1. Data Generation and Exploratory Data Analysis (EDA):

Fabricating data to generate accurate and viable data is always a difficult step.

I sourced my data mostly from using packages out there specifically for data generation and sourcing various websites and blogs for additional information and ideas for having fraudulent claims.

I used *Faker* as well as ***Elizabeth*** in conjunction to get better and appropriate results, I initially started off with an excel database of names and surnames, but due to the malformation of the database, it was tedious and cumbersome to validate and integrate. I also used some random data which I implemented and validated to make the data set more realistic. Here is the list of fraudulent claims reasons with explanations.

No Date of birth	It is important to know the date of birth of a person since it is possible to work out the age of the person and check maturity ages for the policy.
Date of birth calculated Age and Age do not match	Since this is a verification step to make sure that the person is valid and not fabricated, date of birth calculated age and age should match otherwise this claim is invalid.
Claim amount is more than Sum Insured	A person can only claim the maximum amount that the person is insured for.
No Policy start date	This is a mandatory field since without this a claim is incomplete, due to the fact that the policy may not exist.
No Policy end date	Also, another mandatory field since this verifies if the claim was within the policy agreement, due to the fact that the policy may not exist.
Policy end date before start date	This makes an invalid policy since the end date can never be before the start.
Claim Date before loss	A person cannot claim before a loss has happened.
No kind of loss	If the kind loss is not specified the claim cannot be valid.
Invalid kind of loss	If the kind of loss is not one of the options that the insurer provides.
No premium but has claim	If you do not pay a monthly fee, one cannot have a claim. Since nothing is free In this world.
Claim after Policy end date	A Claim cannot be created after the policy has ended.
Claim before Policy start	A Claim cannot be created before a policy has started.
Age is not in requirements	Age is a requirement that each insurer has, you have to have a reasonable age, older than 18 (or turning) and younger that a maximum age of 120.

EDA

From doing exploratory data analysis I have gained an in-depth understating of that data and from plotting the data one can see how distributed the data is and how "Good" the dataset is.

There are a few null values that are in the database and that should be cleaned in the data cleaning steps and the data types identified by ***Pandas*** which is a python library where it gives useful stats on

the data, find numeric data as well as give some nice statistical values on the data like the percentiles as well as the distribution etc.

I sampled the data and got the first 5 and last 5 rows too physically how well the data is scaled and is correct.

For each attributed I then performed pandas value counts that will find the data and perform frequency counts of each data found even caters for null values.

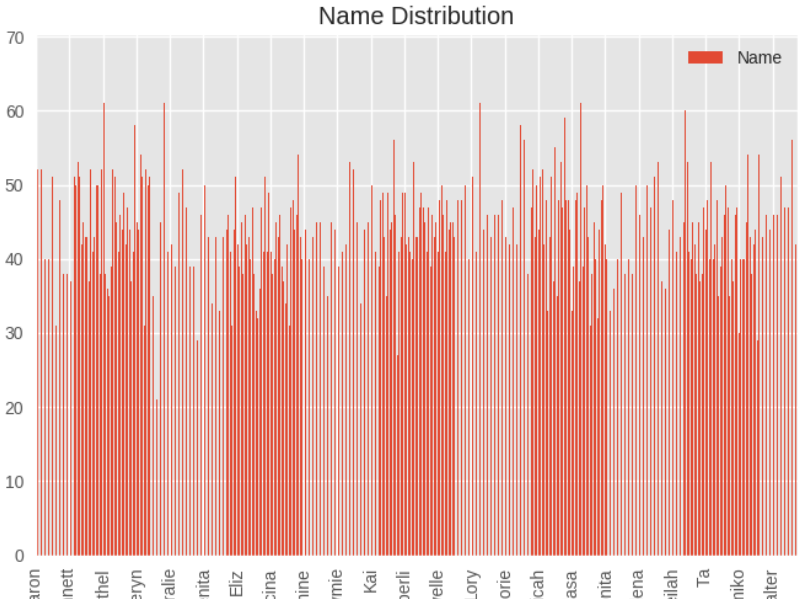
The table below is the correlation for all the numeric data

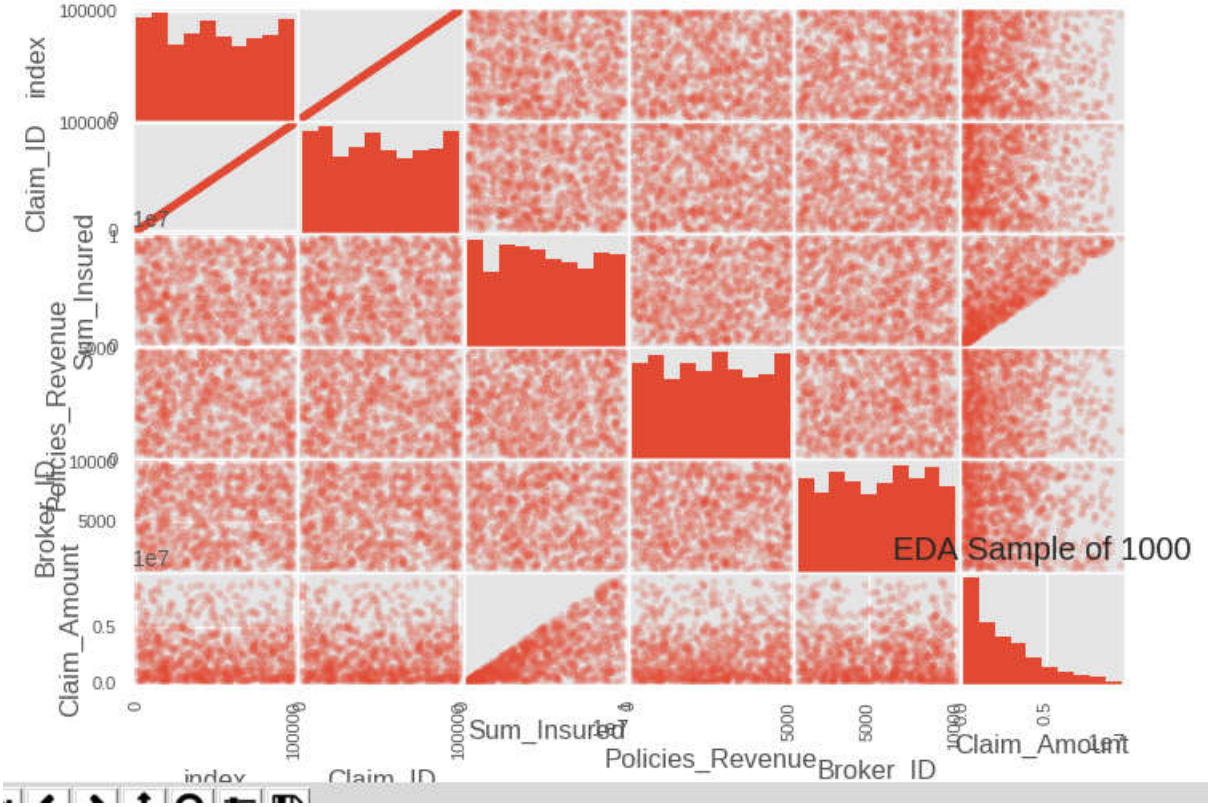
```
Correlate Data:
                  Claim_ID       Age  Sum_Insured  Policies_Revenue  Claim_Amount
Claim_ID          1.000000  0.004639    -0.002261         -0.002160  -0.002018
Age               0.004639  1.000000    -0.000841          0.001008  -0.002539
Sum_Insured      -0.002261 -0.000841     1.000000         -0.001666   0.644548
Policies_Revenue -0.002160  0.001008    -0.001666          1.000000   0.002480
Claim_Amount     -0.002018 -0.002539     0.644548          0.002480   1.000000
```

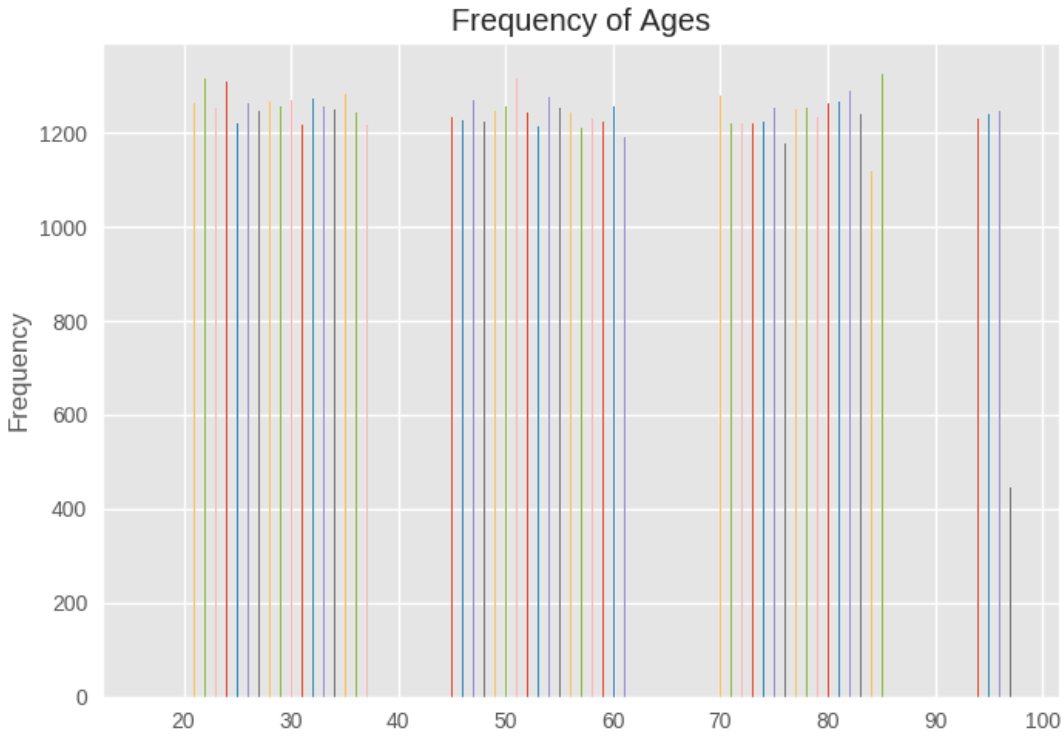
```
Data Type information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100000 entries, 0 to 99999
Data columns (total 31 columns):
Claim_ID                  100000 non-null int64
Name                      100000 non-null object
Surname                   100000 non-null object
Age                       99986 non-null float64
Gender                    100000 non-null object
Marital_Status            100000 non-null object
Date_Of_Birth             99986 non-null datetime64[ns]
Sum_Insured               100000 non-null float64
Policies_Revenue          100000 non-null float64
Policy_Start              99980 non-null datetime64[ns]
Policy_End                99968 non-null datetime64[ns]
Fraudulent_Claim          100000 non-null object
Fraudulent_Claim_Reason   100000 non-null object
Date_Of_Loss              100000 non-null datetime64[ns]
Date_Of_Claim             99980 non-null datetime64[ns]
Broker_ID                 100000 non-null object
Insured_ID                100000 non-null object
Kind_Of_Loss              100000 non-null object
Claim_Amount              100000 non-null float64
Party_Name                100000 non-null object
Party_Surname             100000 non-null object
Service_Provider          100000 non-null object
Policy_Holder_Street      100000 non-null object
Policy_Holder_Province    100000 non-null object
Policy_Holder_City        100000 non-null object
Policy_Holder_Area        100000 non-null object
Policy_Holder_Postal      100000 non-null object
Province                  100000 non-null object
City                      100000 non-null object
Area                      100000 non-null object
Postal_Code               100000 non-null object
dtypes: datetime64[ns](5), float64(4), int64(1), object(21)
memory usage: 23.7+ MB
```

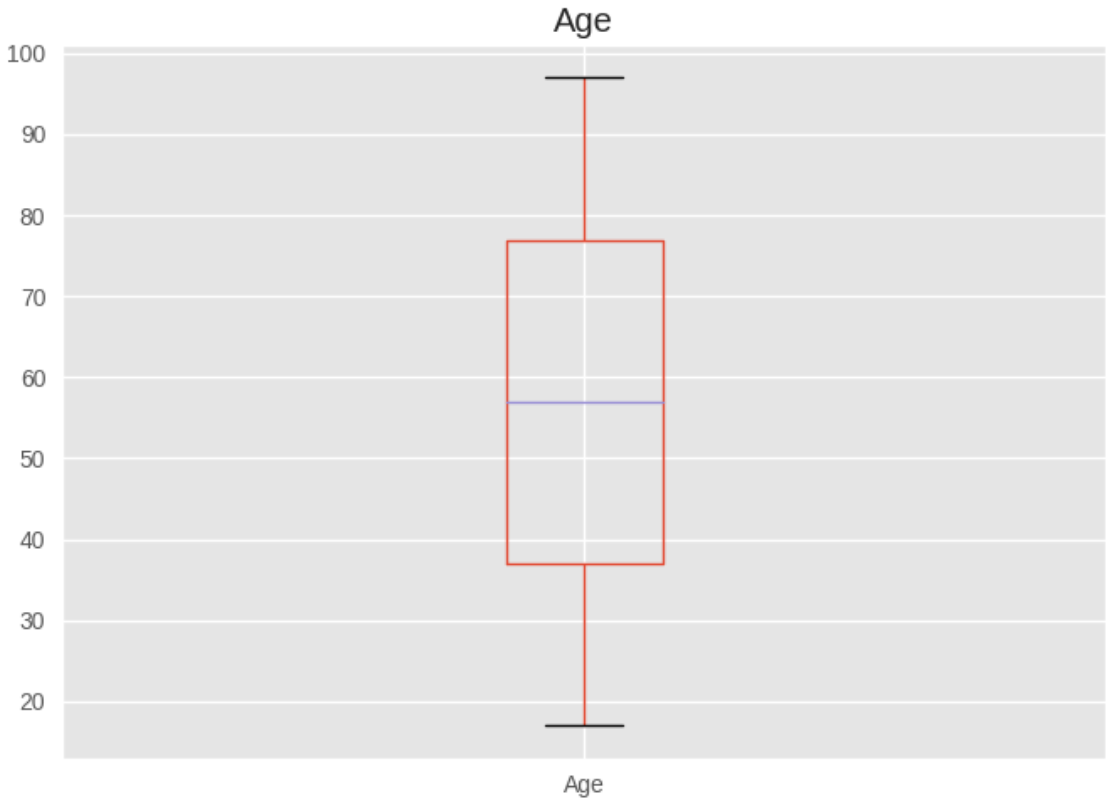
Numeric Data Information:

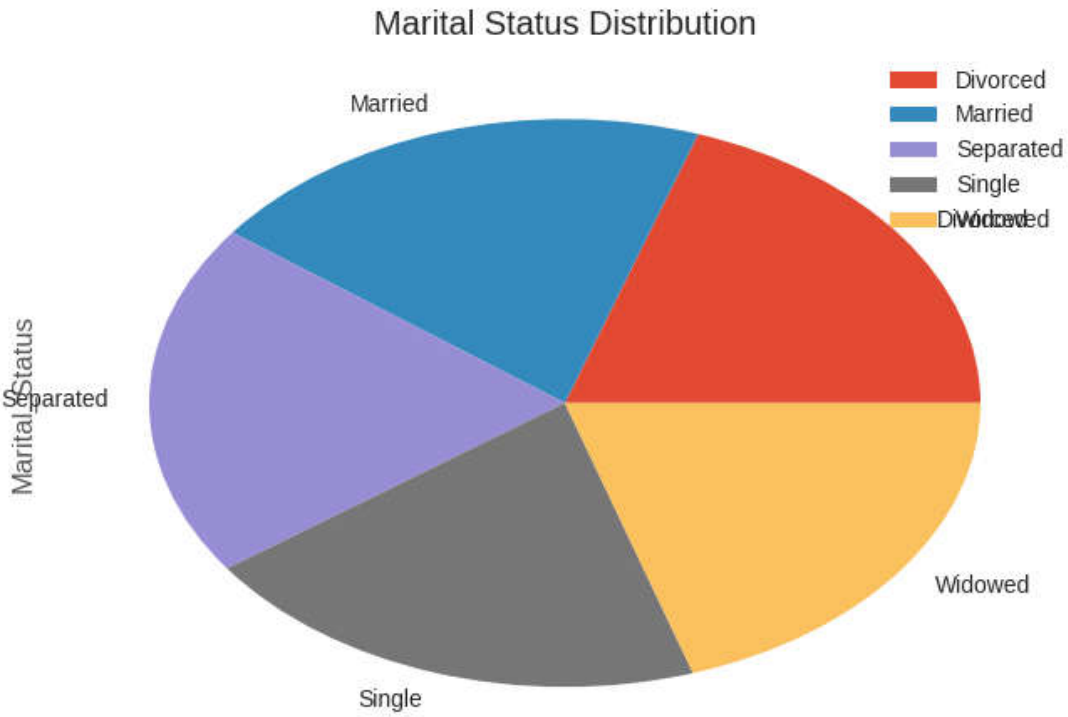
	Claim_ID	Age	Sum_Insured	Policies_Revenue	Claim_Amount
count	100000.000000	99986.000000	1.000000e+05	100000.000000	1.000000e+05
mean	50000.500000	56.754966	5.099169e+06	2547.939607	2.546211e+06
std	28867.657797	23.110670	2.824372e+06	1416.286187	2.196578e+06
min	1.000000	17.000000	2.001569e+05	0.000000	5.780000e+00
25%	25000.750000	37.000000	2.654805e+06	1316.902500	7.334786e+05
50%	50000.500000	57.000000	5.098258e+06	2557.290000	1.924911e+06
75%	75000.250000	77.000000	7.541800e+06	3775.685000	3.879260e+06
max	100000.000000	97.000000	9.999767e+06	4999.910000	9.961545e+06

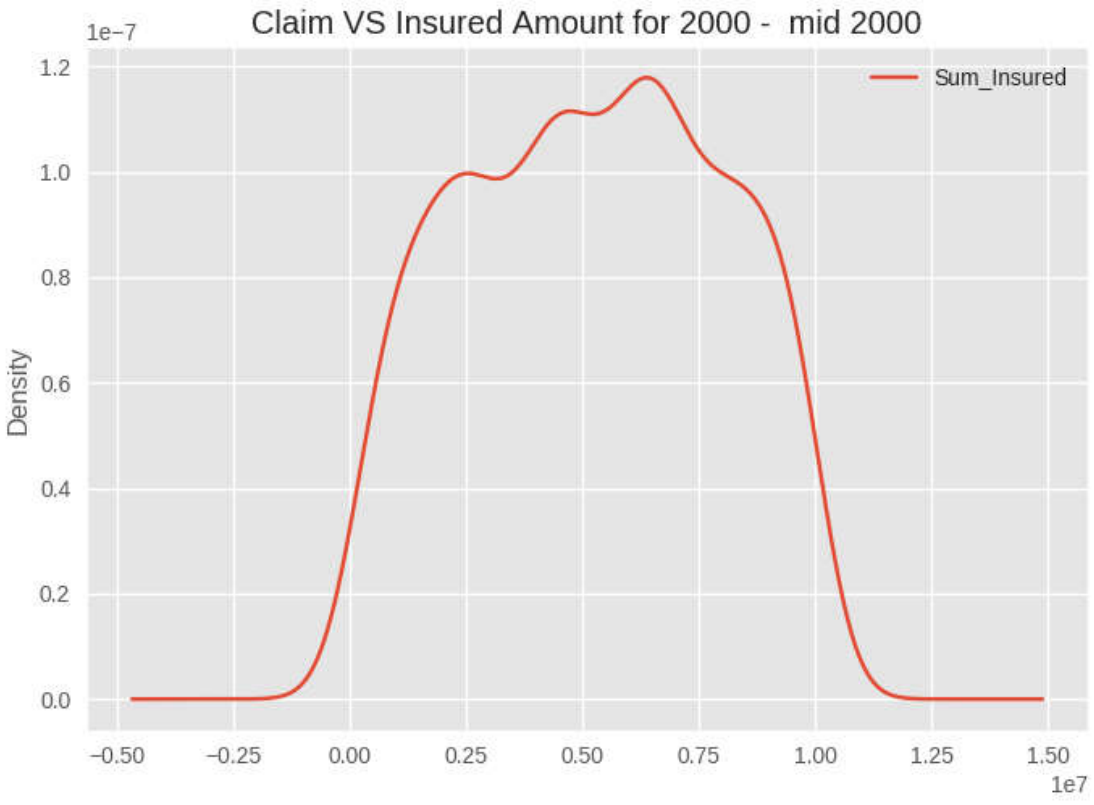


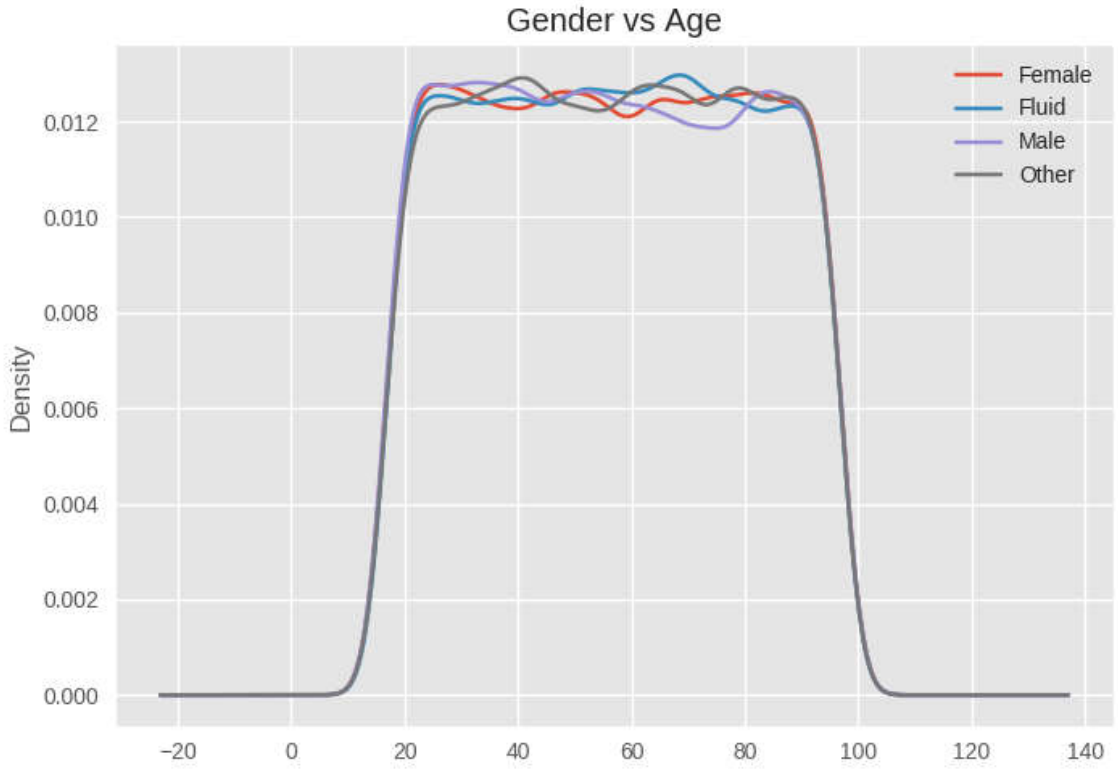


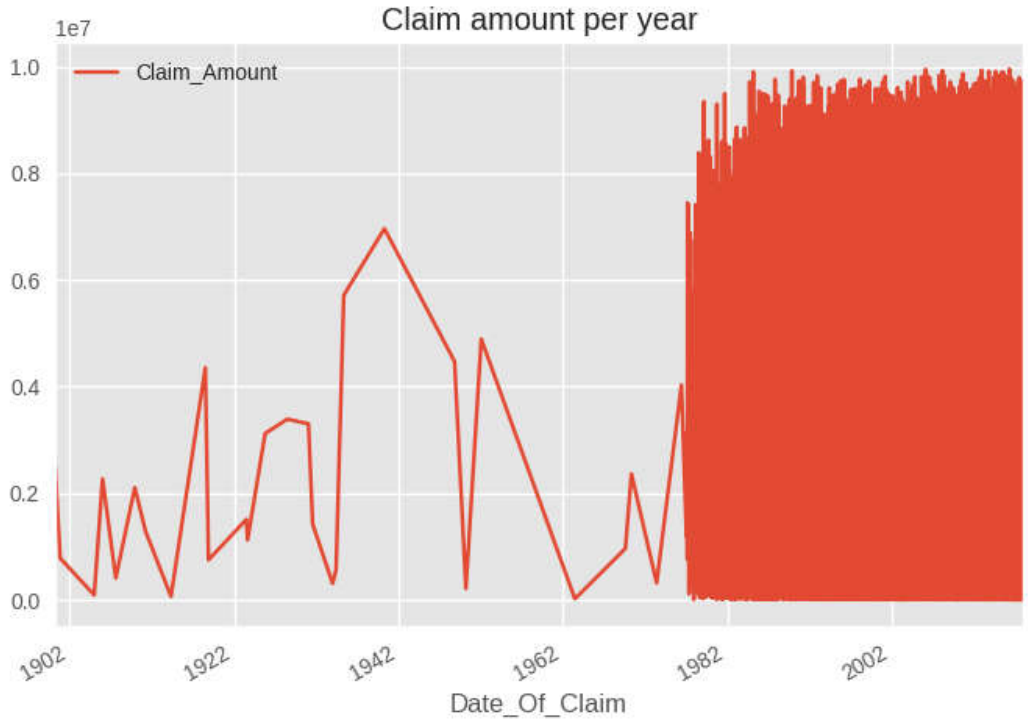

Frequency of Ages
Frequency
0
200
400
600
800
1000
1200
20
30
40
50
60
70
80
90
100

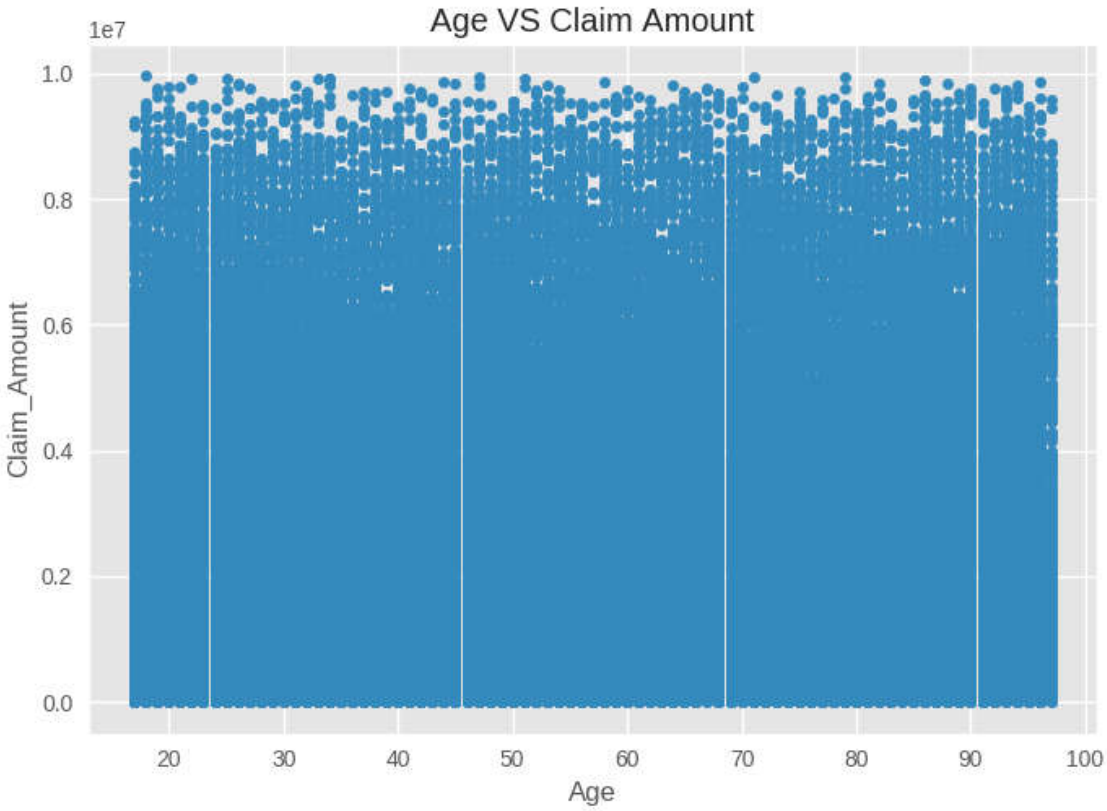

Age
100
90
80
70
60
50
40
30
20
Age

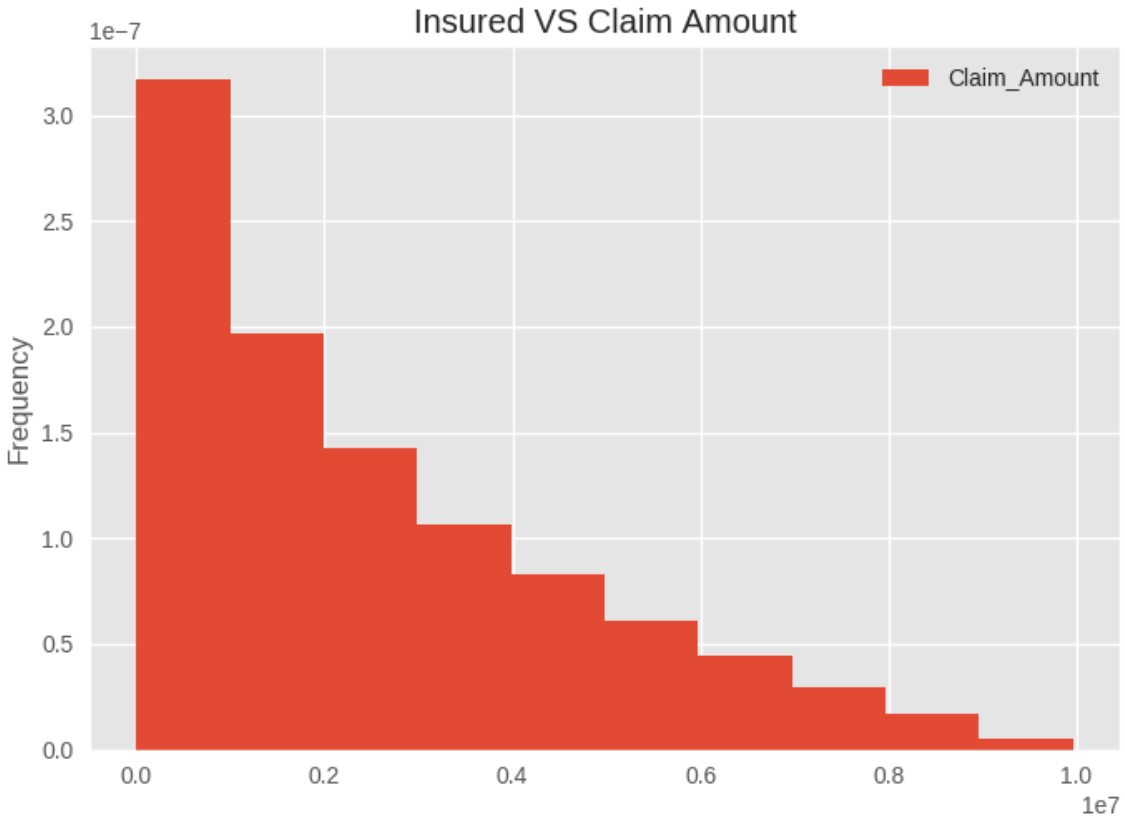

Marital Status Distribution
Divorced
Married
Separated
Single
Widowed
Married
Separated
Widowed
Single
Marital Status

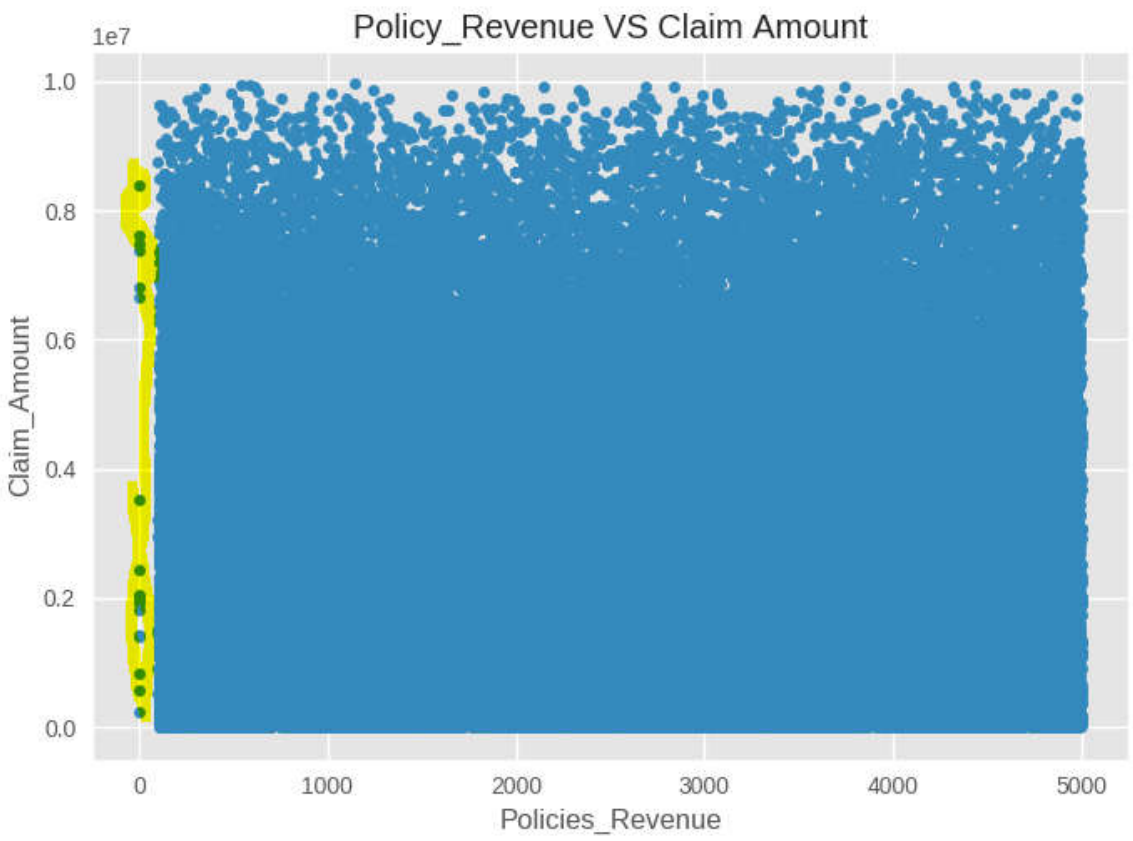

Claim VS Insured Amount for 2000 - mid 2000
1e−7
Sum_Insured
Density
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1e7


Gender vs Age
Female
Fluid
Male
Other
Density
0.012
0.010
0.008
0.006
0.004
0.002
0.000
−20
0
20
40
60
80
100
120
140


Claim amount per year
1e7
Claim_Amount
1.0
0.8
0.6
0.4
0.2
0.0
1902
1922
1942
1962
1982
2002
Date_Of_Claim


Age VS Claim Amount
1e7
1.0
0.8
0.6
0.4
0.2
0.0
Claim_Amount
20
30
40
50
60
70
80
90
100
Age


Insured VS Claim Amount
1e−7
Claim_Amount
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Frequency
0.0
0.2
0.4
0.6
0.8
1.0
1e7



The yellow above shows the outliers which need to be cleaned to provide better machine learning properties.

Data scaling is required since there is random data that was used even tough it was fabricated under certain constraints which then scales it nicely. Only thinking realistically the scaling of Ages would be needed. The standard devieation and means of the numeric fields in the table above need to be scaled for machine learning phase becase the values need to be between 0 and 1 and currently it isn't which would make classification bad for machine learning thus making fraud claims pass easy and having a low accuracy rate on fraud detection.

EDA Questions

1. **What is the distribution of ages? –** The distribution is uniform which is ideal, but not realistic since its random data there too many people with high ages.
2. **What is the correlation of the dataset? –** The table above shows this
3. **What is the statisical infomration about numeric data? –** The table above shows the mean, standard deviation and percentiles which show where the distrubution of data lies.
4. **What does the Box plots say? –** They show the distribution of the feature and we see that the box plot of age is eveny distributed.
5. **How can one detect anomilies and outlies? –** The above graph shows outliers and pandas gave us the null data information [not included in the report]

2. PPDM

a. **Partial Identifiable Attributes / Quasi Identifiers**
   i. Age, Sex,

b. **Sensitive attributes.**
   i. Sum_Insured, Policies_Revenuem Postal Code, BrokerID

Field	Explanation
Name	Replace with *
Surname	Replace with *
Age	Replace last char with * [22 -> 2*]
Gender	Replace duplicates leading with *
Marital_Status	Use only first 2 chars
Date_Of_Birth	Remove day part
Sum_Insured	
Policies_Revenue	
Policy_Start	
Policy_End	
Fraudulent_Claim	Make true claims with *
Fraudulent_Claim_Reason	Replace duplicates leading with *
Date_Of_Loss	
Date_Of_Claim	
Broker_ID	Remove 'BKR' that can makes the ID
Insured_ID	Replace duplicates leading with *
Kind_Of_Loss	Replace duplicates leading with *
Claim_Amount	
Party_Name	Replace with *
Party_Surname	Replace with *
Service_Provider	Replace duplicates leading with *
Policy_Holder_Street	Replace with *
Policy_Holder_Province	
Policy_Holder_City	Replace with *
Policy_Holder_Area	Replace with * - Postal code can give this
Policy_Holder_Postal	Last 2 char *
Province	Replace with *
City	
Area	Replace with * - Postal code can give this
Postal_Code	

I chose to preserve the data in the table above by removing some of the redundant and unclassifiable data such as names etc. The Marital status can be distinguished from the first 2 characters from an admin. The age column I removed the last digit which anonymised it a bit and for several columns I replace leading duplicated with '*' as defined in some research papers. The above table makes use of the k-anonymity model in certain attributes. There is some crucial information that is needed in machine learning which I have left as is and is mostly numerical data. This data cannot be matched or aid in helping the attacker identify the person. However, there is still some info that could lead the person's whereabouts which will be from a look up table within the organisation.

3. Data Cleaning

Data cleaning involves making data relevant and removing elements that are redundant or invalid. There are many criteria that can resolve in cleaning such as removing records with NULL values, invalid records that have invalid data such as incorrect types or data that doesn't match or complement the data in the same row (i.e. having a claim data before a loss data which is not valid).

4. Machine Learning

For this phase I attempted to use a neural network but was too difficult to do. So I used a Decision Tree Classifier which can predict claims based on a tree like structure. I have trained it and the accuracy of prediction is:

Accuracy: **99.852297593**, with training set of **73117** elements and a testing set of **18280** elements.

The training set and testing set is derived from using the original data set after PPDM and Cleaning.

I chose the decision tree because it is easier and faster to detect and train because it relies on logic rather than optimizations as opposed to neural networks, however the accuracy of a neural network can be more and can essentially have a better prediction rate to newer claims opposed to the decision tree since the decision tree makes assumptions on the existing data. This data is accurate because it has been cleaned providing better results. For the model I chose to use 4 features since they had the most relevance and still had its properties after the PPDM unlike some features for privacy concerns had to be anonymised making the feature less relevant because it can give more errors such as age, postal code etc. The 4 features are ['Sum_Insured', 'Policies_Revenue', 'Broker_ID', 'Claim_Amount']. The sum insured can be a useful measure because if the amount is high and your revenue is low then something doesn't correlate, Policies_Revenue is important as mentioned and since it correlates with Sum_Insured they have to have some relationship. BrokerID since this is the person that sells you the package for your insurance, this person can be corrupt and pass you to have a high Sum Insurance and a low Revenue this is a very important factor. Claim_Amount as well needs to correlate to the Sum_Insured amount since they can only pay as much as one is insured.

5. Bonus

Big data is very common in the new age and with this it is easier for fraudsters to get away with crimes. However, due to big data analytics, it is possible to be able to mitigate this using data mining techniques to predict fraud and defer certain events that can lead to fraud or massive attacks that can be malicious. Exploratory data analysis helps in being able to understand the data that one has in order to know what is important and how well the distribution of the data is. This also helps administrators to establish critical decision and analysis of the data in terms of what data is defected, malformed (dirty records or invalid records). If one attaches regression plotting and data correlation between the data collected this can also help administrations additionally decide on what elements depend on another, and decide if data scaling is needed. EDA also enables visual representations of data since its large amounts of data visual EDA is preferred over manually locating, where one can see by the dips in the curves of graphs and through time-series plotting which years were better than others etc. (depending on the data stored). PPDM is very important because if an attacker manages to gain access to a database through a backdoor or through a direct link, peoples information is at risk and since it's the Digital Age everything is in the cloud and this leaves attackers leaching for openings and software vulnerabilities. In such a worst case even if the attacker gains access the data that the attacker receives

should be anonymised so that it is less easy or impossible for the attacker to distinguish what information is what. By using k-anonymity model as well as some other relevant data preserving algorithms will be beneficial since if the attacker gets age groups he cannot directly identify a person. If the attacker gets the province of the member, he won't be able to gain anything however if he manages to get the postal code and some other information this could be dangerous that's why certain information that can lead to identification of a certain individual should be anonymised however not so that a person cannot be identified by the organisation that is holding the data. This also helps for being able to use feature selection for Machine learning. PPDM and EDA will help in deciding what features can be used in the machine learning phase that would bear the best results. Data cleaning is necessary since that will heavily effect the accuracy of the machine learning capabilities. It is also important that these features have good and bad data otherwise the learning would be not helpful, because the algorithm will only know what is correct data won't know what is incorrect so having more good data and little bad is better that having no bad data. Testing is a big part of machine learning since that will classify data to be able to detect what is good and bad. It can also be used to detect malicious instances upon inserting data into a database the machine learning could detect malicious code and reject it. So that it prevents a malicious attack.

Decision Tree Clarifier