**Experiment No.11**

**Aim:** To build a Business Intelligence Mini Project.

**Insurance Lead Prediction(dataset):**

For the data and objective, it is evident that this is a Binary Classification Problem data in the Tabular Data format. A policy is recommended to a person when they land on an insurance website, and if the person chooses to fill up a form to apply, it is considered a Positive outcome (Classified as lead). All other conditions are considered Zero outcomes.

It contains attributes namely ID, City\_Code, Region\_Code, Accomodation\_Type, Reco\_Insurance\_Type, Upper\_Age, Lower\_Age, Is\_Spouse, Health Indicator, Holding\_Policy\_Duration, Holding\_Policy\_Type, Reco\_policy\_Cat, Reco\_policy\_Premium, Response.

Accomodation\_Type, Is\_Spouse, Reco\_Insurance\_Type are categorical attributes while other attributes are numerical attributes. The target attribute that is Response is a binary attribute i.e. 0 / 1.

**Dataset link:** [Insurance Lead Prediction Dataset](https://www.kaggle.com/datasets/owaiskhan9654/health-insurance-lead-prediction-raw-data)

**Problem Statement:**

Insurance companies rely heavily on acquiring new customers to increase their revenue and grow their business. However, identifying potential customers who are likely to purchase insurance policies can be a challenging task. It requires a deep understanding of customer behavior, preferences, and demographics.

In order to address this challenge, the problem statement is to develop a predictive model that can identify potential customers who are likely to purchase insurance policies. The predictive model will help the insurance company to improve their sales performance by targeting leads that are most likely to convert. This will increase the efficiency of the sales process and ultimately lead to increased revenue.

The predictive model will leverage various data sources, such as demographic information, economical condition, and customer behavior, to make accurate predictions about which leads have a higher probability of converting. By identifying the best leads, the insurance company can focus their resources on those leads to improve conversion rates and overall sales performance.

The development of an effective predictive model is crucial for insurance companies to remain competitive in the market. By leveraging business intelligence tools and techniques, the insurance company can gain insights into customer behavior, preferences, and demographics, and develop targeted marketing strategies to increase their customer base and revenue.

**Data Mining Task:**

The data mining task required for the insurance lead prediction project would be a supervised classification task. This involves building a model that can predict the likelihood of a lead converting to a customer based on various attributes such as demographic information, previous purchase history, and customer behavior. The model would be trained on a dataset consisting of labeled instances, where the label indicates whether a lead converted or not.

The goal of the supervised classification task in this project is to develop a predictive model that can accurately identify potential customers who are likely to purchase insurance policies. The model would take into account various factors such as age, gender, income, occupation, and output the probability of a lead converting to a customer.

The performance of the model would be evaluated using metrics such as accuracy, precision, recall, and F1 score. The model with the highest performance would be selected and used to develop a data-driven strategy for the insurance company to target the best leads and improve their sales performance.

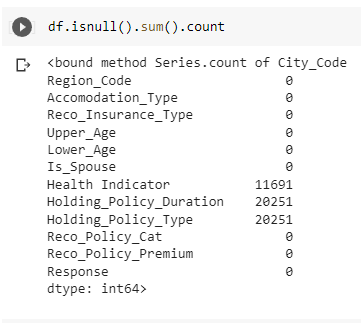
In summary, the data mining task required for the insurance lead prediction project is a supervised classification task, where the goal is to predict the likelihood of a lead converting to a customer based on various attributes. The development of an accurate predictive model can help insurance companies to improve their sales performance by targeting the best leads and ultimately increasing their revenue.

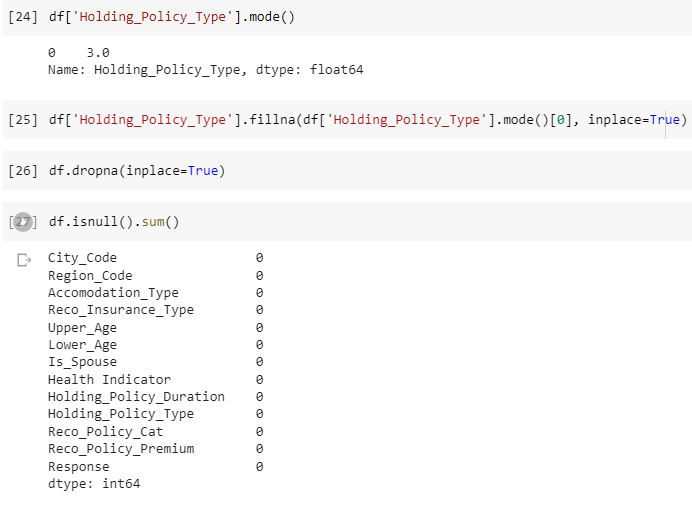
**Preprocessing -**

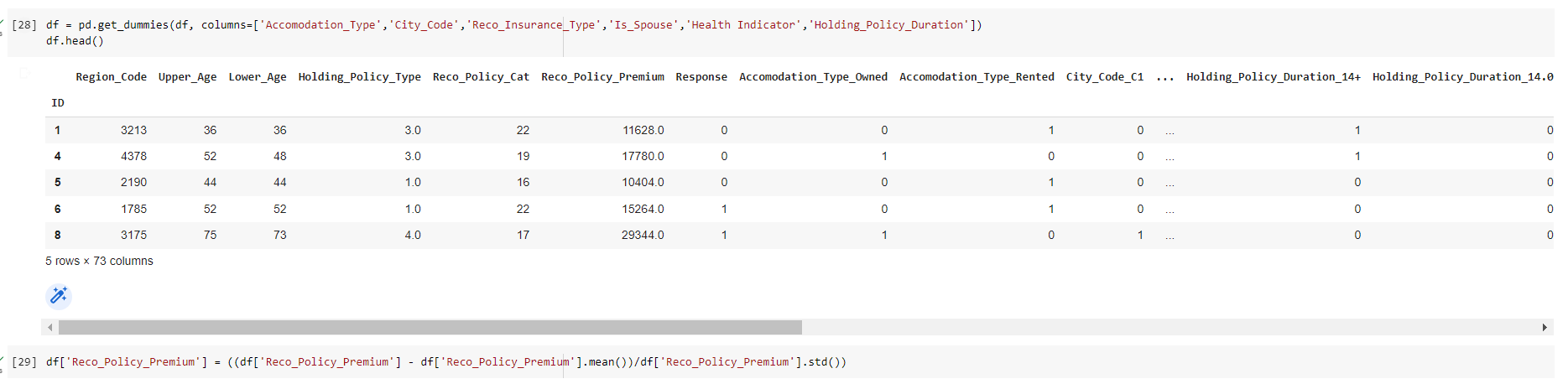
For this problem, we can use the [Insurance Lead Prediction Dataset](https://www.kaggle.com/datasets/owaiskhan9654/health-insurance-lead-prediction-raw-data) available on Kaggle. This dataset includes information about customers with its details along with its response. The dataset contains 50882 records and 14 attributes. The detailed information about the dataset is as follows -

| **No.** | **Column Name** | **Data Type** |
| --- | --- | --- |
| 1 | ID | int64 |
| 2 | City\_Code | object |
| 3 | Region\_Code | int64 |
| 4 | Accomodation\_Type | object |
| 5 | Reco\_Insurance\_Type | object |
| 6 | Upper\_Age | int64 |
| 7 | Lower\_Age | int64 |
| 8 | Is\_Spouse | object |
| 9 | Health Indicator | object |
| 10 | Holding\_Policy\_Duration | object |
| 11 | Holding\_Policy\_Type | float64 |
| 12 | Reco\_Policy\_Cat | int64 |
| 13 | Reco\_Policy\_Premium | float64 |
| 14 | Response | int64 |

****

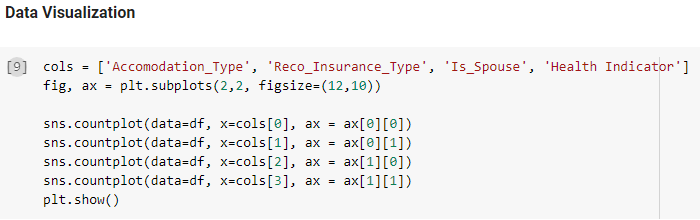
****

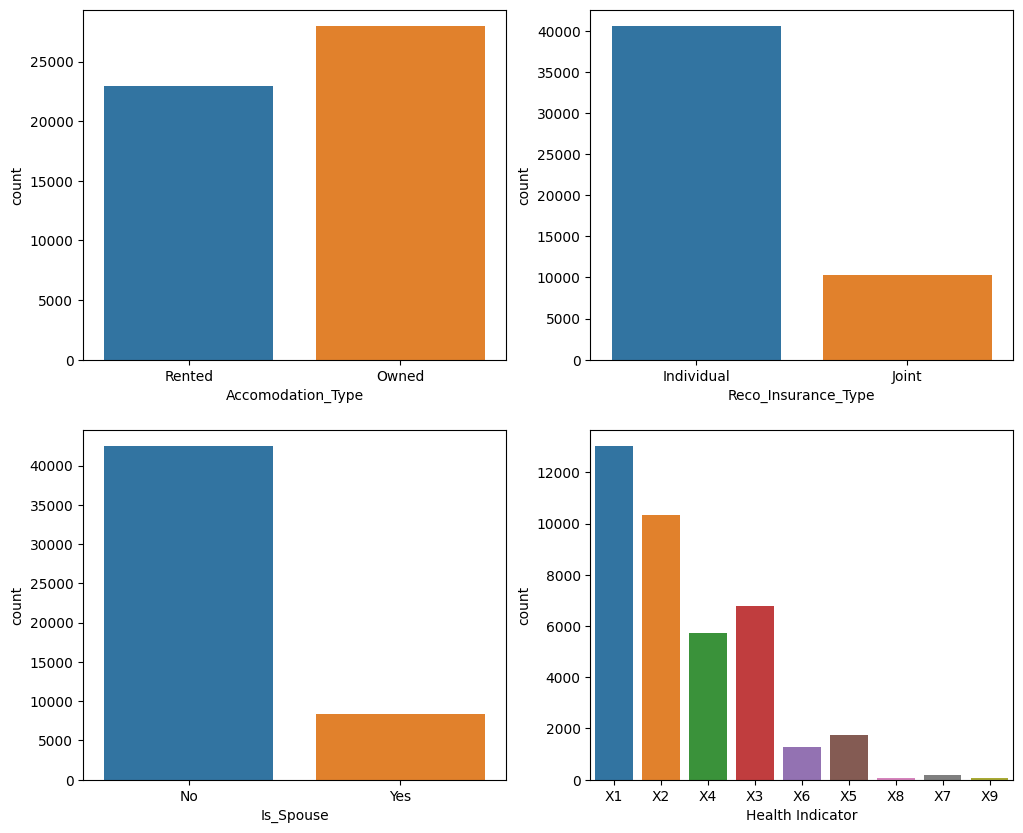
****

****

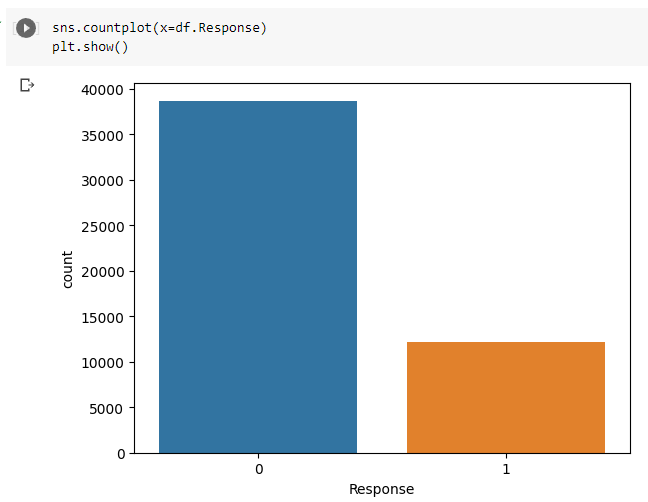
**EDA -**

These four are count plots each showing the number of entries that have the same values. Accommodation\_Type is of two types either Rented or Owned, Reco\_Insurance\_Type is also categorized into two types namely Individual and Joint. Is\_Spuse has two options i.e. Yes / No. While the health indicator is categorized into 9 options ranging from X1 - X9.

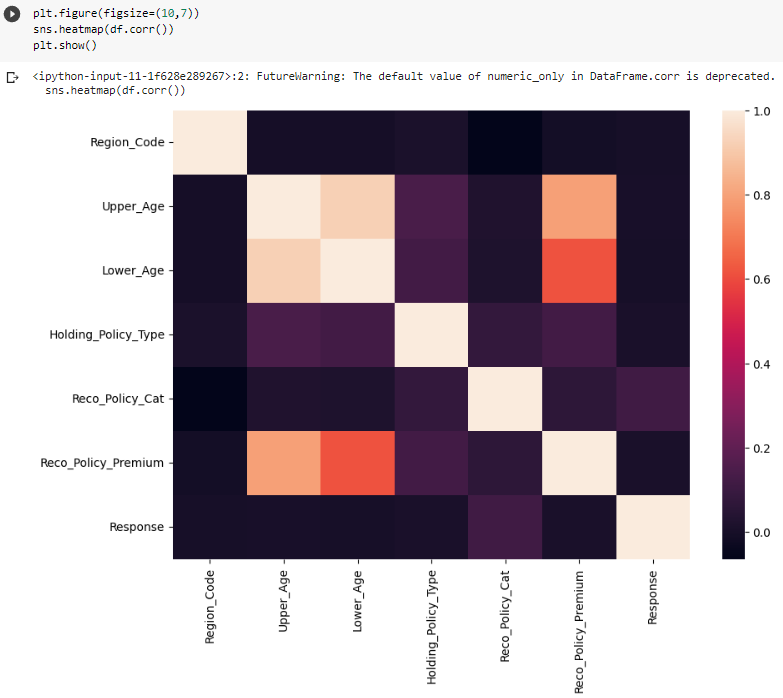




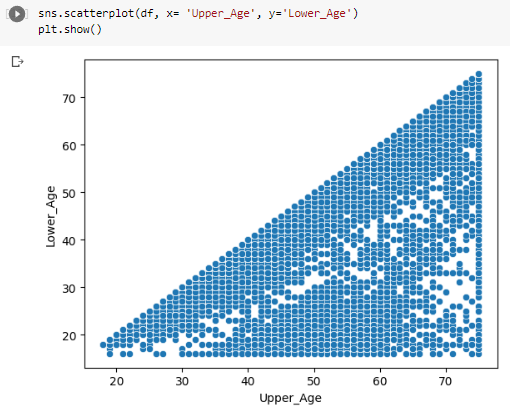
This count plot shows the count of the target attribute i.e. the response attribute in the Health\_Insurance dataset.



This is the heat map showing the amount of correlation in each attribute.

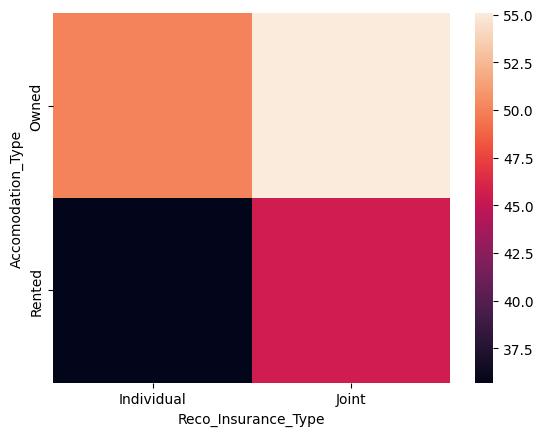


This is a scatterplot between the attributes namely the Upper\_Age and Lower\_Age.

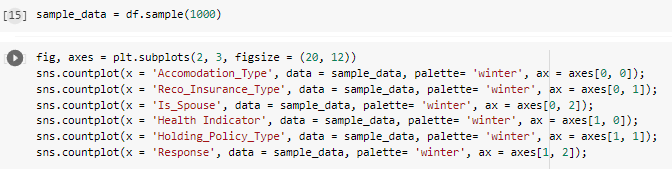


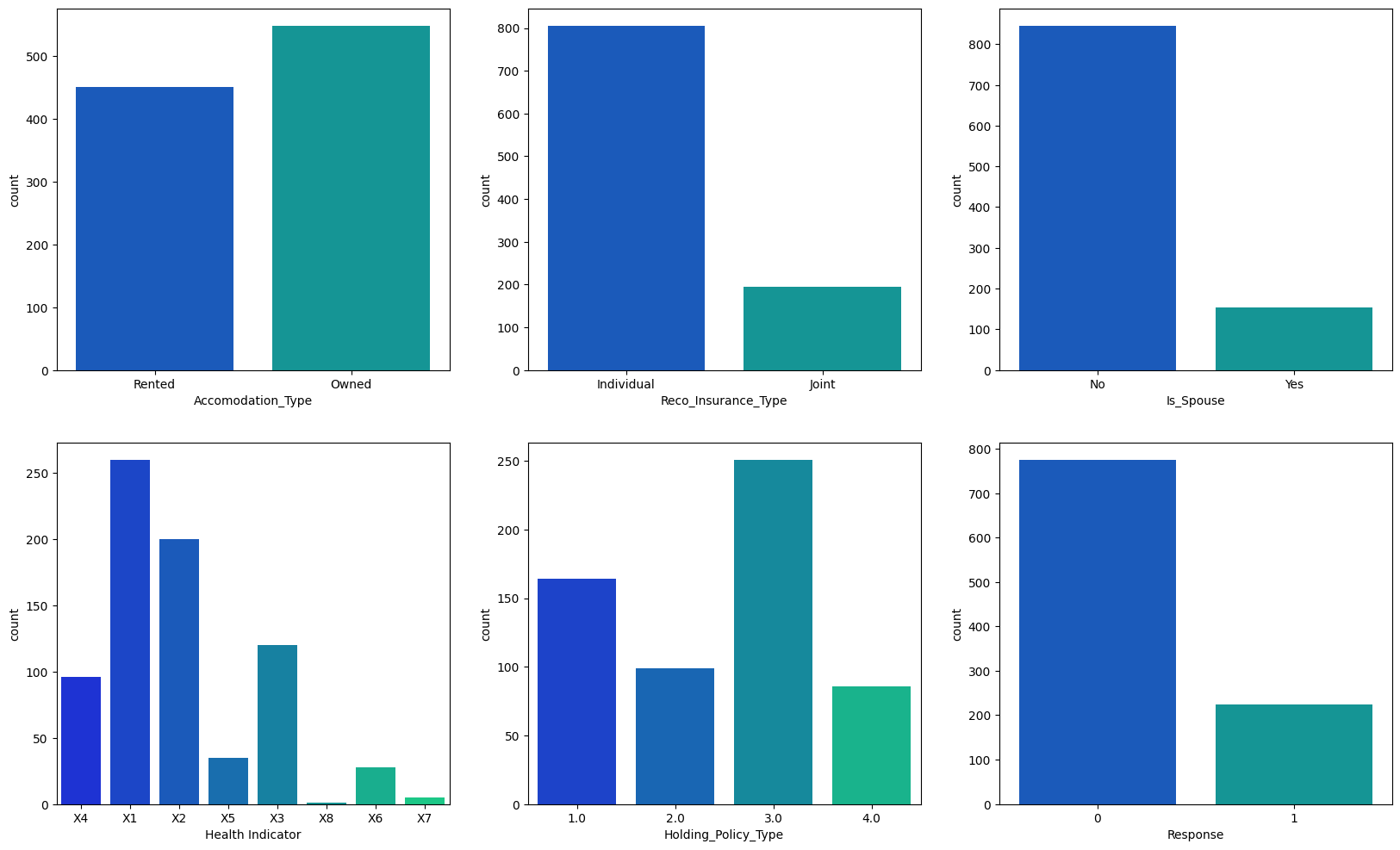
This is a pivot table plotted between three attributes that are Accomdation\_Type as index, Reco\_Insurance\_Type as columns and Upper\_Age as values.



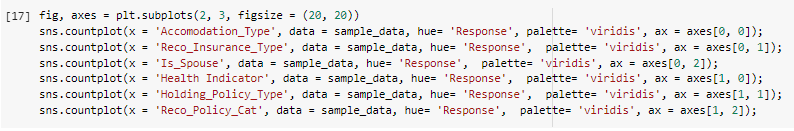


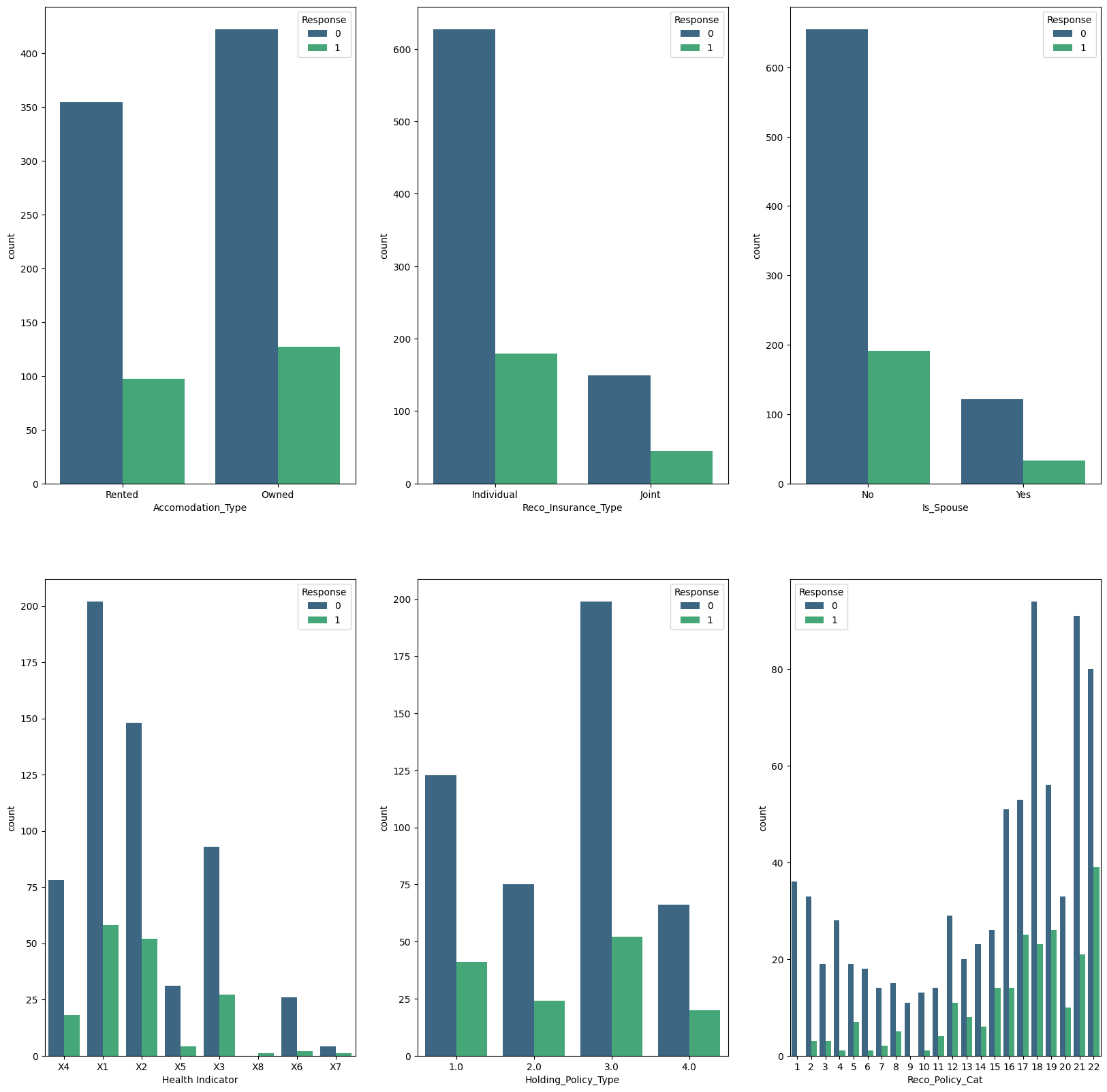
These are the count plots made using the sample data from the dataset. The size of the sample data is 1000.





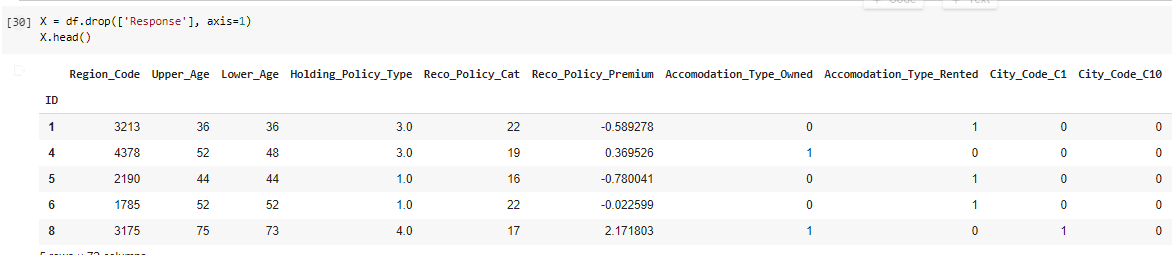
These are also the countplot but by using hue as the response attribute.

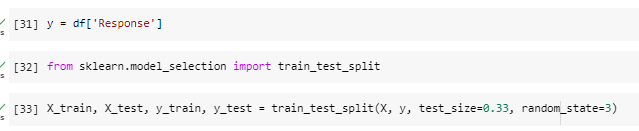




**Data Mining**

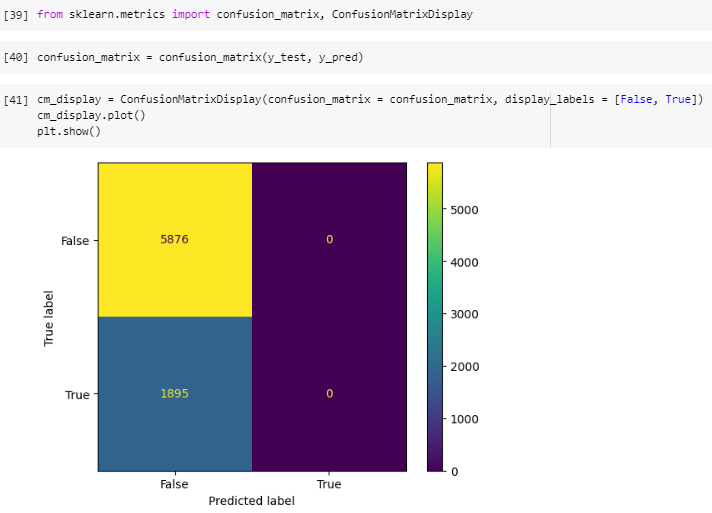
**Model Training**

****

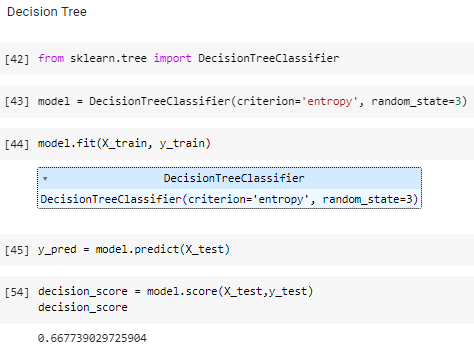
****

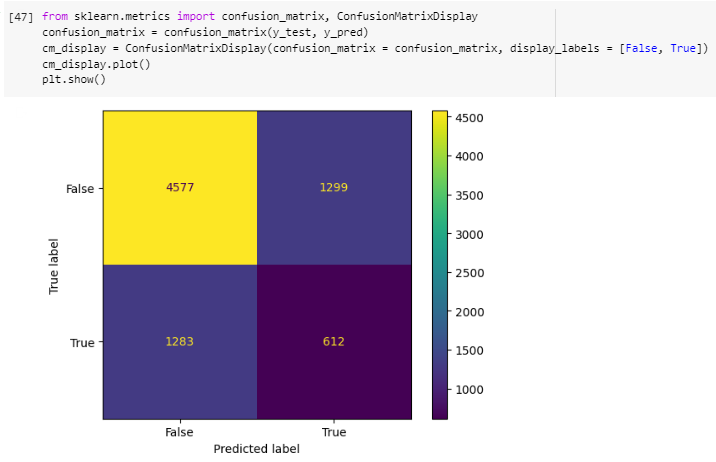
**Logistic regression**

****

****

**Decision Tree**

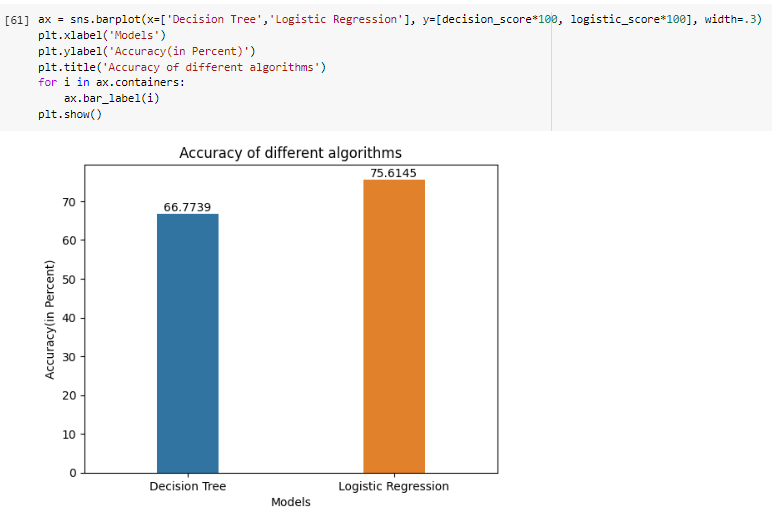
****

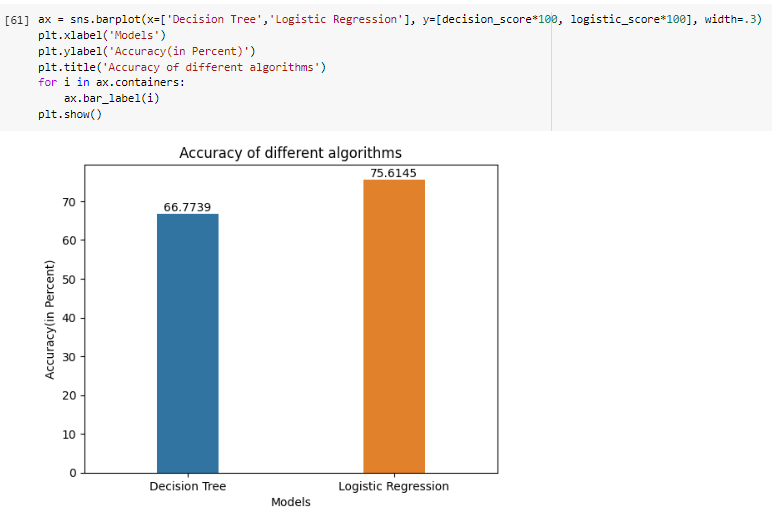
****

**Results**

For determining which customers will take the insurance policy, we can use different classification algorithms. For this problem we have used Decision Tree, and Logistic Regression.

According to the results, the accuracy of the Logistic Regression algorithm is more i.e. 75.6%. Therefore, for such type of dataset we will make use of Logistic Regression model to predict the outcome.

****

****

From the above comparison, we can see that Logistic Regression works best with the accuracy of almost 75.61% for the above dataset.

**Link to Python Notebook -** [**Python Notebook**](https://colab.research.google.com/drive/1Tms3nvW0vojH4-k6RyHmTfrrBkpogFgJ#scrollTo=63_QNtGwjmuw)

**Business Implications -**

The business implications of our project are as follows -

**Targeted marketing:** To identify potential customers based on demographics such as age, location, and type of accommodation. This information can be used to target marketing efforts towards customers who are most likely to be interested in insurance products.

**Customer retention:** Insurance companies can use this project to identify customers who may be at risk of churning and take proactive measures to retain them.

**Product development:** Based on the type of insurance products that customers are interested in and results generated by the project. Insurance companies can use this information to develop new products that are tailored to customer needs and preferences.

**Risk assessment:** Health indicators and holding policy duration can be used to assess the risk of potential customers and adjust insurance premiums accordingly.

**Operational efficiency:** Based on holding policy type, Insurance companies can use this information to optimize their underwriting process and improve operational efficiency.

**Conclusion -**

In conclusion, the insurance lead prediction model is a useful business intelligence project for insurance companies to increase their customer base and revenue. By analyzing customer data and predicting the likelihood of a customer purchasing an insurance policy, companies can focus their marketing efforts on high-potential customers and improve their conversion rates.

However, to ensure the accuracy and reliability of the model, it is important to continuously monitor and update it with new data. Additionally, it is important to maintain data privacy and security, as the model relies on sensitive customer information.