Experiment 4

Aim:

Experiment to perform exploratory data analysis and data visualization using python

Theory:-

A. Descriptive analysis - statistical measures of data (Central tendency)

Descriptive statistics are brief informational coefficients that summarize a given data set, which can be either a representation of the entire population or a sample of a population. Descriptive statistics are broken down into measures of central tendency and measures of variability (spread). Measures of central tendency include the mean, median, and mode, while measures of variability include standard deviation, variance, minimum and maximum variables, kurtosis, and skewness.

Descriptive statistics, in short, help describe and understand the features of a specific data set by giving short summaries about the sample and measures of the data. The most recognized types of descriptive statistics are measures of center: the mean, median, and mode, which are used at almost all levels of math and statistics. The mean, or the average, is calculated by adding all the figures within the data set and then dividing by the number of figures within the set.

The mode of a data set is the value appearing most often, and the median is the figure situated in the middle of the data set. It is the figure separating the higher figures from the lower figures within a data set. However, there are less common types of descriptive statistics that are still very important.

Types of Descriptive Statistics

All descriptive statistics are either measures of central tendency or measures of variability, also known as measures of dispersion.

- **Central Tendency**

Measures of central tendency focus on the average or middle values of data sets, whereas measures of variability focus on the dispersion of data. These two measures use graphs, tables and general discussions to help people understand the meaning of the analyzed data. Measures of central tendency describe the center position of a distribution for a data set. A person analyzes the frequency of each data point in the distribution and describes it using the mean, median, or mode, which measures the most common patterns of the analyzed data set.

- **Measures of Variability**

Measures of variability (or the measures of spread) aid in analyzing how dispersed the distribution is for a set of data. For example, while the measures of central tendency may give a person the average of a data set, it does not describe how the data is distributed within the set. So while the average of the data may be 65 out of 100, there can still be data points at both 1 and 100. Measures of variability help communicate this by describing the shape and spread of the data set. Range, quartiles, absolute deviation, and variance are all examples of measures of variability.

Consider the following data set: 5, 19, 24, 62, 91, 100. The range of that data set is 95, which is calculated by subtracting the lowest number (5) in the data set from the highest (100).

What Is the Main Purpose of Descriptive Statistics?

The main purpose of descriptive statistics is to provide information about a data set. In the example above, there are hundreds of baseball players that engage in thousands of games. Descriptive statistics summarizes the large amount of data into several useful bits of information.

B. Descriptive analysis - statistical measures of data (Dispersion)

The measures of central tendency are not adequate to describe data. Two data sets can have the same mean but they can be entirely different. Thus to describe data, one needs to know the extent of variability. This is given by the measures of dispersion. Range, interquartile range, and standard deviation are the three commonly used measures of dispersion.

1. **Range**

The range is the difference between the largest and the smallest observation in the data. The prime advantage of this measure of dispersion is that it is easy to calculate. On the other hand, it has a lot of disadvantages. It is very sensitive to outliers and does not use all the observations in a data set. It is more informative to provide the minimum and the maximum values rather than providing the range.

2. **Interquartile Range**

Interquartile range is defined as the difference between the 25th and 75th percentile (also called the first and third quartile). Hence the interquartile range describes the middle 50% of observations. If the interquartile range is large it means that the middle

50% of observations are spaced wide apart. The important advantage of interquartile range is that it can be used as a measure of variability if the extreme values are not being recorded exactly (as in case of open-ended class intervals in the frequency distribution). Another advantageous feature is that it is not affected by extreme values. The main disadvantage in using interquartile range as a measure of dispersion is that it is not amenable to mathematical manipulation.

3. **Standard Deviation**

Standard deviation (SD) is the most commonly used measure of dispersion. It is a measure of spread of data about the mean. SD is the square root of the sum of squared deviations from the mean divided by the number of observations.

Appropriate use of Measures of Dispersion

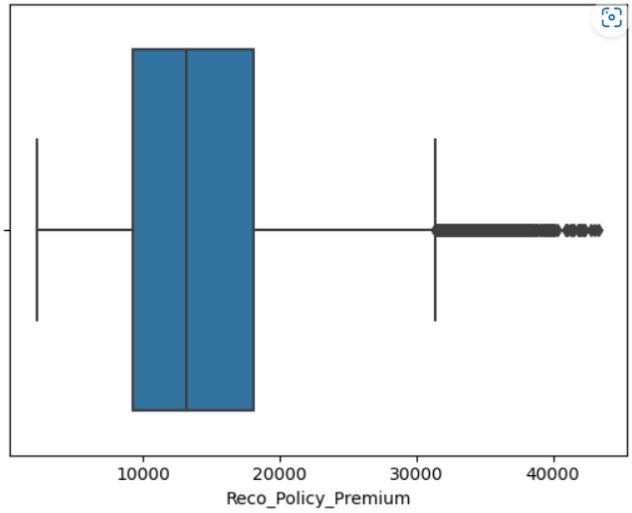
SD is used as a measure of dispersion when mean is used as measure of central tendency (ie, for symmetric numerical data). For ordinal data or skewed numerical data, median and interquartile range are used.

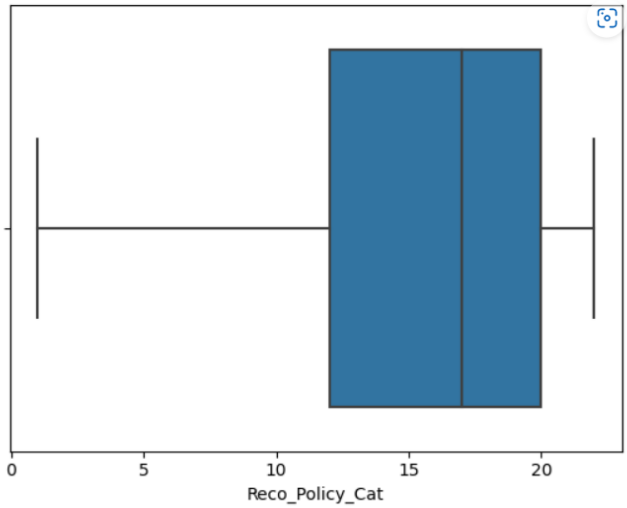
C. Correlation between attributes

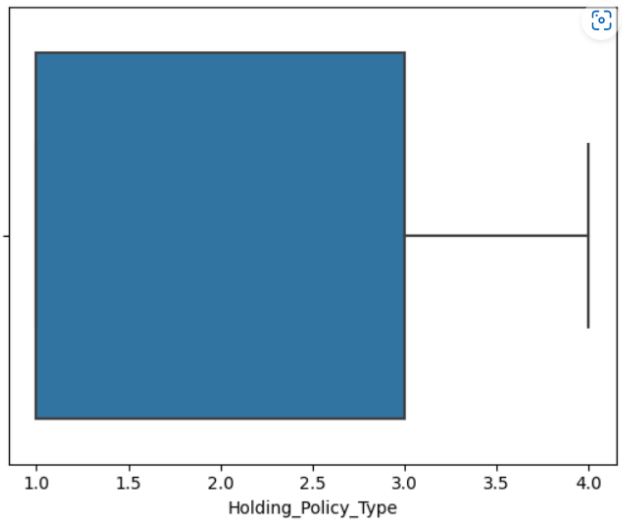
A correlation is a statistical measure of the relationship between two variables. The measure is best used in variables that demonstrate a linear relationship between each other. The fit of the data can be visually represented in a scatterplot. Using a scatterplot, we can generally assess the relationship between the variables and determine whether they are correlated or not.

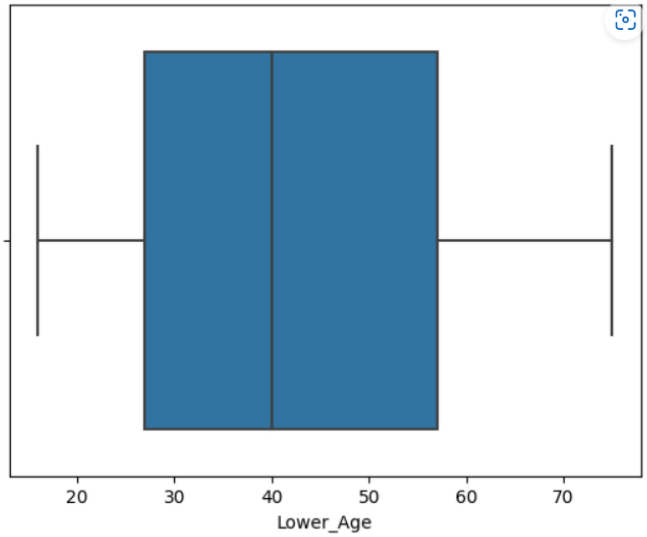
Screenshots of implementation:

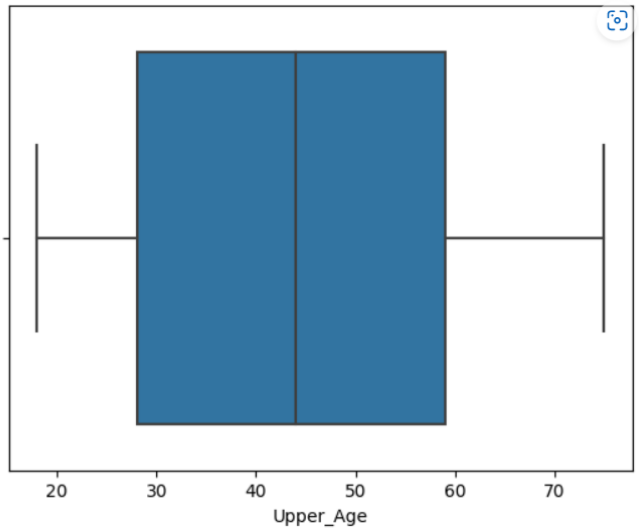
In [26]:
```
cols = df.drop('Response', axis=1).columns
for i in cols:
    try:
        sns.boxplot(df, x=i)
    except:
        pass
    plt.show()
```

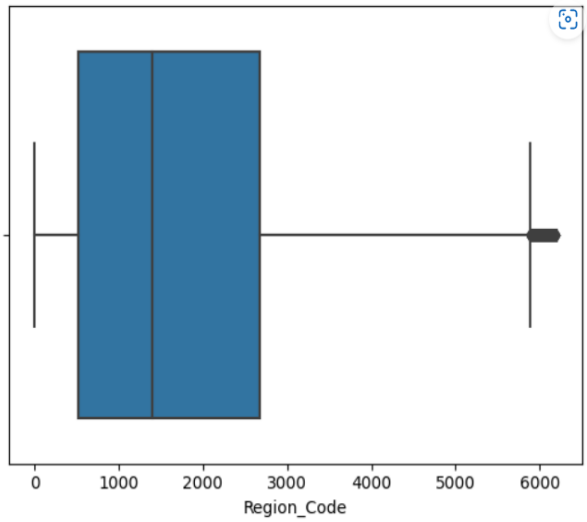












From the boxplots we can see the distribution of the data.

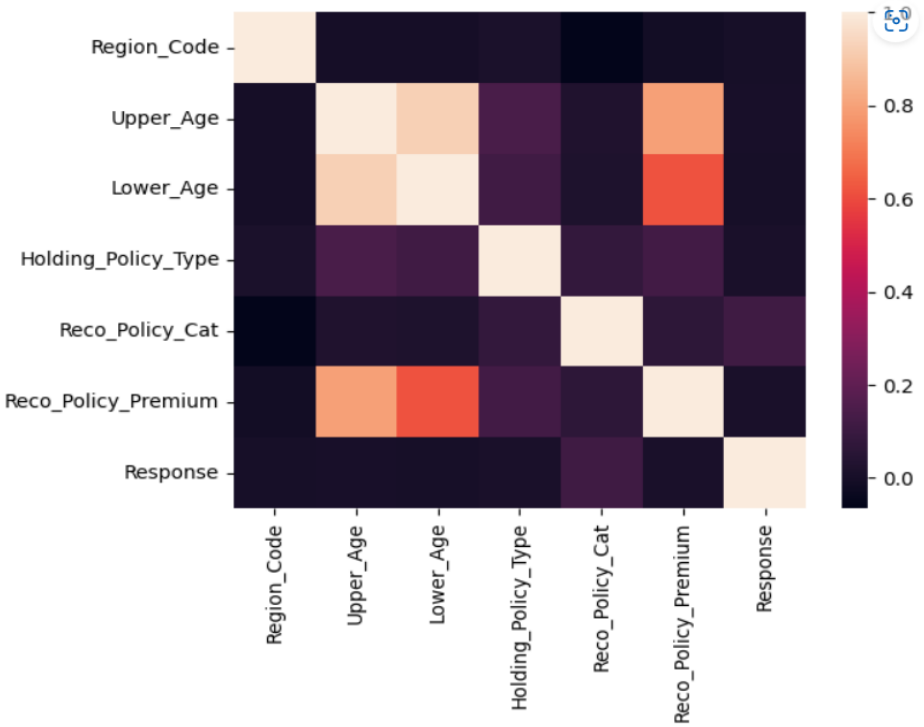
The premium column has a lot of outliers as it can depend on the insurance type and benefits provided by the insurance.

People in the age of 25 to 60 take the most number of insurance policies.

Consumers tend to take policy number 15 to 20.

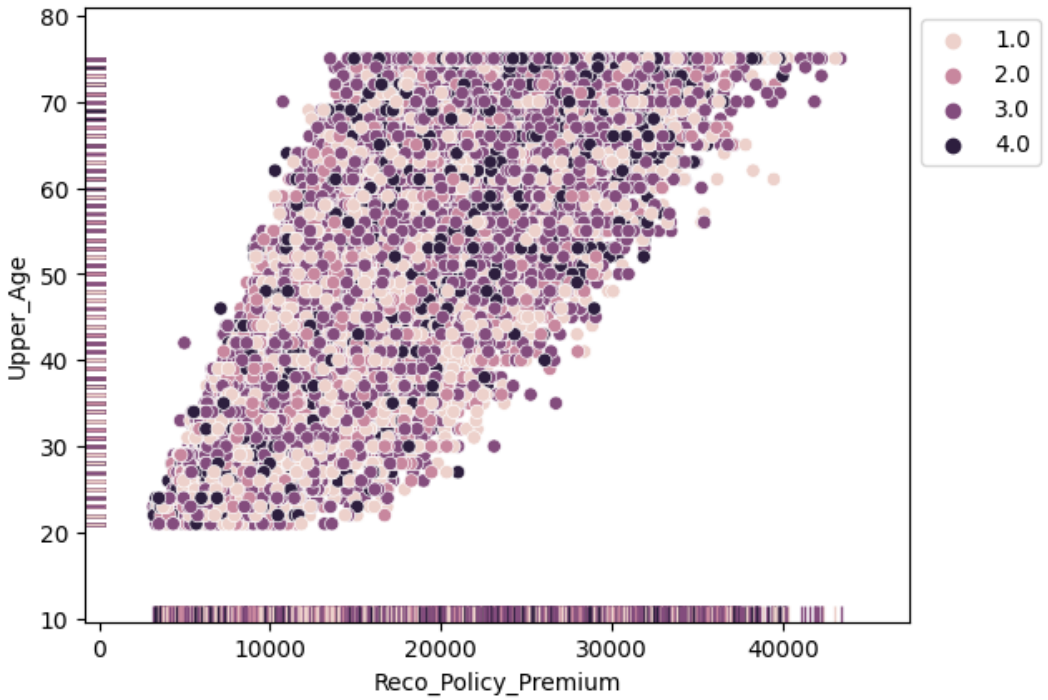
Correlation

```
In [19]: sns.heatmap(df.corr())
         plt.show()
```



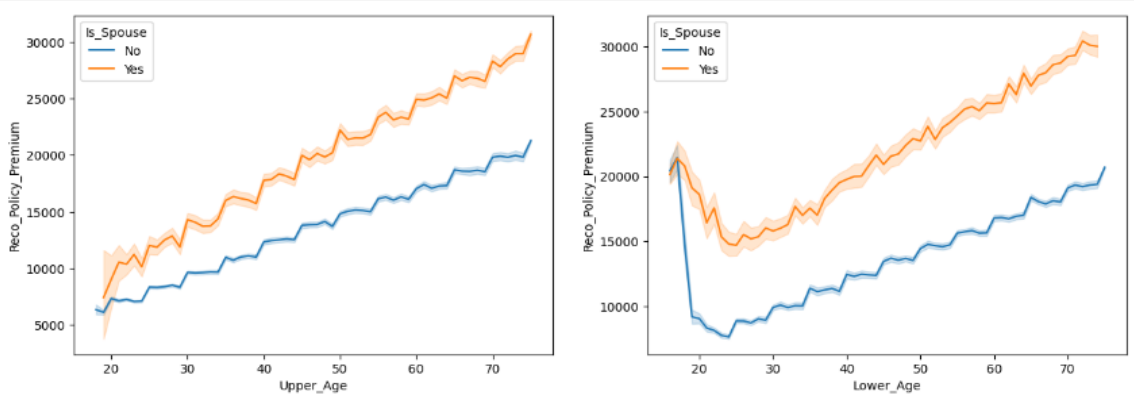
We can clearly see how age and premium are highly correlated.
Upper age and lower age are correlated with each other.
Other factors are not correlated with each other.

In [10]:
```
sns.scatterplot(data=df, y='Upper_Age', x='Reco_Policy_Premium', hue='Holding_Policy_Type')
sns.rugplot(data=df, y='Upper_Age', x='Reco_Policy_Premium', hue='Holding_Policy_Type')
plt.legend(bbox_to_anchor=(1,1))
plt.show()
```



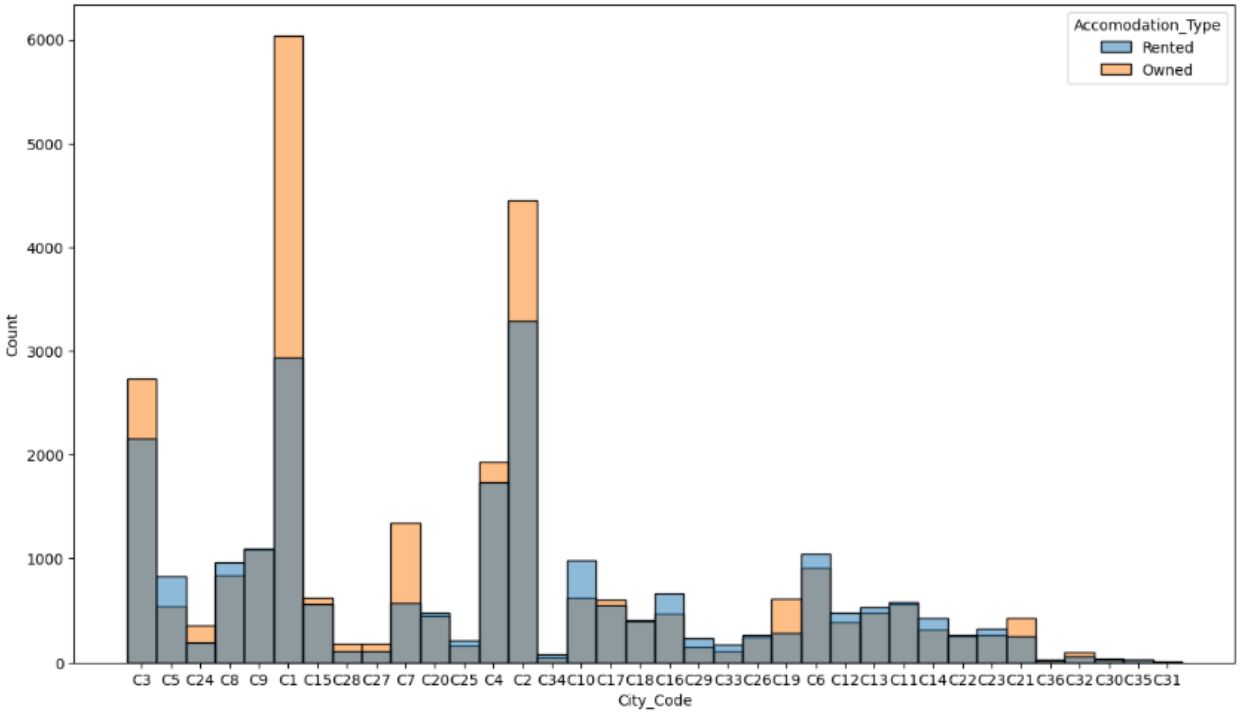
We can conclude that as the age increases the premium paid for insurance increases. Also people prefer to get a higher policy type as age increases. This might be because their salary increases or benefits provided by higher policy types appeal to them.

In [9]:
```
fig, axes = plt.subplots(1,2, figsize=(15,5))
sns.lineplot(data=df, x='Upper_Age', y='Reco_Policy_Premium', hue='Is_Spouse', ax = axes[0])
sns.lineplot(data=df, x='Lower_Age', y='Reco_Policy_Premium', hue='Is_Spouse', ax = axes[1])
plt.show()
```



People with spouses have higher premiums. This might be because the policy type becomes joint. Also age and premium increase simultaneously.

```
In [8]: plt.figure(figsize=(14,8))
        sns.histplot(data=df,x = 'City_Code', hue='Accomodation_Type')
        plt.show()
```



People from cities C1, C2 and C3 take more insurance. They might have more population or the people are more likely to buy insurance.

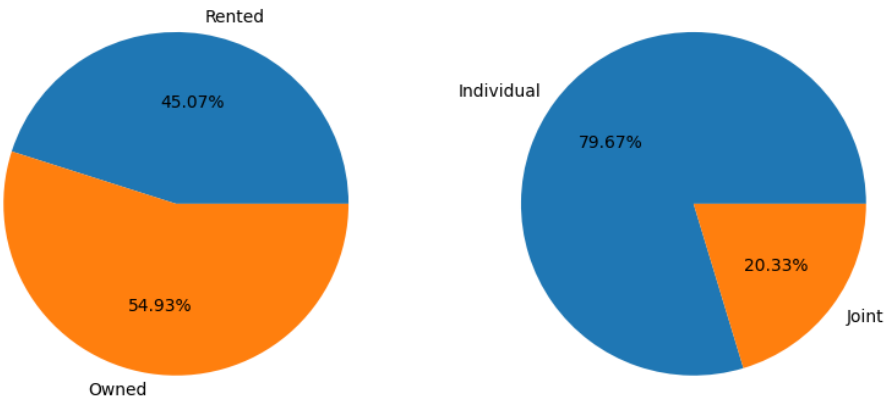
In [6]:

```
fig, ax = plt.subplots(1,2, figsize=(10,10))

accomodation_type = {'Rented':df['Accomodation_Type'][df.Accomodation_Type == 'Rented'].count(),
                     'Owned':df['Accomodation_Type'][df.Accomodation_Type == 'Owned'].count()}
ax[0].pie(accomodation_type.values(), labels=accomodation_type.keys(), autopct='%1.2f%%')

insurance_type = {'Individual':df['Reco_Insurance_Type'][df.Reco_Insurance_Type == 'Individual'].count(),
                  'Joint':df['Reco_Insurance_Type'][df.Reco_Insurance_Type == 'Joint'].count()}
ax[1].pie(insurance_type.values(), labels=insurance_type.keys(), autopct='%1.2f%%')

plt.show()
```

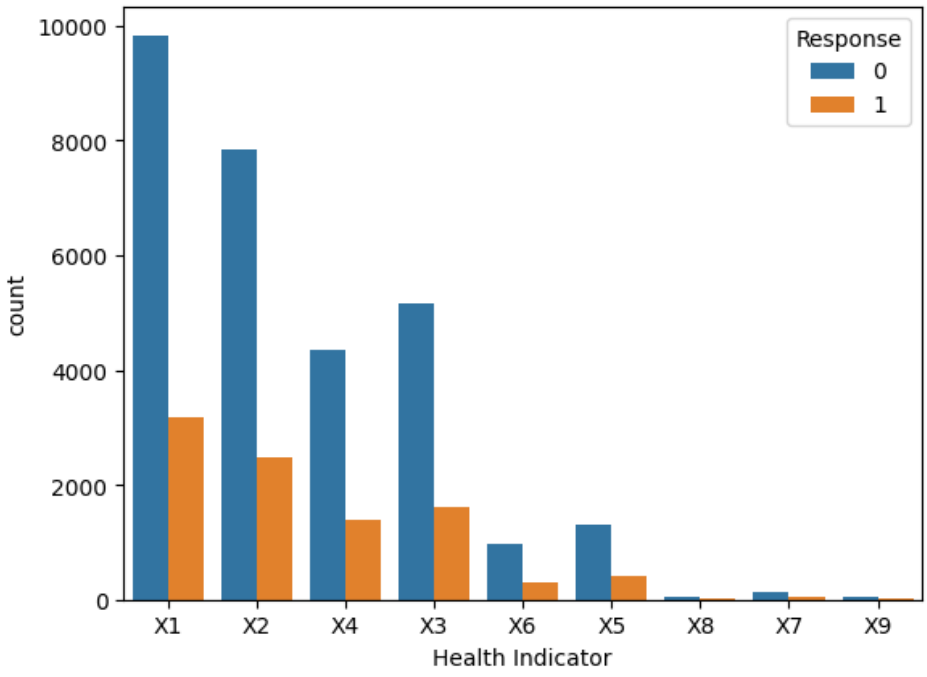


People who own a house are more likely to get insurance as compared to people who rent. Also people tend to get individual insurance over joint insurance.

In [11]:

```
sns.countplot(df, x='Health Indicator', hue='Response')
```

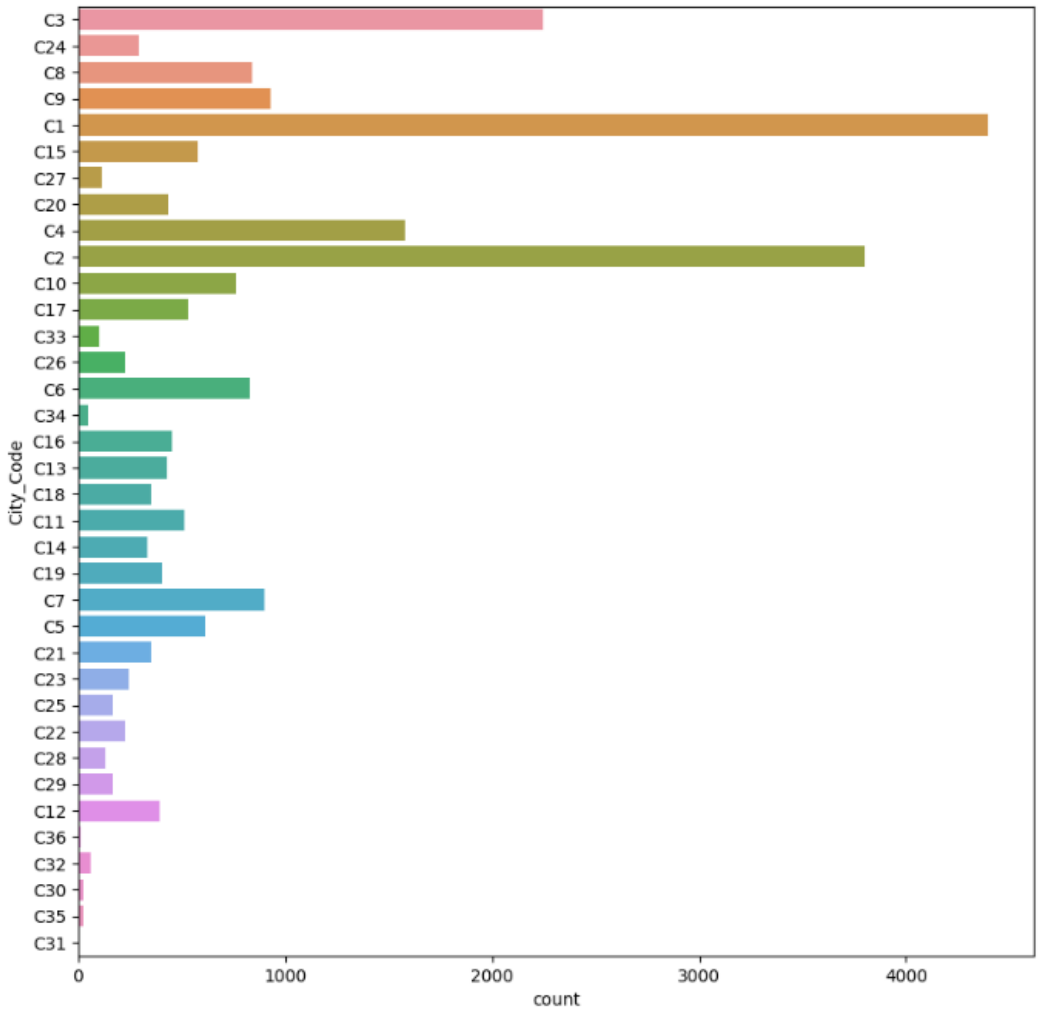
```
<AxesSubplot: xlabel='Health Indicator', ylabel='count'>
```



People with different health conditions might be eligible for taking the insurance or not. Generally, insurance is not given to anyone as they might not be able to fulfill the conditions.

```
In [25]:
plt.figure(figsize=(10, 10))
sns.countplot(df, y='City_Code')

plt.show()
```



Cities C1, C2, C3, C4, C7 and C9 have the most number of insurance holders. Whereas, cities C30, C28, C29 and C31 have the least number of insurance holders.

Conclusion-

Thus we have successfully performed exploratory data analysis and data visualization using python.