Analysis of Insurance Claims Dataset & Detection of Fraudulent Claims

By
Modika Ishwarya
Intern - CHUBB

Abstract:
An insurance claim is a formal request to the insurance provider to provide reimbursement against losses covered under the insurance policy. The need for this is to inform the insurer that to provide compensation as per the policy. But nowadays people are filing fraud claims to be benefited or to get profit. Fraudulent claims can be highly expensive for each insurer. Therefore, it is important to know which claims are correct and which are not. The main aim of this analysis is to get an understanding of insurance claims data, fraudulent claims, and detect which are fraudulent claims.

Data Collection:
The dataset is collected from www.kaggle.com.The dataset includes claims for a car insurance company in the United States. The data consists of 1000 individual claims. The most important variable of interest is fraud_reported. The data consists of numerical as well as categorical variables.

Approach:

- The python libraries used are Pandas, Matplotlib, Seaborn, Numpy, Scikit-learn, plotly.
  - Numpy is an external library in Python. It helps to work with arrays. It has an array object ndarray. The use of Numpy is to perform mathematical operations on arrays.
  - Pandas is used for data cleaning and data analysis purpose.
  - Matplotlib is used for data visualization.
  - Seaborn is a data visualization library formed based on matplotlib. It is used to visualize random distributions.
  - Scikit-learn provides a selection of efficient tools for machine learning and statistical modeling.

- Plotly is an interactive, an open-source plotting library.

❖ As part of data analysis, the data cleaning is performed to remove noise or missing values present if any.

❖ The statistical measures considered for the analysis of data are correlation matrix, mean.

Data Pre-processing

1. Checking for missing values :

- **Purpose:**
  - ➢ As part of data cleaning, it is important to check whether data is having any missing values as the presence of missing values affects the data analysis and may lead to wrong conclusions.

```
In [139]: data.isna().sum()

Out[139]: months_as_customer            0
          age                           0
          policy_number                 0
          policy_bind_date              0
          policy_state                  0
          policy_csl                    0
          policy_deductable             0
          policy_annual_premium         0
          umbrella_limit                0
          insured_zip                   0
          insured_sex                   0
          insured_education_level       0
          insured_occupation            0
          insured_hobbies               0
          insured_relationship          0
          capital-gains                 0
          capital-loss                  0
          incident_date                 0
          incident_type                 0
          collision_type              178
          incident_severity             0
          authorities_contacted         0
          incident_state                0
          incident_city                 0
          incident_location             0
          incident_hour_of_the_day      0
          number_of_vehicles_involved   0
          property_damage             360
          bodily_injuries               0
          witnesses                     0
          police_report_available     343
          total_claim_amount            0
          injury_claim                  0
          property_claim                0
          vehicle_claim                 0
          auto_make                     0
          auto_model                    0
          auto_year                     0
          fraud_reported                0
```

- **Inference:**
  - There are missing values in the data.
  - They are to be filled or removed based on the objective.

2. Filling Missing Values:

- The missing values in the column collision_type, police_report_available, and property_damage are filled with the mode of the respective columns.

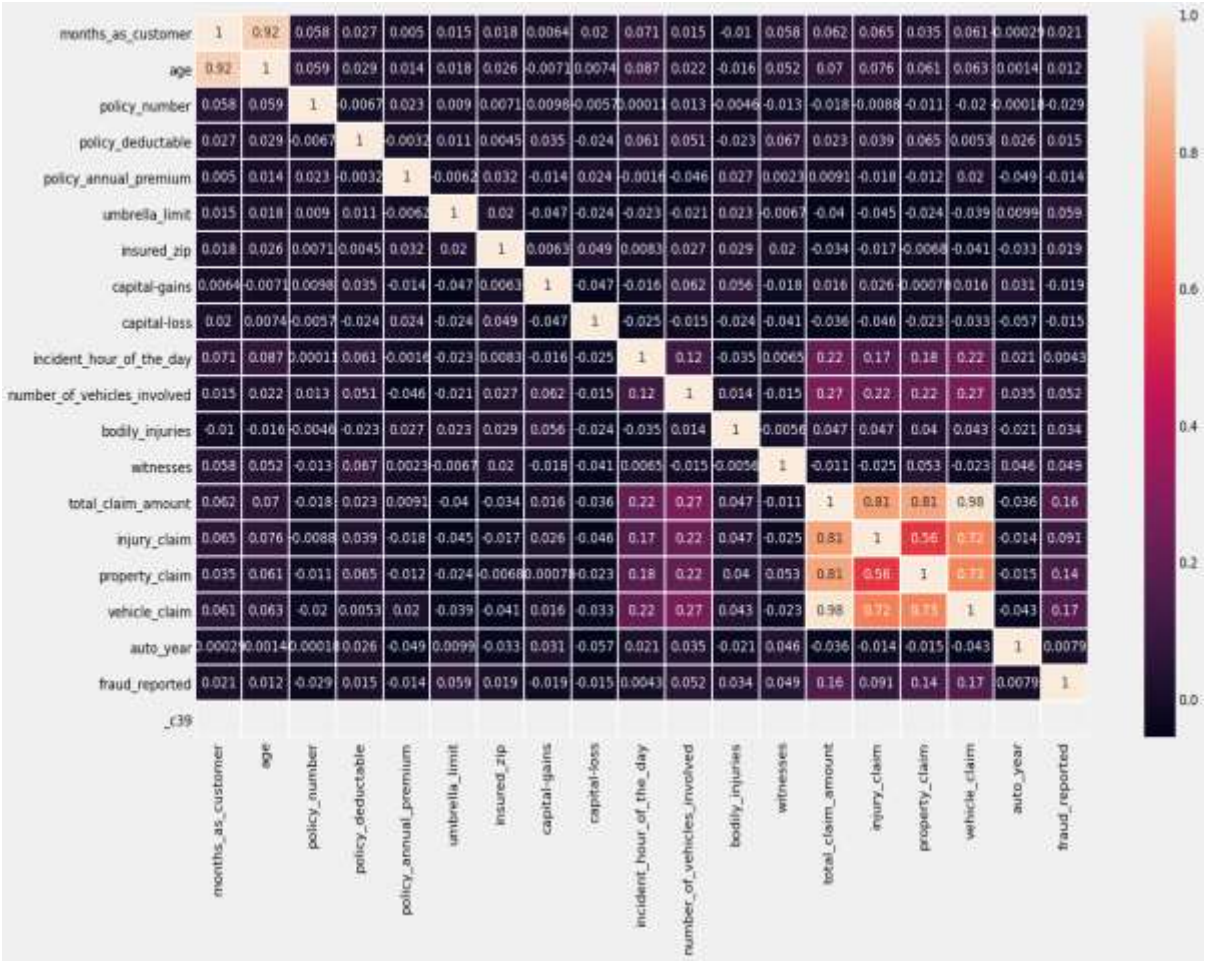
Data Analysis & Visualization:

1. Checking for the duplicate:

- Duplicates lead to overfitting of the model. So, it is necessary to remove them.
- There are no duplicate rows.

2. Descriptive Statistical Analysis:

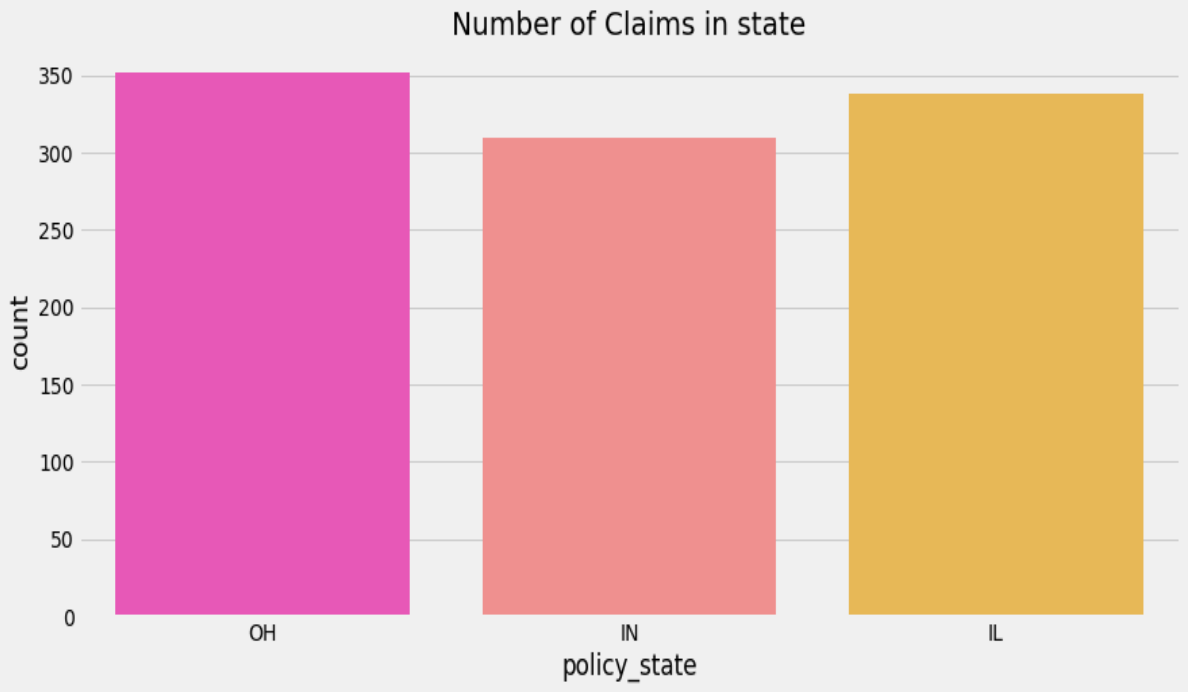
- **Finding Correlation between the variables:**
  - To determine the relationship between each pair of these columns.
  - This is done by plotting the correlation matrix.
  - Each cell in the grid represents the value of the correlation coefficient between two variables.
  - A value that is nearer to 1.0 indicates a strong positive correlation that is if the value of one variable increases, the value of the other variable increases.
  - A value that is nearer to -1.0 indicates a strong negative correlation that is if the value of one variable decreases with the other's increasing.



- There is a high correlation between age and months_as_customer variables.
- There is a high correlation between vehicle claim, total_claim_amount, property_claim, and injury_claim. The reason is that the total_claim_amount is the sum of vehicle claim,property_claim, and injury_claim column values.

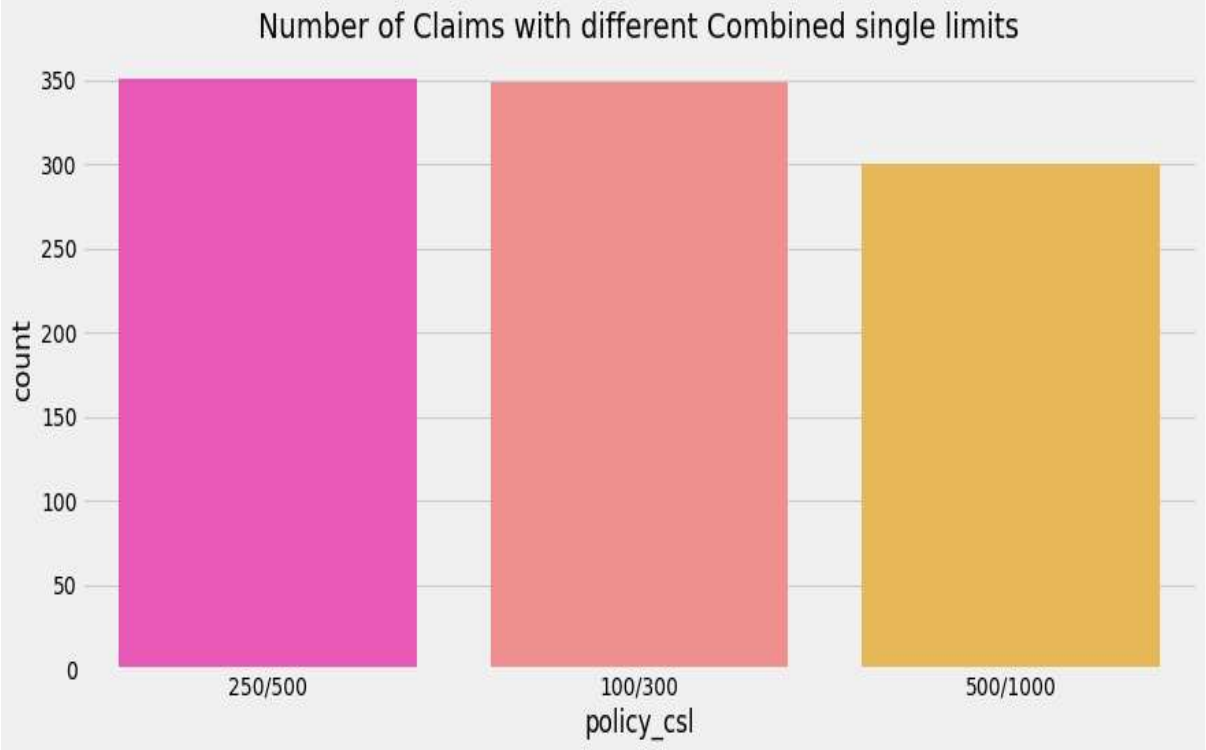
3. Number of Claims State wise:

- **Purpose:**
  - To obtain insight regarding the details of the state-wise claims.



- **Inference:**
  - The claims belong to states - Ohio, Indiana, Illinois.
  - The number of claims is more from Ohio state, by which it can be inferred that more people from Ohio claimed their auto insurance policy, that means people from
  - The number of claims is less from Indiana state.

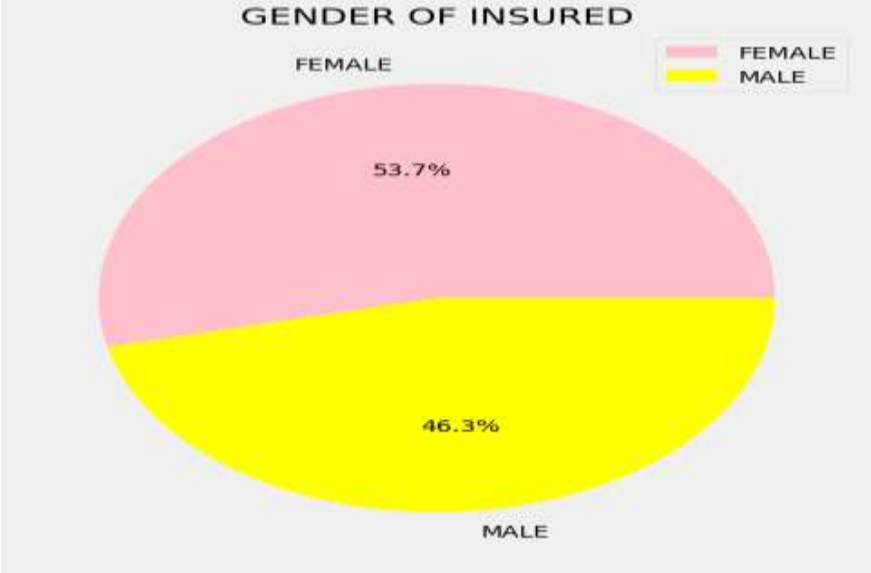
4. **Number of Claims with Combined Single Limit:**
   - **Purpose:**
     - To obtain insight regarding the details of the Combined Single limit liability.



- **Inference:**
  - There are more claims whose combined single limit is $500,000 and $300,000 and there are more claims with a coverage limit per person injury is $250,000,$100,000.

5. Distribution of Gender:

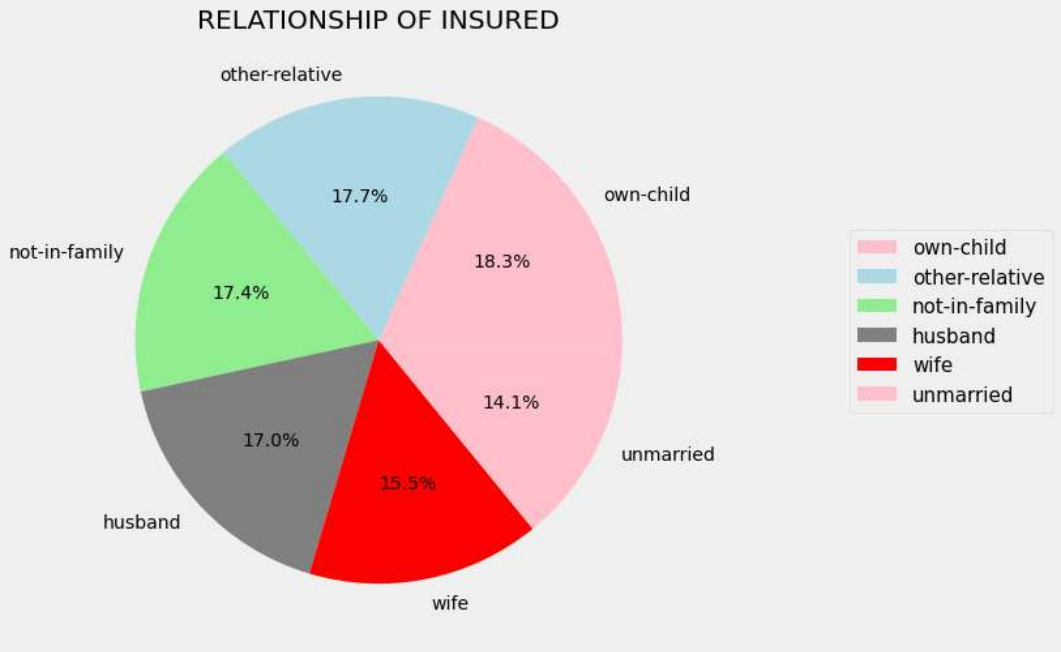
- **Purpose:**
  - To obtain insight regarding the details of the percentage of male & female.



- **Inference:**
  - The percentage of claims made by the male is 46.3%.
  - The percentage of claims made by the female is 53.7%.
  - This indicates that more claims were filed by female.

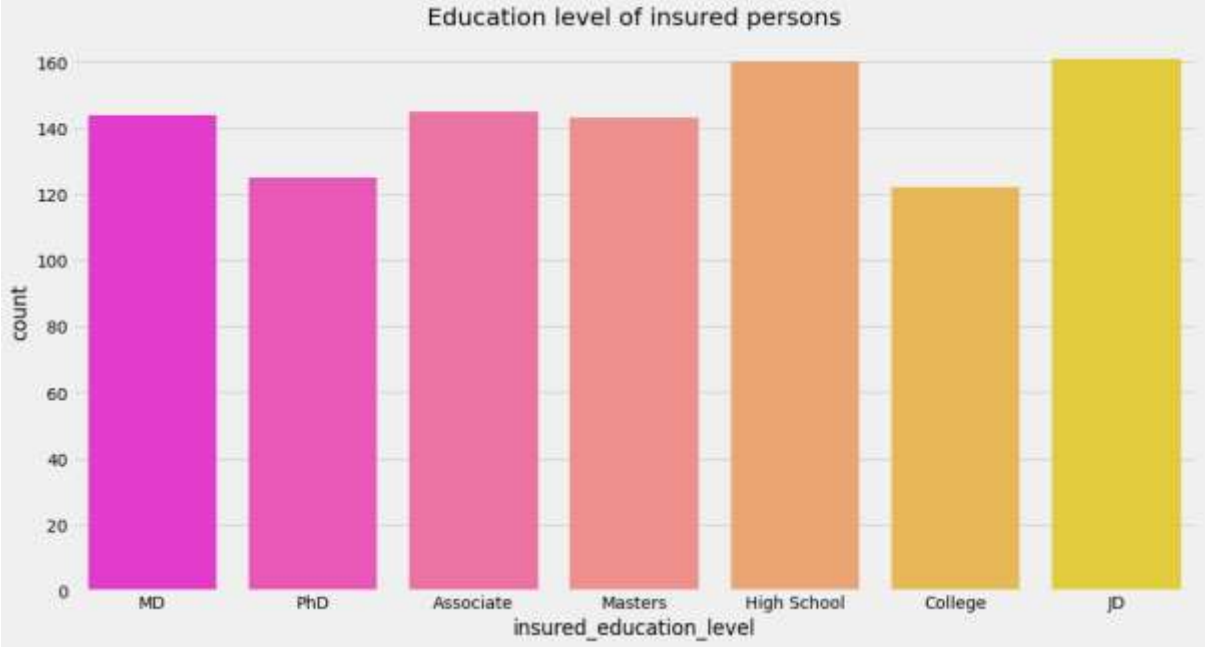
6. Insight on Insured_relationship:

- **Purpose:**
  - To obtain insight regarding the details of the relationship added by the insured in their policy.



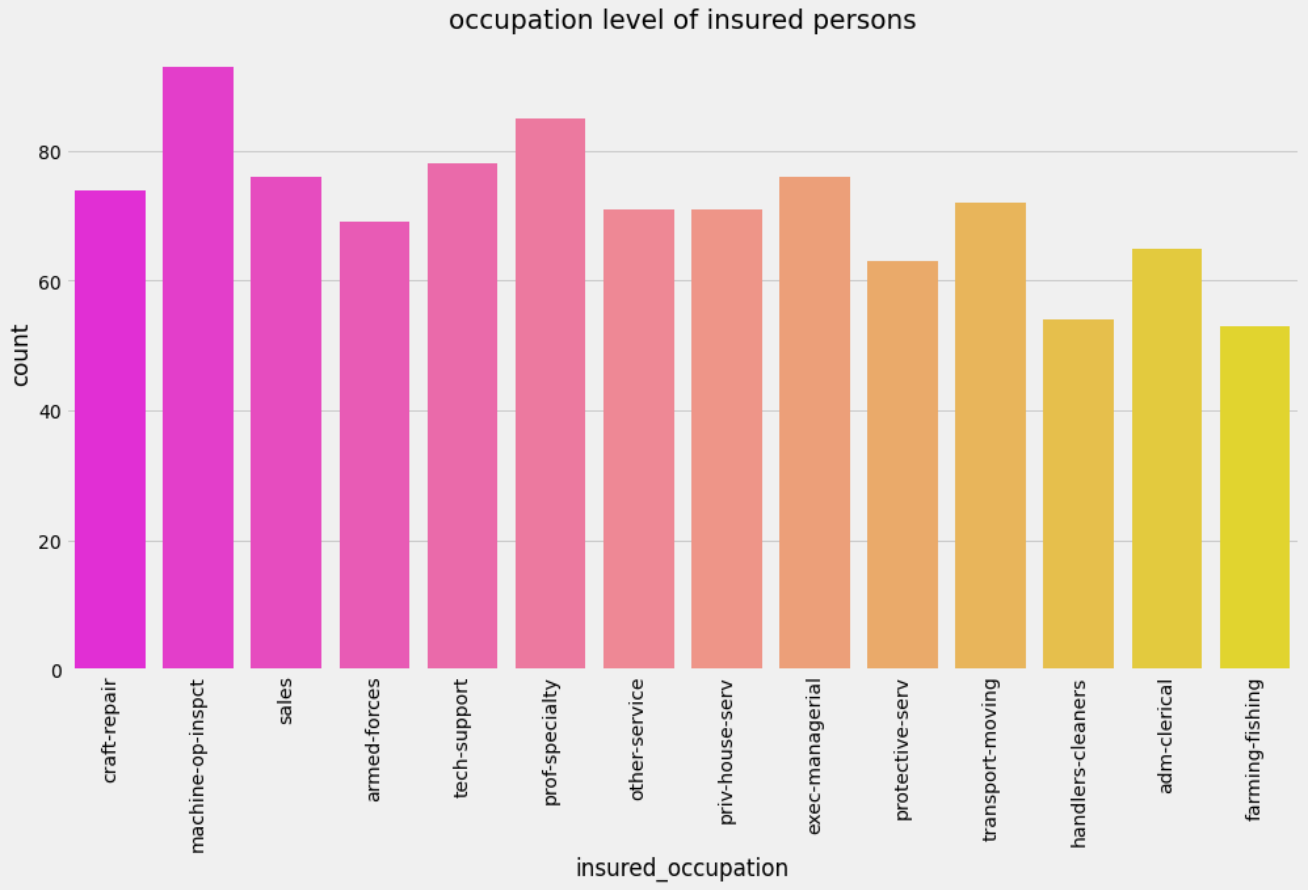
- **Inference:**
  - The 18.3% insured added people to the policy whose relationship with the insured is own-child.
  - The 17.4% insured added people to the policy whose relationship with the insured is not-in-family.
  - The 17.7% insured added people to the policy whose relationship with the insured is other-relative.

7. **Education level of Insured**
   - **Purpose:**
     - To obtain insight regarding the details of the education level of the insured.



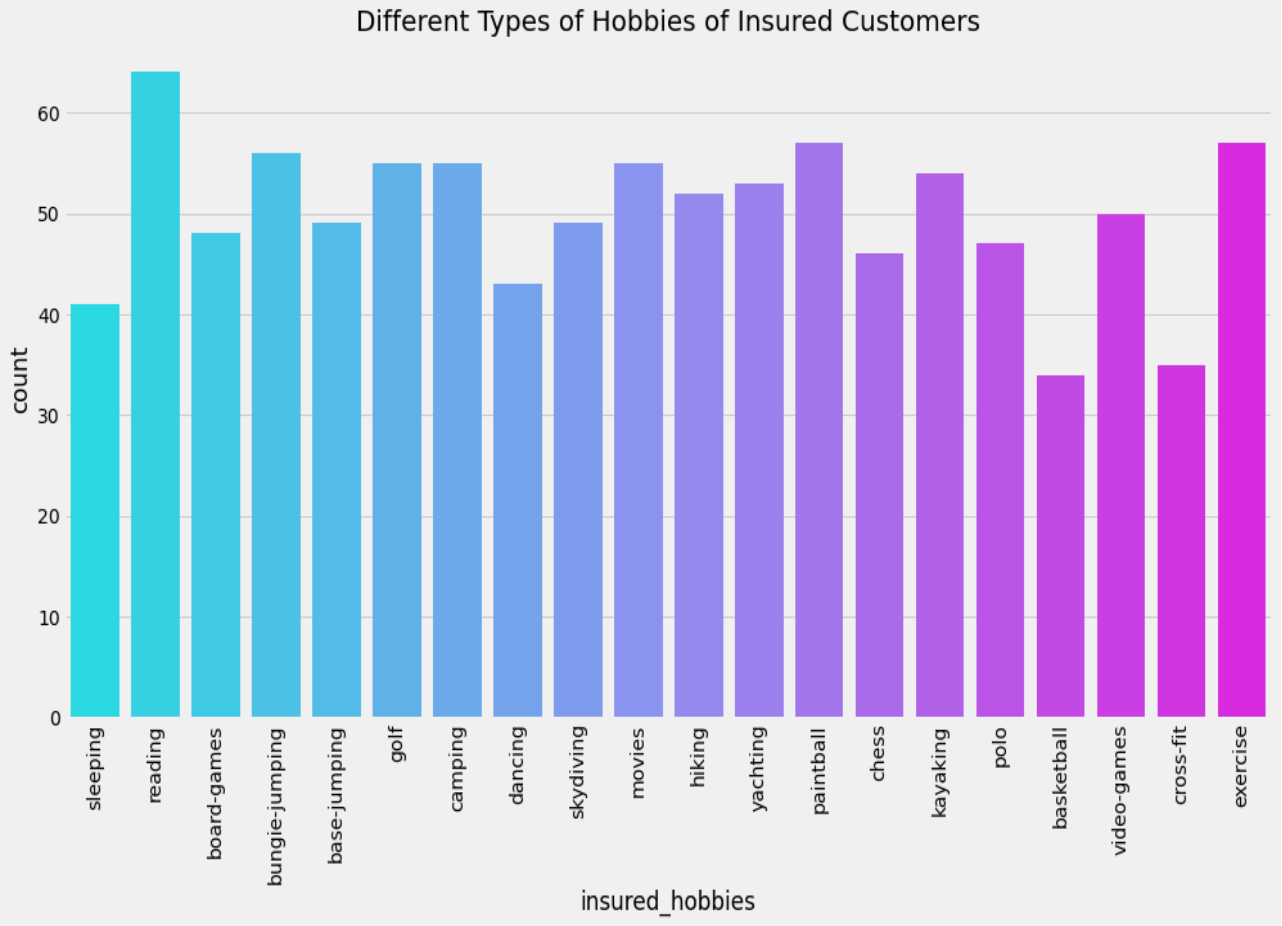
   - **Inference:**
     - Most of the insured people are qualified with education level of High School, JD.

8. **Occupation of the Insured:**
   - **Purpose:**
     - To obtain insight regarding the details of the occupation of the insured.



- **Inference:**
  - Most of the insured person's occupation is machine-op-inspct, prof-specialty.

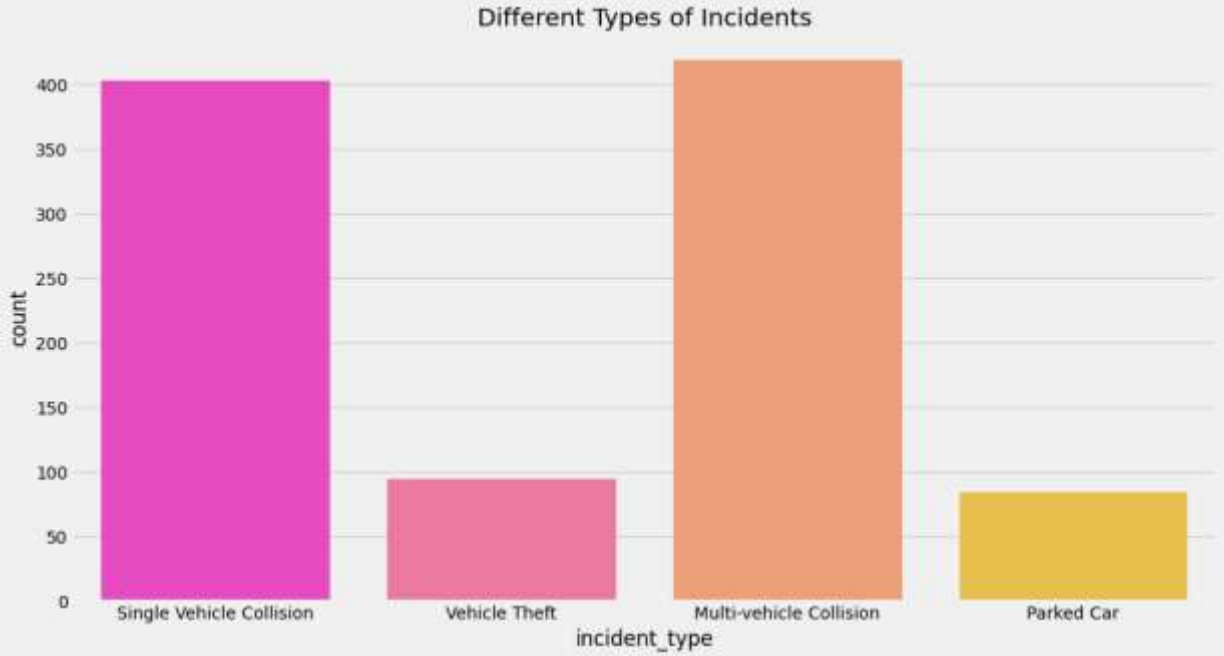
9. **Insured hobbies:**
   - **Purpose:**
     - To obtain insight regarding the details of hobbies of the insured.



- **Inference:**
  - Most of the insured people's hobby is reading, exercise.
  - Very few people have hobbies like basketball.

10. Incident Type:

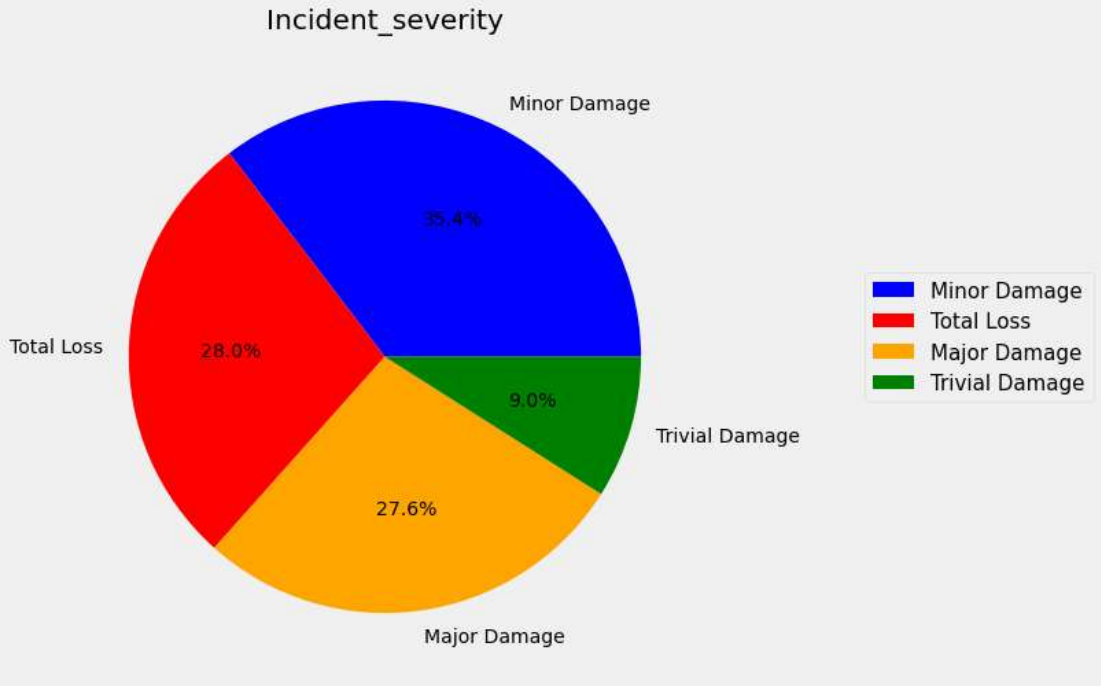
- **Purpose:**
  - To obtain insight regarding the details of hobbies of the insured.



- **Inference:**
  - Most occurred incidents are of type Multiple vehicle collision, Single Vehicle collision.

11. Incident severity

- **Purpose:**

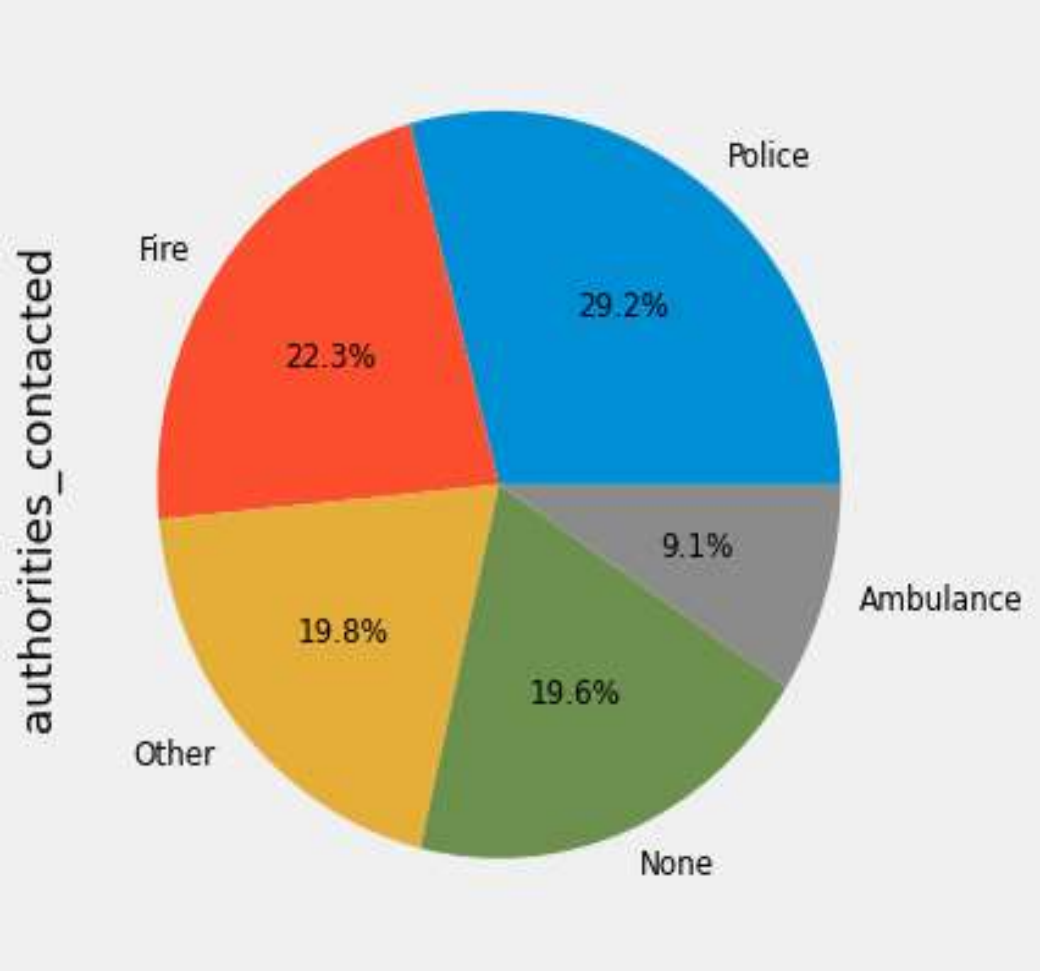


- **Inference:**
  - From the above pie chart, it can be inferred that 35.4% of incidents resulted in minor damage, and 28% of incidents resulted in total loss.

- Only 9% of incidents resulted in trivial damage.
- 27.6% of incidents resulted in major damage.

12. Authorities contacted:

- **Purpose:**
  - To obtain insight regarding the details of authorities contacted after the incident.

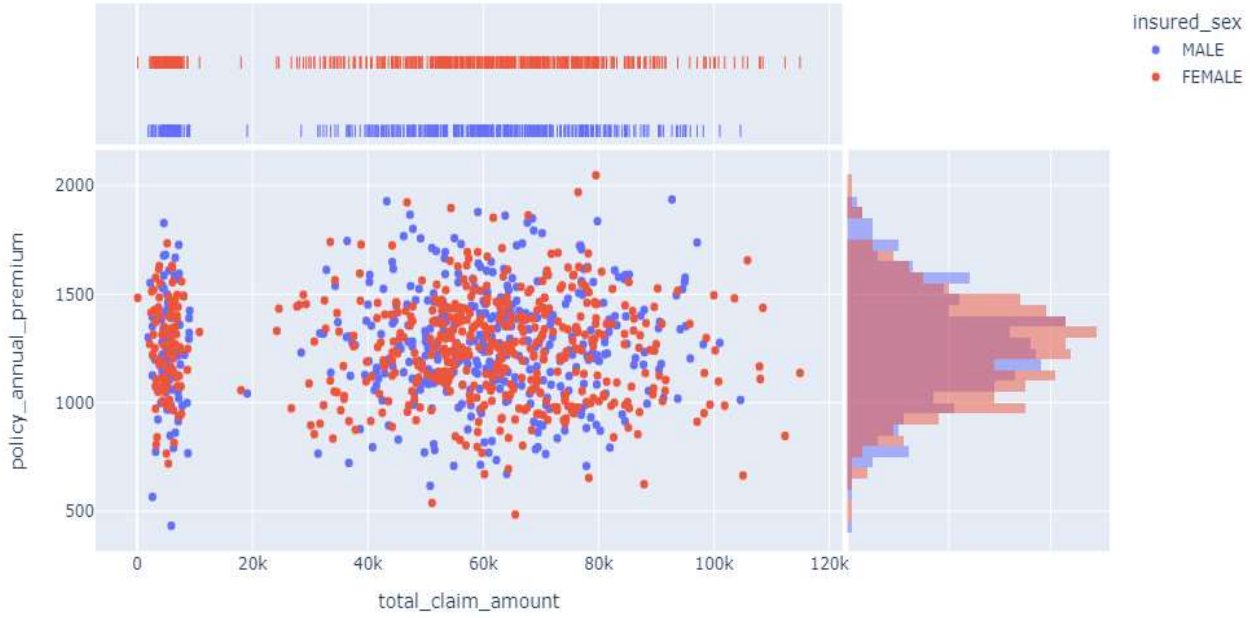


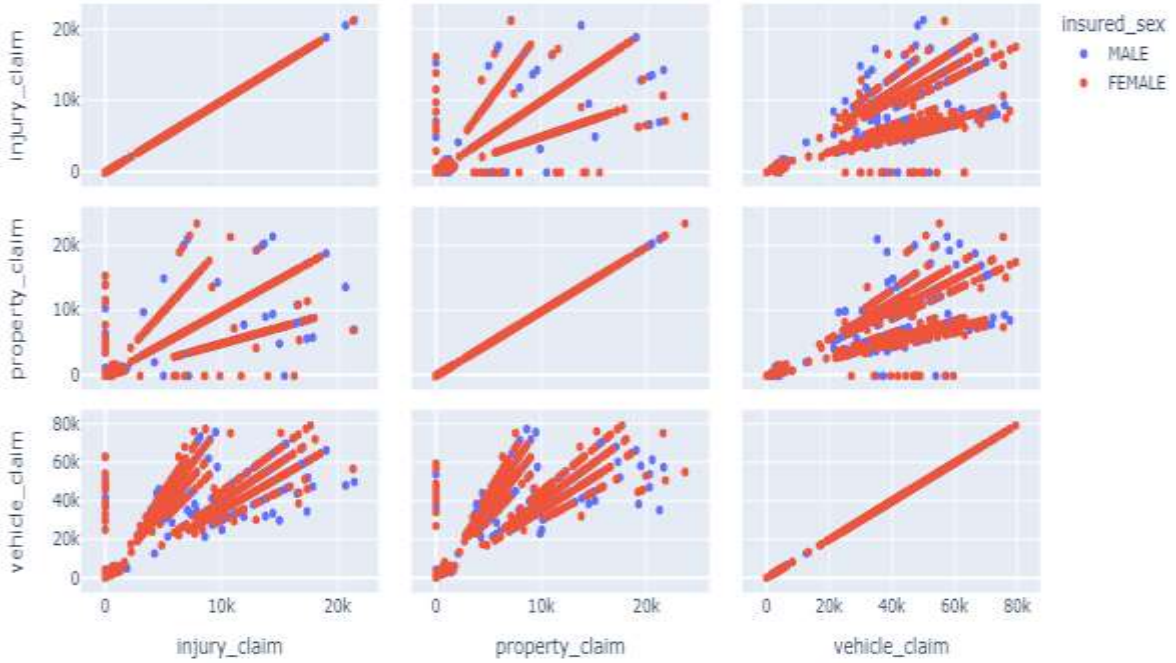
- **Inference:**
  - 29.2% of insured persons contacted police after the incident, this indicates that someone was injured in the other vehicle;
  - 22.3% of insured persons contacted fire authorities, this indicates that the car was damaged due to a fire accident.

- 9.1% of insured persons contacted Ambulance, this indicates that the incident caused bodily injuries.
- The remaining may be the case of minor or trivial damage.

13. Analysis of total claim amount, policy annual premium with respect to gender.

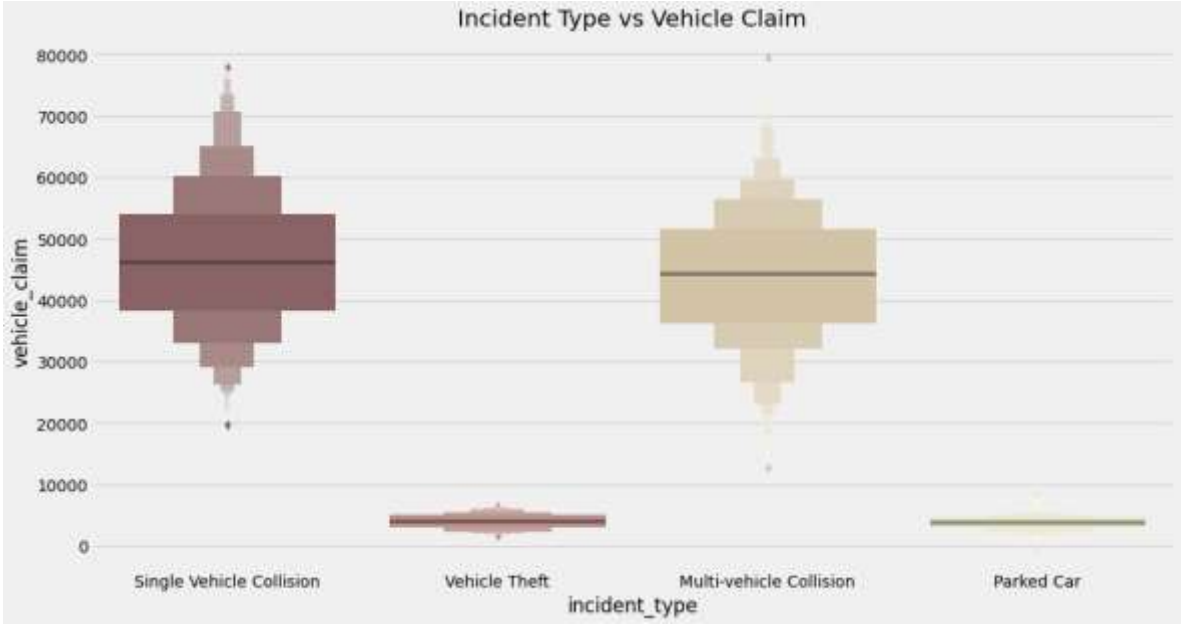
- **Purpose:**
  - The main purpose to file the claim is to get the total claim amount .
  - So, it is important to understand it with respect to premium and gender.





14. Incident type vs Vehicle Claim

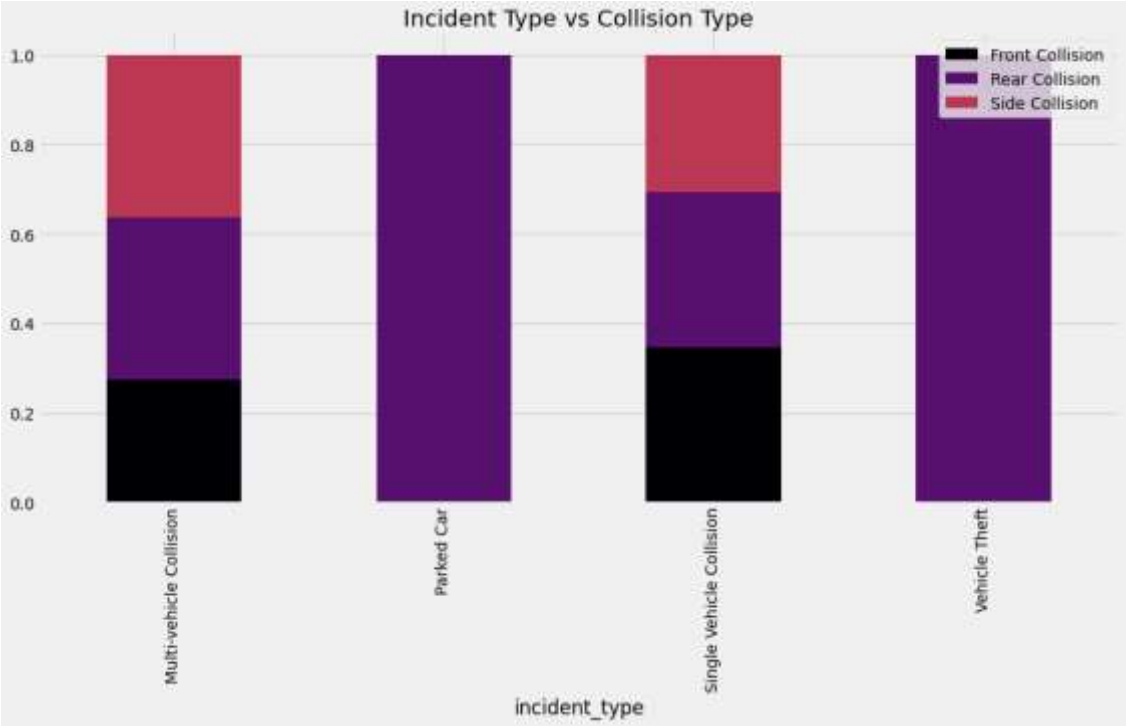
- **Purpose:**
  - To understand how vehicle claim depends on Incident type.



- **Inference:**
  - For single-vehicle collisions & Multi-vehicle collisions, the amount claimed for vehicle is more.

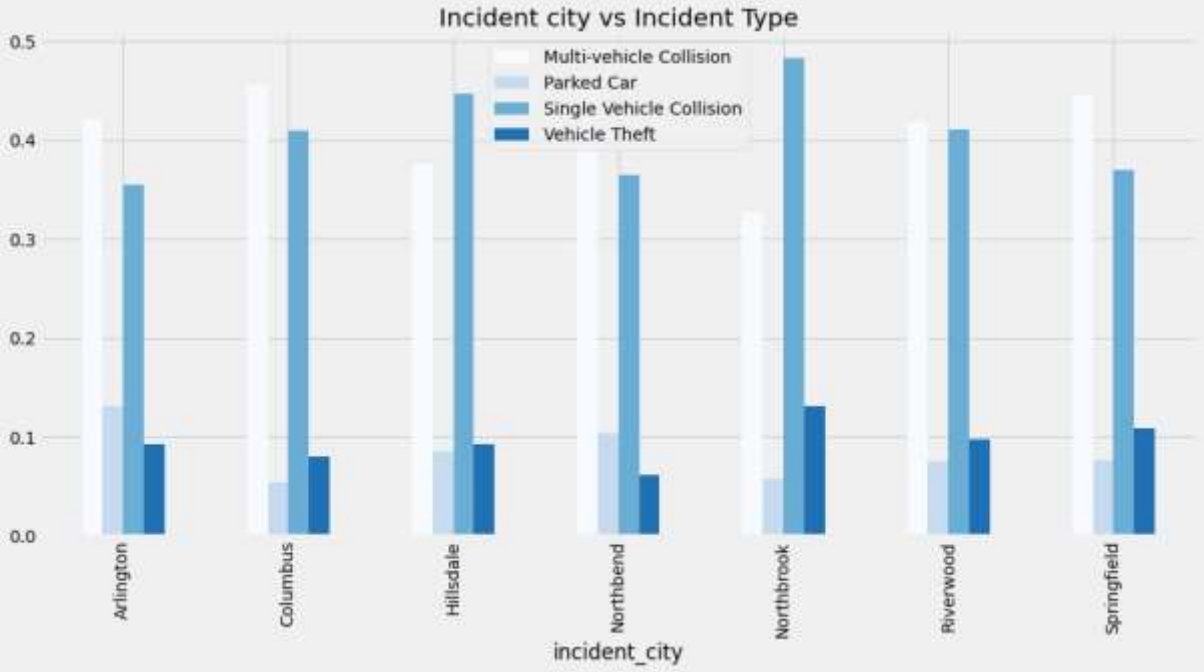
15. Incident type vs Collision type:

- **Purpose:**
  - To understand how incident type & collision type are related.



16. Incident city vs Incident type:

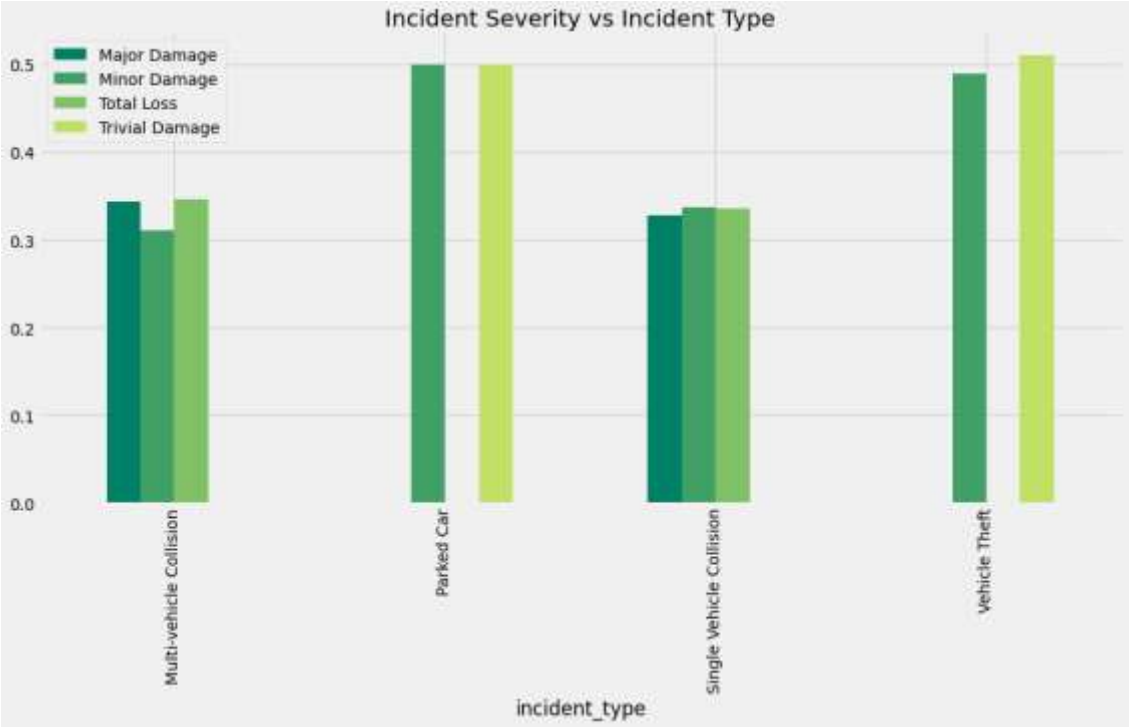
- **Purpose:**
  - To understand how incident type & incident city are related.



- **Inference:**
  - In Northbrook city more incidents have happened involving Single Vehicle collision.

17. **Incident severity vs Incident type:**

- **Purpose:**
  - To understand how incident severity & incident city are related.



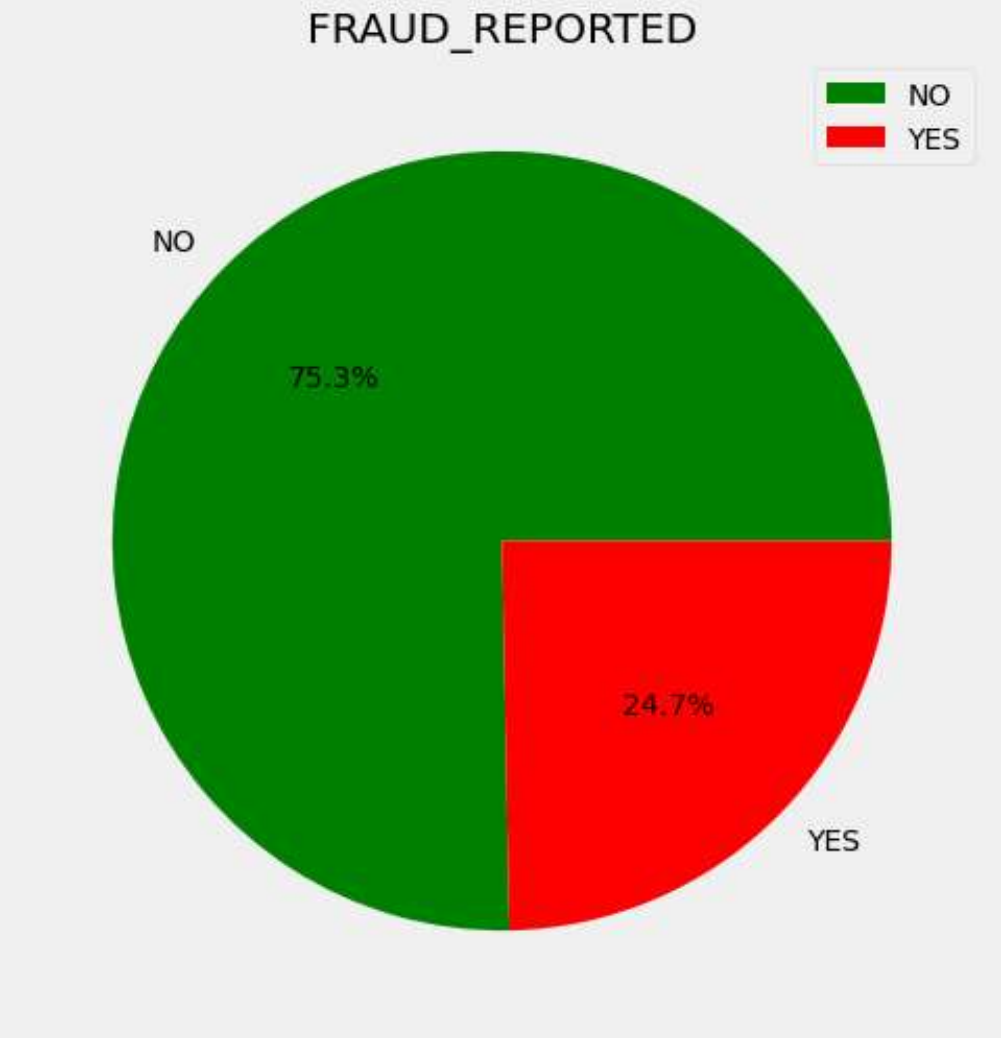
- **Inference:**

- The damage was major in multi-vehicle collision.

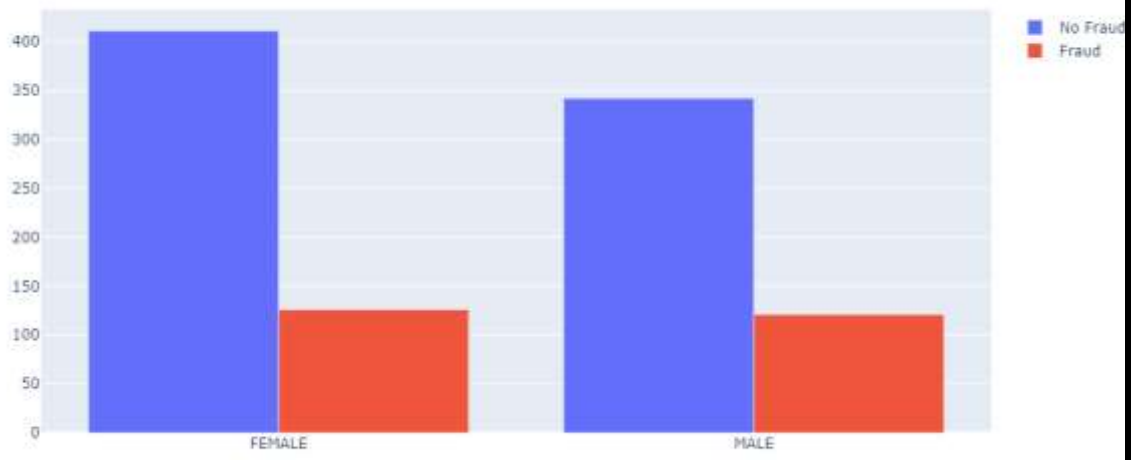
Analysis of variables with respect to target variable(fraud_reported)

1. Analysis of target variable:

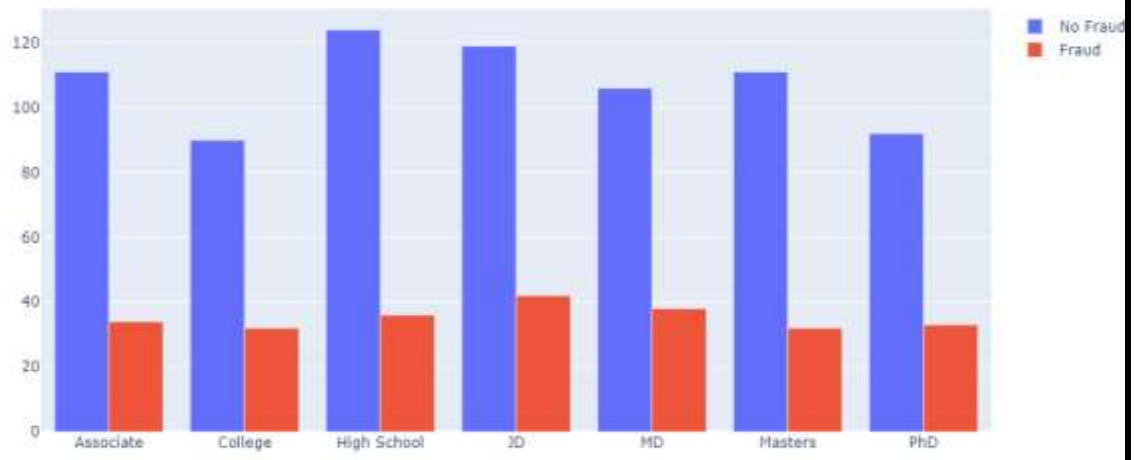
- fraud_reported is going to be our target column. We will encode Y to 1 and N to 0.



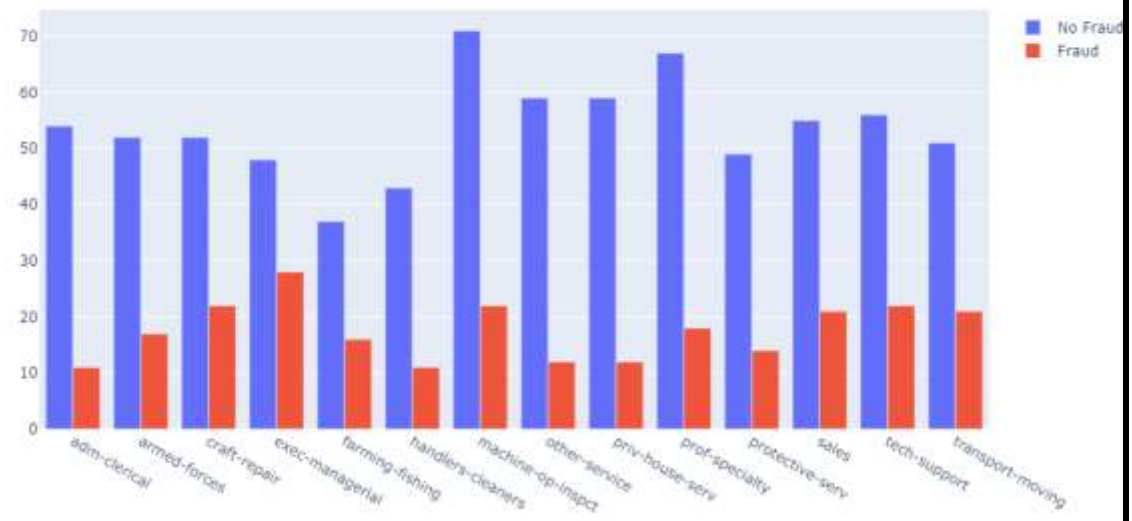
insured_sex vs. fraud_reported



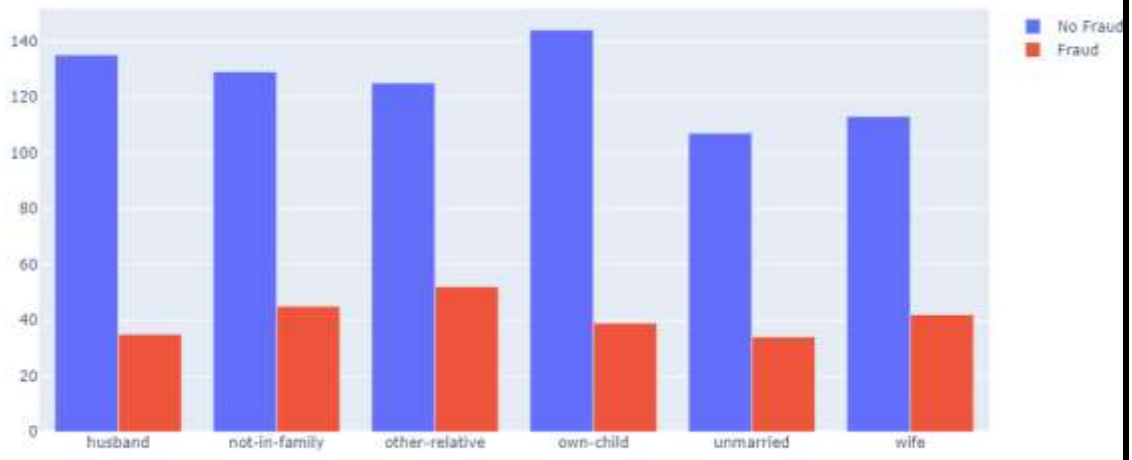
insured_education_level vs. fraud_reported



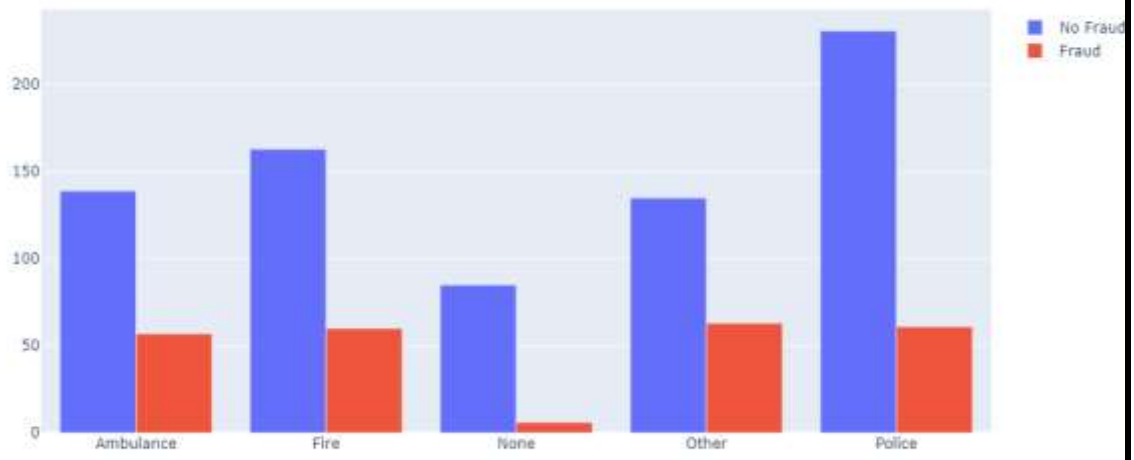
insured_occupation vs. fraud_reported



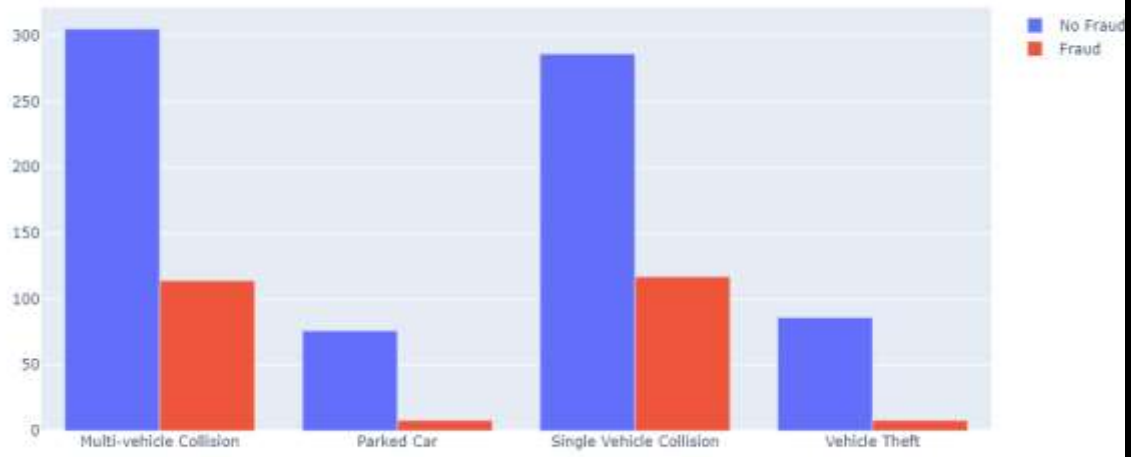
insured_relationship vs. fraud_reported



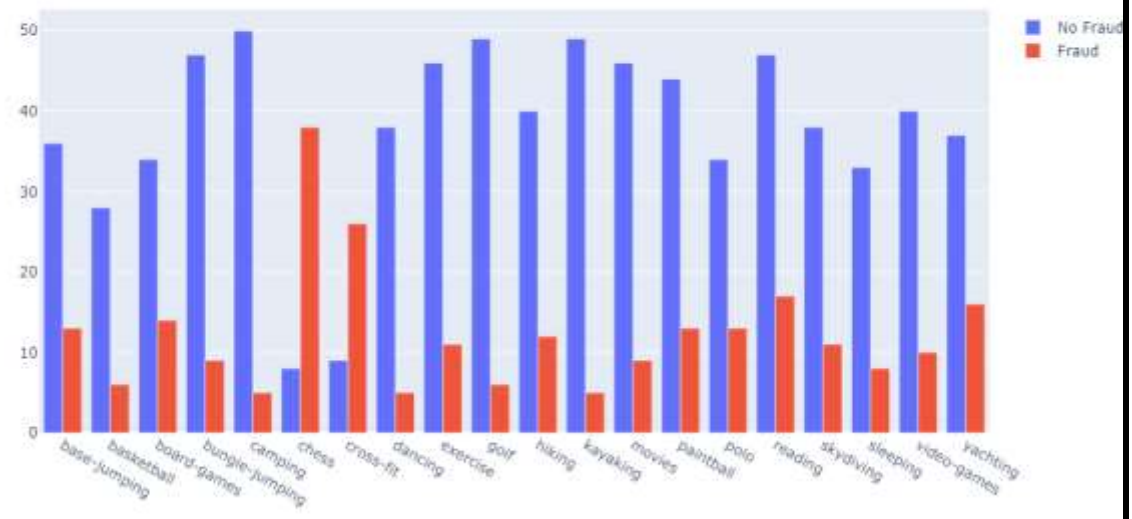
authorities_contacted vs. fraud_reported



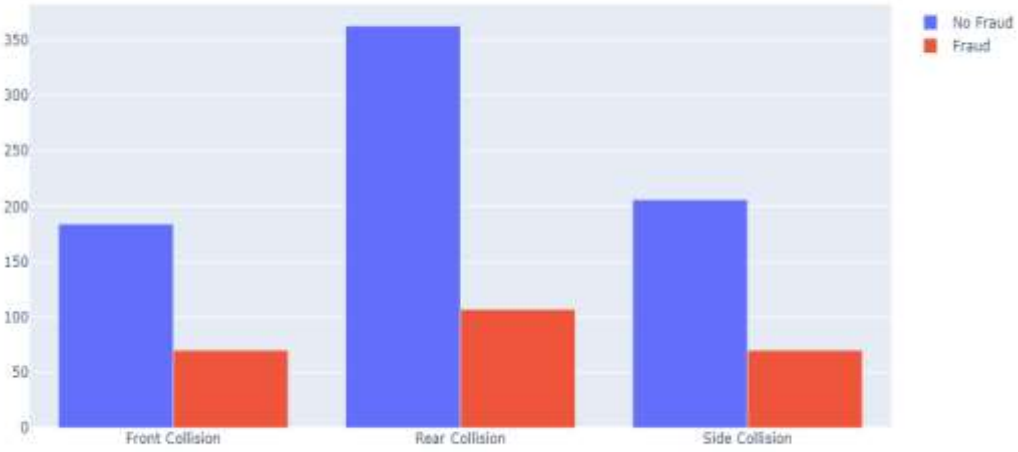
incident_type vs. fraud_reported

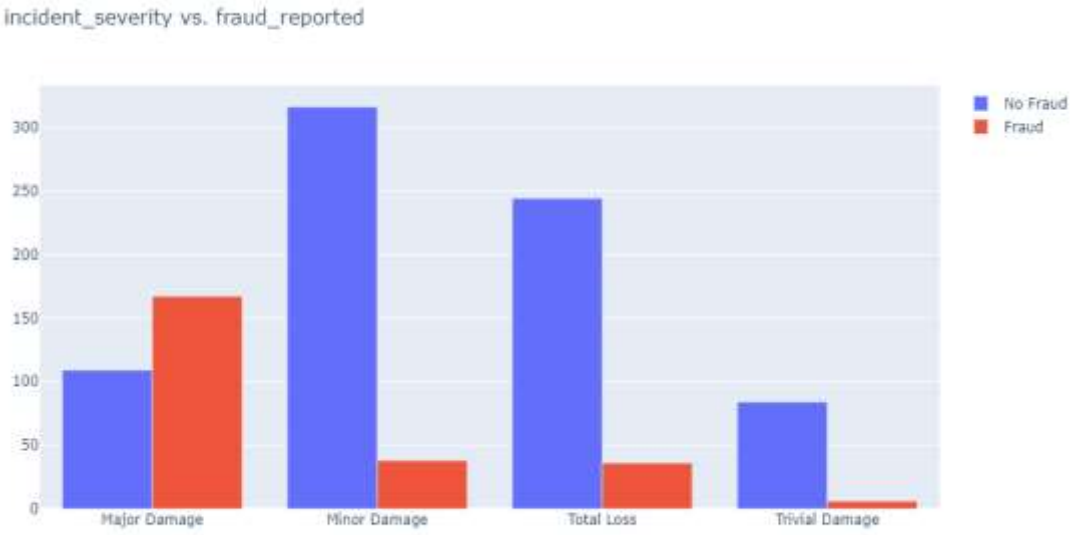


insured_hobbies vs. fraud_reported



collision_type vs. fraud_reported





- **Inference:**
  - ❖ It looks like people in exec-managerial positions have more frauds compared to other occupations. Sales, tech-support, and transport moving also have relatively high cases of fraud.
  - ❖ Multi-vehicle and single-vehicle collisions have more frauds compared to parked and vehicle theft. One of the reasons could be that in a collision, there is a high possibility of more damage to the car, as well as the passengers and hence there is a need to file false insurance claims.
  - ❖ While there are significant numbers of false claims in front and side collisions, rear collisions are the highest. This data is for the US and there, many people use dash cams while driving to record whatever is happening while they drive. In rear collisions, the footage from dashcams is not very helpful to conclusively prove whose mistake it was (insurance owner or other car owner). Maybe that is the reason for more fraudulent claims in rear collisions.
  - ❖ Riverwood city from SC state have claimed maximum amount of fraud for auto insurance.

- ❖ Here, compared to minor damage, total loss, and trivial damage, fraudulent claims are the highest in major damage. One reason could be that the high amount of repair cost which will be incurred by the insurer due to major damage.
- ❖ people with chess and cross-fit as a hobby have an extremely high number of fraudulent claims.
- ❖ People in the age group of 31-35 and 41-45 have more number of frauds.

Detection of fraudulent claims:

- The data given is split into training & testing data.
- K nearest neighbors & Random Forest algorithms are used
- The accuracy achieved with Random Forest is 82%.

Conclusion:

- ➢ The Random Forest Algorithm performed well .

```
Training accuracy of Random Forest is : 0.96
Test accuracy of Random Forest is : 0.82
[[181  10]
 [ 35  24]]
              precision    recall  f1-score   support

           N       0.84      0.95      0.89       191
           Y       0.71      0.41      0.52        59

    accuracy                           0.82       250
   macro avg       0.77      0.68      0.70       250
weighted avg       0.81      0.82      0.80       250
```