VEHICLE INSURANCE CLAIM FRAUD DETECTION

Hannah Lal
November 2023

INSURANCE FRAUD IS A PREVALENT AND COSTLY CHALLENGE FOR BOTH POLICYHOLDERS AND INSURANCE COMPANIES

FRAUDULENT ACTIVITIES MAY ENCOMPASS:

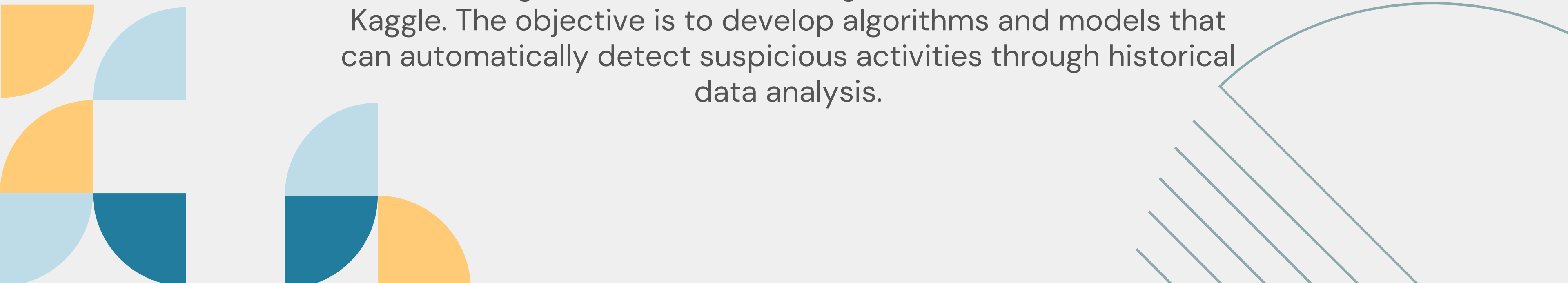FALSE CLAIMS
MISREPRESENTATION OF INFORMATION
ORGANIZED FRAUD SCHEMES

PROJECT INTRODUCTION

Detecting and preventing these deceptive activities is crucial for minimizing financial losses for issuers as well as safeguarding policyholders. Given its significant impact, insurance fraud detection has emerged as a prominent research area in data science and machine learning. This capstone project aims to concentrate on identifying and preventing fraudulent or misleading insurance claims using the dataset available at Kaggle. The objective is to develop algorithms and models that can automatically detect suspicious activities through historical data analysis.

1 - DEFINE SUSPICIOUS

What is suspicious behavior in the world of insurance?

2 - IDENTIFY PATTERNS

Spot the patterns and anomalies in real world data that raises suspicion in insurance claims.

3 - PREVENT FRAUDS

Use machine learning methods, combined with some human monitored methods (mostly to prepare accurate data, having the least false positives) to prevent fraudulent activities.
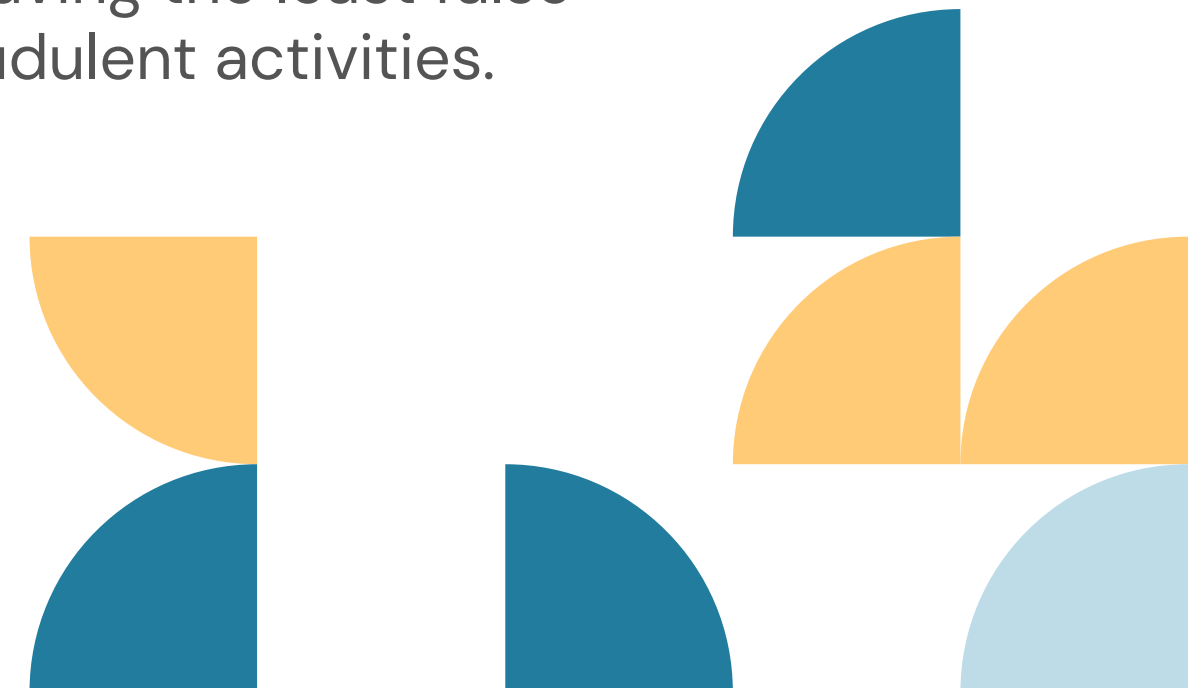
BENEFITS OF FRAUD DETECTION

Stakeholders involved are

POLICYHOLDERS

INSURANCE COMPANIES

FINANCIAL INSTITUTIONS

POLICYHOLDERS STAND TO GAIN HEIGHTENED SECURITY AND PROTECTION AGAINST FRAUDULENT CLAIMS, FORTIFYING THEIR CONFIDENCE IN THE INSURANCE PROCESS.

INSURANCE COMPANIES AND FINANCIAL INSTITUTIONS REAP THE BENEFITS OF DIMINISHED FINANCIAL LOSSES RESULTING FROM FRAUDULENT ACTIVITIES, THEREBY BOLSTERING CUSTOMER TRUST AND ELEVATING THEIR STANDING AND REPUTATION WITHIN THE INDUSTRY

DATASET OVERVIEW
KAGGLE

DATE OF ACCIDENT (BROKEN INTO YEAR, MONTH, WEEK OF MONTH, DAY OF WEEK)

DATE OF CLAIM (SAME AS ACCIDENT DATE)   AGE OF VEHICLE OWENER

MAKE OF CAR   AGE OF CAR   VEHICLE PRICE

ACCIDENT AREA   SEX   MARITAL STATUS

FAULT   POLICY TYPE (COLLISION. LIABILITY, ALL PERILS)

VEHICLE CATEGORY (SEDAN, SPORT, UTILITY)   DEDUCTIBLE

DRIVARE RATING   AGE OF POLICY HOLDER   POLICE REPORT FILED

WITNESS PRESENT   NUMBER OF PAST CLAIMS   AGENT TYPE

NUMBER OF CARS   NUMBER OF SUPPLIMENTS   BASE POLICY

ADDRESS CHANGE CLAIM   DAYS POLICY CLAIM   DAYS POLICY ACCIDENT

FRAUDFOUND_P (FAULT)

FEATURE SET

TARGET VERIABLE

EDA

Progress of FraudFound_P over accident time
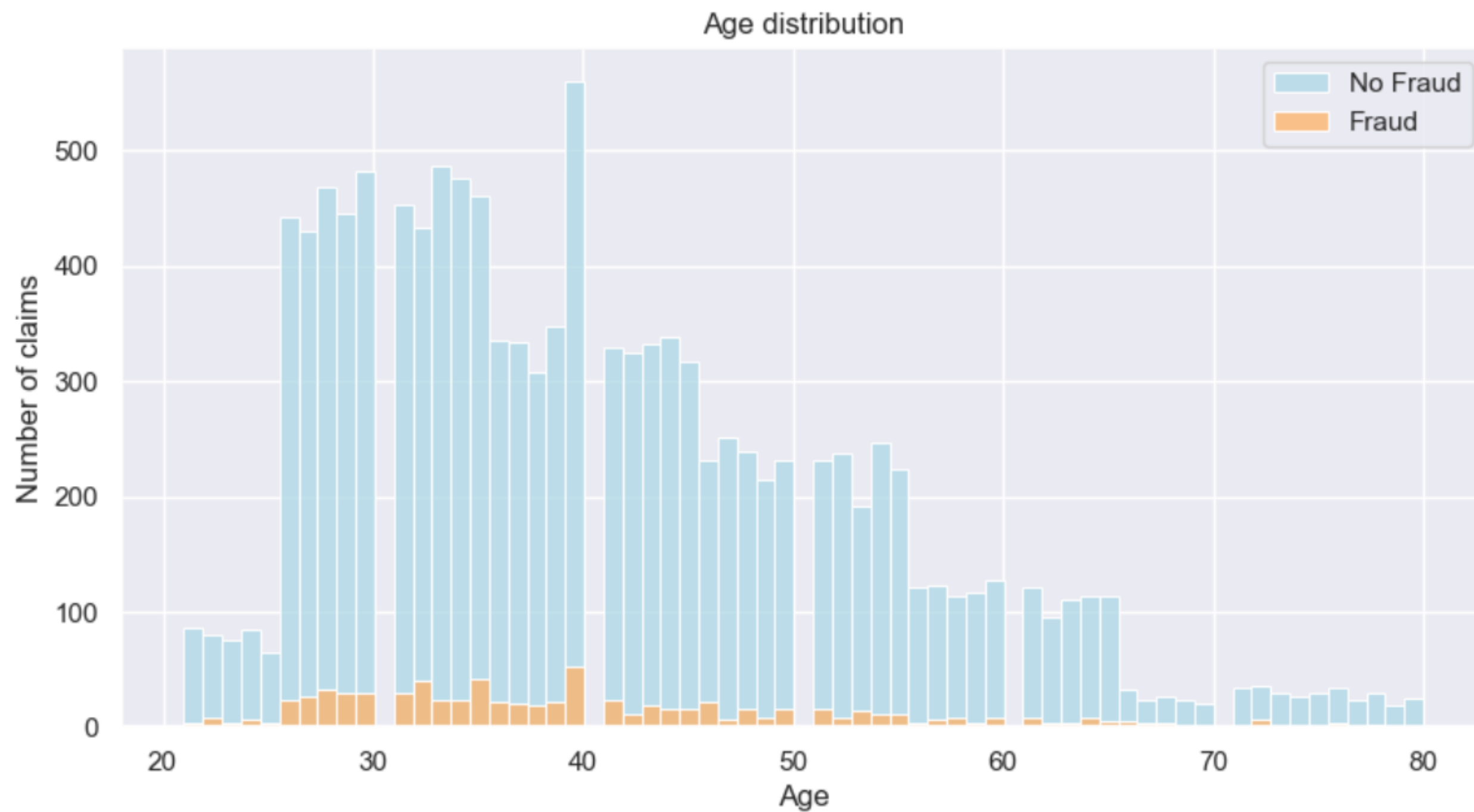

FraudFound_P Over Time
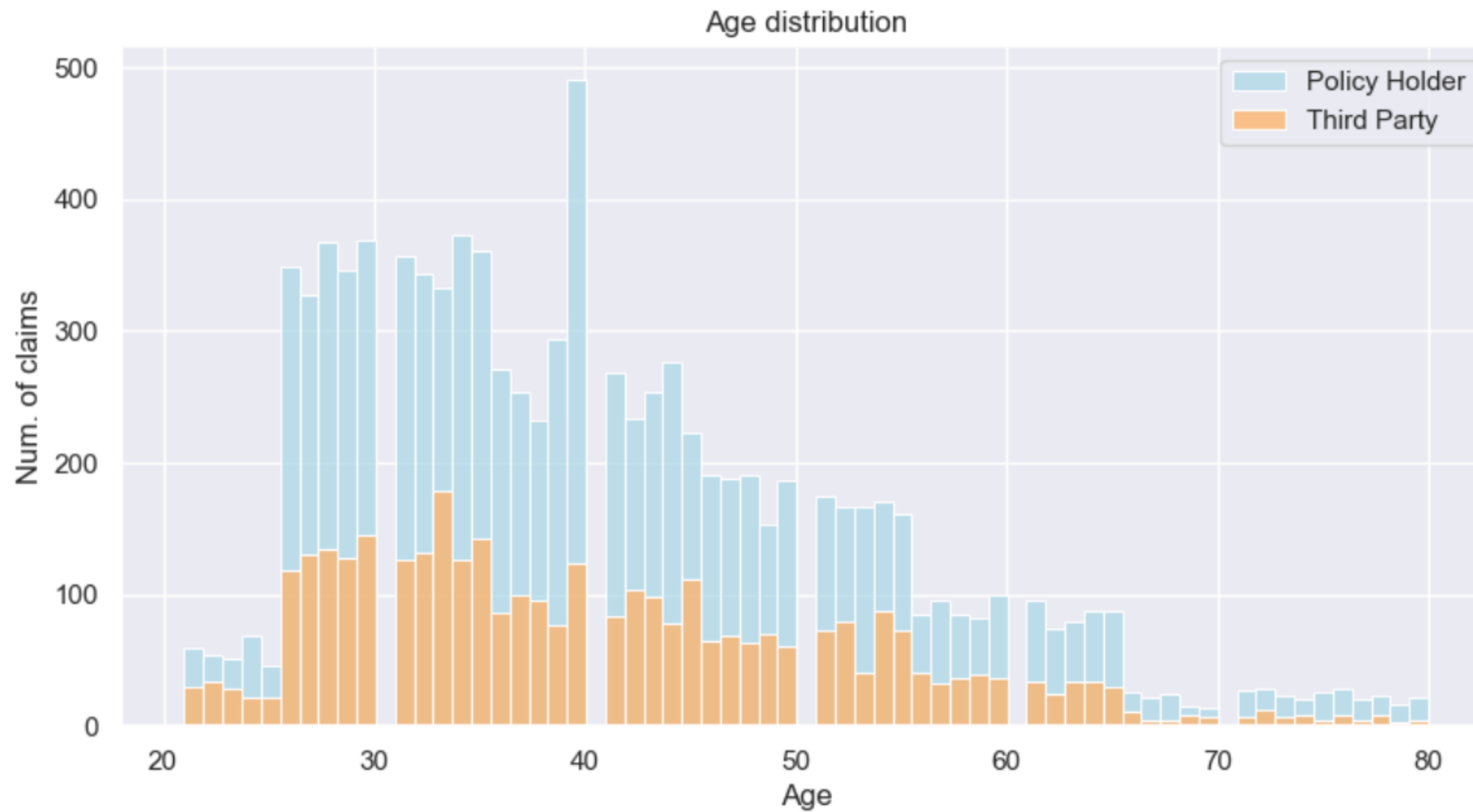
EDA
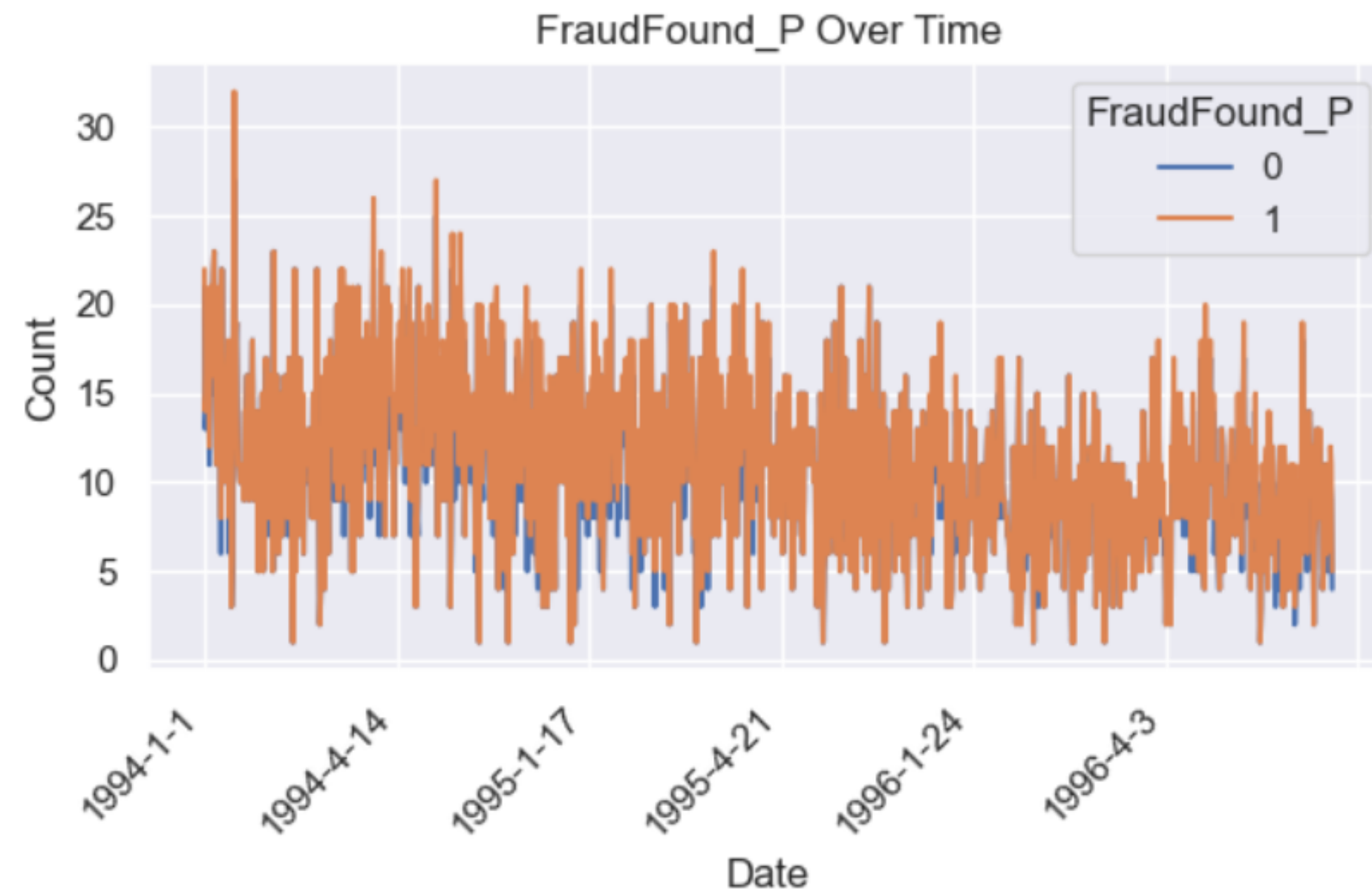
Progress of FraudFound_P over insurance claim time

EDA
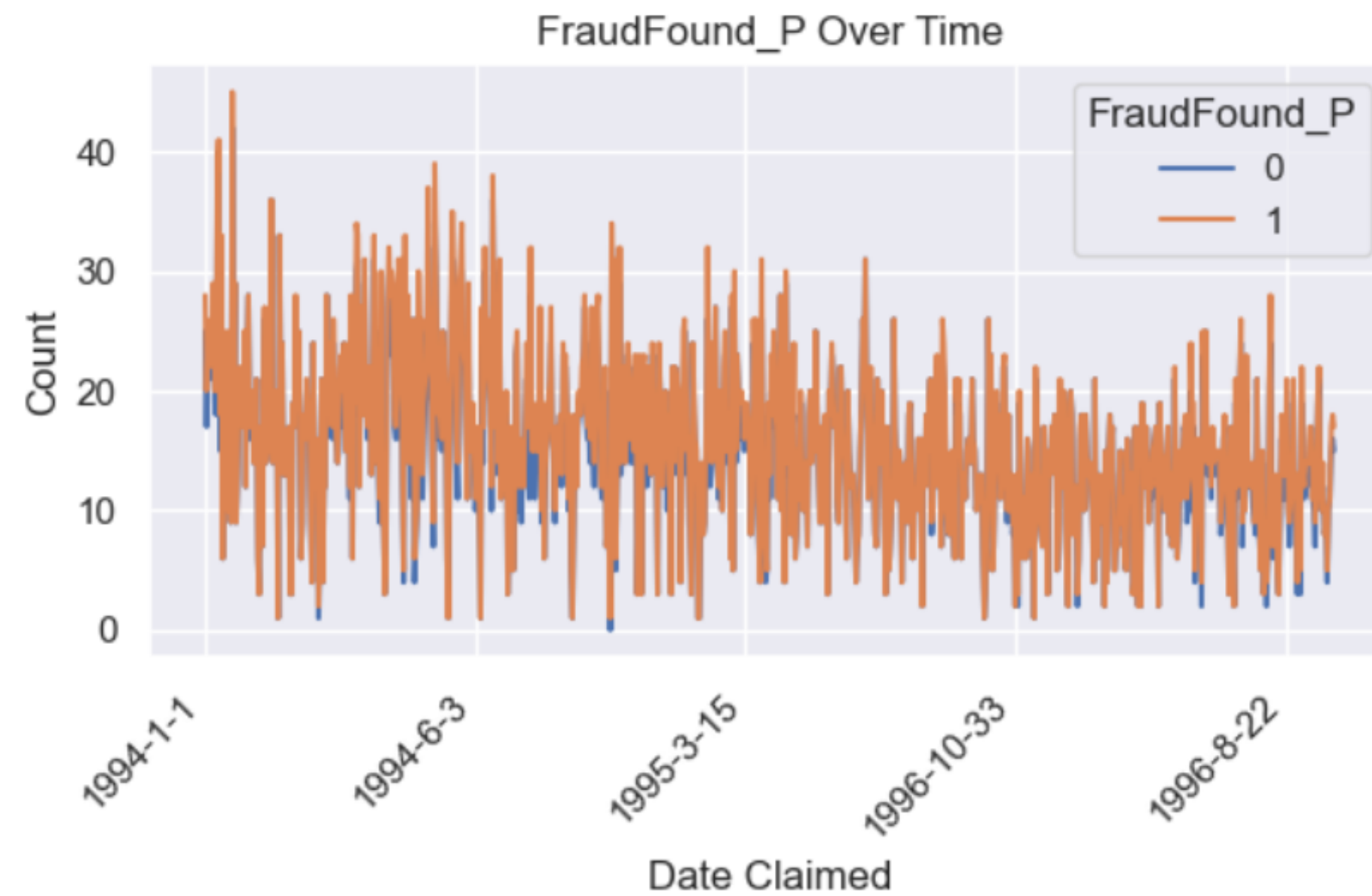
Relationship between Age and number of claims made for each category in our target variable (Fraud Found, and Fraud not found)

Age distribution

EDA

Relationship between Age and number of claims made for each category in 'Fault' variable (Policy holder, and third party)
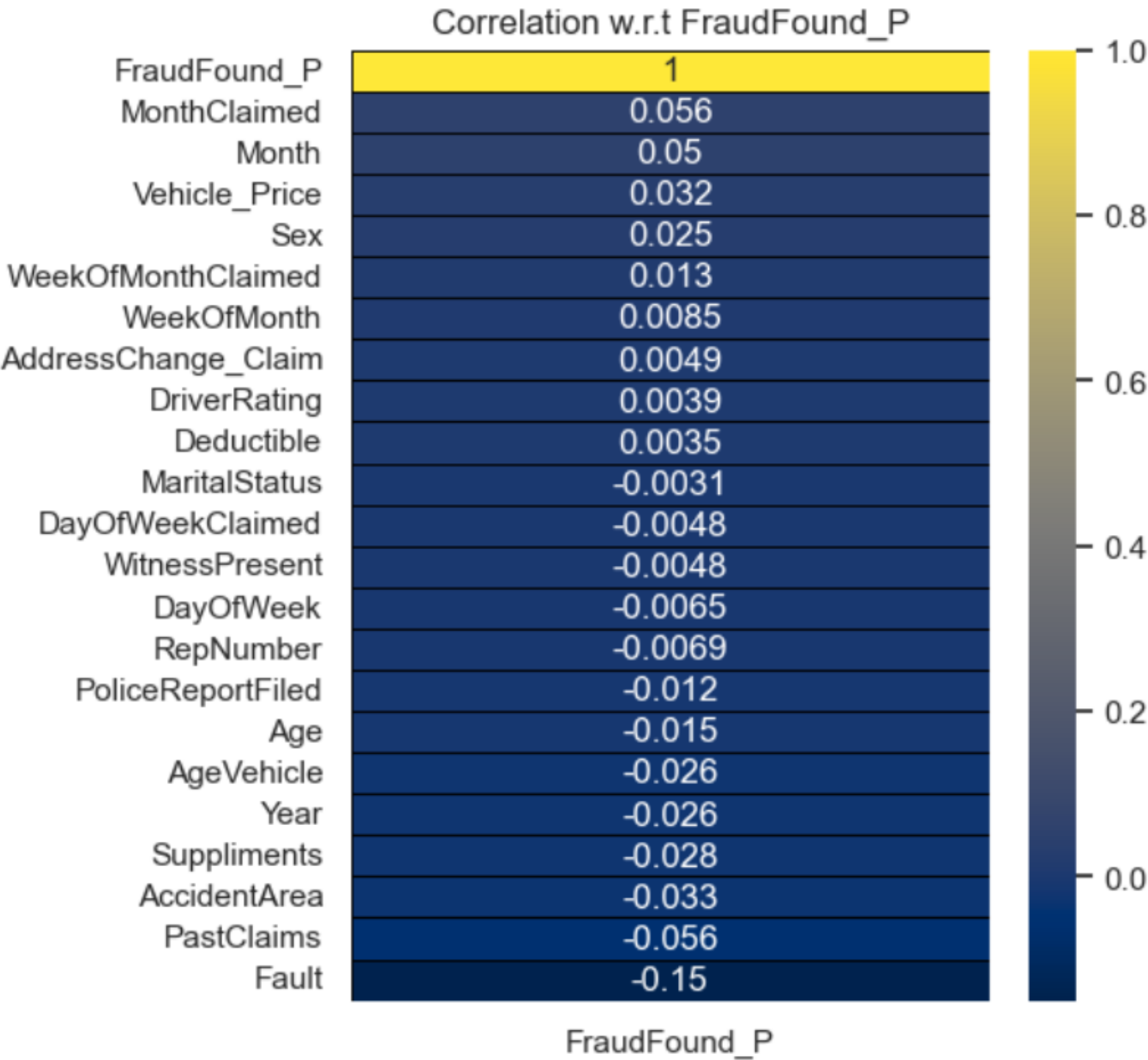
EDA

Correlation of feature set:

EDA

Correlation of feature set with respect to the target variable FraudFound_P:

Correlation w.r.t FraudFound_P

Feature	Correlation
FraudFound_P	1
MonthClaimed	0.056
Month	0.05
Vehicle_Price	0.032
Sex	0.025
WeekOfMonthClaimed	0.013
WeekOfMonth	0.0085
AddressChange_Claim	0.0049
DriverRating	0.0039
Deductible	0.0035
MaritalStatus	-0.0031
DayOfWeekClaimed	-0.0048
WitnessPresent	-0.0048
DayOfWeek	-0.0065
RepNumber	-0.0069
PoliceReportFiled	-0.012
Age	-0.015
AgeVehicle	-0.026
Year	-0.026
Suppliments	-0.028
AccidentArea	-0.033
PastClaims	-0.056
Fault	-0.15

FraudFound_P

EDA

Correlation of feature set with respect to Fault Variable:

Correlation w.r.t Fault

	Fault
Fault	1
WitnessPresent	0.058
PoliceReportFiled	0.026
AgeVehicle	0.019
Year	0.016
Month	0.014
AccidentArea	0.012
Suppliments	0.01
MonthClaimed	0.0098
AddressChange_Claim	0.0073
Sex	0.0042
WeekOfMonthClaimed	0.0031
Deductible	-0.0051
RepNumber	-0.0053
DriverRating	-0.0061
DayOfWeekClaimed	-0.0083
Age	-0.0085
DayOfWeek	-0.014
MaritalStatus	-0.014
WeekOfMonth	-0.019
Vehicle_Price	-0.039
PastClaims	-0.13
FraudFound_P	-0.15

T VALUES AND P VALUES

PREDICTORS WITH HIGHER CORRELATION

MONTH
MONTH CLAIMED
YEAR
PAST CLAIMS
(FOR FRAUDFOUND_P)

WITNESS PRESENT
VEHICLE PRICE
PAST CLAIMS
(FOR FAULT)

statistic (FraudFound_P vs predictors)

statistic (Fault vs predictors)

SUMMARY REPORT

FraudFound_P:

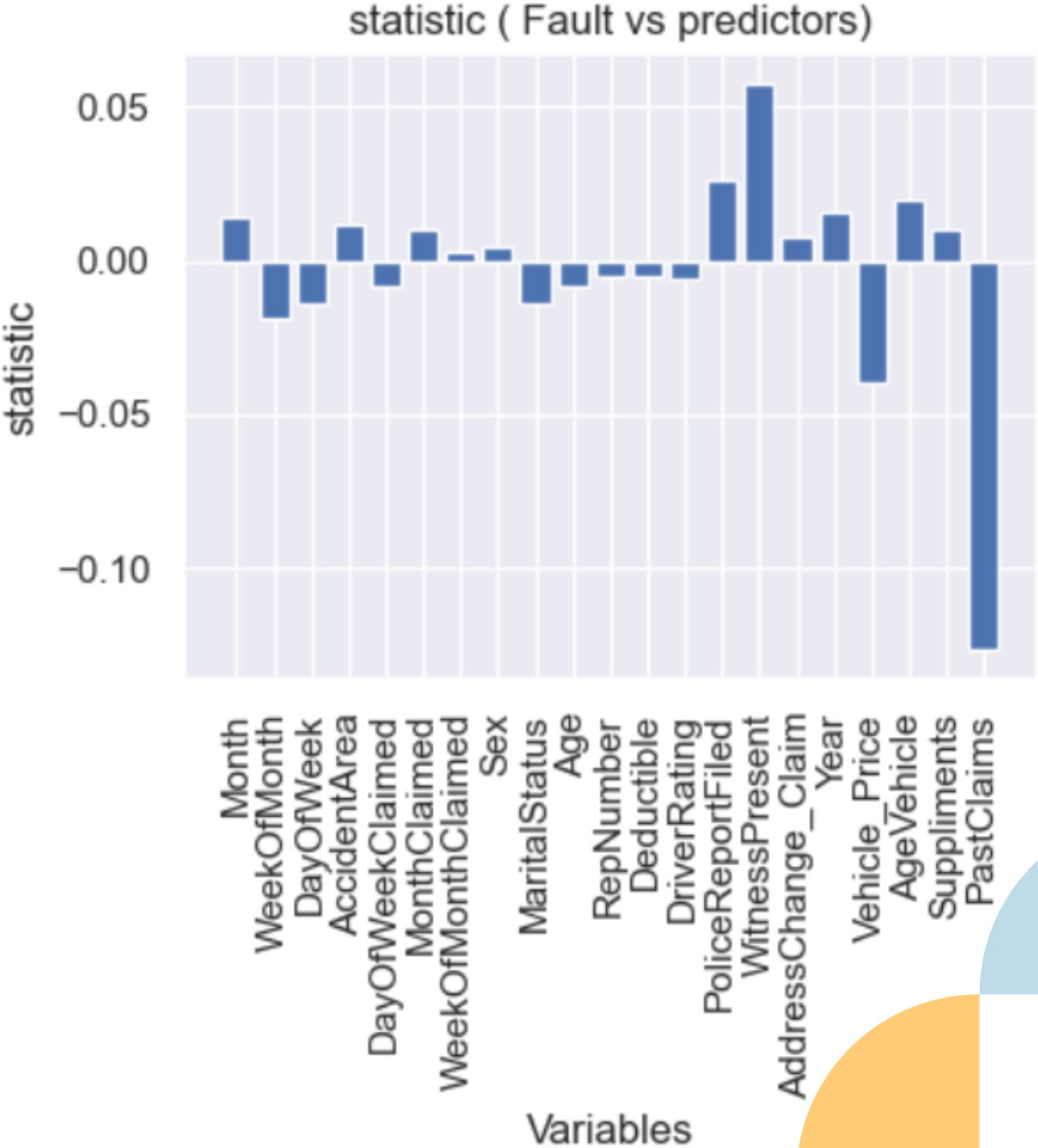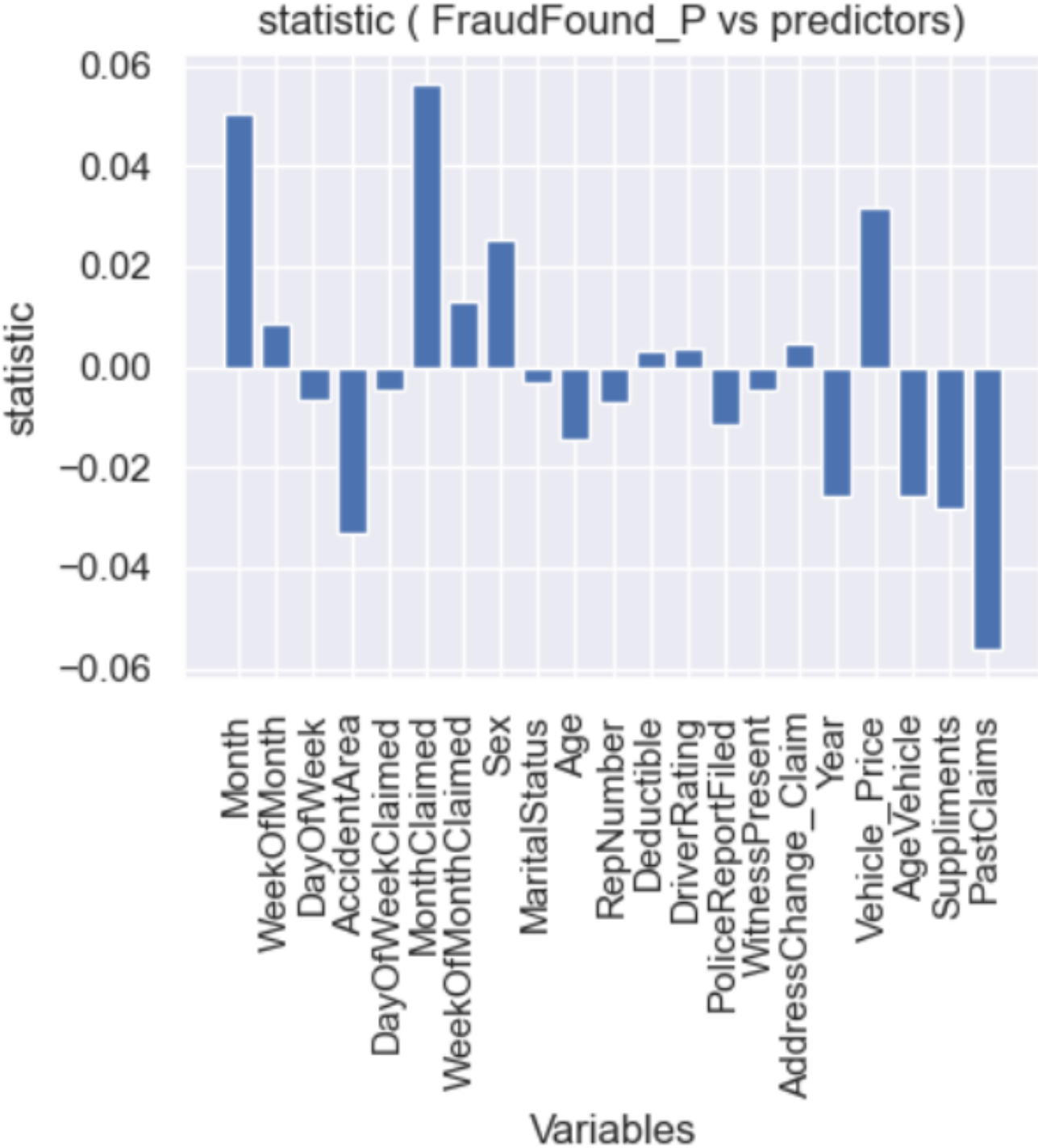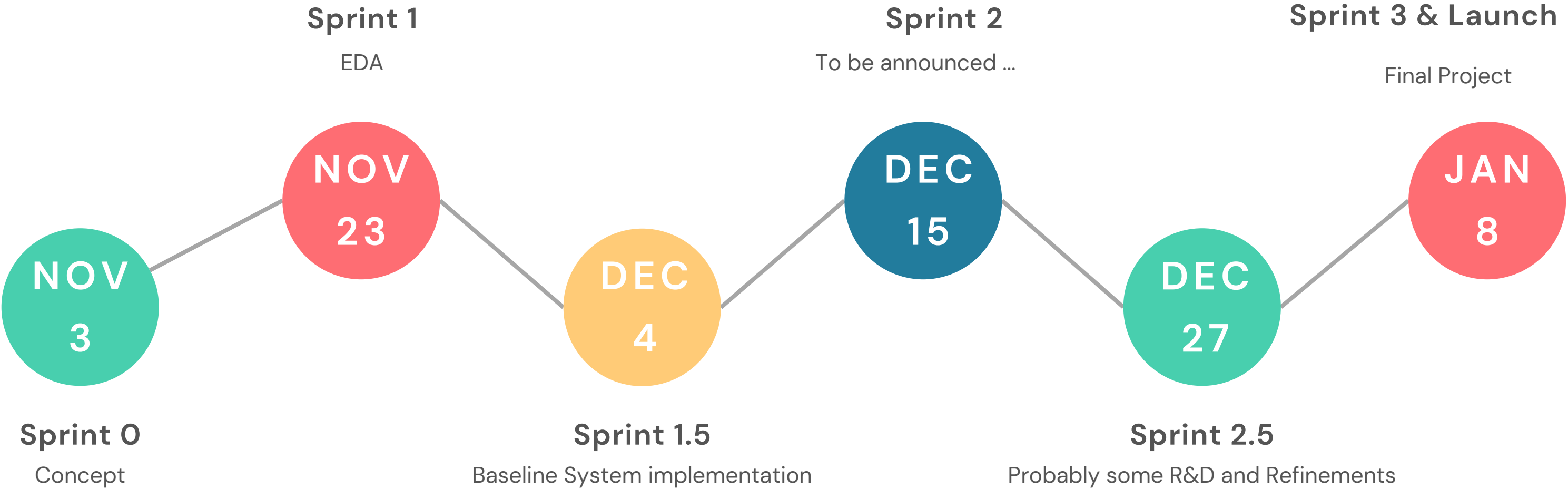- Month, WeekOfMonth, MonthClaimed, WeekOfMonthClaimed, Sex, Dedictible, DriverRating, AddressChange_Claim and Vehicle_Price are positively correlated with our target variable FraudFound_P.
- DayOfWeek, AccidentArea, DayOfWeekClaimed, MaritalStatud, Age, RepNumber, PoliceReportFiled, WitnessPresent, Year, Age of Vehicle and Number of suppliments are negatively correlated with our target variable FraudFound_P.|

Fault:

- Month, AccidentArea, MonthClaimed, WeekOfMonthClaimed, Sex, PoliceReportFiled, WitnessPresent, AddressChange_Claim, Year, Age of Vehicle and Number of suppliments are positively correlated with Fault variable.
- WeekOfMonth, DayOfWeek, DayOfWeekClaimed, MaritalStatud, Age, RepNumber, Dedictible, DriverRating, Vehicle_Price and Past Number of Claims are negatively correlated with number of t_registered rentals.

PROJECT TIMELINE

Sprint 1

EDA

Sprint 2

To be announced ...

Sprint 3 & Launch

Final Project

NOV 23

NOV 3

DEC 4

DEC 15

DEC 27

JAN 8

Sprint 0

Concept

Sprint 1.5

Baseline System implementation

Sprint 2.5

Probably some R&D and Refinements

THANK YOU

Questions?