[**Structure: 2**](#_pd5rhabell4i)

[KEY: 5](#_goclzy5asxm)

[**DATASET 1 CLEANING MAPPING 5**](#_5b7fwmpj8bby)

[Source: 5](#_nsed63dr8ri9)

[Raw Columns: 5](#_7h55bt406q0d)

[Dataset 1 Columns after Cleaning & Enriching: 7](#_4o5vxmojwugp)

[Dataset 1 Column Creations Logic: 8](#_k2fcdiivw1j8)

[**DATASET 2 CLEANING MAPPING 9**](#_1mveemub53z8)

[Source: 9](#_bdng39lqlg2z)

[Raw Columns: 9](#_bzwqg4864ots)

[Dataset 2 Columns after Cleaning & Enriching: 10](#_4ypsktl4i71j)

[Dataset 2 Column Creations Logic: 12](#_zm7nx1do26g)

[**DATASET 3 CLEANING MAPPING 13**](#_63ked9mt3cz)

[Source: 13](#_sqxxyg7i25va)

[Raw Columns: 13](#_lp6fqtszfgha)

[Dataset 3 Columns after Cleaning & Enriching: 14](#_si3wm9ptchax)

[Dataset 3 Column Creations Logic: 14](#_55g2ky6ipgm5)

[**ORGANISATION: 15**](#_52t5ceazshql)

[**Raw Files: 15**](#_tofp4suxc2po)

[**Cleaning: 15**](#_xfjc7ubsgbz)

[**Enriching: 16**](#_z13n98xfa4fz)

[**Synthesising: 16**](#_915mnpw5hqh4)

[**Merging: 17**](#_5fjtv2utvjok)

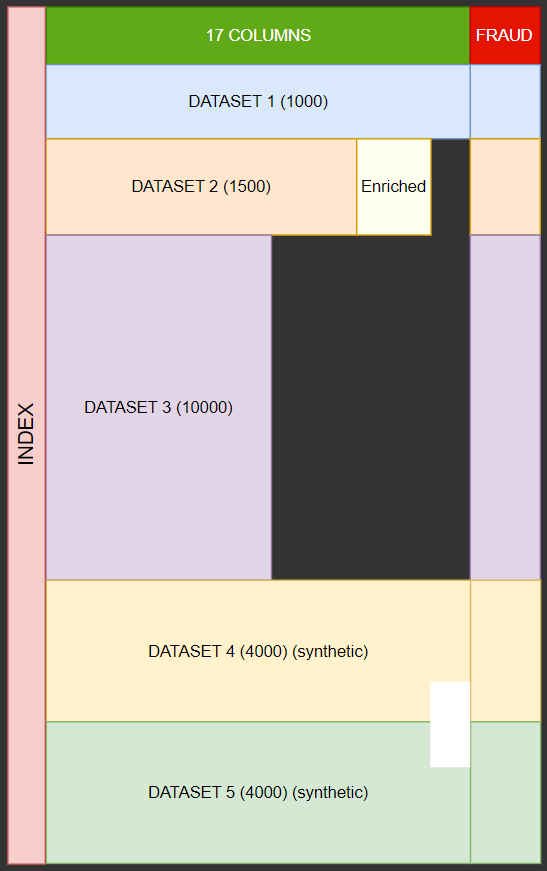
[**JSON: 17**](#_7z3avaqusr6s)

[Statistical Run Results: 18](#_u1fv7d1negar)

[Archive Notes: 20](#_l44mmsh63fmz)

Final Dataset Structure:

# Structure:



Column Names:

* index
* timeAsCustomer
* driverAge
* insuranceAccess
* insurancePremium
* driverGender
* educationLevel
* accidentType
* incidentSeverity
* authoritiesInvolved
* incidentTime
* numVehiclesInvolved
* numBodilyInjuries
* policeReportBool
* totalClaimAmount
* Fraud
* vehicleAge
* driverExperience
* licenseType

THEORY & PLANNING

For dataset 1 it will be **fully complete** with all columns populated and cleaned.

Dataset 2 will also be **made perfect** by enriching it with 3 additional columns based on dataset 1. It will not have a populated Gender column.

Dataset 3 will be a very large **incomplete** but cleaned dataset.

All datasets will contain a **fraud column** which will indicate whether an entry is fraudulent or not.

Additional data (8000 rows) will also be synthesised using 2 data synthesization techniques (4000 rows each).

1 technique used will be **Remapping and Mathematical Sampling**. It basically remaps the data then samples that data to generate rows.

The second technique used for synthesising was **CTGAN** (Conditional Tabular Generative Adversarial Network). It is basically a deep learning method which uses two separate neural networks to generate realistic data.

DATA CLEANING MAPPING

## KEY:

Our Column Name - Source Data Column Name - Source Data Column Key - Column Metadata

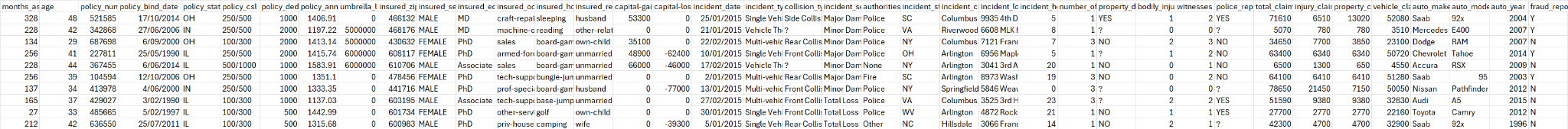
# DATASET 1 CLEANING MAPPING

## Source:

<https://github.com/mwitiderrick/insurancedata/blob/master/insurance_claims.csv>

<https://www.kaggle.com/datasets/sumansuhag/insurance-dataset-csv>

## Raw Columns:

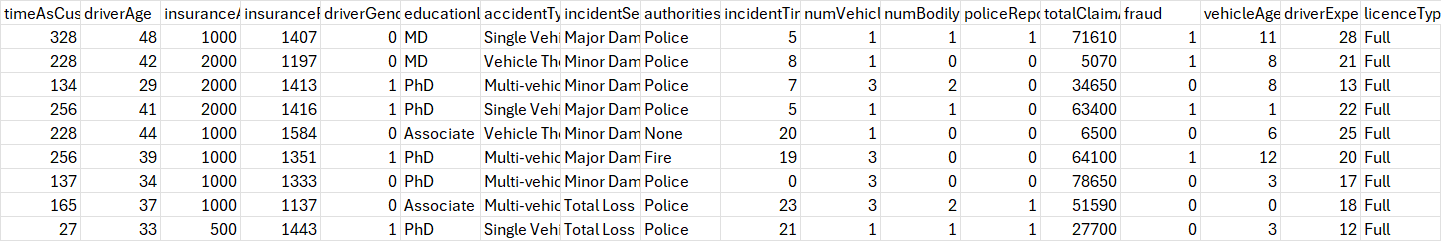


| Column Name | Data Type | Description |
| --- | --- | --- |
| months\_as\_customer | Numerical | Number of months the individual has been a customer |
| age | Numerical | Age of the customer |
| policy\_number | Categorical | Unique identifier for the insurance policy |
| policy\_bind\_date | Date | The date when the policy was created/bound |
| policy\_state | Categorical | The state where the policy was issued |
| policy\_csl | Categorical | Combined Single Limit for the policy |
| policy\_deductable | Numerical | The amount of deductible in the policy |
| policy\_annual\_premium | Numerical | The annual premium paid for the insurance policy |
| umbrella\_limit | Numerical | The umbrella limit in the policy |
| insured\_zip | Categorical | Zip code of the insured person |
| insured\_sex | Categorical | Gender of the insured person |
| insured\_education\_level | Categorical | Education level of the insured person |
| insured\_occupation | Categorical | Occupation of the insured person |
| insured\_hobbies | Categorical | Hobbies of the insured person |
| insured\_relationship | Categorical | Relationship status of the insured person |
| capital-gains | Numerical | Capital gains of the insured person |
| capital-loss | Numerical | Capital loss of the insured person |
| incident\_date | Date | Date of the incident |
| incident\_type | Categorical | Type of incident |
| collision\_type | Categorical | Type of collision |
| incident\_severity | Categorical | Severity of the incident |
| authorities\_contacted | Categorical | Whether authorities were contacted |
| incident\_state | Categorical | State where the incident occurred |
| incident\_city | Categorical | City where the incident occurred |
| incident\_location | Categorical | Specific location of the incident |
| incident\_hour\_of\_the\_day | Numerical | Hour of the day when the incident occurred |
| number\_of\_vehicles\_involved | Numerical | Number of vehicles involved in the incident |
| property\_damage | Categorical | Whether property damage occurred |
| bodily\_injuries | Numerical | Number of bodily injuries |
| witnesses | Numerical | Number of witnesses to the incident |
| police\_report\_available | Boolean | Whether a police report is available |
| total\_claim\_amount | Numerical | The total amount claimed |
| injury\_claim | Numerical | Claim amount for injuries |
| property\_claim | Numerical | Claim amount for property damage |
| vehicle\_claim | Numerical | Claim amount for vehicle damage |
| auto\_make | Categorical | Make of the vehicle involved |
| auto\_model | Categorical | Model of the vehicle involved |
| auto\_year | Numerical | Year of the vehicle involved |
| fraud\_reported | Boolean | Whether fraud was reported |

## 

## Dataset 1 Columns after Cleaning & Enriching:

| NO. | Column Name | Column Letter | Data Type | Description |
| --- | --- | --- | --- | --- |
| 1 | timeAsCustomer | A | Numerical | Number of months the individual has been a customer |
| 2 | driverAge | B | Numerical | Age of the driver |
| 3 | insuranceAccess | C | Numerical | The level of access to insurance |
| 4 | insurancePremium | D | Numerical | The premium amount for the insurance policy |
| 5 | driverGender | E | Boolean | Gender of the driver (0 = Male, 1 = Female) |
| 6 | educationLevel | F | Categorical | Education level of the driver |
| 7 | accidentType | G | Categorical | Type of accident |
| 8 | incidentSeverity | H | Categorical | Severity of the incident |
| 9 | authoritiesInvolved | I | Categorical | Whether authorities were involved |
| 10 | incidentTime | J | Numerical | Time of the incident |
| 11 | numVehiclesInvolved | K | Numerical | Number of vehicles involved in the accident |
| 12 | numBodilyInjuries | L | Numerical | Number of bodily injuries |
| 13 | policeReportBool | M | Boolean | Whether a police report was available (0 = No, 1 = Yes) |
| 14 | totalClaimAmount | N | Numerical | The total claim amount |
| 15 | fraud | O | Boolean | Whether the claim was fraudulent (0 = No, 1 = Yes) |
| 16 | vehicleAge | P | Numerical | Age of the vehicle |
| 17 | driverExperience | Q | Numerical | Number of years of driving experience |
| 18 | licenceType | R | Categorical | Type of licence held by the driver |



## 

## Dataset 1 Column Creations Logic:

**Vehicle Age**

Source:  
Column auto\_year - AL

2015

Formula:

VA = 2015 - auto\_year

**DriverExperience**

Source:

Column age - B

16

R - (Random weighted variable between 0 and 6)

Formula:

DriverExperience = (Age − 16) − R

**LicenceType**

Source: DriverExperience (Generated Column)

Formula:

IF(L2<1, "Learners", IF(AND(L2>=1, L2<3), "Provisional P1", IF(AND(L2>=3, L2<5), "Provisional P2", IF(L2>=5, "Full", ""))))

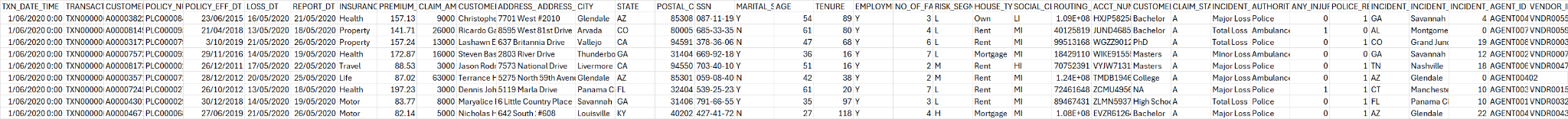
# DATASET 2 CLEANING MAPPING

## Source:

<https://www.kaggle.com/datasets/mastmustu/insurance-claims-fraud-data?select=insurance_data.csv>

**NOTE: FILTER FOR ONLY MOTOR CLAIMS**

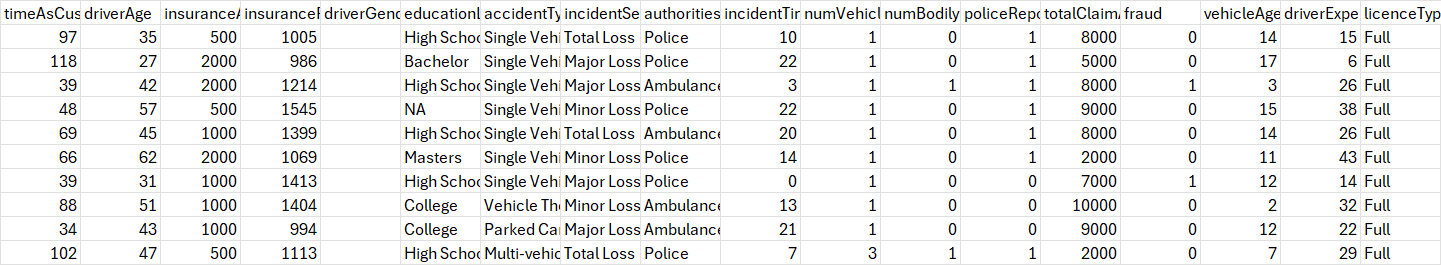
## Raw Columns:



| Column Name | Data Type | Description |
| --- | --- | --- |
| TXN\_DATE\_TIME | Date/Time | The date and time of the transaction |
| TRANSACTION\_ID | Categorical | Unique identifier for the transaction |
| CUSTOMER\_ID | Categorical | Unique identifier for the customer |
| POLICY\_NUMBER | Categorical | Unique identifier for the policy |
| POLICY\_EFF\_DT | Date | The policy effective date |
| LOSS\_DT | Date | Date when the loss occurred |
| REPORT\_DT | Date | Date when the incident was reported |
| INSURANCE\_TYPE | Categorical | Type of insurance |
| PREMIUM\_AMOUNT | Numerical | Premium amount for the insurance policy |
| CLAIM\_AMOUNT | Numerical | Amount claimed |
| CUSTOMER\_NAME | Categorical | Name of the customer |
| ADDRESS\_LINE1 | Categorical | Address line 1 of the customer |
| ADDRESS\_LINE2 | Categorical | Address line 2 of the customer |
| CITY | Categorical | City of the customer |
| STATE | Categorical | State of the customer |
| POSTAL\_CODE | Categorical | Postal code of the customer |
| SSN | Categorical | Social Security Number |
| MARITAL\_STATUS | Categorical | Marital status of the customer |
| AGE | Numerical | Age of the customer |
| TENURE | Numerical | Number of years the customer has been with the insurance company |
| EMPLOYMENT\_STATUS | Categorical | Employment status of the customer |
| NO\_OF\_FAMILY\_MEMBERS | Numerical | Number of family members |
| RISK\_SEGMENTATION | Categorical | Risk category assigned to the customer |
| HOUSE\_TYPE | Categorical | Type of house the customer owns or rents |
| SOCIAL\_CLASS | Categorical | Social class of the customer |
| ROUTING\_NUMBER | Categorical | Customer’s bank routing number |
| ACCT\_NUMBER | Categorical | Customer’s bank account number |
| CUSTOMER\_EDUCATION\_LEVEL | Categorical | Education level of the customer |
| CLAIM\_STATUS | Categorical | Status of the claim |
| INCIDENT\_SEVERITY | Categorical | Severity of the incident |
| AUTHORITY\_CONTACTED | Boolean | Whether the authorities were contacted |
| ANY\_INJURY | Boolean | Whether any injury was reported |
| POLICE\_REPORT\_AVAILABLE | Boolean | Whether a police report is available |
| INCIDENT\_STATE | Categorical | State where the incident occurred |
| INCIDENT\_CITY | Categorical | City where the incident occurred |
| INCIDENT\_HOUR\_OF\_THE\_DAY | Numerical | Hour of the day when the incident occurred |
| AGENT\_ID | Categorical | Unique identifier for the insurance agent |
| VENDOR\_ID | Categorical | Unique identifier for the vendor |

## Dataset 2 Columns after Cleaning & Enriching:

| NO | Column Name | Column Letter | Data Type | Description |
| --- | --- | --- | --- | --- |
| 1 | timeAsCustomer | A | Numerical | Number of months the individual has been a customer |
| 2 | driverAge | B | Numerical | Age of the driver |
| 3 | insuranceAccess | C | Numerical | The level of access to insurance |
| 4 | insurancePremium | D | Numerical | The premium amount for the insurance policy |
| 5 | driverGender | E | Boolean (0 = Male, 1 = Female, empty) | Gender of the driver (0 = Male, 1 = Female) EMPTY |
| 6 | educationLevel | F | Categorical | Education level of the driver |
| 7 | accidentType | G | Categorical | Type of accident |
| 8 | incidentSeverity | H | Categorical | Severity of the incident |
| 9 | authoritiesInvolved | I | Categorical | Whether authorities were involved |
| 10 | incidentTime | J | Numerical | Time of the incident |
| 11 | numVehiclesInvolved | K | Numerical | Number of vehicles involved in the accident |
| 12 | numBodilyInjuries | L | Numerical | Number of bodily injuries |
| 13 | policeReportBool | M | Boolean (0 = No, 1 = Yes) | Whether a police report was available (0 = No, 1 = Yes) |
| 14 | totalClaimAmount | N | Numerical | The total claim amount |
| 15 | fraud | O | Boolean (0 = No, 1 = Yes) | Whether the claim was fraudulent (0 = No, 1 = Yes) |
| 16 | vehicleAge | P | Numerical | Age of the vehicle |
| 17 | driverExperience | Q | Numerical | Number of years of driving experience |
| 18 | licenceType | R | Categorical | Type of licence held by the driver |



## 

## Dataset 2 Column Creations Logic:

FRAUD REPORTED:  
Will use the Claim\_Status (AC) column and assign A (approved) as Not Fraud, and D (denied) as Fraud. Then convert to numerical with Fraud as 1 and Not Fraud as 0.

driverExperience:

Source:

Column age - S

16

R - (Random weighted variable between 0 and 6)

Formula:

driverExperience = (Age − 16) − R

accidentType:

Generate based on the statistical commonhood of each accident type that exists in Dataset 1.

numVechileInvolved:  
Generate based on the statistical likelihood of numvehiclesinvolved in Dataset 1 based on accident type.

vehicleAge:

Generate based on likelihood/occurrence of vehicle age in dataset 1.

insuranceAccess:

Generate based on correlation between InsurancePremium and InsuranceAccess prevalent in Dataset 1, and then apply that to Dataset2.

# DATASET 3 CLEANING MAPPING

## Source:

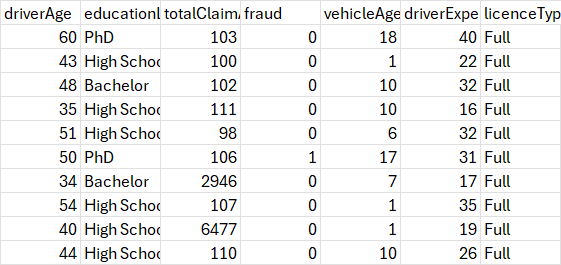
<https://www.kaggle.com/datasets/xiaomengsun/car-insurance-claim-data>

## Raw Columns:

| Column Name | Data Type | Description |
| --- | --- | --- |
| ID | Categorical | Unique identifier for the customer |
| KIDSDRIV | Numerical | Number of kids driving in the household |
| BIRTH | Date | Birth date of the customer |
| AGE | Numerical | Age of the customer |
| HOMEKIDS | Numerical | Number of kids living in the household |
| YOJ | Numerical | Years on the job |
| INCOME | Numerical (Currency) | Income of the customer |
| PARENT1 | Categorical | Whether the customer is a single parent |
| HOME\_VAL | Numerical (Currency) | Value of the home |
| MSTATUS | Categorical | Marital status |
| GENDER | Categorical | Gender of the customer |
| EDUCATION | Categorical | Education level of the customer |
| OCCUPATION | Categorical | Occupation of the customer |
| TRAVTIME | Numerical | Travel time to work |
| CAR\_USE | Categorical | Whether the car is used for private or commercial purposes |
| BLUEBOOK | Numerical (Currency) | Value of the car |
| TIF | Numerical | Time in force of the insurance policy |
| CAR\_TYPE | Categorical | Type of car |
| RED\_CAR | Boolean | Whether the car is red |
| OLDCLAIM | Numerical (Currency) | Amount claimed in prior incidents |
| CLM\_FREQ | Numerical | Frequency of claims |
| REVOKED | Boolean | Whether the driver's licence has been revoked |
| MVR\_PTS | Numerical | Motor Vehicle Record points |
| CLM\_AMT | Numerical (Currency) | Amount claimed in the current incident |
| CAR\_AGE | Numerical | Age of the car |
| CLAIM\_FLAG | Boolean | Whether a claim was filed |
| URBANICITY | Categorical | Urban/rural classification |

## Dataset 3 Columns after Cleaning & Enriching:

| No. | Column Name | Column Letter | Data Type | Description |
| --- | --- | --- | --- | --- |
| 1 | driverAge | A | Numerical | Age of the driver |
| 2 | educationLevel | B | Categorical | Education level of the driver |
| 3 | totalClaimAmount | C | Numerical | The total claim amount |
| 4 | fraud | D | Boolean | Whether the claim was fraudulent (0 = No, 1 = Yes) |
| 5 | vehicleAge | E | Numerical | Age of the vehicle |
| 6 | driverExperience | F | Numerical | Number of years of driving experience |
| 7 | licenceType | G | Categorical | Type of licence held by the driver |



## Dataset 3 Column Creations Logic:

FRAUD:  
Will use the REVOKED (V) column and assign YES (DENIED) as Fraud, and NO (NOT REVOKED) as NOT Fraud. Then convert to numerical with Fraud as 1 and Not Fraud as 0.

# ORGANISATION:

# Raw Files:

Raw Dataset 1 - [dataset1.csv](https://github.com/AlanDataPortfolio/ey-azure-fn-pipeline/blob/master/assets/data/raw/dataset1.csv)

Raw Dataset 2 - [dataset2.csv](https://github.com/AlanDataPortfolio/ey-azure-fn-pipeline/blob/master/assets/data/raw/dataset2.csv)

Raw Dataset 3 - [dataset3.csv](https://github.com/AlanDataPortfolio/ey-azure-fn-pipeline/blob/master/assets/data/raw/dataset3.csv)

# Cleaning:

Scripts:

Dataset 1:

[cleaning\_Dataset1.py](https://github.com/AlanDataPortfolio/ey-azure-fn-pipeline/blob/master/src/local_pipeline/cleaning/cleaning_Dataset1.py)-> Filling NULL Values, Renaming Relevant

Dataset 2:

[cleaning\_Dataset2.py](https://github.com/AlanDataPortfolio/ey-azure-fn-pipeline/blob/master/src/local_pipeline/cleaning/cleaning_Dataset2.py) -> Filling NULL Values, Renaming Relevant

Dataset 3:

[cleaning\_Dataset3.py](https://github.com/AlanDataPortfolio/ey-azure-fn-pipeline/blob/master/src/local_pipeline/cleaning/cleaning_Dataset3.py) -> Filling NULL Values, Renaming Relevant Columns

Data Files:  
 [cleaned\_dataset1.csv](https://github.com/AlanDataPortfolio/ey-azure-fn-pipeline/blob/master/assets/data/cleaned/cleaned_dataset1.csv)

[cleaned\_dataset2.csv](https://github.com/AlanDataPortfolio/ey-azure-fn-pipeline/blob/master/assets/data/cleaned/cleaned_dataset2.csv)

[cleaned\_Dataset3.csv](https://github.com/AlanDataPortfolio/ey-azure-fn-pipeline/blob/master/assets/data/cleaned/cleaned_Dataset3.csv)

# 

# Enriching:

Scripts:

Converting Categorical/Booleans to Numerical -> Adding/Creating Columns, Dropping Columns

[enriching\_Dataset1.py](https://github.com/AlanDataPortfolio/ey-azure-fn-pipeline/blob/master/src/local_pipeline/enriching/enriching_Dataset1.py)

[enriching\_Dataset2.py](https://github.com/AlanDataPortfolio/ey-azure-fn-pipeline/blob/master/src/local_pipeline/enriching/enriching_Dataset2.py)

[enriching\_Dataset3.py](https://github.com/AlanDataPortfolio/ey-azure-fn-pipeline/blob/master/src/local_pipeline/enriching/enriching_Dataset3.py)

Data Files:

[enriched\_dataset1.csv](https://github.com/AlanDataPortfolio/ey-azure-fn-pipeline/blob/master/assets/data/enriched/enriched_dataset1.csv)

[enriched\_dataset2.csv](https://github.com/AlanDataPortfolio/ey-azure-fn-pipeline/blob/master/assets/data/enriched/enriched_dataset2.csv)

[enriched\_dataset3.csv](https://github.com/AlanDataPortfolio/ey-azure-fn-pipeline/blob/master/assets/data/enriched/enriched_dataset3.csv)

# Synthesising:

Scripts:

Technique 1:

[synthesising\_method\_1.py](https://github.com/AlanDataPortfolio/ey-azure-fn-pipeline/blob/master/src/local_pipeline/synthesising/synthesising_method_1.py)

Technique 2:

[synthesising\_method\_2.py](https://github.com/AlanDataPortfolio/ey-azure-fn-pipeline/blob/master/src/local_pipeline/synthesising/synthesising_method_2.py)

Synthetic Data Files:

[synthesised\_method1.csv](https://github.com/AlanDataPortfolio/ey-azure-fn-pipeline/blob/master/assets/data/synthesised/synthesised_method1.csv)

[synthesised\_method2.csv](https://github.com/AlanDataPortfolio/ey-azure-fn-pipeline/blob/master/assets/data/synthesised/synthesised_method2.csv)

# 

# 

# Merging:

Scripts:

Merged Dataset 1 & 2:

[merging\_dataset\_1\_2.py](https://github.com/AlanDataPortfolio/ey-azure-fn-pipeline/blob/master/src/local_pipeline/merging/merging_dataset_1_2.py)

Merged 6500 rows:

[merging\_6500\_rows.py](https://github.com/AlanDataPortfolio/ey-azure-fn-pipeline/blob/master/src/local_pipeline/merging/merging_6500_rows.py)

Merged 10500 rows:

[merging\_10500\_rows.py](https://github.com/AlanDataPortfolio/ey-azure-fn-pipeline/blob/master/src/local_pipeline/merging/merging_10500_rows.py)

Merged 20000 rows:

[merging\_20000\_rows.py](https://github.com/AlanDataPortfolio/ey-azure-fn-pipeline/blob/master/src/local_pipeline/merging/merging_20000_rows.py)

Merged Data Files:

[merged\_dataset\_1\_2.csv](https://github.com/AlanDataPortfolio/ey-azure-fn-pipeline/blob/master/assets/data/merged/merged_dataset_1_2.csv)

[merged\_6500\_rows.csv](https://github.com/AlanDataPortfolio/ey-azure-fn-pipeline/blob/master/assets/data/merged/merged_6500_rows.csv)

[merged\_10500\_rows.csv](https://github.com/AlanDataPortfolio/ey-azure-fn-pipeline/blob/master/assets/data/merged/merged_10500_rows.csv)

[merged\_20000\_rows.csv](https://github.com/AlanDataPortfolio/ey-azure-fn-pipeline/blob/master/assets/data/merged/merged_20000_rows.csv)

# JSON:

Script:

[to\_json.py](https://github.com/AlanDataPortfolio/ey-azure-fn-pipeline/blob/master/src/local_pipeline/formatting/to_json.py)

Data File:

[20000\_rows.json](https://github.com/AlanDataPortfolio/ey-azure-fn-pipeline/blob/master/assets/data/json/20000_rows.json)

ARCHIVE (NOT RELEVANT):

## Statistical Run Results:

**accidentType distribution in cleaned dataset:**

{'Multi-vehicle Collision': 0.419, 'Single Vehicle Collision': 0.403, 'Vehicle Theft': 0.094, 'Parked Car': 0.084}

**Correlation between accidentType and numVehiclesInvolved:**

accidentType categories found: ['Single Vehicle Collision' 'Vehicle Theft' 'Multi-vehicle Collision' 'Parked Car']

Data points per accidentType:

accidentType

Multi-vehicle Collision 419

Single Vehicle Collision 403

Vehicle Theft 94

Parked Car 84

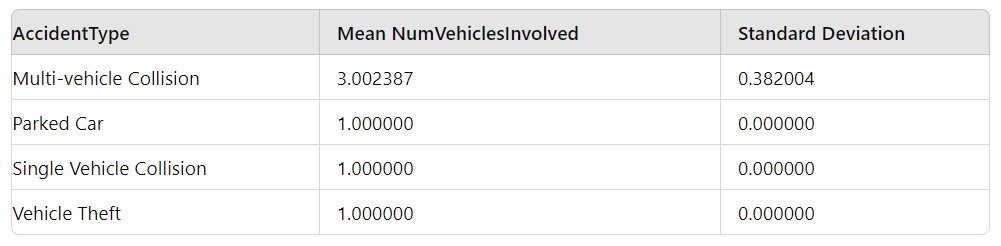
ANOVA results:

F-value = 5312.650815087258

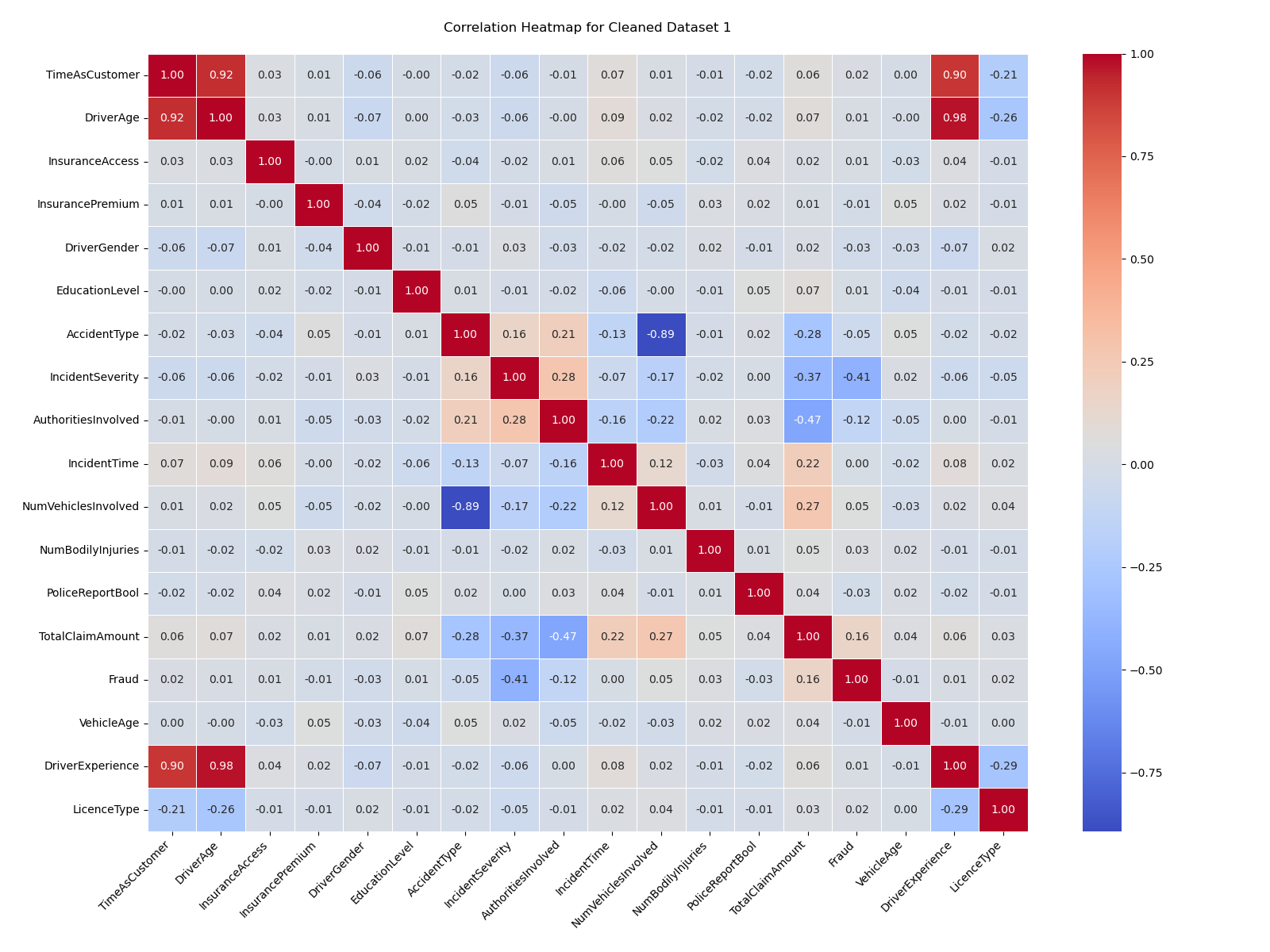
p-value = 0.0

Conclusion: There is a significant relationship between accidentType and numVehiclesInvolved.

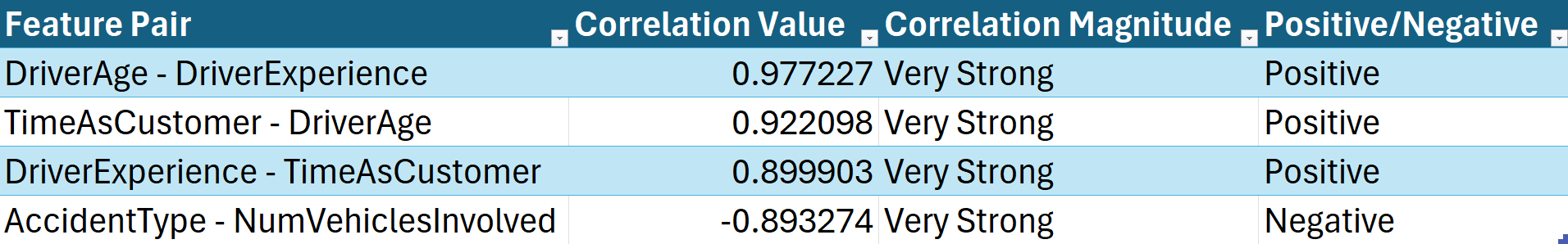
accidentType statistics for numVehiclesInvolved:



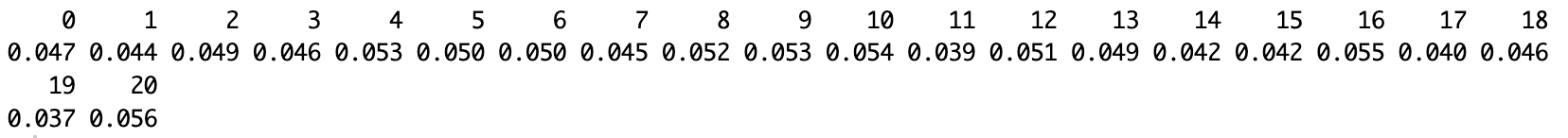
**Correlation Heatmap of Cleaned & Enriched Dataset 1:**



**Significant Correlations:**



**Distribution of vehicleAge in Cleaned & Enriched Dataset 1:**



GETTING RID OF VEHICLE MAKE AND VEHICLE MODEL BECAUSE DATASETS ARE FROM AMERICA AND AMERICAN CARS AREN’T RELEVANT TO NRMA PROJECT

## Archive Notes:

Got rid of netcapital

**NetCapital**

**Source:**

**Capital\_loss - P**

**capital\_gains - Q**

**Formula:**

**NC =(Capital Gains − Capital Loss)**

**Notes for Cleaning Dataset 3:**

Columns to Drop:

YOJ

Occupations

Car type

PIF

Blue BOOk

Red Car

MVR Points

Claim Flag

Urbancity

Columns to Clean (Fill):

Income

Home Value

Gender - Make it numerical

Education - Standardise between datasets

Revoked WIll Become Fraud COlumn

Claim Amount needs to be filled (replace 0 with values)  
Car Age - Fill Missing Values

Correlation between Revoked and Claim\_Flag