



The Analysis of

MB BANK

Churn Prediction

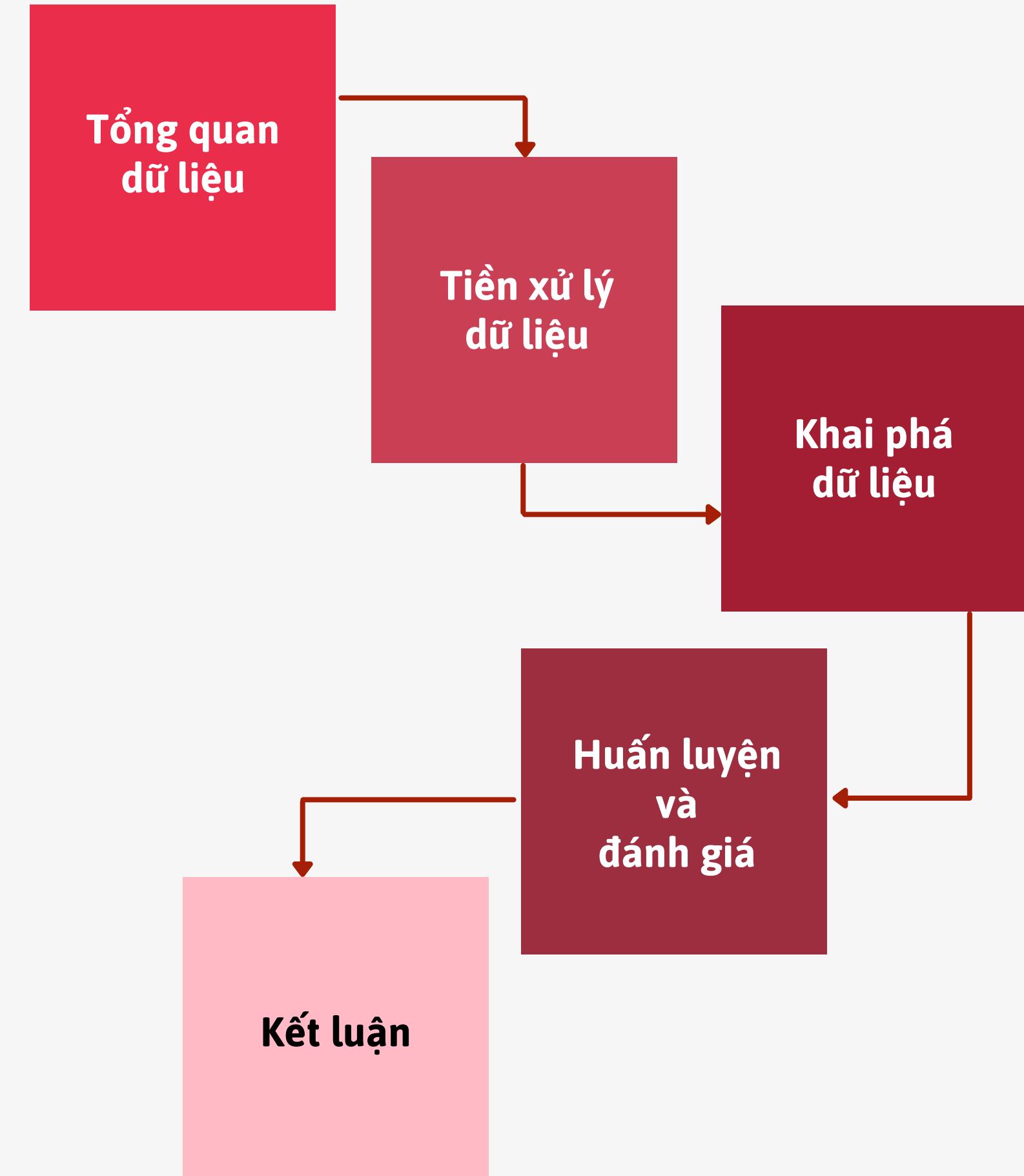
FDC105 - Group 1





Presentation Agenda

MB





Timeline

12/2020

- Latest 3 month
- Latest 1 month
- Latest 1 week →

- amount_week
- max_amount_week
- min_amount_week
- distinct_payment_code_week
- count_payment_code_week
- distinct_trans_group_week
- distinct_ref_no_week



03/2021

- most_act_mar2021_count
- most_act_mar2021
- total_act_mar2021
- total_amt_mar2021

06/2021

- most_act_juin2021_count
- most_act_juin2021
- total_act_juin2021
- total_amt_juin2021

| Data columns (total 50 columns): | | | |
|----------------------------------|------------------------------|----------------|------------------|
| # | Column | Non-Null Count | Dtype |
| 0 | local_ref_1 | 97266 | non-null object |
| 1 | vn_marital_status | 96971 | non-null object |
| 2 | resid_province | 93431 | non-null float64 |
| 3 | resid_district | 93435 | non-null float64 |
| 4 | residwards | 93433 | non-null float64 |
| 5 | birth_incorp_date | 99990 | non-null float64 |
| 6 | amount_week | 54800 | non-null float64 |
| 7 | max_amount_week | 54800 | non-null float64 |
| 8 | min_amount_week | 54800 | non-null float64 |
| 9 | distinct_payment_code_week | 56850 | non-null float64 |
| 10 | count_payment_code_week | 56850 | non-null float64 |
| 11 | distinct_trans_group_week | 56850 | non-null float64 |
| 12 | distinct_ref_no_week | 56850 | non-null float64 |
| 13 | amount_month | 77000 | non-null float64 |
| 14 | max_amount_month | 77000 | non-null float64 |
| 15 | min_amount_month | 77000 | non-null float64 |
| 16 | distinct_payment_code_month | 82223 | non-null float64 |
| 17 | count_payment_code_month | 82223 | non-null float64 |
| 18 | distinct_trans_group_month | 82223 | non-null float64 |
| 19 | distinct_ref_no_month | 82223 | non-null float64 |
| 20 | amount_3month | 86520 | non-null float64 |
| 21 | max_amount_3month | 86520 | non-null float64 |
| 22 | min_amount_3month | 86520 | non-null float64 |
| 23 | distinct_payment_code_3month | 100000 | non-null int64 |
| 24 | count_payment_code_3month | 100000 | non-null int64 |
| 25 | distinct_trans_group_3month | 100000 | non-null int64 |
| 26 | distinct_ref_no_3month | 100000 | non-null int64 |
| 27 | most_act_mar2021_count | 73477 | non-null float64 |
| 28 | most_act_mar2021 | 73476 | non-null object |
| 29 | total_act_mar2021 | 73477 | non-null float64 |
| 30 | total_amt_mar2021 | 72855 | non-null float64 |
| 31 | most_act_juin2021_count | 71377 | non-null float64 |
| 32 | most_act_juin2021 | 71375 | non-null object |
| 33 | total_act_juin2021 | 71377 | non-null float64 |
| 34 | total_amt_juin2021 | 70664 | non-null float64 |
| 35 | rd_id | 100000 | non-null int64 |
| 36 | savingValueMar2021_heoSo | 5039 | non-null float64 |
| 37 | savingValueJuin2021_heoSo | 7146 | non-null float64 |
| 38 | totalLoginMar2021_heoSo | 7315 | non-null float64 |
| 39 | totalLoginJuin2021_heoSo | 7571 | non-null float64 |
| 40 | totalSavings2021_heoSo | 7592 | non-null float64 |
| 41 | balanceJuin2021 | 29956 | non-null float64 |
| 42 | nominal_interestJuin2021 | 29956 | non-null float64 |
| 43 | real_interestJuin2021 | 29956 | non-null float64 |
| 44 | nhomno_xhtdJuin2021 | 29956 | non-null float64 |
| 45 | categoryJuin2021 | 29956 | non-null float64 |
| 46 | sub_productJuin2021 | 20750 | non-null float64 |
| 47 | loaikyhanJuin2021 | 29956 | non-null object |
| 48 | sectorJuin2021 | 29956 | non-null float64 |
| 49 | product_codeJuin2021 | 29956 | non-null float64 |

dtypes: float64(40), int64(5), object(5)

Tổng quan Dữ liệu

100.000
quan sát

50
thuộc tính

Data Type:
Float, Integer, Object

Churn customer?

Không có giao dịch trong 3/2021 và 6/2021

2020 : dữ liệu cho huấn luyện mô hình

Data Wrangling

Xoá cột không sử dụng

Cột về Heoso

Cột về nhóm nợ

Cột Id

Chuẩn hoá dữ liệu

Cột giới tính

Tình trạng hôn nhân

Năm sinh/ Tuổi

Giao dịch 2020

Giữ cột amount_week

Dán nhãn có giao dịch (1) và không giao dịch (0)

Drop max_amount_week, min_amount_week

distinct_trans_group

Drop 'distinct_ref_no_week'

Drop 'distinct_ref_no_month'

Drop 'distinct_ref_no_3month'

distinct_payment_code

Tịnh tiến nhãn lên 1 đơn vị

Đặt các nhãn null bằng 0

count_payment_code

Drop 'count_payment_code_week'

Drop 'count_payment_code_month'

Xử lý Outliers bằng quantile

Tạo cột churn

8 giá trị trong cả tháng 3 và tháng 6 năm 2021 cùng Null

Sau bước tiền xử lý

```
→ <class 'pandas.core.frame.DataFrame'>
RangeIndex: 100000 entries, 0 to 99999
Data columns (total 14 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   gender            100000 non-null   object  
 1   marital_status    100000 non-null   object  
 2   distinct_payment_code_week  100000 non-null   float64 
 3   distinct_trans_group_week  100000 non-null   float64 
 4   distinct_payment_code_month 100000 non-null   float64 
 5   distinct_trans_group_month 100000 non-null   float64 
 6   distinct_payment_code_3month 100000 non-null   int64  
 7   count_payment_code_3month   100000 non-null   int64  
 8   distinct_trans_group_3month 100000 non-null   int64  
 9   age                100000 non-null   float64 
 10  check_trans_amount_week   100000 non-null   int32  
 11  check_trans_amount_month  100000 non-null   int32  
 12  check_trans_amount_3month 100000 non-null   int32  
 13  churn               100000 non-null   int32  
dtypes: float64(5), int32(4), int64(3), object(2)
memory usage: 9.2+ MB
```

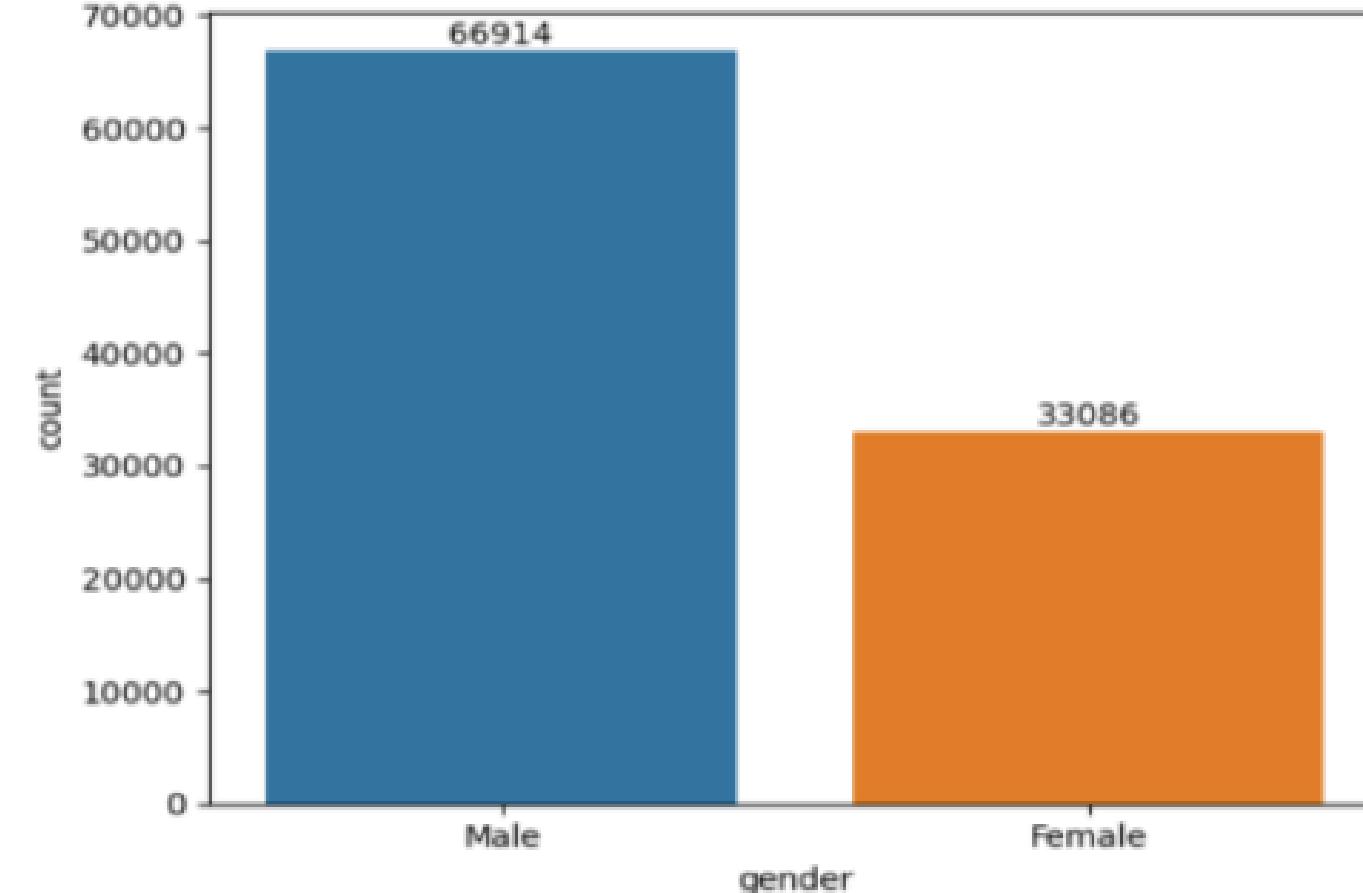
100.000
non-null

14
thuộc tính

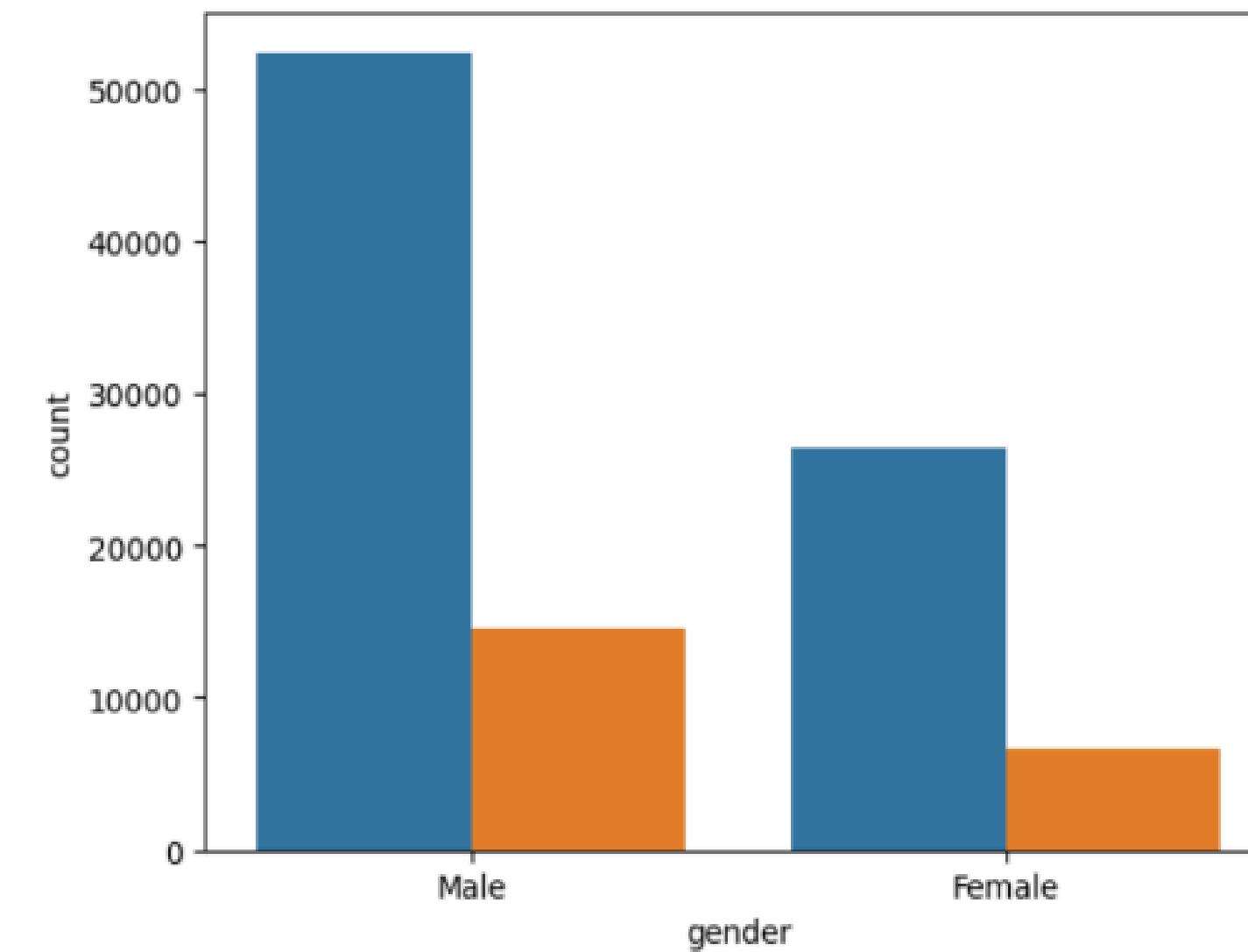
Data Type:
Float, Integer, Object

Count Gender

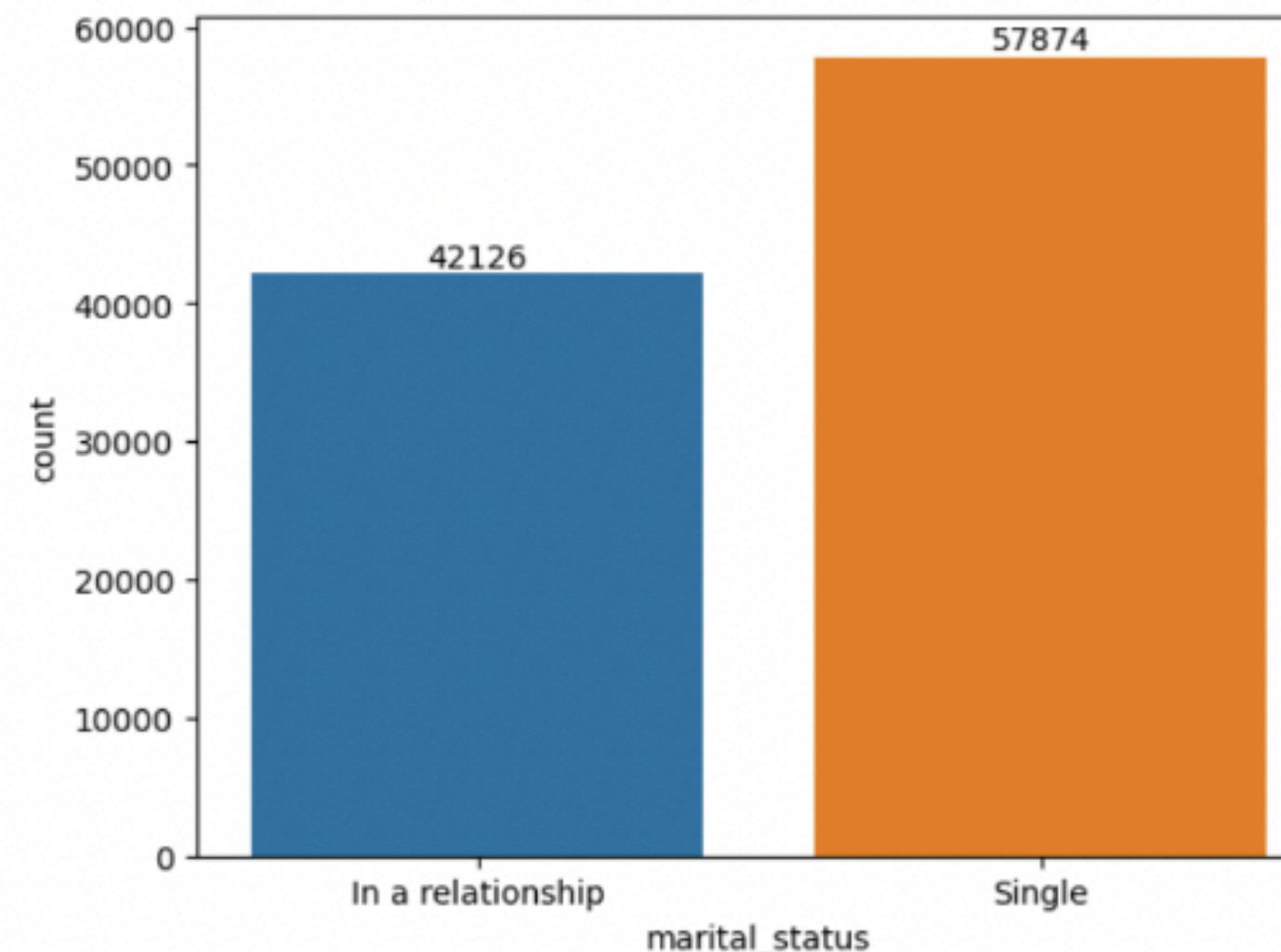
```
ax = sns.countplot(df, x="gender")
ax.bar_label(ax.containers[0])  
[Text(0, 0, '66914'), Text(0, 0, '33086')]
```



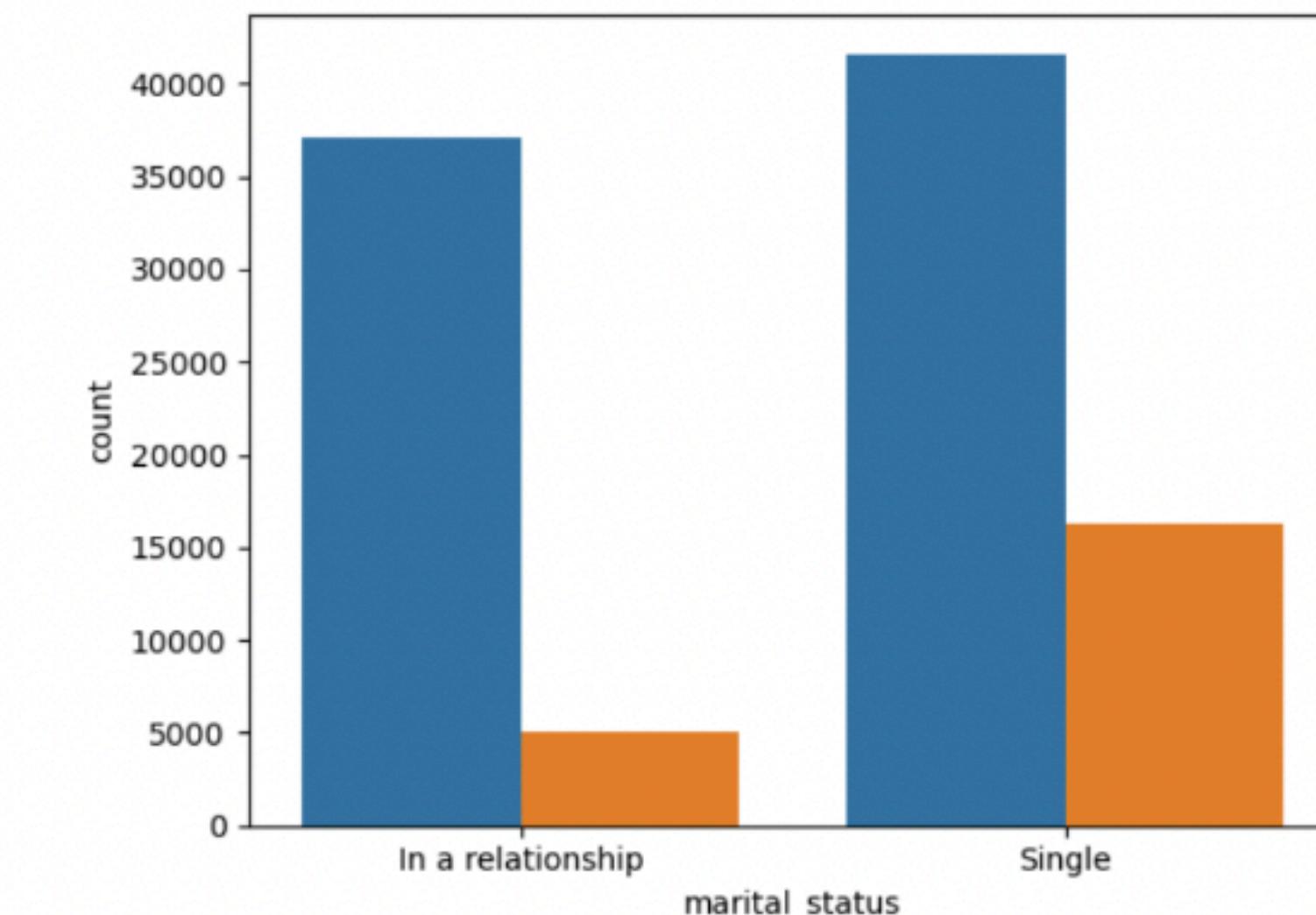
Count Gender by Churn



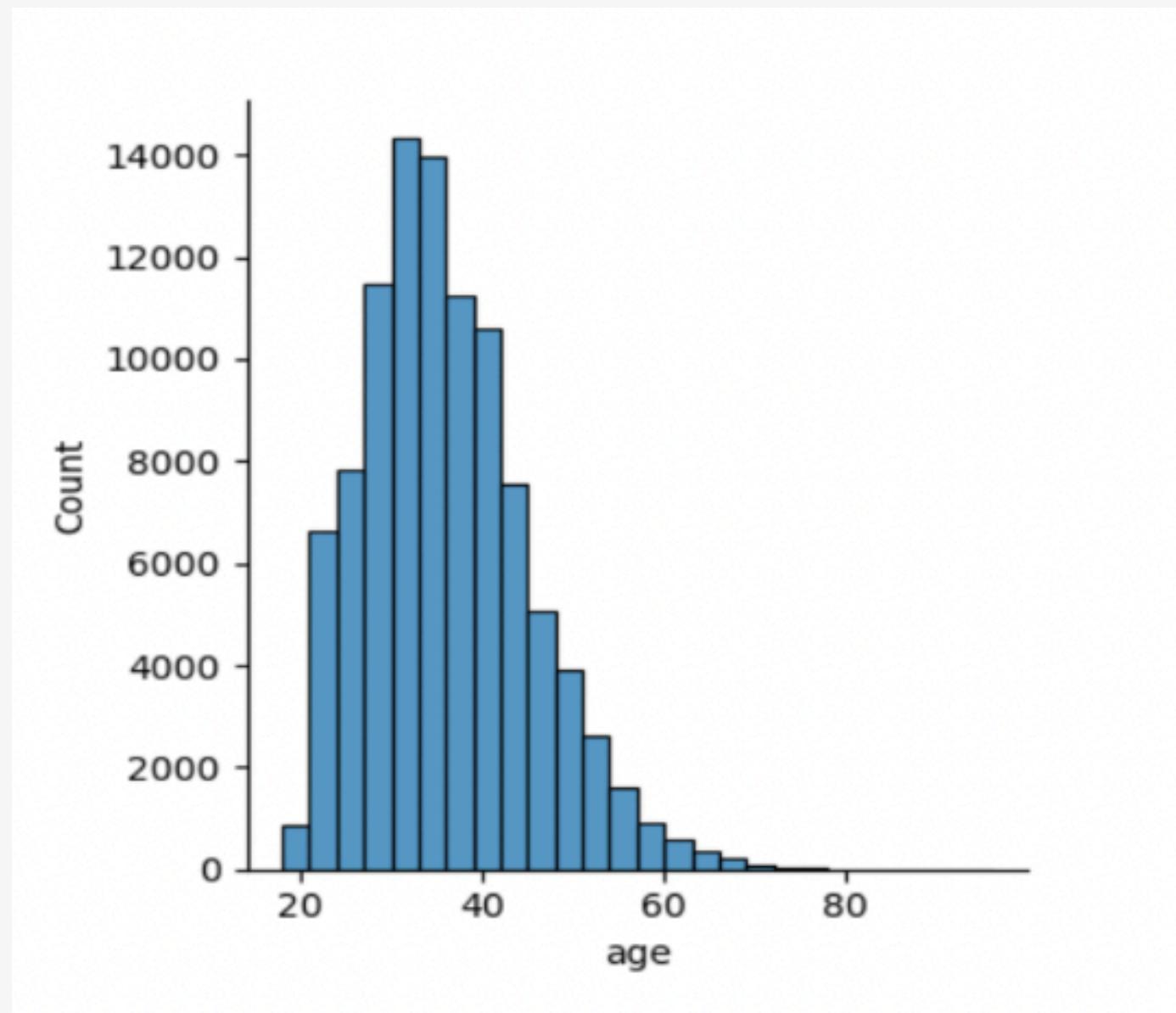
Count Marital_status



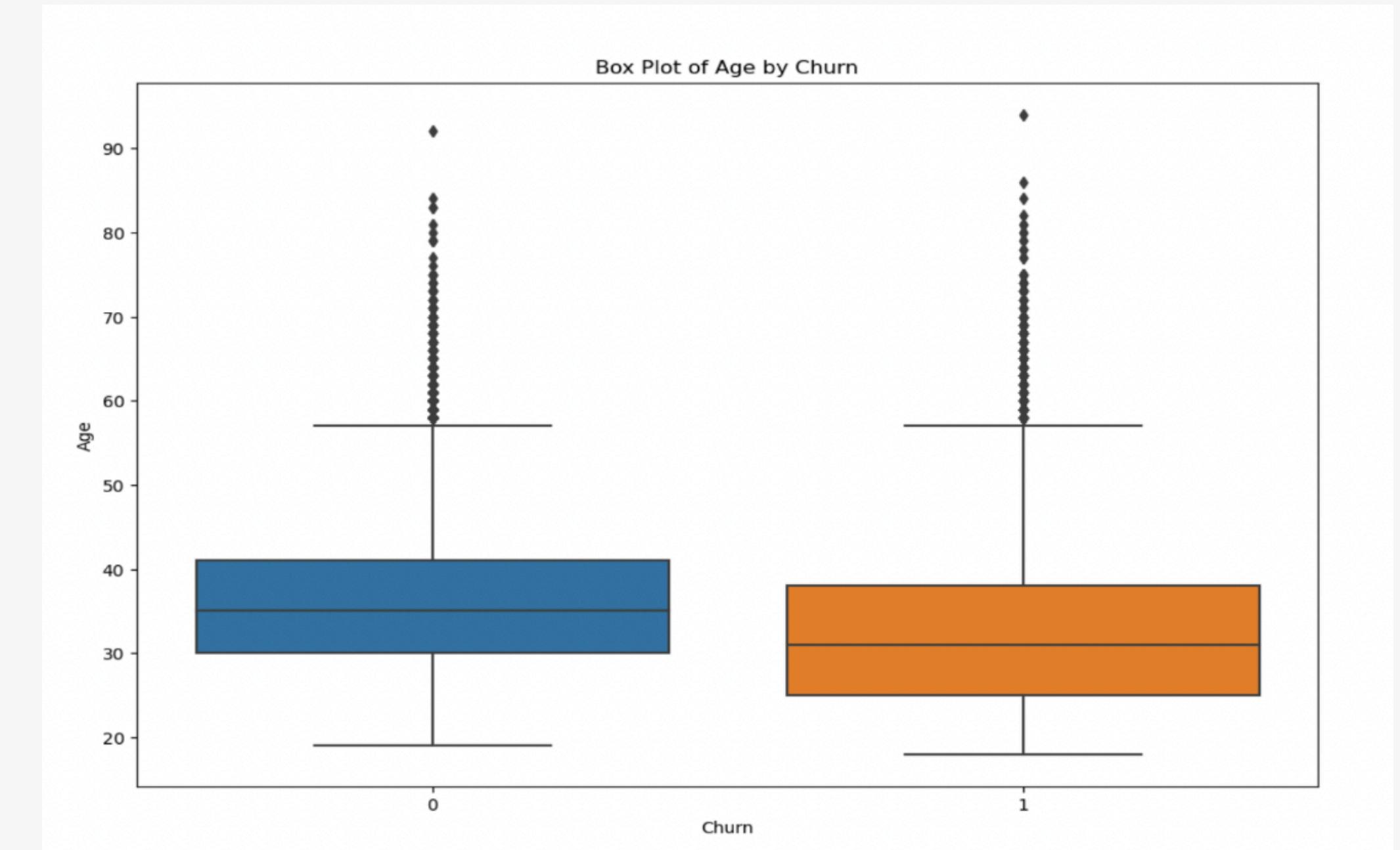
Count Marital_status by Churn



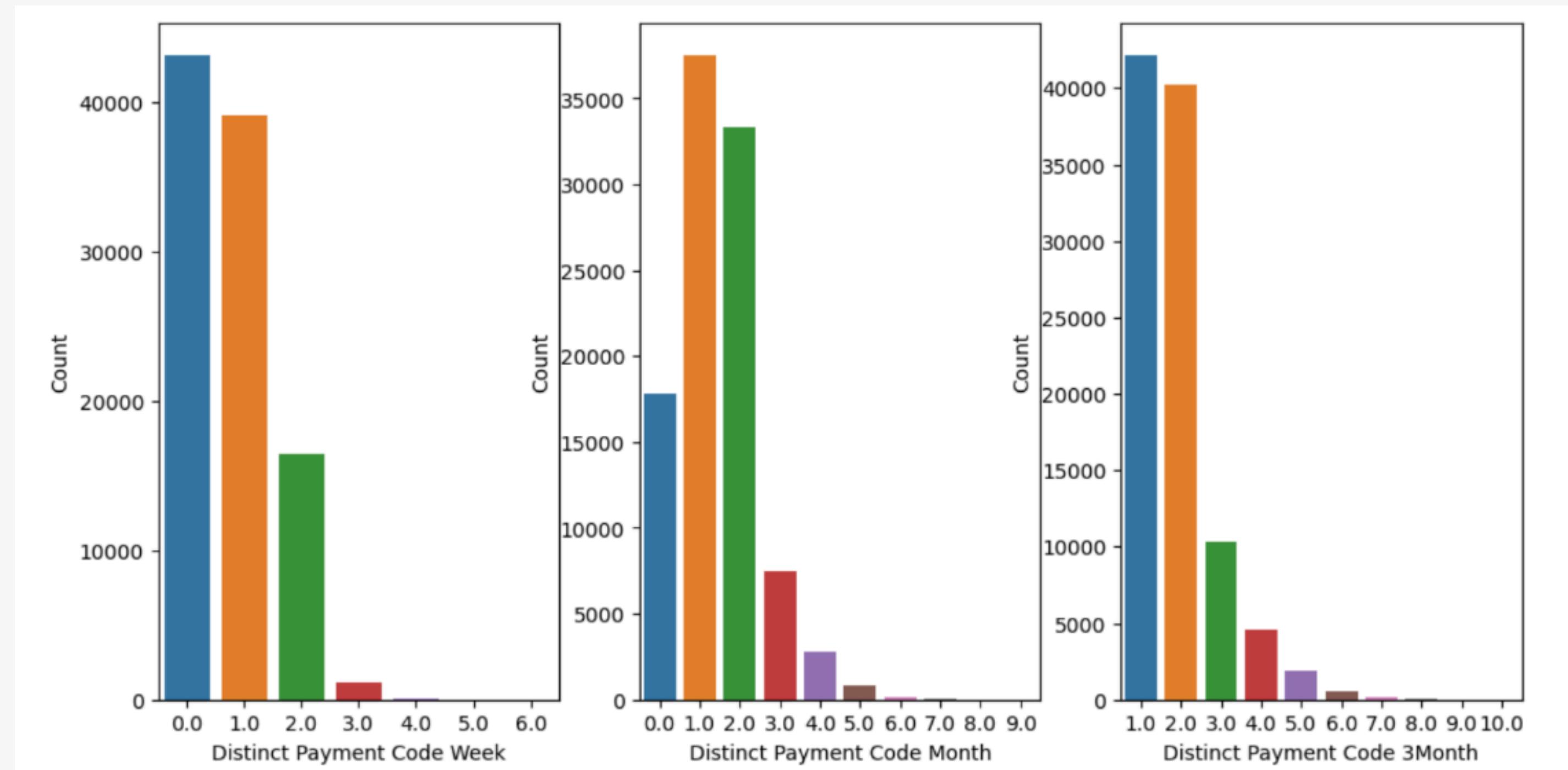
Count Age



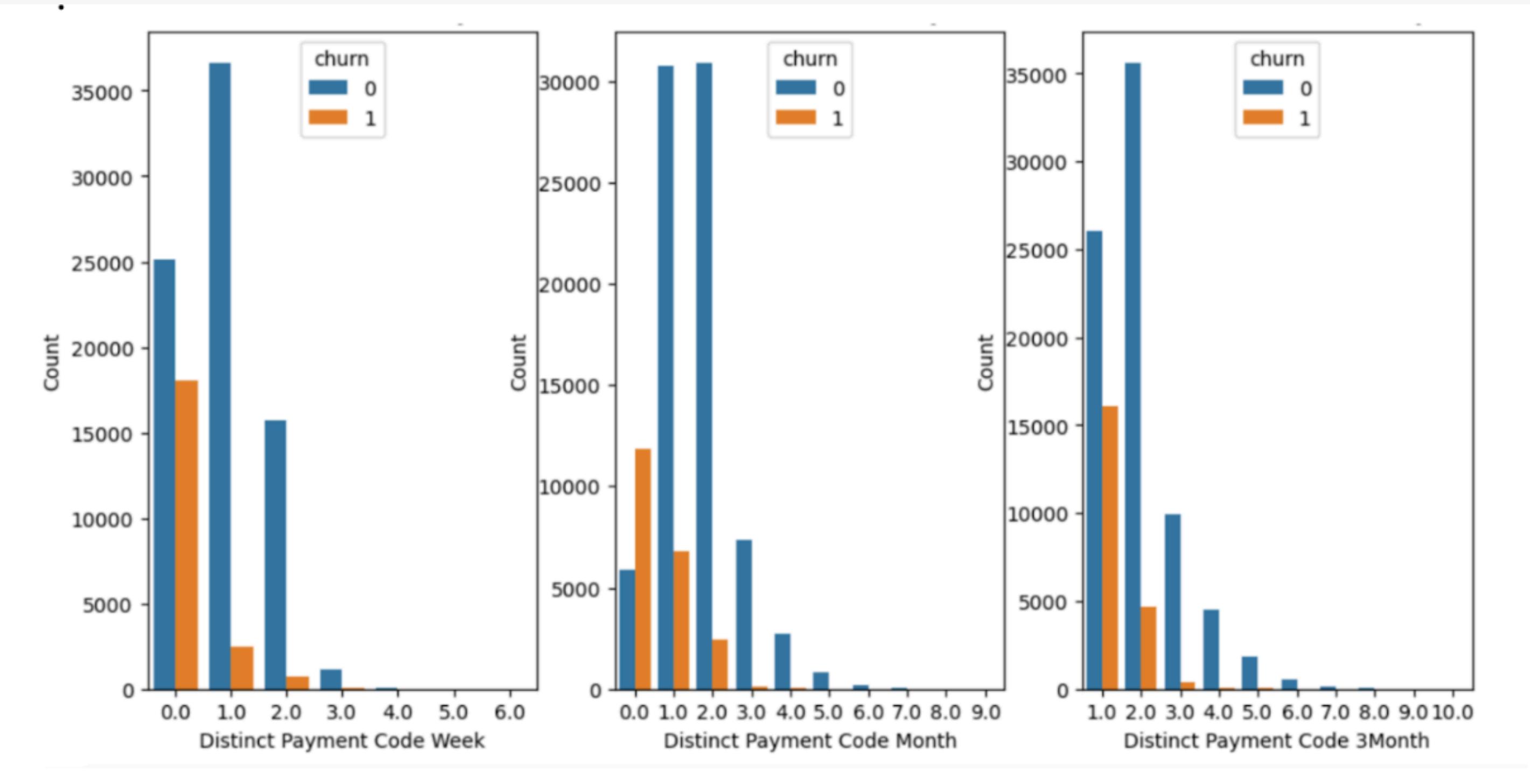
Boxplot Age by Churn



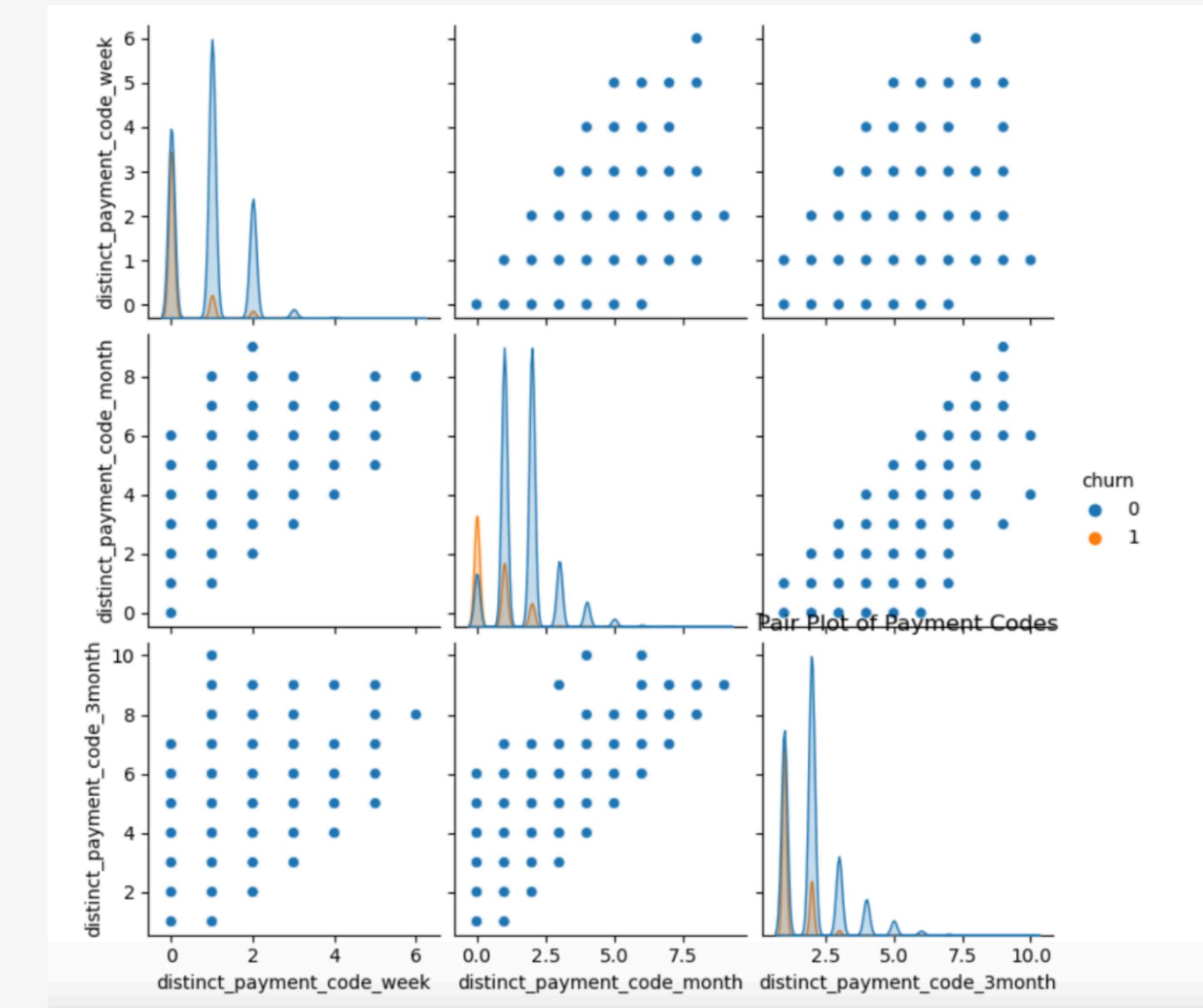
Count Distinct_payment_code



Count Distinct_payment_code by Churn

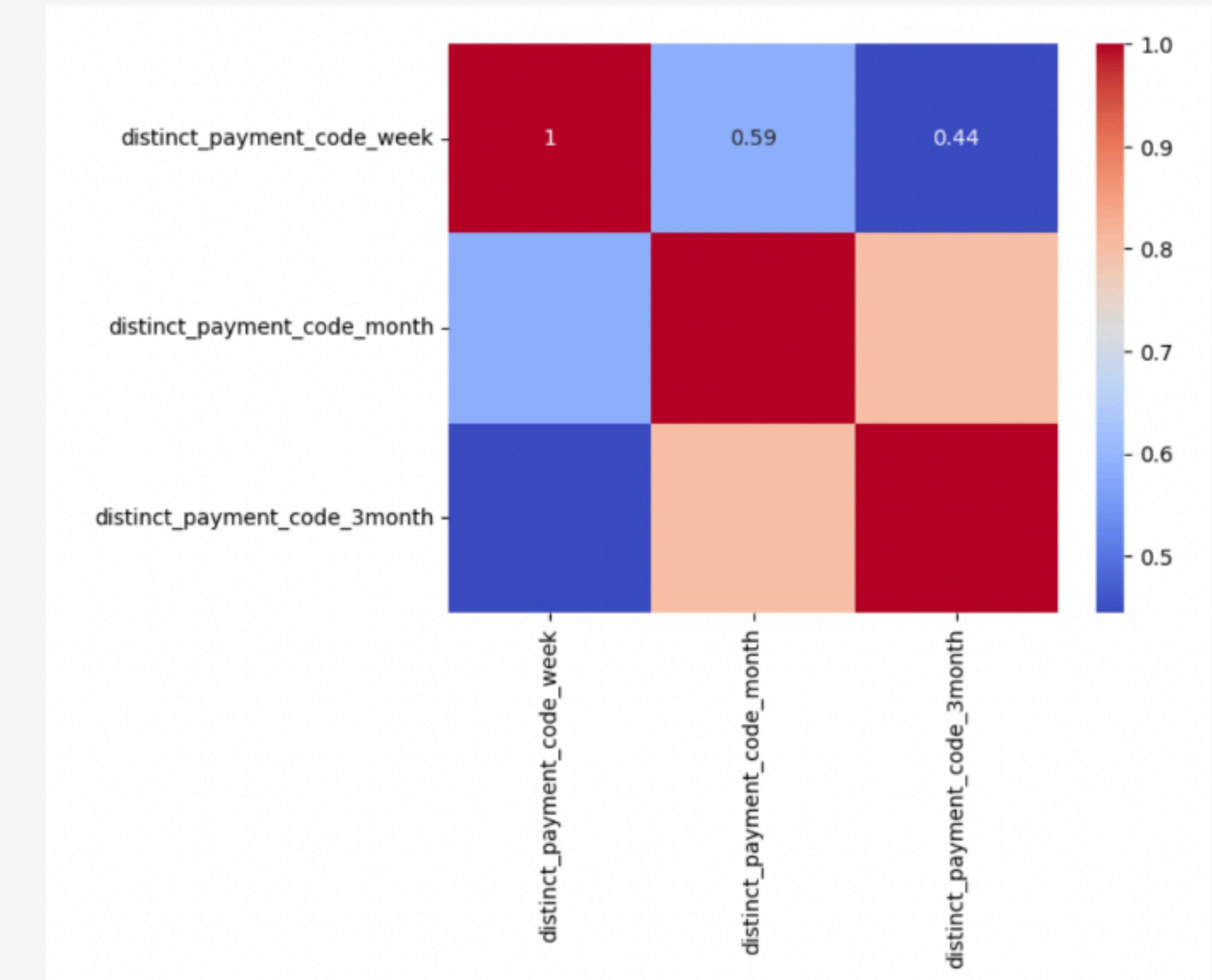


Distinct_payment_code Pairplot

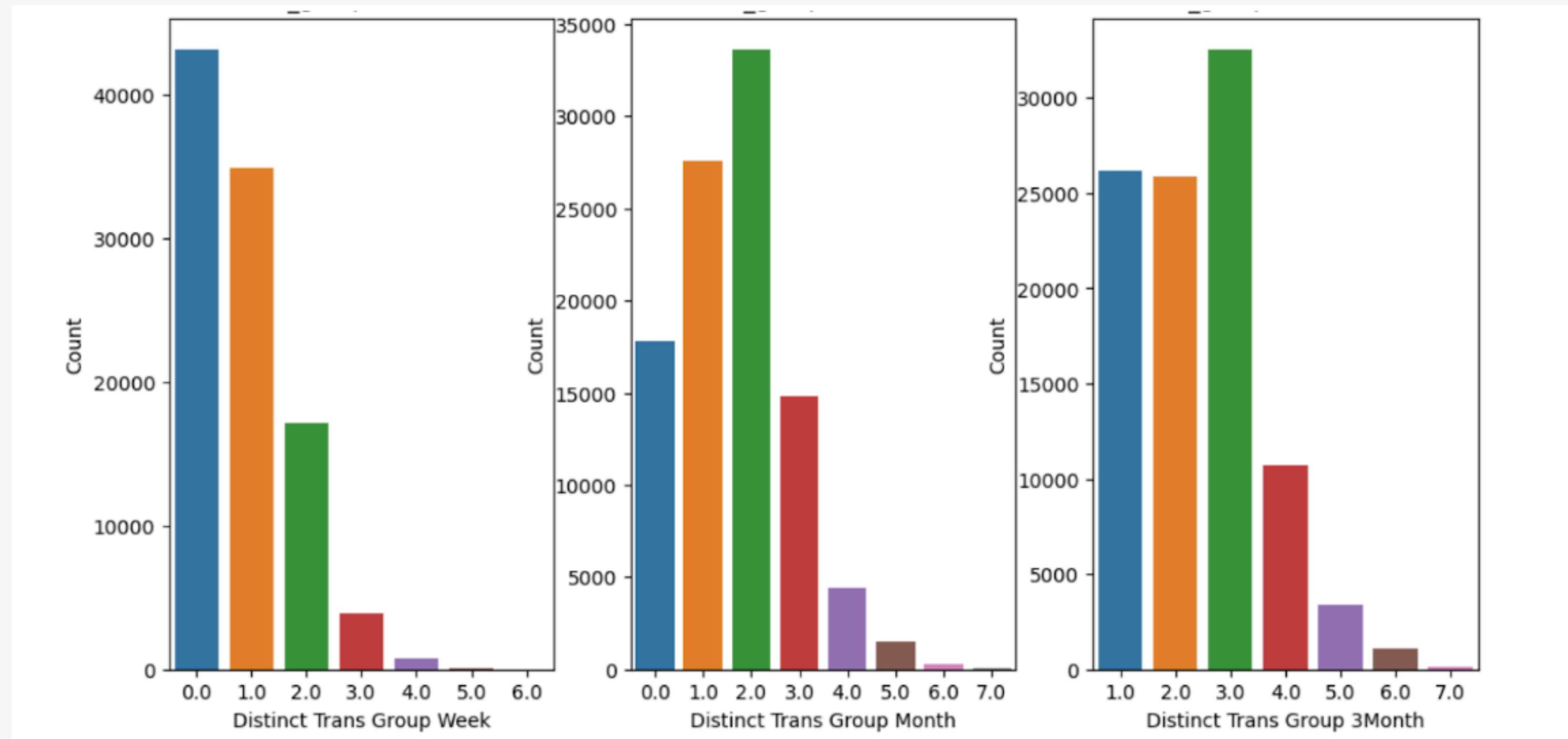


Correlation Distribution

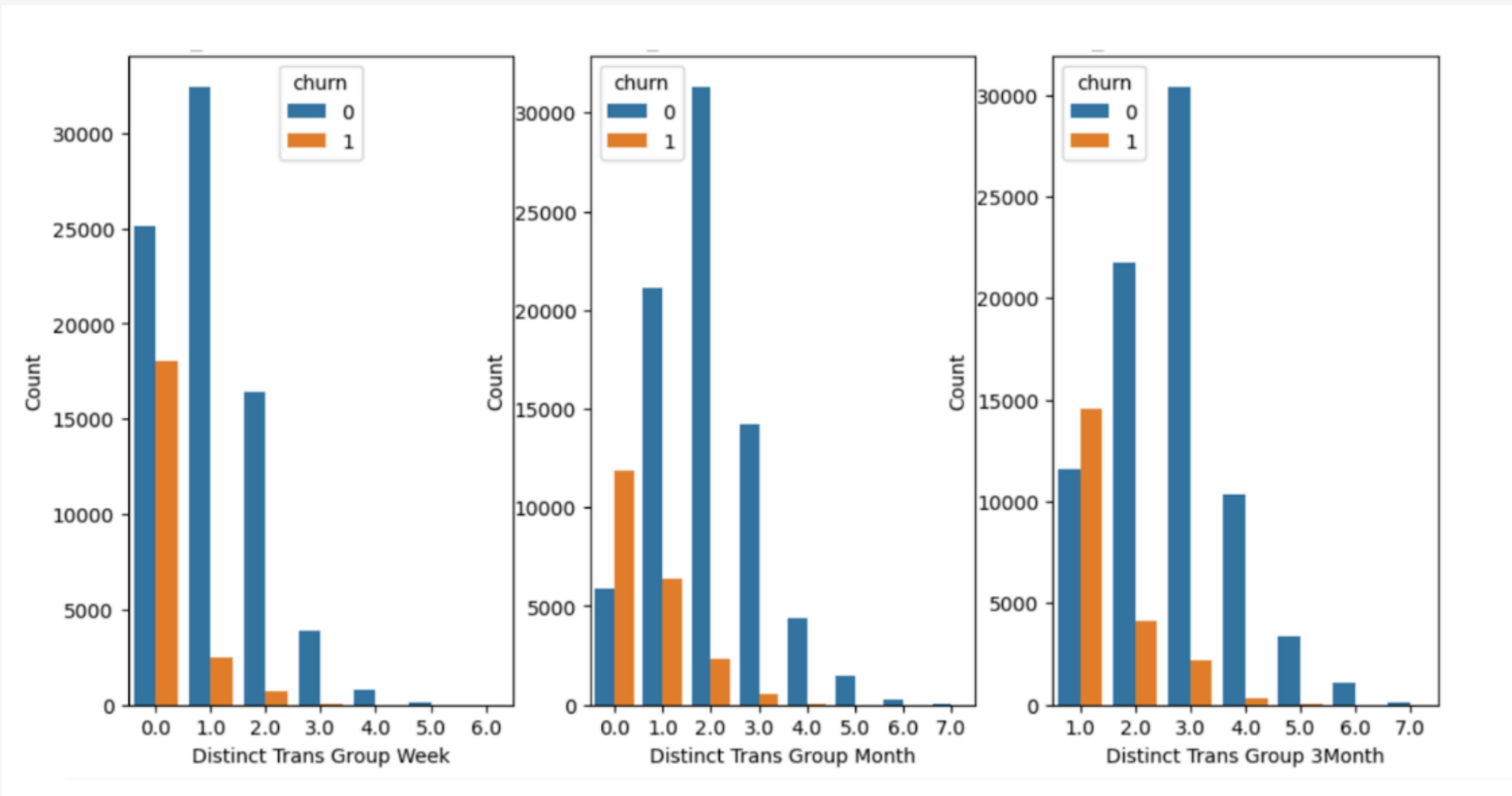
Distinct_payment_code



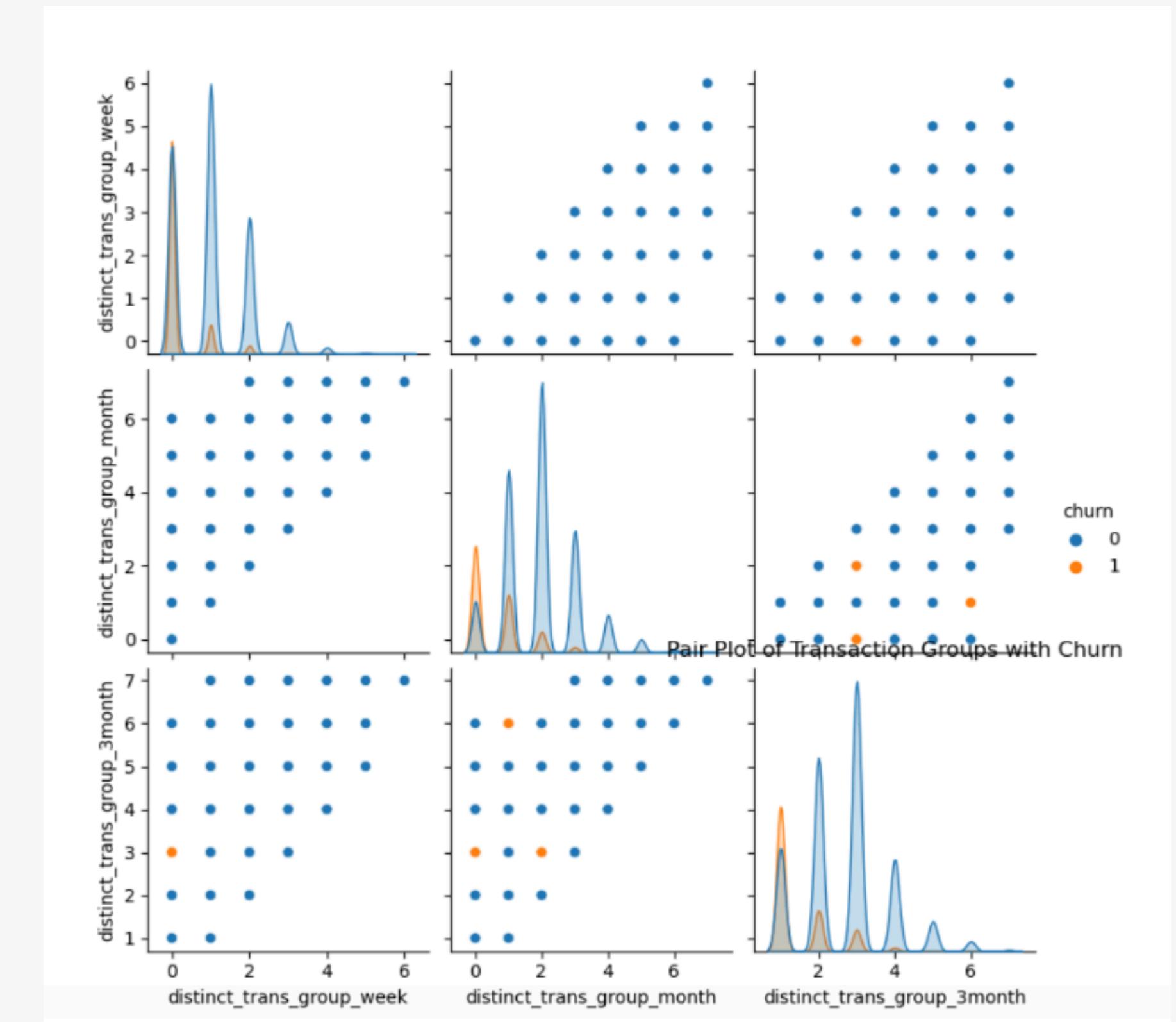
Count Distinct_trans_group



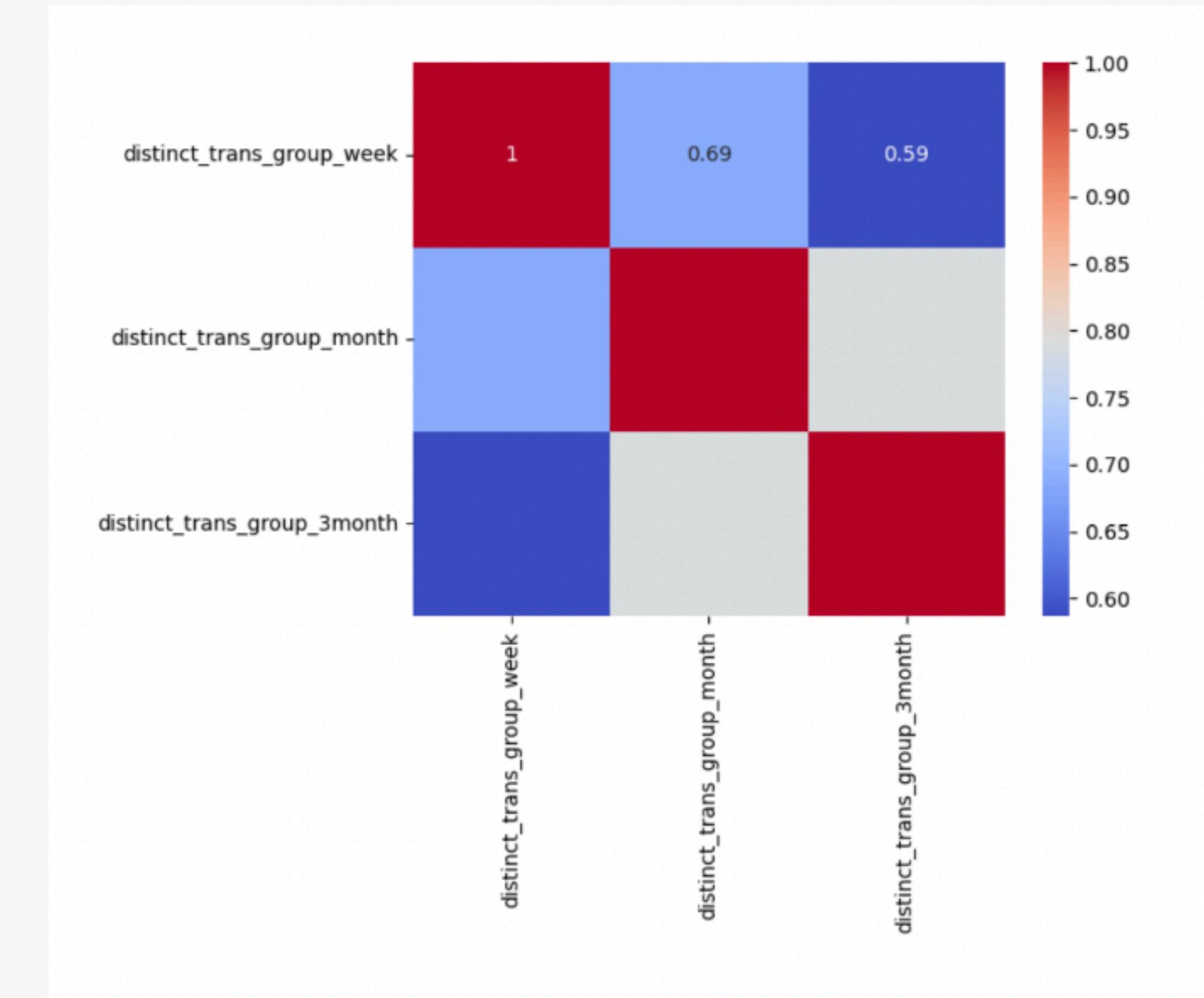
Count Distinct_trans_group by Churn



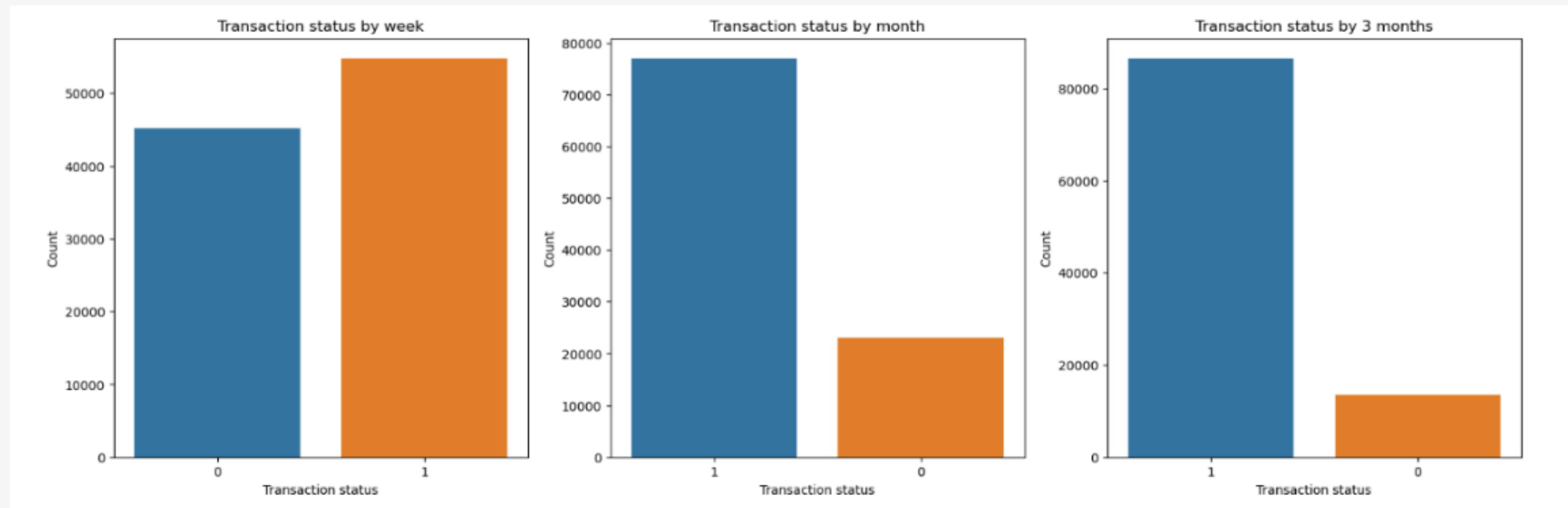
Pairplot Distinct_trans_group by Churn



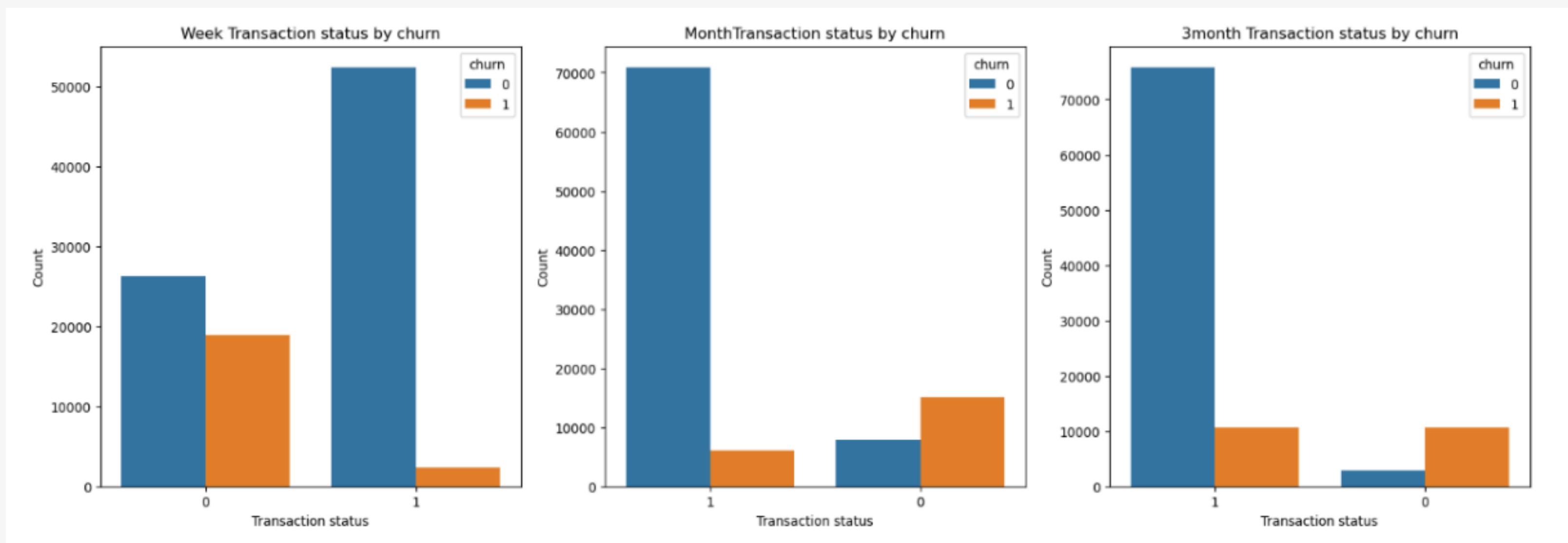
Correlation Distribution Distinct_trans_group



Count Check_trans_amount

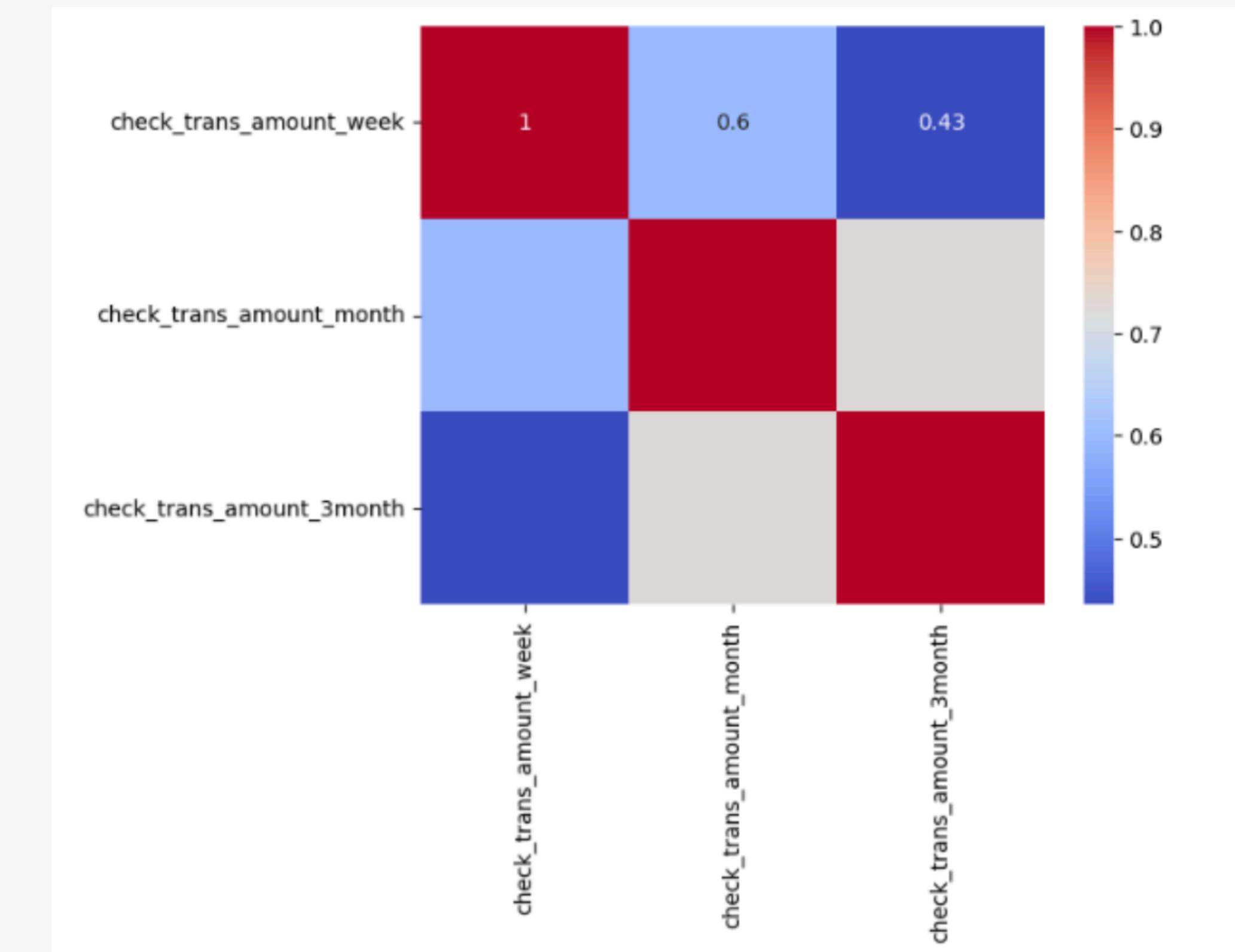


Count Check_trans_amount by Churn

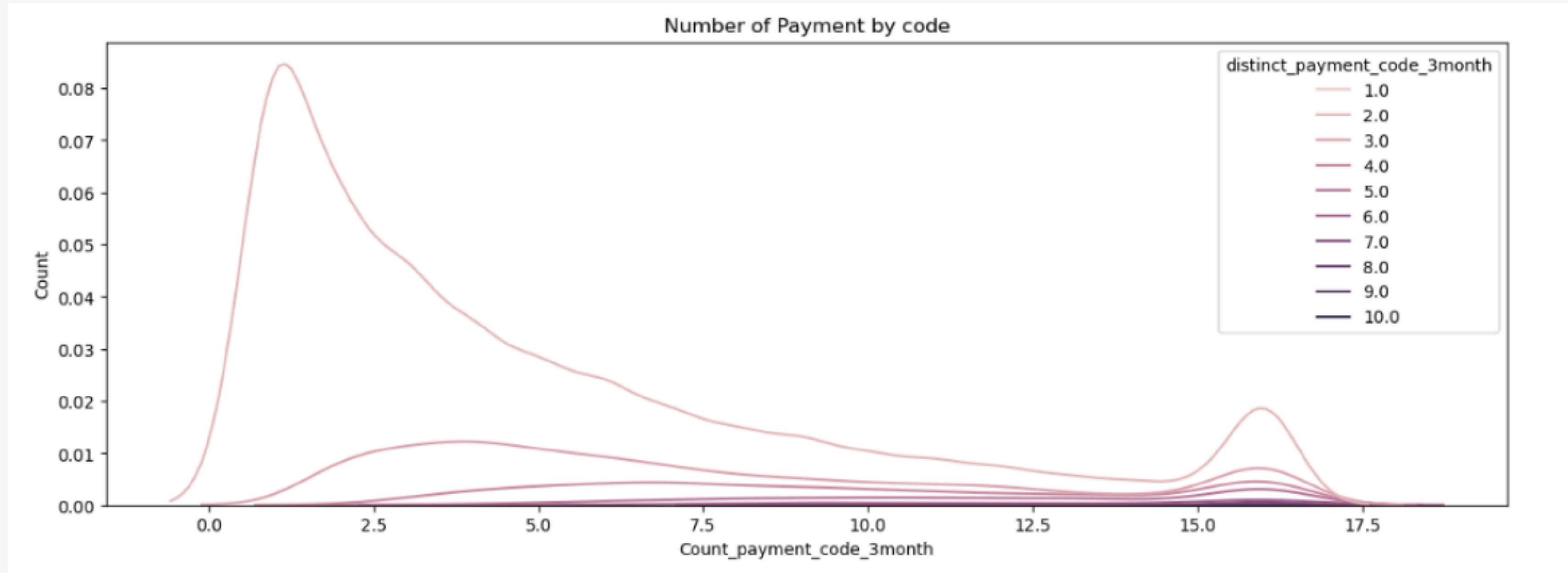


Correlation Distribution

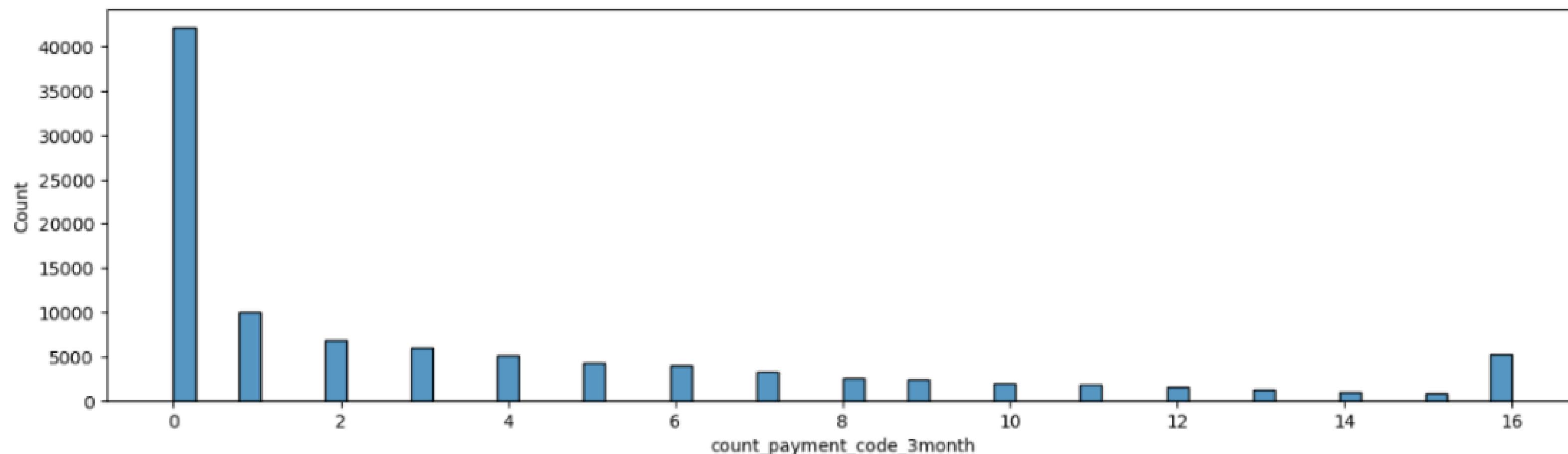
Check_trans_amount



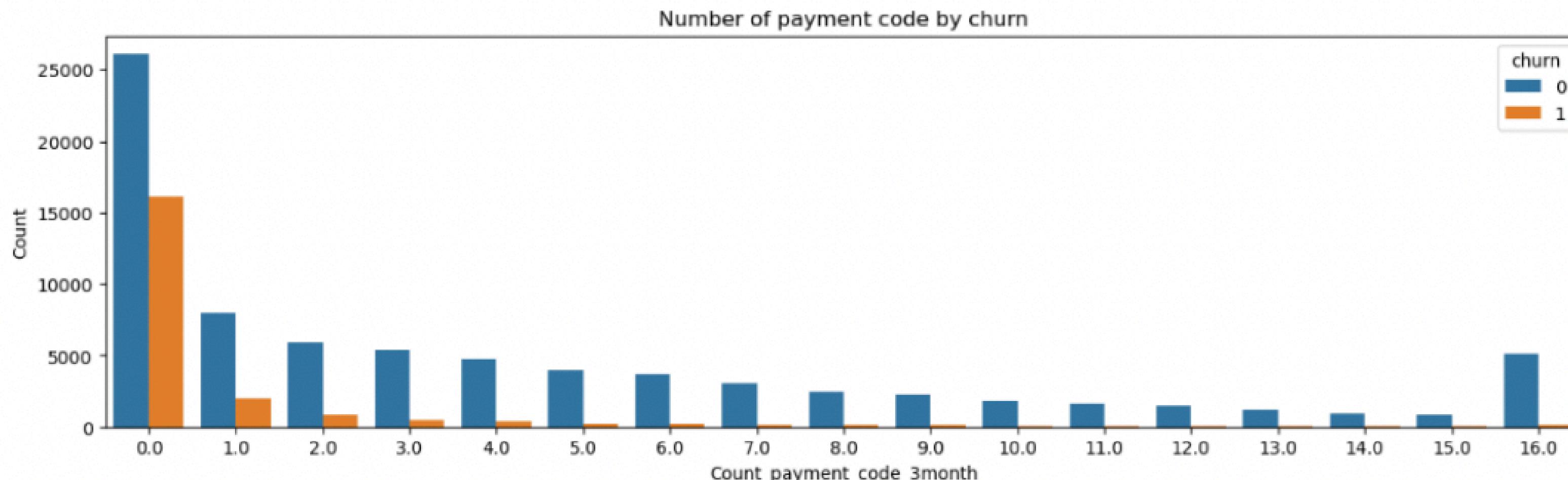
kdeplot count_payment_code_3month by code



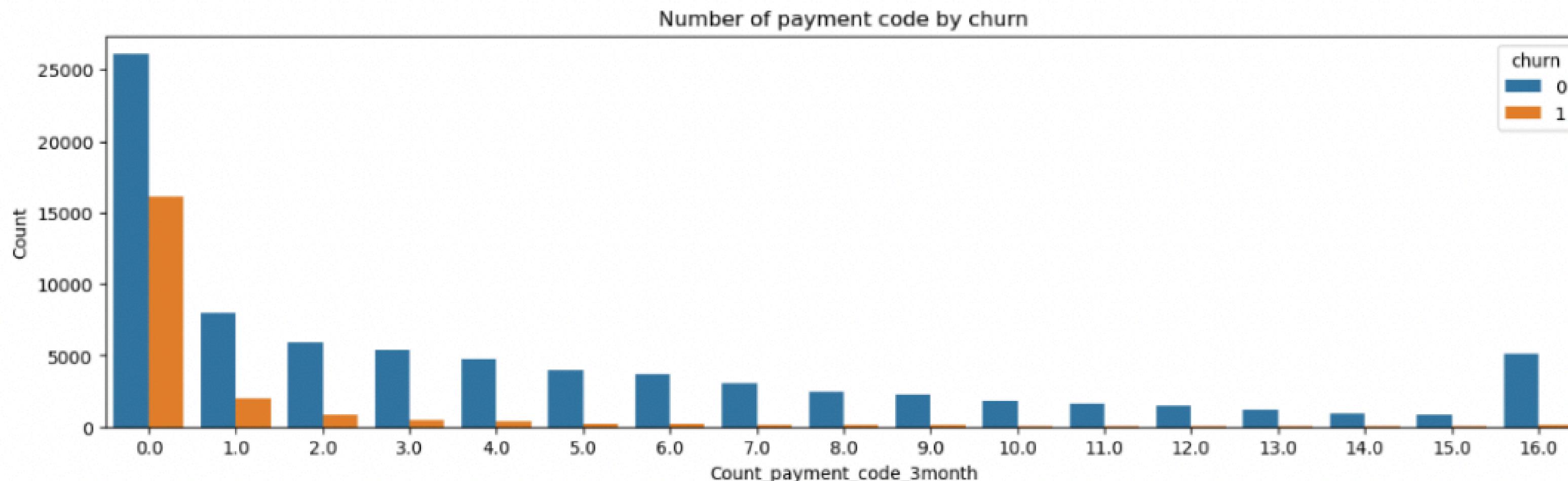
Count **count_payment_code_3month**



Count count_payment_code_3month by Churn



Count count_payment_code_3month by Churn



Model Development

Chia tách dữ liệu

Tập Train 70%

Tập Validation 15%

Tập Test 15%

Xử lý dữ liệu

Standard Scaler Age

Get Dummy gender

Get Dummy marital_status

11 biến được đưa vào mô hình

'gender'

'marital_status'

'distinct_payment_code_week'

'distinct_trans_group_week'

'distinct_payment_code_3month'

'count_payment_code_3month'

'distinct_trans_group_3month'

'age'

'check_trans_amount_week'

'check_trans_amount_month'

'check_trans_amount_3month'

Khai báo mô hình vào các lưới siêu tham số

Logistic regression

```
# L1 Regularization (Lasso) - Logistic Regression
logistic_l1 = LogisticRegression(solver='saga', penalty='l1', max_iter=10000)
param_grid_logistic_l1 = {
    'C': [0.1, 1, 10] # Regularization strength
}

# L2 Regularization (Ridge) - Logistic Regression
logistic_l2 = LogisticRegression(solver='lbfgs', penalty='l2', max_iter=10000)
param_grid_logistic_l2 = {
    'C': [0.1, 1, 10] # Regularization strength
}
```

Decision Tree

```
# Decision Tree
decision_tree = DecisionTreeClassifier(random_state=42)
param_grid_decision_tree = {
    'max_depth': [10, 20],
    'min_samples_split': [2, 10],
    'min_samples_leaf': [1, 5]
}
```

Random Forest

```
# Random Forest
random_forest = RandomForestClassifier(random_state=42)
param_grid_random_forest = {
    'n_estimators': [50, 100],
    'max_depth': [10, 20],
    'min_samples_split': [2, 10],
    'min_samples_leaf': [1, 5]
}
```

Tổng hợp các mô hình vào lưới siêu tham số

```
# Tổng hợp các mô hình vào lưới siêu tham số
models = [
    (logistic_l1, param_grid_logistic_l1),
    (logistic_l2, param_grid_logistic_l2),
    (decision_tree, param_grid_decision_tree),
    (random_forest, param_grid_random_forest)
]
```

Tuning trên từng model bằng loop for

```
for model, param_grid in models:
    grid_search = GridSearchCV(model, param_grid, cv=2)
    grid_search.fit(X_train, y_train)
    best_model = grid_search.best_estimator_
    val_score = f1_score(y_val, best_model.predict(X_val))
    best_models.append(best_model)
    val_scores.append(val_score)

# Lựa chọn best model dựa trên tập validation
best_model_index = np.argmax(val_scores)
best_model = best_models[best_model_index]
print(f"Best Model: {best_model}")
print(f"Validation Set Score: {val_scores[best_model_index]})

# Combine tập train + validation để retrain lại cho mô hình tốt hơn nữa
X_combined = np.vstack((X_train, X_val))
y_combined = np.hstack((y_train, y_val))
best_model.fit(X_combined, y_combined)

# Đánh giá final model trên tập test
test_score = f1_score(y_test, best_model.predict(X_test))
print("Test Set Score: ", test_score)
```

Kết quả Model Tuning

Best Model

RandomForestClassifier
(max_depth=10, min_samples_leaf=5,
random_state=42)

Validation Set Score

0.6263265686091974

Test Set Score

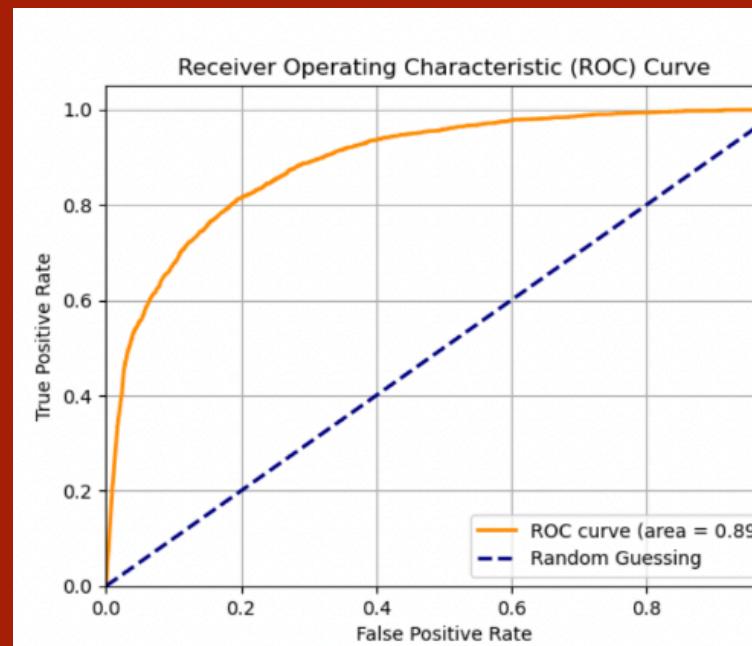
0.6322075013995148

Đánh giá mô hình

Classification Report

| | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 0 | 0.88 | 0.96 | 0.92 | 11811 |
| 1 | 0.78 | 0.53 | 0.63 | 3189 |
| accuracy | | | 0.87 | 15000 |
| macro avg | 0.83 | 0.75 | 0.78 | 15000 |
| weighted avg | 0.86 | 0.87 | 0.86 | 15000 |

ROC
curve



Mô hình Random Forest hoạt động tốt trong việc dự đoán non-churn (0) với các chỉ số Precision, Recall và F1-score cao.

Tuy nhiên, mô hình gặp khó khăn trong việc dự đoán churn (1), với Recall thấp (0.53) và F1-score trung bình (0.63).

=> Oversampling bằng SMOTE với mô hình XGBoost.

Đánh giá mô hình sau khi Oversampling

Classification Report

Trên tệp Validation

Validation F1 Score: 0.654936974789916

Validation Classification Report:

| | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 0 | 0.93 | 0.84 | 0.88 | 11811 |
| 1 | 0.56 | 0.78 | 0.65 | 3189 |
| accuracy | | | 0.82 | 15000 |
| macro avg | 0.75 | 0.81 | 0.77 | 15000 |
| weighted avg | 0.86 | 0.82 | 0.83 | 15000 |

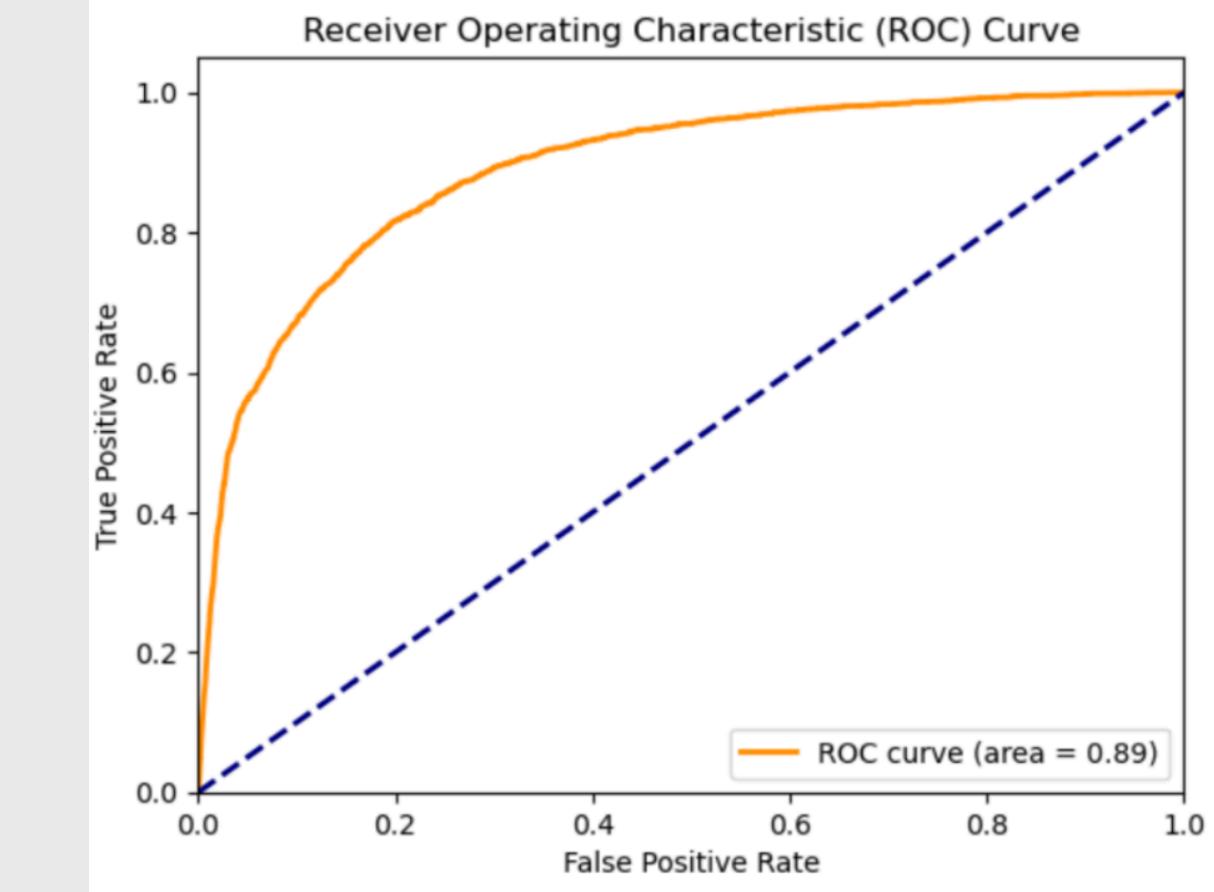
Trên tệp Test

Test F1 Score: 0.6521332085060854

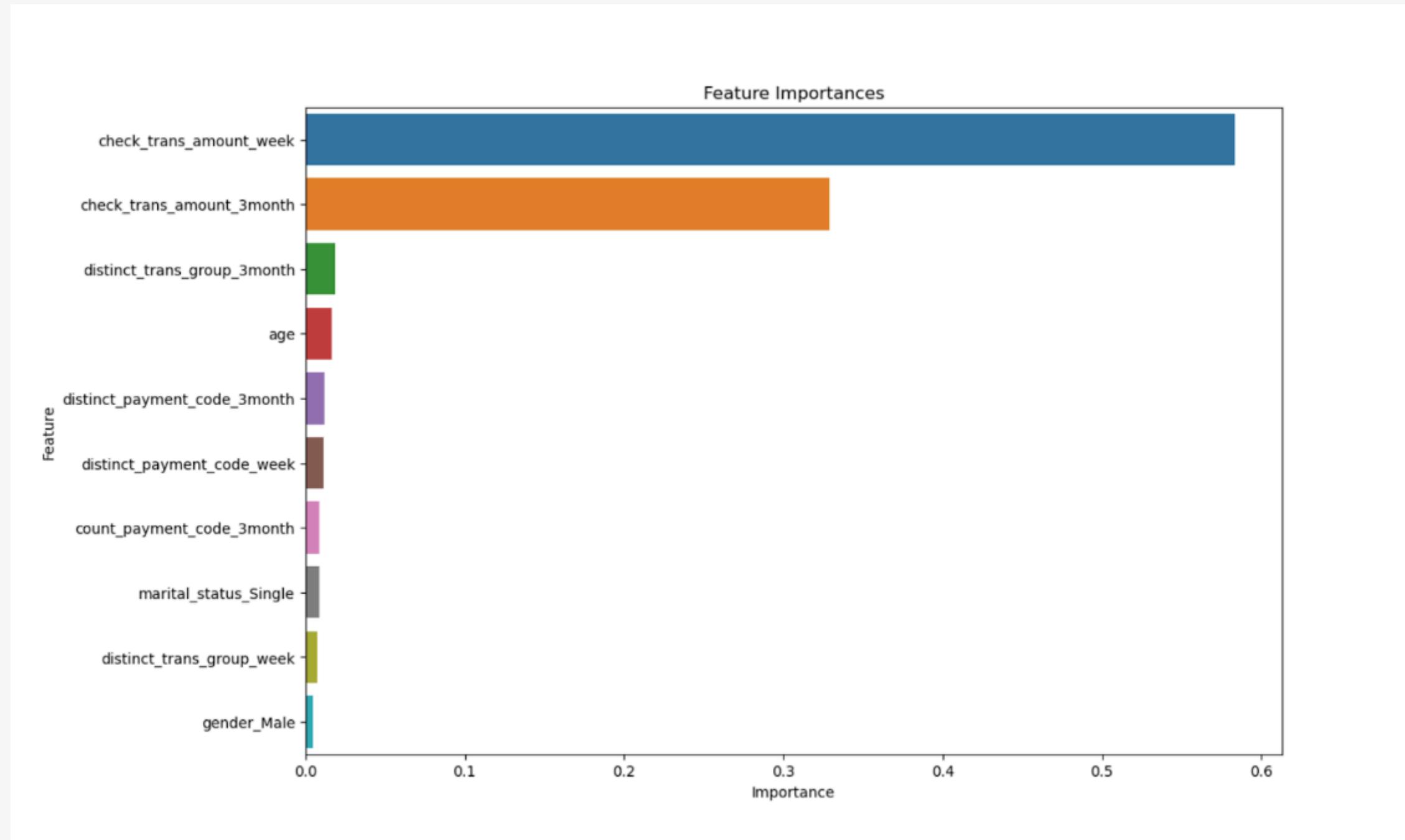
Test Classification Report:

| | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 0 | 0.93 | 0.84 | 0.88 | 11811 |
| 1 | 0.57 | 0.76 | 0.65 | 3189 |
| accuracy | | | 0.83 | 15000 |
| macro avg | 0.75 | 0.80 | 0.77 | 15000 |
| weighted avg | 0.85 | 0.83 | 0.84 | 15000 |

ROC curve



Model's Feature Importances



High Importance

check_trans_amount_week
check_trans_amount_3month

Medium Importance

distinct_transgroup_3month
age

Low Importance

payment_code
Others...

Hạn chế và định hướng

Hạn chế

- Bộ dữ liệu chứa nhiều outliers và missing values, việc xử lý dựa trên các giả định có thể chưa phù hợp với bản chất của dữ liệu do mô tả dữ liệu còn mơ hồ.
- Bộ dữ liệu có ít biến với tương quan cao, nhóm chỉ sử dụng biến có tính đại diện cao nhất nên mô hình có thể ít biến.
- Chỉ số đánh giá của mô hình chưa được cải thiện, đặc biệt trong việc xác định đúng khách hàng churn, ảnh hưởng đến việc đưa ra chiến lược giữ chân khách hàng hiệu quả.

Định hướng

- Tìm hiểu rõ đặc điểm của bộ dữ liệu và nguyên nhân thực tế để áp dụng các cách xử lý phù hợp hơn.
- Phát triển thêm các trường dữ liệu khác liên quan đến hành vi giao dịch của khách hàng.
- Tìm hiểu và áp dụng các thuật toán khác để phát triển và tinh chỉnh mô hình.
- Tìm hiểu thêm các thang đo đánh giá khác để có cái nhìn toàn diện về mô hình.
- Tránh can thiệp quá mức vào dữ liệu gây méo mó mô hình và kết quả phản ánh sai lệch thực tế.



THANK YOU

FDC105 - Group 1

