Data Analysis Report On Vehicle Insurance



**Prepared by:** Rutuja Harer

**Guide by:** Vishal sir

## Introduction:



This dataset provides a detailed overview of insurance claim data, likely used for fraud detection purposes. It contains 15,420 records with 33 features, encompassing information about claims such as the accident's date, vehicle details, driver characteristics, and claim specifics. Key columns include Make (vehicle manufacturer), Accident Area (urban or rural) and various details about the claimant's demographics and policy status.

## Executive summary:

### Key Findings:

In a fraud detection dataset, key findings often include spotting risky transactions that stand out

because of unusual amounts, times, or places. Fraud may happen more often in certain regions

or during specific times of the year. Fraudulent transactions also tend to share traits, like certain

types of purchases or user behaviors. Looking closely at user profiles can reveal that new accounts

or accounts with specific characteristics are more likely to be involved in fraud.

### Conclusion:

By understanding on which day accident happen we can suggest them to avoid car on that day,

or do not use car whose age is more than 7 years, and we also suggest the better policy for the

accident people.

### Recommendations:

* **Frequent Address Changes:** Claims where policyholders changed addresses within the last 6 months or 2-3 years show a slightly higher likelihood of fraud**.**
* **Higher or Lower Vehicle Prices:** Vehicles priced either very high (over 69,000) or low (under 20,000) have a modest association with fraud cases.
* **Policy Type:** Claims on "Sport" or "Utility" policy types, especially with collision coverage, are more correlated with fraud.

## Objective:

1. **Detect Fraud Claims:** Use the features to identify patterns that indicate fraud.
2. **Analyze Claim Patterns:** Study relationships between attributes (e.g., accident area, vehicle type) and fraud occurrence.
3. **Predictive Modeling :** Develop models to predict the likelihood of fraud in future claims.
4. **Identify High-Risk Factors:** Highlight factors (e.g., age, vehicle price) most correlated with fraud.

## Data overview:

#### Data source:

The dataset is publicly available on the Kaggle website and focuses on vehicle engagement with a Fraud detection.

* **Data Description :**

The analysis focuses on following key columns: make , Vehicle category , VehiclePrice , and many more things.

* + **Records**: [Number of rows -15421]
  + **Features:** [Number of columns - 33]

#### Data Pre-processing:

**Not Required columns:**

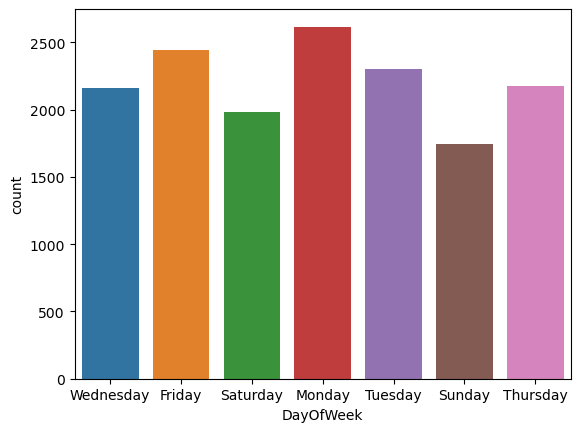
* There are total 33 columns are there.
* For categorical columns I have remove Fraud Found P and Rep Number

#### Methodology:

Load and clean the dataset using pandas, then perform numerical analysis with descriptive statistics and correlations using pandas and numpy. Visualize key patterns make , age , marriage status using seaborn.

**Numerical findings & Analysis using Visualisation:**

# Finding DayofWeek:



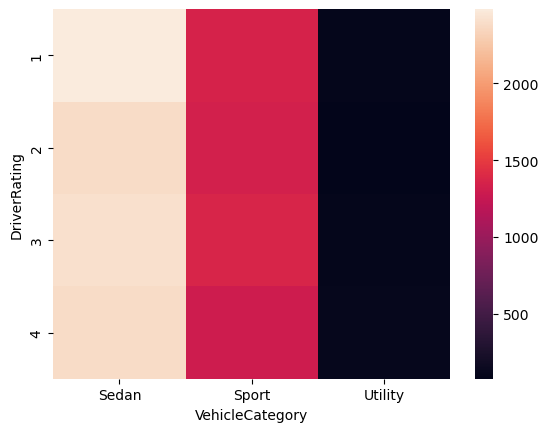
#### Key findings :

#### Findings Day of week in the column of day of week column present in dataset.

#### Plotting count plot graph for the day of week

#### After analysis this graph we get to know that more accident are happen on Monday value count is above 2500 in the comparison of other.

* **Analysis on DriverRating Vs vehicle Category :**



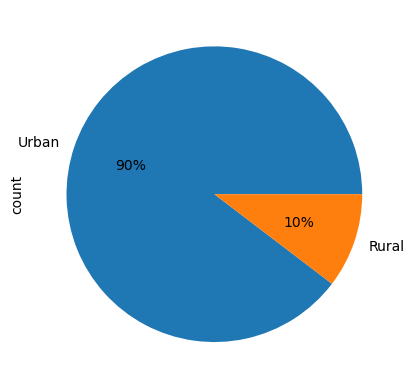
#### Key findings:

* Comparison between driver ratings and vehicle category.
* After working or analysis this driver rating vs vehicle category we can figure out that **sedan category** have more rating in the comparison on another 2

**Conclusion:**

We can suggest customer to use sedan category vehicle most or they should give first preference to sedan category cars then another.

**Top Accident location :**



**Key Findings :**

* Top areas where accident are happen most.
* IN urban areas most accident have happen in the comparison of other location we can see here **90% accident happen in urban area**
* In Rural area **only 10 % accident are happen**.
* **Analysis on Vehicle Price and Vehicle category:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Vehicle Price** | **20000 to 29000** | **30000 to 39000** | **40000 to 59000** | **60000 to 69000** | **less than 20000** | **more than 69000** |
| **Vehicle Category** |  |  |  |  |  |  |
| **Sedan** | **5435** | **2044** | **237** | **42** | **1010** | **903** |
| **Sport** | **2643** | **1480** | **183** | **43** | **86** | **923** |
| **Utility** | **1** | **9** | **41** | **2** | **0** | **338** |

**Key Finding:**

Which category how cheaper price and which category have higher price for this chart we can analyze that .

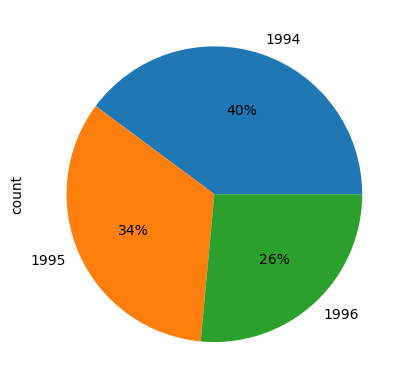
**Conclusion:**

Sedan category are securing highest number that category are used by customer most the value count is **5435 in the price range of 20000 t0 29000**

* **Analysis on Vehicle Price and Make**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | |  |  |  |  |  | | --- | --- | --- | --- | --- | |  |  |  |  |  | |  |  |  |  |   Vehicle |
| Category Sedan Sport Utility |
| Make |
| Accura 298 97 77 |
| BMW 9 6 0 |
| Chevrolet 971 579 131 |
| Dodge 67 42 0 |
| Ferrari 0 0 2 |
| Ford 283 131 36 |
| Honda 1570 1222 9 |
| Jaguar 3 3 0 |
| Lexus 1 0 0 |
| Mazda 1384 935 35 |
| Mecedes 4 0 0 |
| Mercury 58 17 8 |
| Nisson 18 6 6 |
| Pontiac 2564 1191 82 |
| Porche 0 2 3 |
| Saab 81 27 0 |
| Saturn 48 10 0 |
| Toyota 2148 972 1 |

**Years wise Accident pie chart**



**Key Findings :**

Year wise checking accident here we can see in **1994 more accident has happen 40%** and the accident cycle begin decrease year by year

**Conclusion:**

From this we can get idea that the Accident has been decreasing year by year.

**From dataset figure out Top 5 make and cheaper 5 make:**

**Top 5Make**

|  |
| --- |
| Pontiac 3837 |
| Toyota 3121 |
| Honda 2801 |
| Mazda 2354 |
| Chevrolet 1681 |

**Cheaper 5 Make**

|  |
| --- |
| Jaguar 6 |
| Porche 5 |
| Mecedes 4 |
| Ferrari 2 |
| Lexus 1 |

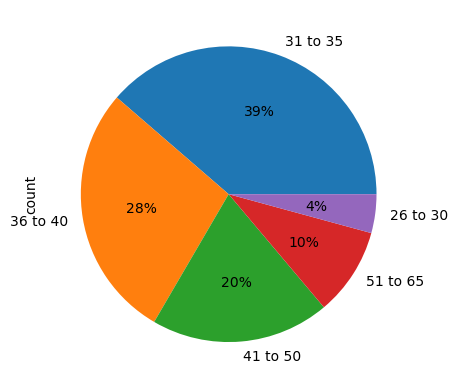
#### Key findings:

#### From Make column which are the top most 5 make are there and on the other hand top 5 cheaper make are there.

#### Conclusion:

#### The analysis reveals that which are top 5 make customer used most and what are the cheaper make so we can suggest customer what to buy.

**Policy analysis based on their Accident Happen across Age ,policy’s holder**



#### Key Findings :

#### The age between 31 to 35 have created more policy we can see percent her is 39%

#### 

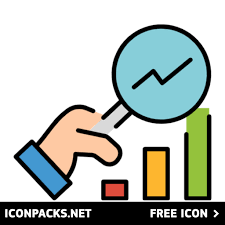
#### Conclusion:

#### The analysis shows that more accident happen in the age between 31 to 35 , and also send age group 36 to 40 accident are happen is 28%.

#### .

#### Which car sell most then what is the price of that and car is sell less then what is the price for that?

#### Sell Most



#### 

#### Toyota

|  |
| --- |
| Vehicle Price |
| 20000 to 29000 1958 |
| 30000 to 39000 530 |
| less than 20000 405 |
| more than 69000 208 |
| 40000 to 59000 19 |
| 60000 to 69000 1 |

#### 

#### Pontiac

|  |
| --- |
| Vehicle Price |
| 20000 to 29000 2263 |
| 30000 to 39000 884 |
| less than 20000 436 |
| more than 69000 187 |
| 40000 to 59000 65 |
| 60000 to 69000 2 |

#### Sell Less

#### Jaguar

|  |
| --- |
| Vehicle Price |
| 30000 to 39000 3 |
| 40000 to 59000 3 |

#### Porche

|  |
| --- |
| Vehicle Price |
| more than 69000 3 |
| 40000 to 59000 2 |

#### Ferrari

|  |
| --- |
| Vehicle Price |
| 40000 to 59000 2 |

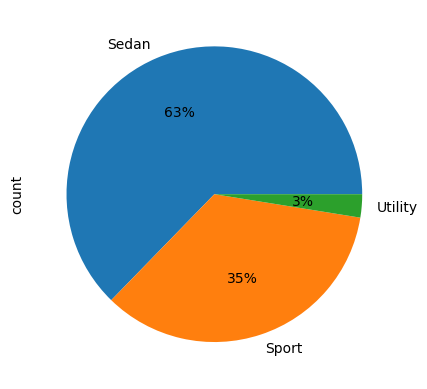
#### Key Findings : we can analyze that which car are sell most used by customers and price of that and which car sell less then price of that also.

#### Conclusion:

#### The analysis indicates that the car which have minimum Price are sell most and used by more customer in the comparison of high rated car are sell less.



#### Analysis of vehicle category are most used by customer.



#### Key Finding :

Here we can figure out that vehicle category are used by most user are which one and there percentage.

**Conclusion:**

The analysis reveals that users use the most category are sedan which one is **a 63%** are that next sport are used by **35%** and reaming utility is **3%.**

* **Recommendation**

**Frequent Address Changes**: Claims where policyholders changed addresses within the last 6 months or 2-3 years show a slightly higher likelihood of fraud.

**Higher or Lower Vehicle Prices**: Vehicles priced either very high (over $69,000) or low (under $20,000) have a modest association with fraud cases.

**Policy Type**: Claims on "Sport" or "Utility" policy types, especially with collision coverage, are more correlated with fraud.

**Enhanced Vetting**: Focus more rigorous checks on new or recent address changes.

**Higher Scrutiny on High/Low-Value Vehicles**: Claims on very high or low-priced vehicles could benefit from more comprehensive verification processes.

**Utilize Internal Agents**: Consider handling more claims with internal agents, as this may deter fraudulent submissions.



* **Conclusions**

The dataset contains 15,420 records and 33 columns, with a focus on various aspects of insurance claims, including policy details, vehicle information, claimant demographics, and indicators of potential fraud. By understanding on which day accident happen we can suggest them to avoid car on that day,

or do not use car whose age is more than 7 years, and we also suggest the better policy for the

accident people.

