ANALYZING INFORMATION ON INSURANCE COMPLAINTS

Naquibuddin Shaik (11602741)
Sneha Reddy Nayini (11618990)
Snehitha Donthi (11627234)
Hemanth Narayanan Sathiya (11606107)
Sai Charan Siddagoni (11694074)
Professor: Dr. Zeynep Orhan

INTRODUCTION

Since insurance offers stability and protection during regular operations, it is essential for both individuals and organizations. Insurance companies can find issues that need to be addressed for the entire industry by receiving and looking into complaints. Analyzing the dataset of complaints made against Texas insurance companies is the main goal. Insurance firms can learn a lot about areas that require development in order to ensure customer pleasure, regulatory compliance, and general service improvements from this investigation.

ABSTRACT

In our project we are aiming to study and analyse the Insurance details data, and our main objective is to comprehend grievances throughout a insurance period, detect trends, assess client demands, and investigate connections between insurance companies and complaint attributes. For this purpose we are using the data from The Texas Department of Insurance (TDI) to examine insurance complaints in Texas. Data on NAIC codes, firm names, product categories, complaint dispositions, resolution dates, complaint IDs, receipt dates, and justifications are all included in the dataset. "Complaint Reason," which reflects the reasons given by clients for making complaints, is the primary variable. Timeliness for resolution and complaint statuses are revealed by secondary goal variables. By providing insurance companies and regulatory bodies with pertinent insights, the project seeks to enhance customer happiness and regulatory compliance.

PROBLEM STATEMENT

- The project statement is to use a dataset from The Texas Department of Insurance (TDI) to evaluate insurance complaints in Texas and find patterns, trends, and connections between complaint features and insurance providers. By disseminating practical insights to pertinent parties, the objective is to use this study to enhance consumer happiness and regulatory compliance within the insurance sector.

OBJECTIVES

- Through in-depth research, have a thorough grasp of insurance complaints in Texas.
- Over time, look for patterns or trends in the insurance complaints that have been filed.
  Examine and evaluate the different requests that customers have made using the complaint information.
- Examine any connections that might exist between insurance providers and the quantity or nature of complaints they receive.
- Examine the present state of complaints and the average time it takes to resolve them.
- To improve consumer happiness and guarantee compliance, share pertinent data with insurance firms and regulatory agencies.

DATASET

- The Texas Department of Insurance (TDI) provided the dataset utilized in this investigation, which was obtained from data.Texas.gov. The dataset contains comprehensive details on every insurance complaint, such as the NAIC code, name of the company, kind of product, disposition of the complaint, date of resolution, complaint ID, date of receipt, and complaint explanation. "Complaint Reason," which reflects the client's rationale for submitting the complaint, is the main variable. The secondary target variables "Disposition" and "Resolution Date" include details on the complaint's state and the timeframe for its resolution, giving insights into the way insurance companies manage and handle complaints.

METHODOLOGY

Descriptive Statistics and Exploratory Data Analysis (EDA):

The main characteristics of the dataset, such as measures of central tendency, dispersion, and frequency distributions, will be compiled and presented using descriptive statistics. Visual exploration, pattern recognition, and anomaly detection will be achieved by the application of Exploratory Data Analysis (EDA) methodologies

Data Transformation and Encoding:

In order to make modeling and analysis easier, data pretreatment procedures will require converting categorical variables using encoders

Model Training and Predictive Analytics:

There will be training and testing sets within the dataset.
To forecast the "Complaint Reason" variable, supervised learning approaches such as logistic regression and XGBoost (XBbooster) will be utilized. This variable, which indicates the justification provided by customers for their complaints, will be useful in examining the many kinds and classifications of grievances lodged against insurance providers.

Analysis of Complaint Status and Resolution:

In order to comprehend the status of complaints and the timeliness of their response, secondary target variables like "Disposition" and "Resolution Date" will be investigated

Predictive Modeling :

Anticipating which businesses would receive the most complaints is an extra goal. This information will help determine which businesses receive complaints more frequently

Evaluation:

Then we use the different evaluation metrics namely F1 score, accuracy, Precision to find the model performance

```
Null_data=Input_Data.isnull().sum()
Null_data
```

```
Complaint number                0
Complaint filed against         0
Complaint filed by              0
Reason complaint filed          6
Confirmed complaint             0
How resolved                 1028
Received date                   0
Closed date                     0
Complaint type                  1
Coverage type                   0
Coverage level                  0
Others involved             28191
Respondent ID                   0
Respondent Role                 2
Respondent type                 0
Complainant type                0
Keywords                    49540
dtype: int64
```

```
Input_Data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 249169 entries, 0 to 249168
Data columns (total 17 columns):
 #   Column                  Non-Null Count   Dtype 
---  ------                  --------------   ----- 
 0   Complaint number        249169 non-null  int64 
 1   Complaint filed against 249169 non-null  object
 2   Complaint filed by      249169 non-null  object
 3   Reason complaint filed  249163 non-null  object
 4   Confirmed complaint     249169 non-null  object
 5   How resolved            248141 non-null  object
 6   Received date           249169 non-null  object
 7   Closed date             249169 non-null  object
 8   Complaint type          249168 non-null  object
 9   Coverage type           249169 non-null  object
 10  Coverage level          249169 non-null  object
 11  Others involved         220978 non-null  object
 12  Respondent ID           249169 non-null  int64 
 13  Respondent Role         249167 non-null  object
 14  Respondent type         249169 non-null  object
 15  Complainant type        249169 non-null  object
 16  Keywords                199629 non-null  object
dtypes: int64(2), object(15)
memory usage: 32.3+ MB
```

```
# Impute missing values in 'How resolved' column with mode
mode_reason_complaint_filed = Input_Data['How resolved'].mode()[0]
Input_Data['How resolved'] = Input_Data['How resolved'].fillna(mode_reason_complaint_filed)

# Impute missing values in 'Others involved' column with mode
mode_complaint_type = Input_Data['Others involved'].mode()[0]
Input_Data['Others involved'] = Input_Data['Others involved'].fillna(mode_complaint_type)

# Impute missing values in 'Keywords' column with mode
mode_respondent_role = Input_Data['Keywords'].mode()[0]
Input_Data['Keywords'] = Input_Data['Keywords'].fillna(mode_respondent_role)

# Impute missing values in 'Reason complaint filed' column with mode
mode_respondent_role = Input_Data['Reason complaint filed'].mode()[0]
Input_Data['Reason complaint filed'] = Input_Data['Reason complaint filed'].fillna(mode_respondent_role)
```

```
null_d=Input_Data.isnull().sum()
null_d
```

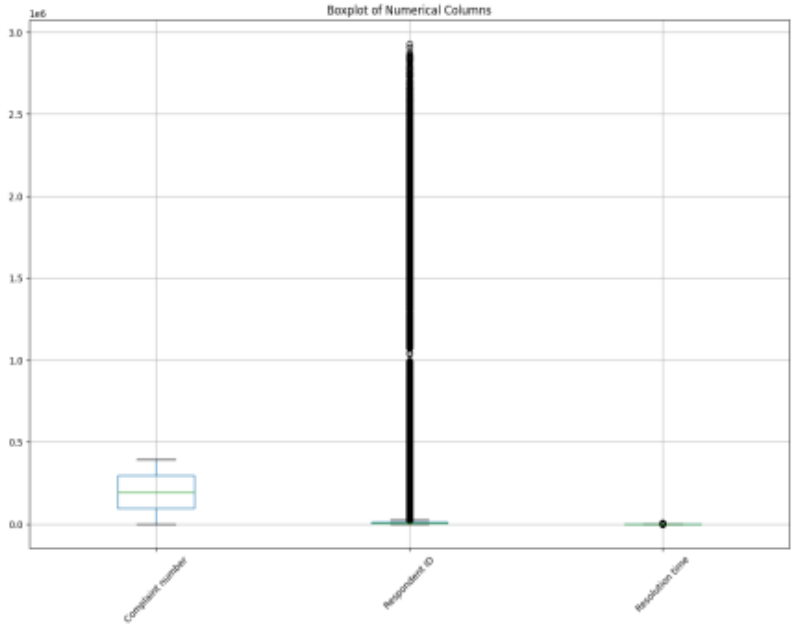
```
Complaint number            0
Complaint filed against     0
Complaint filed by          0
Reason complaint filed      0
Confirmed complaint         0
How resolved                0
Received date               0
Closed date                 0
Complaint type              0
Coverage type               0
Coverage level              0
Others involved             0
Respondent ID               0
Respondent Role             0
Respondent type             0
Complainant type            0
Keywords                    0
dtype: int64
```

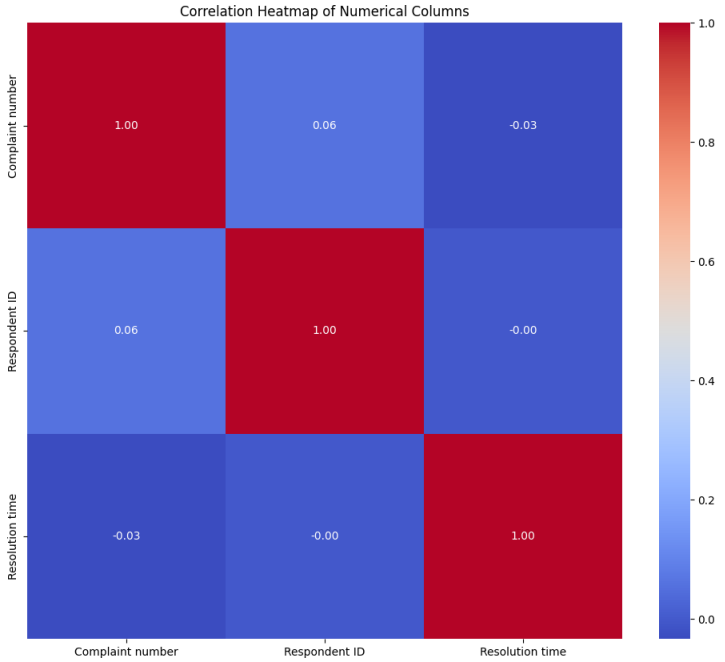
EDA

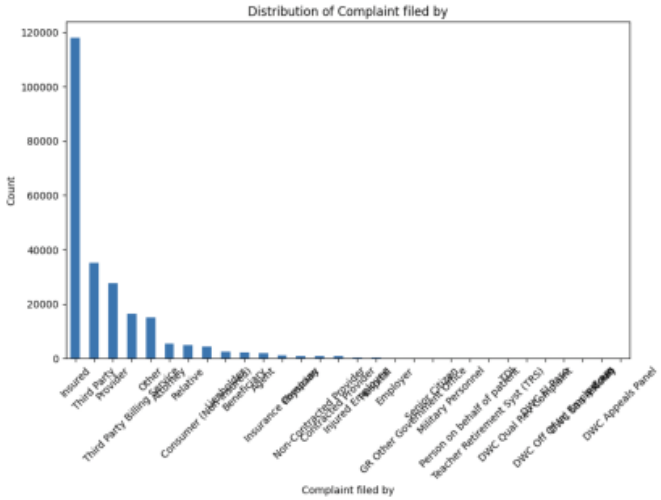
WE HAVE NULLS IN DIFFERENT COLUMNS, WE HAVE HANDLED THEM BY REPLACING THE MISSING VALUES WITH MODE.

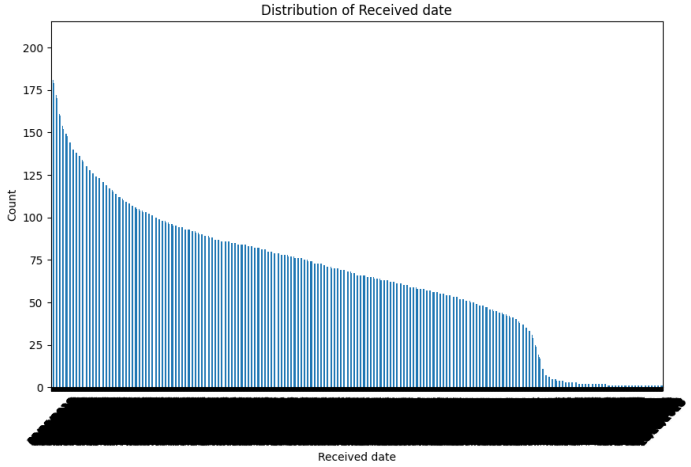
EDA

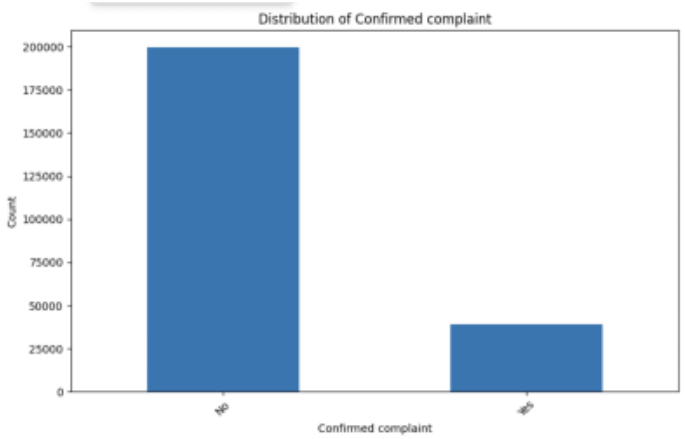
- WE HAVE DRAWN HEATMAPS TO FIND THE CORRELATION BETWEEN NUMERICAL VARIABLES.
- THEN DRAWN THE BOXPLOTS TO OBSERVE THE OUTLIERS IN NUMERICAL COLUMNS.



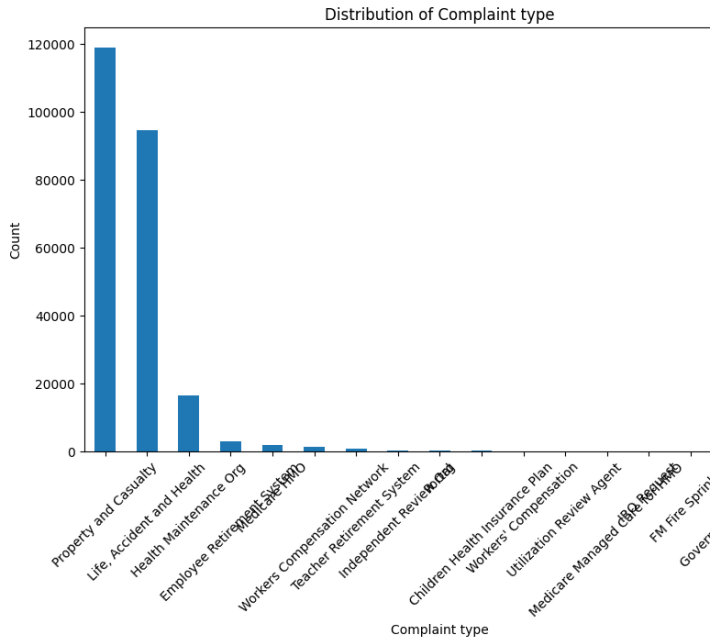


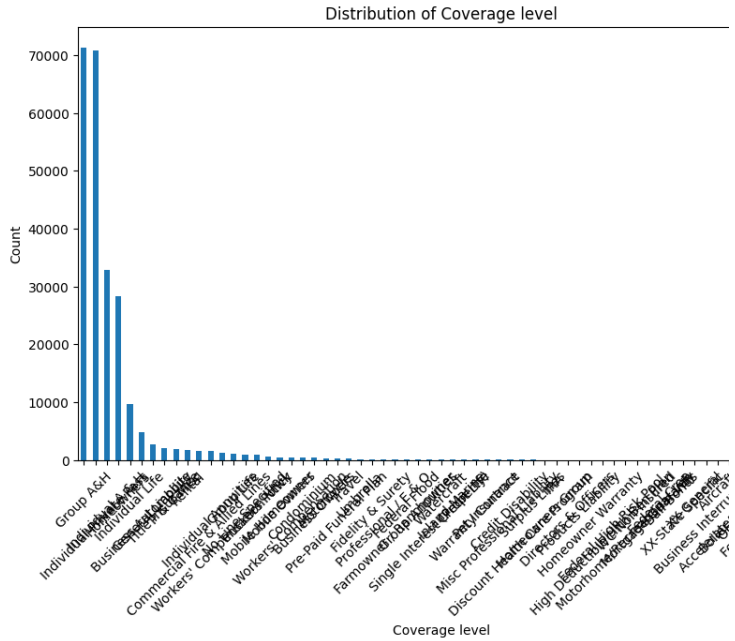

Distribution of Complaint filed by
Count
Complaint filed by
Insured
Third Party
Provider
Relative
Employer
DWC Appeals Panel

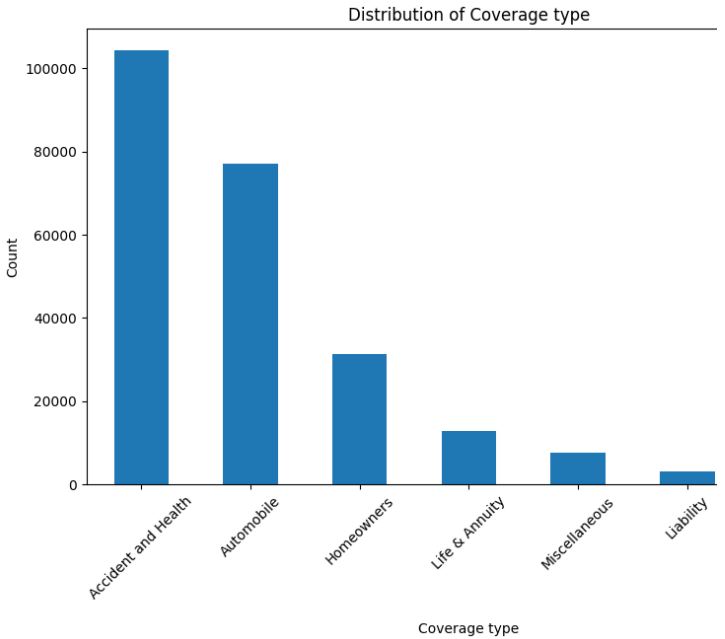

Distribution of Received date
Count
Received date


Distribution of Confirmed complaint
Count
Confirmed complaint
No
Yes

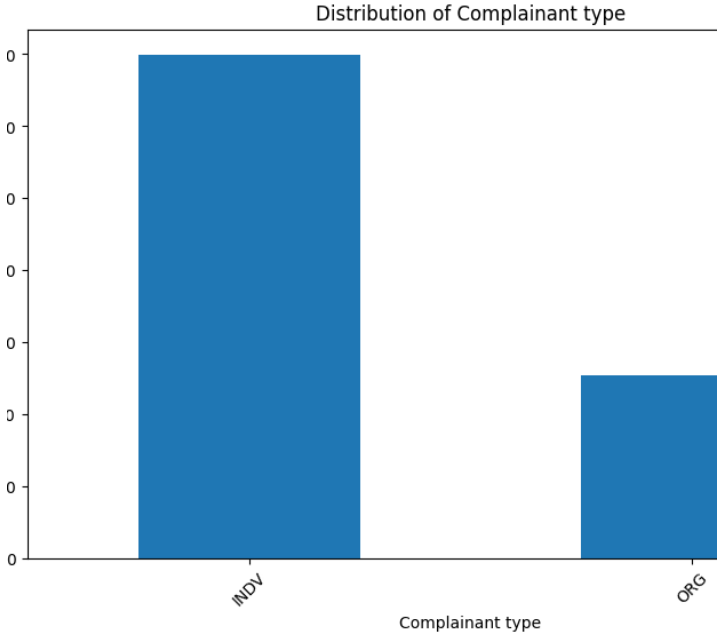
EDA

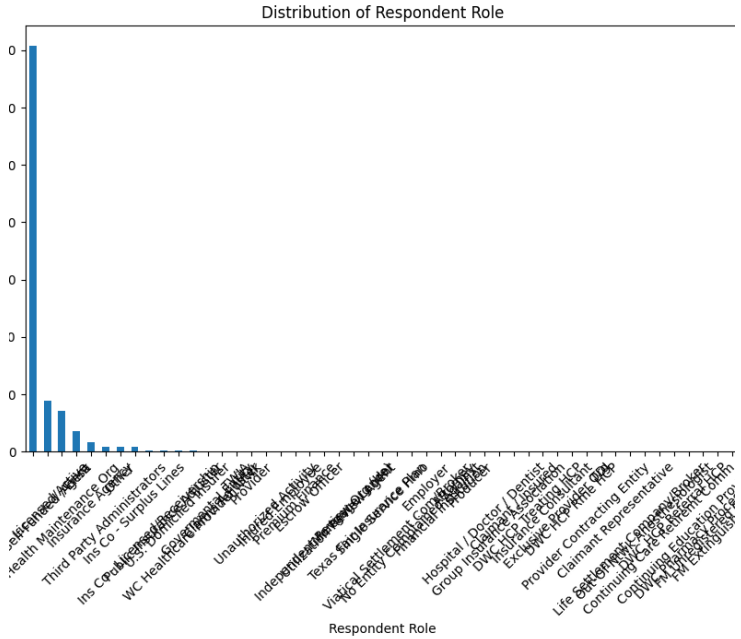

Distribution of Complaint type
Count
Complaint type
Property and Casualty
Life, Accident and Health
Health Maintenance Org
Workers Compensation Network
Teacher Retirement System
Independent Review Org
Children Health Insurance Plan
Workers' Compensation
Utilization Review Agent
Medicare Managed Care Referral
FM Fire Spri

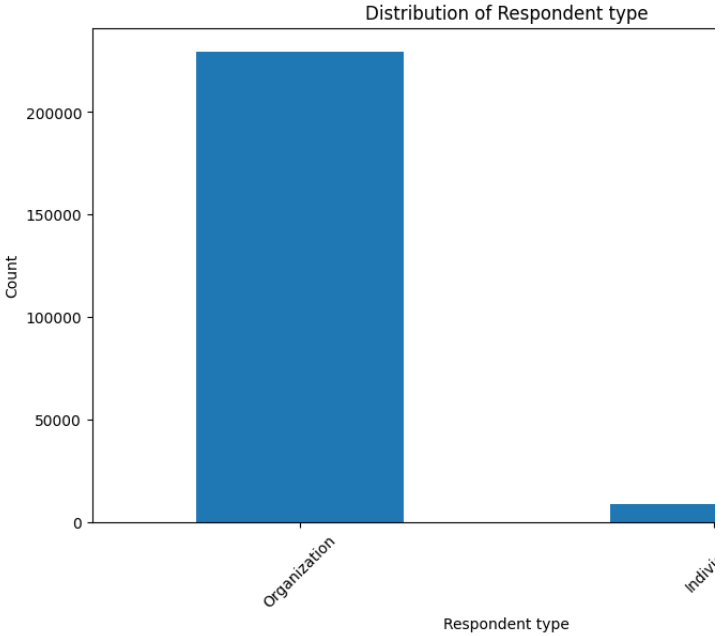

Distribution of Coverage level
Count
Coverage level
Group A&H


Distribution of Coverage type
Count
Coverage type
Accident and Health
Automobile
Homeowners
Life & Annuity
Miscellaneous
Liability

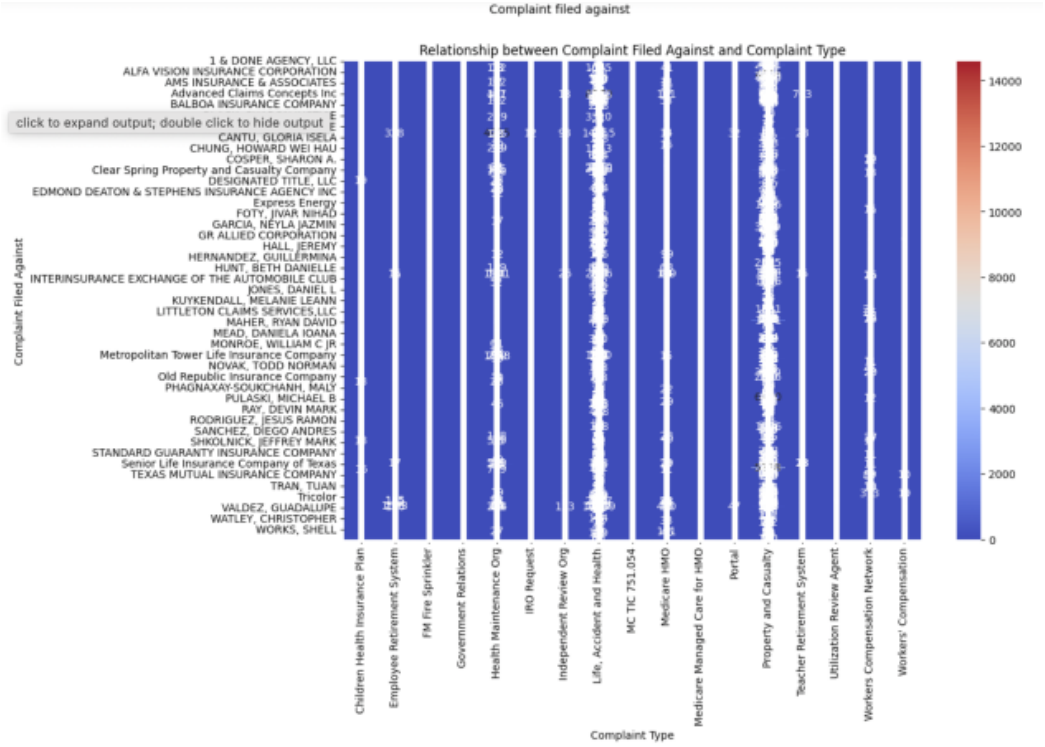
EDA

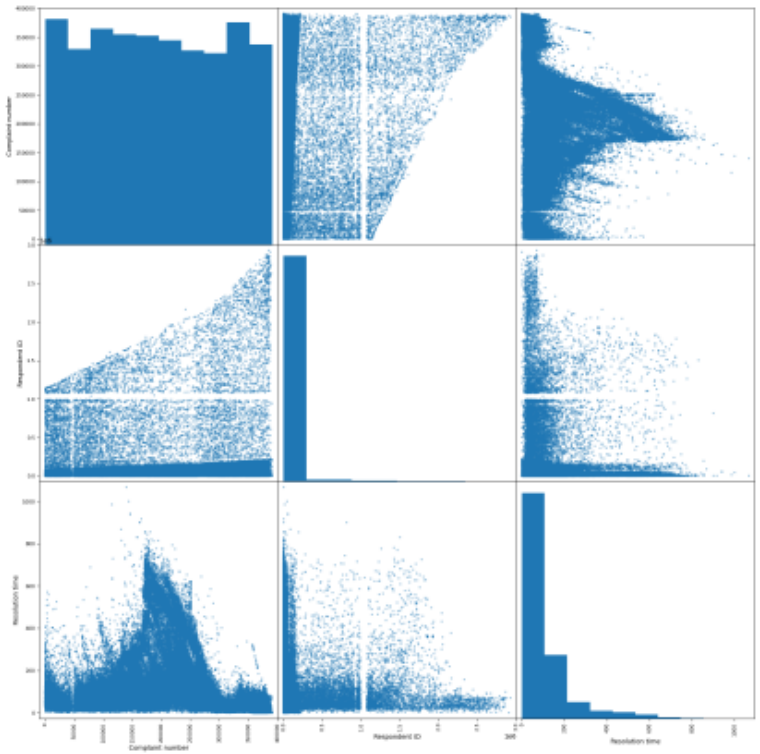

Distribution of Complainant type
INDV
ORG
Complainant type


Distribution of Respondent Role
Respondent Role


Distribution of Respondent type
Count
200000
150000
100000
50000
0
Organization
Respondent type

EDA





FEATURE ENGINEERING

PREDICTIVE MODELLING

- Random Forest : I have performed random forest to find the companies that are more likely to be complained against. Whose performance metrics are as given below

```
Accuracy: 65.14999999999999 %
Precision: 58.617715919854554 %
F1 Score: 60.5210195565312 %
```

We have also performed other models namely Neural Networks, Naïve Bayes, Decision trees, SVM and Xgboost. However, these models didn't show us any better performance then Random Forest model.

Link of code: https://colab.research.google.com/drive/15EbeC4fdMW_WExl-xJAszjJre0q5qaJz

THANK YOU