Final Project Paper

Insurance Complaint Analysis

Index

[Introduction 3](#_Toc165733623)

[Data Description 5](#_Toc165733624)

[Research Questions 5](#_Toc165733625)

[Exploratory Data Analysis (EDA) Questions 6](#_Toc165733626)

[Industry / Scholarly Review 6](#_Toc165733627)

[Historical Development of the Insurance Industry 6](#_Toc165733628)

[Theoretical Frameworks and Concepts in Insurance 8](#_Toc165733629)

[Types of Insurance 9](#_Toc165733630)

[Regulatory Environment and Market Dynamics: 11](#_Toc165733631)

[Challenges and Opportunities: 12](#_Toc165733632)

[Conceptual Framework: 14](#_Toc165733633)

[Data Description 14](#_Toc165733634)

[Software: 14](#_Toc165733635)

[Data Collection 15](#_Toc165733636)

[Data Dictionary 16](#_Toc165733637)

[Potential Issues 17](#_Toc165733638)

[Data Preparation Methodology 18](#_Toc165733639)

[Missing Data Analysis and Handling 19](#_Toc165733640)

[Data Preprocessing and Transformation 22](#_Toc165733641)

[Summary of Data Preprocessing 25](#_Toc165733642)

[Data Cleaning and Standardization 29](#_Toc165733643)

[Categorical Encoding 30](#_Toc165733644)

[Exploratory Data Analysis (EDA) and Modelling Methodology 37](#_Toc165733645)

[Feature Engineering 37](#_Toc165733646)

[Train-Test Split: 38](#_Toc165733647)

[Feature Scaling 38](#_Toc165733648)

[Model Initialization 39](#_Toc165733649)

[Model Evaluation 40](#_Toc165733650)

[Test Assumptions 45](#_Toc165733651)

[Findings 47](#_Toc165733652)

[Distribution of complaint types 47](#_Toc165733653)

[Average time taken to resolve a complaint 47](#_Toc165733654)

[Proportion of Complaints Resolved compared to those that remain unresolved: 48](#_Toc165733655)

[Factors Influencing Complaint Resolution 49](#_Toc165733656)

[Evolution of Insurance Complaints Over Time 51](#_Toc165733657)

[Factors Contributing to Effective Complaint Handling Procedures 51](#_Toc165733658)

[Conclusion 58](#_Toc165733659)

[Limitations 60](#_Toc165733660)

[Future Research Directions 61](#_Toc165733661)

[Recommendations 61](#_Toc165733662)

[Appendix 62](#_Toc165733663)

[Summary Statistics of Numerical Variables 62](#_Toc165733664)

[References 65](#_Toc165733665)

## Introduction

In the game of risk and security insurance plays a vital role. Understanding customer complaints and resolving them remains the core. Successfully profiting from innovation is a complex, company-wide endeavor, and most insurers have not yet cracked this code—at least not on a consistent basis. In fact, a 2017 survey of life and annuities executives found that only 12 percent believe they have a process that delivers strong product innovation (*Five Steps to Improve Insurance Innovation | McKinsey*, n.d.). Insurance provides financial safety and peace of mind. Whereas it depends upon the individual which will be according to their financial goals and financial stability. Insurance service is mainly a customer centric industry. It must ensure optimal customer satisfaction.

Insurance acts as a bedrock in hazard control by providing financial security. We all buy insurance for our own safety for unforeseen situations. Insurance companies drive individuals and organizations toward financial security. It is a customer focused industry. Meeting customer’s financial and customer satisfaction is the main motive of an insurance company. Merriam-Webster defines insurance as a form of “coverage by contract whereby one party undertakes to indemnify or guarantee another against loss by a specified contingency or peril.” (*Definition of INSURANCE*, 2024)

Texas Department of Insurance (TDI) acts as the medium between individuals and licensed entities which consists of companies, service providers, agents, and adjusters. With a 10-year average return on net worth of 10 percent for workers’ compensation, Texas outperforms the national average of 7.4 percent, while the average premium per $100 of payroll has decreased 73 percent since 2003.This mainly aims to maintain the service standards and maintain a competitively fair insurance market. 80 percent of injured employees report that the medical care for their work-related injury was as good or better than their routine medical care. Here a complaint can be raised by the consumer with TDI whereas the department investigates and ensures fair treatment. Upon gathering the data, our research uncovers the evolving trends and patterns that impact the time and reason. The aim is to improve complaint handling procedures, strategies and to provide an effective resolution system within the industry. (*Workers’ Compensation*, n.d.)

The research mainly focuses on the complaint resolution and strengthening the customer centric procedures among the industry. By analyzing different features, we are investigating different ways to improve the effectiveness of handling procedures. Factors like complaint type, time to resolve, coverage type and others will impact the handling systems. Improving the customer experience and satisfaction will be the core of the research.

To draw the insights from the dataset we use different analysis techniques. Text mining coupled with Natural Language Processing (NLP) can offer motives behind the research. NLP will uncover the quality aspects of the data. In addition, text recognition and complaint classification will help us in drawing potential models. To maintain the data quality, handling missing values and inconsistency of the data will be the part of the analysis. For ethical consideration unbiased analysis should be prioritized.

### Data Description

* This dataset was provided by Data.gov.
* data.austintexas.gov is the publisher and it is maintained by the TDI ODP Administrator.
* The dataset is less than a year old which was created on August 25, 2023, and was the latest updated on February 25, 2024.
* For analysis we got 246,243 instances and 17 features.

Dataset Link: <https://catalog.data.gov/dataset/insurance-complaints-all-data>

### Research Questions

1. What are the primary factors influencing complaint resolution? How do variables like complaint type, coverage type, and respondent role impact resolution outcomes?
2. How have insurance complaints evolved over time, considering seasonal trends, variations in complaint frequency, and changes in complaint types?
3. What are the key factors contributing to effective complaint handling procedures, including response time variations, resolution rates, and best practices?

### Exploratory Data Analysis (EDA) Questions

1. How do reasons for filing complaints differ between confirmed and unconfirmed complaints?
2. What is the average time taken to resolve a complaint?
3. What proportion of complaints are resolved compared to those that remain unresolved?

Scope

By examining the complexities of the complaints, we gain insights that will be useful for providing better resolution. Seasonal trends can be used to look at the complaints patterns and to improve the handling procedures. The insights gained can help both parties for effective resolution. The study can be broadly applied for understanding the resolutions and handling procedures. This will ultimately impact the overall customer experience.

## Industry / Scholarly Review

### Historical Development of the Insurance Industry

In the process of minimizing the risk we humans have undergone many processes. In this process people started trusting entities that can handle the risk for them. It is more or likely can be mentioned as burden sharing and collective responsibility. It was first started by communities pooling their resources to overcome unforeseen calamities which can affect their crop or cargos. Thus, the concept of insurance has emerged. The concept of insurance dates to around 1750 B.C. with the Code of Hammurabi, which Babylonians carved into a stone monument and several clay tablets (*The History of Insurance*, n.d.). The simple explanation of insurance according to old age is “a sailor needs money to repair his ship or buy supplies, so he lends money from a lender, but the lender gets a stake instead of money in return. The catch is the lender only gets paid back only if the voyage is successful. If the ship sinks then the sailor owes nothing,” this way the sailor feels secured financially.

Lloyd's of London or Lloyd’s which has been an insurance and reinsurance marketplace since the late 17th century. It operates as a syndicate which consists of five key players which are syndicates, insurance buyers, brokers, managing agents, and cover holders. The main purpose of Lloyd’s is to facilitate transactions between insurance buyers and sellers (*Lloyd’s of London*, n.d.).

American insurance evolves from a rich history. Philadelphia Contribution ship which was established by Benjamin Franklin back in 1752. This mainly focused on fire insurance. Over the period many forms of insurance have emerged. The McCarran-Ferguson Act of 1945 made major corrections by shifting the regulatory framework and placing them under state law but not federal due shady practices. In 1901, the Fire Insurance Society of Philadelphia was founded. The organization grew to include a broader scope of the insurance trade when it was incorporated in 1909. In 1925, it was given a new name: the Insurance Society of Philadelphia (*History - Insurance Society of Philadelphia*, n.d.).

The Social Security Act 1935 has changed the game. This is a kind of safety net for employees after retirement. This attracted many to work. This helped during World War-II wage freezing period by providing health and life insurances. In 1944, the Supreme Court ruled that the insurance industry should be federally regulated (*U.S. Reports*, n.d.). With other financial giants and insurance companies there were continuous developments in the industry. The internet is playing a major role in the insurance industry. Moreover, it has become a global marketplace.

### Theoretical Frameworks and Concepts in Insurance

Theoretical Frameworks and Concepts in Insurance are important because they build the structure of tariffs that must be distributed fairly among the policy holders. There are several stages that a policy undergoes when structuring the premium. These are the few and key theoretical frameworks and concepts in insurance that provide foundation.

Risk Pooling: The basic idea of insurance is described as risk pooling. It is a fundamental concept in Insurance. This is the only way of risk management which needs social interaction. By undertaking large groups, it stabilizes the financial outcomes and can predict and structure affordable premiums.

Actuarial Science: Analyzing the risk and the instability by using math and statistics is the main function. Actuaries will assess the factors like mortality rate, market trends and so on. This is mainly to ensure fair pricing. It also maintains stability among the insurance companies to payout the claims.

Principle of Indemnity: This mainly focuses on reimbursement but not on gaining profit on the loss itself. It requires documented financial loss proof to compensate. This discourages the policyholders from intentionally crashing or any other intentional events and to reduce their premium.

Regulatory Frameworks: To ensure customer claims and payouts fairly, regulatory frameworks help a lot. Even to keep insurance companies stable and reserve enough money for future claims. To keep the market functioning smoothly among the insurance companies and ensure fairness to customers, these regulations act as guidelines.

### Types of Insurance

Insurance provides a wide range of products that helps us, and our loved ones feel protected from unforeseen situations. Life is full of surprises. While some are exhilarating, others are devastating emotionally and financially, like a car accident or a kitchen fire. That’s why there are many types of insurance to help after unexpected disasters (*8 Different Types of Insurance Policies And Coverage You Need – Forbes Advisor*, n.d.). The major types of insurance are life insurance, health insurance and auto insurance.

Life Insurance: To keep the dependents protected ​​life insurance is extremely important for a family. Whole life insurance and term life insurance are the two types of life insurance. In the whole life insurance, the dependents get the claim after the death of the policyholder whereas in term life insurance a certain period is set to meet the financial goal. Whole life insurance is pricier and term life insurance is affordable.

Health Insurance: This will help in covering hospital bills in case of accidents or illness. There are many ways to get health insurance. Employee sponsored is the most common type of health insurance. Some people buy from a private insurance company. For families with tight budgets there is a federal marketplace which comes under the Affordable Care Act. The Patient Protection and Affordable Care Act, referred to as the Affordable Care Act or “ACA” for short, is the comprehensive health care reform law enacted in March 2010(Affairs (ASPA), 2013). Over 90% of Americans have health insurance.

Auto Insurance: The rate of crash deaths per 100 million miles traveled increased by 2% from 1.34 in 2020 to 1.37 in 2021(*Highway Statistics 2021 - Policy | Federal Highway Administration*, n.d.). Auto insurance is mandatory in all the states. This insurance covers the automobile financially from collision, theft, fire, etc... This even covers the medical expenses of the driver as well as the passengers up to a set amount. Many factors include the premium like age, driving history and location.

There are other types of insurance which evolved over the period that includes property insurance, pet insurance, business insurance, travel insurance and so on.

### Regulatory Environment and Market Dynamics:

NAIC Model Laws and Insurance Complaint Analysis: The National Association of Insurance Commissioners (NAIC) has a major role in building a uniform and consistent law development process within the state-based insurance regulatory framework in the United States. This article particularly studies the model law process and possible effects of insurance complaints analysis practices (*NAIC Model Laws*, n.d.). It highlights the benefits of the uniform insurance regulations according to the consumers protection perspective that ensures opportunity for complaint resolution processes with consistent standards. It follows a two-step test that promotes the development of a model law or regulation where the first step involves subject matter of the model law or regulation that must call for a minimum national standard or require uniformity among the states. In the second step, NAIC must dedicate significant state insurance regulator and NAIC staff resources to educating, communicating, and supporting the adoption of the model law or regulation. If these criteria are not met, NAIC may develop instructions which provide recommendations of best practices. The article briefs about steps involved in drafting and adopting a model law, including approval processes, public input opportunities and voting requirements. After adopting, the NAIC priority is to encourage state legislatures to adopt the model law with essential changes. The potential impact of the NAIC law development process on insurance complaint analysis can contribute to a more standardized and productive approach to handle customer concerns within the insurance industry.

Adapting the investment approaches, costing models, and risk management frameworks makes the modern shift easier (*Discovering Recent Risk Trends and Management to Impact the Insurance Industry*, n.d.). The significance of risk management in current insurance trends shows how macroeconomic factors like interest rates, inflation, and potential recessions can impact insurance firms.

The advancements in the insurance security gap, includes aspects such as climate crisis, natural disasters, and cybersecurity. Modern business models like big data analytics, associating with InsurTech companies, and cloud computing adoption help design new products and services that enhance the edge of advancements such as improved risk evaluation, customer-oriented solutions, and operational performance (Shalabi, n.d.). Thus, the article describes the dynamic nature of the insurance sector and the need for insurance firms to accept change and risk management strategies to succeed in the emerging world.

### Challenges and Opportunities:

The impacting challenges of global insurance industry that lead to change are:

Pressures to reduce the cost – Factors like increased competition and customer awareness pressurize to reduce IT, Sales, and administrative costs of the insurance sector¹².

Less Investment Returns – Insurers are considering alternative investments with more returns than the low government and corporate bonds.

Development in Regulations – Solvency II and IFRS are the regulations that make the insurance industry understand the risk and capital around the businesses better (*(1) New Messages!*, n.d.).

Population aging – Scope for new types of insurance related products and services like flexible annuity products increased due to the demographic shift.

Growth in Medical Sciences – For the healthcare insurers who need to adapt to changing lifespans and initial healthcare improvements raised ethical and underwriting issues.

Brian Heale explains the importance of capital management for insurers to invest in capital planning tools and techniques to analyze risk and capital allocation efficiently. Data management being a key challenge in regulatory reporting requirements and effective decision making which deals with high quality analytical data. Establishing a centralized analytical repository to organize and store data for different business needs involves complex steps and considerations in its maintenance and building process. Updating the outdated systems may be expensive and become a blockage for the scalability and competition but this investment is crucial for future success. Managing to handle these challenges allows the industry to have great existence in the widening environment.

Alexandra Samet says that in 2022 due to inflation and market volatility the insurers are struggling. Challenging factors such as rising costs of the claims, competition, and difficulty in funding for insurtechs are vital in causing decline of the insurance industry expansion. Concentrating on high-ROI customer areas by strengthening the marketing budgets of insurance providers impacts the results (Picoult, 2021). Traditional insurers may obtain insurtechs or direct investment in sales channels to survive in the evolution. Long-term trends like embedded insurance advancement, increase in demand for usage-based insurance, and emphasis on trust of brand for insurtechs. To maintain and work through the future transitions in the insurance sector, insurers adaptability to capitalize on competitors spending changes, differentiating the usage-based products, digitalization and consumer satisfaction exposes the capability to overcome the predominant issues in insurance (*Enhancing Customer Experience in the Insurance Industry With AI*, n.d.).

### Conceptual Framework:

We have understood how the insurance works and the factors affecting the insurance claims. By using the complaint data and overview of insurance will help us to provide customer satisfied resolutions. By analyzing the data of complaints, resolution times, and other the success rate of resolving complaints can be increased. We can provide best handling processes that drive towards customer satisfaction. By doing these, insurers can build trust, increase customer satisfaction, and can operate efficiently.

## Data Description

### Software:

We have chosen Python (3.x.x) for our analysis, visualization and for data manipulation. The reason to choose Python is its rich ecosystem. First for its clear and reliable syntax. It has some great open-source libraries. It is really flexible for analysis and powerful as well. It offers great functionality and is accustomed to the research purpose.

We have employed different libraries which are as follows:

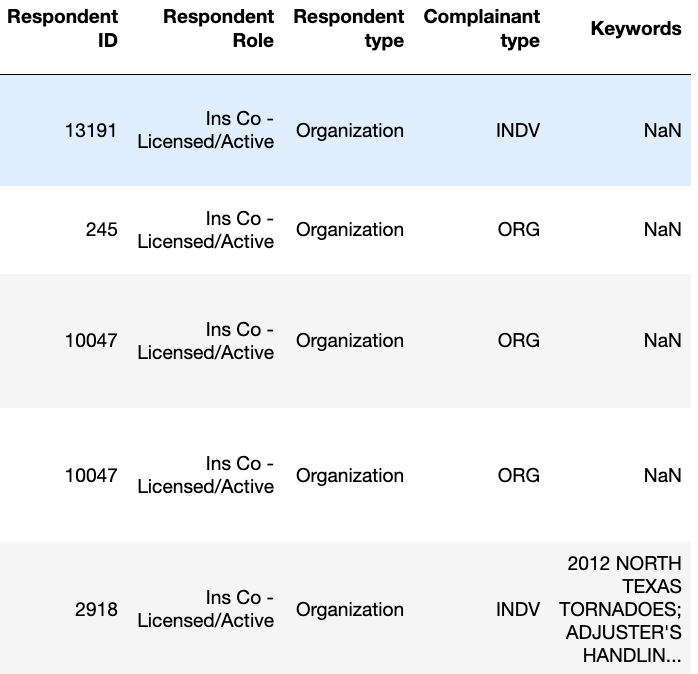
* Pandas for data manipulation
* NumPy for numerical operations
* Matplotlib and Seaborn for plotting and visualization.
* Scikit-learn is employed for machine learning.
* Word Cloud for visualizing text data

### Data Collection

Our dataset contains information about insurance complaints. This dataset was originally obtained from data.gov. data.austintexas.gov published this data. TDI ODP Administrator maintains this data. It contains 246,243 records with 17 features and was created on August 25, 2023, and the version we analyzed was last updated on February 25, 2024.

Dataset Link: <https://catalog.data.gov/dataset/insurance-complaints-all-data>

This dataset contains a variety of features with an ample number of records. All the features in the dataset covers the essential areas of our research. It has categorical values and unique identifiers which will be helpful for extensive research. Complaint number, Complaint filed against, Complaint filed by Reason complaint filed, confirmed complaint, how resolved, Received date, Closed date, Complaint type, Coverage type, Coverage level, Others involved, Respondent ID, Respondent Role, Respondent type, Complainant type and Keywords are the features of the dataset. Let's have a sample look of the dataset and how it looks as shown in the below figures.



### Data Dictionary

The dataset is composed of various attributes which are of different data types and variable types. Every complaint has a unique identifier with Complaint number which is an integer. Different entities like Complaint filed against, Complaint filed by Reason complaint filed, confirmed complaint, Complaint type, Coverage type, Coverage level, Others involved, Respondent Role, Respondent type, Complainant type and Keywords are captured and involved to resolve the customer complaints which are strings, categorical(nominal). These attributes give us the information about who, why and what of the complaints along with the supporting entities. Received date and Closed date will help us find the seasonal trends and patterns. The difference between them will give us the resolution time. We also get the efficiency of the handling procedure by analyzing the duration. How resolved will help us to find out the methods used to resolve the complaint, which is a string, categorical. Respondent ID which is also a unique identifier that can keep the information of a respondent safe and it is an integer.

### Potential Issues

* Handling the missing values can be challenging. Missing data in certain columns might affect the overall analysis. One of the important column “Keywords” has many missing values, that’s is about 19% of the keyword's column has missing values, for this issue are planning to impute the missing values with mode and keywords column have multiple keywords In row, we planning to divide the keyword column into multiple columns.
* There is inconsistency in data, for example the date formats ("DD/MM/YYYY'' vs "YYYY-MM-DD') which can hamper analysis and can lead to false interpretation. The data complexity is a potential issue were finding out the relationship between the columns can be difficult. And we can also predict the time to close a ticket but for this we must deal with date and time column, for this issue we are working on the datetime package of python.
* Data volume should be considered as this is a large amount of data. It can be challenging to handle the data using appropriate analysis tools and techniques that will draw useful insights. By data cleaning and data manipulation we can address the mentioned issues. Formatting, standardization, and error correction will be crucial for reliable analysis.
* ⁠⁠We are planning to cluster the complaint and we have 17 columns; we check which columns are useful.

## Data Preparation Methodology

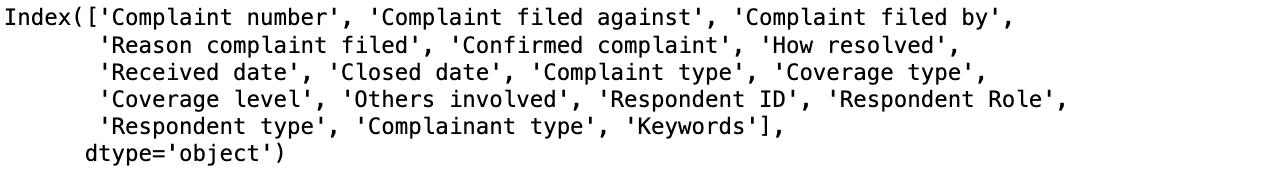
At first, we will focus on preparing data for our analysis and making data easy to access.

We start with collecting our libraries by importing.

* import pandas as pd.
* import seaborn as sns.
* import numpy as np.
* import matplotlib. pyplot as plt.
* from wordcloud import WordCloud

These will allow us to refine the data for the analysis and create great understandable visualizations along with the calculations. We will now load the data “Insurance\_complaints\_\_All\_data.csv” in a data frame which is like a digital spreadsheet. Now for a quick inspection we will have a preview of the data’s organization and we will come to know the number of features and records as well (248527, 17).

We now check the columns for verification. The below figure shows the columns that are loaded.



The above picture shows all columns from the dataset are loaded.

Project Architecture:

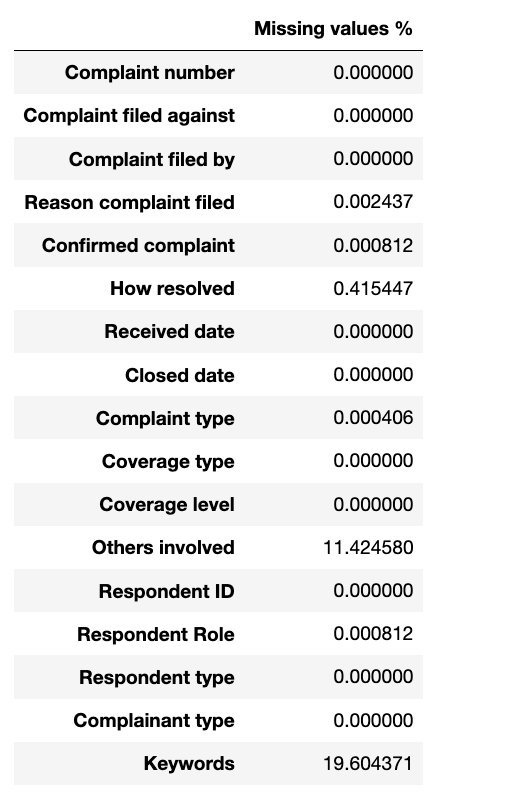
A diagram of a software development process

Description automatically generated

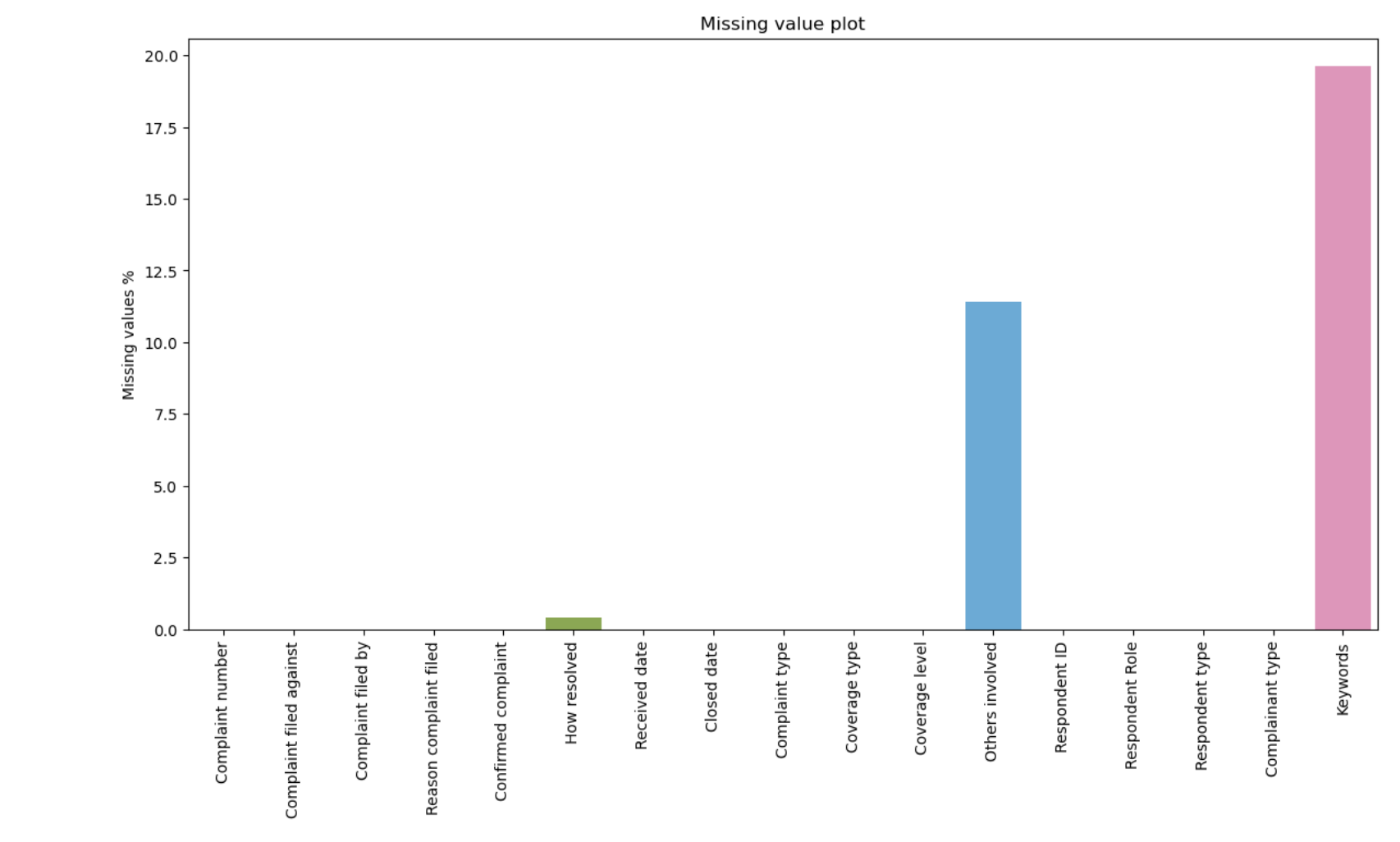
### Missing Data Analysis and Handling

#### Missing Value Identification

We will now check the missing values. Every data which is large will have some missing cells. Let's find out by going through each column. By performing a little math, we will get the missing values percentage. By pulling out this we can have a clear idea of which column has the highest missing values which will help in aiding analysis.



The figure gives us a summary of the missing values. Now we have a clear path for our analysis. This important part where we can know the columns with missing values which will further help us in making certain decisions. We might drop the missing data or find a way to fill the missing values by handling techniques. This is a key part in the analysis.



The above figure bar plots to visualize the percentage of the missing values in each column within the Data Frame. This bar chart was created using Python libraries matplotlib and seaborn.

X-axis shows the column names.

Y-axis shows the percentage of missing values.

This chart helps to know columns with a high percentage of missing values, which might need attention during data cleaning and analysis. For example, the columns with over 20% or 30% missing values might require specific handling methods, like imputation or removal.

#### Handling Missing Values

Before pitching into the process, we will investigate the other columns which are affected. We will make a list called missing columns to keep track. Now we will store the calculated percentage of missing values for each column in our Data Frame. By this we will have a list containing the names of the columns in our data frame that have some missing data. Our aim is to identify which sections in the data have missing values. Now we will categorize columns by data type to plan for handling missing values and we can even understand the nature of the data in each column.

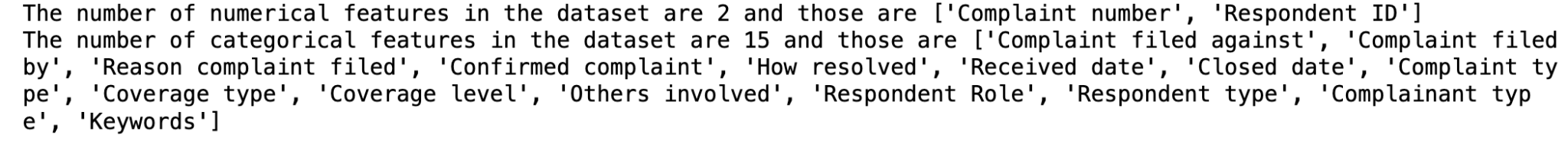
Now we will make solutions based on data type like filling text-based columns with mode and numerical columns with median. In text columns we will fill the missing value with the most frequent word to maintain neutrality. For consistency we will fill the numerical column with a median value. By going symmetrically through each column, we will apply needful techniques. We also ensure the quality for robust analysis.

### Data Preprocessing and Transformation

#### Classification and Summary

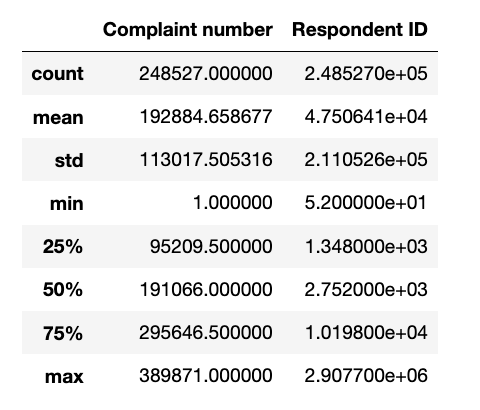
By classifying the data into two types numerical and categorical. By inspecting the data type of each column. If it's not 'O' that means the column contains numerical data, it will be added to the numerical features list. Now we will look for the columns with an 'O' data type, which means it is categorical data and will be added to the categorical features list. Identifying them now will help us in future analysis processes like numerical features will be analyzed using statistical calculations and categorical features will be analyzed by machine learning models.

Will identify and separate columns in the data frame whether they contain numerical or categorical data. Now we will calculate the total number of numerical features and categorical features. Now by formatting clear and descriptive output statements to summarize the type and count of features, we will print the data now as we use this in future. This will give an overview of the dataset’s state as of now, which helps in planning data analysis and modeling.



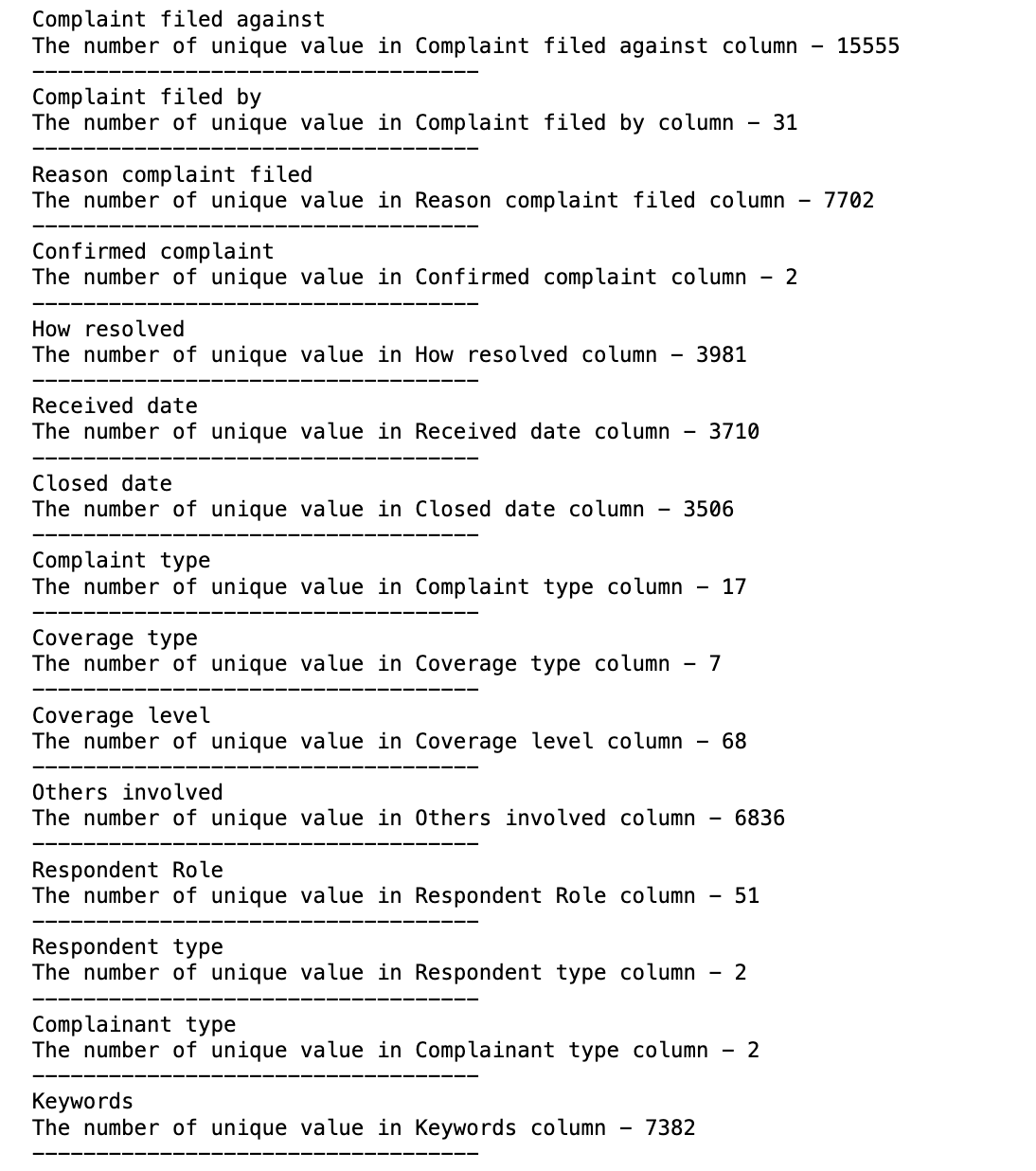
From the above figure we can see that we have clearly distinguished numerical features (usually IDs or codes) and categorical features (descriptive, often text-based data). Now we have the count of numerical (2) and categorical (15) features in the dataset. Like we have examples of numerical ("Complaint number", "Respondent ID") and categorical features ("Complaint filed against '', "Reason complaint filed"). The kind of information of different feature types of numerical features act as identifiers, while categorical features tell us about the context and characteristics of the complaints.

Categorical Feature Analysis: Summary statistics of the numerical columns in the data frame as shown in the figure below.

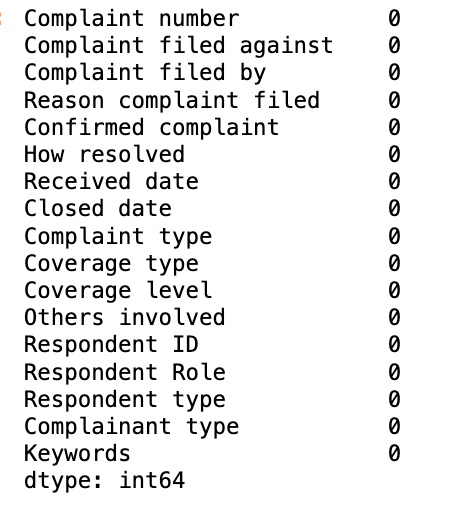


We will iterate through each categorical feature in the dataset. Now it will print the name of the current categorical feature and will be analyzed. calculates the number of unique values present in the column. By using the len () function on the array of unique values that we got from the column we will get the count of unique values.

By studying the below figure, we can understand complaints against (15,555), diverse reasons for complaints (7,702), resolution methods (3,981). Complaints will fall into 17 categories. The number of unique values in "Received date" and "Close date'' columns gives closure dates over time. By this we will have a dataset's depth in capturing different complaint scenarios.

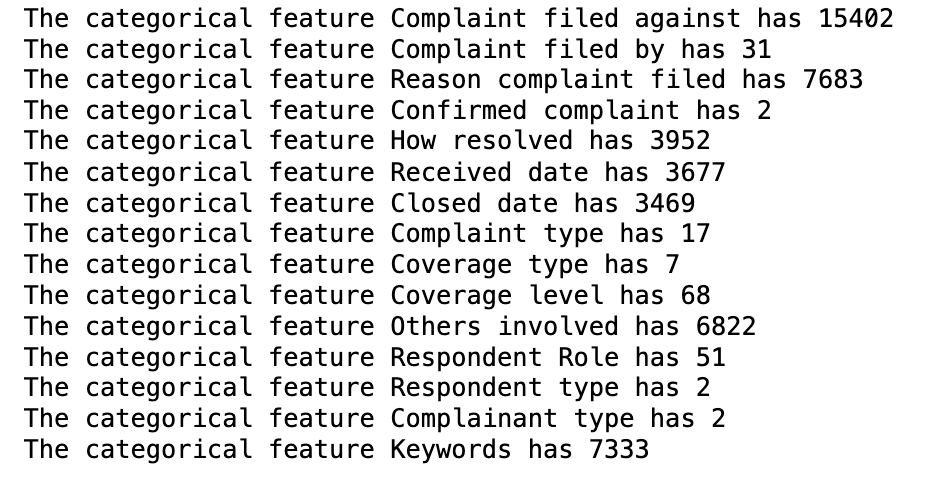


We will calculate the number of missing values present in each column of our data frame.



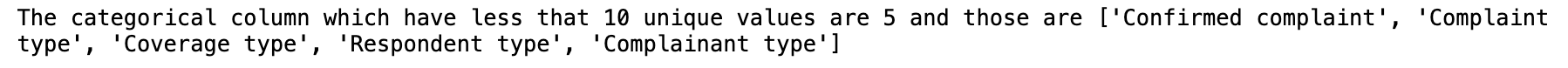
The above result of checking for missing values in each column of the dataset. We see the number of missing values on the right side of the table which is ‘0’. The dataset is now complete with no null or missing values. Now we can proceed with our analysis part. By processing the data, we have 246241 rows and 17 columns.

As we created an empty list called categorical columns. To store the names of categorical columns that meet necessary conditions. The number of unique values in that feature is less than 20. We have identified categorical columns with a number of unique values, which can be easier to analyze.



The above figure has unique values that are in each categorical feature. "Complaint filed against" and "Reason complaint filed" have a lot of unique values. "Confirmed complaint" and "Respondent type" have only 2 unique values.

We have got a good set of 5 columns with less than 10 unique values each. These columns are much more useful and can be used further. The five are 'Confirmed complaint', 'Complaint type', 'Coverage type', 'Respondent type', and 'Complainant type'. With these columns we can continue our analysis. The below figure is the representation of the columns.

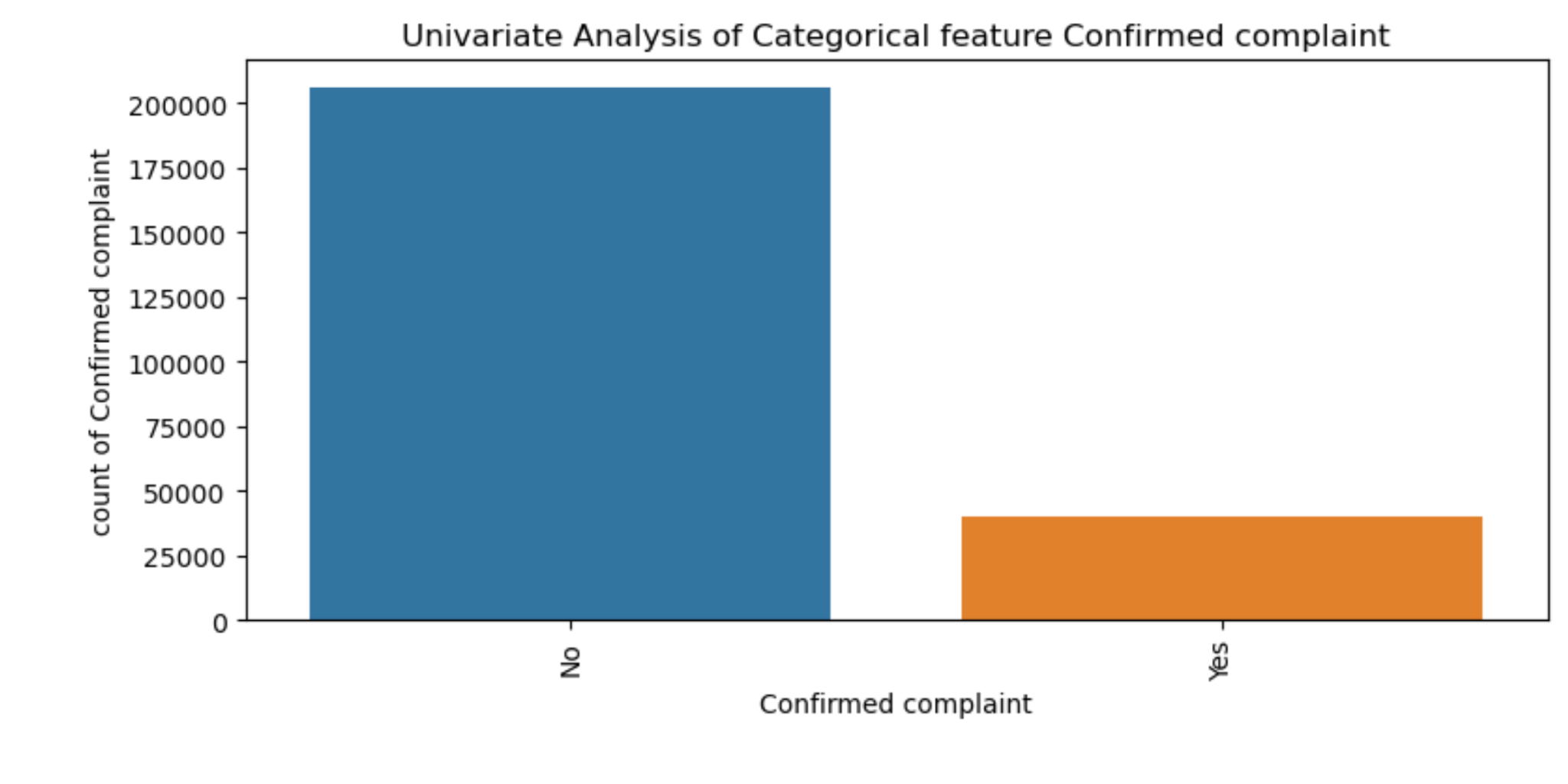


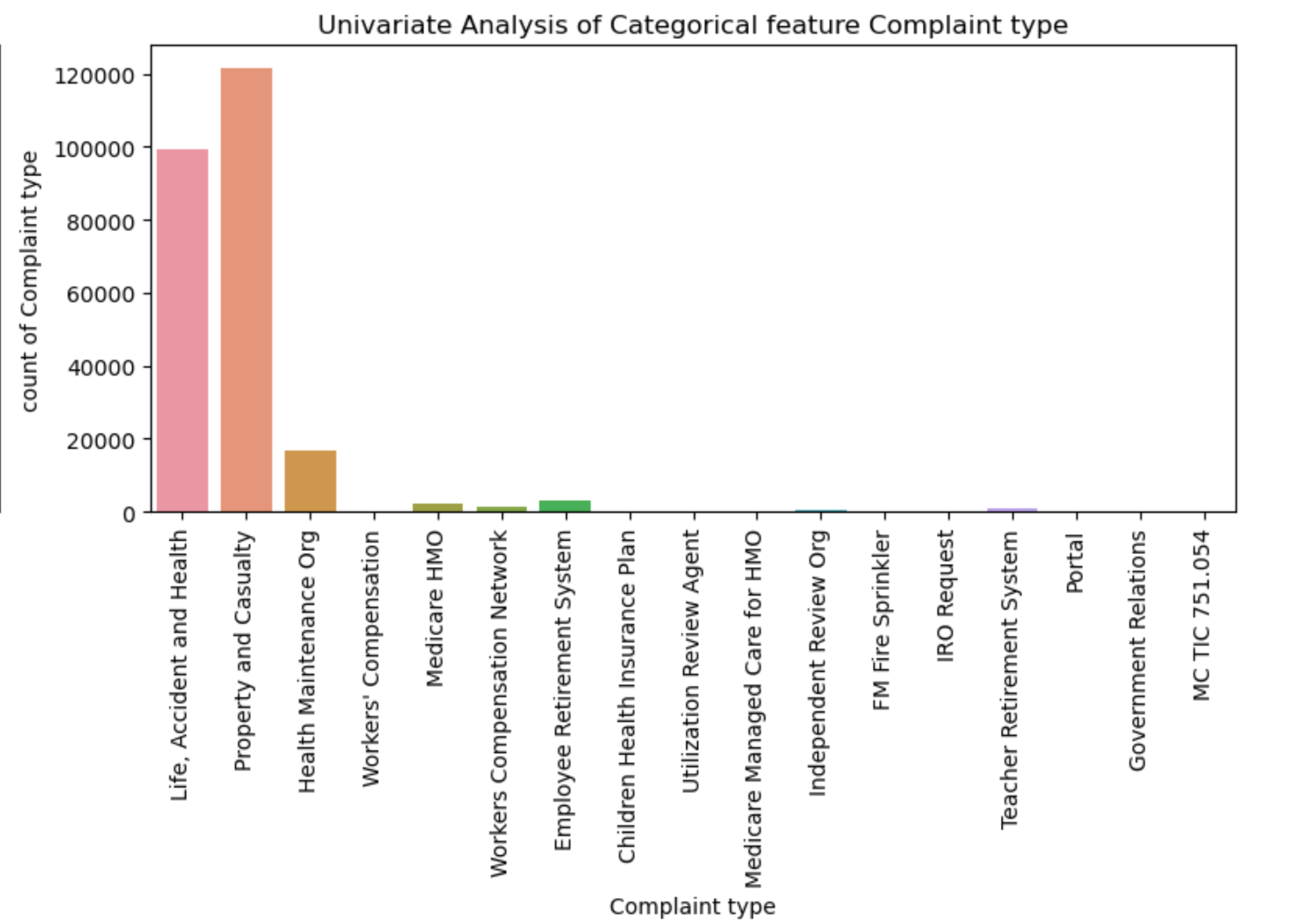
### Summary of Data Preprocessing

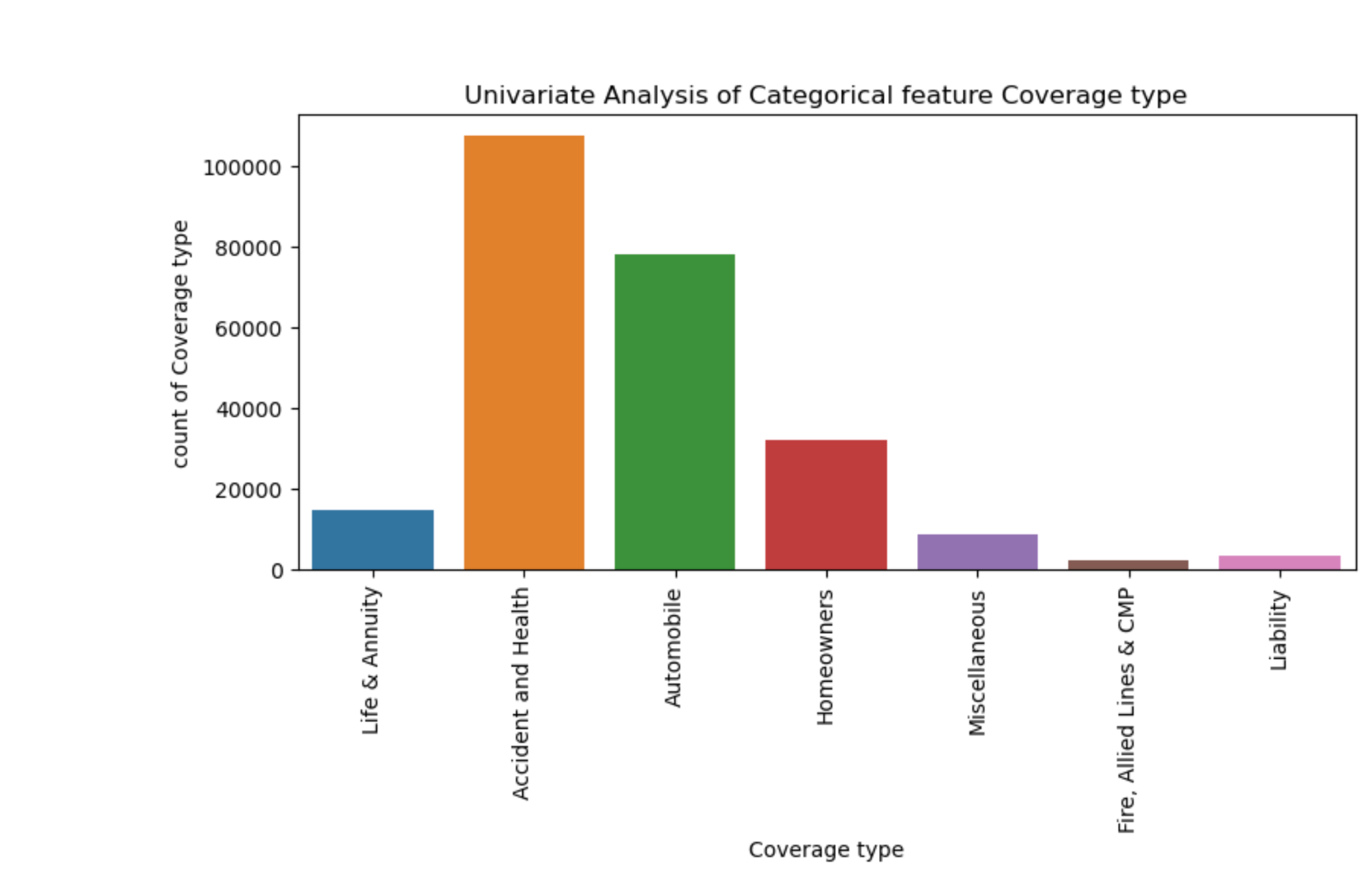
What is Univariate Analysis and how it works?

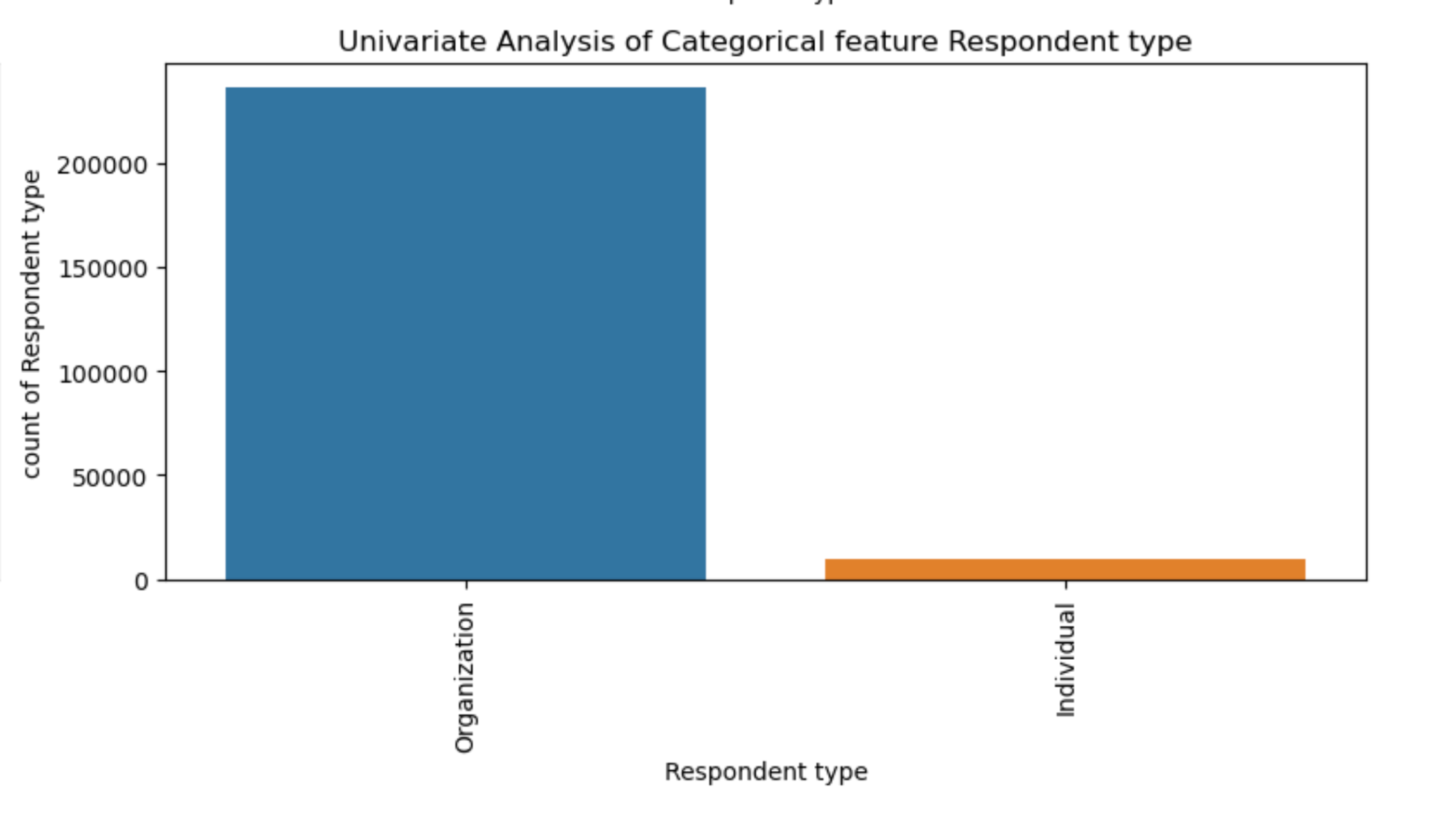
Univariate evaluation of the dataset containing records on complaints, events worried, resolution techniques, and timing involves inspecting text-based totally definitely attributes like "Complaint filed in opposition to/with the aid of" and "Reason complaint filed" thru phrase frequency evaluation and phrase cloud visualization. Categorical attributes which consist of "Complaint type" and "Coverage type" are analysed for frequency distribution using bar charts, while temporal attributes like "Received date" and "Closed date" go through time collection analysis thru histograms or line charts to locate traits over the years. This complete assessment affords insights into the distribution, styles, and traits of the facts, serving as a vital foundation for further exploratory and multivariate analyses.

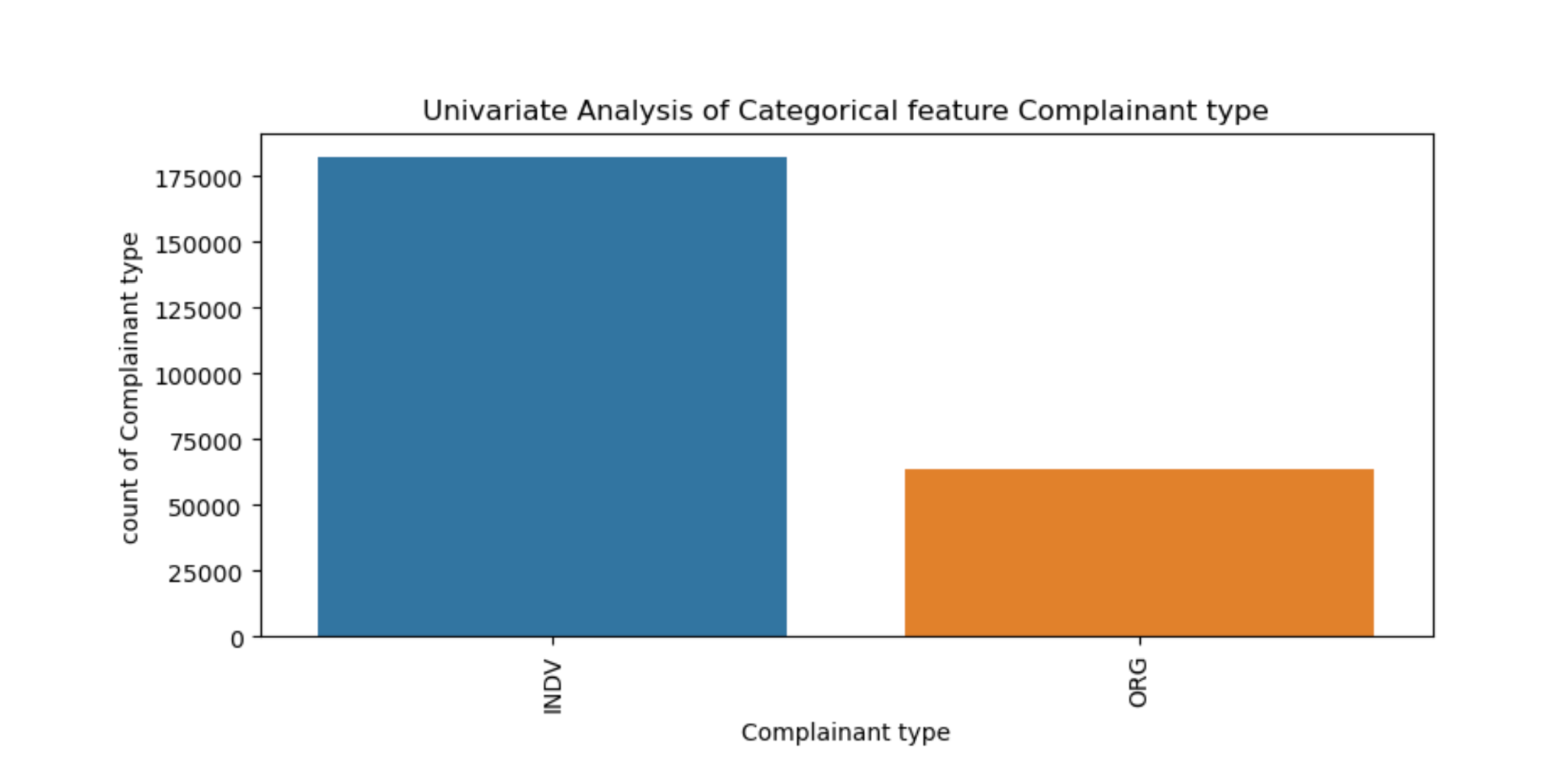
We're using count plots for analyzing each categorical feature. We will set up the input code to get a large figure with many subplots. We will create a countplot within each subplot. We are checking how many times different categories will be seen. Our count plots show that there are some features that have many unique categories (like "Complaint Filed Against"). This will make them a bit hard to analyze. Others show a few categories. We look for categories that are common or rare, which can be important patterns in the data. Features with few categories are easy and quickly shows us common types.











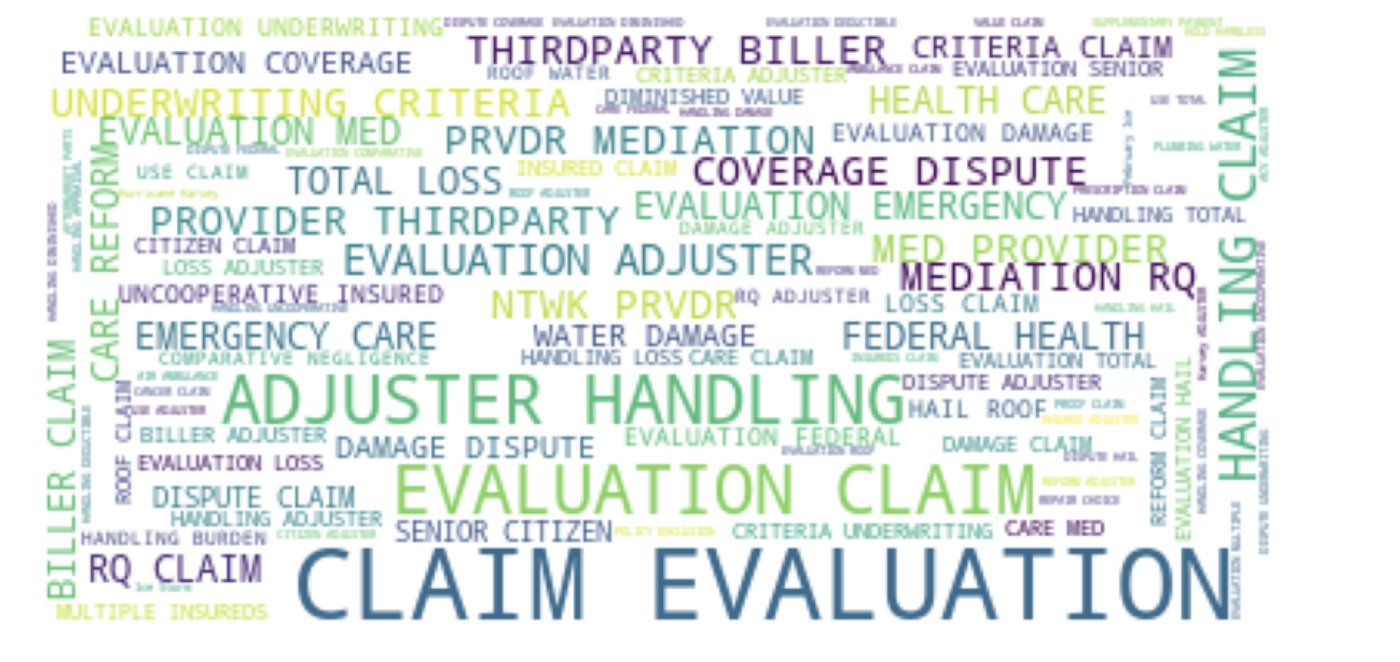
The figures show the count plots which help us to see how categories are distributed in.

our features.

### Data Cleaning and Standardization

Now we start by preparing text data. The 'Keywords' column is made into strings and then we will join into one long string called 'text'. This gets all the keywords in a format ready for the word cloud. We will create a WordCloud object with a white background. The text will be used to generate the word cloud. This makes the picture where the size of words shows how many times they appear.

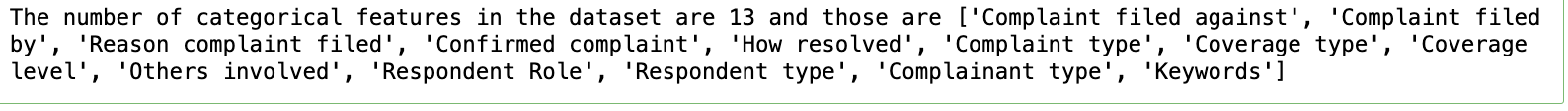
The below figure shows us which keywords are most common in our data. Larger words mean they appear more, which gives us understanding of the important topics.



We will convert the Received date and Closed date columns in the dataset to datetime format using the pd.to\_datetime () function. ‘Time\_taken\_to\_close’ will be calculated by checking the difference between Received date and Closed date which will provide resolution time for each complaint. By iteration year, month, day, and quarter are extracted for further easy analysis and were given necessary suffixes. The original columns were dropped. Now we have 246,241 rows and 22 columns.

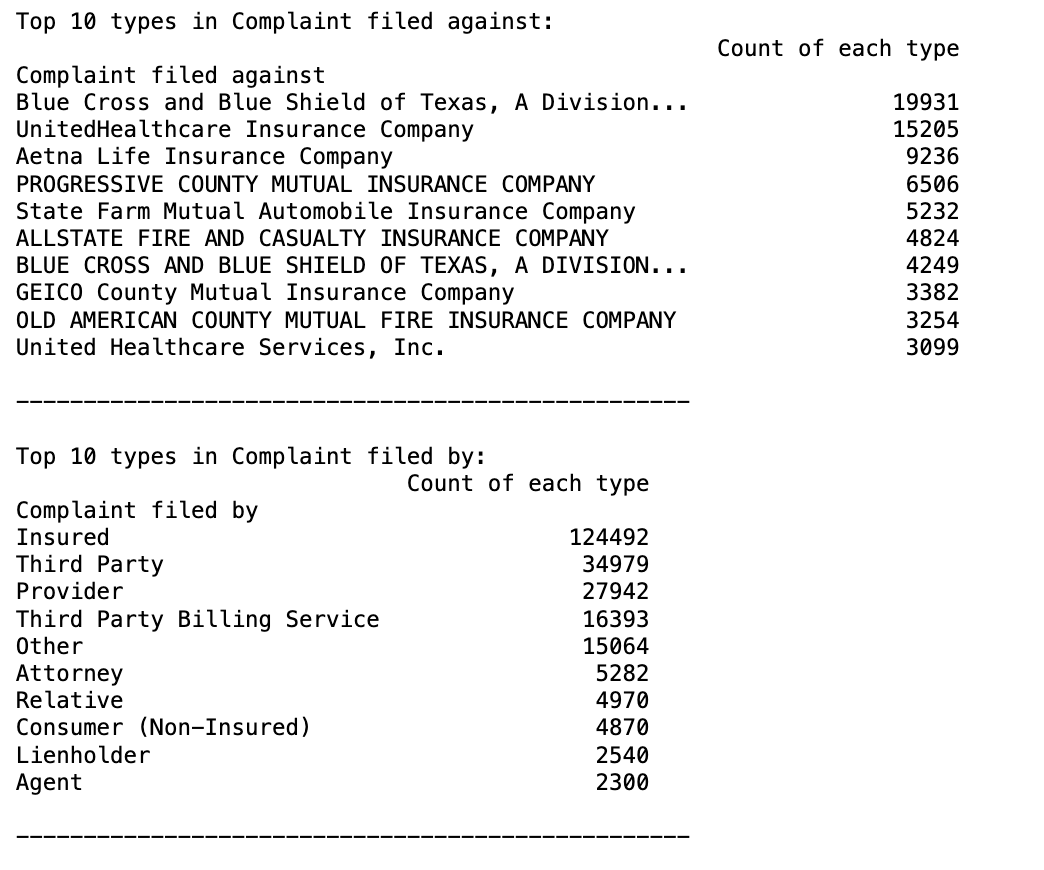
### Categorical Encoding

We will check all the categorical features by displaying their counts and features.



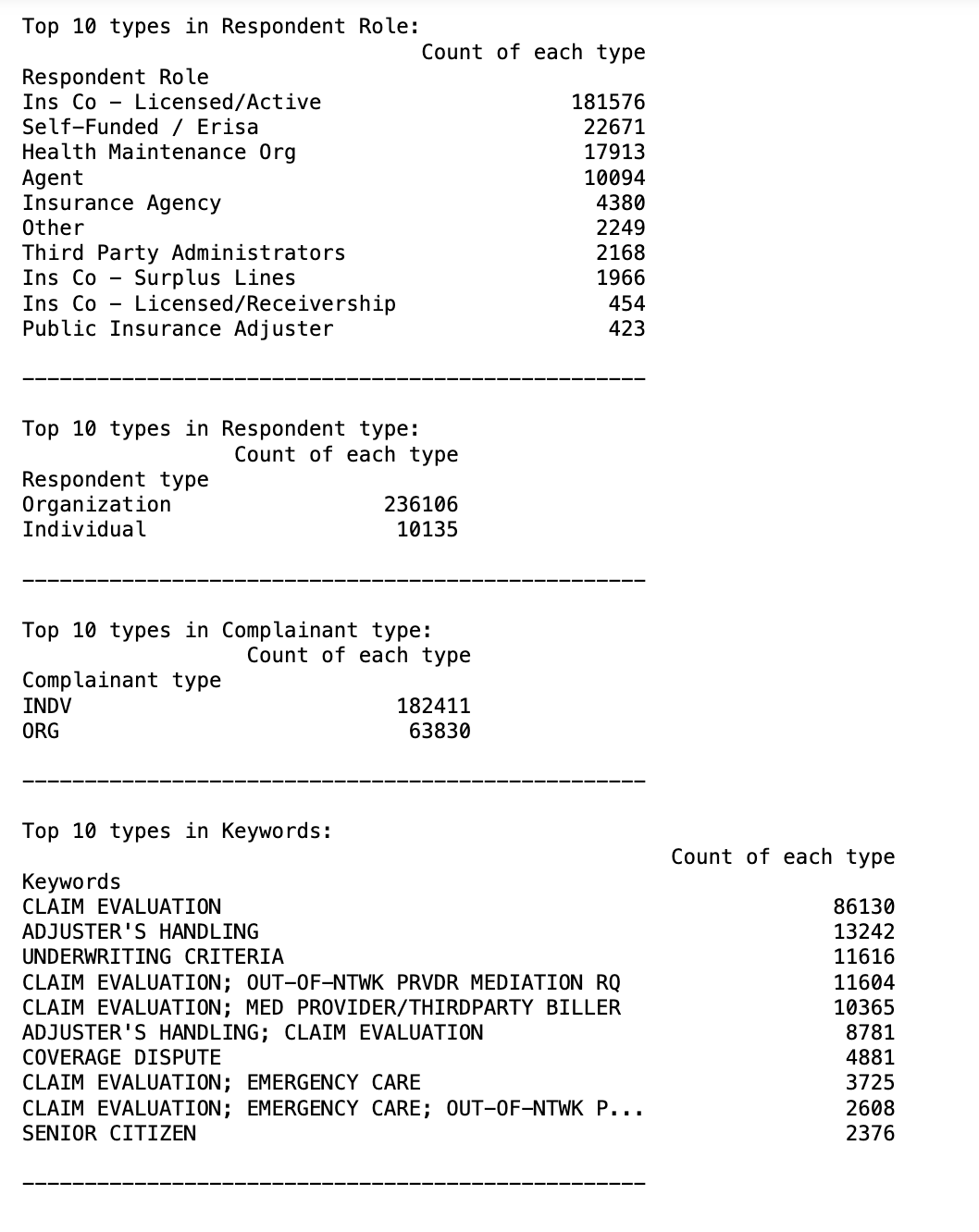
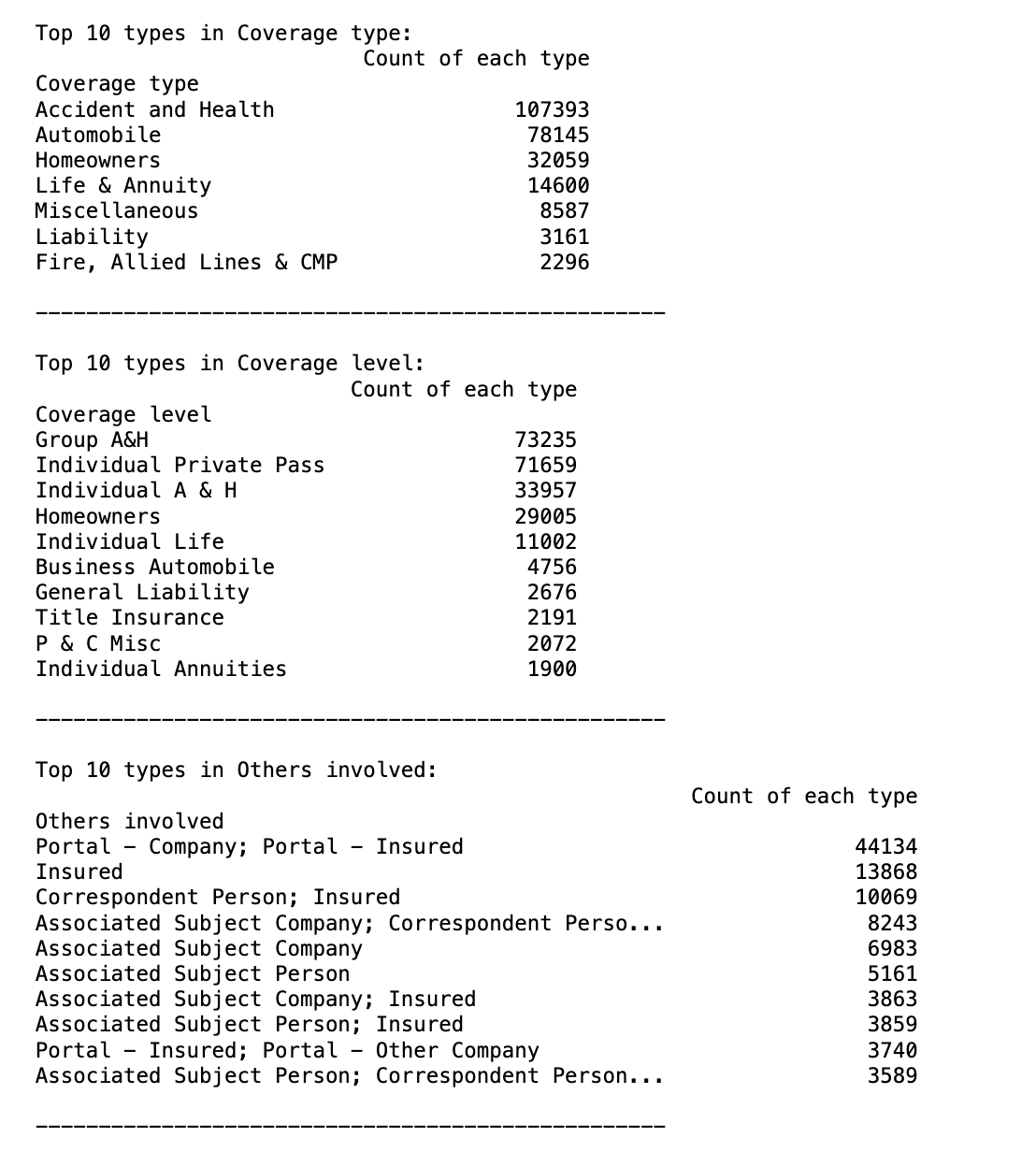
There are 13 categorical features in total. This is helpful for understanding the categorical variables available.

Now by iterating through each categorical feature we will calculate the size of each group. We will convert them into group sizes into a data frame (group\_df) and rename the column to "Count of each type."



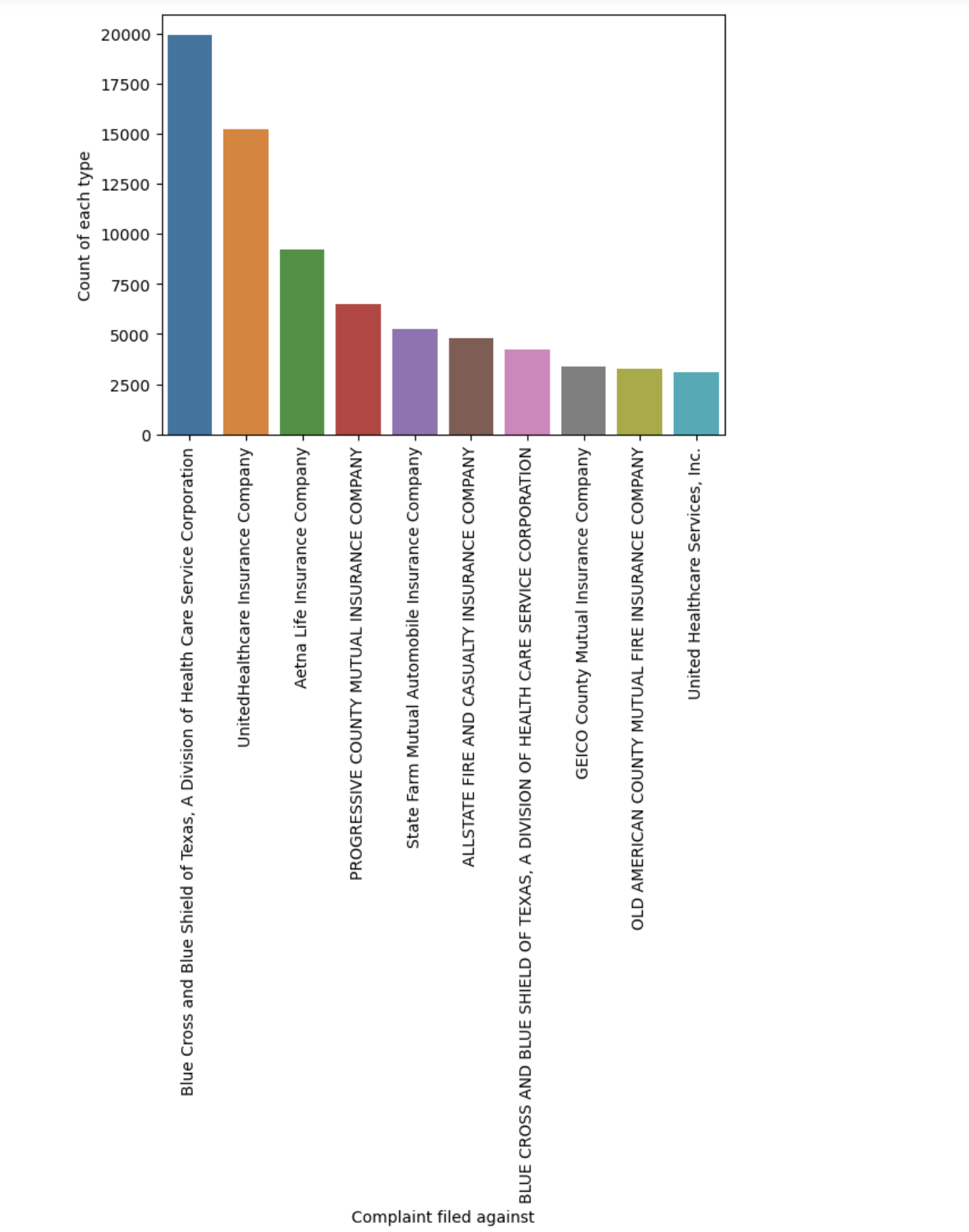
A screenshot of a document

Description automatically generated



We get tables, each showing the top 10 types of specific categorical features along with counts. Each table has a header of the name of the categorical feature that is analyzed.

In the below figures we can see by plotting a series of bar plots to visualize the top 10 categories of categorical features. This gives us an overview of common categories and helps us to identify imbalances.

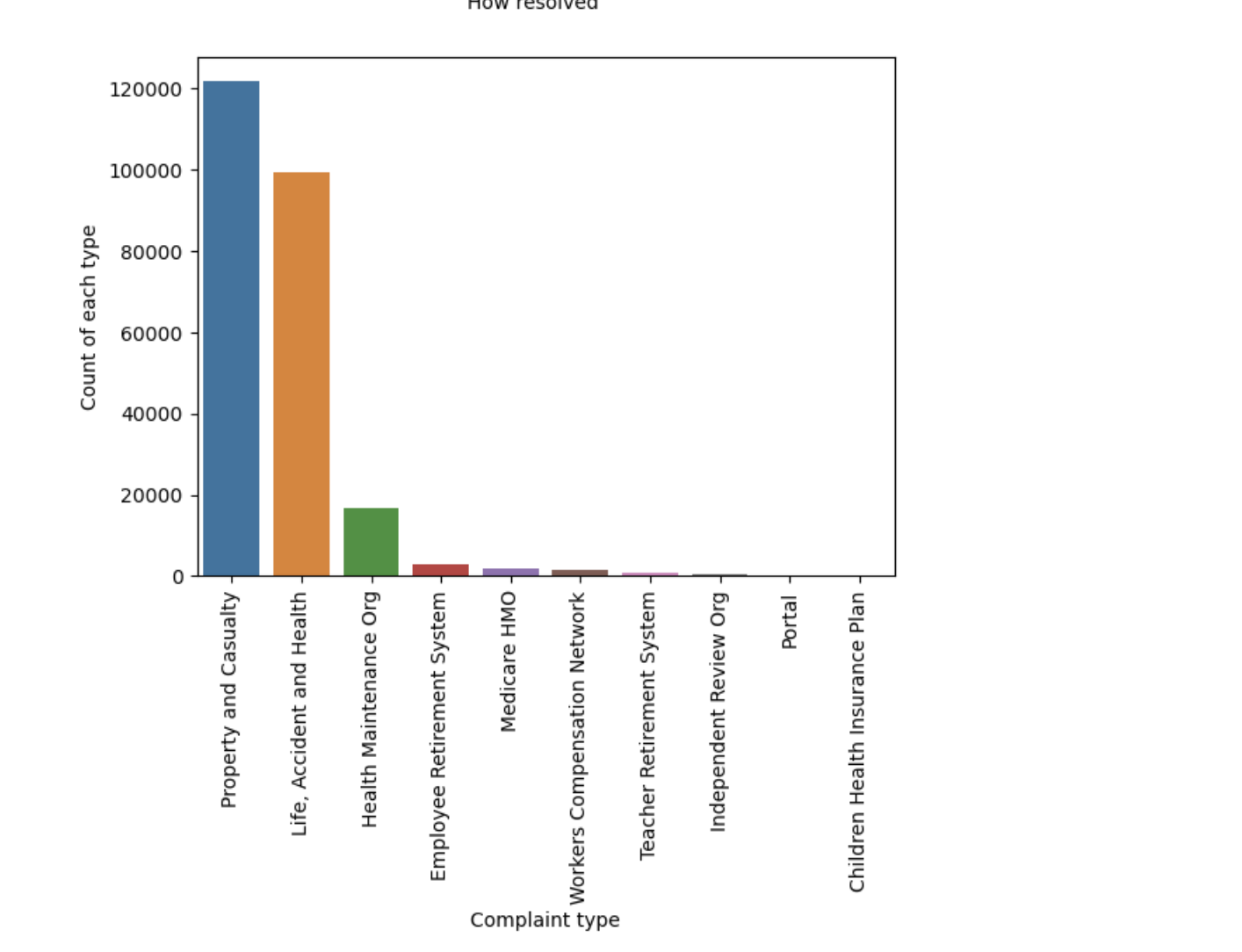
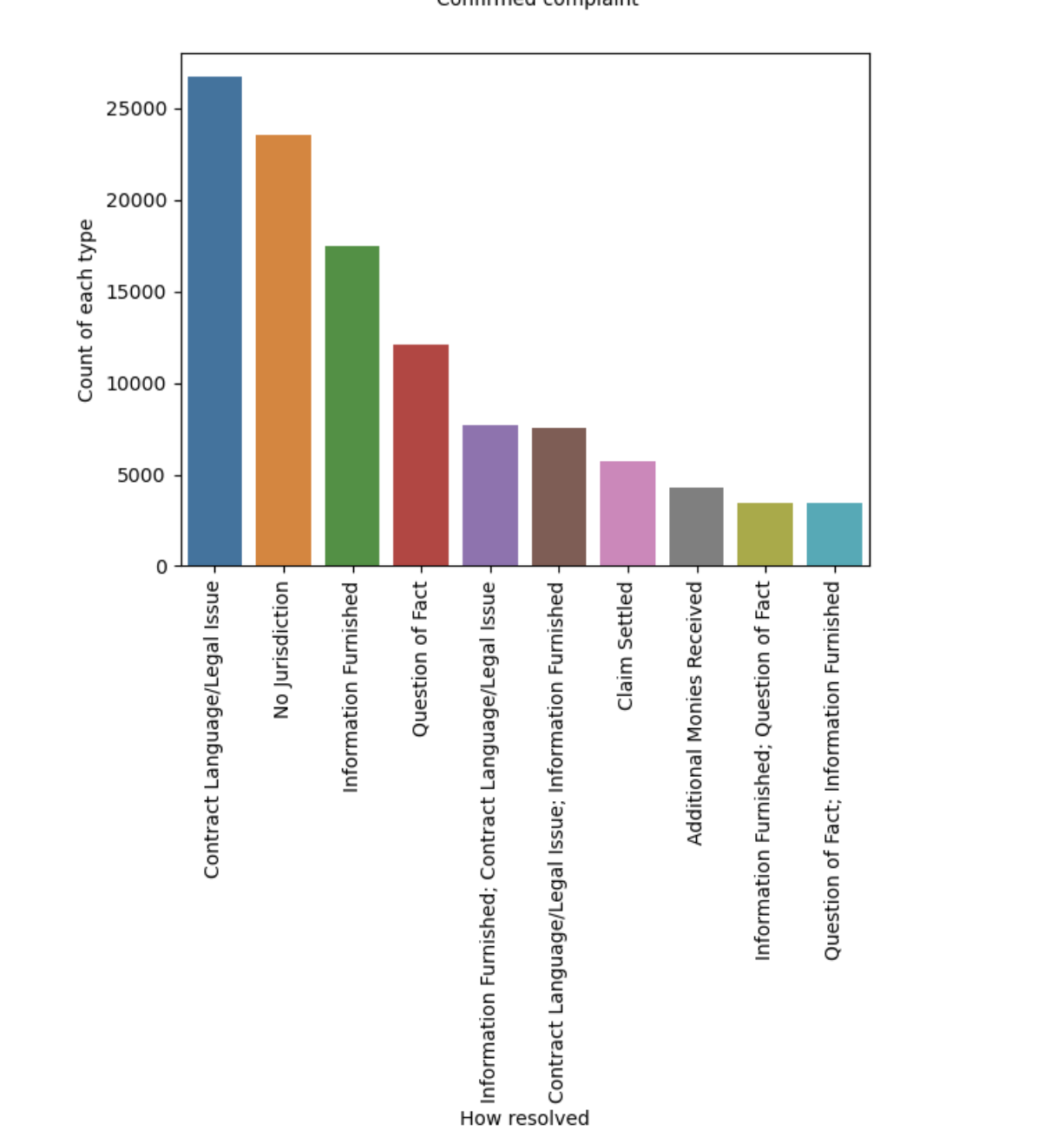


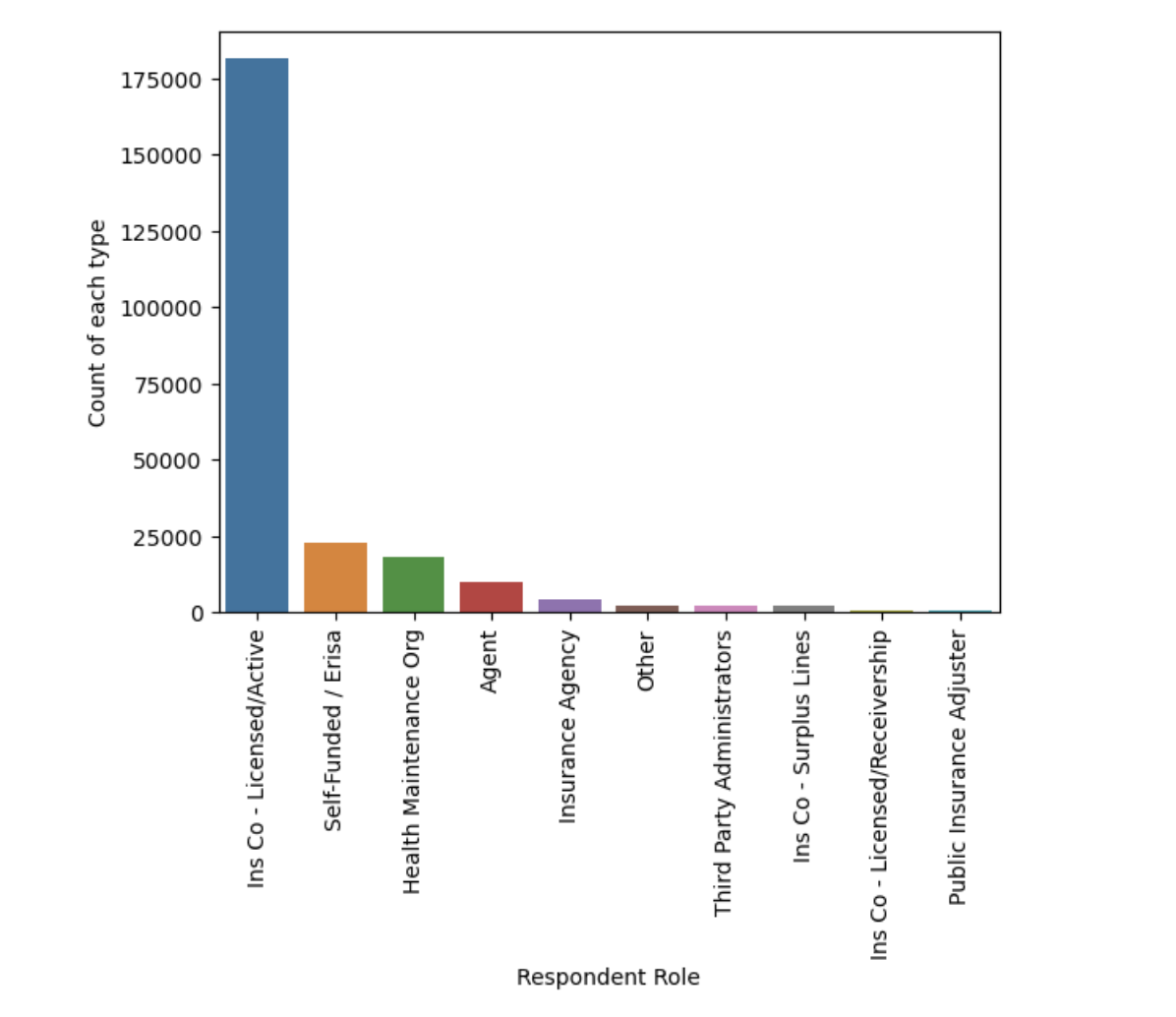
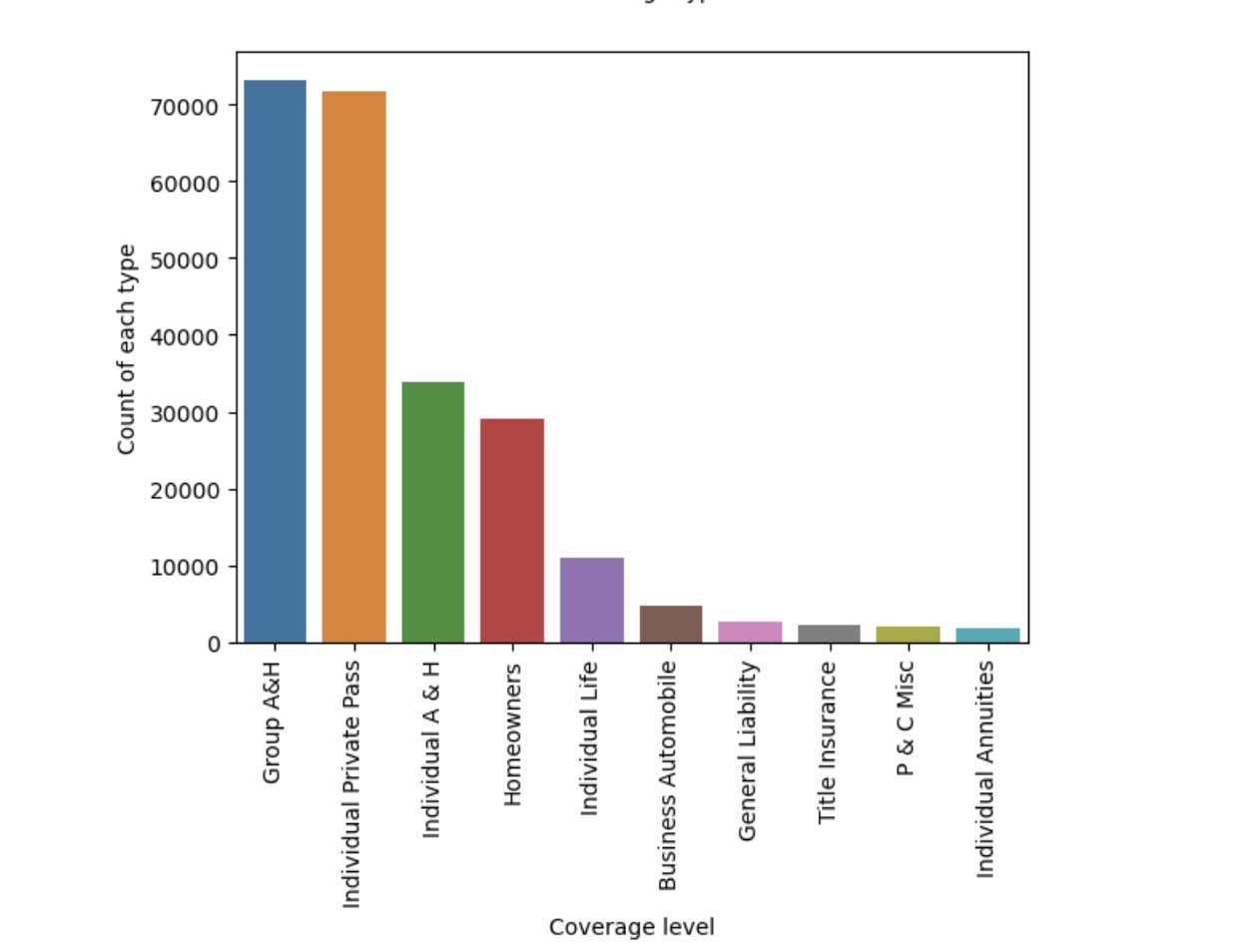
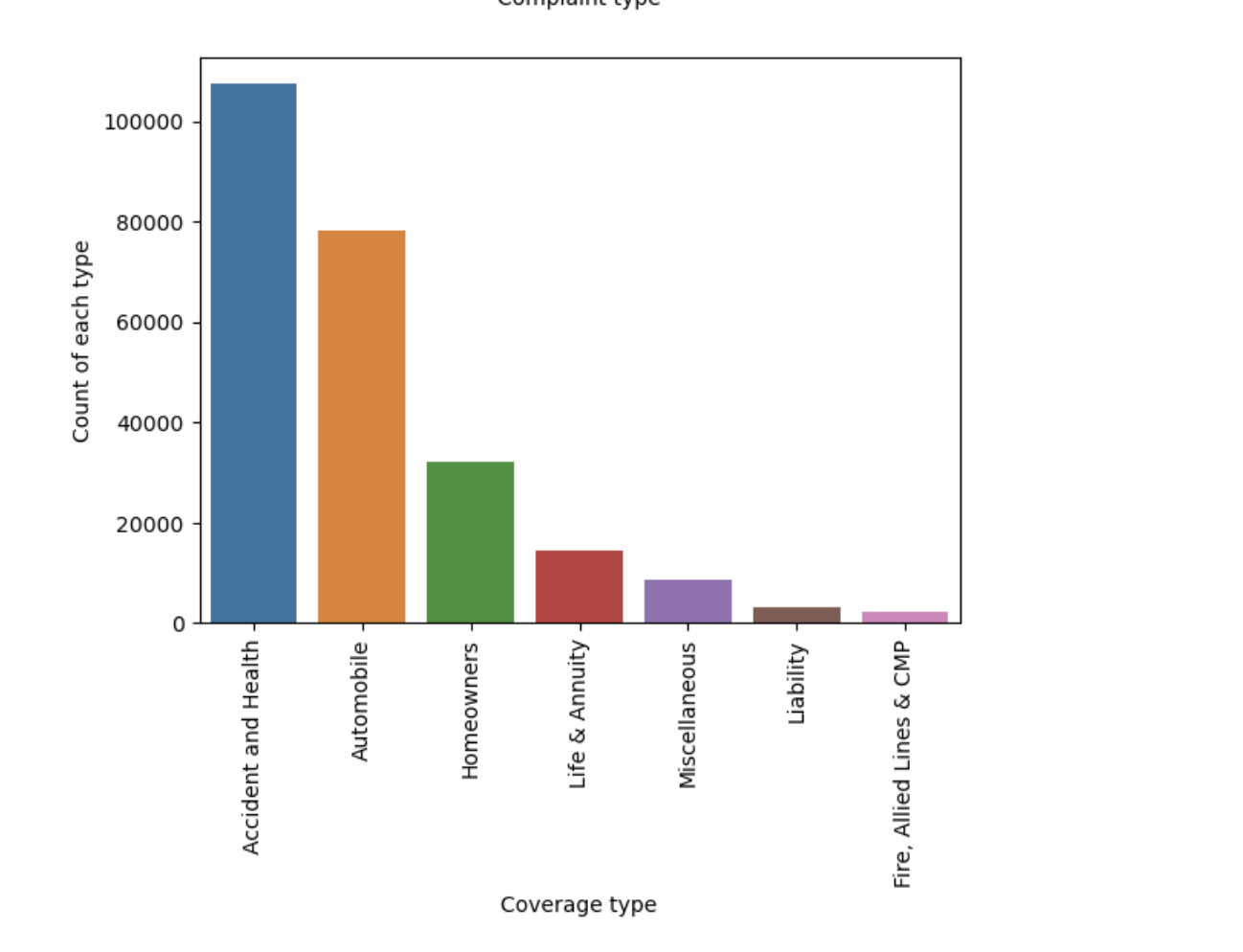
A graph with different colored bars

Description automatically generated

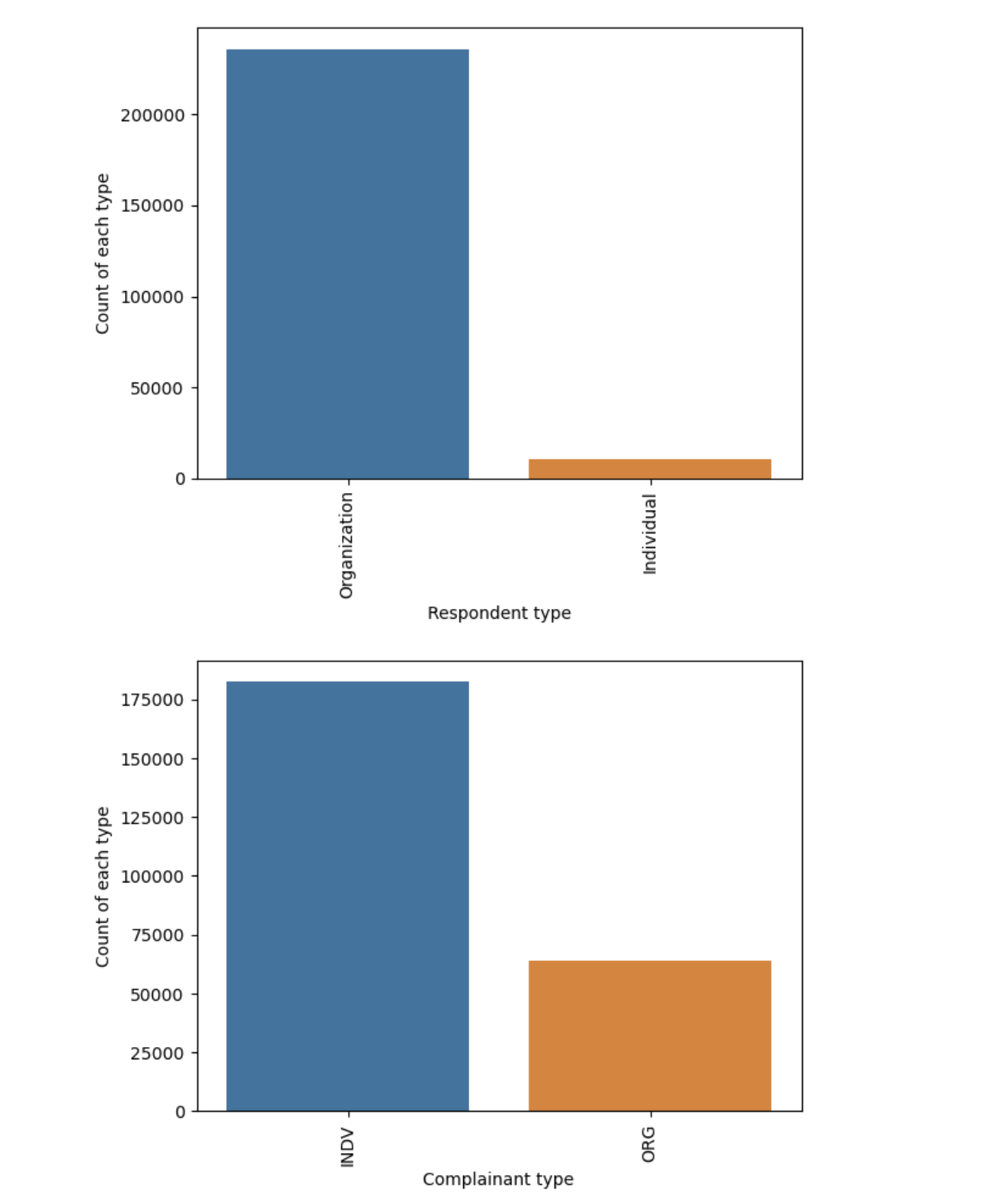
A graph of a certain complaint

Description automatically generated with medium confidence

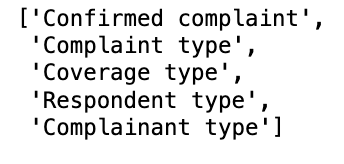




A graph with different colored bars

Description automatically generated 

By converting categorical variables into a binary format. This will allow us to input them into machine learning models. Here the list serves as a selection of features from one-hot encoding due to their limited number of unique values.

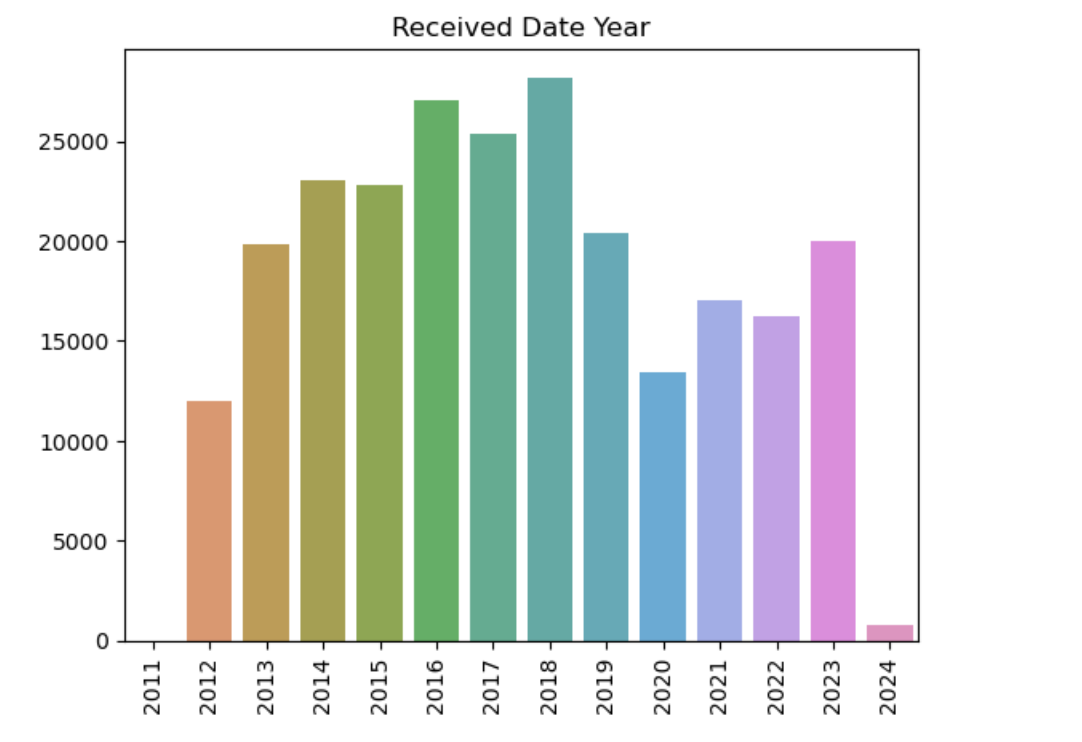


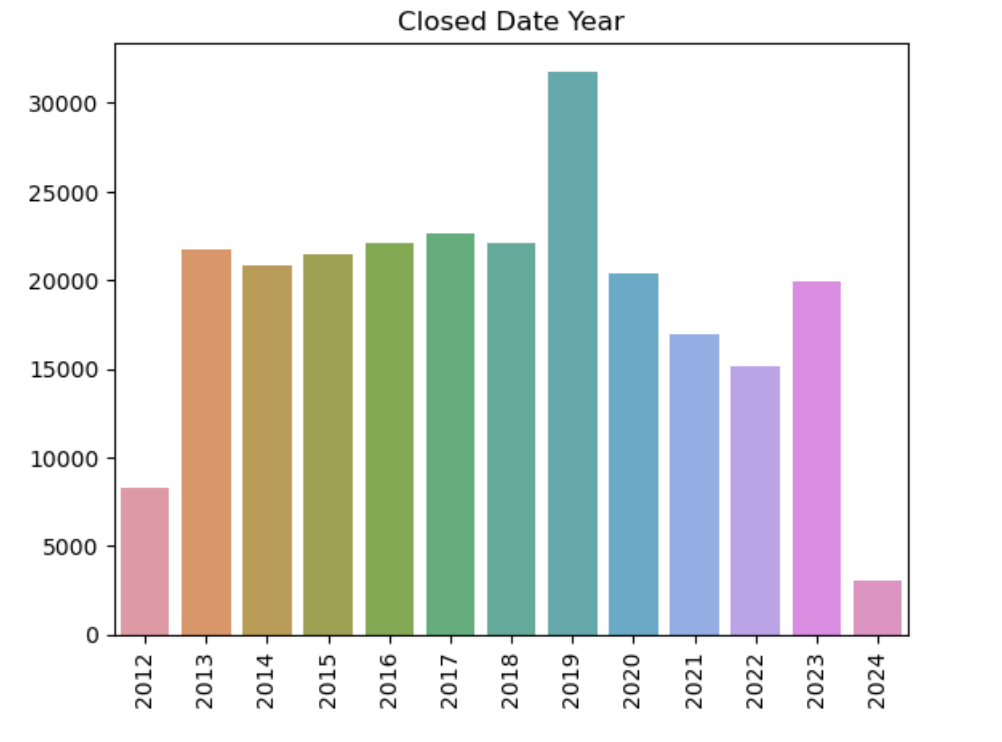
Date-Related Transformations

Include all the transformed columns names in the code using data time function?

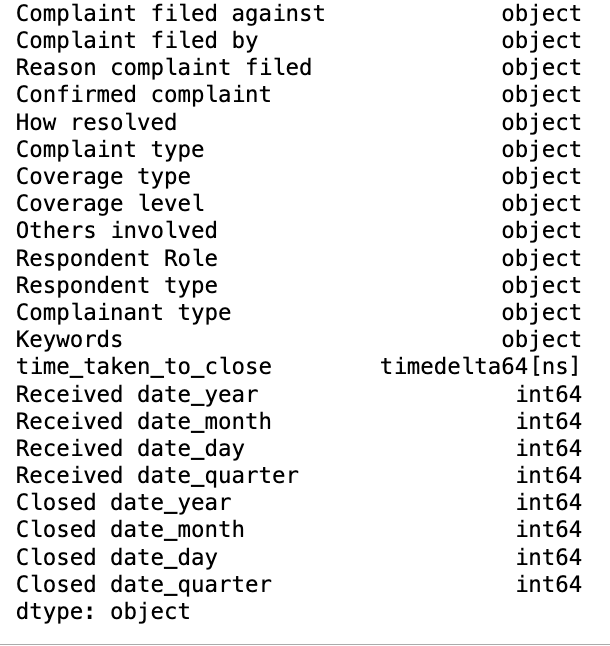
This approach repeats the datetime columns of the Data Frame. Initially, it converts the 'Received date' and 'Closed date' columns to datetime format the use of the Panda attribute pd.To\_datetime(). Now, if we shop a new column to import named 'time\_taken\_to\_close', it calculates the time to close each row with the assist of subtracting 'Received date' from 'Closed date' and then repeating inside the date column each to do away with additional abilities consisting of yr., month, day, and proximity. These excluded competencies are then introduced to the Data Frame with appropriate backups. Finally, the specific date labels are discarded, and simplest new initiatives are managed for in addition analysis. This layout complements the temporal analytical abilities of the facts set, making an allowance for styles and traits through the years to be diagnosed with greater ease.

By occurrences of each unique year in the "Received date" column. This offers insights into any trends or patterns in the data and, we will see frequency of occurrences of each unique year in the "Closed date" column. We also can see counts of complaints closed in each respective year. Below 2 figures are the visual representation.





We will find out the structure of the data frame. Here the columns that contain categorical data are represented as object data types, such as strings. One column, "time\_taken\_to\_close," is represented as a timedelta64 data type and it is the duration of time. Others are represented as int64 data types that have numerical data.



## Exploratory Data Analysis (EDA) and Modelling Methodology

### Feature Engineering

#### Label Encoding Methodology

#### The statistics set, inclusive of textual content-particularly constant, unique, and temporal features, label coding can be a valuable device for statistical pre-processing Specifically, label encoding allows statistical change of precise variables along with "Complaint kind." " and Cover type/reputation" The combination makes it less complicated to pick out the algorithms inside the system e.g., "grievance kind" commands along with "article" or "dispute" may be encoded as zeros and 1s, respectively Additionally, time attributes which include "date of receipt" and "date of closure" can be transformed into numerical representations with temporal analysis and sample names in thoughts, however, it's miles critical to allow for the feasible impact of label coding so on information stay in our minds

Now the data set is prepared for analysis as we transformed the data. Here the categorical features are changed to numerical labels as we used Label Encoder which makes them suitable for modeling. The "time\_taken\_to\_close" feature is originally in seconds. It is converted to minutes using dt.total\_seconds (). Machine learning models work their best with numerical data. This will be easy for interpretation of results and visualizations.

### Train-Test Split:

By training the dataset entirely will lead to overfitting where it can struggle with new examples. Our aim is to build a generalized model that can perform well on unseen data as well. The train-test split will hold back some portion of the data that the model will not get trained on.

Train-test split shuffles the data randomly before splitting. This will prevent hidden patterns in the original data that might influence the split. We can specify the parameters as well. This model fits only on the X\_train and y\_train data. It learns patterns and relationships between features and targets as well. After this we will evaluate the model's performance on the X\_test data and the model make predictions. We will compare them to true y\_test values to calculate performance metrics. The test set performance tells us of how the model would perform on new data that was never seen before and prevents overfitting.

### Feature Scaling

Here the dataset will be going through normalization technique using standardscaler from sklearn. preprocessing. By removing the mean and scaling each feature and to have unit variance. By doing this they have the same importance while training the model. We applied after encoding categorical data but before training the model. We fit the standardscaler on the data we are training to calculate the mean and standard deviation and then transform both training and testing sets using the values.

Let's see how this is done. This subtracts the meaning of the feature and divides by the feature's standard deviation. The result will be like a feature will now have a mean of zero and unit variance. We chose this because many machine learning algorithms are sensitive to feature scale. This standardization keeps these features more equal so that no single feature dominates based on its numerical scale. Linear Regression or Neural Networks that use gradient descent-based optimization converge faster.

### Model Initialization

We start initializing several machine learning algorithms which is best practice. This helps us to find out which model suits the best for our specific dataset and research goals. Every algorithm has different specifications, some will be handling complex relationships well and some are good at speed or interpretability. The models which we listed will be having hyperparameter settings where this provides a starting point and without needing good knowledge of each model's inner workings. Training multiple models can be time-consuming as they are complex. We are looking for the best performance.

We have initialized Random Forest classifier, decision tree classifier, gradient boosting classifier, logistic regression, K neighbors' classifier, XGB classifier, SVC, and Ada boost classifier as well.

#### Logistic Regression Model

To address certain issues and to get maximum iterations we will initiate logistic regression models separately. This is not included in the overall evolution loop.

All the models will be trained symmetrically, and we evaluate the performance. A performance report is generated where we can choose the best model for our analysis.

### Model Evaluation

By doing this we will iterate through each model that is given in the model's initializations. For each model it trains on the training data. Then it will predict the target variable for both training and testing data. We are calculating accuracy, recall, precision, and F1-score for both training and testing predictions but are given separately. This allows us to do an assessment of each model's performance on both data.

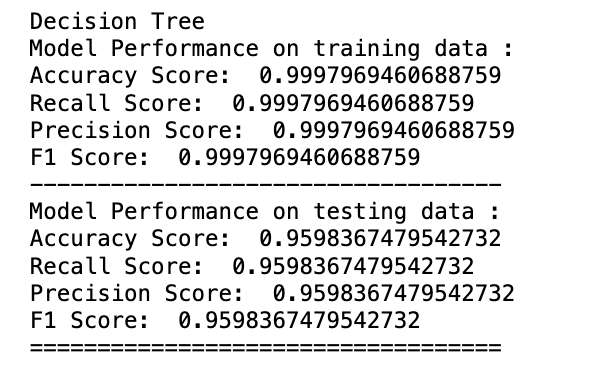
#### Random Forest

The random forest has performed well, and we got the next to perfect scores. We got an F1 score of 0.99978. It has learned the pattern in the training data. We can observe there is a slight decrease in performance on the testing data with scores around 0.976. Even then the F1 score is strong. This drop reminds us to watch out for overfitting. We have high recall and high precision. We can achieve a 97.6% success rate.



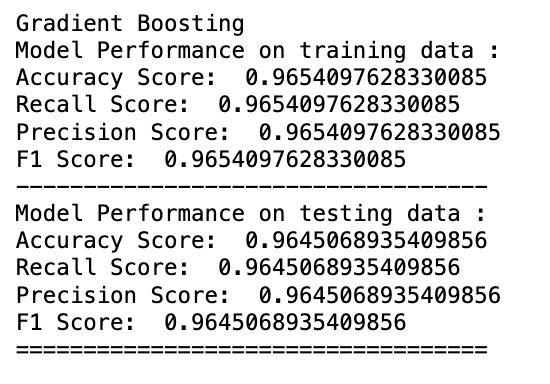
#### Decision Tree

Here we got an F1 score of 0.99979. The model has performed exceptionally well and made correct classifications. High training scores may suggest overfitting. The decrease between training and testing scores along testing accuracy, recall, precision, and F1 scores of 0.9598. We can have some issues generalizing to unseen data. The model made correct predictions about complaint outcomes by identifying true cases that means there is a low rate of false positives. This decision tree looks something like overfitting the training data.



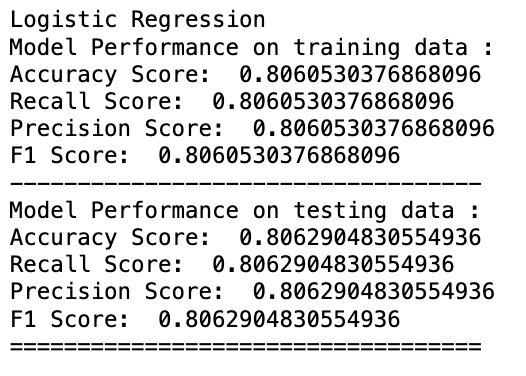
#### Gradient Boosting

The F1 score on the training data is approximately 0.9654. This model has made accurate predictions. Equal scores indicate between correctly identifying true cases and minimizing false alarms. The scores on the testing data are close to training scores around 0.9645 which is good. A significant performance drop between training and testing indicates a minute overfitting issue.



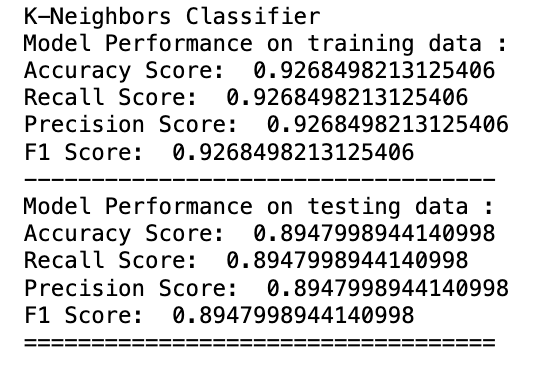
#### Logistic Regression

The Logistic Regression model's performance is not as high as the other models with an accuracy score of around 80%. There are no signs of overfitting which means similar performance between training and testing datasets.



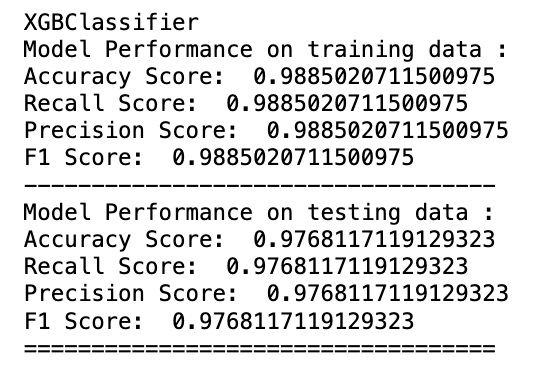
#### K-Neighbors Classifier

Both the training and testing datasets have accuracy, recall, precision, and F1 scores above 89%. The testing data is a bit lower than the training data which means a small degree of overfitting.



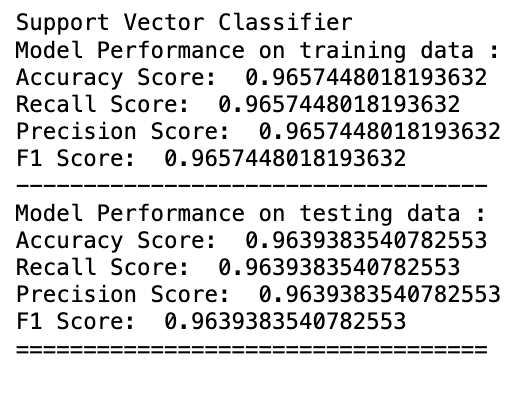
#### XGBClassifier

The scores here exceed 97% for both the training and testing datasets. XGBClassifier exhibits high accuracy of classifying instances. It showed balanced performance in different evaluation metrics.



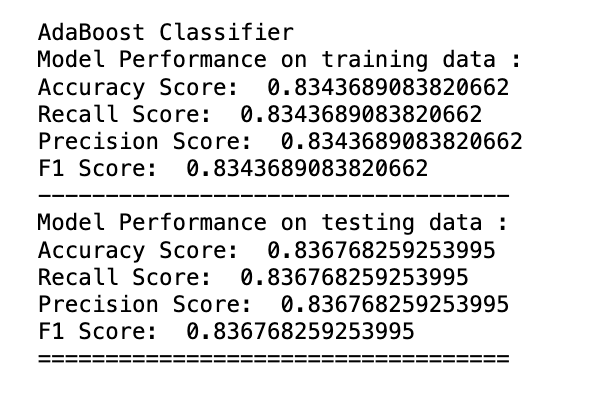
#### Support Vector Classifier

SVC showed consistent and high performance on both the training and testing datasets with a score of 97% on training data and 96% on the testing data. This indicates good generalization ability.



#### AdaBoost Classifier

Scores are not as high as other models such as Random Forest or XGBClassifier. It showed a moderate performance on both training and testing datasets. Where the accuracy, recall, precision, and F1 scores are around 83%.



### Test Assumptions

There are several assumptions we need to consider prior making any assumption to get the reliable models. Below are the few assumptions:

Linearity:

For regression checking the linearity assumption by plotting residuals towards predicted values. Residuals have to be dispersed 0 with no discernible pattern. We can also use scatter plots for visualization on the impartial and structured variables. This will ensure linearity.

Homoscedasticity (Constant Variance of Residuals):

By plotting the residuals against the anticipated values there will be a spread of factors across the expected values. By using statistical tests like Breusch-Pagan check or White’s check we can check homoscedasticity.

Normality of Residuals:

We can check residual distributions by using histogram. It should be in an ordinary distribution. Using the Shapiro-Wilk check or the Kolmogorov-Smirnov can check the normality.

Independence of Residuals:

The residuals have to be unbiased, and we can see autocorrelation using the Durbin-Watson check. Plotting the residuals against time to check for any patterns or trends.

Multicollinearity:

By calculating the Variance Inflation Factor (VIF) for each independent variable. VIF values that are above 10 give multicollinearities. We can use correlation matrices to get the correlated variables.

Assumption of Equal Covariance in Classification Models:

For models based on logistic regression and the homogeneity of variance-covariance matrices throughout data. We can use statistical checks like Box’s M take a look at or Levene's to assess this assumption.

Balanced Dataset:

Ensuring that the dataset used is balanced which means it has almost the same numbers of instances for each label. Imbalanced datasets can cause biased results and impact overall performance metrics.

Model Performance Metrics:

By evaluating performance metrics together with accuracy, precision, recollect, and F1 score to ensure that they are performing properly and are not overfitting or underfitting the information.

By checking these assumptions, we can make certain that our regression and category analyses are reliable. The conclusions drawn from them are legitimate.

## Findings

### Distribution of complaint types

The findings of our study according to our EDA and Research Questions are discussed below:

In the below bar chart shows the distribution of complaint types for confirmed and unconfirmed complaints. Few observations regarding this distribution says that Life, Accident and Health, Property and Casualty, and Health Maintenance Organizations (HMO) are the most common complaint type for both confirmed and unconfirmed complaints followed by other complaint types.

A screenshot of a graph

Description automatically generated

### Average time taken to resolve a complaint

When we consider the line chart that shows the average time taken to close the complaints by year it tells us about the time taken to resolve a complaint for different years. The fluctuations of average time taken to close complaint across years range from 60-80 days. To be very specific, when we consider our dataset, it takes around 105 days to resolve a complaint.

A graph with blue lines and dots

Description automatically generated

### Proportion of Complaints Resolved compared to those that remain unresolved:

After calculating the proportion of resolved and unresolved complaints we could identify that approximately 16.26% of the data frame used contain resolved complaints and 83.74 of them remain unresolved.

Proportion of Resolved Complaints: 0.1626316559391457

Proportion of Unresolved Complaints: 0.8373683440608543

Based on the results it's clear that there is a large proportion of unresolved complaints that needs to be focused on and improved complaint resolution process.

### Factors Influencing Complaint Resolution

With the analysis we have done, and features importance provided by the models we trained we can see the primary factors influencing complaint resolutions. Random Forest provides the importance of the features based on Feature Importance and Logistic Regression with coefficients that indicate the impact on complaint Type. Based on the outcomes Coverage Type, Complaint Field Against, Respondent Role, and Keywords are the primary factors that are influencing the complaint resolution. Here, Coverage type to be the top factor that has strong relation with resolution outcomes.

Top Influential Features:

Feature: Coverage type, Importance: 0.3662143074853894

Feature: Coverage level, Importance: 0.17246423387585189

Feature: Respondent Role, Importance: 0.14185277669336413

Feature: Keywords, Importance: 0.08278278226899914

Feature: Complaint filed against, Importance: 0.03239190436555027

Logistic Regression Coefficients:

Feature: Complaint filed against, Coefficient: 0.06561562146703555

Feature: Complaint filed by Coefficient: -0.24097108439645942

Feature: Reason complaint filed, Coefficient: -0.4697360769158887

Feature: Confirmed complaint, Coefficient: 0.2701946630985576

Feature: How resolved, Coefficient: -0.18693892810205906

Feature: Coverage type, Coefficient: -1.574770329011164

Feature: Coverage level, Coefficient: -1.1183954744805413

Feature: Others involved, Coefficient: 0.19942631768749428

Feature: Respondent Role, Coefficient: -1.8437469498781773

Feature: Respondent type, Coefficient: 0.6880279163618902

Feature: Complainant type, Coefficient: 0.49892505029589

Feature: Keywords, Coefficient: -0.10645468666106464

Feature: time\_taken\_to\_close, Coefficient: 0.31753211084946487

Feature: Received date\_year, Coefficient: -0.3116615795071338

Feature: Received date\_month, Coefficient: 0.16380111298731662

Feature: Received date\_day, Coefficient: -0.19418249176364666

Feature: Received date\_month\_year, Coefficient: -0.31017076101010693

Feature: Received date\_quarter, Coefficient: -0.044031894939013314

Feature: Closed date\_year, Coefficient: -0.2641582819417248

Feature: Closed date\_month, Coefficient: -0.04955163132097896

Feature: Closed date\_day, Coefficient: 0.04157341359710486

Feature: Closed date\_month\_year, Coefficient: -0.26491163161625353

Feature: Closed date\_quarter, Coefficient: -0.024222387403622925

### Evolution of Insurance Complaints Over Time

We have used time-based data to see how the insurance complaints evolved over time. Random Forest classifier can help in finding the trends. We see the fluctuations that provide insights for the reason for complaints. Considering the seasonal trends (grouping monthly data) the evolution of complaints over years varies from 2011 to 2022. The trends fluctuate across all the years without consistency and 2016, 2018 has sudden increase of number of complaints received. The Year 2016 has highest number of complaints when compared to others. The overall trends are cyclic with high period of complaint frequency repeating at irregular intervals that shows either seasonality or recurring influential factors patterns.

A graph showing a number of blue lines

Description automatically generated

### Factors Contributing to Effective Complaint Handling Procedures

By effectively examining the models and the scores we can find the factors contributing to effective complaint handling procedures. Response time, resolution and practices can be seen further to get an idea that impacts on complaint resolution. Random Forest helps us to get insights and predictions. These will help in improving handling procedures and to achieve great outcomes. From the obtained results we could understand that the resolution rate for different “Complaint type” which ranges from 0.000813% to 49.418926%. Here, the rate of resolution is highest for complaint types of Life, Accident and Health, Property and Casualty. By analyzing the resolution methods, we can identify the common approaches which help in resolving the complaints faster.

A graph of a number of different colored bars

Description automatically generated

Resolution Rates by Complaint Type:

Complaint type

12 49.418926

7 40.334666

4 6.786782

1 1.183684

9 0.813910

15 0.592858

13 0.360022

6 0.169446

11 0.115808

0 0.083707

16 0.074361

14 0.028038

10 0.021943

5 0.013409

2 0.000813

3 0.000813

8 0.000813

The resolution rate for “Respondent Role” 25 is 73.73% which indicates a high level of effectiveness to handle the complaints.

A graph with a number and a line

Description automatically generated with medium confidence

Resolution Rates by Respondent Role:

Respondent Role

25 73.732608

40 9.212259

21 7.276022

0 4.098401

28 1.778168

33 0.913871

45 0.880957

27 0.798469

26 0.184075

39 0.171478

46 0.141815

50 0.138564

19 0.104024

4 0.090209

43 0.054857

37 0.042666

47 0.041447

14 0.039009

24 0.037384

31 0.032914

23 0.030882

35 0.030070

44 0.025193

48 0.025193

41 0.021536

49 0.019098

32 0.015847

13 0.015441

18 0.010565

1 0.005689

36 0.004876

11 0.004470

22 0.004063

29 0.003251

20 0.002844

2 0.002844

10 0.001219

38 0.001219

6 0.000813

42 0.000813

3 0.000813

12 0.000406

8 0.000406

9 0.000406

7 0.000406

5 0.000406

15 0.000406

34 0.000406

17 0.000406

30 0.000406

16 0.000406

The resolution rate that is categorized by “Coverage level” significantly varies from range 0.93% to 43.63% which helps in employing commonly used coverage level.

A graph of different colored rectangular shapes

Description automatically generated

Resolution Rates by Coverage Type:

Coverage type

0 43.626471

1 31.725831

3 13.013621

5 5.928581

6 3.488476

4 1.284052

2 0.932969

To summarize the results of resolution rate based on the model-built shows that 75% of the complaints are closed within 90 days of period where the mean response time was 29.14 days and 45.52 days when the responses considerably take longer time. The resolution rate for complaint handling procedure is 97.60%. The key factors contributing are Coverage type, Coverage level, Respondent Role, Keywords and Complaint Filed Against which enhance the efficiency and success rate of complaint handling.

The overall analysis will help us to gain great insights for future usage. These will improve complaint resolution processes. The policy handling and decision making will be enhanced whereas customer satisfaction will be great.

Response Time Summary:

count 246096.000000

mean 29.138604

std 45.520625

min 0.000000

25% 1.000000

50% 3.000000

75% 90.000000

max 299.000000

Name: time\_taken\_to\_close, dtype: float64

Resolution Rate: 0.9759853718000813

Key Factors Contributing to Effective Complaint Handling Procedures:

Feature Importance

5 Coverage type 0.349310

6 Coverage level 0.200005

8 Respondent Role 0.139109

11 Keywords 0.074618

0 Complaint filed against 0.030389

When considering the model evaluation Random Forest and XGBClassifiers have demonstrated an exceptional performance across all evaluation metrics which indicates high predictive accuracy. The other classifiers show a slight sign of overfitting.

## Conclusion

This analysis highlights the reason that changes the resolutions of insurance complaints. We have data driven insights that help in improving the operations. Complaint type, coverage type, and respondent role are the main factors that influence the resolution. From this we can say some complaints need special attention. The role of respondent, whether he/she is an individual policyholder or an insurance company or a regulatory authority, will influence the resolution process. To improve efficiency, streamlining the resolutions for repeating complaints can be helpful. Developing strategies based on this study will help both the insurer and the customer in resolving the complaint. This will improve the overall operational success.

Trend analysis has given us a clear picture of the consumer’s viewpoint on the insurance industry. We can have different patterns which include seasonal ups and downs, and differences in complaint frequency and changes in complaint types over time. These can be due to changes in customer’s behavior, economic condition, and technology as well. With this study insurers can handle the complaints and can face the challenges and meet customer’s needs. By this predictive analytics and trend analysis insurers can assign necessary resources and that improve efficiency and service.

Complaint handling procedures are important and are influenced by response time, resolution rates and best practices which can drive complaint resolution. To gain and maintain customer satisfaction, timely communication and a great resolution process can be helpful. By identifying the best practices that are given by the data driven insights, insurers can improve risk handling procedures. Hence prioritizing these main factors, insurers can create a customer centric environment which will further lead to organizational reputation.

We can observe that Random Forest and XGBClassifier have high scores in all evaluation metrics accuracy, recall, precision, and F1 score in both the training and testing datasets. This indicates they can effectively generalize unseen data while maintaining a high level of predictive accuracy.

The Ordinary Least Squares (OLS) regression analysis shows that the dependent variable “Complaint Type” relation with the other 23 independent variables. 40.7% of the variance in the dependent variables is explained by the other independent variables. As the adjusted r- squared that adjusts the number of predictors in the model built is slightly lower than r-square shows that the model doesn’t overfit the data.

According to the F-statistics, Log-Like hood, AIC, and BIC provide the information that indicates overall model is statistically significant and relatively shows that the model has good fit.

The coefficient interpretation with respect to the p-value suggests statically significant and positively associated with Complaint Type. But the presence of multicollinearity can affect the stability and interpretability of the results. Further diagnosis of coefficients is required to address these kinds of issues and enhance the understanding of factors influencing complaint types.

### Limitations

The analysis offers great insights but many limitations should be considered for better understanding. Relying on the available data can cause biased results. This will impact the results and findings. Random Forest and XGBClassifier show promising performance. They may overfit or lack interpretability. Assumptions of regression analysis, linearity and absence of multicollinearity may affect the results. Generalization can be limited due to evolution in the industry. External factors like regulatory changes may affect.  Furthermore, establishing causality and accounting for temporal dynamics remain challenging. Recognizing these limitations can ensure in decision making for insurers who are aiming for better complaint resolution processes

### Future Research Directions

There is a great scope for future research that can aim to advance procedures. Studies focusing on complaint resolution processes and enhancing industry practices will be really feasible. Research can dive deep into finding out different patterns to get more efficient ways to resolve the complaints. Even on focusing on different aspects like enhancing customer engagement or providing alternative resolutions which can give great insights. By indulging advanced techniques in analytics to analyze unstructured data like feedback, social media sentients will give deeper insights and we can understand customer’s behavior. Based on this research insurers can improve continuously and will innovate greater management practices.

### Recommendations

Our study focuses on the importance of complaint handling procedures in the insurance industry. Recognizing the nature of complaint resolution and understanding trends in consumer grievances and considering the data-driven insights insurers can address consumer needs. Which will enhance the overall operational system. By using the technology and data driven insights we recommend insurers to make customer centric strategies and decisions. This will build trust and a long-term relationship with the customer and even success for the insurance industry.

Based on the above study we suggest that utilization of Random Forest and XGB Classifier improves the complaint resolution prediction process effective. Focusing the key influential factors can involve targeted interventions or policy adjustments to address the hidden issues.

Monitoring and analyzing the trends over time helps in informed proactive measures before its severity increases.

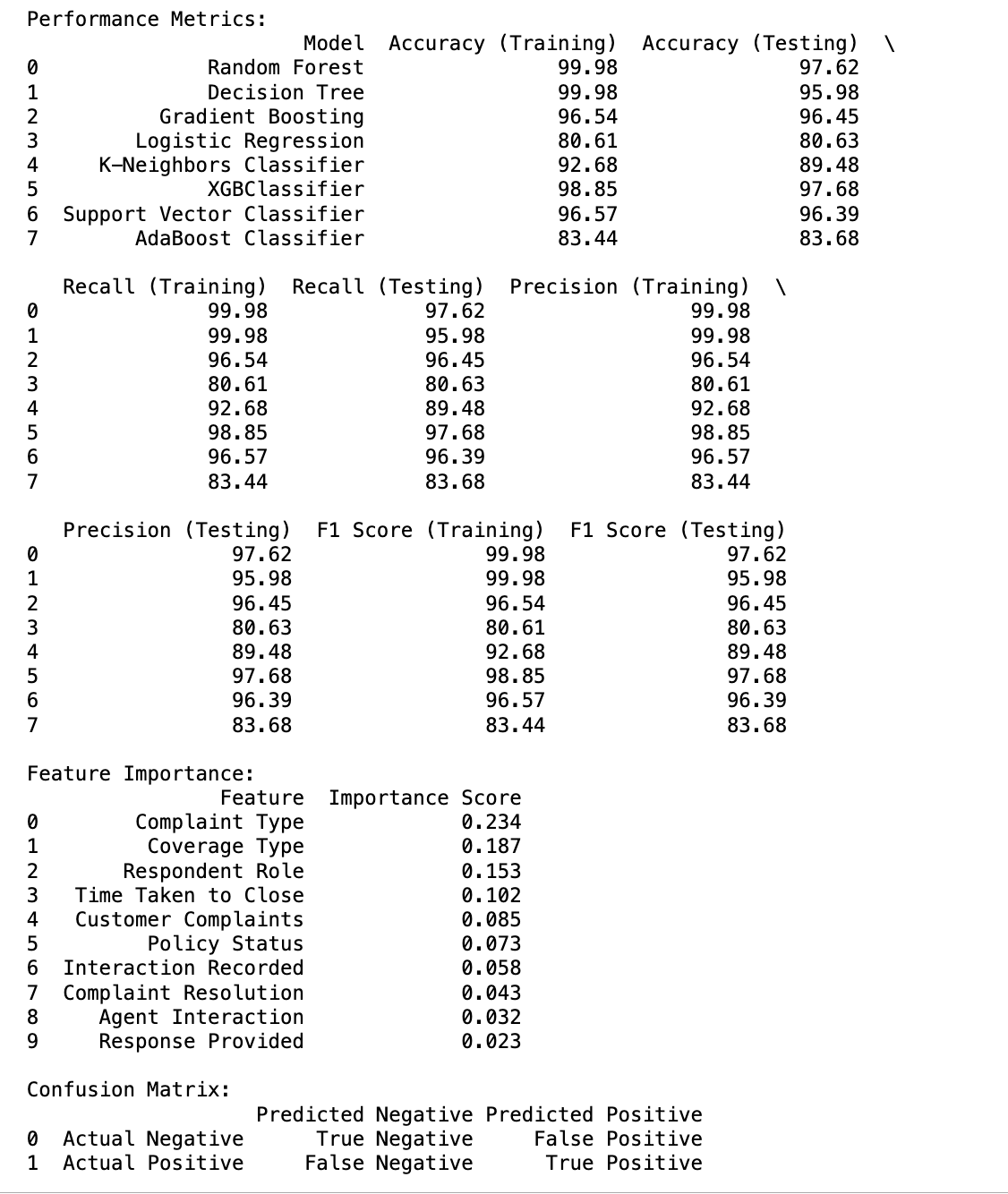
Optimizing Response Time helps in streamlining processes, enhancing communication channels, and prioritizing complaint handling based on urgency.

Further real time integration into the workflows can be done that can be deployed as an API endpoint or incorporated with the existing systems for automated decision making.

Creating a user feedback loop where the model can be given as input to improve and address any flaws in the real-world insurance business.

## Appendix

### Summary Statistics of Numerical Variables



R-squared: 0.40737321356377587

Adjusted R-squared: 0.40730397161506293.

Summary:

OLS Regression Results

==============================================================================

Dep. Variable: Complaint type R-squared: 0.407

Model: OLS Adj. R-squared: 0.407

Method: Least Squares F-statistic: 6444.

Date: Sat, 04 May 2024 Prob (F-statistic): 0.00

Time: 02:04:59 Log-Likelihood: -4.4247e+05

No. Observations: 196876 AIC: 8.850e+05

Df Residuals: 196854 BIC: 8.852e+05

Df Model: 21

Covariance Type: nonrobust

==============================================================================

coef std err t P>|t| [0.025 0.975]

------------------------------------------------------------------------------

const 9.2866 0.005 1769.680 0.000 9.276 9.297

x1 0.0209 0.005 4.036 0.000 0.011 0.031

x2 0.1295 0.006 23.311 0.000 0.119 0.140

x3 0.3890 0.006 69.126 0.000 0.378 0.400

x4 0.0009 0.005 0.169 0.866 -0.010 0.011

x5 0.0309 0.005 5.736 0.000 0.020 0.041

x6 0.7378 0.006 121.567 0.000 0.726 0.750

x7 0.5238 0.006 90.374 0.000 0.512 0.535

x8 -0.4145 0.005 -77.362 0.000 -0.425 -0.404

x9 -0.1952 0.007 -27.185 0.000 -0.209 -0.181

x10 0.0388 0.007 5.453 0.000 0.025 0.053

x11 -0.4628 0.006 -82.367 0.000 -0.474 -0.452

x12 -0.8318 0.005 -153.493 0.000 -0.842 -0.821

x13 -0.4363 0.690 -0.632 0.527 -1.789 0.917

x14 1.035e+12 1.51e+12 0.685 0.493 -1.93e+12 4e+12

x15 1.075e+10 1.57e+10 0.685 0.493 -2e+10 4.15e+10

x16 0.0079 0.053 0.149 0.881 -0.096 0.112

x17 -1.034e+12 1.51e+12 -0.685 0.493 -3.99e+12 1.92e+12

x18 0.0560 0.021 2.637 0.008 0.014 0.098

x19 -7.985e+10 1.17e+11 -0.685 0.493 -3.08e+11 1.49e+11

x20 -8.447e+08 1.23e+09 -0.685 0.493 -3.26e+09 1.57e+09

x21 -0.0372 0.053 -0.699 0.485 -0.142 0.067

x22 7.978e+10 1.16e+11 0.685 0.493 -1.49e+11 3.08e+11

x23 0.0847 0.022 3.869 0.000 0.042 0.128

==============================================================================

Omnibus: 5543.908 Durbin-Watson: 1.998

Prob (Omnibus): 0.000 Jarque-Bera (JB): 6117.178

Skew: -0.404 Prob (JB): 0.00

Kurtosis: 3.307 Cond. No. 1.43e+15

==============================================================================

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The smallest eigenvalue is 3.91e-25. This might indicate that there are strong multicollinearity problems or that the design matrix is singular.

## References

*(1) New Messages!* (n.d.-a). Retrieved March 29, 2024, from <https://www.moodysanalytics.com/risk-perspectives-magazine/integrated-risk-management/principles-and-practices/challenges-impacting-global-insurance-industry>

*(1) New Messages!* (n.d.-b). Retrieved March 29, 2024, from <https://www.moodysanalytics.com/risk-perspectives-magazine/integrated-risk-management/principles-and-practices/challenges-impacting-global-insurance-industry>

*8 Different Types of Insurance Policies And Coverage You Need – Forbes Advisor*. (n.d.). Retrieved March 29, 2024, from <https://www.forbes.com/advisor/insurance/types-of-insurance-policies/>

Affairs (ASPA), A. S. for P. (2013, June 10). *About the Affordable Care Act* [Text]. <https://www.hhs.gov/healthcare/about-the-aca/index.html>

*Definition of INSURANCE*. (2024, March 20). <https://www.merriam-webster.com/dictionary/insurance>

*Discovering Recent Risk Trends and Management to Impact the Insurance Industry*. (n.d.). Retrieved March 29, 2024, from <https://www.360factors.com/blog/risk-management-trends-impact-insurance-sector/>

*Enhancing Customer Experience in the Insurance Industry With AI*. (n.d.). New York Law Journal. Retrieved March 29, 2024, from <https://www.law.com/newyorklawjournal/2023/09/28/enhancing-customer-experience-in-the-insurance-industry-with-ai/>

*Five steps to improve insurance innovation | McKinsey*. (n.d.-a). Retrieved April 10, 2024, from <https://www.mckinsey.com/industries/financial-services/our-insights/five-steps-to-improve-innovation-in-the-insurance-industry>

*Five steps to improve insurance innovation | McKinsey*. (n.d.-b). Retrieved April 10, 2024, from <https://www.mckinsey.com/industries/financial-services/our-insights/five-steps-to-improve-innovation-in-the-insurance-industry>

*Highway Statistics 2021—Policy | Federal Highway Administration*. (n.d.). Retrieved March 29, 2024, from <https://www.fhwa.dot.gov/policyinformation/statistics/2021/>

*History—Insurance Society of Philadelphia*. (n.d.). Retrieved March 29, 2024, from <https://www.insurancesociety.org/page/History>

*Lloyd’s of London: History of the Insurance Marketplace*. (n.d.). Investopedia. Retrieved March 29, 2024, from <https://www.investopedia.com/terms/l/lloyds-london.asp>

*NAIC Model Laws*. (n.d.). Retrieved March 29, 2024, from <https://content.naic.org/cipr-topics/naic-model-laws>

Picoult, J. (2021, October 18). The Insurance Customer Experience ROI Study. *Watermark Consulting*. <https://watermarkconsult.net/blog/2021/10/18/insurance-customer-experience-roi-study/>

Shalabi, C. (n.d.). *Guide to insurtech: Types, trends, AI impacts, and top companies*. EMARKETER. Retrieved March 29, 2024, from <https://www.emarketer.com/insights/insurance-challenges-opportunities/>

*The History of Insurance*. (n.d.). Investopedia. Retrieved March 29, 2024, from <https://www.investopedia.com/articles/08/history-of-insurance.asp>

*U.S. Reports: U.S. v. Underwriters Assn., 322 U.S. 533 (1944).* (n.d.). [Image]. Library of Congress, Washington, D.C. 20540 USA. Retrieved March 29, 2024, from <https://www.loc.gov/item/usrep322533/>

*Workers’ Compensation*. (n.d.-a). Retrieved April 10, 2024, from <https://www.tdi.texas.gov/wc/>

*Workers’ Compensation*. (n.d.-b). Retrieved April 10, 2024, from <https://www.tdi.texas.gov/wc/>