Data Analysis Report

by
team Linear-Aggression

Aakash Sharma
Ayushi Mittal
Tanish Chugh

Key questions, objectives and insights

The key objective of this report is to analyze the given dataset and generate insights to figure out correlations different variables and Own Damage values calculated in the dataset. The analysis will be largely focused on Own Damage and factors associated with it.

Feature Engineering

Two new columns were defined in the dataset:

(1) Own Damage Ratio = (TOTALPAID_OD * MotorOD_IBNR * MotorOD_OCR) / Earned_OD *100

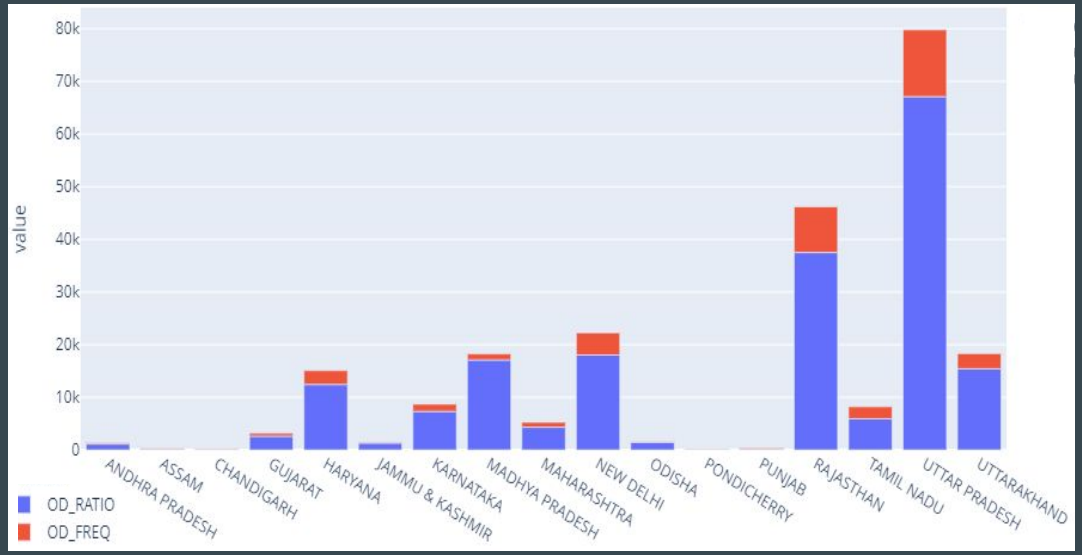
(2) Own Damage Frequency = (Intimated_ODClaims) / Earnedpolicies *100

Methodology and Rationale behind Analysis

- We have used Spearman Rank Correlation Coefficient to check correlation between two values. We have also calculated the P-value for the same.
- Spearman correlation test is used to check for a monotonic relationship between two variables which may be non-linear , but still increasing or decreasing. We decided to use this test as this test is quite useful when the data is non-normal or has outliers which both of which is the case with the data we have.
- In hypothesis testing, P-value is the probability of observing a test statistic at least as extreme as the one calculated from the data, assuming the null hypothesis is that there is no association between the two values being compared.

1.1 Trends of Own Damage across different states

Graph: Net value of OD Ratio and OD frequency across different states



Key Observations:

Uttar Pradesh, Rajasthan, New Delhi, Madhya Pradesh, Uttarakhand and Haryana are the leading states in highest values of both OD Ratio and OD Frequency, both of which are closely related.

States with lower OD values include Arunachal Pradesh, Bihar, Chhattisgarh, West Bengal Telangana, Jharkhand

Hypotheses:

1) Population: High OD Values belong to highly populated states like Uttar Pradesh and Rajasthan so population may be correlated with these values.

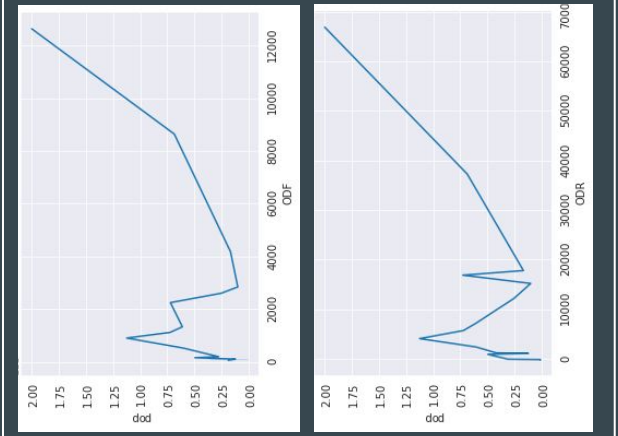
2) Theft reports: States like Delhi and Haryana, which have highest reports of theft are also states with high OD values thus theft report may also correlate with these statistics.

3) Road accidents: Rate of road accidents can be correlated to high OD value since states as Madhya Pradesh affirm this hypothesis.

4) Natural disaster: States more prone to property damage from natural disasters can have higher OD values.

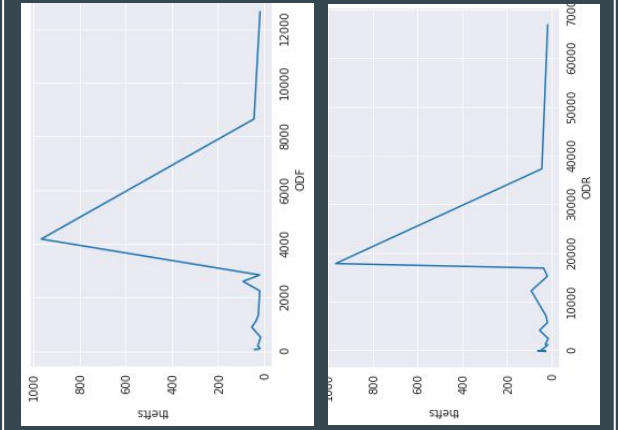
1.2.1 Testing out our hypotheses and drawing inferences

Hypothesis 1: Population of the state is correlated to sum of OD values.



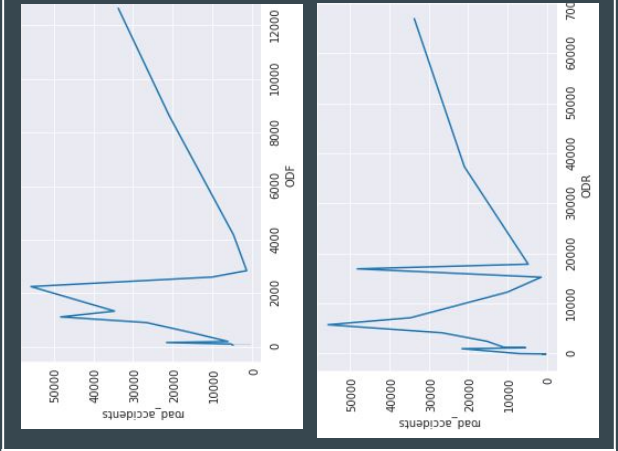
Statistical Significance	OD freq	OD ratio
correlation coefficient	0.434040	0.52451
P-value	0.081724	0.030654

Hypothesis 2: Number of reported thefts is correlated to OD values



Statistical Significance	OD freq	OD ratio
correlation coefficient	-0.049323	0.002451
P-value	0.850887	0.992551

Hypothesis 3: Annual number of road accidents is correlated to OD values



Statistical Significance	OD freq	OD ratio
correlation coefficient	0.383484	0.468137
P-value	0.128636	0.058071

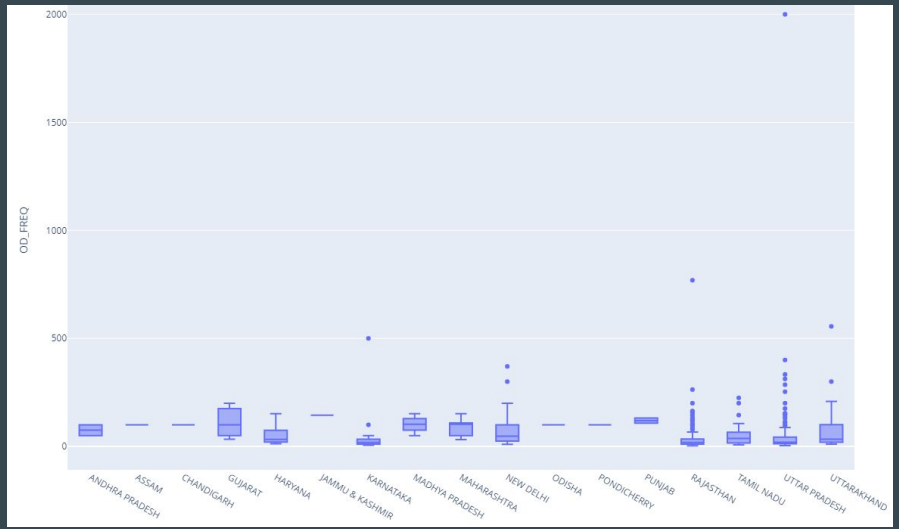
1.2.2 Results of our Hypotheses and possible biases

Population and OD values	Our statistical values indicate a somewhat strong correlation between the population of the state and the net Own Damage from that state. This is a somewhat obvious inference that higher population leads increases magnitude of OD values as well. People who get insurance are not a perfect sample of the overall state populations and the fact that the data collected is also not a perfect sample introduces some biases.
Theft and OD values	Number of thefts reported in states vs OD values comes out to have statistically no significant correlation. This is a surprising result but maybe explained with the help of biases and inaccuracies that might have piled up. The theft reports include thefts of all personal items and not just vehicles. Biases in data collection have also lead to this result as some states with high theft reports do not have enough representation to give clearer insights
Road accidents and OD values	Number of road accidents are also well correlated with the OD values and the simple explanation for this is that more road accidents lead to filing more insurance claims. We also get the insight that people from states with more road accidents are more likely to file for insurance damage.
Natural disaster and OD values	A statistical analysis of the correlation between Natural disaster and OD values could not be calculated due to unavailability of data as well as the highly unpredictable nature of natural disasters. Moreover, some states that are highly prone to property damage by natural disasters like the north-eastern states do not have enough data about them. But we can safely assume based on pure deduction that people from disaster prone area are more likely to file insurance claims.

2.0 Why is the company facing high OD frequency ? : Insights

- Entertaining customers who are more probable to file insurance claims, such as customers that live in accident prone or disaster prone areas.

- Not recognising customers whose payout sum can be a potential outlier. A handful of customers with very large insurance claims has lead to a net high value of OD frequency as evident from the graph below.



3.1 For the CEO: Implementations to balance the losses that the company is incurring and recommendations to increase policy sales

- Creating a system to predict the possible number of Initiated OD claims by a customer can be a powerful tool.

- Such a system, which will work on data analytics and machine learning should be able to predict OD frequency of a customer based on their profile. Attributes such as their income levels and their probability to encounter accident or property damage by natural disaster should be taken into account.

- Customers should be presented with customised policies. Customers who are labelled as “risky” by the above system need to be presented with lesser payout options on property damages

- Advertising the insurance company to people who are less likely to file claims maybe an excellent way to increase policy sales while avoiding losses that may occur with it.

- Heavily populated states which are less disaster and accident prone should be targeted for the same.

Thank you

The team

Aakash Sharma

Engineering Physics

2nd Year

IIT (BHU) Varanasi

Ayushi Mittal

Mathematics and Computing

2nd Year

IIT (BHU) Varanasi

Tanish Chugh

Chemical Engineering

2nd Year

IIT (BHU) Varanasi