Health Insurance Cross Sale Prediction

Student: Nguyen Anh Tai - FX13245
Mentor: Nguyen Huy Thanh

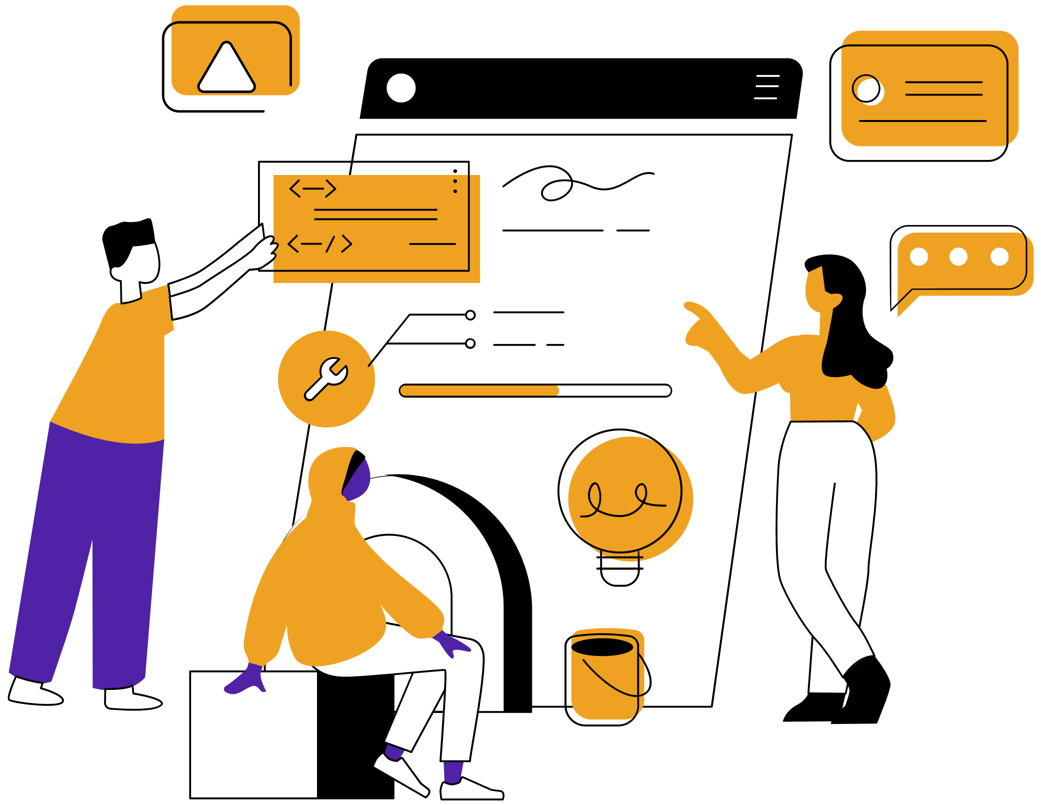
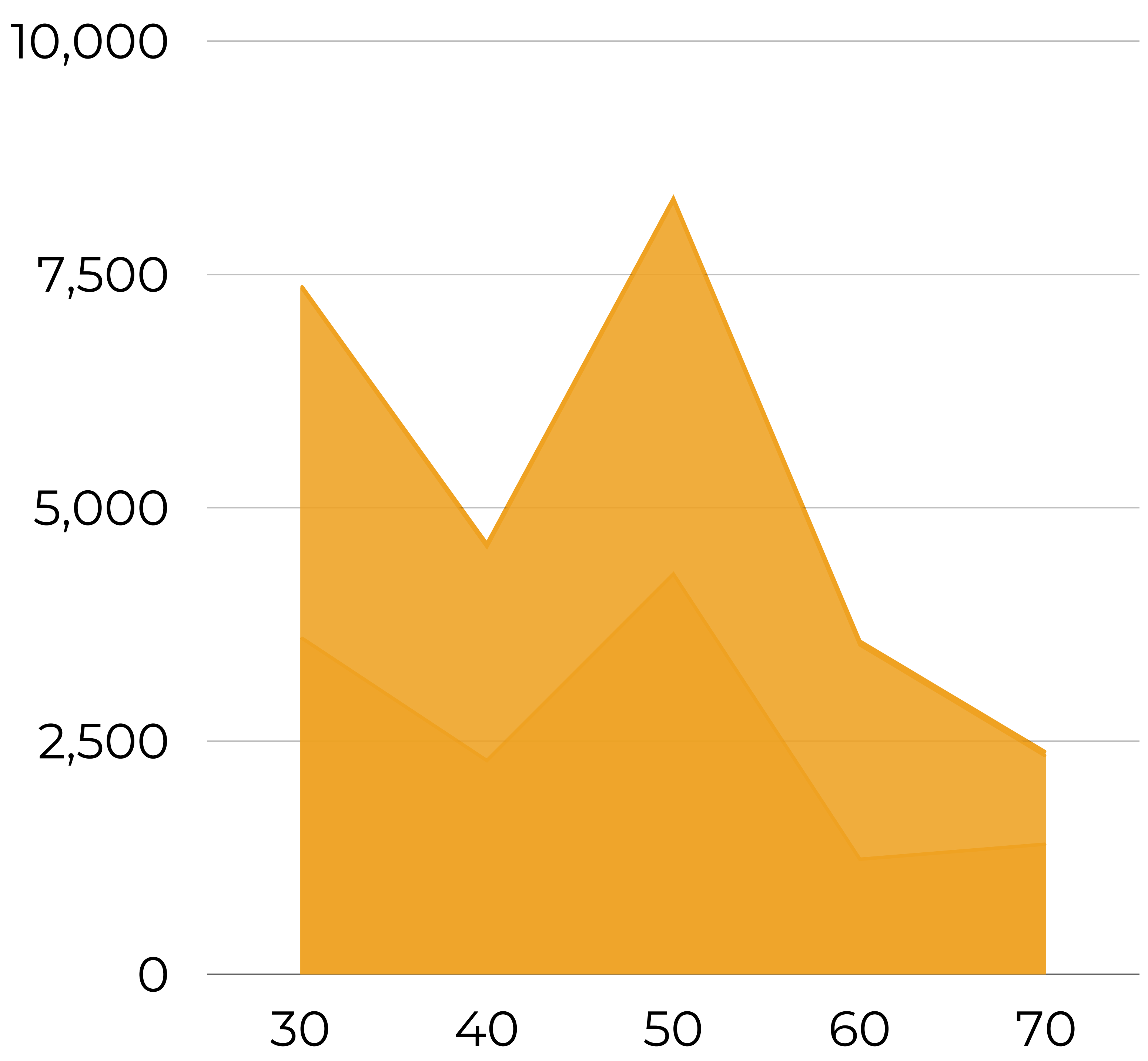


Table of Contents

The presentation would go through the following main points

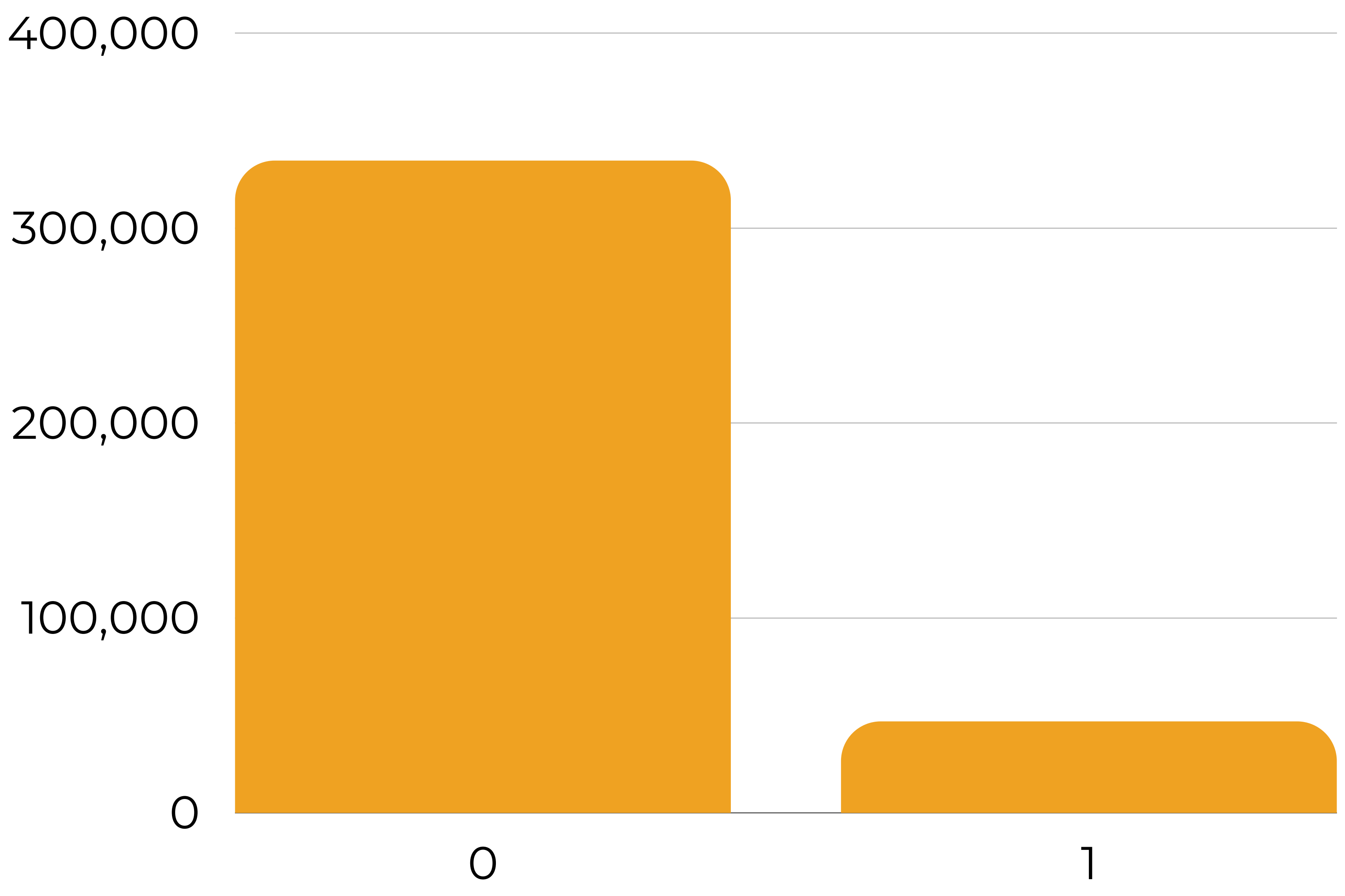
Problem Statement

The client in the context is an insurance company which has provided health insurance to the customers. Now, they need a model which helps predict the likelihood that the customers from the previous year will choose the vehicle insurance offered by the company.



Objectives

- Identify the driving forces behind the conversion to vehicle insurance among the customers
- Build a model which predicts the potential customers who would be interested in vehicle insurance



Metrics

- A good model should obtain high f1 score (at least 0.4)
- For the purpose of profit generation, high recall rate (at least 0.8) is preferable

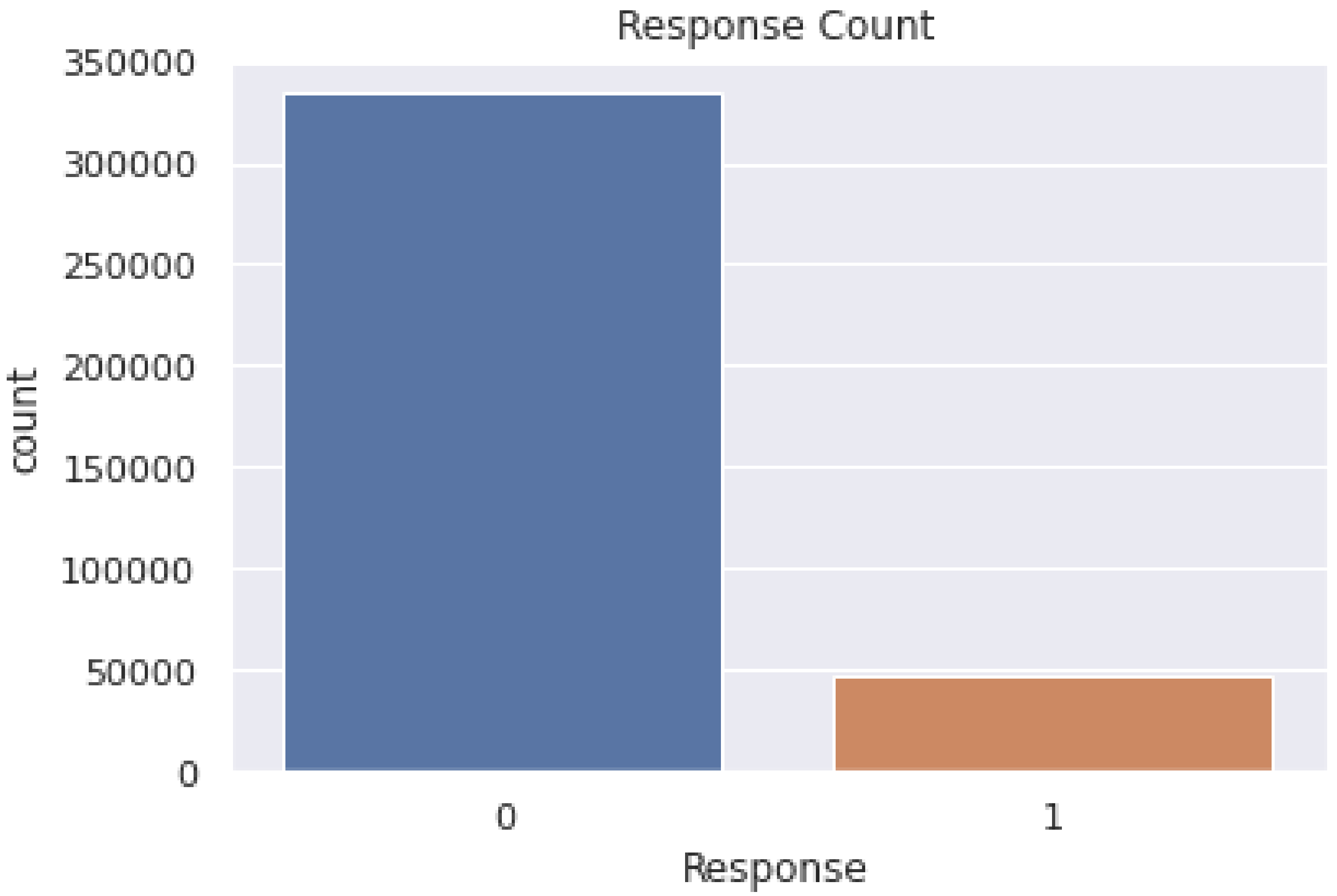
- ID: Unique ID for the customers
- Gender: Gender of the customers
- Age: Age of the customers
- Driving license: 0 - no DL, 1 - Already have
- Region Code: Unique code for the customers' region
- Previously Insured: 0 - no vehicle insurance, 1 - Already have
- Vehicle Age: Age of the vehicle
- Vehicle Damage: 0 - not get vehicle damaged in the past, 1 - get vehicle damaged in the past
- Annual Premium: The amount of premium paid in a year
- Policy Sales Channel: Channel of outreaching to the customers
- Vintage: Number of days associated with the company
- Response: 0 - customer not interested, 1 - customer is interested

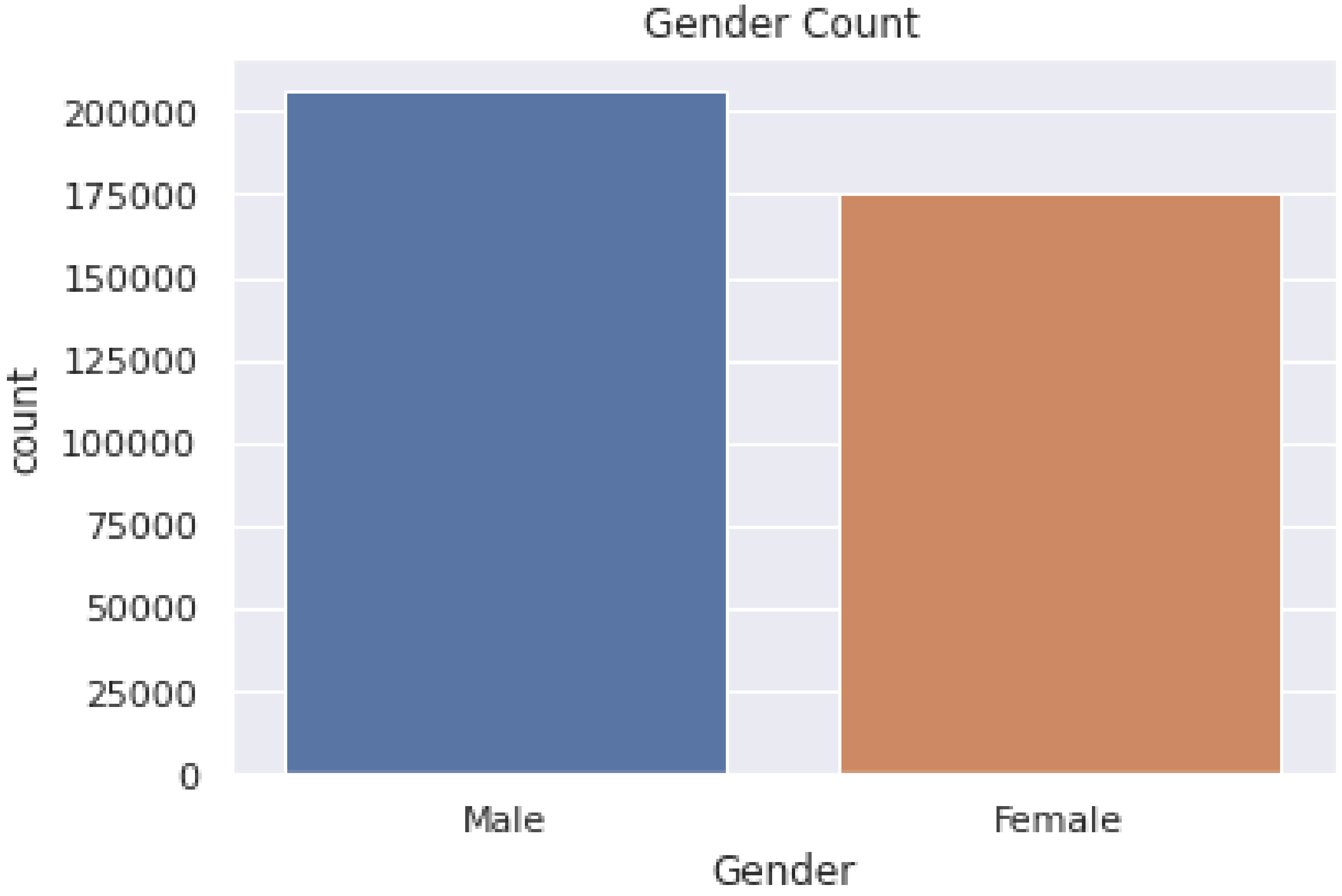
Data Understanding

Data distribution

Dataset is **imbalanced** towards the **negative** target (**uninterested** responses)

Dataset has a **balanced gender** distribution

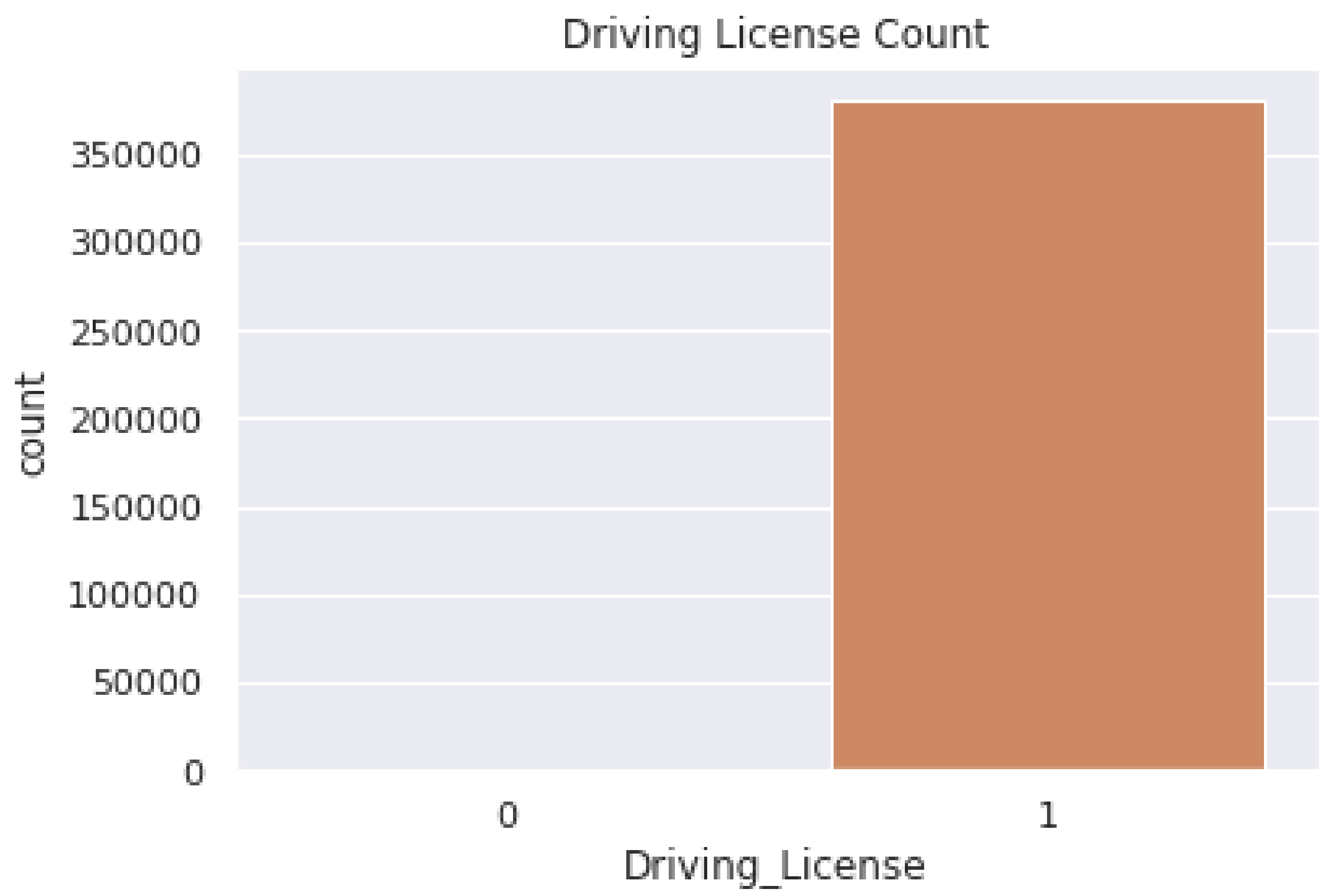


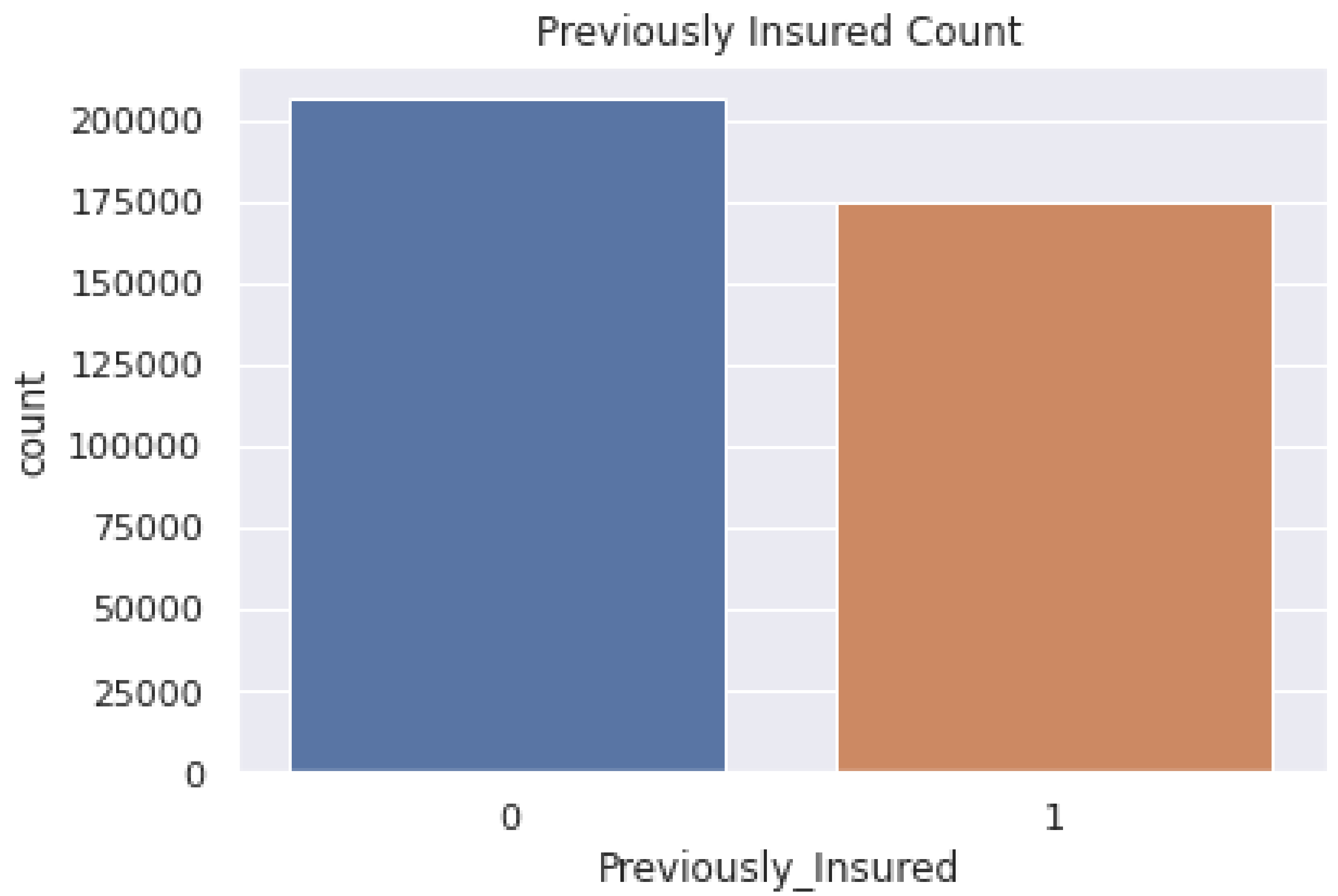


Data distribution

A great majority of customers have **driving license**

A great number of customers are **not previously insured --> potential customers**

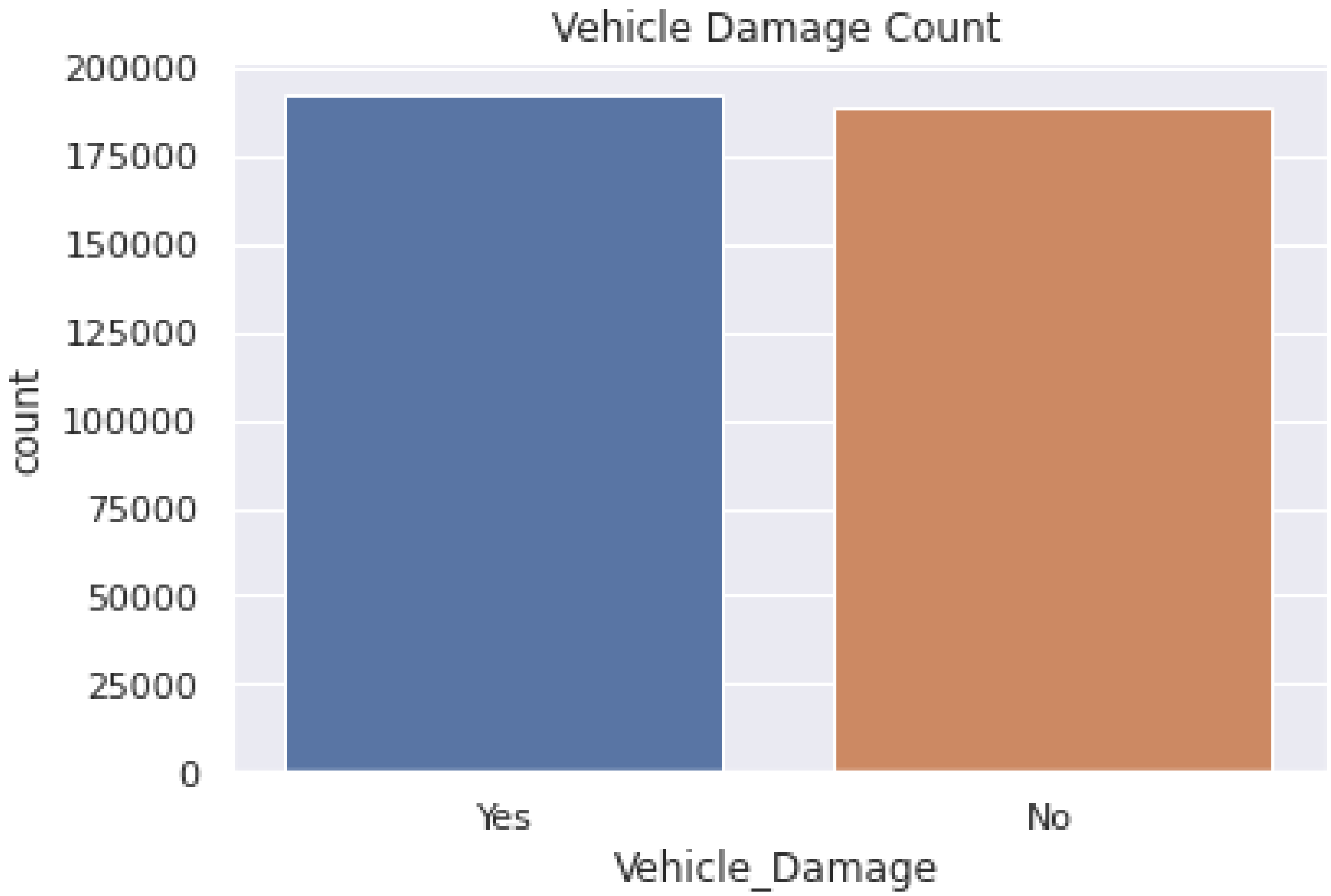


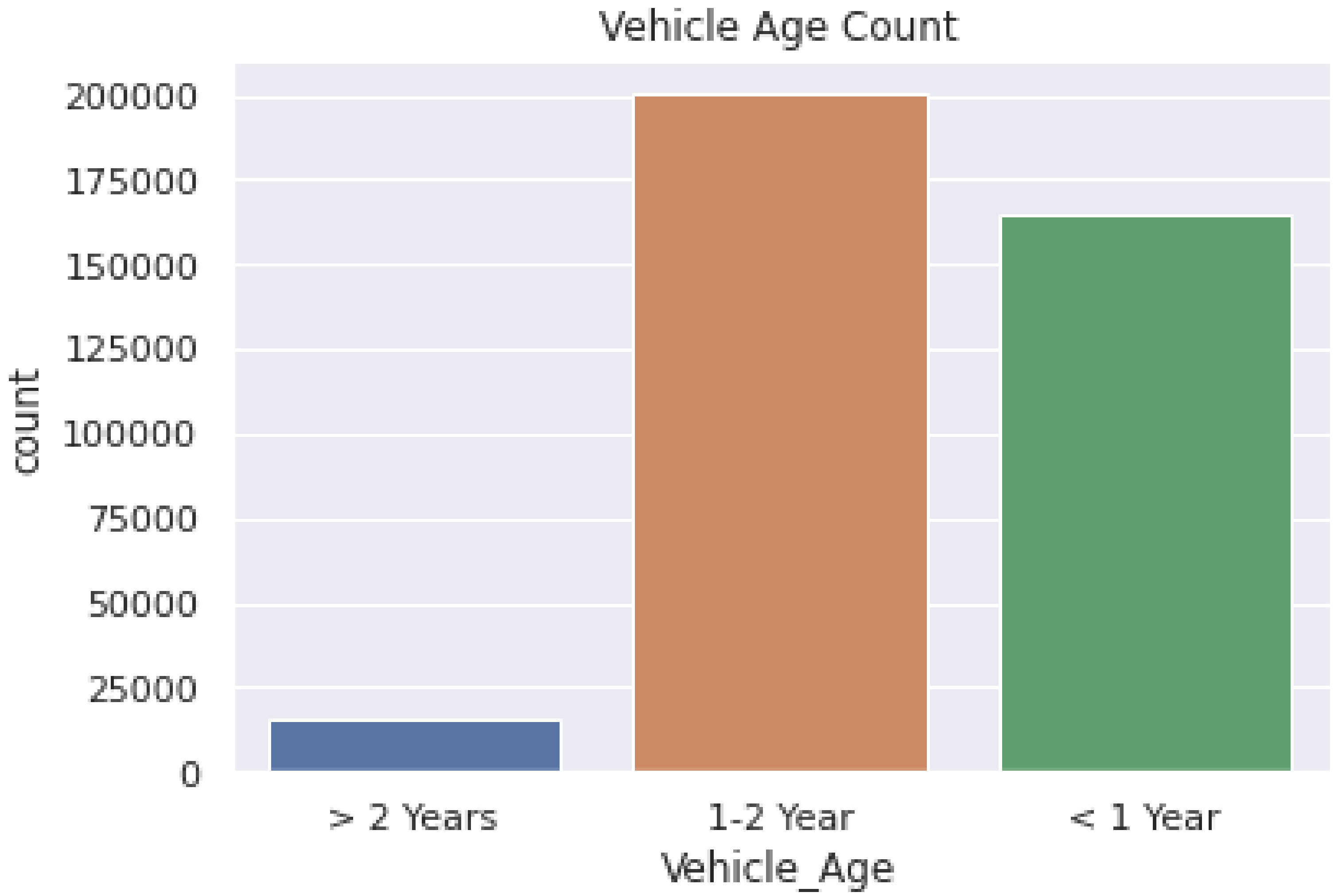


Data distribution

The dataset is **balanced** regarding the number of customers previously **having vehicle damage**

The dataset is **biased** towards the customers owning a new car (**fewer than 2 years**) --> not generalize for customers owing a car **more than 2 years**

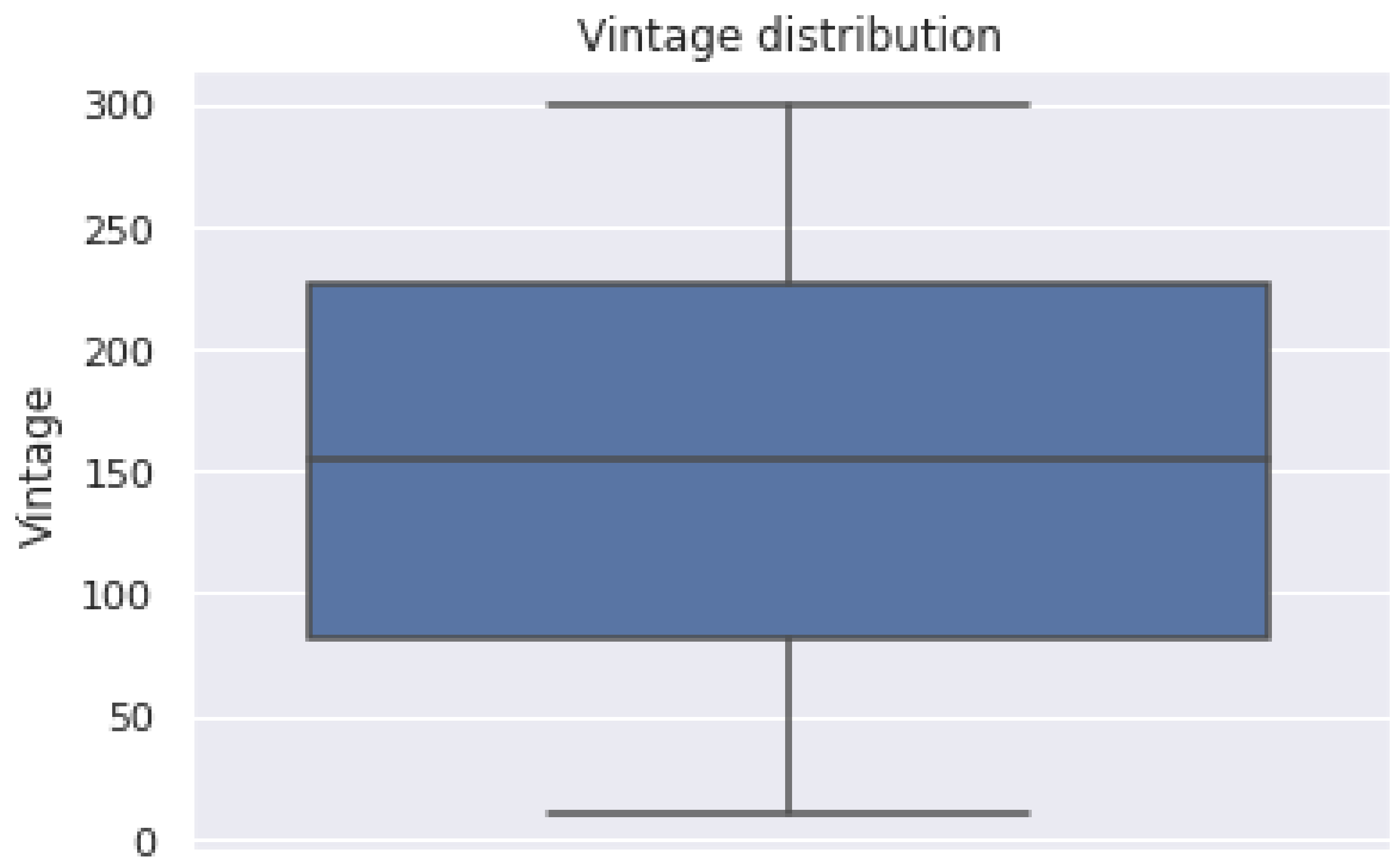


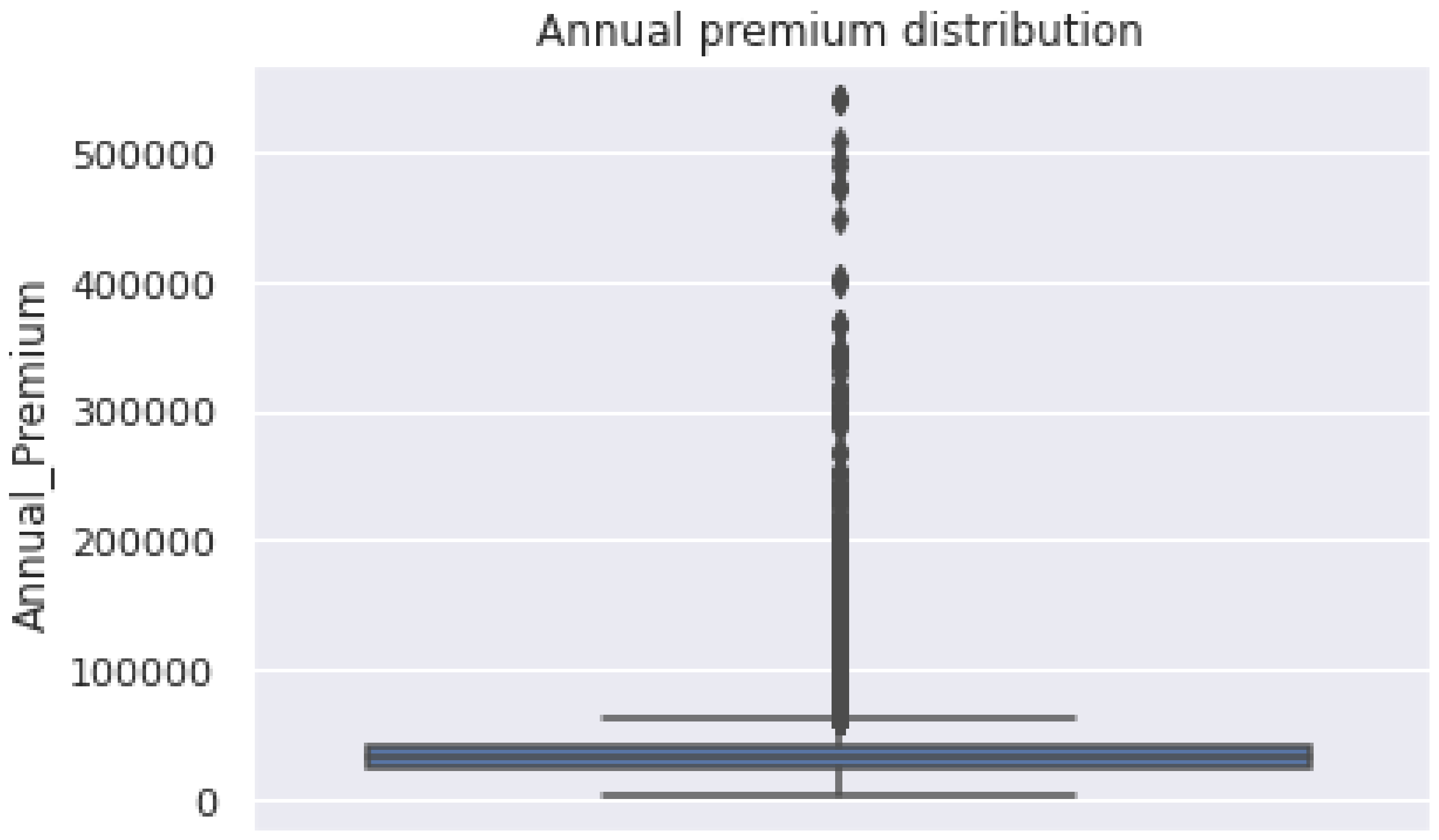


Data distribution

The feature vintage appears to have no outliers with a **wide range of associated days**

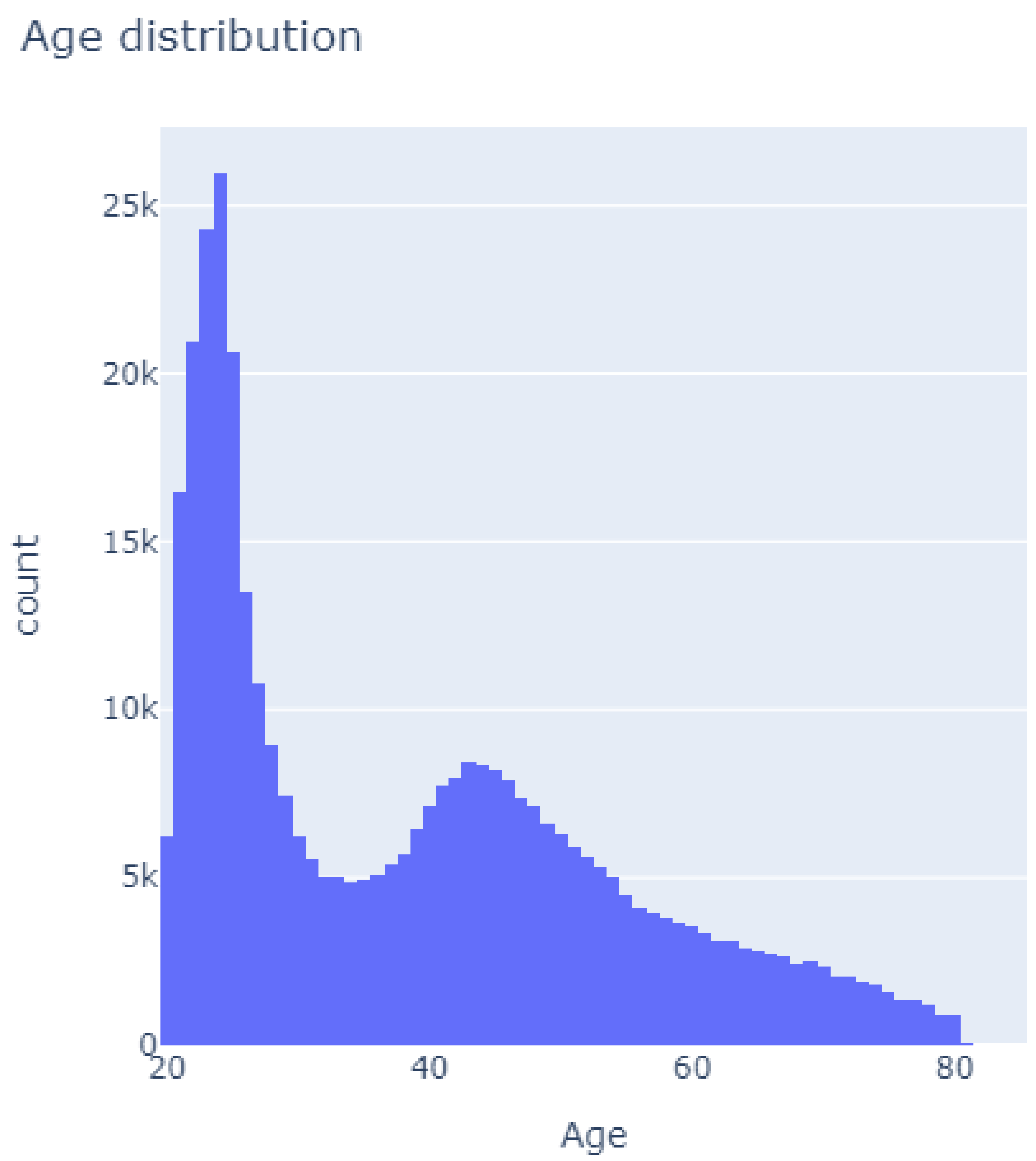
The annual premium has outliers and most of the customers pay **less than 100000 ruppee**





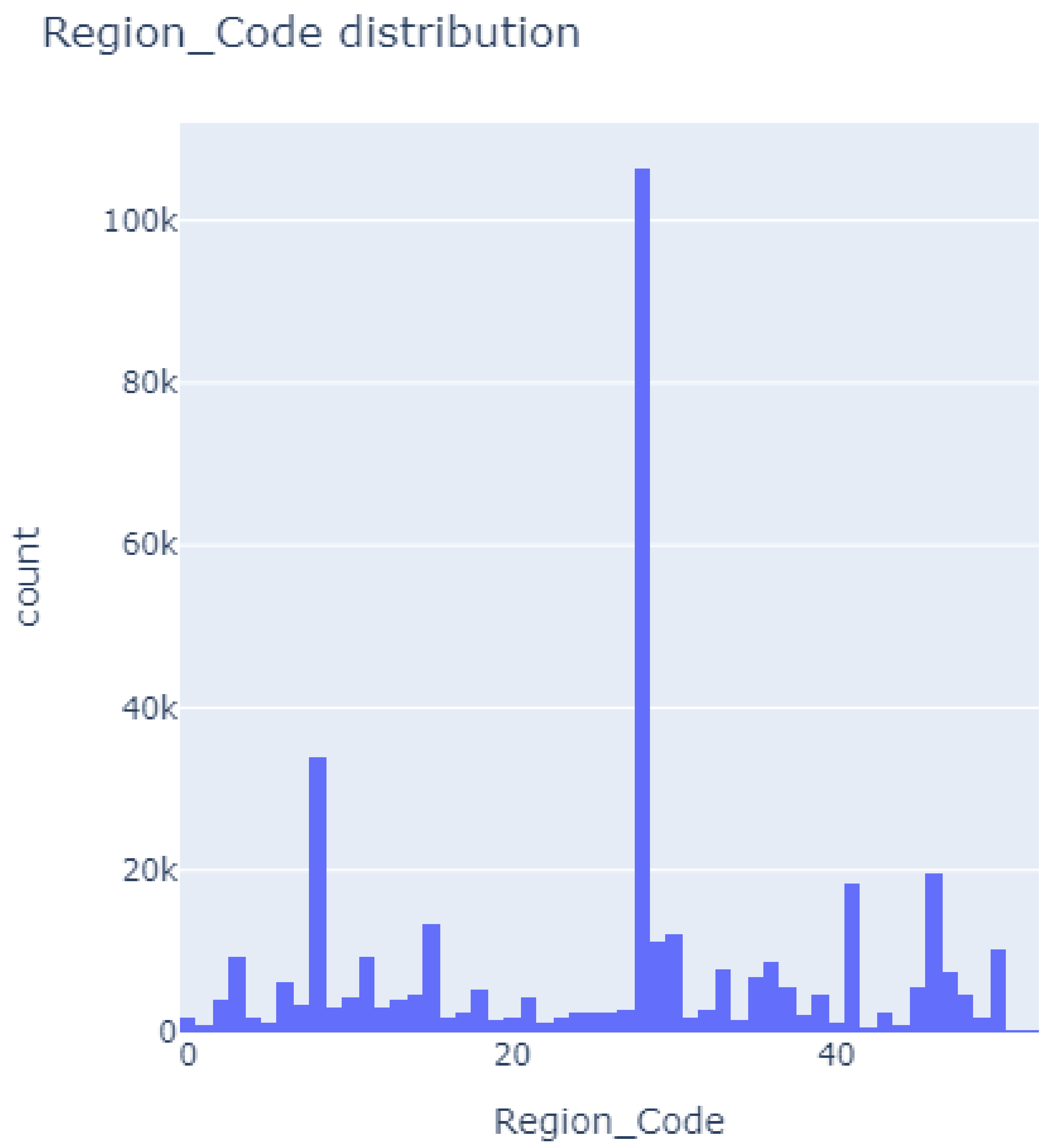
Data distribution

The dataset contains mostly the customers **at a young age** (younger than 35 years old)



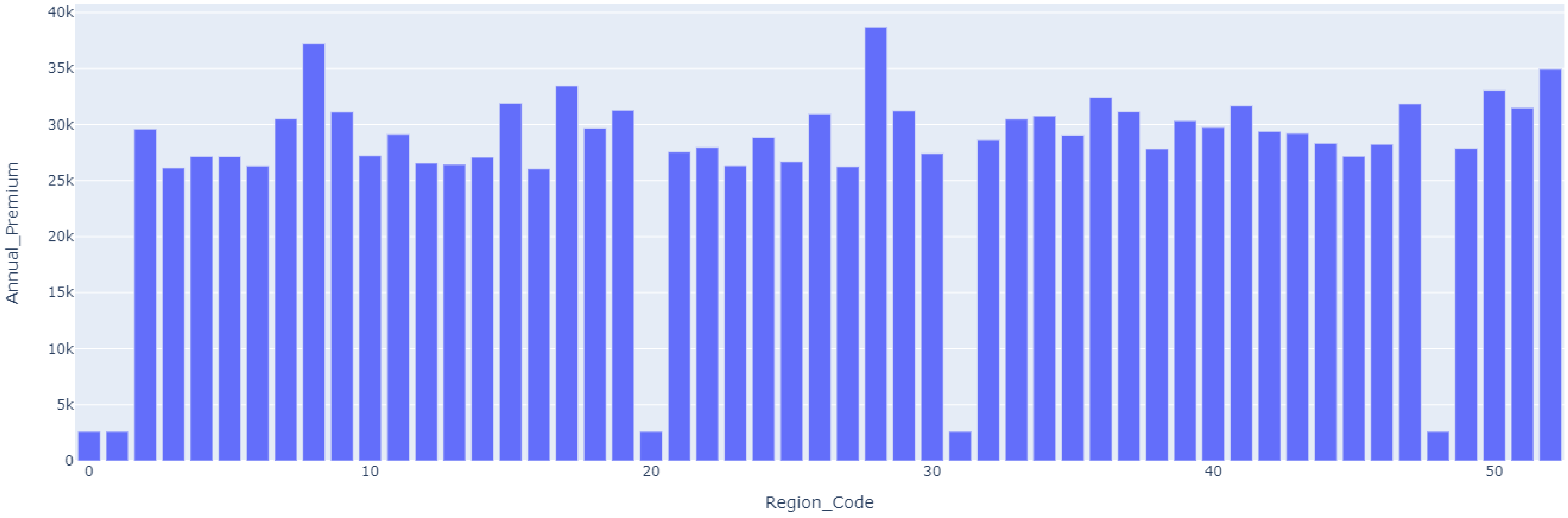
Data distribution

The majority of the customers come from region code 8 and 28



Data distribution

Region code **8 and 28** also have the ***highest median/average annual premium***



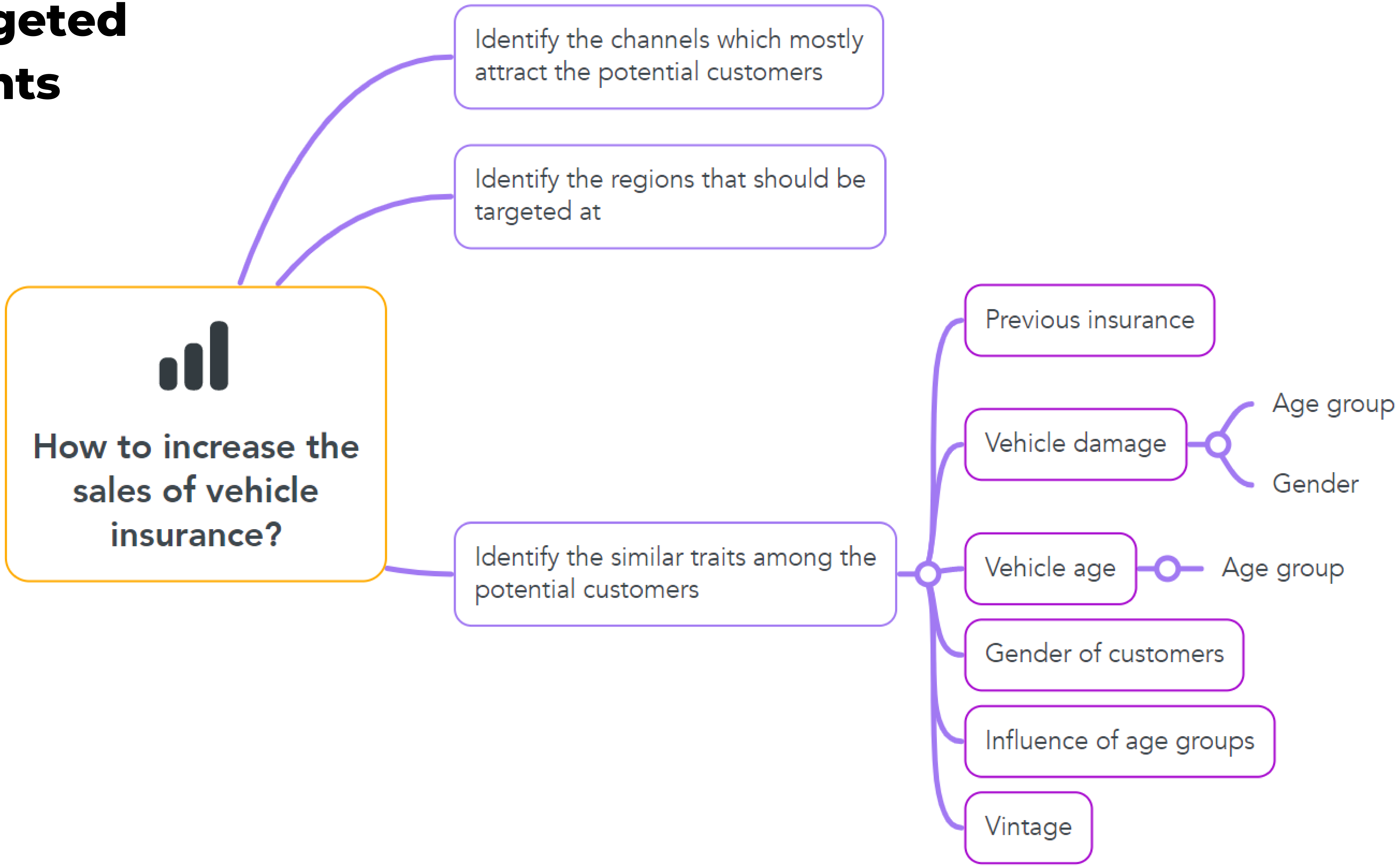
Key Findings:

- The dataset is highly **imbalanced**, biased towards the **uninterested responses**
- Nearly all the customers in the dataset **have a driving license**
- Majority of the customers were **not previously insured** for vehicle damage
- All the customers seem to be the new vehicle owner with **less than 2 years** of ownership
- The customers seem to **pay very low** for **annual premium**
- The customers come primarily from the **young age group** and the **region code 8, 28**

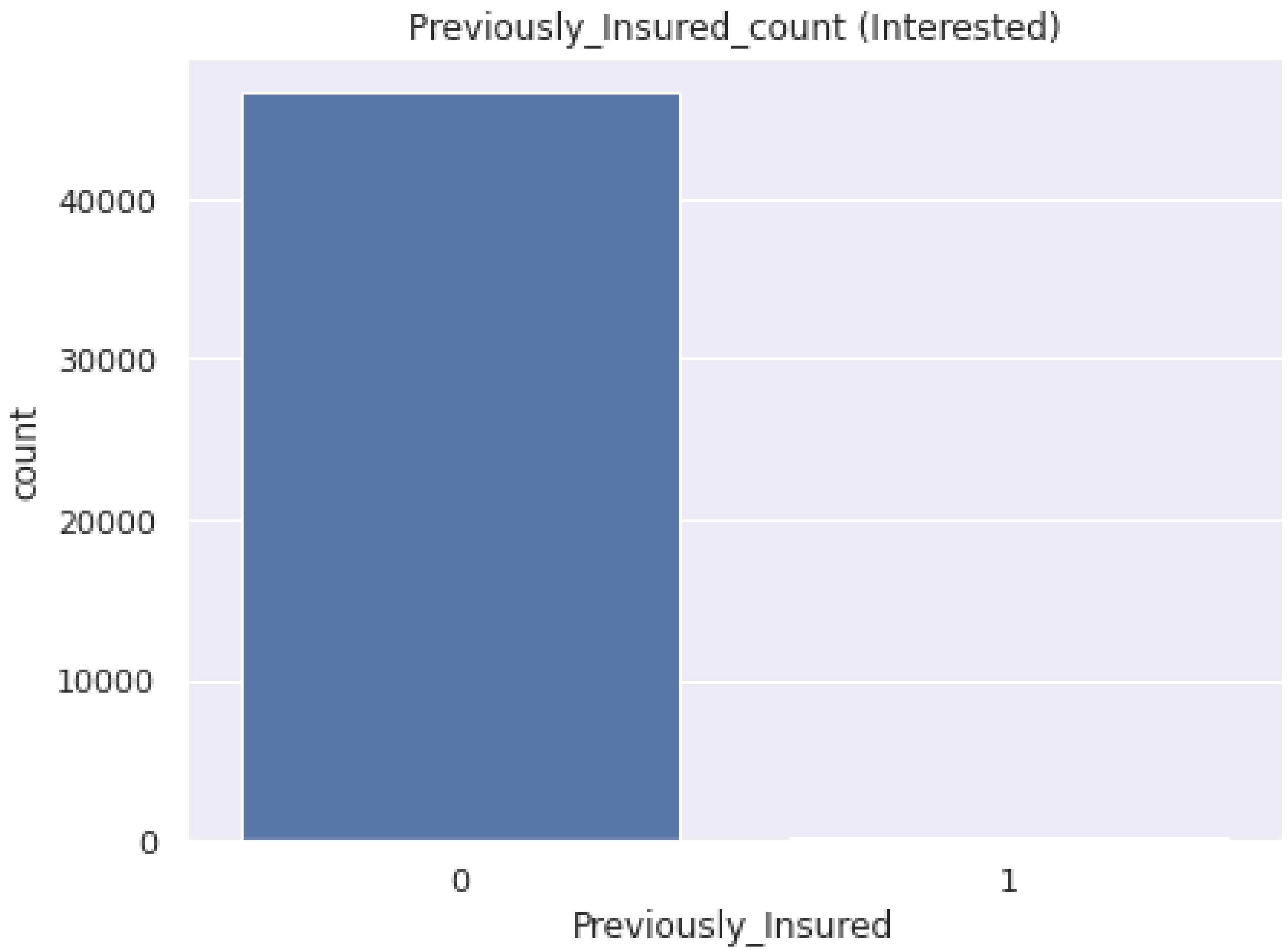
Exploratory Data Analysis

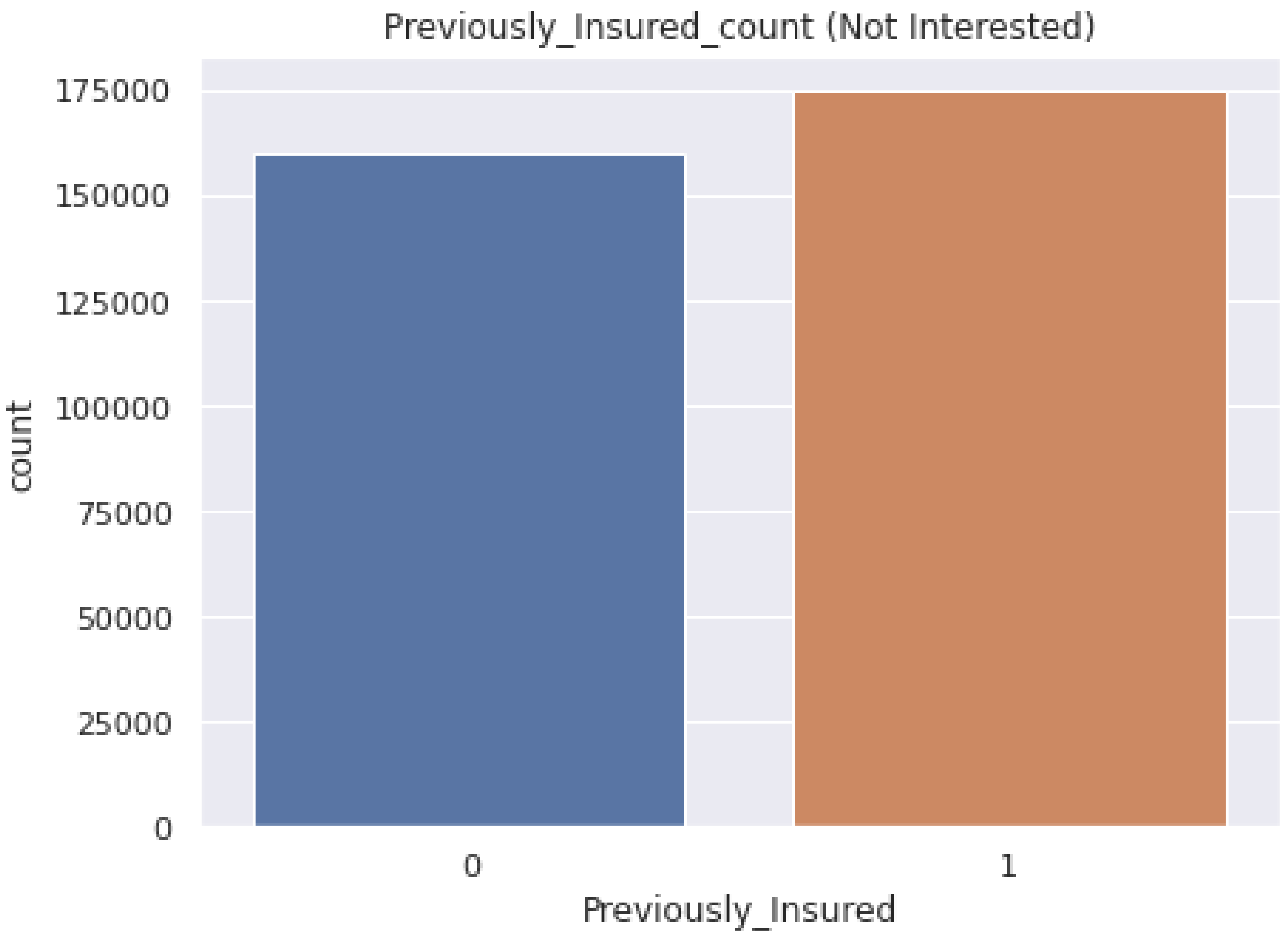
How to increase conversion rate for vehicle insurance?

Targeted points

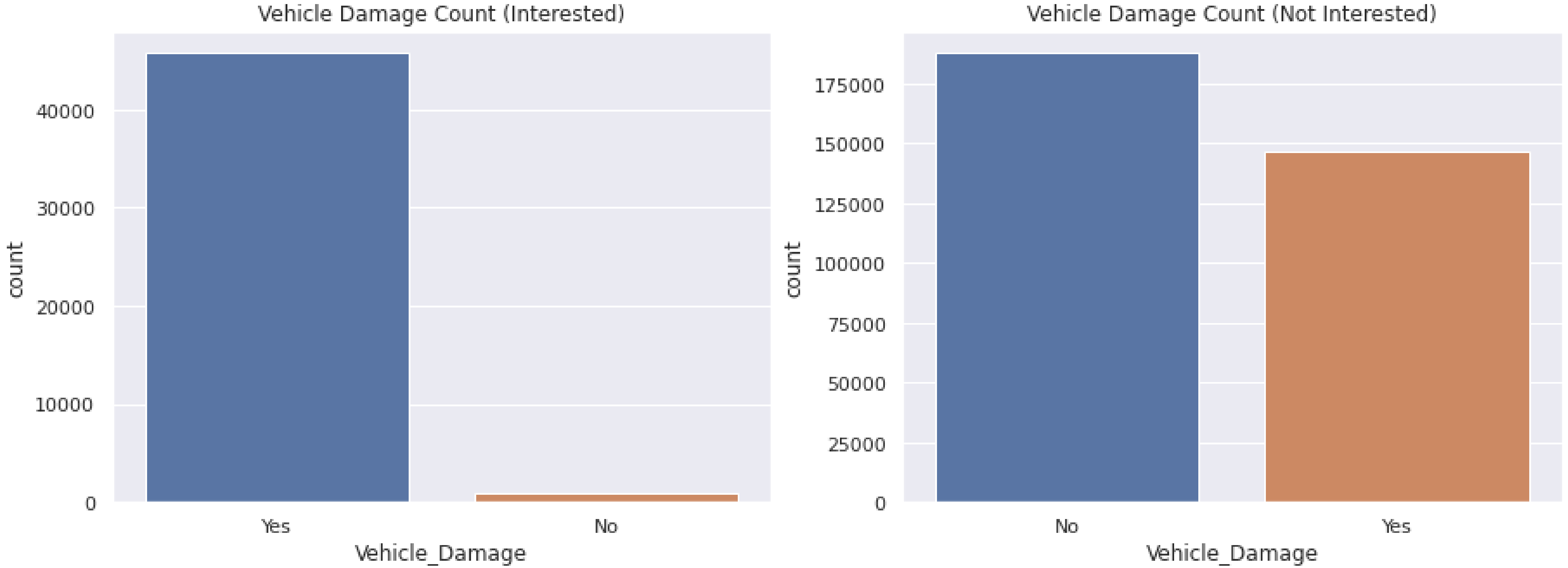


Attention to the previously insured customers?

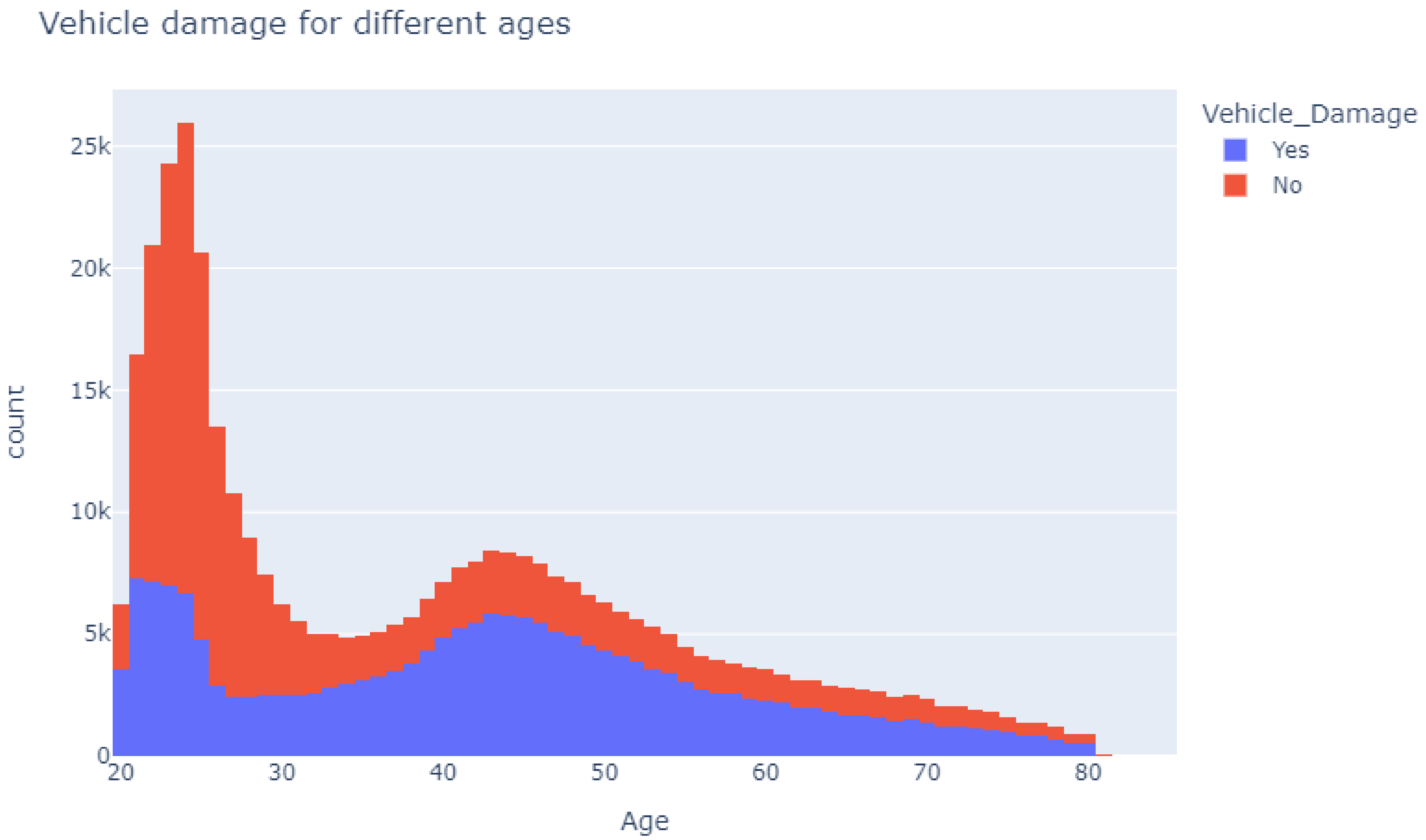




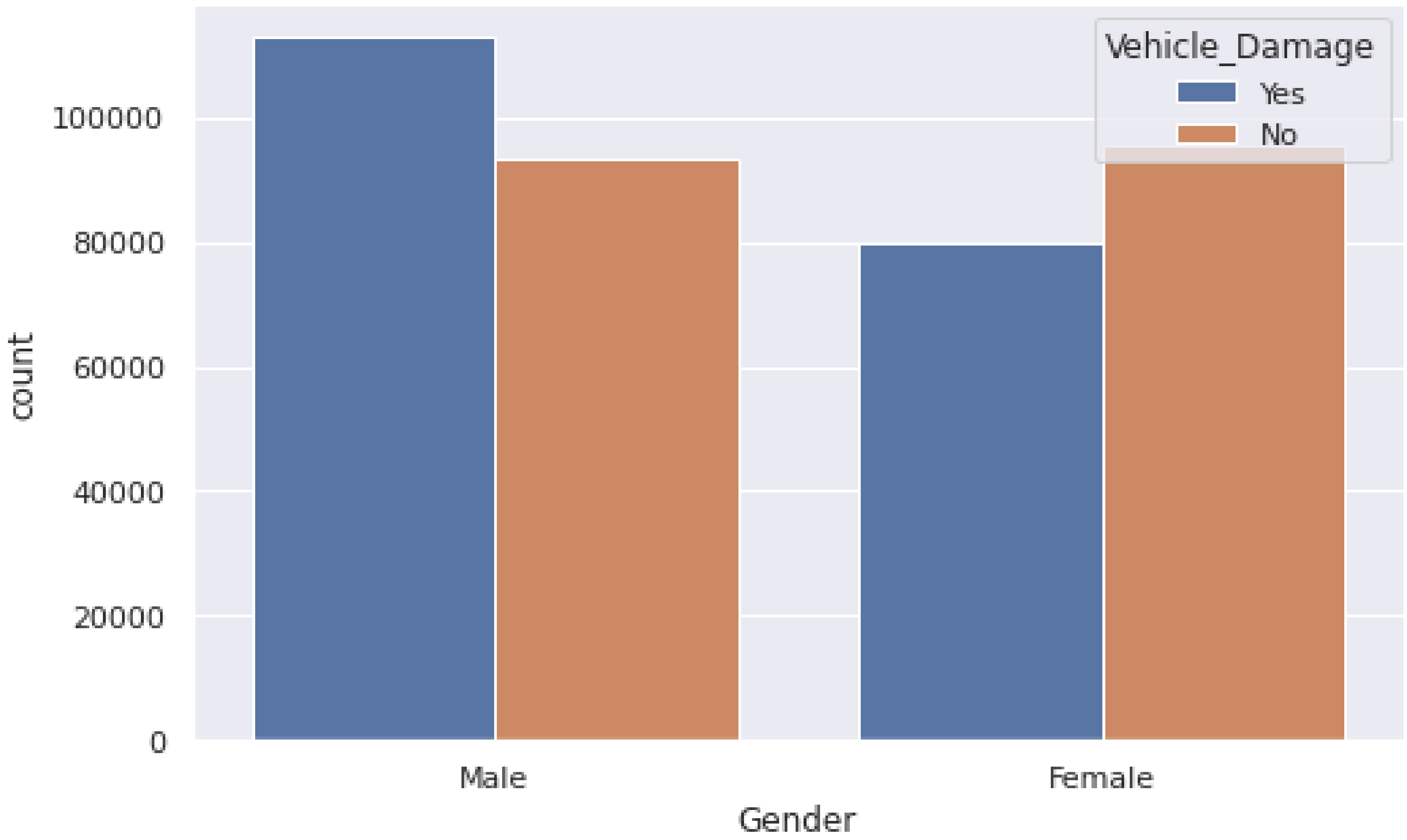
Attention to the customers previously having vehicle damage?



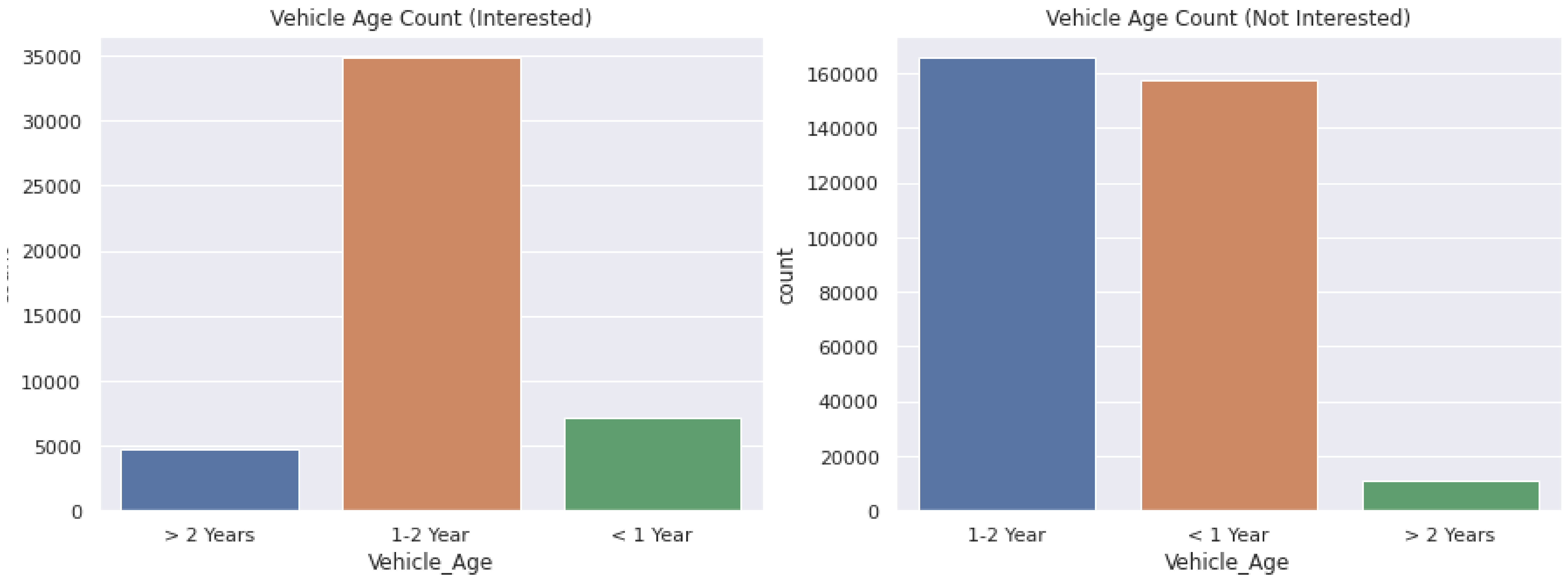
Who tends to be involved in vehicle damage and at which age?



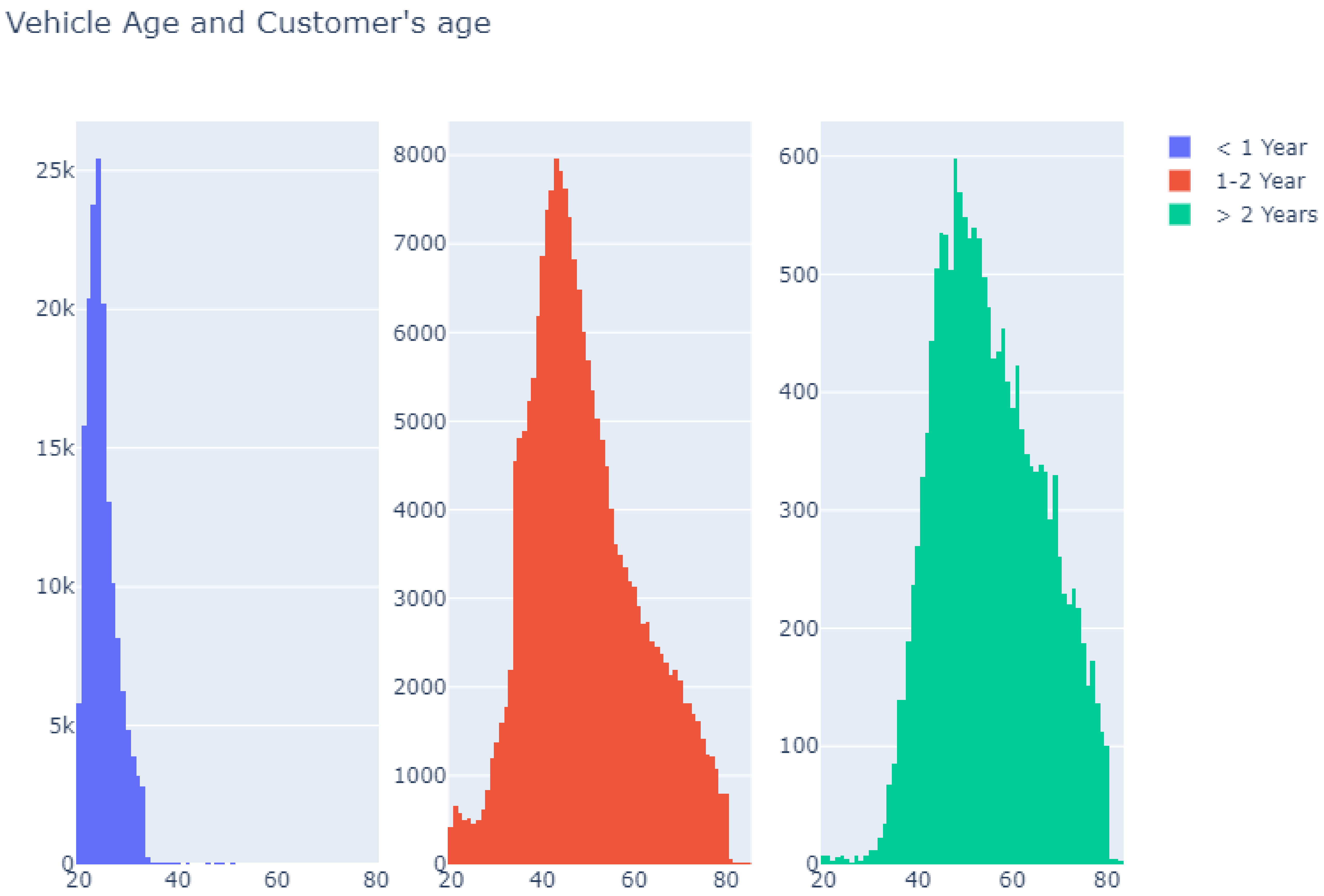
Does gender influence the likelihood of vehicle damafe?



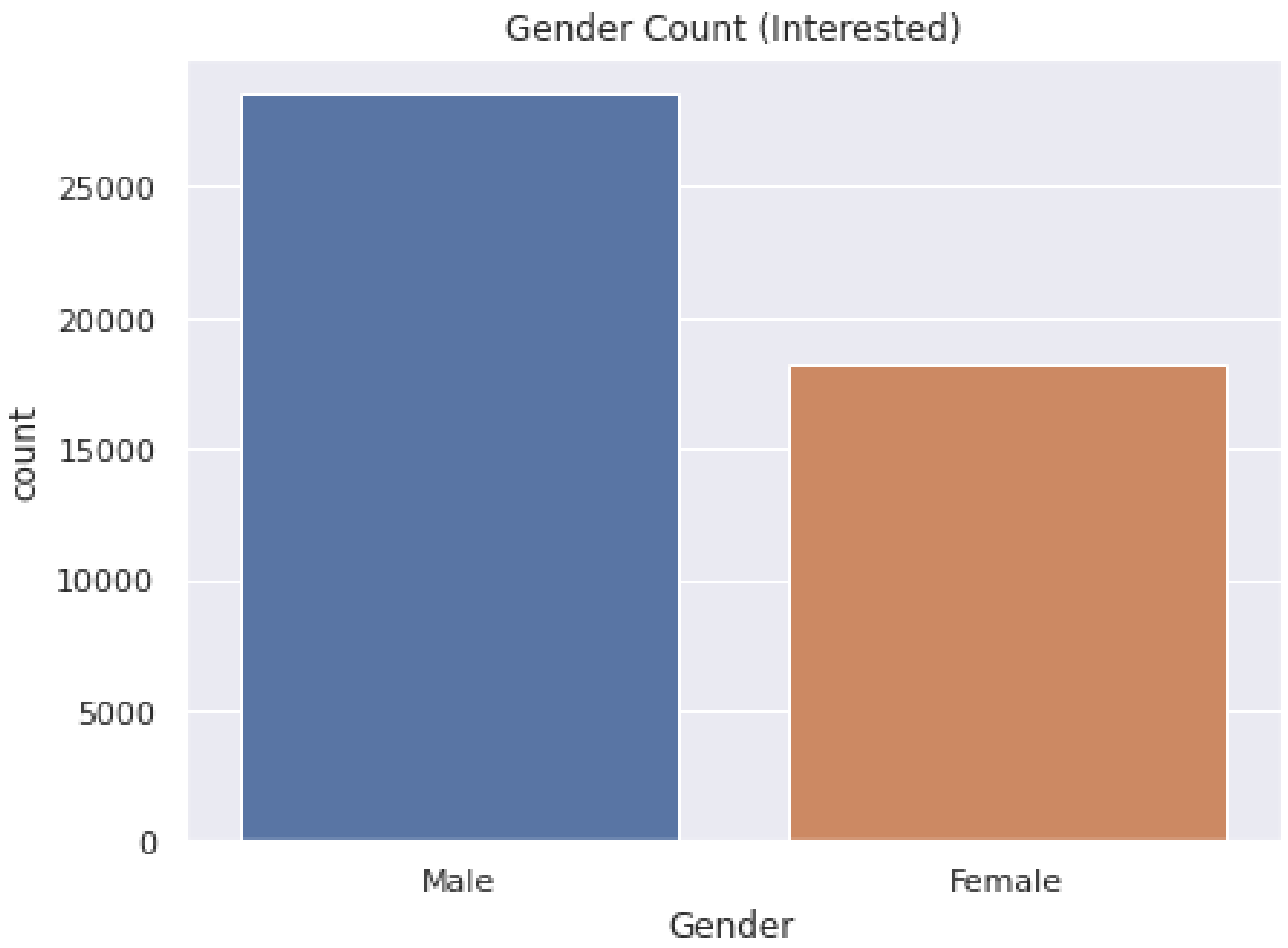
Does vehicle age influence the decision of the customers to buy vehicle insurance?

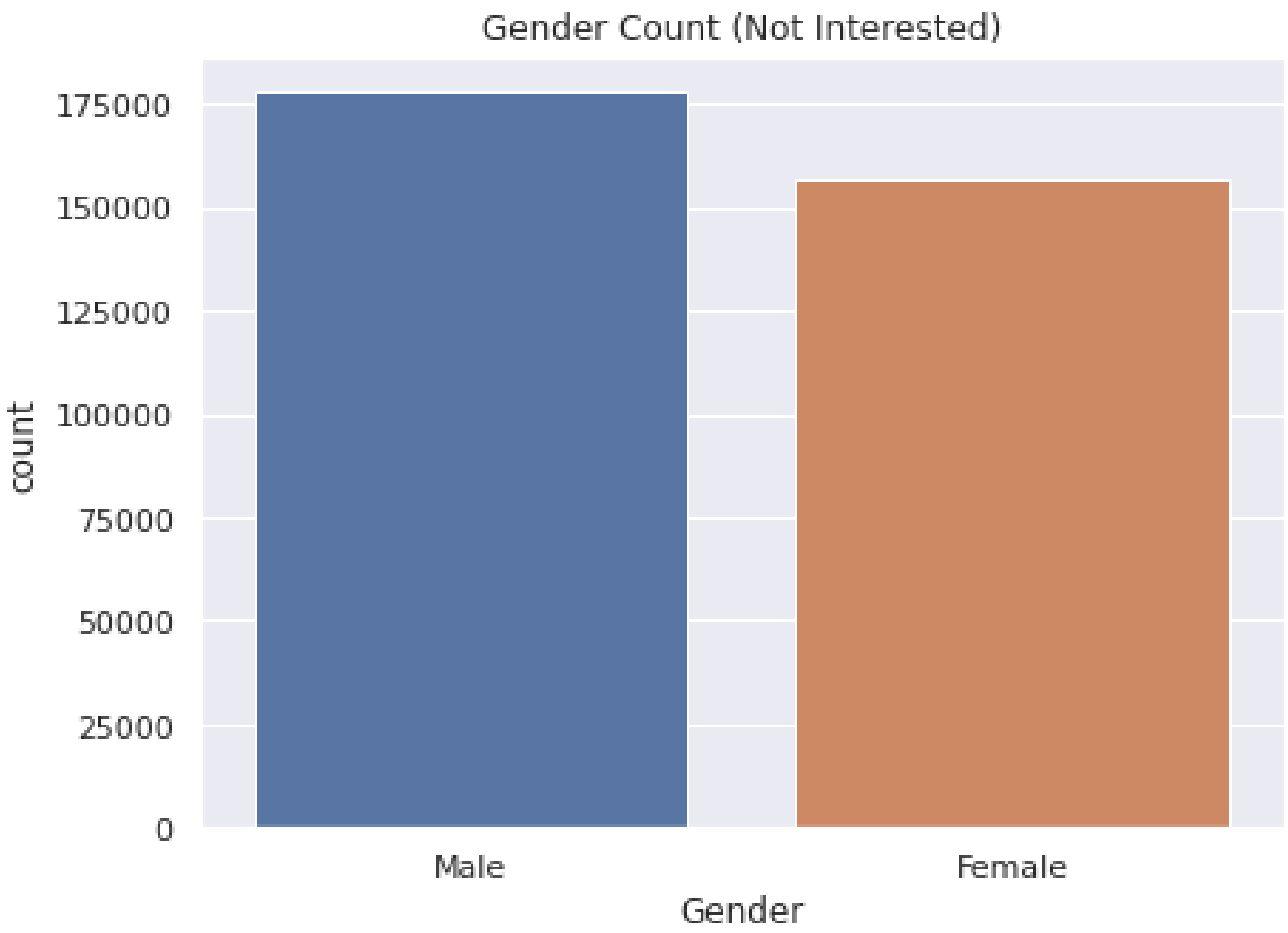


Does age correlate with the vehicle age among the customers?

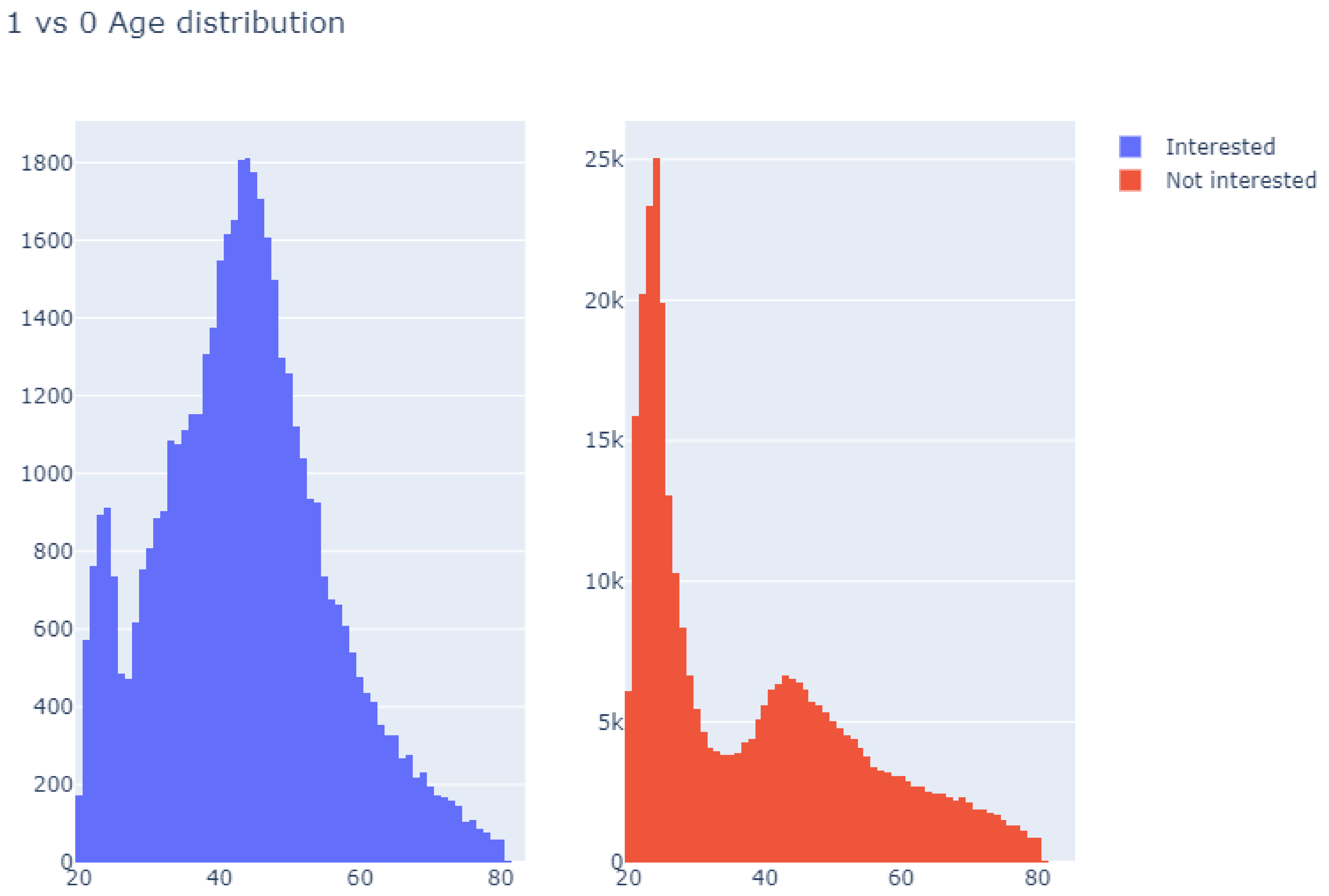


Does gender influence the decision to buy vehicle insurance?

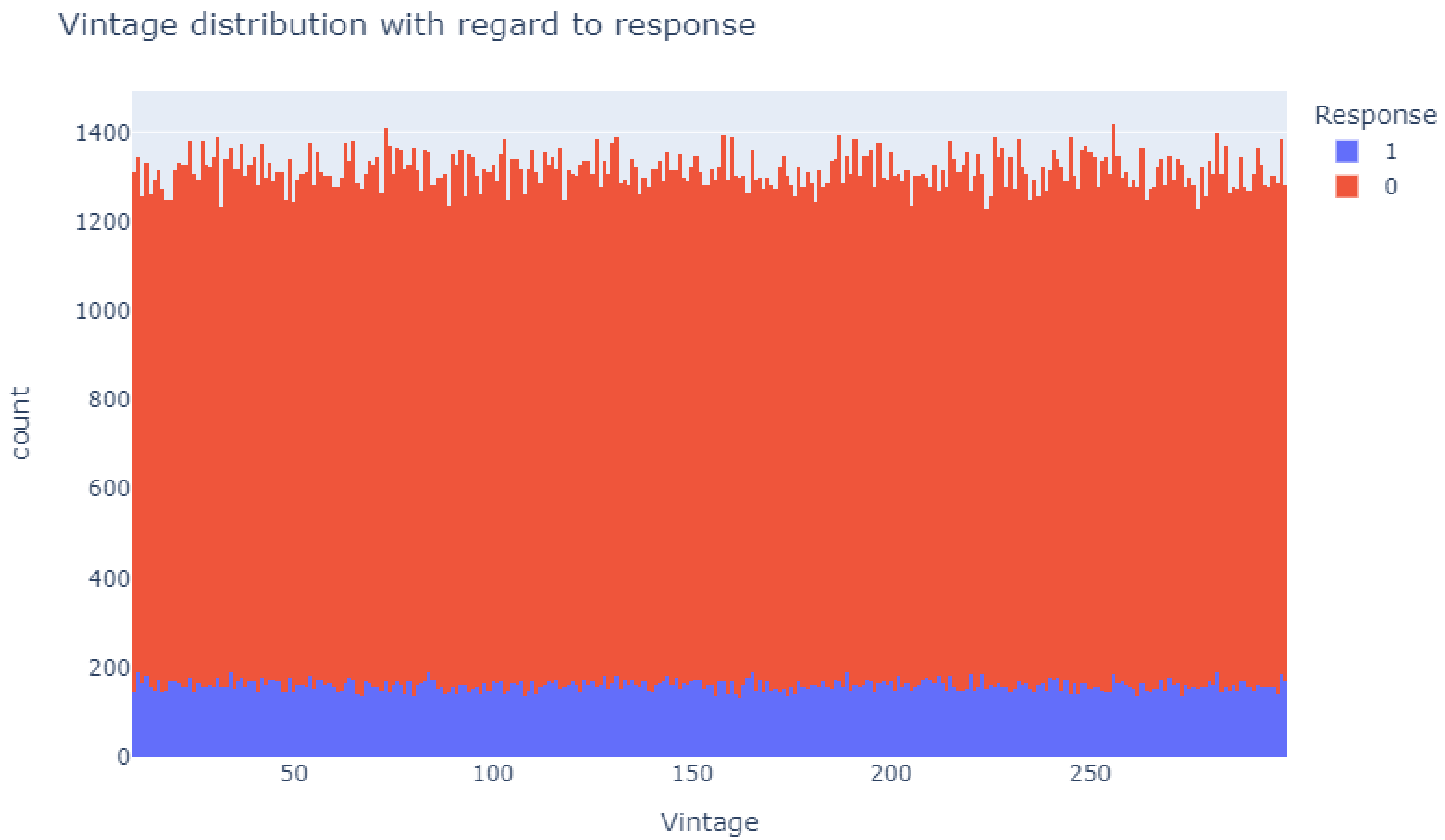




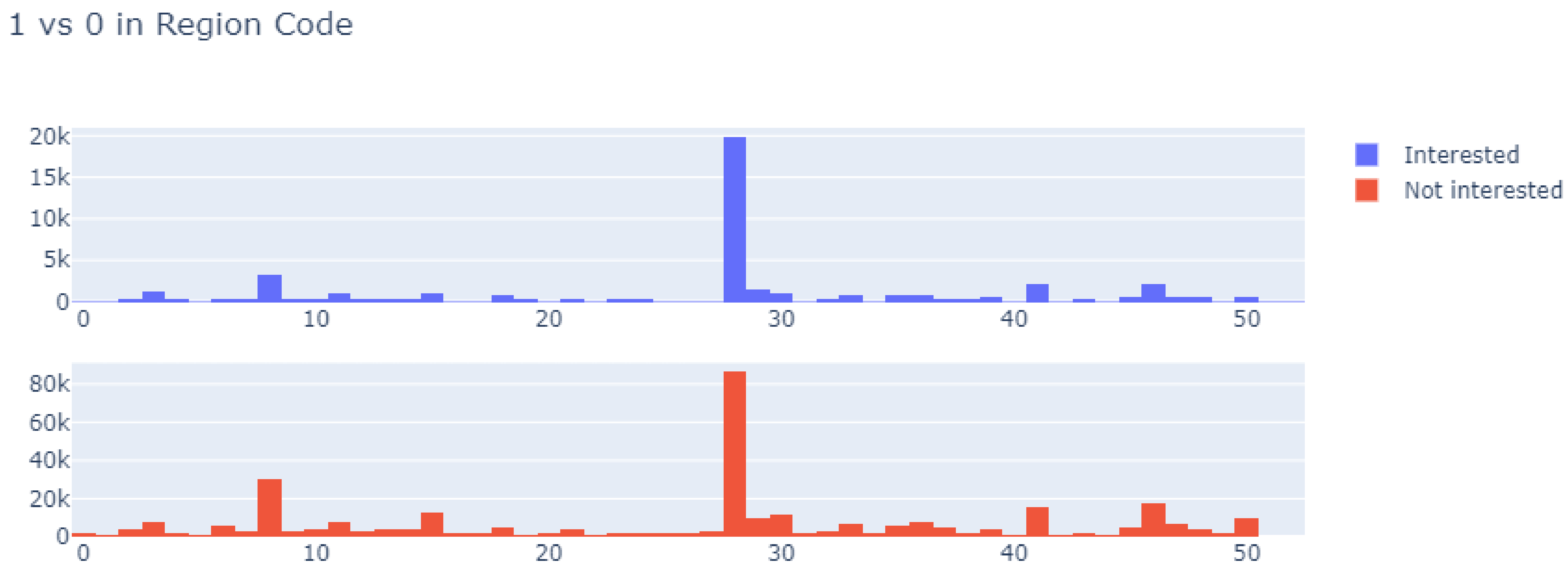
Do different age group have different pattern regarding the decision to buy vehicle insurance?



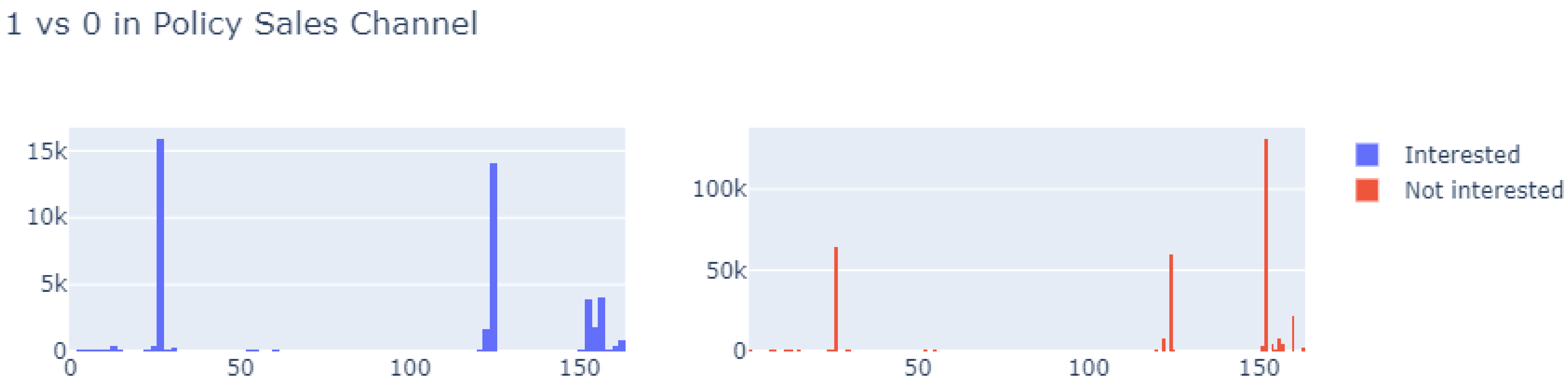
Does longer vintage correlate with higher likelihood to buy vehicle insurance?



Which region codes would the campaign for increasing sales of vehicle insurance be targeted?



Which channel would be the optimal one for increasing the sales of vehicle insurance?



Key Findings:

- Traits of potential customers:

- Previously **not granted** insurance policy and had **vehicle damage**

- **Male** customers in the **middle-aged** group with vehicle age ranging from **1-2 years**

- Customers coming from **region code 28**

Exploratory Data Analysis

Data Preprocessing

<u>Methods:</u>

1. Convert to **numerical value**

- *region code*

2. **One-hot** encode the categorical features

- *Vehicle Age, Vehicle_Damage, Gender*

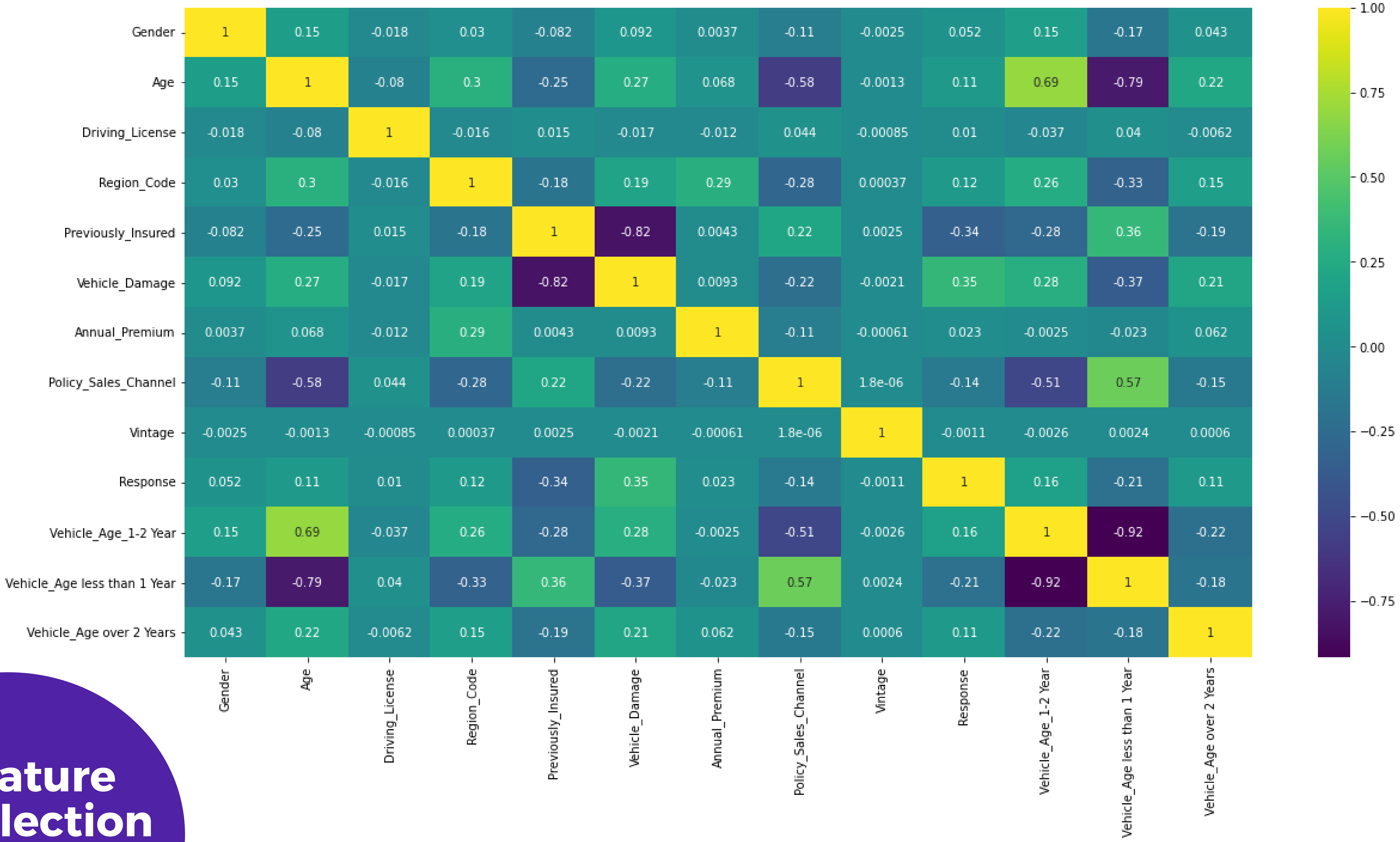
3. Apply **min-max scaler** to the continuous values

- *Age, vintage, Annual_Premium*

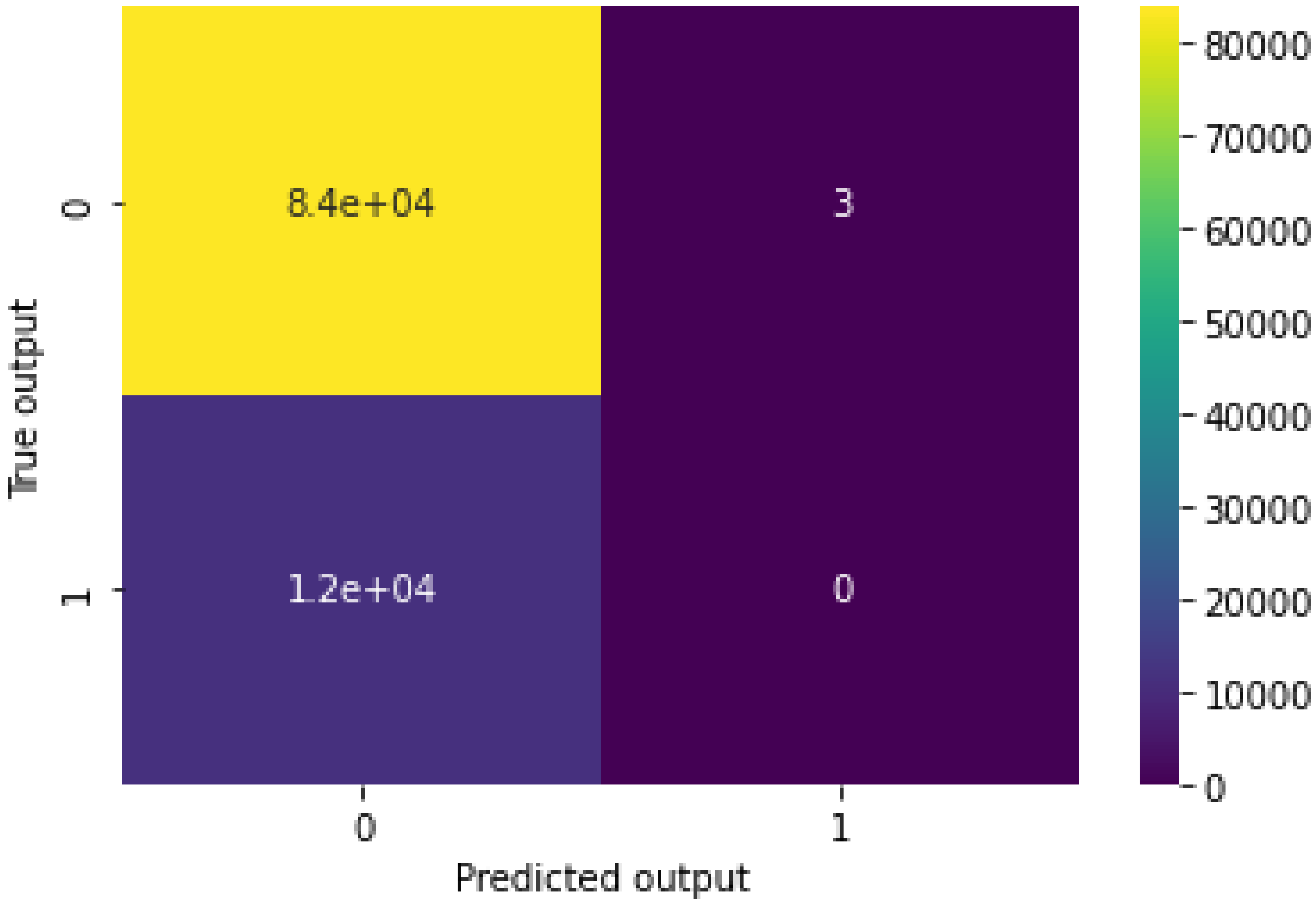
4. Drop insignificant columns

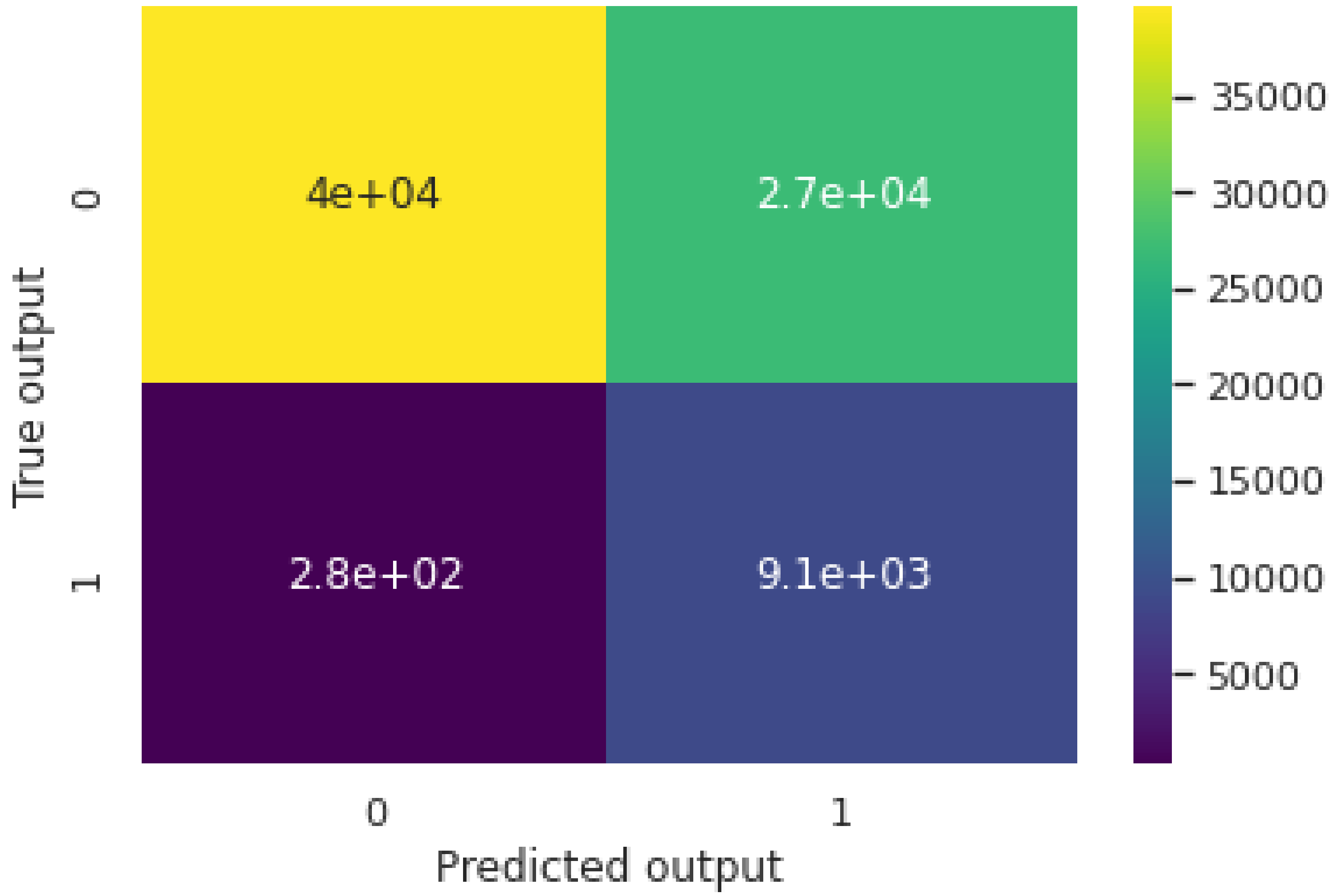
- id, Policy_Sales_Channel, Driving_License

Feature Selection



Build Model

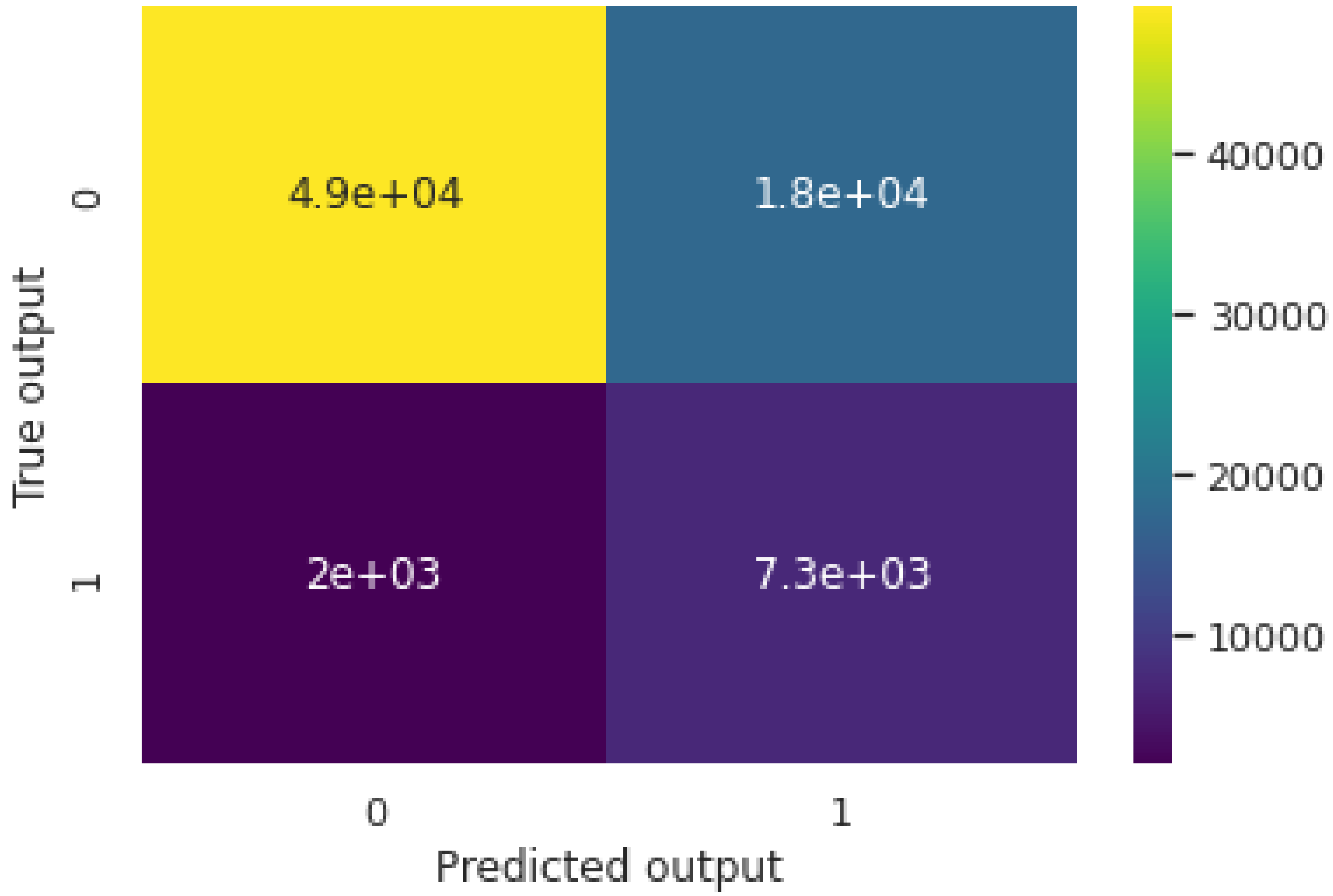

8.4e+04
3
1.2e+04
0
True output
Predicted output
0
1
80000
70000
60000
50000
40000
30000
20000
10000



Baseline Model

Logistic Regression is chosen as the baseline model

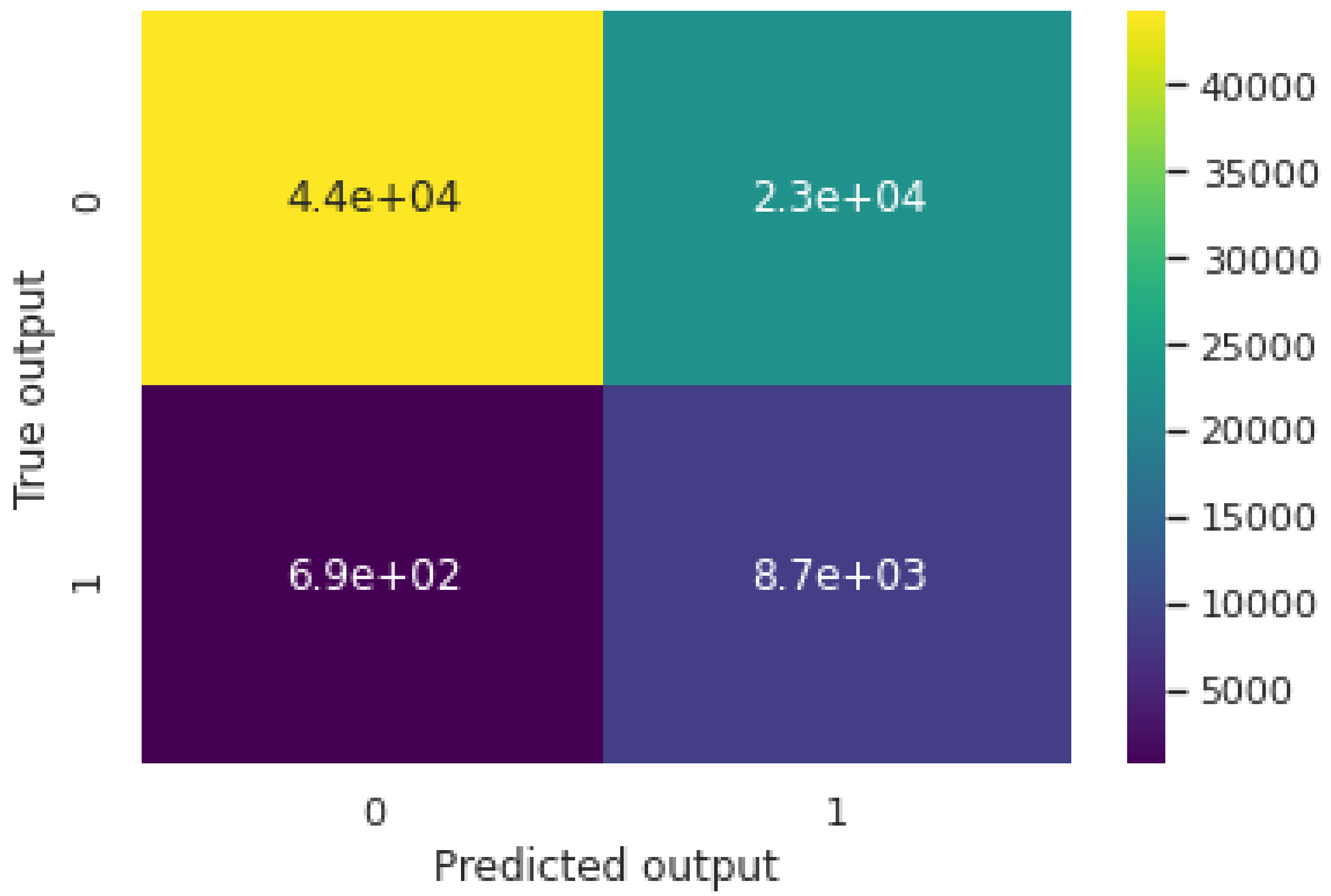
F1 score: 0.397
Recall: 0.97
Precision: 0.249



Ensemble Method

XGBoost Classifier is firstly chosen

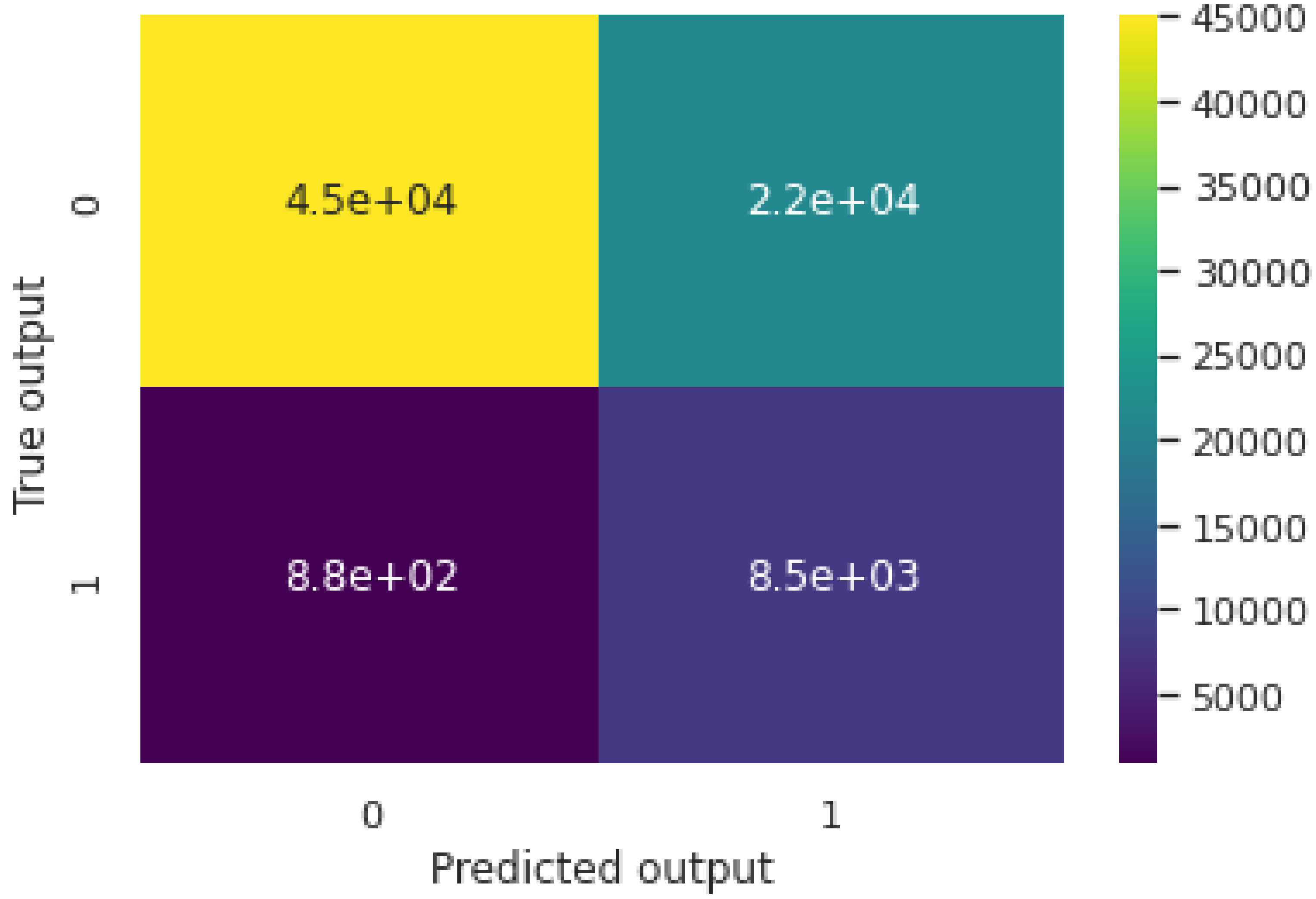
F1 score: 0.427
Recall: 0.785
Precision: 0.293



Ensemble Method

Random Forest Classifier is then chosen

F1 score: 0.427
Recall: 0.926
Precision: 0.28

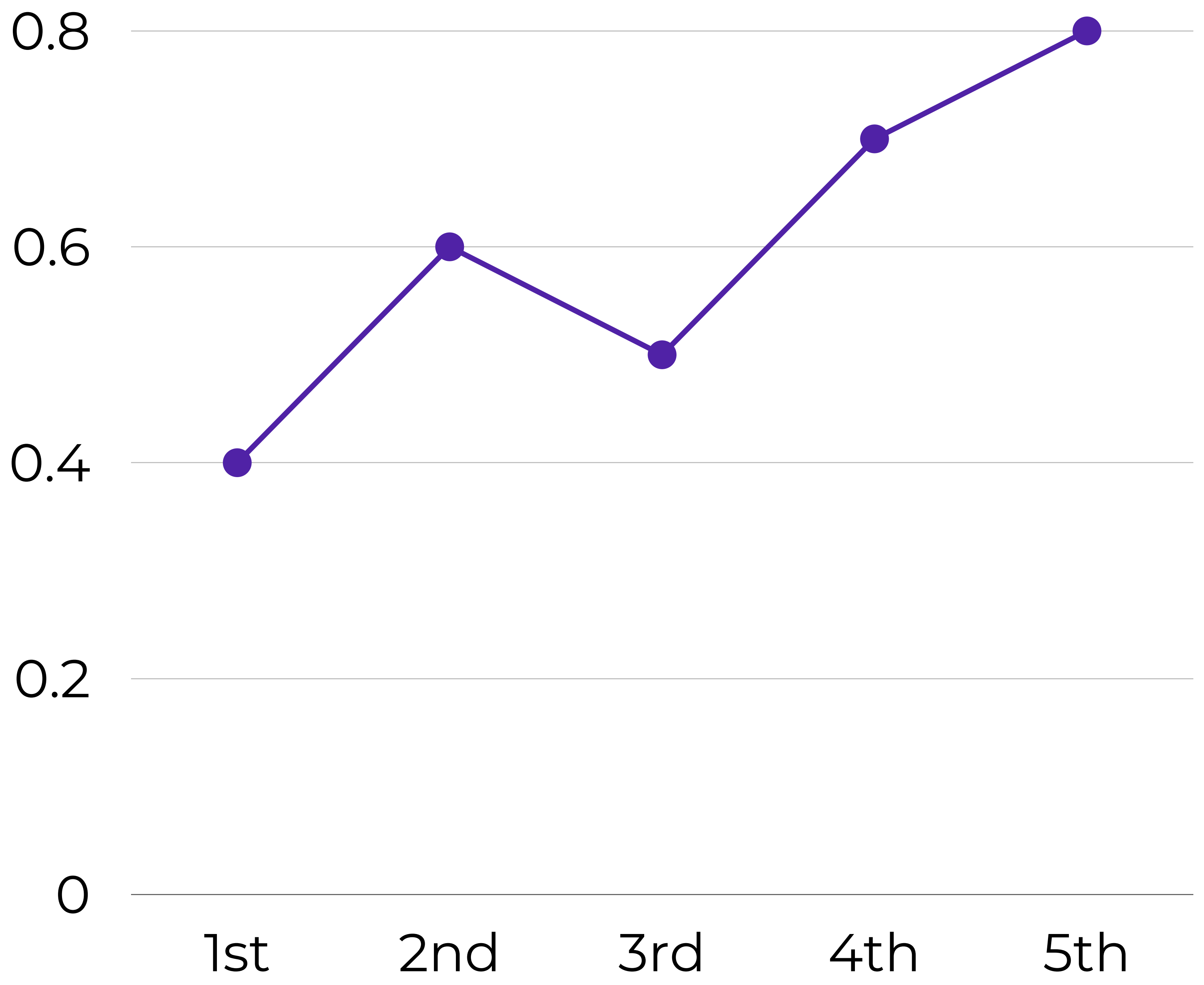


Ensemble Method

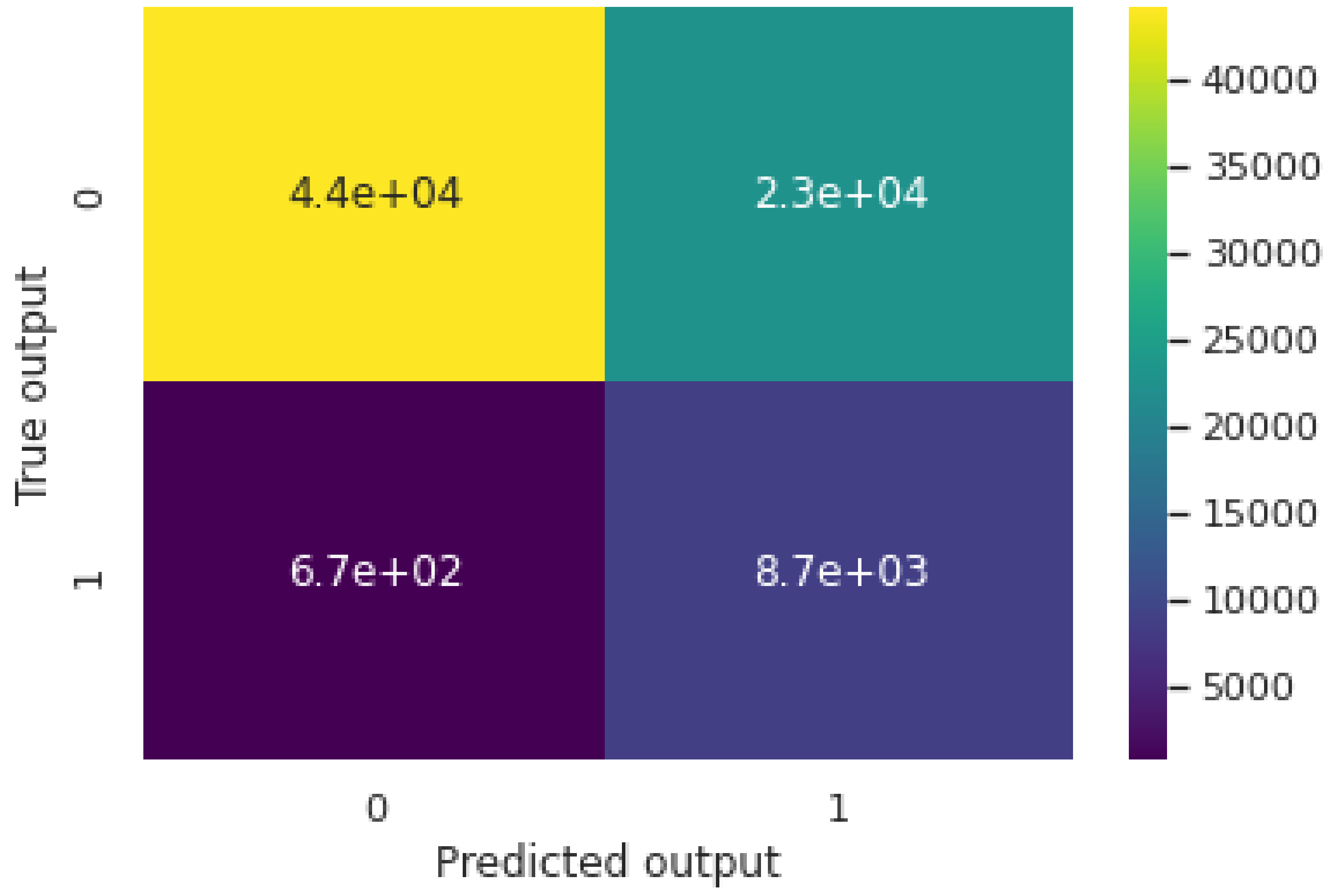
Light Gradient Boosting Classifier is then chosen

F1 score: 0.428
Recall: 0.91
Precision: 0.28

Improvement to the model

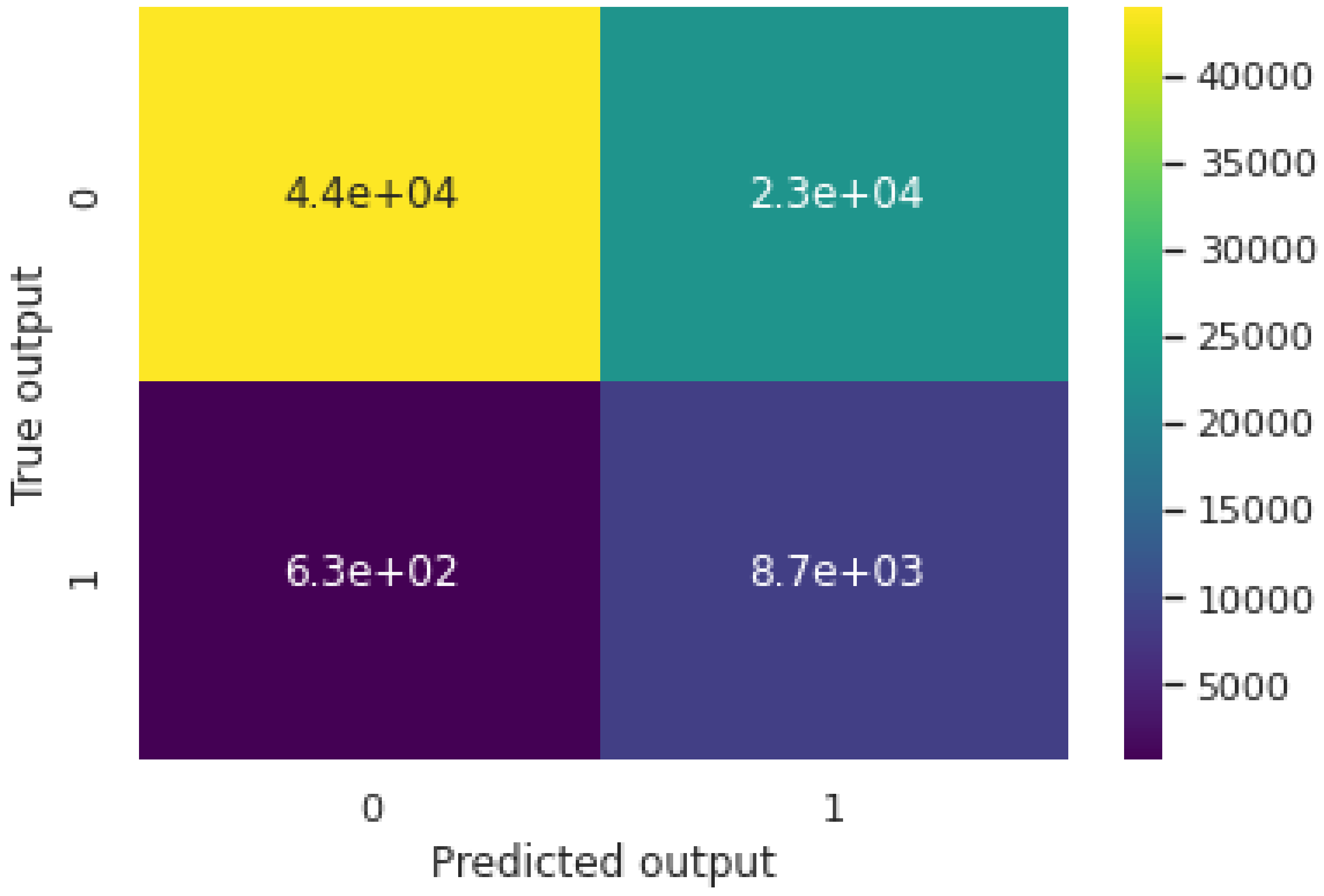


Oversampling (RFC)



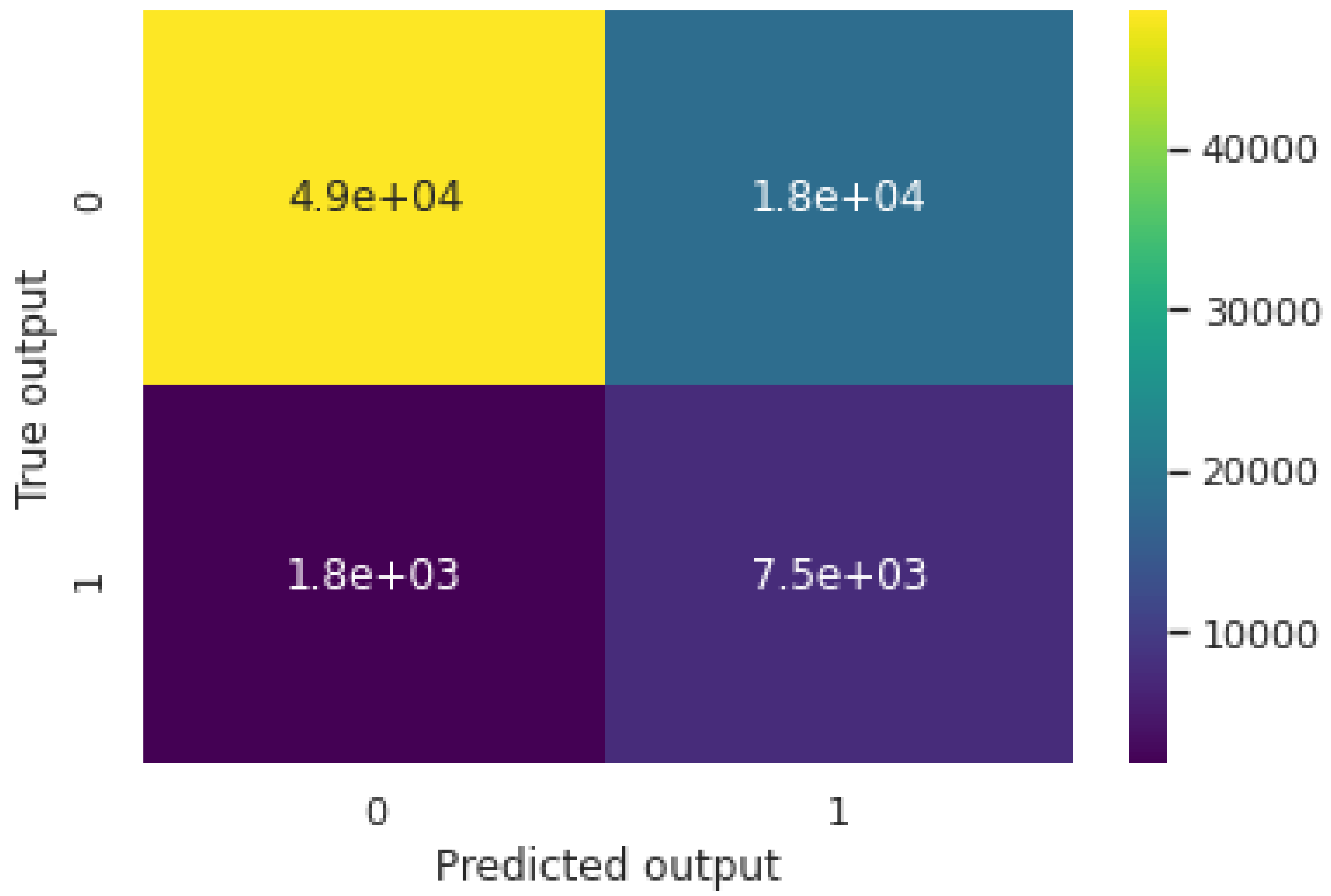
F1 score: 0.426
Recall: 0.93
Precision: 0.276

Undersampling (RFC)



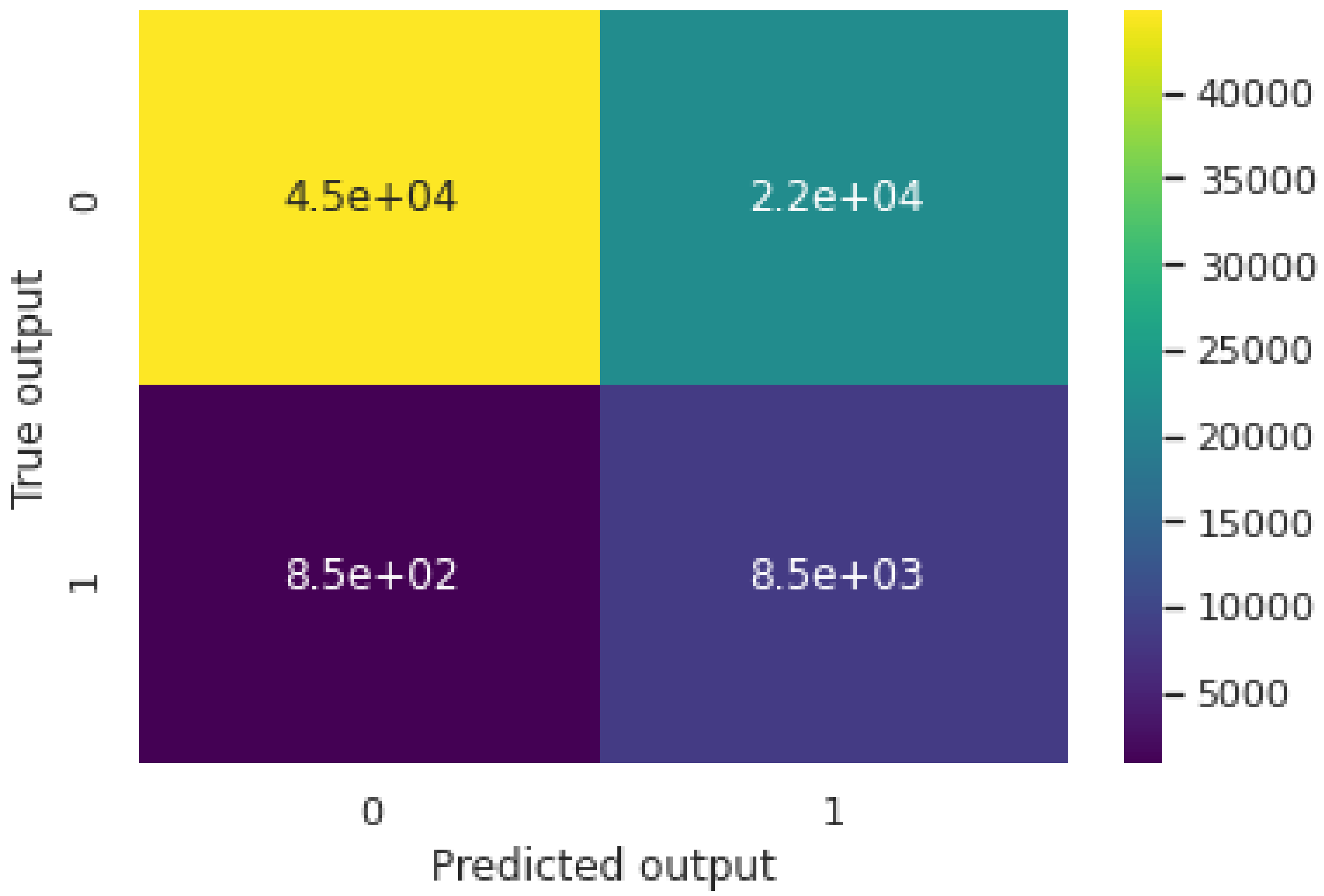
F1 score: 0.425
Recall: 0.932
Precision: 0.275

Oversampling (XGBoost)



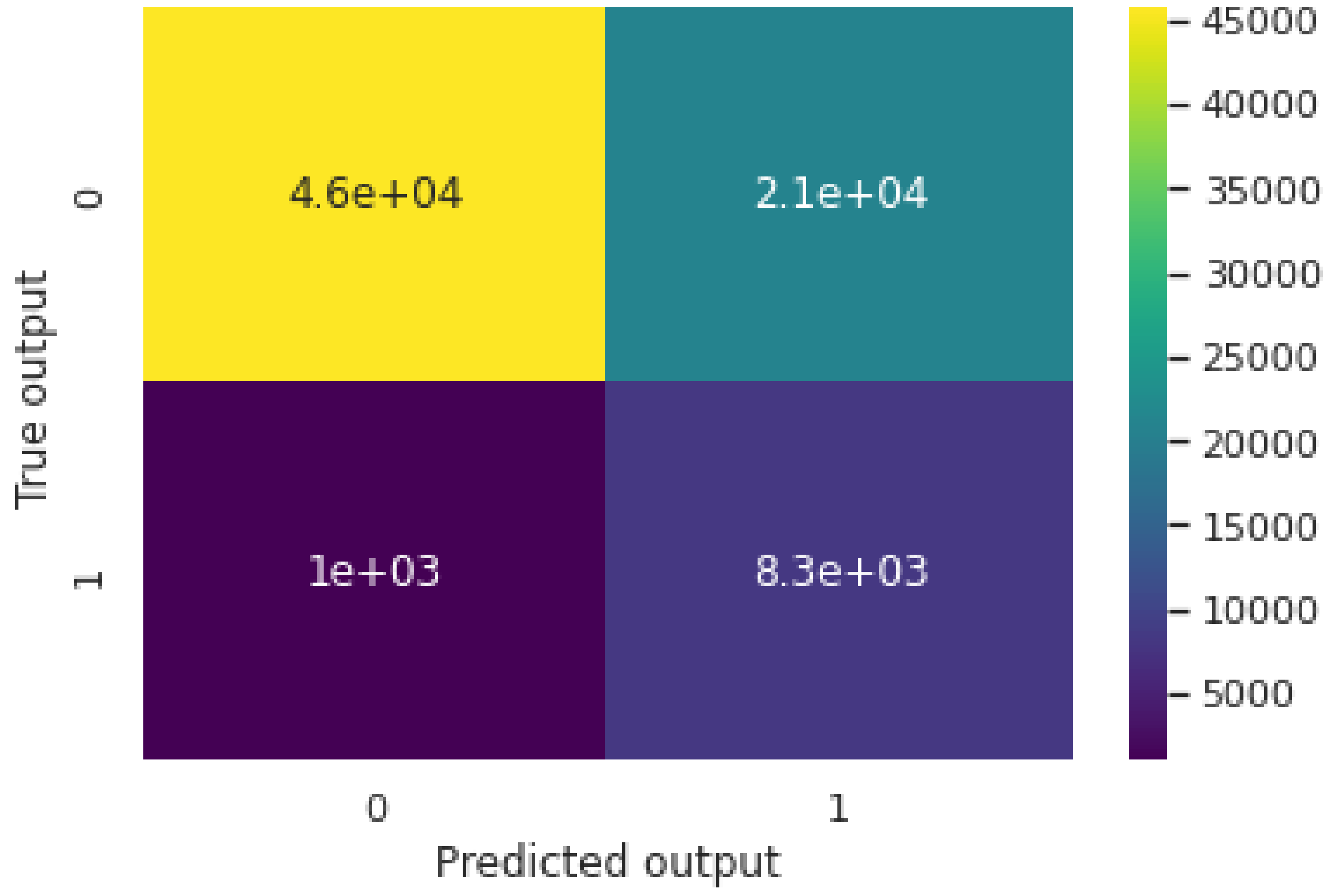
F1 score: 0.428
Recall: 0.907
Precision: 0.29

Undersampling (XGBoost)



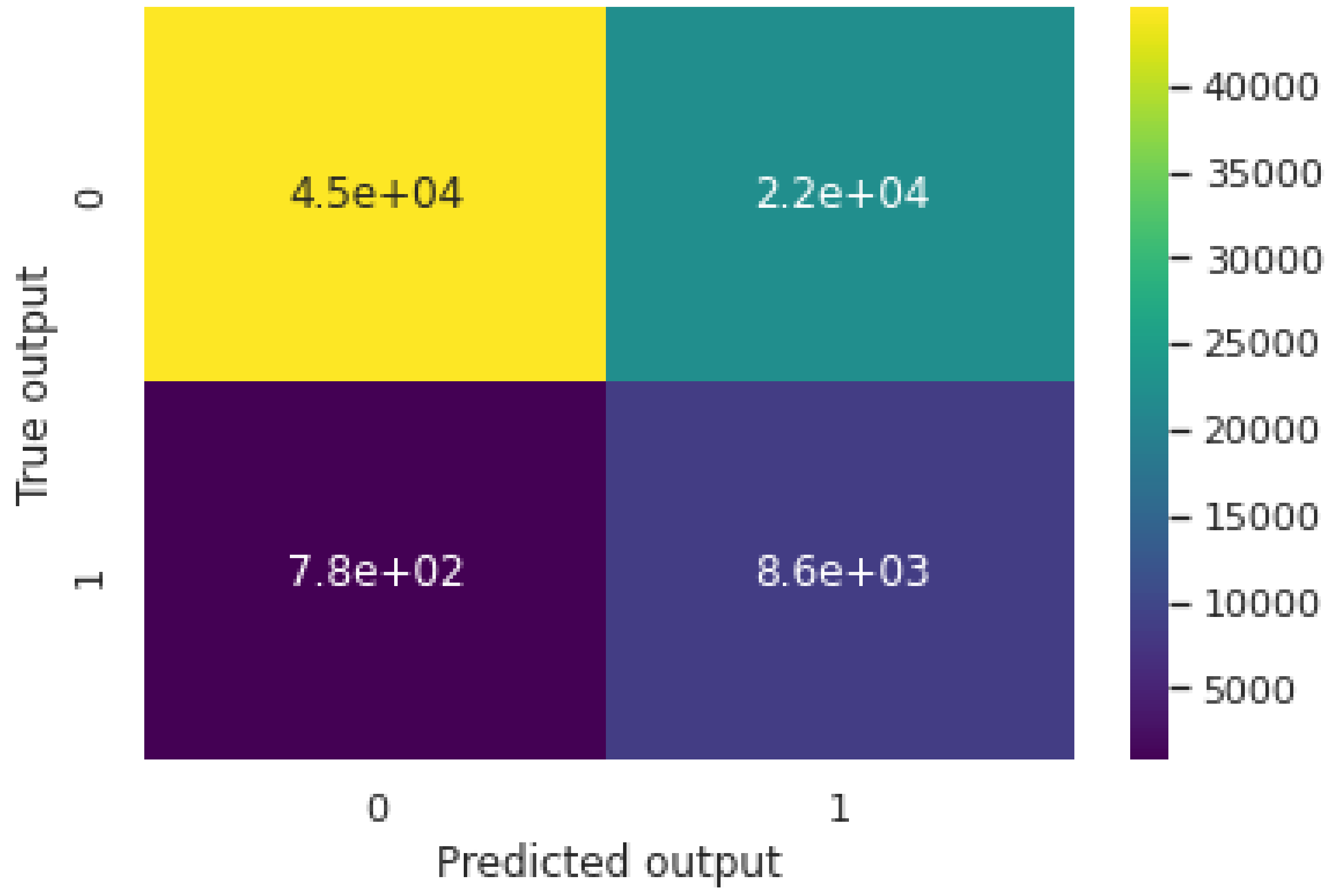
F1 score: 0.426
Recall: 0.91
Precision: 0.278

Oversampling (LGBM)



F1 score: 0.43
Recall: 0.892
Precision: 0.283

Undersampling (LGBM)



F1 score: 0.426
Recall: 0.917
Precision: 0.278

Thank You

Wish you a happy and enjoyable day!