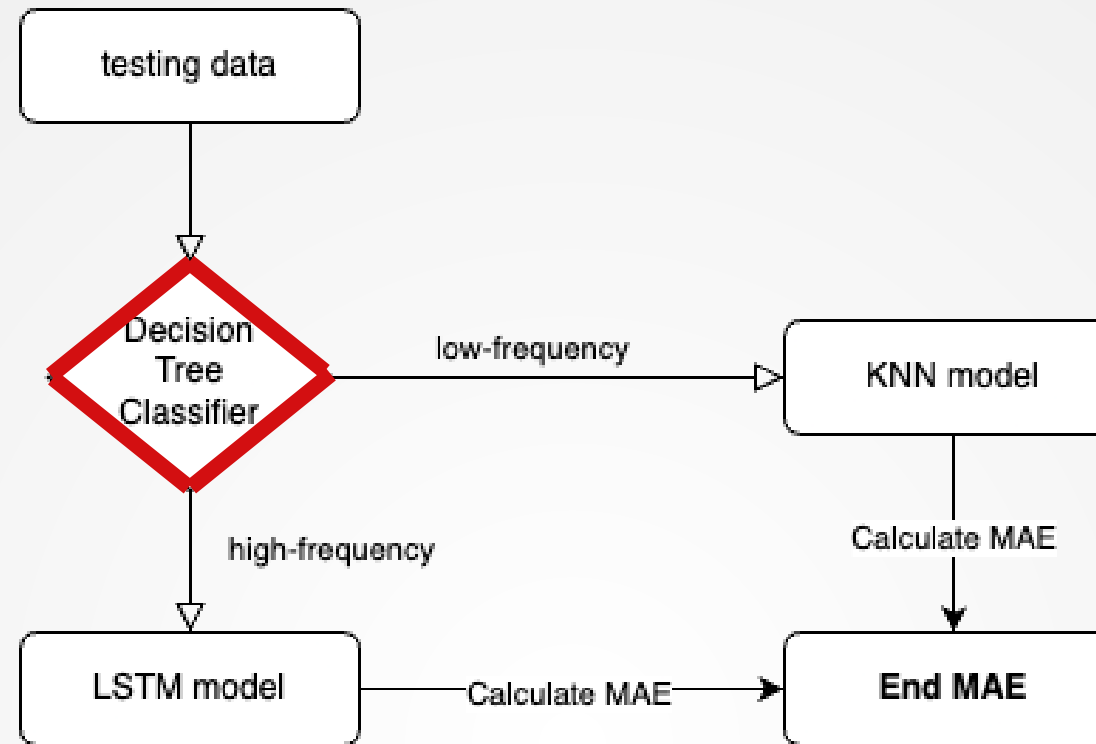
Model Evaluation (KNN, LSTM)



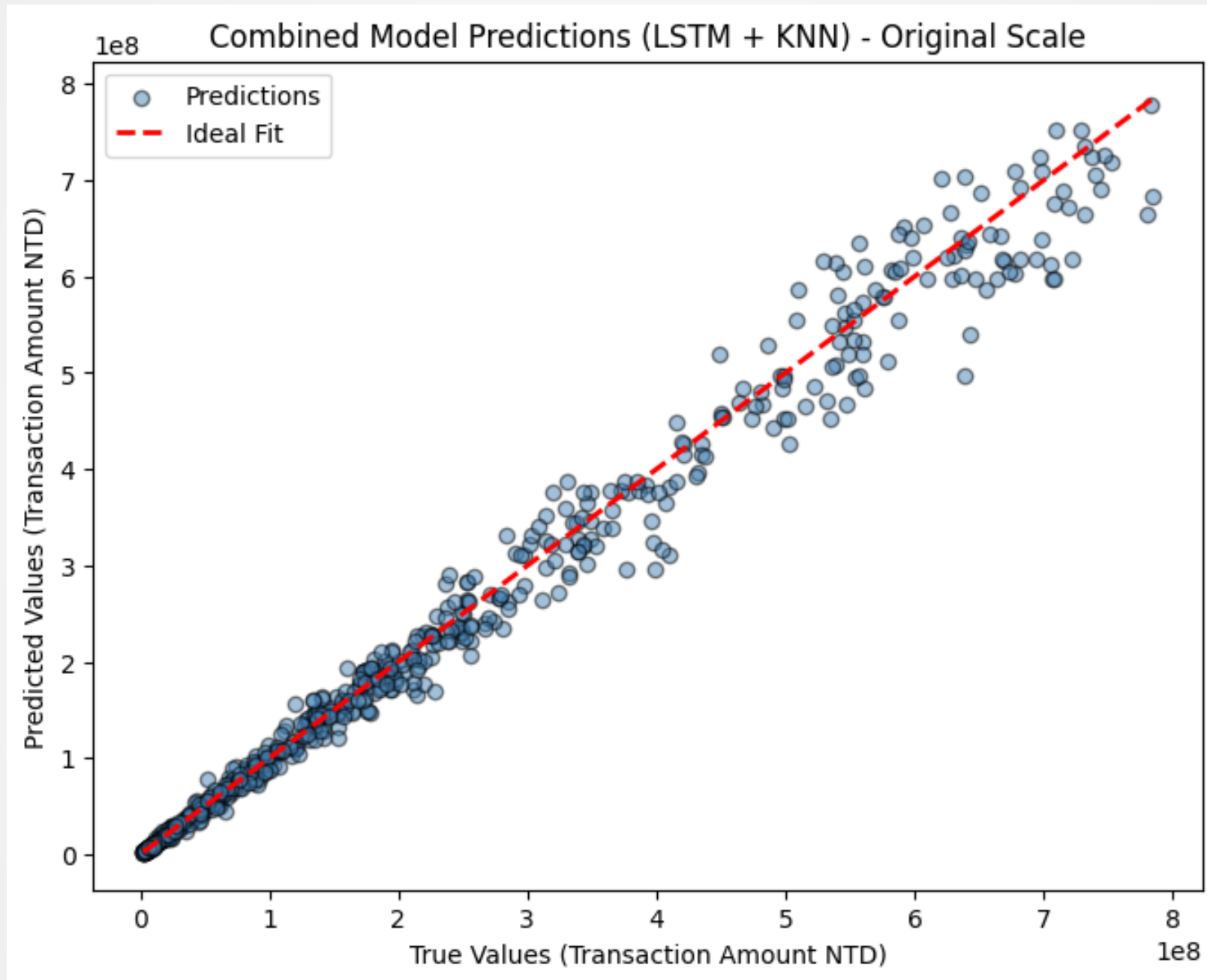
Model Evaluation (KNN, LSTM)



Model Evaluation (KNN, LSTM)



Model Evaluation (KNN+LSTM)



MAE LSTM: 0.086



Model Evaluation (ALL)

Model	MAE
Linear regression	0.727
KNN	0.167
LSTM	0.301
KNN and LSTM	0.086

