**Unbalanced dataset**

**Data Profiling**

***Unique Values***

*insured\_relationship: ['wife' 'husband' 'not-in-family' 'unmarried' 'own-child' 'other-relative']*

*Insured\_hobbies: ['camping' 'cross-fit' 'movies' 'reading' 'board-games' 'kayaking' 'polo''base-jumping' 'video-games' 'skydiving' 'chess' 'bungie-jumping'*

*'basketball' 'exercise' 'paintball' 'golf' 'sleeping' 'dancing' 'hiking'*

*'yachting']*

*insured\_occupation: ['transport-moving' 'machine-op-inspct' 'craft-repair' 'adm-clerical'*

*'sales' 'handlers-cleaners' 'exec-managerial' 'tech-support'*

*'prof-specialty' 'armed-forces' 'farming-fishing' 'protective-serv'*

*'priv-house-serv' 'other-service']*

*insured\_education\_level: ['JD' 'College' 'MD' 'High School' 'PhD' 'Masters' 'Associate']*

*insured\_sex: ['MALE' 'FEMALE']*

*umbrella\_limit: [ 5000000 0 6000000 3000000 4000000 9000000 8000000 7000000*

*10000000 2000000 -1000000]*

*incident\_type: ['Parked Car' 'Multi-vehicle Collision' 'Single Vehicle Collision'*

*'Vehicle Theft']*

*collision\_type: ['?' 'Side Collision' 'Rear Collision' 'Front Collision']*

*incident\_severity: ['Trivial Damage' 'Major Damage' 'Minor Damage' 'Total Loss']*

*authorities\_contacted: ['Police' 'Fire' 'Other' 'Ambulance' nan]*

*incident\_state: ['NY' 'WV' 'VA' 'SC' 'OH' 'NC' 'PA']*

*incident\_city: ['Columbus' 'Northbrook' 'Springfield' 'Riverwood' 'Northbend' 'Arlington'*

*Incident\_location : sample ['3094 Best Lane']*

incident\_hour\_of\_the\_day: [ 5 20 23 3 7 16 12 17 15 11 6 9 8 10 0 13 21 14 4 22 19 18 2 1]

number\_of\_vehicles\_involved: [1 3 4 2]

property\_damage: ['?' 'YES' 'NO']

bodily\_injuries: [0 1 2]

column witnesses: [2 3 1 0]

police\_report\_available: ['NO' '?' 'YES']

Injury\_claim: continuous

property\_claim : continuous with 0

Vehicle\_claim : continuous

Auto\_model: ['F150' 'Wrangler' 'CRV' 'Forrestor' 'RAM' 'Ultima' 'M5' 'Camry' 'Neon'

'MDX' 'RSX' 'TL' 'Civic' 'Corolla' 'Highlander' 'A5' '93' 'Pathfinder'

'3 Series' 'Malibu' 'X5' 'A3' 'X6' 'Maxima' 'E400' 'Legacy' 'Escape'

'Silverado' 'Grand Cherokee' 'Impreza' '92x' 'Passat' 'Jetta' 'ML350'

'Tahoe' '95' 'Accord' 'Fusion' 'C300']

auto\_year: [1997 2010 2011 2007 2009 2013 2003 2000 1996 2008 2014 2001 2004 2006

1995 2002 2012 2005 1999 1998 2015]

fraud\_reported: ['N' 'Y']

Unique values in column \_c39: [nan]

**Missing Data/Questionable data includes;**

*Collision type : ‘?’*

*Authorities\_contacted: ‘Nan’*

*property\_damage : “?”*

*policie\_report \_avalibilty: “?”*

*Umbrella\_limit : -100,000 not conceivable*

*Columns to drop:*

* *Unnamed ).1*
* *Unamed 0*
* *Capital\_loss*
* *\_c39*

*Based on the size of the data we need to experiment with the data, and since ? can add values to our project, thus we need to account during our transformation stage*

***Data distribution graphs***

***Data Transformation:***

*KNN/’?’ for authority\_contected*

*Since all fraud detection are either “N” or “Y” then it is important to consider if the authority was contected, since the authority\_cnotact values are* ['Police' 'Fire' 'Other' 'Ambulance' nan], therefore Nan is important.

Important we considered two options ‘nan’ as it own values and ‘nan’ a wrong imputed that

Based upon technical advise we consduer nan as it own variable,

* Potential knn

For should be imputed using We consider Knn, intuitively it consider all neighbors and we make the assumption nearest neighbor. But we must consider first standardizing/normlising the dataset.

When consider the Knn we need to check the optimal parameter for Knn we consider 3, 5, 10

*For Presentation*

1. *Objectives*

* *Identify key areas where fraudulent claims are most likely to occur*
* *Develop a sol to mitigate these financial losses*
* *Ethics (Gender, race, location ?)*

*2. Audience*

* *A semi-technical marketing manager*

***Data transformation encoding***

*Id : changes to frequency*

*months\_as\_customer : currently the data distribution is not normalli distributed thus transform*

age : transform distribution

Policy\_deductible: just encoder into 0 1 2

policy anulam premium: good

umbrella\_limit : zero inflated

Capital gain: zero inflated

Capital loss: drop since it is inverse of capital loss

incident *hour*of\_the\_day : uniform distribution → keep same

Number of vbehicle: unbalcned discrete

bodily \_injuries: discrete -> unchanged

Witness: discrete -> unchanged

Total claim amount : skewed

Injury claim : skewed

proeprty\_claim : skewed

Vehicle claim: zero-inflated

Distribution auto\_year: leave it

Count of policy\_states: encode nominal

policy \_csl : nominal data transform

Insured\_sex : nominal data transform

insured\_education\_level : encode into 2 levels above undergraduate/below undergraduate

insured \_ocputation + Insured hobbies : high cardinality

Insured\_relationship: nominal

Incident\_type: nominal

collsition : nominal

incident \_serviertty: nominal

Authorities\_contacted : nominal

Incidnet\_state : nominal

Incident\_city: nominal

property \_damage: nominal -> encode

polci\_report\_avalibalie : nominal

auto\_make : nominal

Count of auto\_model: nominal

Ethical considerations

**Embedding ethical risk management into work**

* Did we ethical assessment checkpoints
* Seeking independent or domain-expert advice on assumptions and risks? Yes
* industrial ethical policies
  + GDPR, PIPL, and the HIPAA (Health Insurance Portability and Accountability Act)
    - GDPR grants individuals enhanced privacy rights, including the right to access their data held by insurance companies, the right to request corrections or deletions of inaccurate or outdated data, and the right to object to the processing of their data for certain purposes.

**Ethical considerations in data science: Balancing privacy and utility**

* Algorithmic bais : males vs female , undergraduate vs overgraudate
* Once a fraugilence is found, how many times should it be done for it to be flagged potential damaging?