

CAR INSURANCE

DỰ ĐOÁN RỦI RO XÁC NHẬN SỬ
DỤNG BẢO HIỂM TRONG 6
THÁNG TỚI





CONTENT

1

DATASET INTRODUCTION

2

DATA CLEANING

3

EDA

4

MACHINE LEARNING

DATASET INTRODUCTION



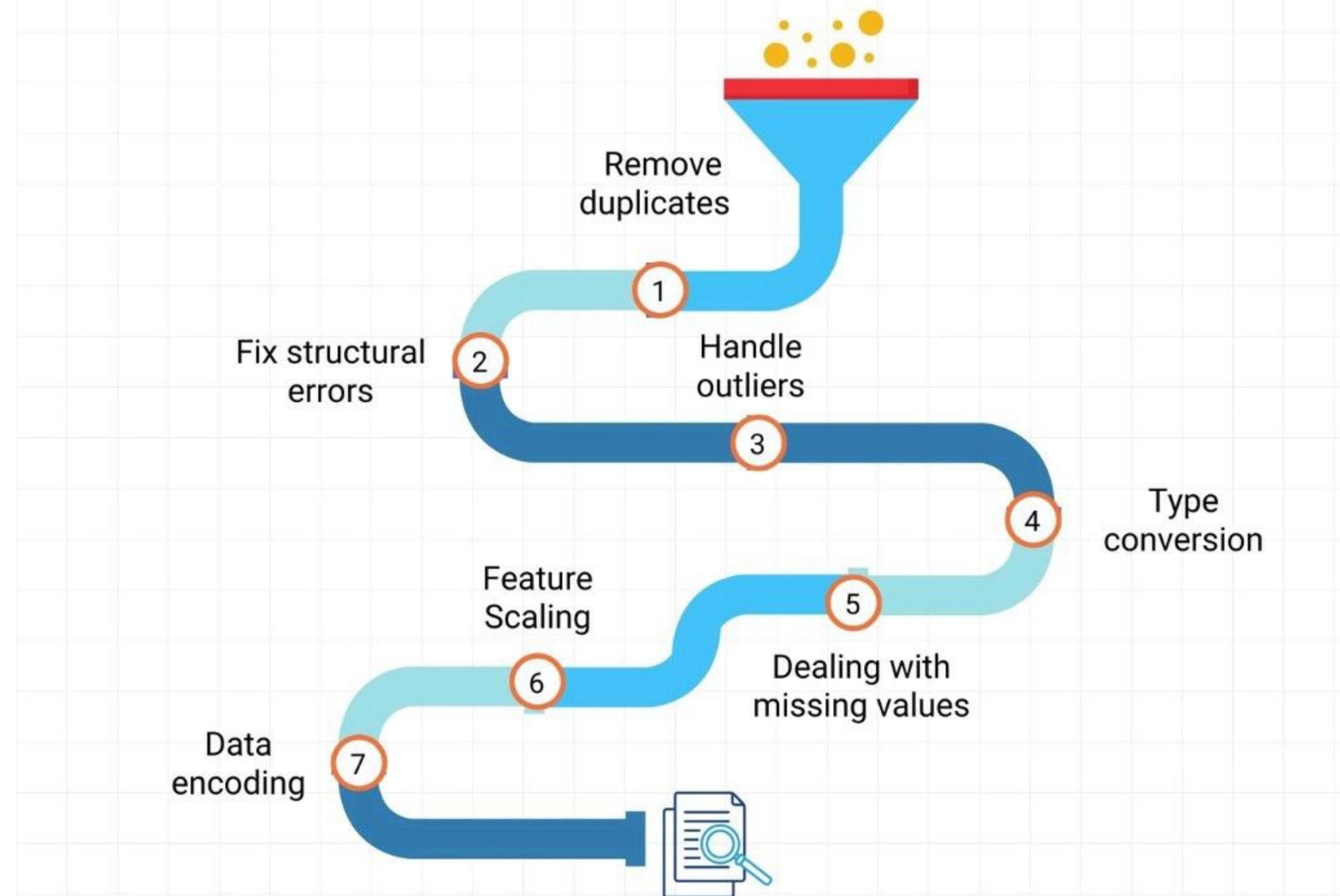
WHY CHOOSE IT

Increasingly, insurance companies are paying attention to:
Predicting customers with high risk tendencies => high probability of using insurance.
=> Maximizing profits.
=> Providing appropriate strategies.



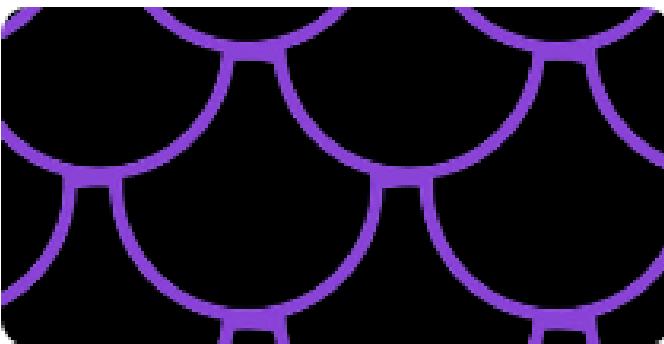
Data preprocessing

The foundation of data science solution



INTRODUCTION

Car Insurance Claim Data



Data Card Code (5) Discussion (3)

About Dataset

No description available

Usability ⓘ

2.94

License

Unknown

Expected update frequency

Not specified

Insurance

[car_insurance_claim.csv](#) (1.58 MB)

Download

Detail Compact Column

ID	KIDSDRV	BIRTH	AGE	HOMEKIDS	YOJ
----	---------	-------	-----	----------	-----

10 of 27 columns

Data Explorer

[Version 1 \(1.58 MB\)](#)

car_insurance_claim.csv

KAGGLE

INTRODUCTION

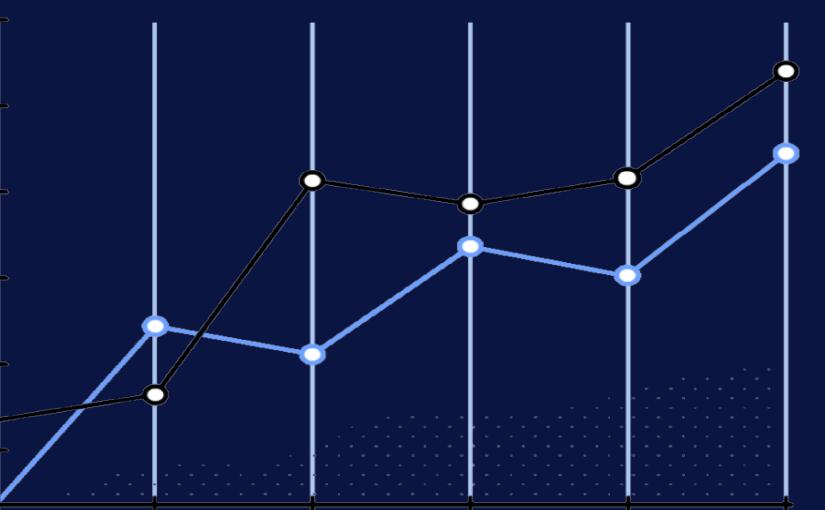
ID	CUSTOMER ID
KIDSDRIV	DRIVING TEENAGER => THE NUMBER OF TEENAGERS USING CAR
AGE	CUSTOMER AGE
HOMEKIDS	THE NUMBER OF KIDS
YOJ	YEAR ON JOB - THE LONGEST TIME CUSTOMER STAY A JOB
INCOME	SALARY EACH YEAR
PARENT1	SINGLE PARENTS (YES OR NO)
HOME_VAL	VALUE OF CUSTOMER HOUSE => IF VALUE EQUAL 0 MEANS CUSTOMER DON'T OWN ANY HOUSE
MSTATUS	MARITAL STATUS
GENDER	GENDER
EDUCATION	THE HIGHEST LEVEL OF EDUCATION
OCCUPATION	NAME OF JOB
TRAVTIME	TIME TO WORK
CAR_USE	THIS IS THE PRIVATE OR COMMERCIAL VEHICLE USED BY CUSTOMER
BLUEBOOK	THE VALUE OF CARS
TIF	TIME IN FORCE - HOW LONG THEY BECOME CUSTOMER WITH COMPANY

INTRODUCTION

CAR_TYPE	TYPE OF CAR
RED_CAR	COLOR OF CAR IS RED OR NOT
OLDCLAIM	THE PREVIOUS PAYOUTS IN LAST 5 YEARS
CLM_FREQ	TOTAL NUMBER OF CLAIMS
REVOKED	LICENSE REVOKED IN 7 YEARS
MVR PTS	MOTOR VEHICLE RECORD POINTS => GET MORE TRAFFIC TICKETS (GIẤY XỬ PHẠT)
CLM_AMT	USED CLAIM PAYMENT
CAR_AGE	AGE OF CAR
CLAIM_FLAG	WAS CAR IN CRASH (1 OR 0)
URBANICITY	HOME/WORK AREA

10302 observations
27 variables

DATA CLEANING





CONTENT

**MISSING
VALUES**

**DATA
INCONSISTENT**

OUTLIERS

DATA INCONSISTENT

INCOME

\$67,349

\$91,449

\$52,881

\$16,039

\$114,986

\$125,301

\$18,755

EDUCATION

PhD

z_High School

Bachelors

z_High School

<High School

PhD

CAR_TYPE

Minivan

Minivan

Van

z_SUV

Minivan

z_SUV

DATA INCONSISTENT

```
df['INCOME'] = df['INCOME'].str.replace('$', '').str.replace(',', '').astype(float)
df['HOME_VAL'] = df['HOME_VAL'].str.replace('$', '').str.replace(',', '').astype(float)
df['OLDCLAIM'] = df['OLDCLAIM'].str.replace('$', '').str.replace(',', '').astype(float)
df['CLM_AMT'] = df['CLM_AMT'].str.replace('$', '').str.replace(',', '').astype(float)
df['GENDER'] = df['GENDER'].str.replace('z_', '')
df['MSTATUS'] = df['MSTATUS'].str.replace('z_', '')
df['EDUCATION'] = df['EDUCATION'].str.replace('z_', '').str.replace('<', '')
df['OCCUPATION'] = df['OCCUPATION'].str.replace('z_', '')
df['BLUEBOOK'] = df['BLUEBOOK'].str.replace('$', '').str.replace(',', '').astype(int)
df['CAR_TYPE'] = df['CAR_TYPE'].str.replace('z_', '')
df['CAR_TYPE'] = df['CAR_TYPE'].str.replace('z_', '')
df['URBANICITY'] = df['URBANICITY'].str.replace('Highly Urban/ ', '').str.replace('z_Highly Rural/ ', '')
```

MISSING VALUES

MODE IMPUTATION

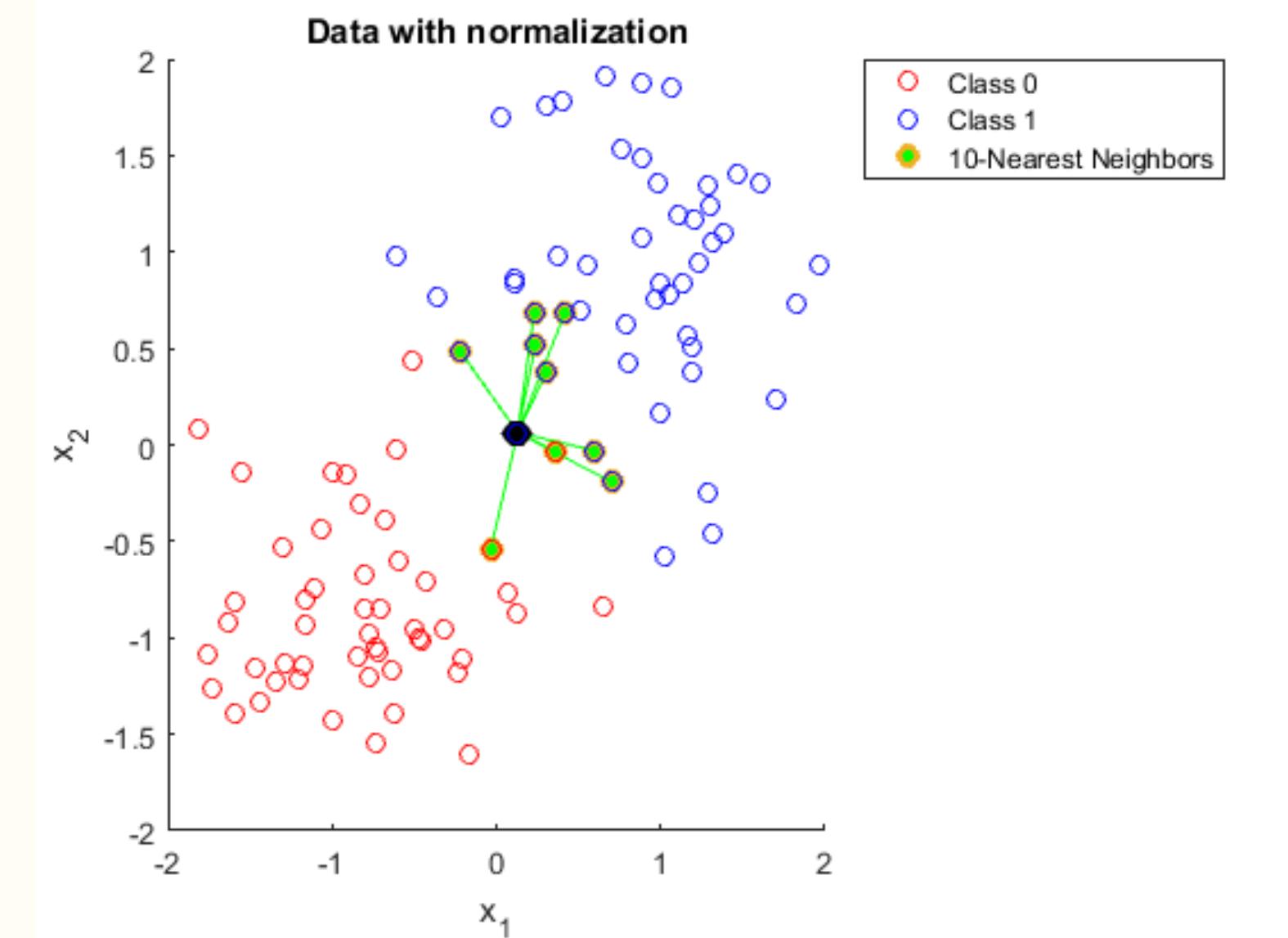
Make	Price
Ford	Ford
Ford	Ford
Fiat	Fiat
BMW	BMW
Ford	Ford
Kia	Kia
Fiat	Fiat
Ford	Ford
Kia	Kia

Mode = Ford



Price
Ford
Ford
Fiat
BMW
Ford
Kia
Fiat
Ford
Kia

KNN IMPUTATION



MISSING VALUES

```
missing_value_table(df)
```

These dataframe has 26 columns

These dataframe has 6 columns which has null values

Missing Value	Percentage of Missing
---------------	-----------------------

OCCUPATION	665	6.5
CAR_AGE	639	6.2
HOME_VAL	575	5.6
INCOME	570	5.5
YOJ	548	5.3
AGE	7	0.1

CATEGORY

OCCUPATION

NUMERIC

CAR_AGE

HOME_VAL

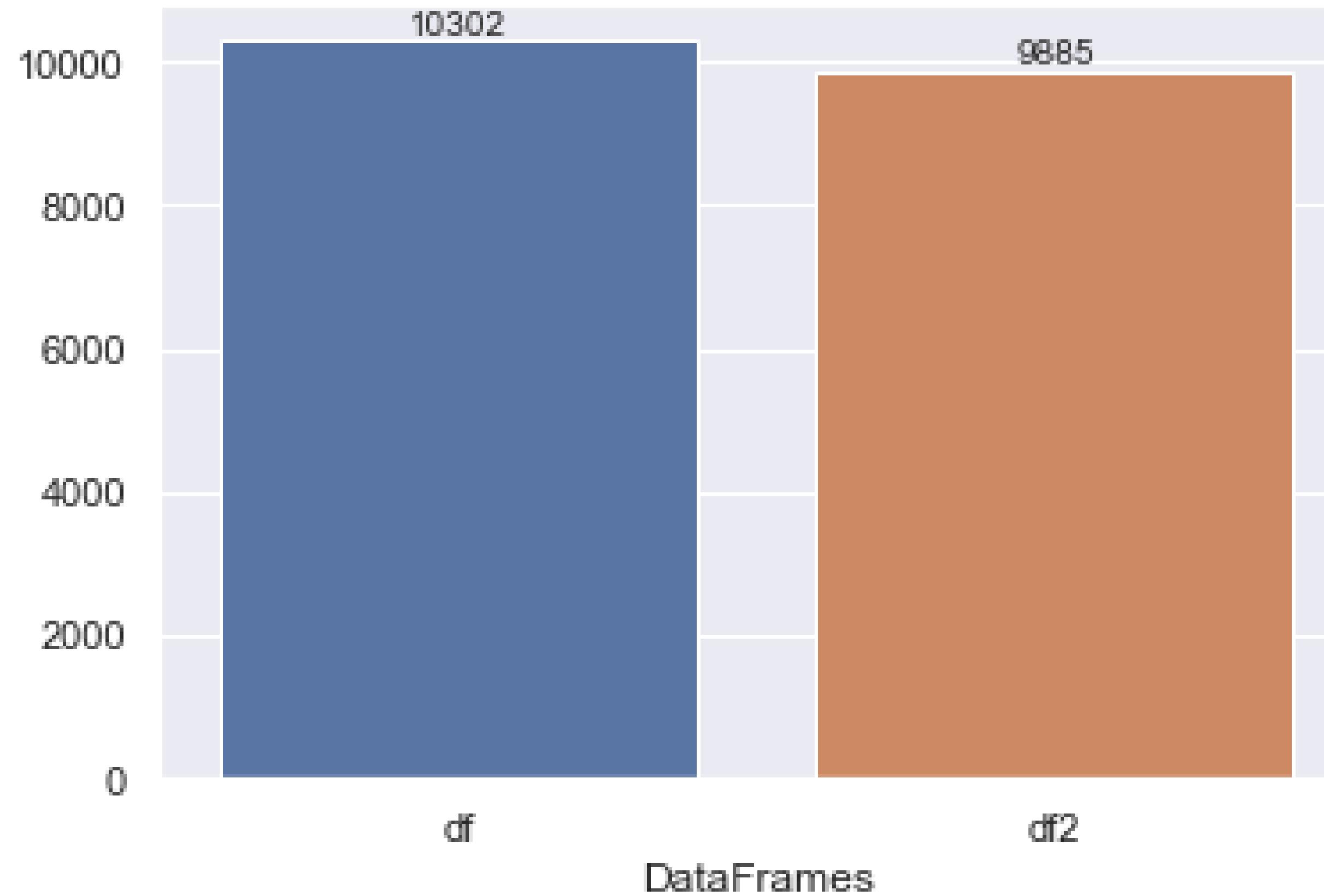
INCOME

YOJ

AGE

OUTLIERS

Lengths of DataFrames



417 OUTLIERS

EXPLORATORY DATA ANALYSIS





CONTENT

**HISTORY OF
USING CAR
INSURANCE**

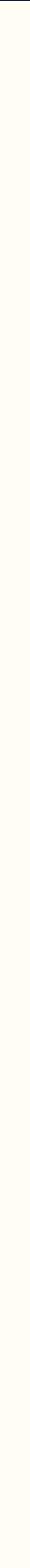
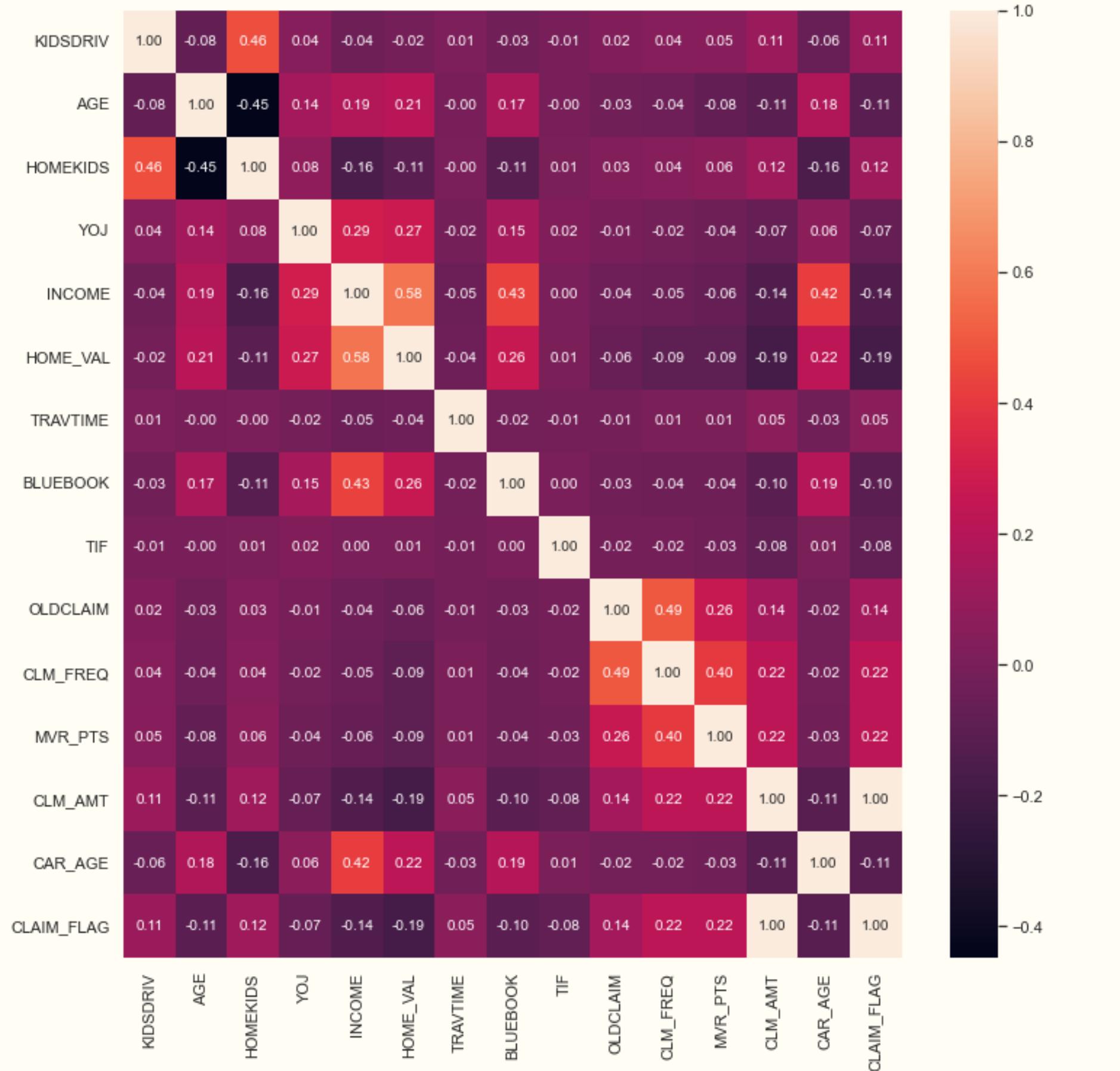
**CUSTOMER
PROFILE**

VEHICLE

CUSTOMER'S PROFILE



CORRELATION

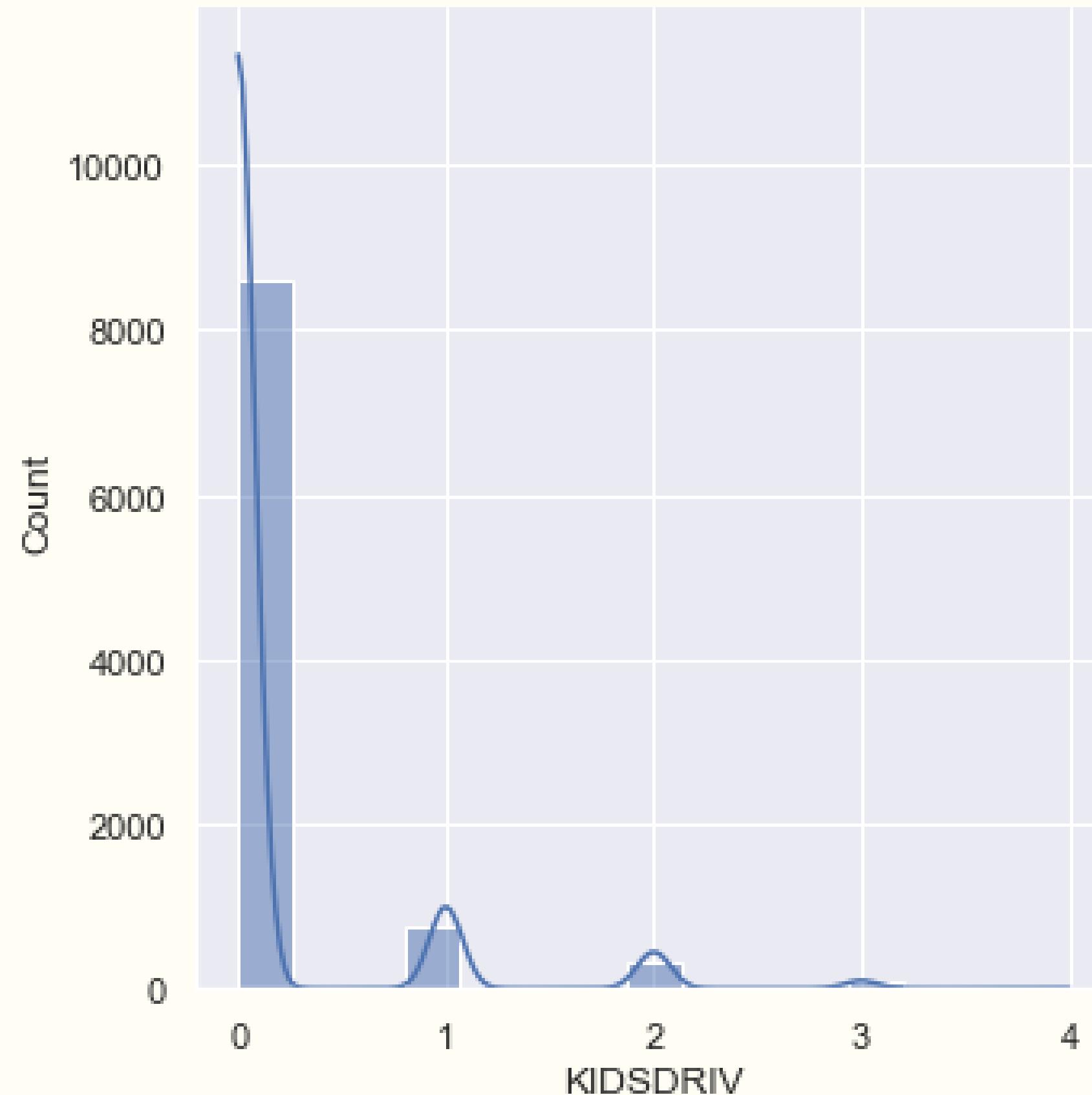


The correlation between CLM_AMT and CLAIM_FLAG is 1

ALL CUSTOMERS INVOLVED IN CAR ACCIDENTS USE INSURANCE.

KIDS DRIVING

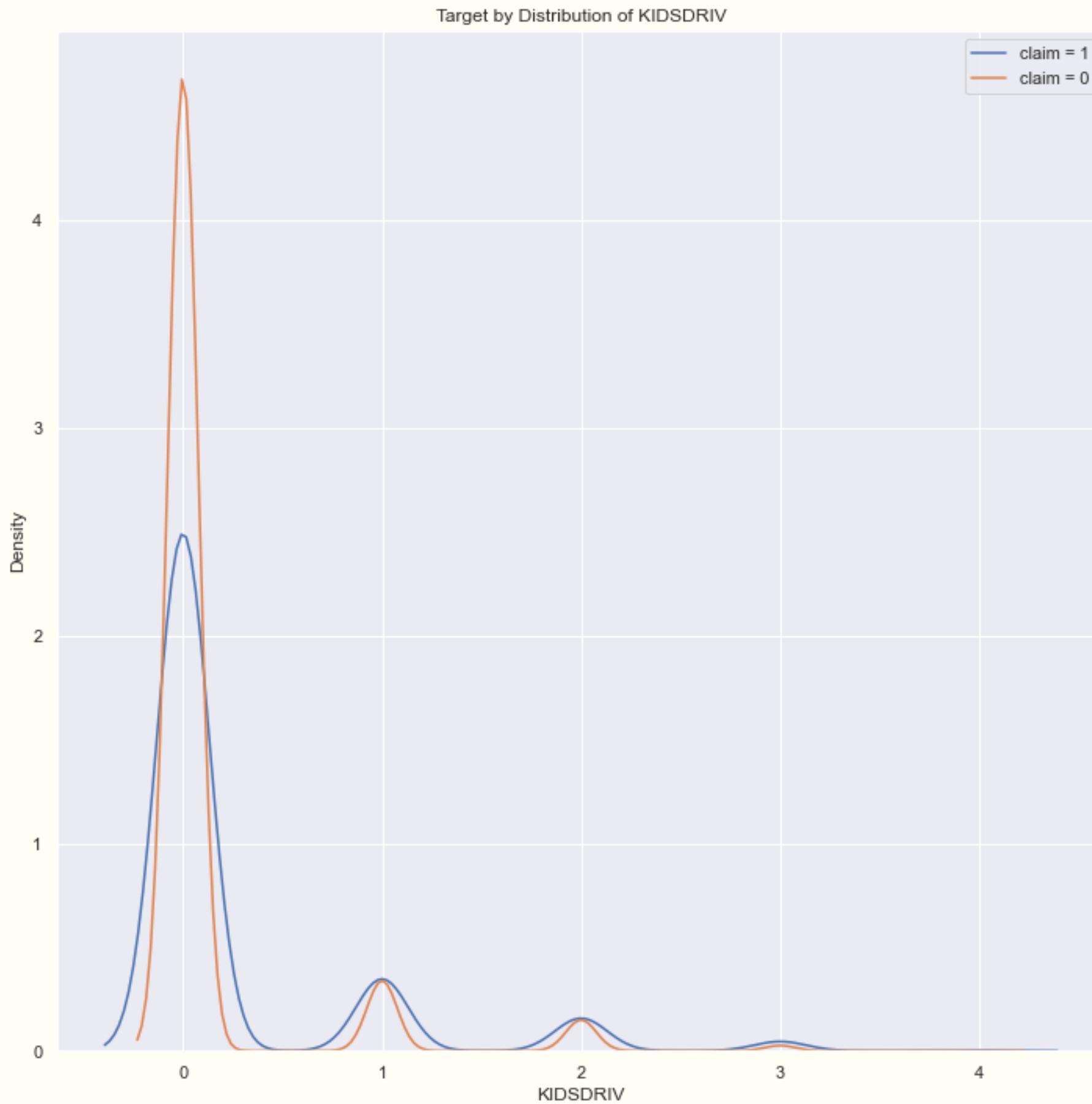
Distribution by Driving Children



83.7%

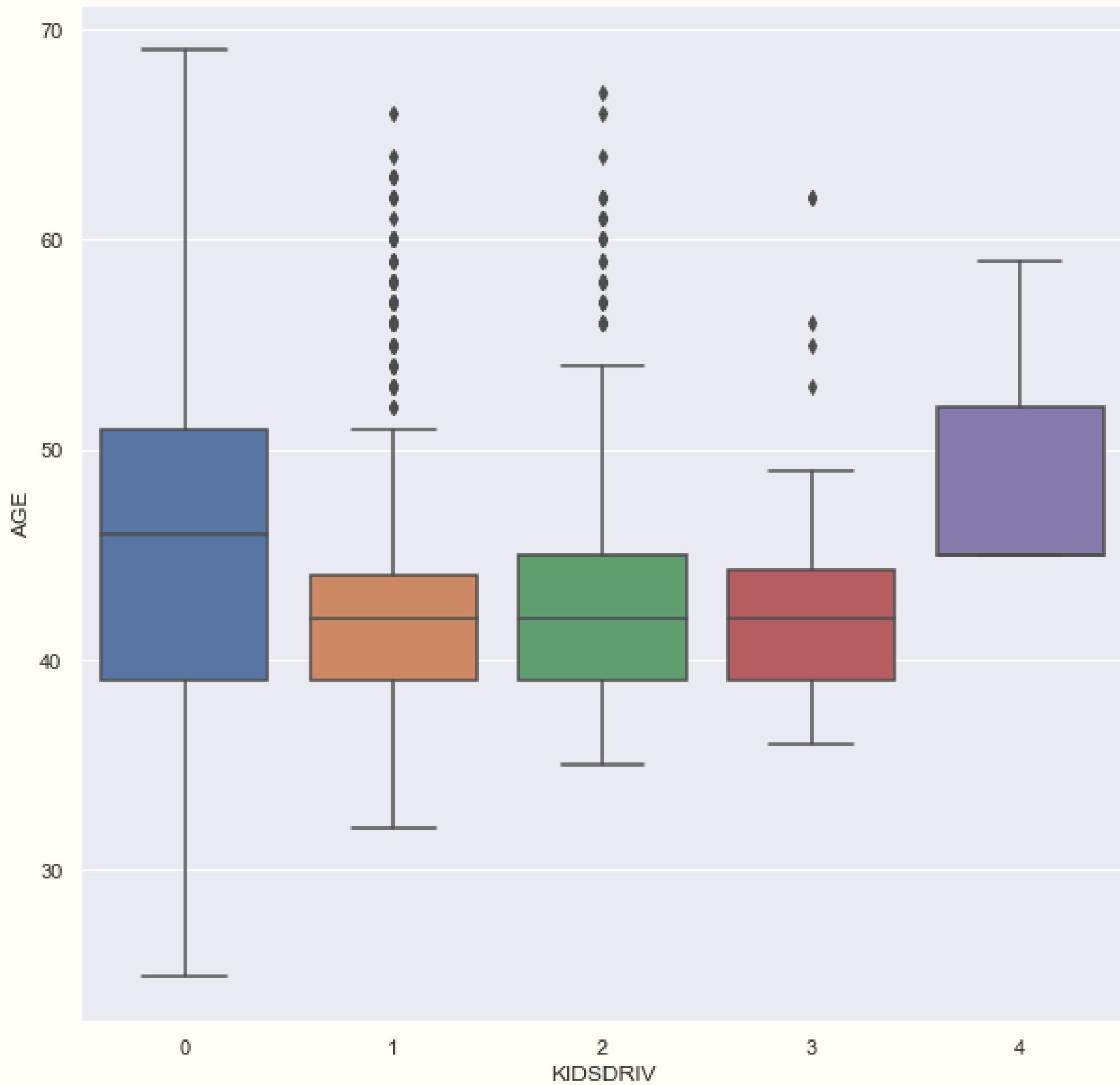
The group of customers who do not
allow teenagers to use their cars
have the highest proportion

KIDS DRIVING



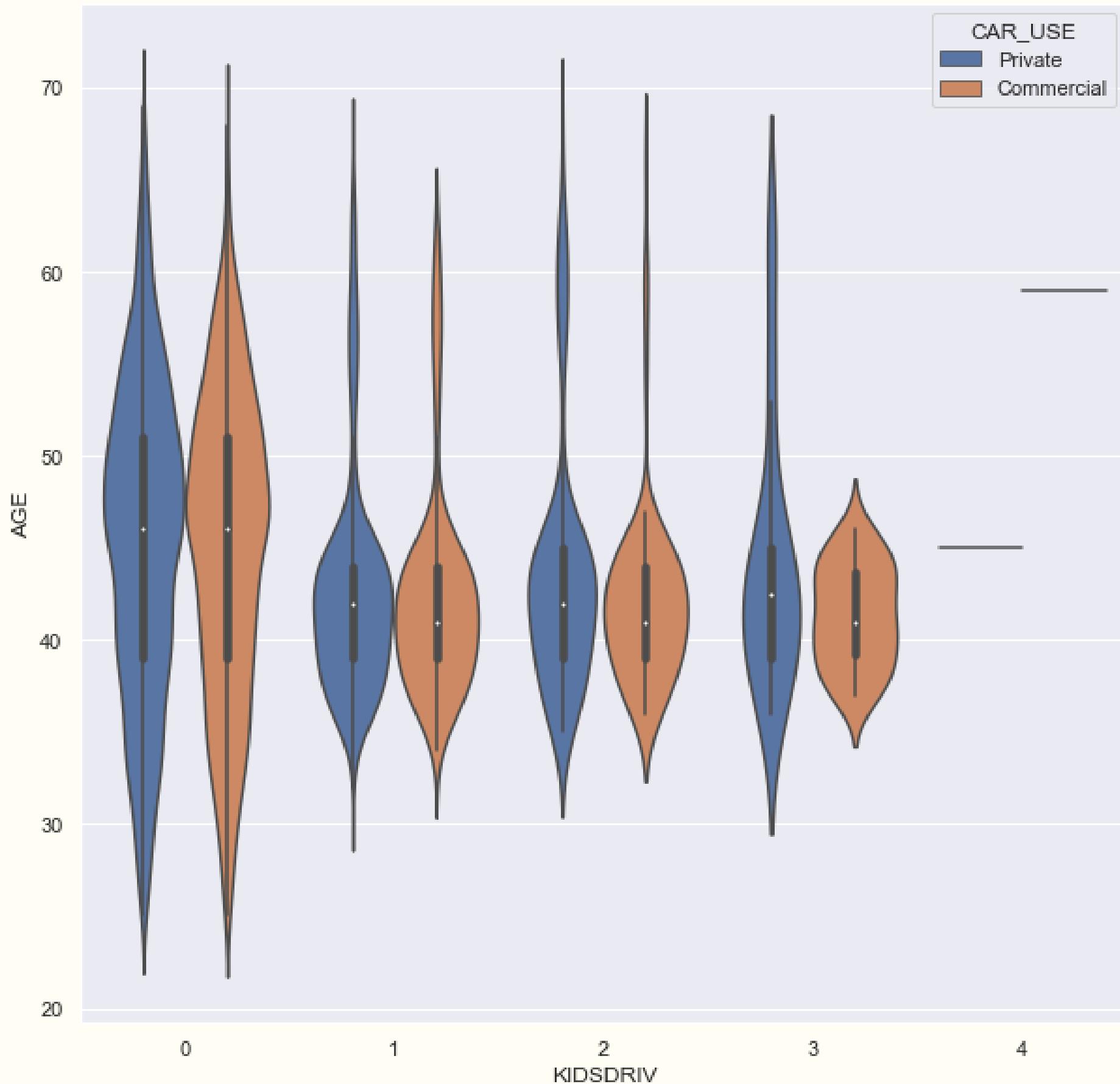
**THE GROUP OF CUSTOMERS WHO
ALLOW TEENAGERS TO USE THEIR
VEHICLES TEND TO HAVE A HIGHER
TENDENCY TO USE CAR
INSURANCE.**

KIDS DRIVING



THE AGE GROUP OF CUSTOMERS
ALLOWING TEENAGERS TO USE
THEIR VEHICLES IS CONCENTRATED
FROM 35-55.

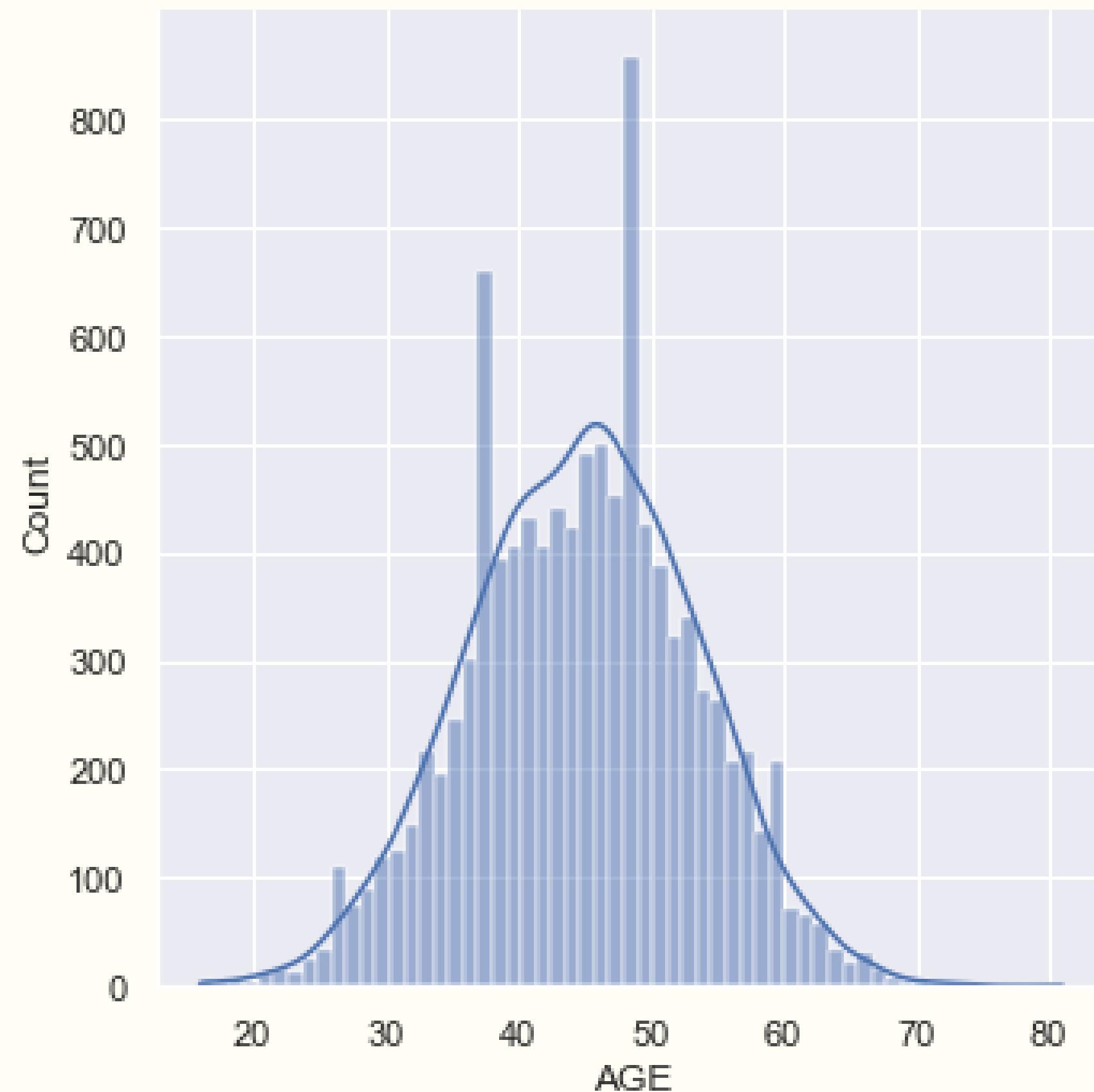
KIDS DRIVING



**THERE IS NO AGE DIFFERENCE
BETWEEN PERSONAL OR
COMMERCIAL VEHICLES WITHIN
THE KIDSDRIV GROUP**

AGE

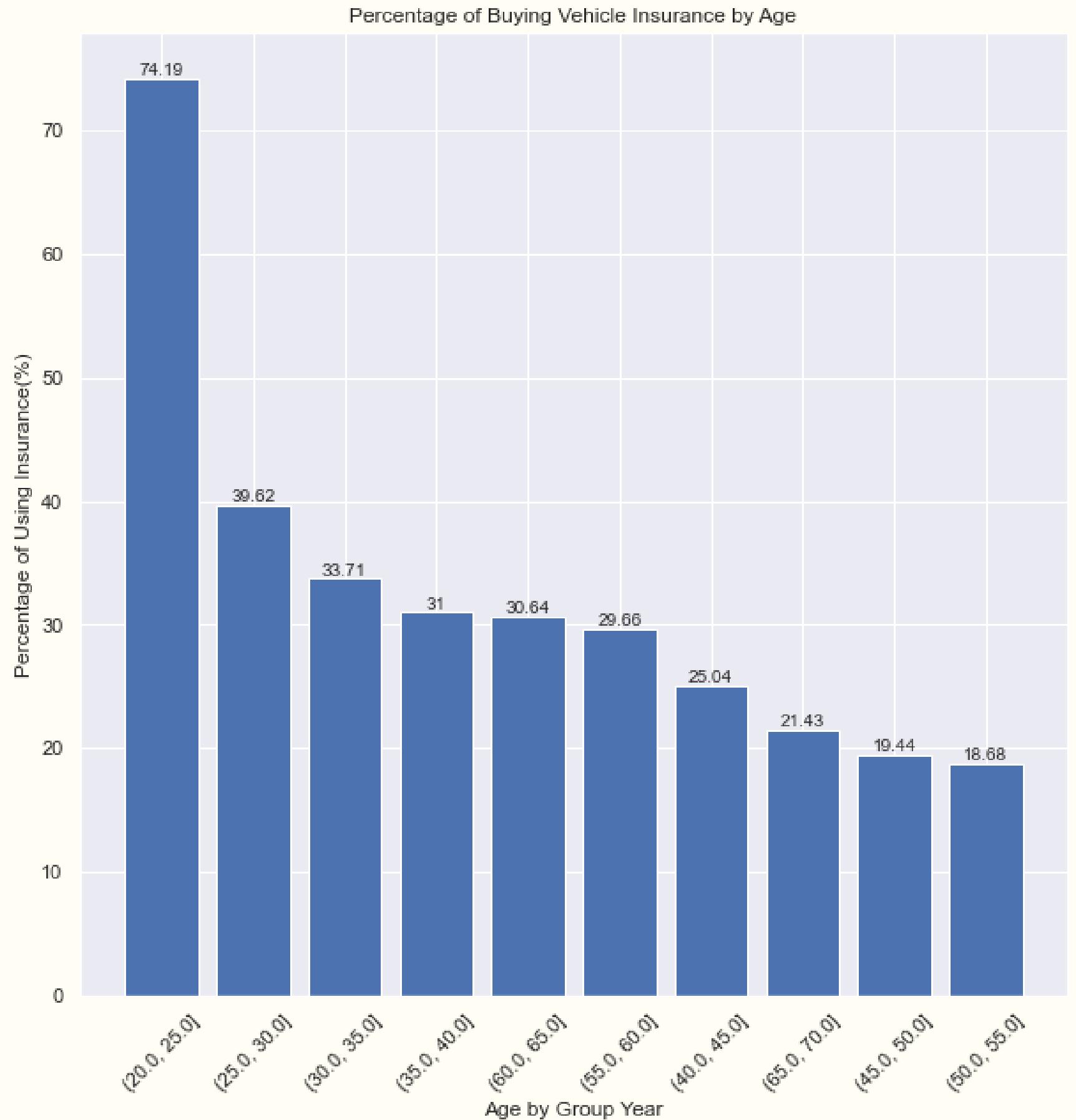
Distribution by AGE



35-55

**THE MAIN AGE GROUP OF THE
DATA**

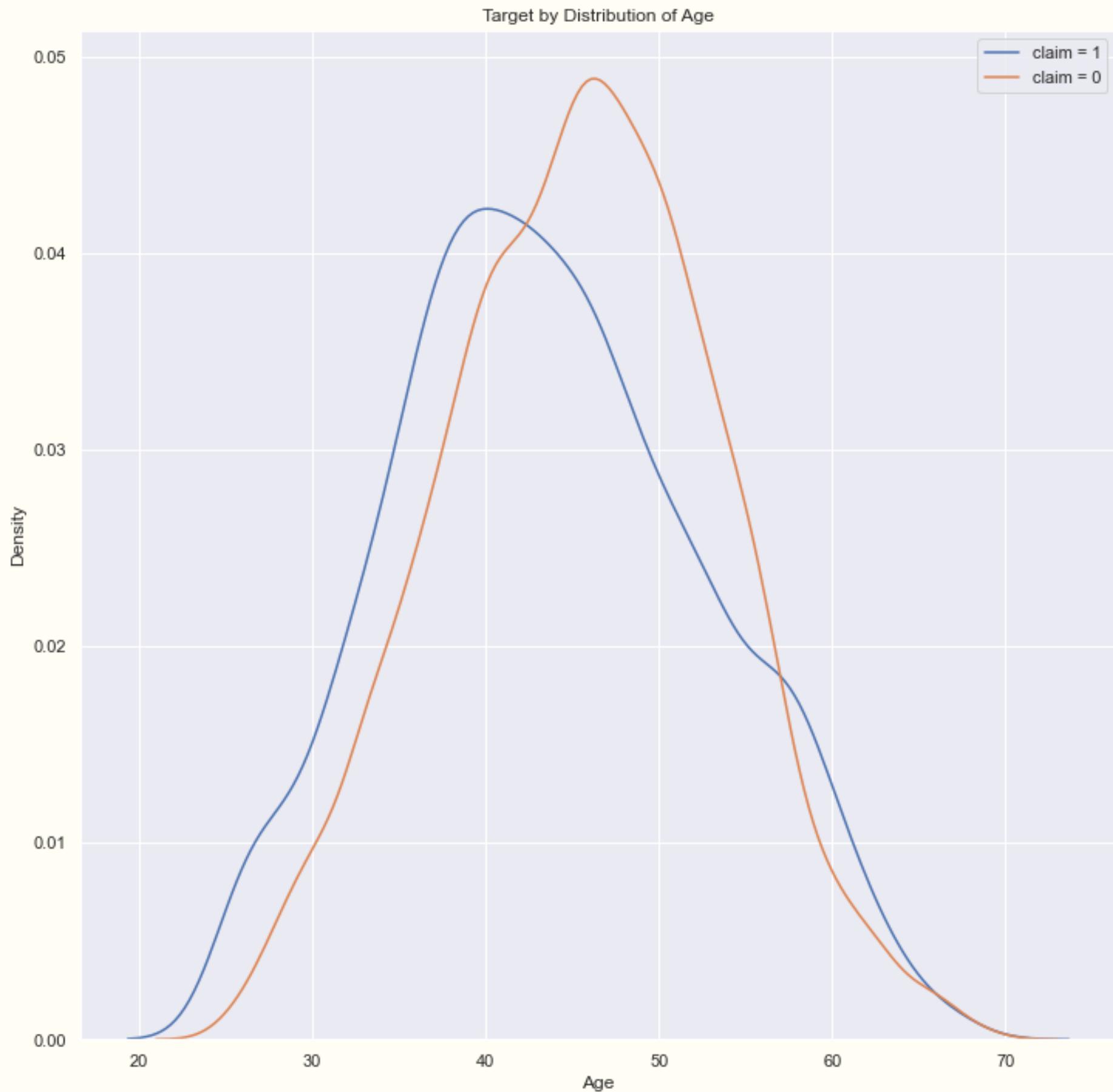
AGE



19%-31%

**PERCENTAGE OF THE AGE GROUP
35-55 USING CAR INSURANCE**

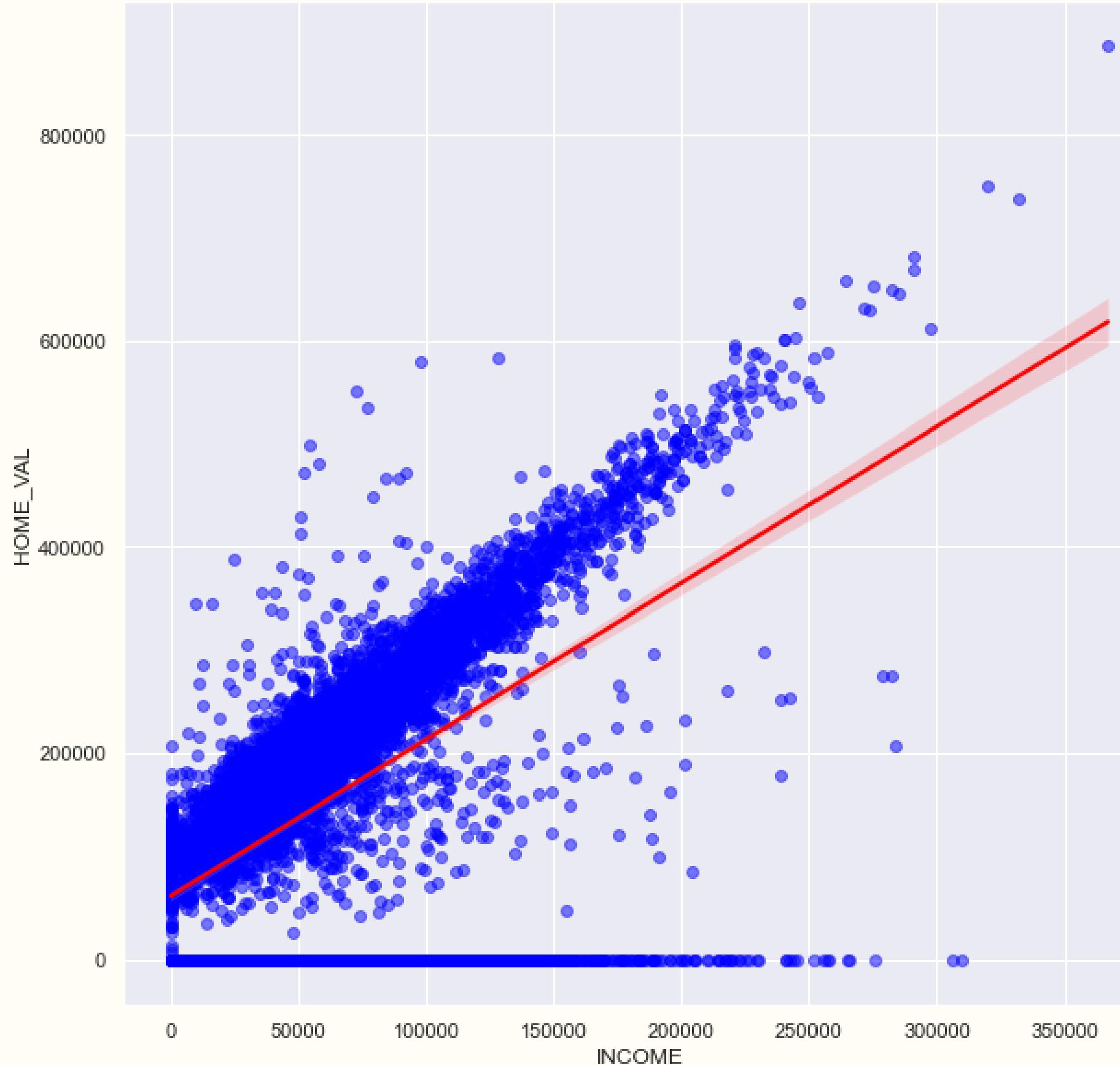
AGE



**THE AGE GROUPS THAT MAKE UP A
LARGE NUMBER OF THE DATASET
ARE ALSO THE GROUPS WITH THE
HIGHEST CAR INSURANCE USAGE**

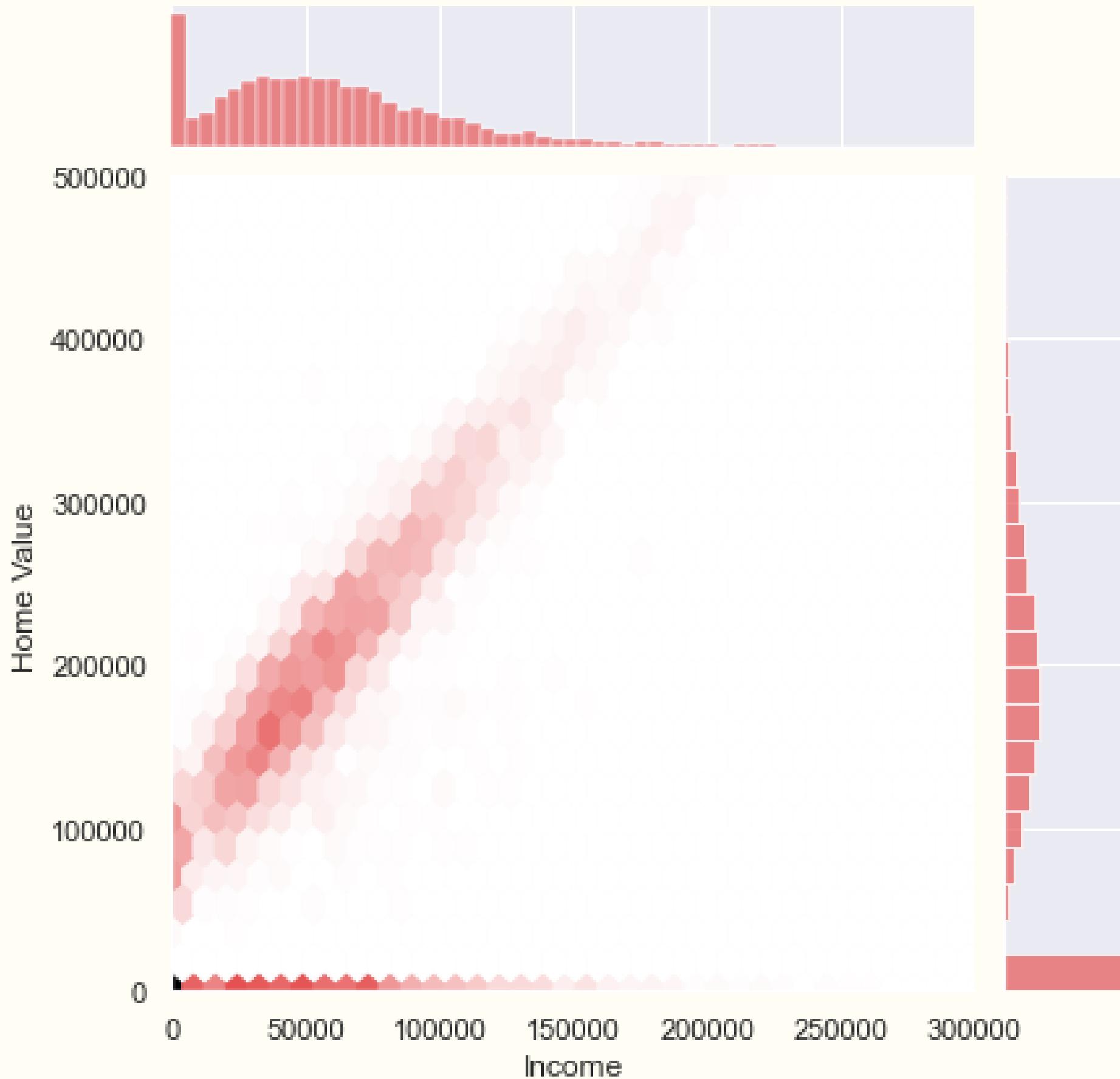
INCOME AND HOME VALUE

Relationship between Home Value and Income



**THE RELATIONSHIP BETWEEN
INCOME AND HOME VALUE IS
POSITIVE.**

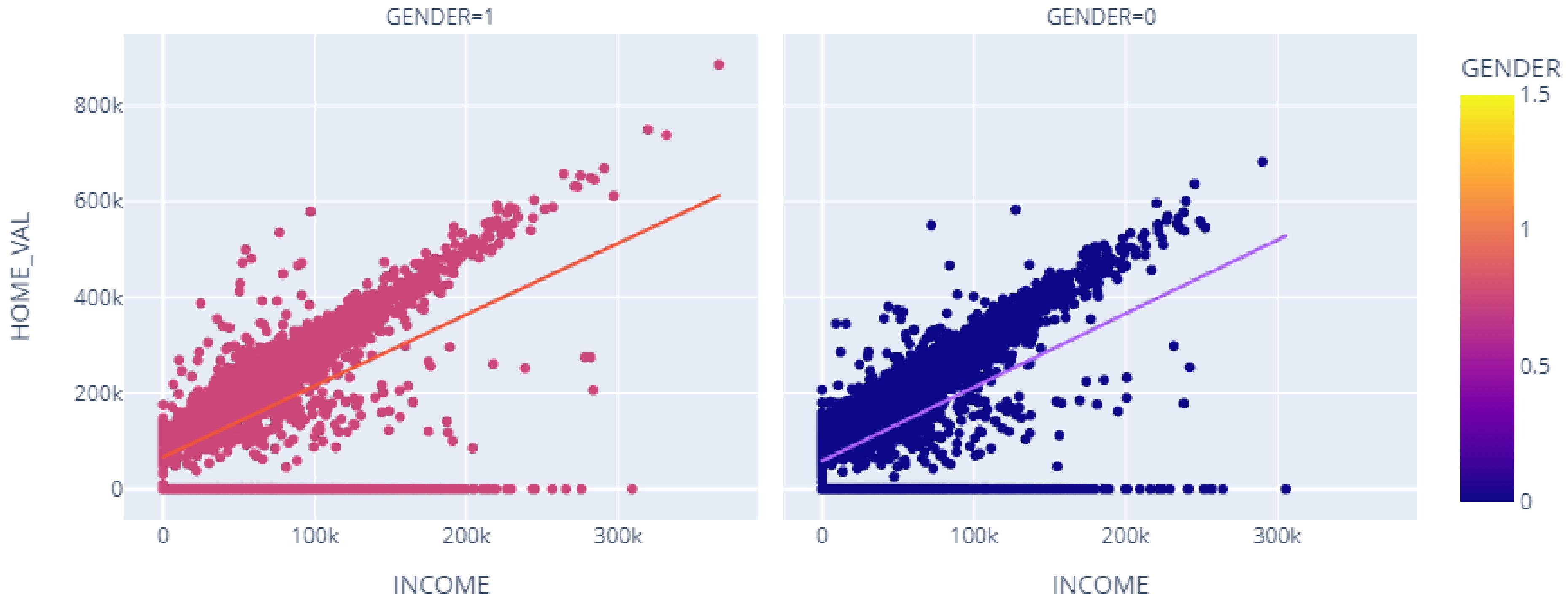
INCOME AND HOME VALUE



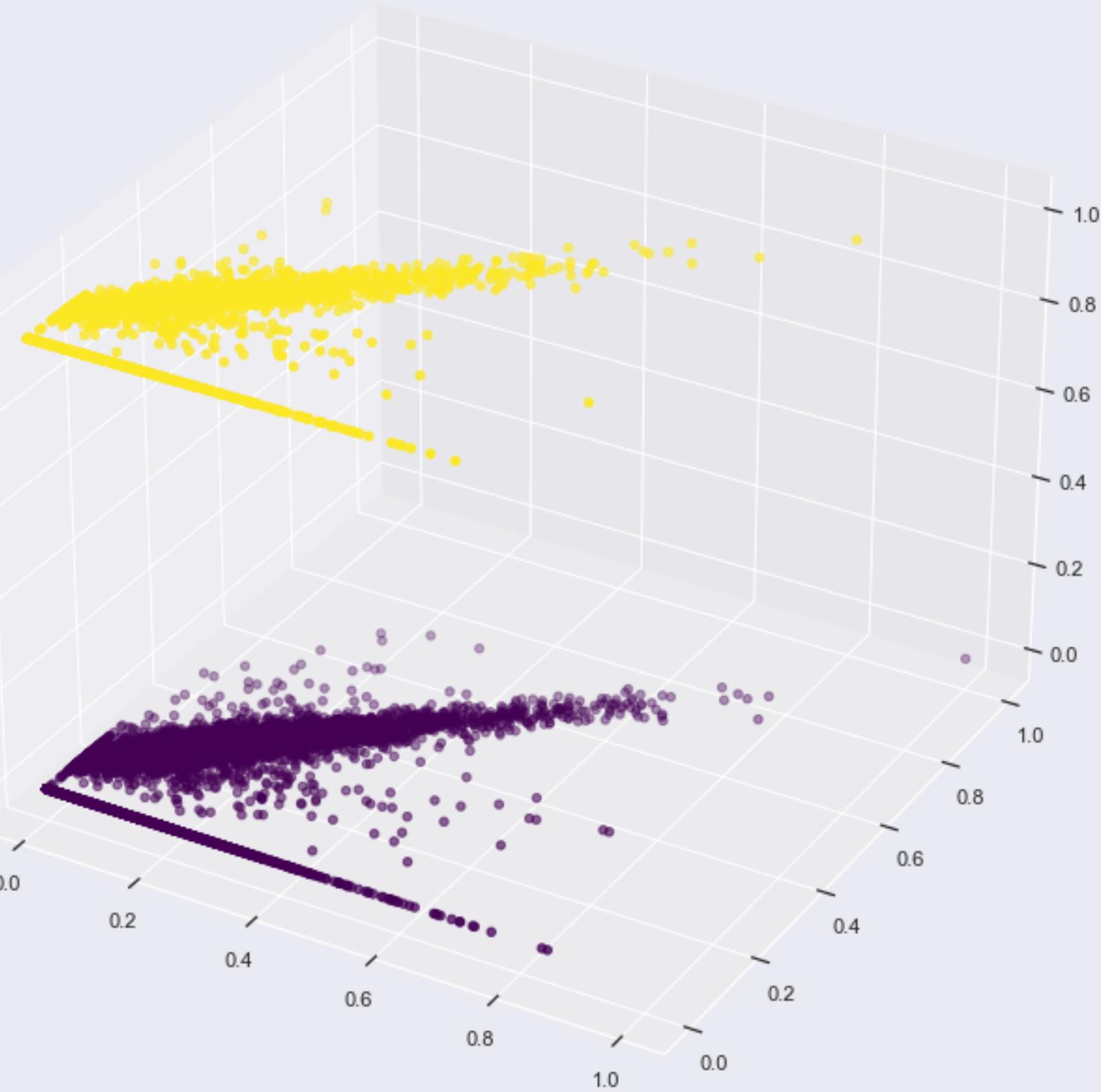
**THE RELATIONSHIP BETWEEN
INCOME AND HOME VALUE IS
POSITIVE.**

INCOME AND HOME VALUE

Relationship with Income and Home Value



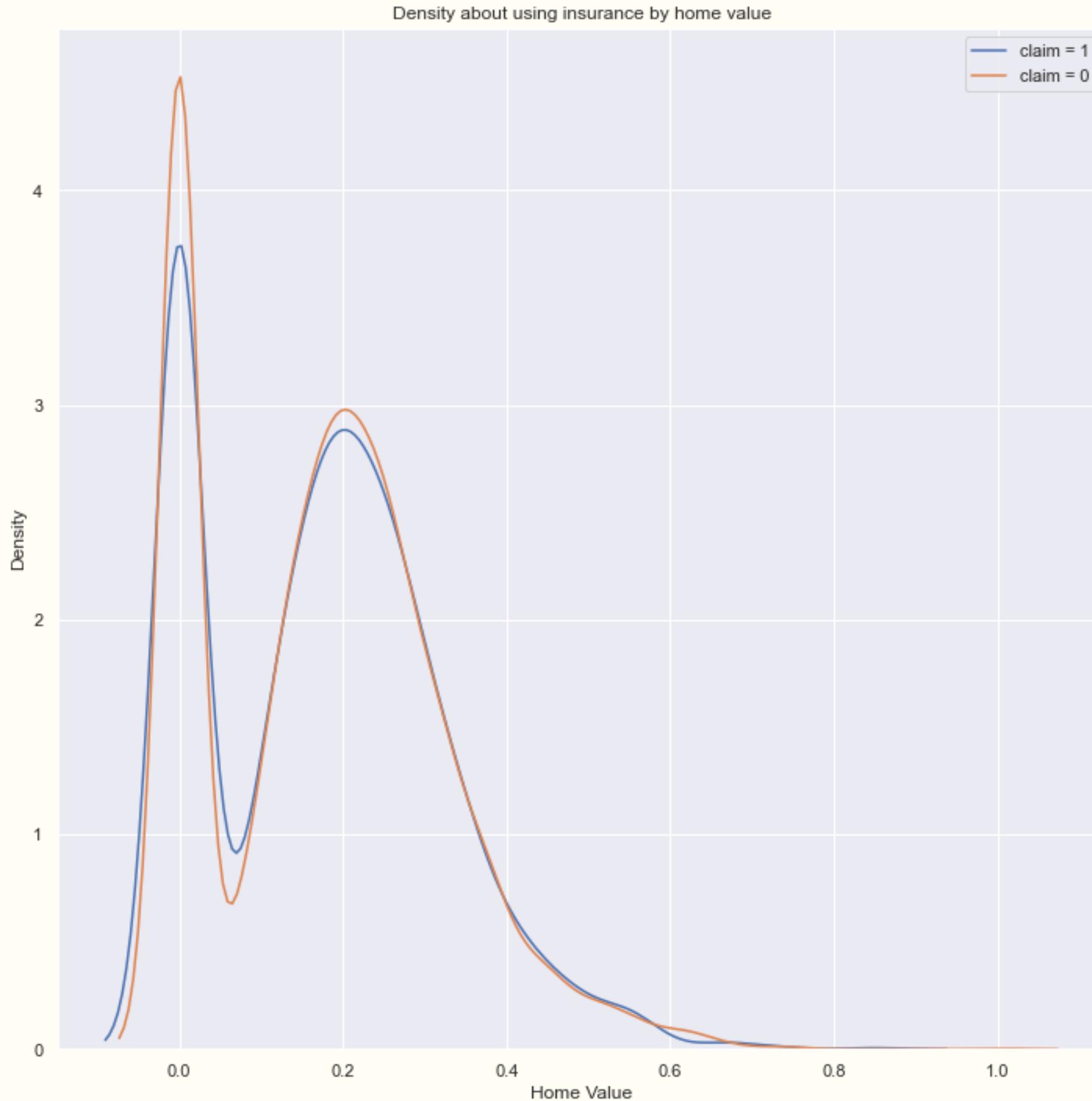
INCOME AND HOME VALUE



**THERE IS VIRTUALLY NO
DIFFERENCE IN THE IMPACT OF
INCOME AND HOME VALUE ON THE
USE OF INSURANCE**

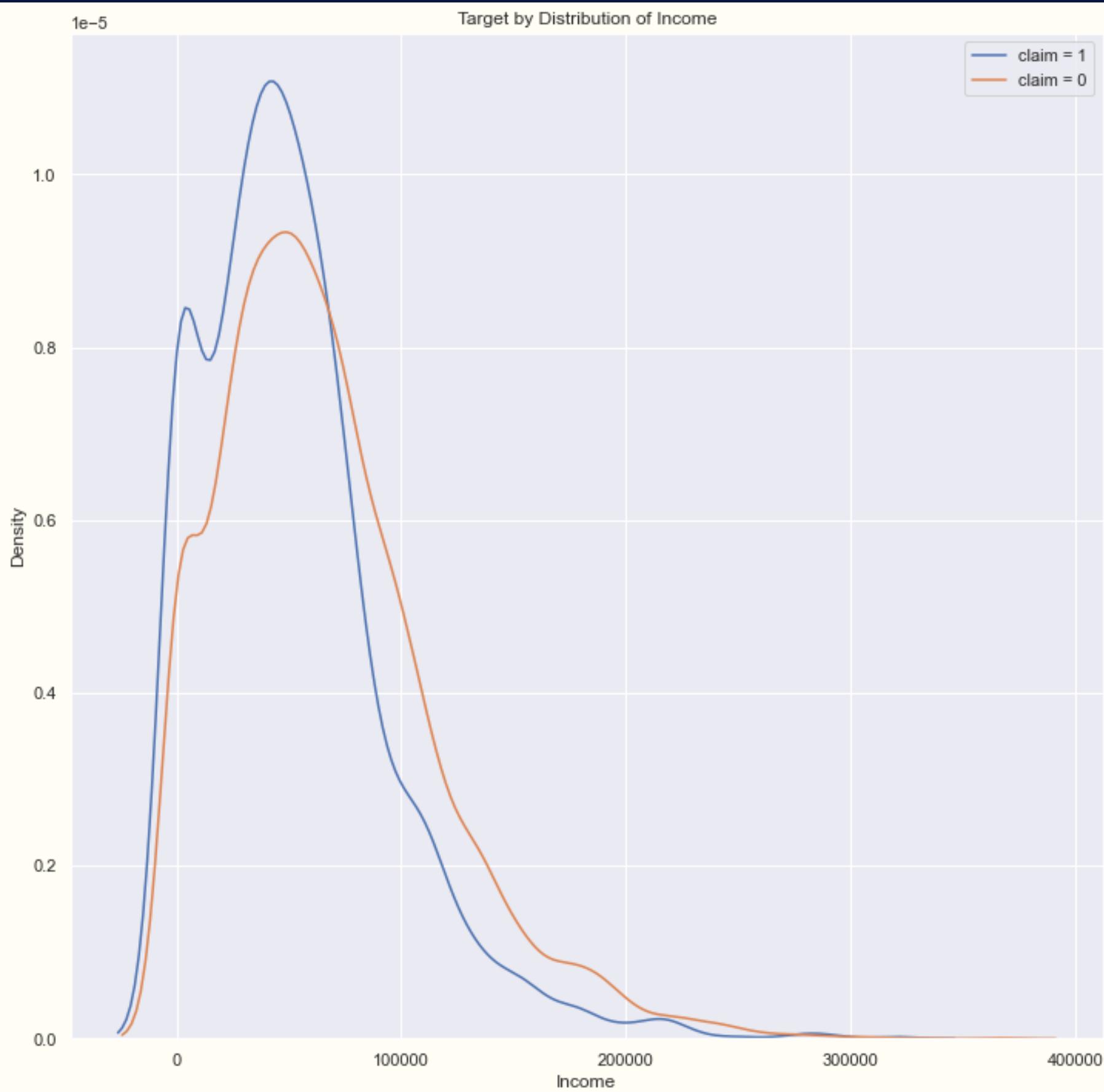
**THE DATA POINTS SHOW THAT
INDIVIDUALS WITH HIGH INCOME
AND HOME VALUES OFTEN DO NOT
USE INSURANCE.**

INCOME AND HOME VALUE



**THE GROUP OF CUSTOMERS WHO
DO NOT OWN HOMES TEND TO
HAVE A HIGHER TENDENCY TO USE
INSURANCE THAN THE GROUP OF
CUSTOMERS WHO DO OWN
HOMES.**

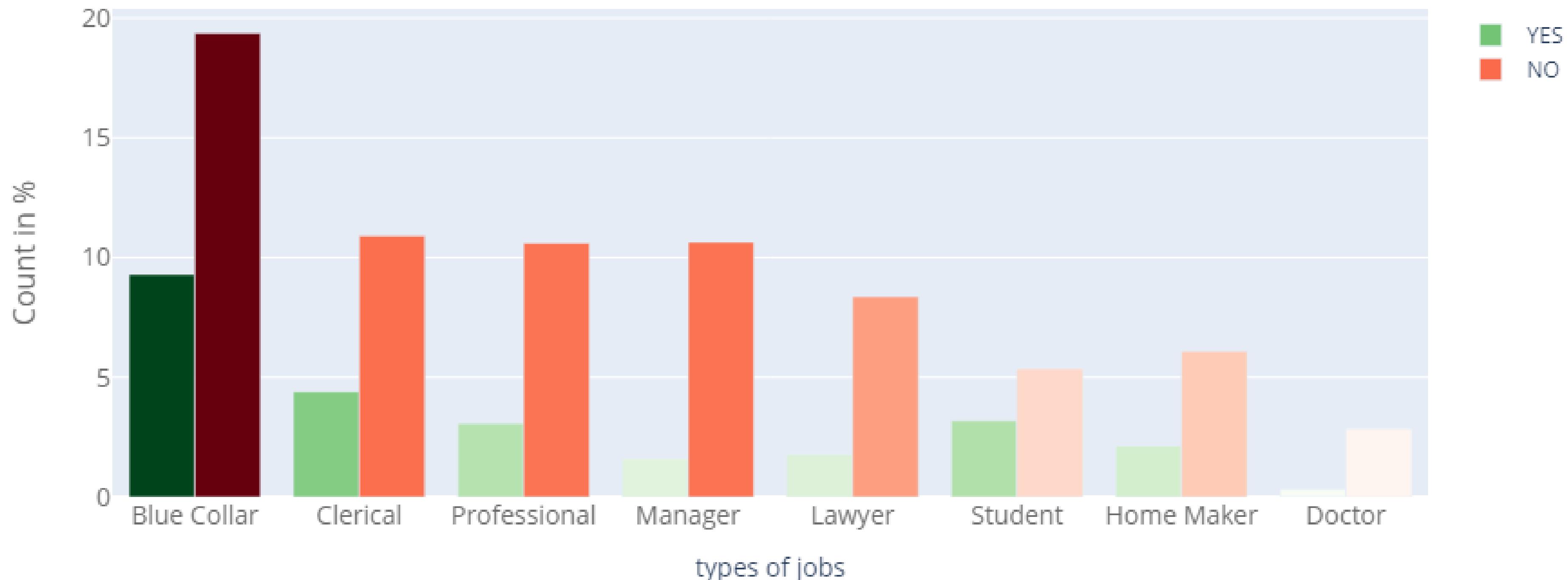
INCOME AND HOME VALUE



**THE GROUP OF CUSTOMERS WITH
LOW INCOME TEND TO HAVE A
HIGHER TENDENCY TO USE
INSURANCE BUT IT'S NOT CLEAR**

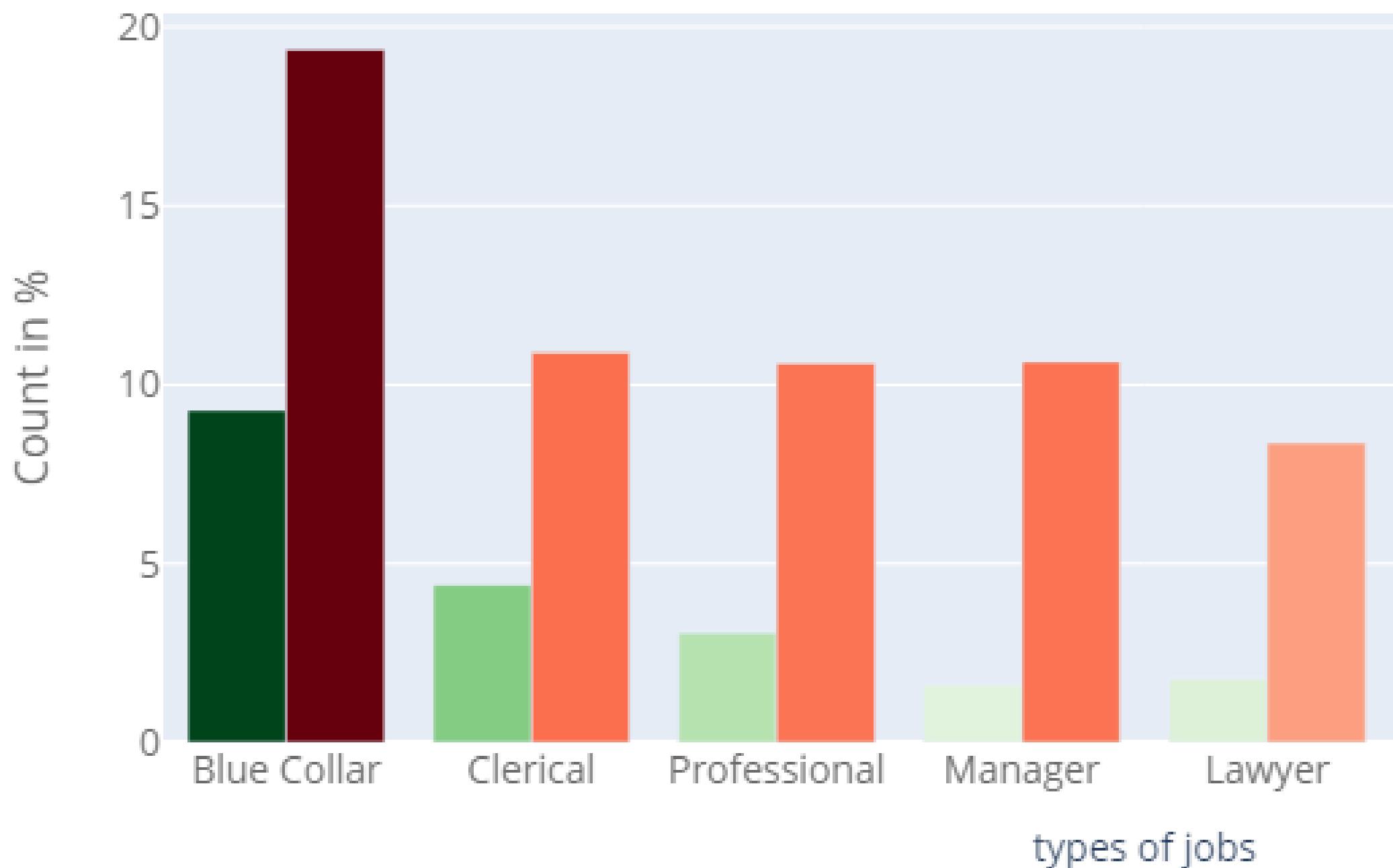
OCCUPATION

For which jobs - applicant's applied for insurance in terms of insurance used or not %



OCCUPATION

For which jobs - applicant's applied for insurance in terms of ins



9.3%

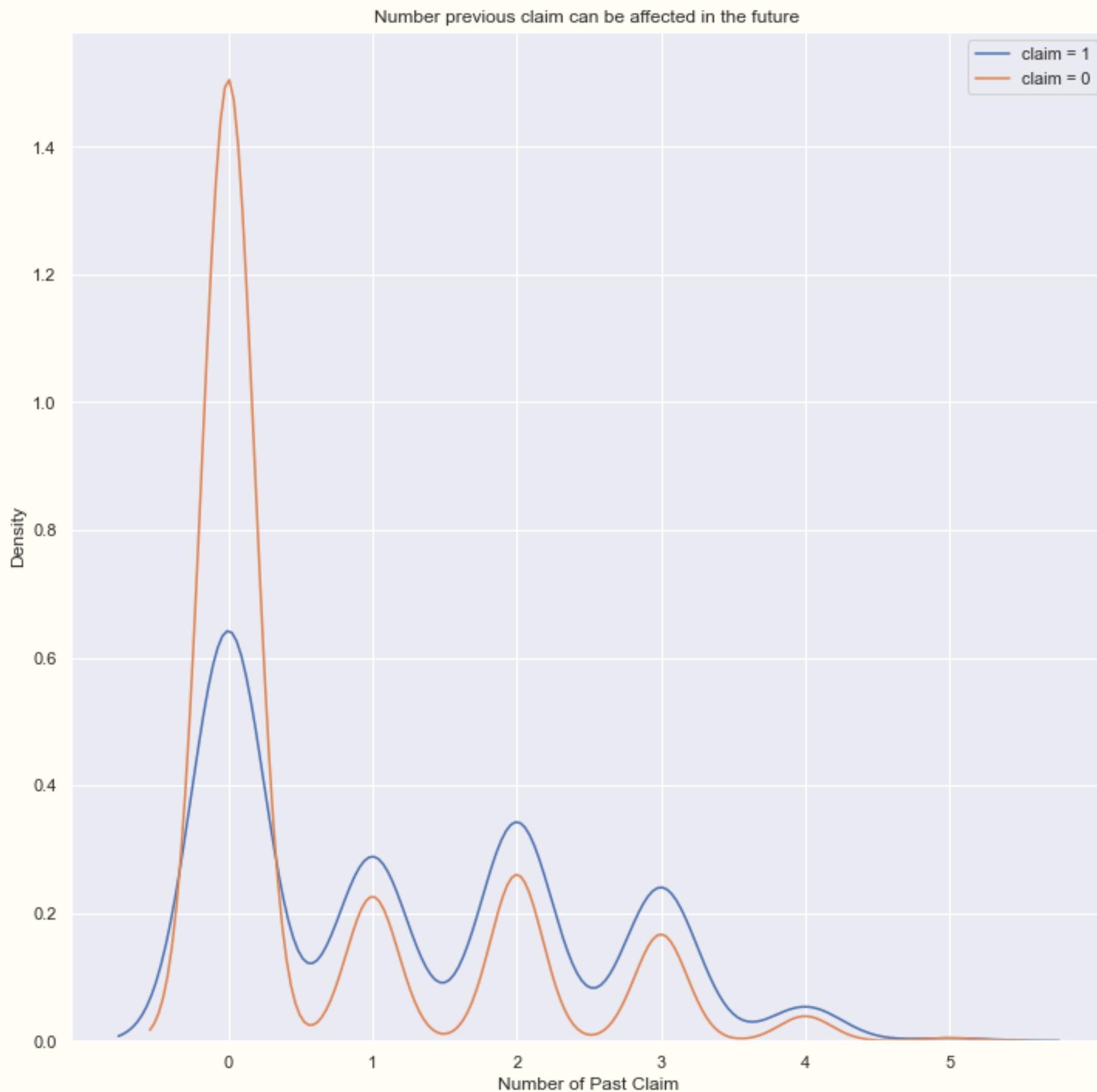
**THE CUSTOMER GROUP WITH
FEWER OFFICE JOBS HAS THE
HIGHEST INSURANCE USAGE RATE**

48.0%

CUSTOMER'S HISTORY



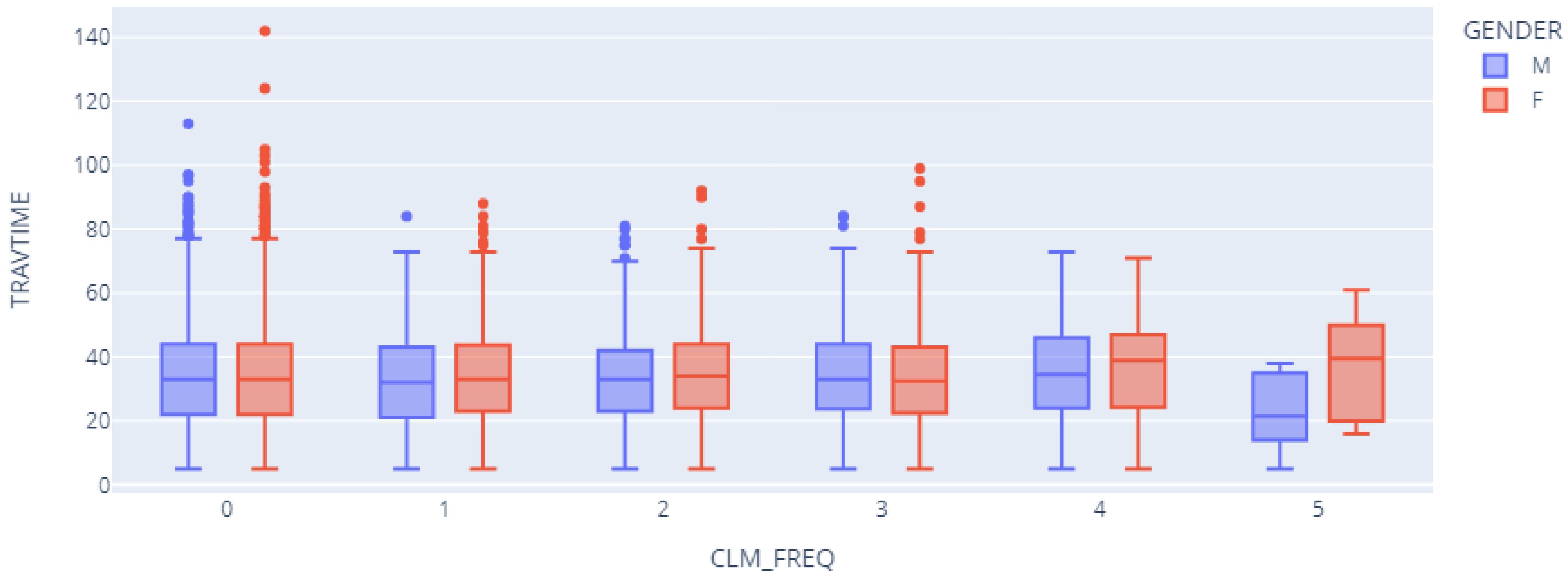
PREVIOUS CLAIM FREQUENCY



**THE GROUP OF CUSTOMERS WHO
HAVE A HISTORY OF USING
INSURANCE HAVE A MUCH HIGHER
LIKELIHOOD OF CONTINUING TO
USE IT IN THE FUTURE COMPARED
TO THOSE WHO HAVE NEVER USED
IT BEFORE.**

PREVIOUS CLAIM FREQUENCY

Box Plot about Time to work with total claim in past by gender

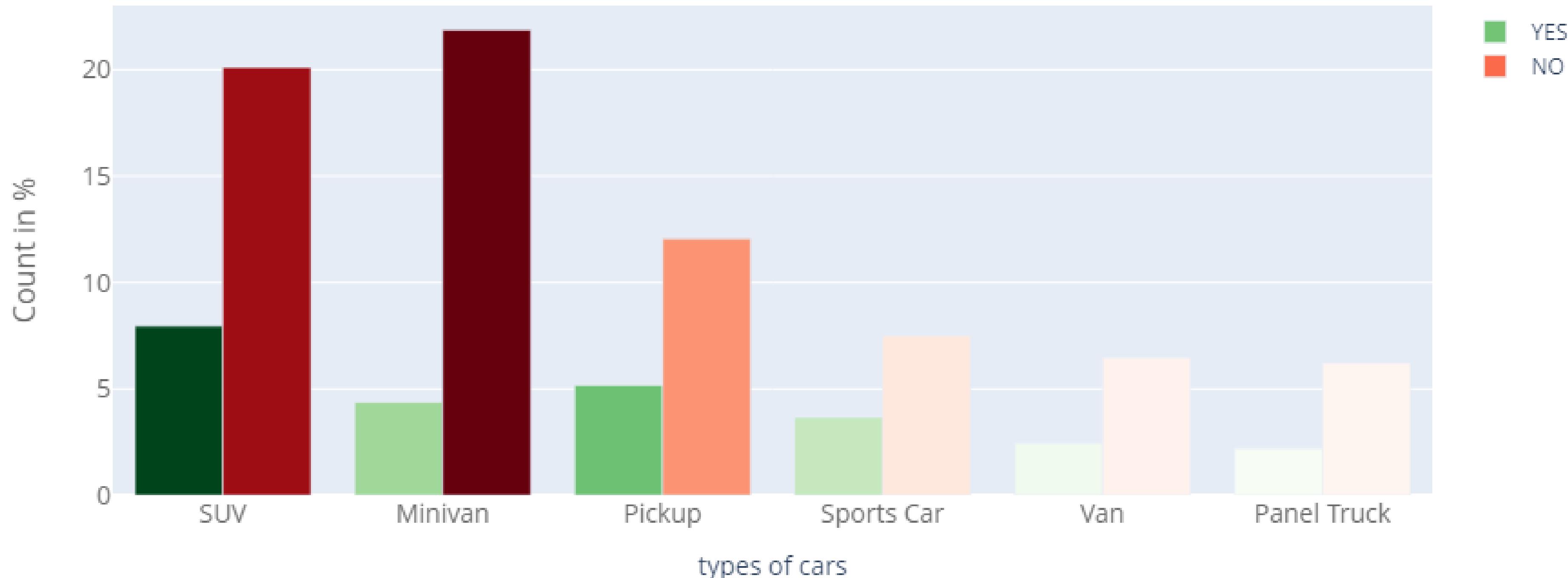


VEHICLE



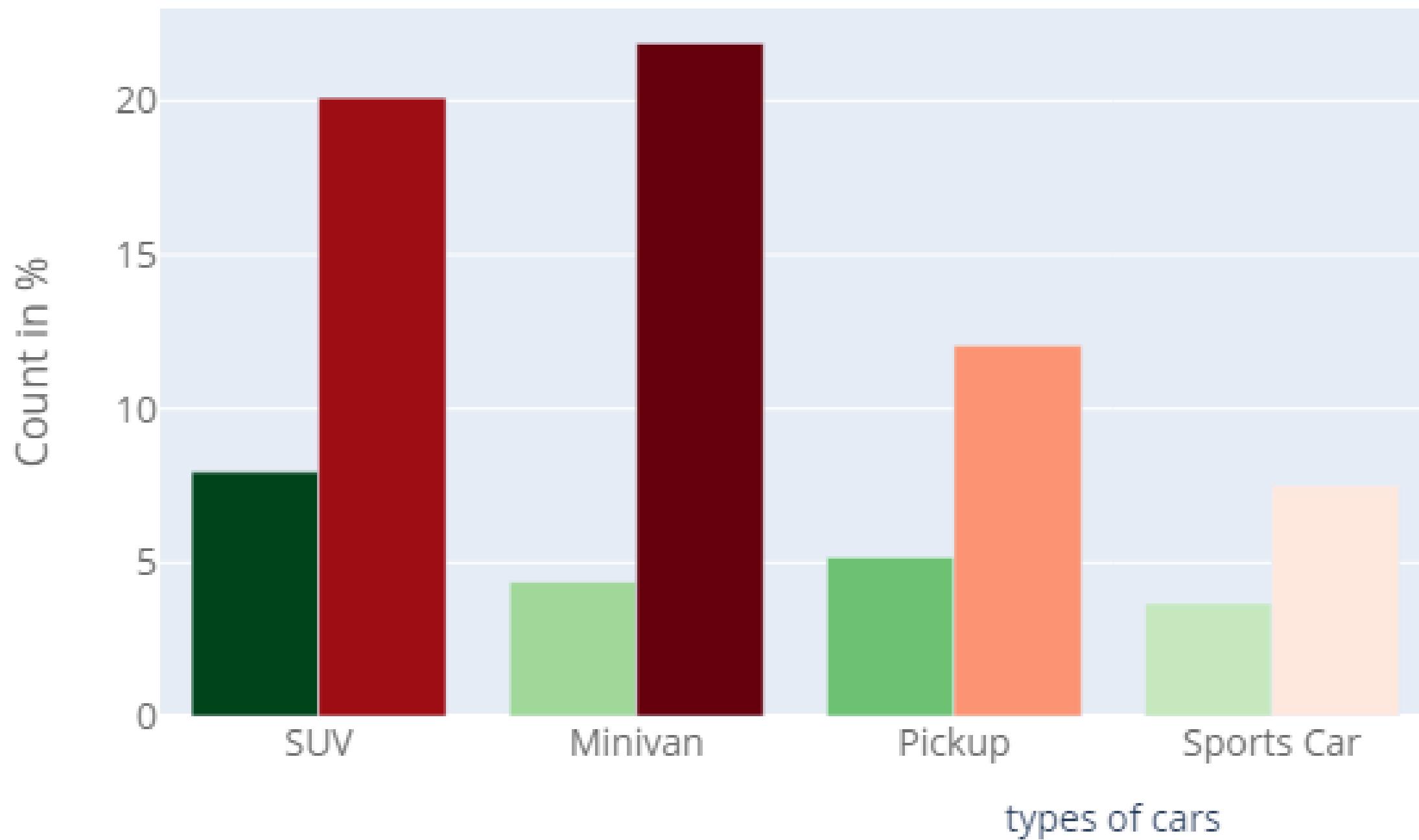
TYPE OF CARS

For which types of car higher applicant's applied for insurance in terms of insurance used or not %



TYPE OF CARS

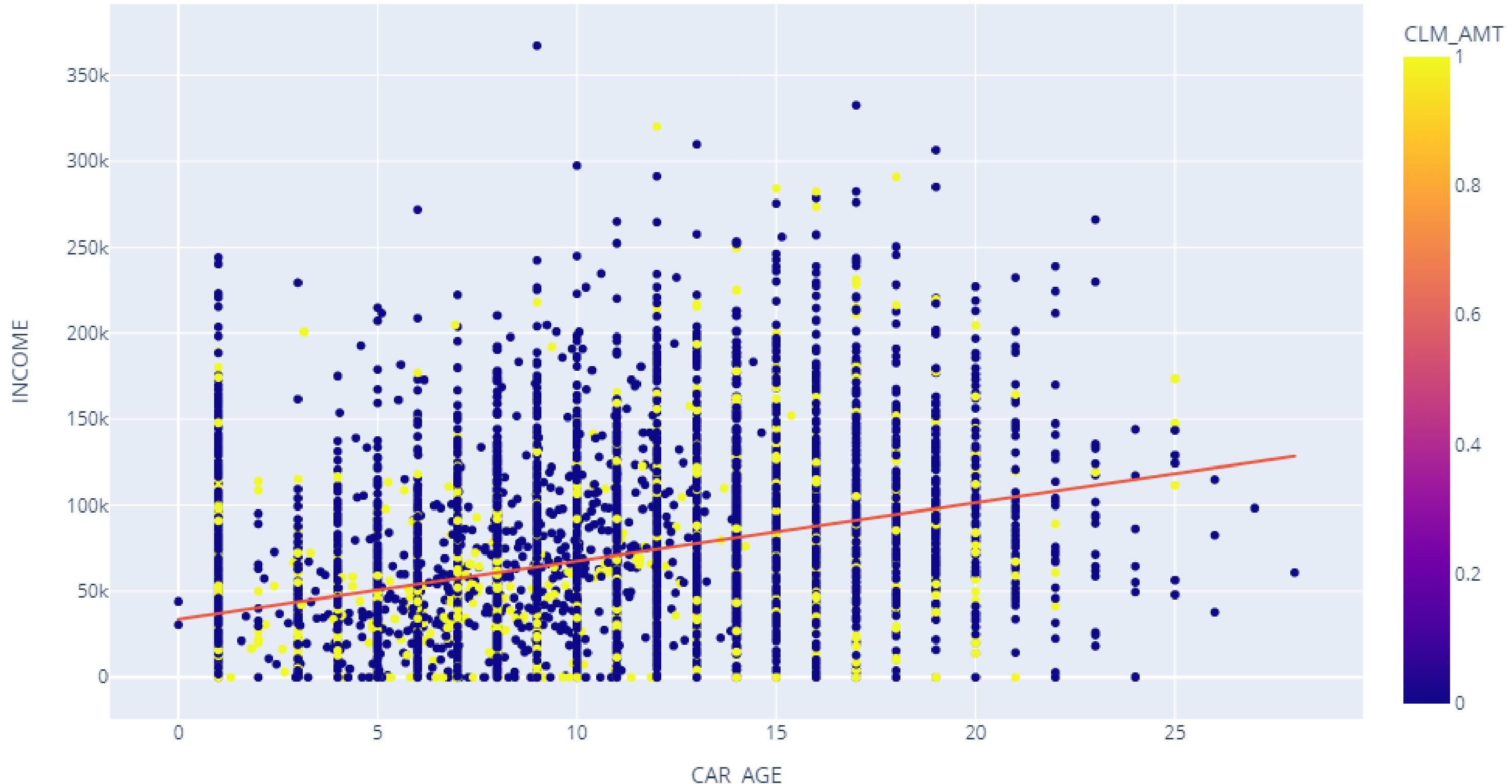
For which types of car higher applicant's applied for insurance in te



7.9%

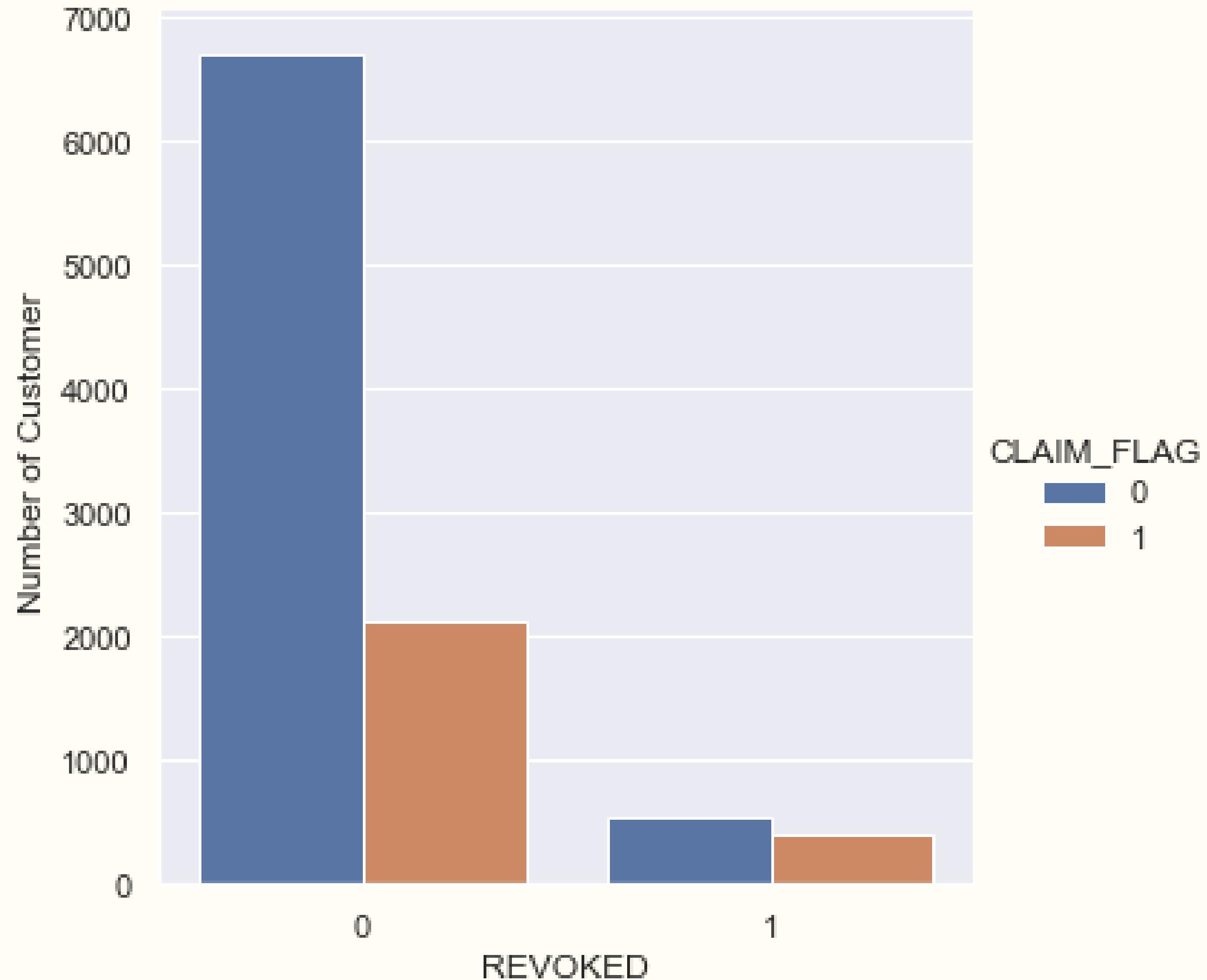
**THE GROUP OF
CUSTOMERS USING SUV
CARS HAS THE HIGHEST
BUYING RATE**

CAR AGE



**THE OLDER THE
CAR, THE LOWER
THE INSURANCE
USAGE RATE**

REVOKE



50-50

THE GROUP OF CUSTOMERS
WHO HAVE HAD THEIR
LICENSE REVOKED IN THE
PAST TEND TO USE
INSURANCE IN THE FUTURE

MACHINE LEARNING





CONTENT

**HANDLE
IMBALANCE
DATA**

**FEATURE
ENGINEERING**

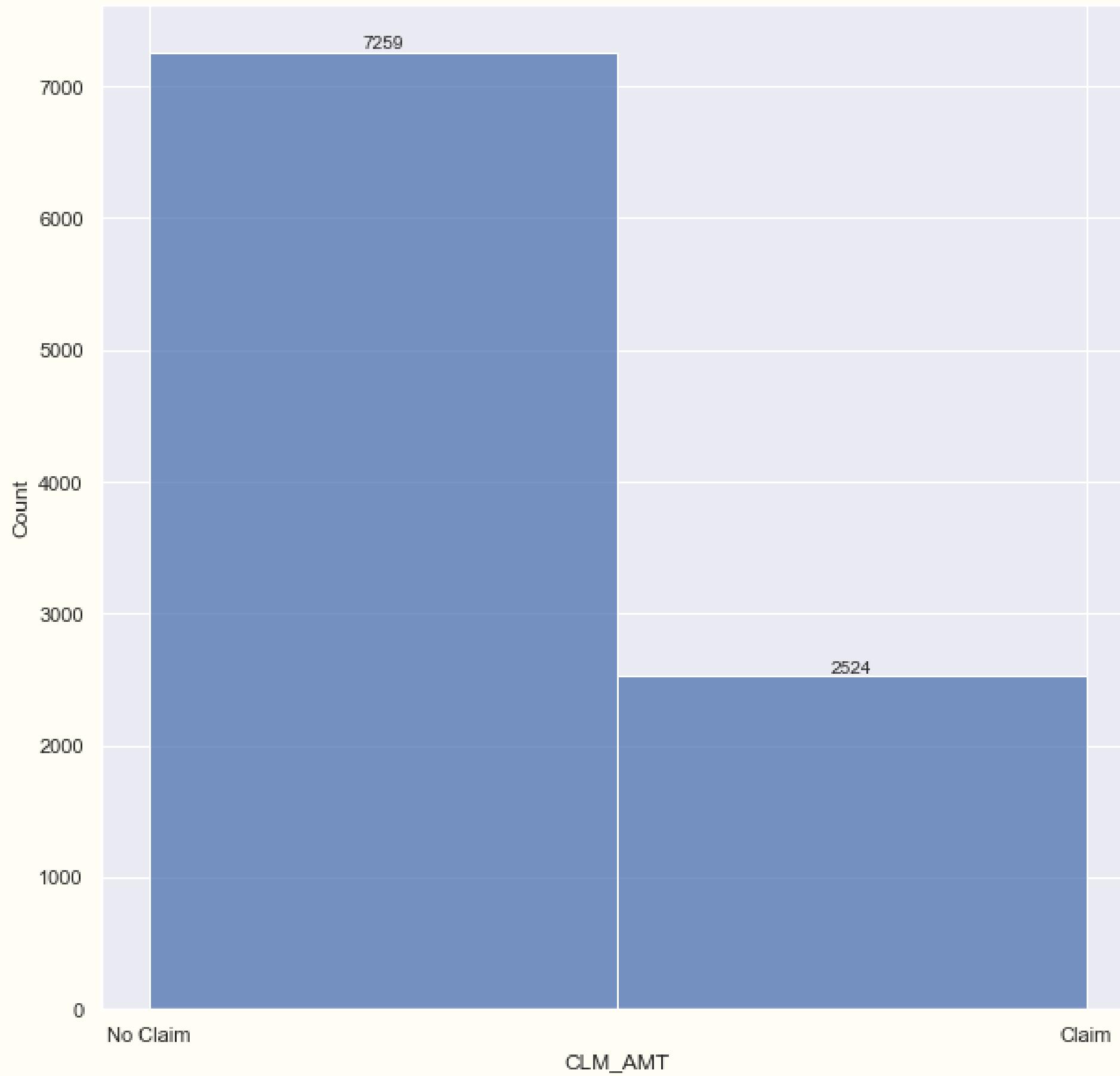
**MODEL
TRAINING**

FEATURE ENGINEERING

2-VALUE CATEGORY

3 OR MORE VALUES CATEGORY

HANDLE IMBALANCE DATA

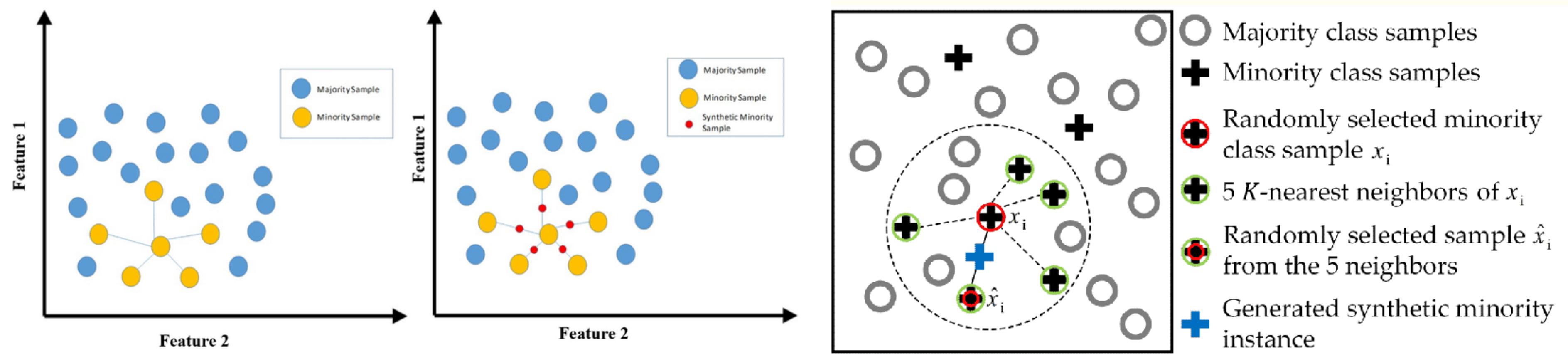


74.2%

THE RATE OF
CUSTOMERS WHO WILL
NOT USE CAR
INSURANCE IN THE
FUTURE

HANDLE IMBALANCE DATA

SMOTE



ALGORITHM

XGBOOST

**LOGISTIC
REGRESSION**

**RANDOM
FOREST**



ALGORITHM

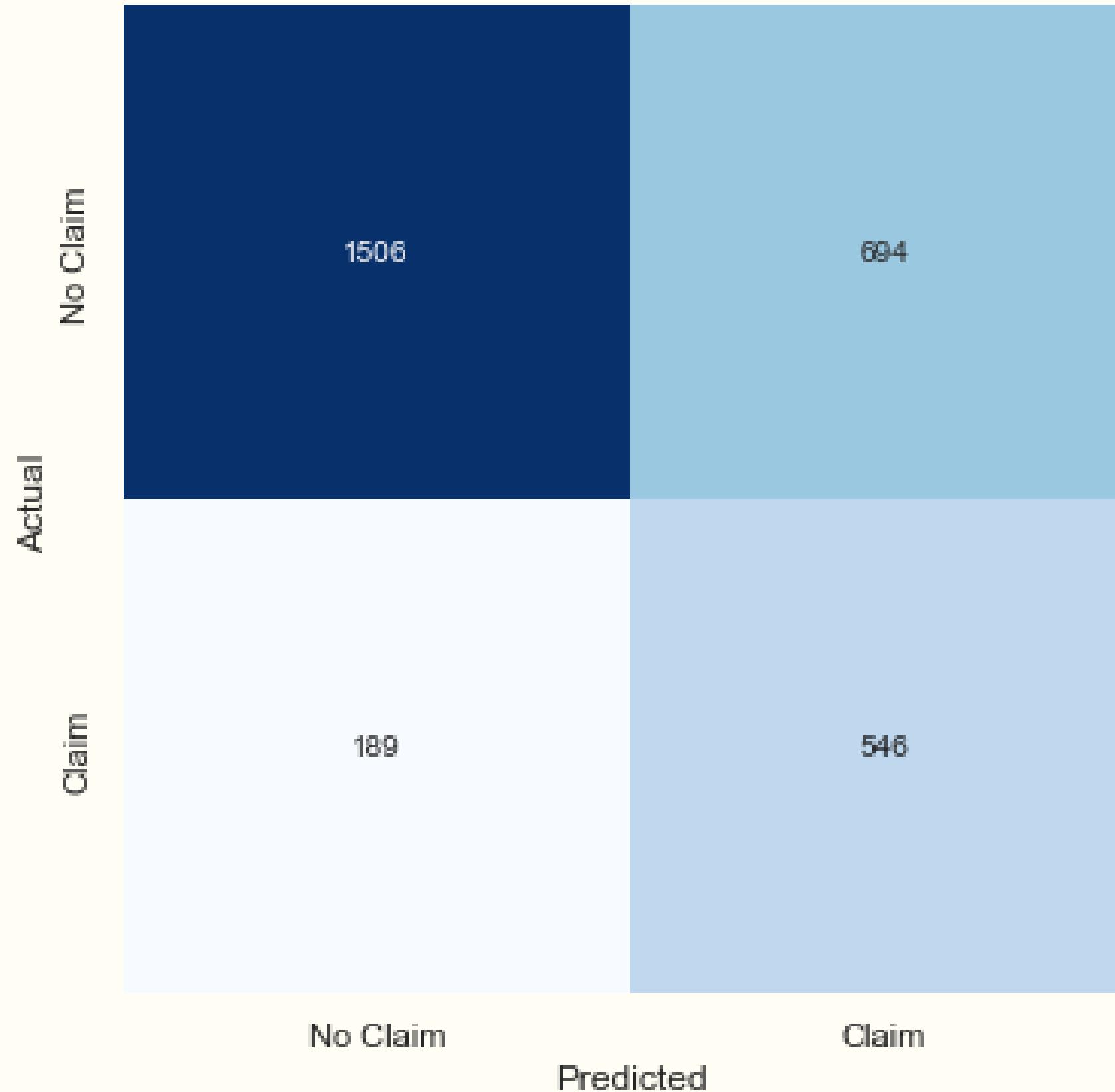
ADA BOOST

DECISION TREE

**GRADIENT
BOOSTING**

LOGISTIC REGRESSION

Confusion Matrix
Cohen's kappa=0.348



ACCURACY: 0.699
PRECISION: 0.44
RECALL: 0.743
F1 SCORE: 0.553

DECISION TREE

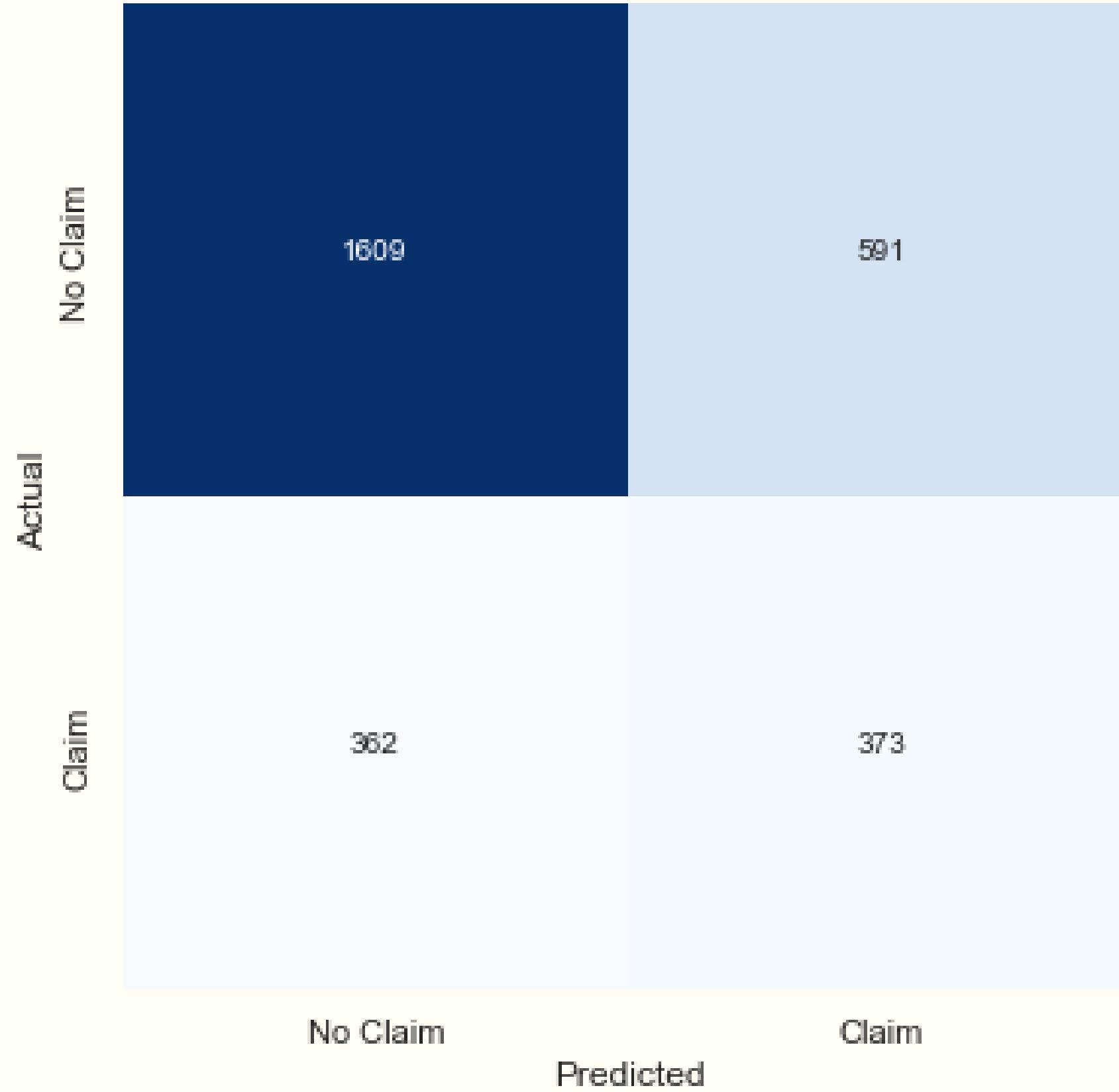
Confusion Matrix
Cohen's kappa=0.216



ACCURACY: 0.675
PRECISION: 0.387
RECALL: 0.507
F1 SCORE: 0.439

XGBOOST

Confusion Matrix
Cohen's kappa=0.216



ACCURACY: 0.734
PRECISION: 0.478
RECALL: 0.653
F1 SCORE: 0.552

RANDOM FOREST

Confusion Matrix
Cohen's kappa=0.365



ACCURACY: 0.738
PRECISION: 0.482
RECALL: 0.626
F1 SCORE: 0.545

ADA BOOST

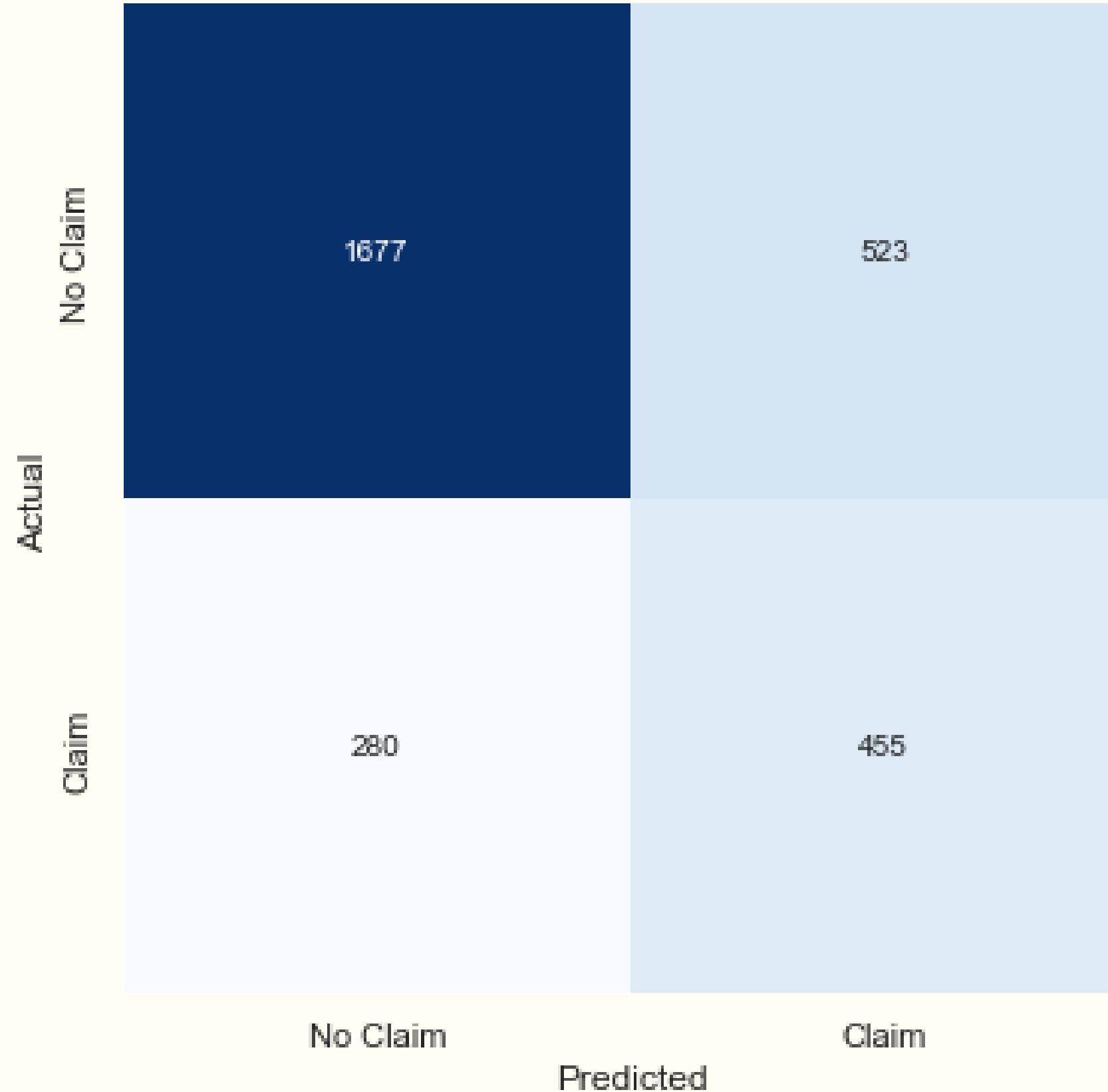
Confusion Matrix
Cohen's kappa=0.328

		Predicted
Actual	No Claim	Claim
	No Claim	Claim
No Claim	1698	502
Claim	305	430

ACCURACY: 0.725
PRECISION: 0.461
RECALL: 0.585
F1 SCORE: 0.516

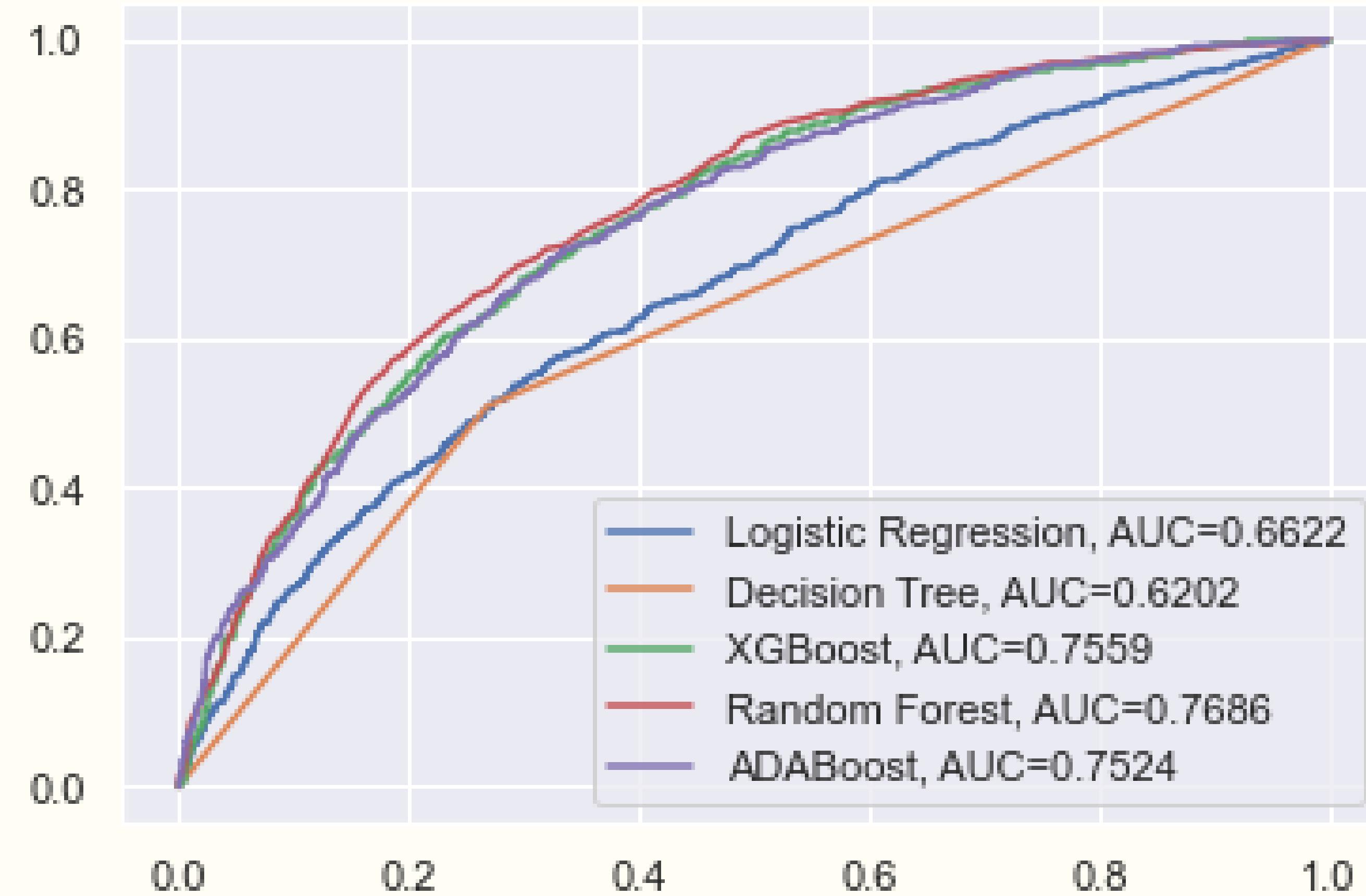
GRADIENT BOOSTING

Confusion Matrix
Cohen's kappa=0.344



ACCURACY: 0.726
PRECISION: 0.465
RECALL: 0.619
F1 SCORE: 0.531

AUC AND ROC

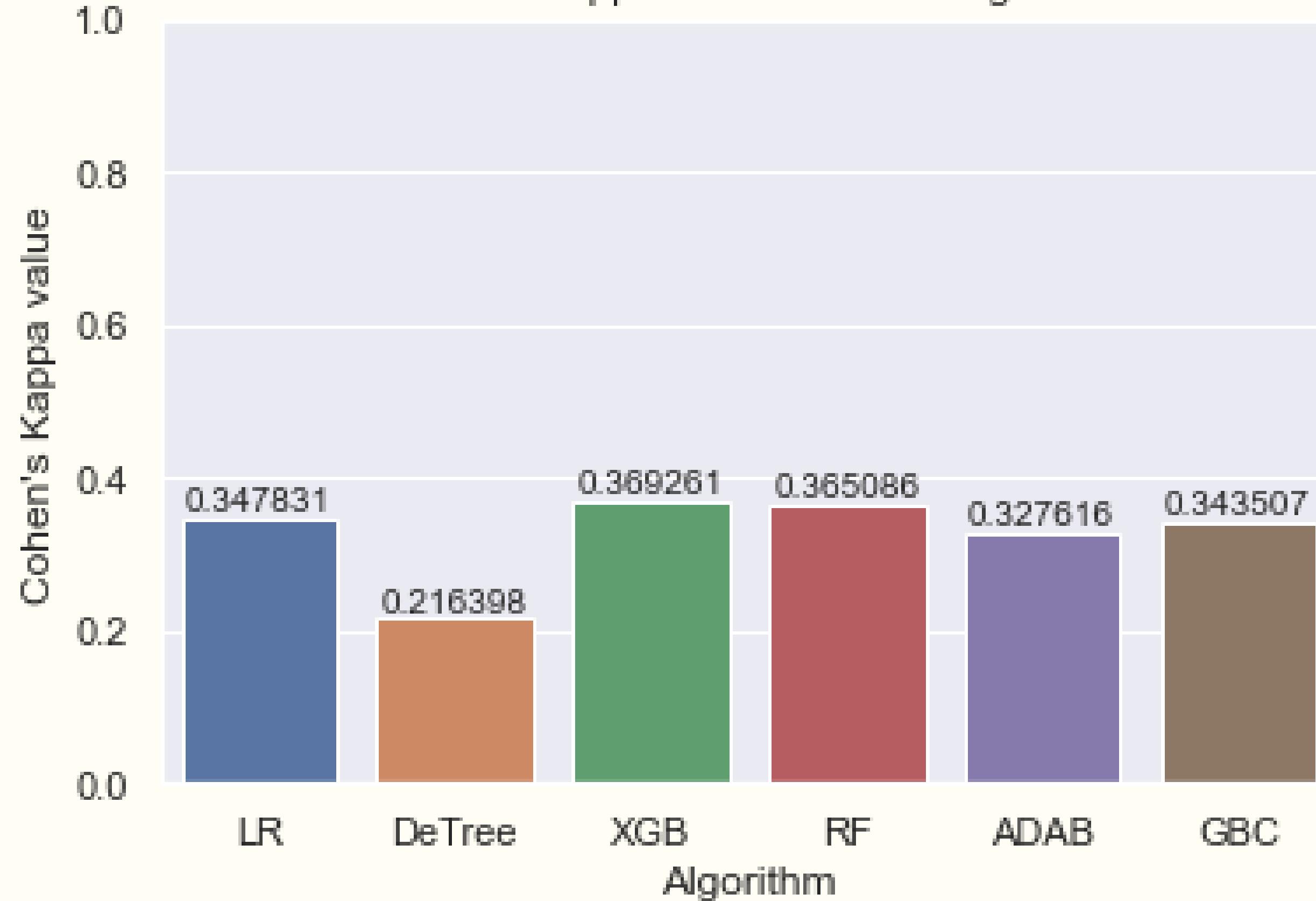


0.7686

THE AUC SCORE OF
RANDOM FOREST
ALGORITHM

COHEN'S KAPPA

Cohen's Kappa Value in Some Algorithm

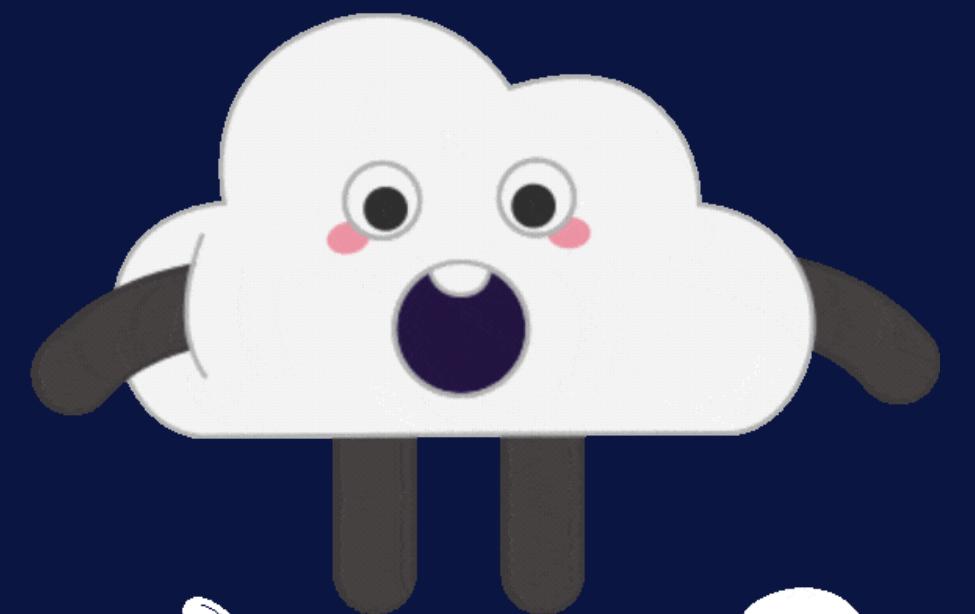


0.369

**COHEN'S KAPPA SCORE
OF XGBOOST
ALGORITHM**

THE END

BYE



THE END

