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

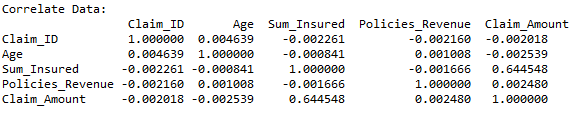
From doing exploratory data analysis I have gained an in-depth understating of that data and from plotting the data one can see how distributed the data is and how “Good” the dataset is.

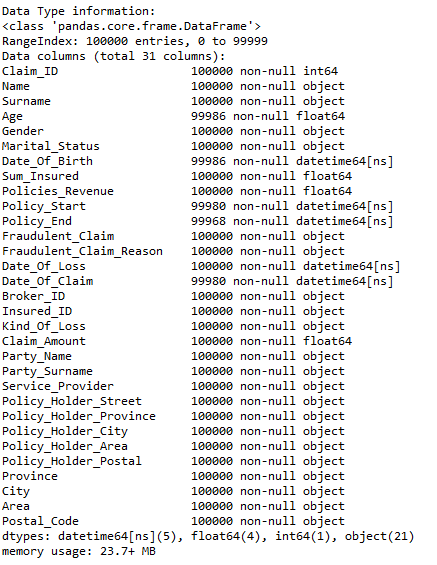
There are a few null values that are in the database and that should be cleaned in the data cleaning steps and the data types identified by *Pandas* which is a python library where it gives useful stats on the data, find numeric data as well as give some nice statistical values on the data like the percentiles as well as the distribution etc.

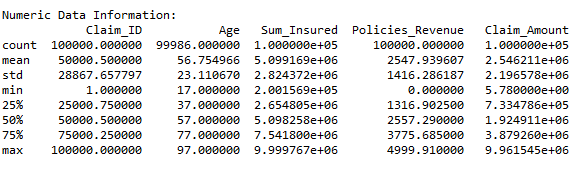
I sampled the data and got the first 5 and last 5 rows too physically how well the data is scaled and is correct.

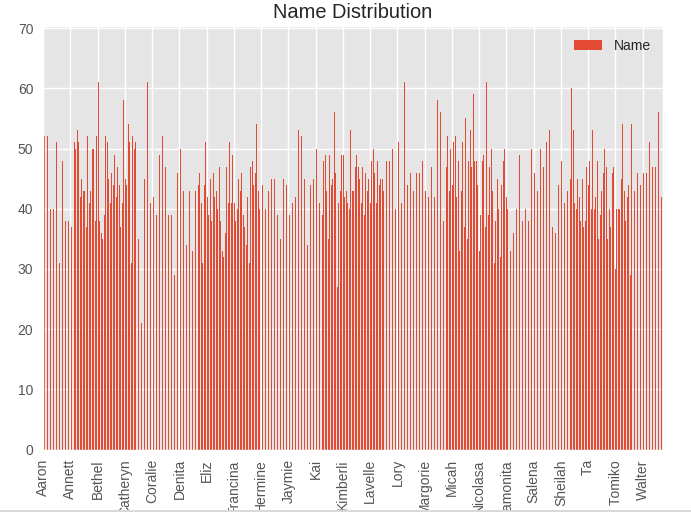
For each attributed I then performed pandas value counts that will find the data and perform frequency counts of each data found even caters for null values.

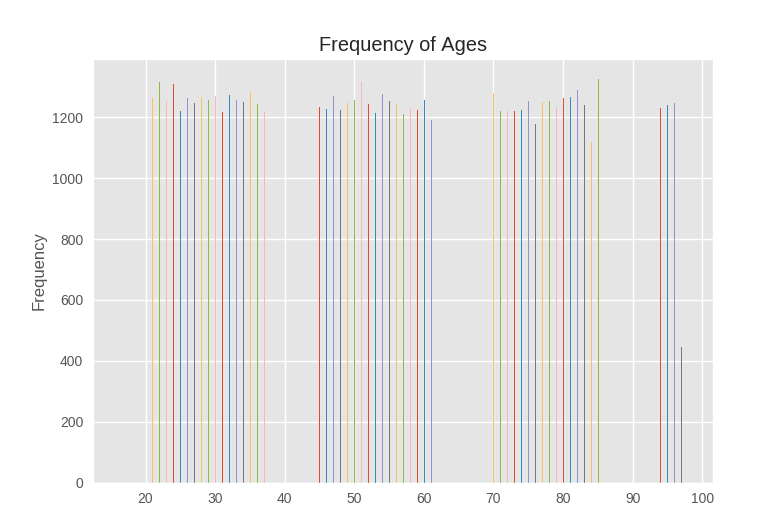
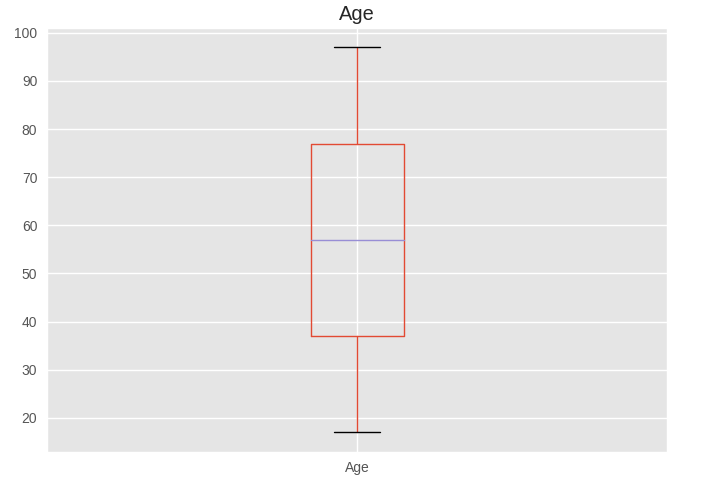
The table below is the correlation for all the numeric data

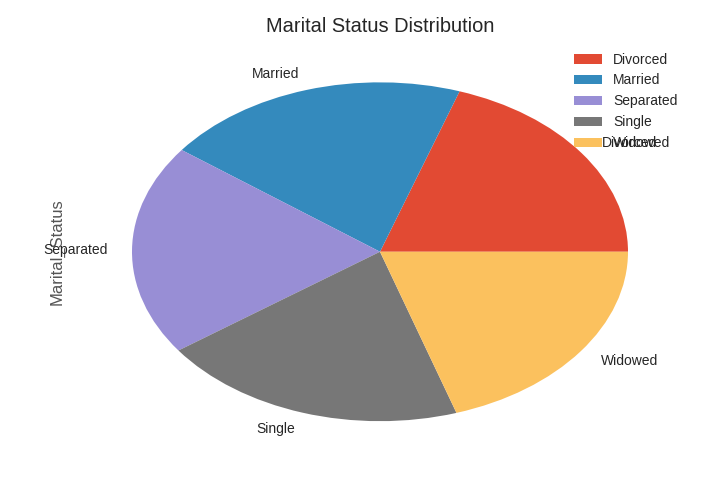
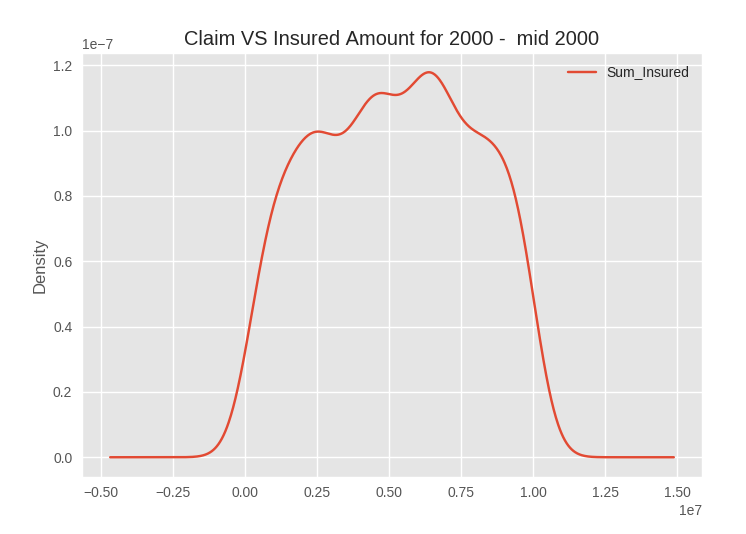


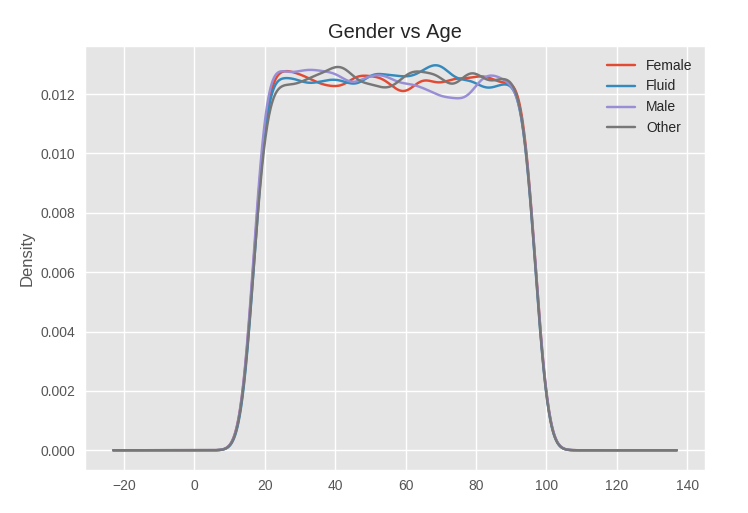


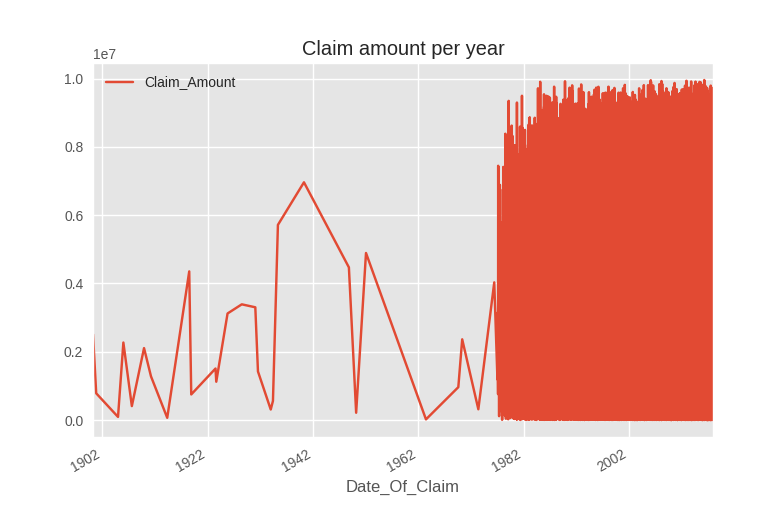


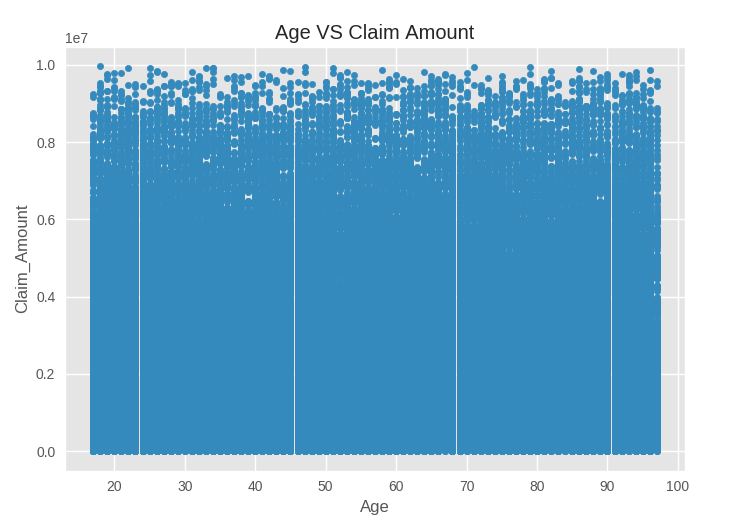


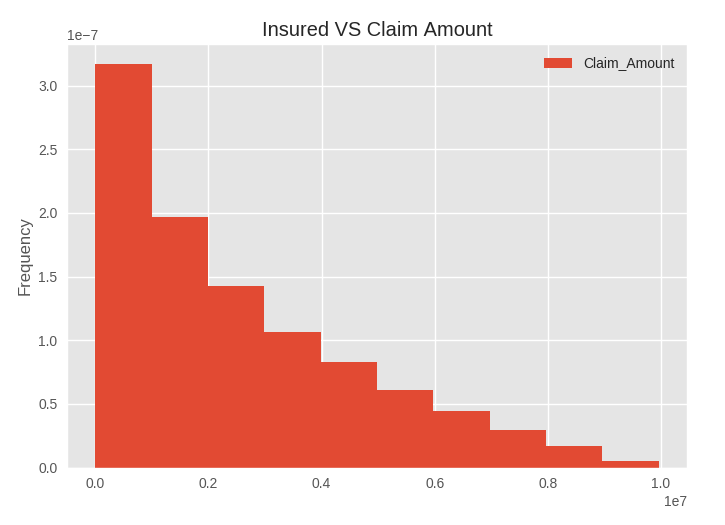
**** 

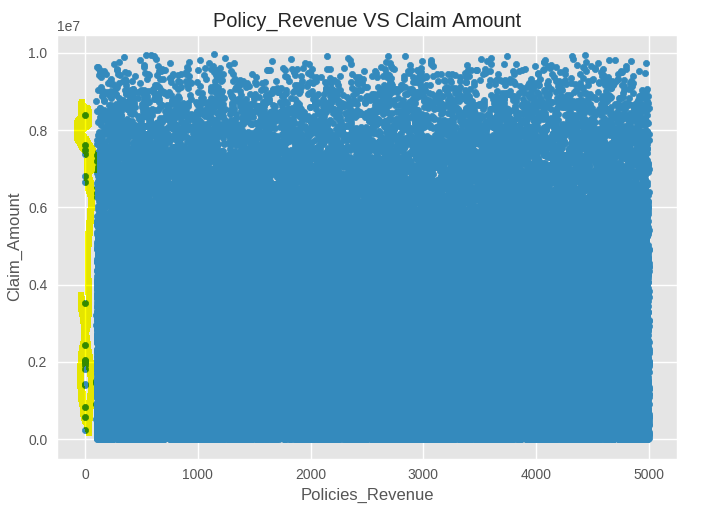
 











The yellow above shows the outliers which need to be cleaned to provide better machine learning properties.

1. **PPDM**
2. **Data Cleaning**
3. **Machine Learning**
4. **Bonus**