EDA-Elo Merchant Category Recommendation

上原　蔵人

2024-10-5

1. 各ファイルに含まれる行数・列数、データ型、欠損値

以下は各ファイルのデータの行数・列数とデータの型

```
train.csv
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 201917 entries, 0 to 201916
Data columns (total 6 columns):
 #    Column                Non-Null Count     Dtype
---   ------                --------------     -----
 0    first_active_month    201917 non-null    object
 1    card_id               201917 non-null    object
 2    feature_1             201917 non-null    int64
 3    feature_2             201917 non-null    int64
 4    feature_3             201917 non-null    int64
 5    target                201917 non-null    float64
dtypes: float64(1), int64(3), object(2)
memory usage: 9.2+ MB
```

```
test.csv
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 123623 entries, 0 to 123622
Data columns (total 5 columns):
 #    Column                Non-Null Count     Dtype
---   ------                --------------     -----
 0    first_active_month    123622 non-null    object
 1    card_id               123623 non-null    object
 2    feature_1             123623 non-null    int64
 3    feature_2             123623 non-null    int64
 4    feature_3             123623 non-null    int64
dtypes: int64(3), object(2)
memory usage: 4.7+ MB
```

```
historical_transactions.csv
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 29112361 entries, 0 to 29112360
Data columns (total 14 columns):
 #    Column                Dtype
---   ------                -----
 0    authorized_flag       object
 1    card_id               object
 2    city_id               int64
 3    category_1            object
 4    installments          int64
 5    category_3            object
 6    merchant_category_id  int64
 7    merchant_id           object
 8    month_lag             int64
 9    purchase_amount       float64
 10   purchase_date         object
 11   category_2            float64
 12   state_id              int64
 13   subsector_id          int64
dtypes: float64(2), int64(6), object(6)
memory usage: 3.0+ GB
```

```
new_merchant_transactions.csv
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1963031 entries, 0 to 1963030
Data columns (total 14 columns):
 #    Column                Dtype
---   ------                -----
 0    authorized_flag       object
 1    card_id               object
 2    city_id               int64
 3    category_1            object
 4    installments          int64
```

```
 5    category_3                  object
 6    merchant_category_id    int64
 7    merchant_id                 object
 8    month_lag                   int64
 9    purchase_amount             float64
 10   purchase_date               object
 11   category_2                  float64
 12   state_id                    int64
 13   subsector_id                int64
dtypes: float64(2), int64(6), object(6)
memory usage: 209.7+ MB
```

```
merchants.csv
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 334696 entries, 0 to 334695
Data columns (total 22 columns):
 #    Column                          Non-Null Count     Dtype
---   ------                          --------------     -----
 0    merchant_id                     334696 non-null    object
 1    merchant_group_id               334696 non-null    int64
 2    merchant_category_id            334696 non-null    int64
 3    subsector_id                    334696 non-null    int64
 4    numerical_1                     334696 non-null    float64
 5    numerical_2                     334696 non-null    float64
 6    category_1                      334696 non-null    object
 7    most_recent_sales_range         334696 non-null    object
 8    most_recent_purchases_range     334696 non-null    object
 9    avg_sales_lag3                  334683 non-null    float64
 10   avg_purchases_lag3              334696 non-null    float64
 11   active_months_lag3              334696 non-null    int64
 12   avg_sales_lag6                  334683 non-null    float64
 13   avg_purchases_lag6              334696 non-null    float64
 14   active_months_lag6              334696 non-null    int64
 15   avg_sales_lag12                 334683 non-null    float64
 16   avg_purchases_lag12             334696 non-null    float64
 17   active_months_lag12             334696 non-null    int64
```

```
 18   category_4                  334696 non-null   object
 19   city_id                     334696 non-null   int64
 20   state_id                    334696 non-null   int64
 21   category_2                  322809 non-null   float64
dtypes: float64(9), int64(8), object(5)
memory usage: 56.2+ MB
```

以下は各ファイルのカラムの欠損値

```
Train Null Values:
first_active_month    0
card_id               0
feature_1             0
feature_2             0
feature_3             0
target                0
dtype: int64
```

```
Test Null Values:
first_active_month    1
card_id               0
feature_1             0
feature_2             0
feature_3             0
dtype: int64
```

```
History Null Values:
authorized_flag              0
card_id                      0
city_id                      0
category_1                   0
installments                 0
category_3              178159
merchant_category_id         0
merchant_id             138481
month_lag                    0
purchase_amount              0
```

```
purchase_date                0
category_2             2652864
state_id                     0
subsector_id                 0
dtype: int64
```

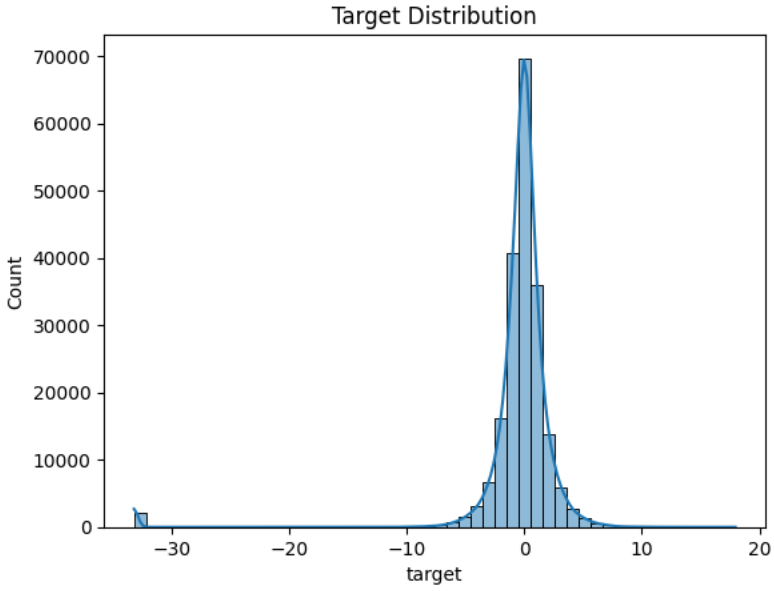
```
New Merchant Null Values:
authorized_flag              0
card_id                      0
city_id                      0
category_1                   0
installments                 0
category_3               55922
merchant_category_id         0
merchant_id              26216
month_lag                    0
purchase_amount              0
purchase_date                0
category_2              111745
state_id                     0
subsector_id                 0
dtype: int64

Merchants Null Values:
merchant_id                        0
merchant_group_id                  0
merchant_category_id               0
subsector_id                       0
numerical_1                        0
numerical_2                        0
category_1                         0
most_recent_sales_range            0
most_recent_purchases_range        0
avg_sales_lag3                    13
avg_purchases_lag3                 0
active_months_lag3                 0
```

```
avg_sales_lag6                 13
avg_purchases_lag6              0
active_months_lag6              0
avg_sales_lag12                13
avg_purchases_lag12             0
active_months_lag12             0
category_4                      0
city_id                         0
state_id                        0
category_2                  11887
dtype: int64
```

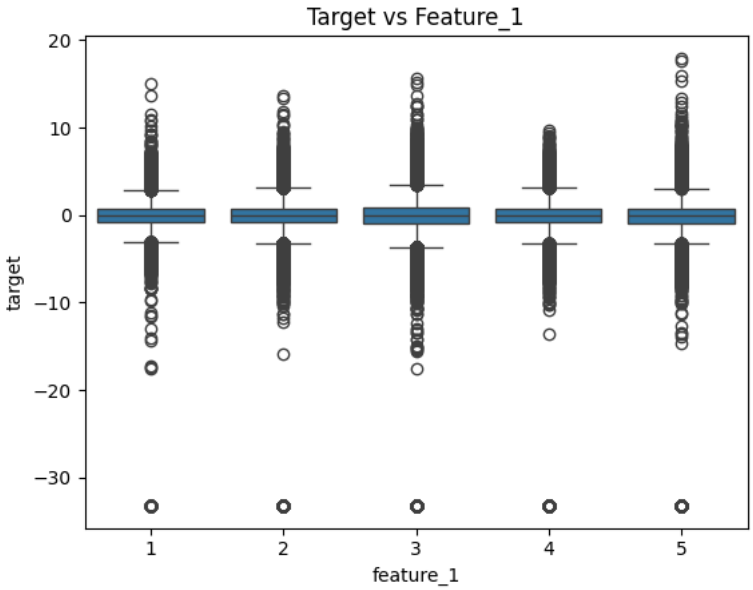
2. ターゲット変数の分布

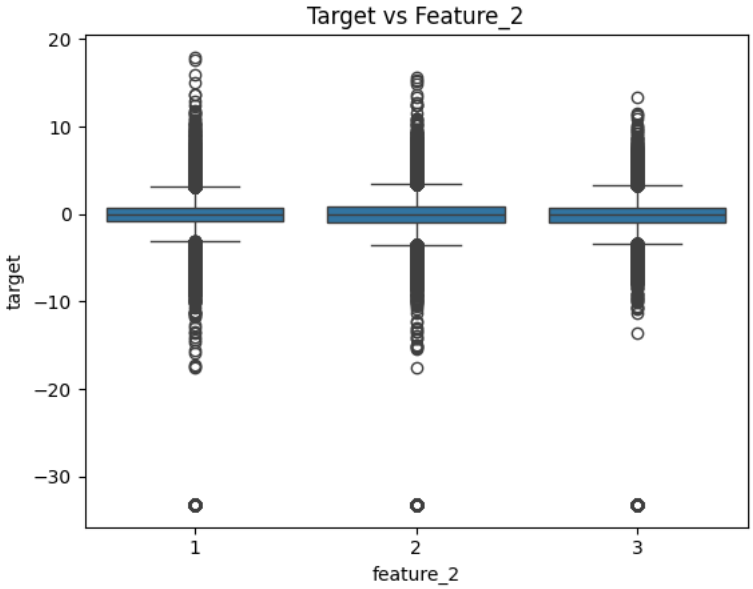
以下は train.csv 内にあるターゲット変数の分布。外れ値が負の方向にあるが、大体のデータは 0 を中心とした尖った正規分布の形になっている。

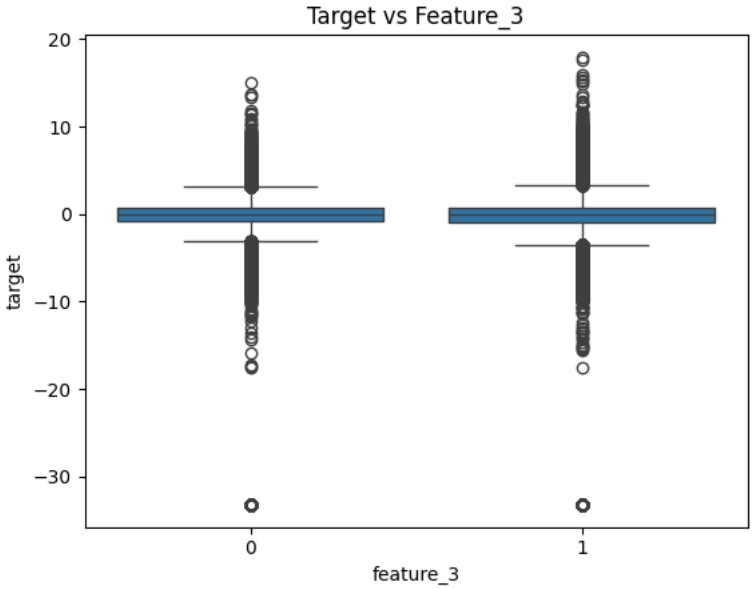


3. カテゴリ変数とターゲットの関係

以下の 3 つ図は train.csv の中にある特微量とターゲットの関係を箱ヒゲ図にしたもの。2 章で見たように負の方向に外れ値が存在するが、0 を中心にまとまっている。その傾向はどの特微量も同じようだ。







4. 新しい特徴量の生成と結合

以下のコードはカード ID を軸に取引データを集約している。取引回数、取引額の合計、平均、最小値・最大値などを特徴量として追加しています。

```
# カード ID ごとに取引データを集約
history_agg = history.groupby('card_id').agg({
    'purchase_amount': ['sum', 'mean', 'min', 'max', 'count'],
    'month_lag': 'mean'
}).reset_index()

new_merchant_agg = new_merchant.groupby('card_id').agg({
    'purchase_amount': ['sum', 'mean', 'min', 'max', 'count'],
    'month_lag': 'mean'
}).reset_index()
```

以下は上記で作った新しい特徴量を’card_id’をキーにして train に結合したもの。

```
# データの結合
train_merged = pd.merge(train, history_agg, on='card_id', how='left')
train_merged = pd.merge(train_merged, new_merchant_agg, on='card_id', how='left',
suffixes=('_history', '_new'))
```

train_merged データで'first_active_month'のみ時系列データだったため、扱いやすいように時間の情報を個別に取得して int 型の新しい特徴量を作った。

```
# 例: 'first_active_month'を日付型に変換し、各成分を抽出
train_merged['first_active_month'] =
pd.to_datetime(train_merged['first_active_month'])
train_merged['year'] = train_merged['first_active_month'].dt.year
train_merged['month'] = train_merged['first_active_month'].dt.month
train_merged['day'] = train_merged['first_active_month'].dt.day
train_merged['dayofweek'] = train_merged['first_active_month'].dt.dayofweek
```

以下は train_merged の特微量の情報。

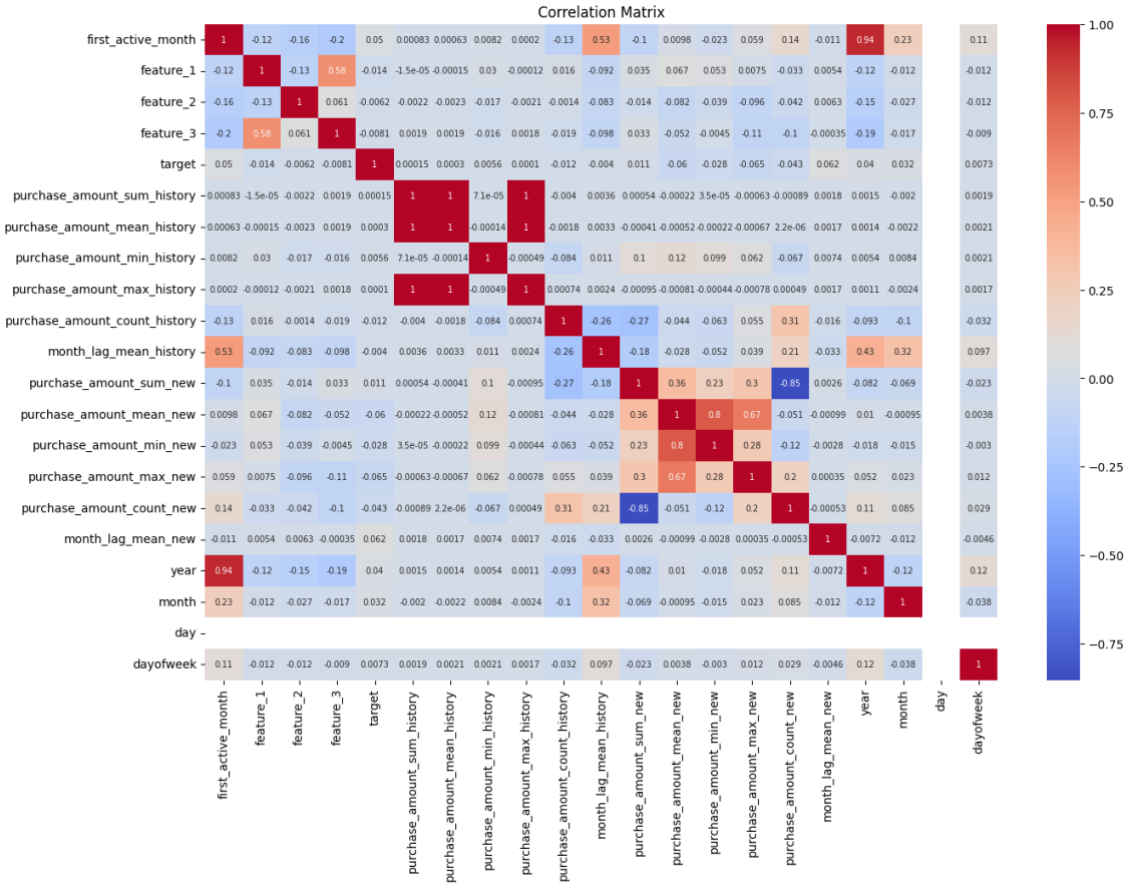
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 201917 entries, 0 to 201916
Data columns (total 22 columns):
 #    Column                          Non-Null Count     Dtype
---   ------                         --------------     -----
 0    first_active_month              201917 non-null    datetime64[ns]
 1    card_id                          201917 non-null    object
 2    feature_1                        201917 non-null    int64
 3    feature_2                        201917 non-null    int64
 4    feature_3                        201917 non-null    int64
 5    target                          201917 non-null    float64
 6    purchase_amount_sum_history      201917 non-null    float64
 7    purchase_amount_mean_history     201917 non-null    float64
 8    purchase_amount_min_history      201917 non-null    float64
 9    purchase_amount_max_history      201917 non-null    float64
 10   purchase_amount_count_history    201917 non-null    int64
 11   month_lag_mean_history           201917 non-null    float64
 12   purchase_amount_sum_new          179986 non-null    float64
 13   purchase_amount_mean_new         179986 non-null    float64
 14   purchase_amount_min_new          179986 non-null    float64
 15   purchase_amount_max_new          179986 non-null    float64
 16   purchase_amount_count_new        179986 non-null    float64
 17   month_lag_mean_new               179986 non-null    float64
 18   year                            201917 non-null    int32
 19   month                            201917 non-null    int32
```

```
 20    day                                   201917 non-null    int32
 21    dayofweek                             201917 non-null    int32
dtypes: datetime64[ns](1), float64(12), int32(4), int64(4), object(1)
memory usage: 30.8+ MB
```

5. 相関関係の分析

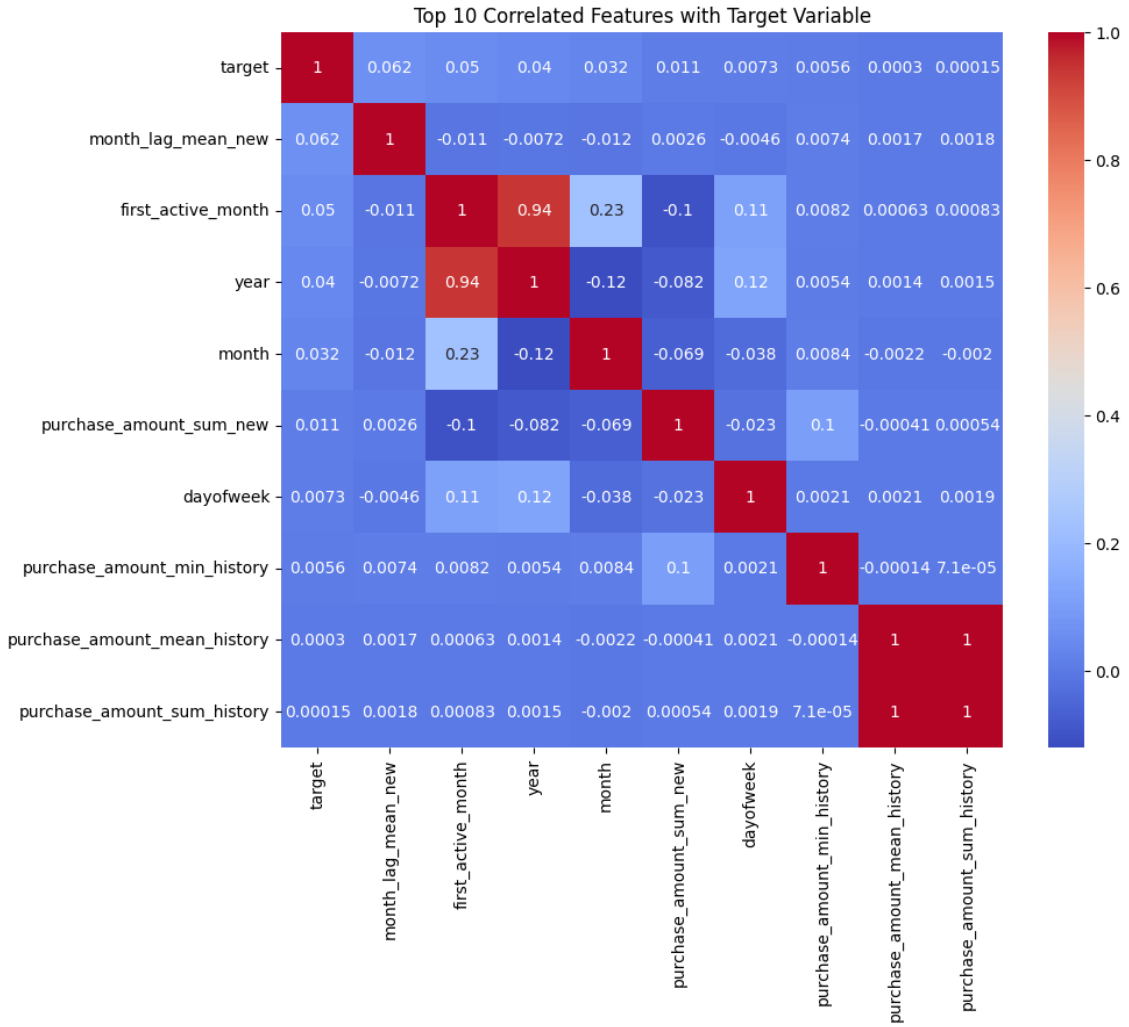
この章では 4 章で作った train_merged の’card_id’を取り除いた new_train_merged の特徴量の相関関係を可視化する。

以下は全ての特徴量のヒートマップ。

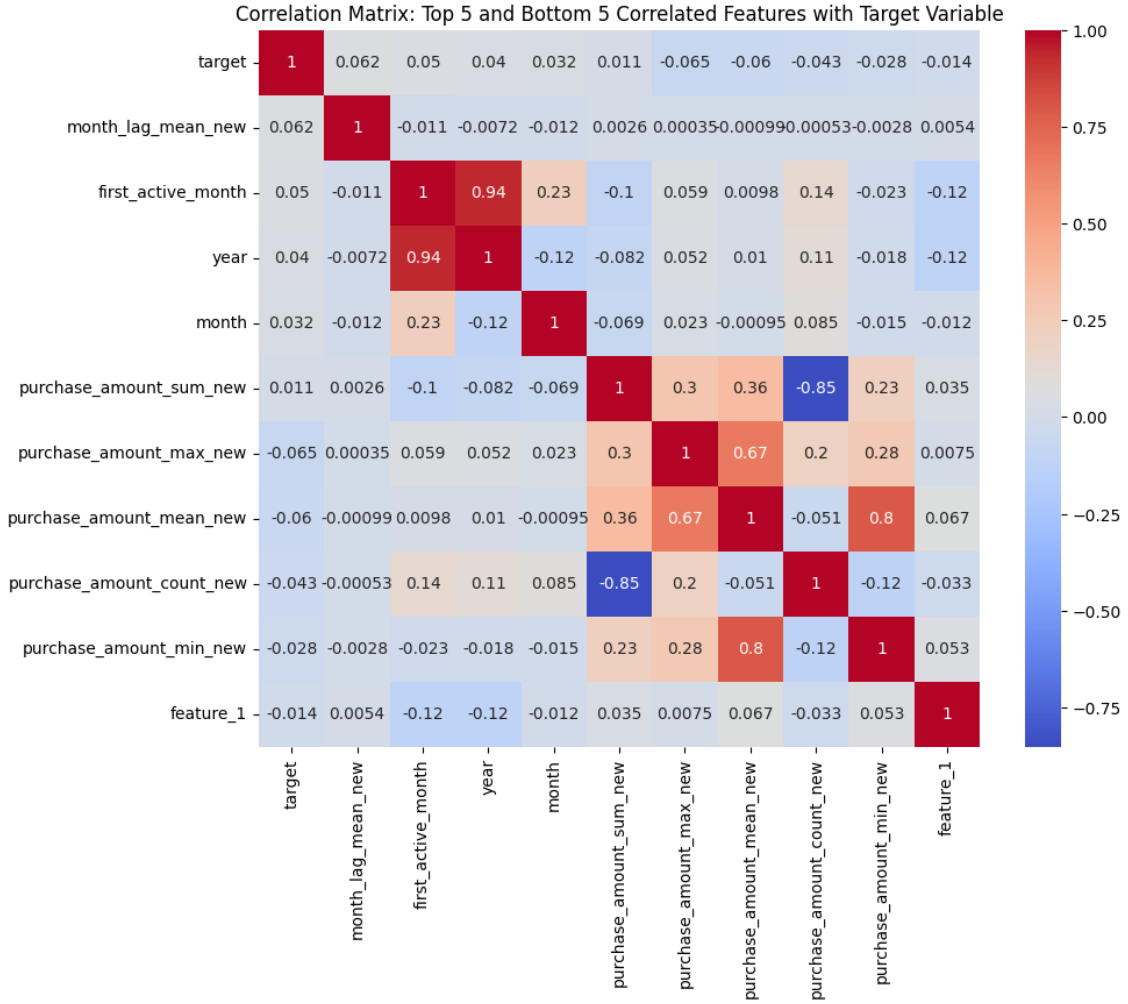


かなり見にくいため下記に図を 2 つ補足する。

以下は相関が高い上位 10 個の特微量のヒートマップ。



以下は相関が高い上位 5 個と低い下位 5 個を合わせたヒートマップ。



target との相関が一番高いのは’month_lag_mean_month’だが、その数値は 0.062 でかなり小さい。