Capstone Project 3

Supervised Machine Learning(Classification): Health Insurance Cross sell Prediction

Akifuddin Kashif | Zeeshan Ahmed

All about this presentation:

1. Defining problem statement.
2. Overview of data.
3. Performing exploratory data analysis.
4. Model preparation.
5. Building different models.
6. Evaluation of all models.
7. Extracting the best model.

Problem statement

- We are tasked with predicting whether a customer who previously purchased a company's Health Insurance will opt for Vehicle Insurance or not. This is a supervised machine learning classification problem with many independent variables and dependent variable namely Response which comprises responses of customers regarding vehicle insurance. Our model helps understand the behaviors of customers and build a model around that behavior.

Overview of given data

1. ID : Unique ID for the customer
2. Gender : Gender of the customer
3. Age : Age of the customer
4.Driving_License 0 : Customer does not have DL, 1 : Customer already has DL
5.Region_Code : Unique code for the region of the customer
6.Previously_Insured 1 : Customer already has Vehicle Insurance, 0 : Customer doesn't have Vehicle Insurance
7. Vehicle_Age : Age of the Vehicle
8.Vehicle_Damage 1 : Customer got his/her vehicle damaged in the past. 0 : Customer didn't get his/her vehicle damaged in the past.
9.Annual_Premium : The amount customer needs to pay as premium in the year
10.PolicySalesChannel : Anonymized Code for the channel of outreaching to the customer ie. Different Agents, Over Mail, Over Phone, In Person, etc.
11.Vintage : Number of Days, Customer has been associated with the company
12.Response : 1 : Customer is interested, 0 : Customer is not interested

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 381109 entries, 0 to 381108
Data columns (total 12 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   id                    381109 non-null  int64
 1   Gender                381109 non-null  object
 2   Age                   381109 non-null  int64
 3   Driving_License       381109 non-null  int64
 4   Region_Code           381109 non-null  float64
 5   Previously_Insured    381109 non-null  int64
 6   Vehicle_Age           381109 non-null  object
 7   Vehicle_Damage        381109 non-null  object
 8   Annual_Premium        381109 non-null  float64
 9   Policy_Sales_Channel  381109 non-null  float64
 10  Vintage               381109 non-null  int64
 11  Response              381109 non-null  int64
dtypes: float64(3), int64(6), object(3)
memory usage: 34.9+ MB
```

Description of data

	id	Age	Driving_License	Region_Code	Previously_Insured	Annual_Premium	Policy_Sales_Channel	Vintage	Response
count	381109.000000	381109.000000	381109.000000	381109.000000	381109.000000	381109.000000	381109.000000	381109.000000	381109.000000
mean	190555.000000	38.822584	0.997869	26.388807	0.458210	30564.389581	112.034295	154.347397	0.122563
std	110016.836208	15.511611	0.046110	13.229888	0.498251	17213.155057	54.203995	83.671304	0.327936
min	1.000000	20.000000	0.000000	0.000000	0.000000	2630.000000	1.000000	10.000000	0.000000
25%	95278.000000	25.000000	1.000000	15.000000	0.000000	24405.000000	29.000000	82.000000	0.000000
50%	190555.000000	36.000000	1.000000	28.000000	0.000000	31669.000000	133.000000	154.000000	0.000000
75%	285832.000000	49.000000	1.000000	35.000000	1.000000	39400.000000	152.000000	227.000000	0.000000
max	381109.000000	85.000000	1.000000	52.000000	1.000000	540165.000000	163.000000	299.000000	1.000000

Sample of data

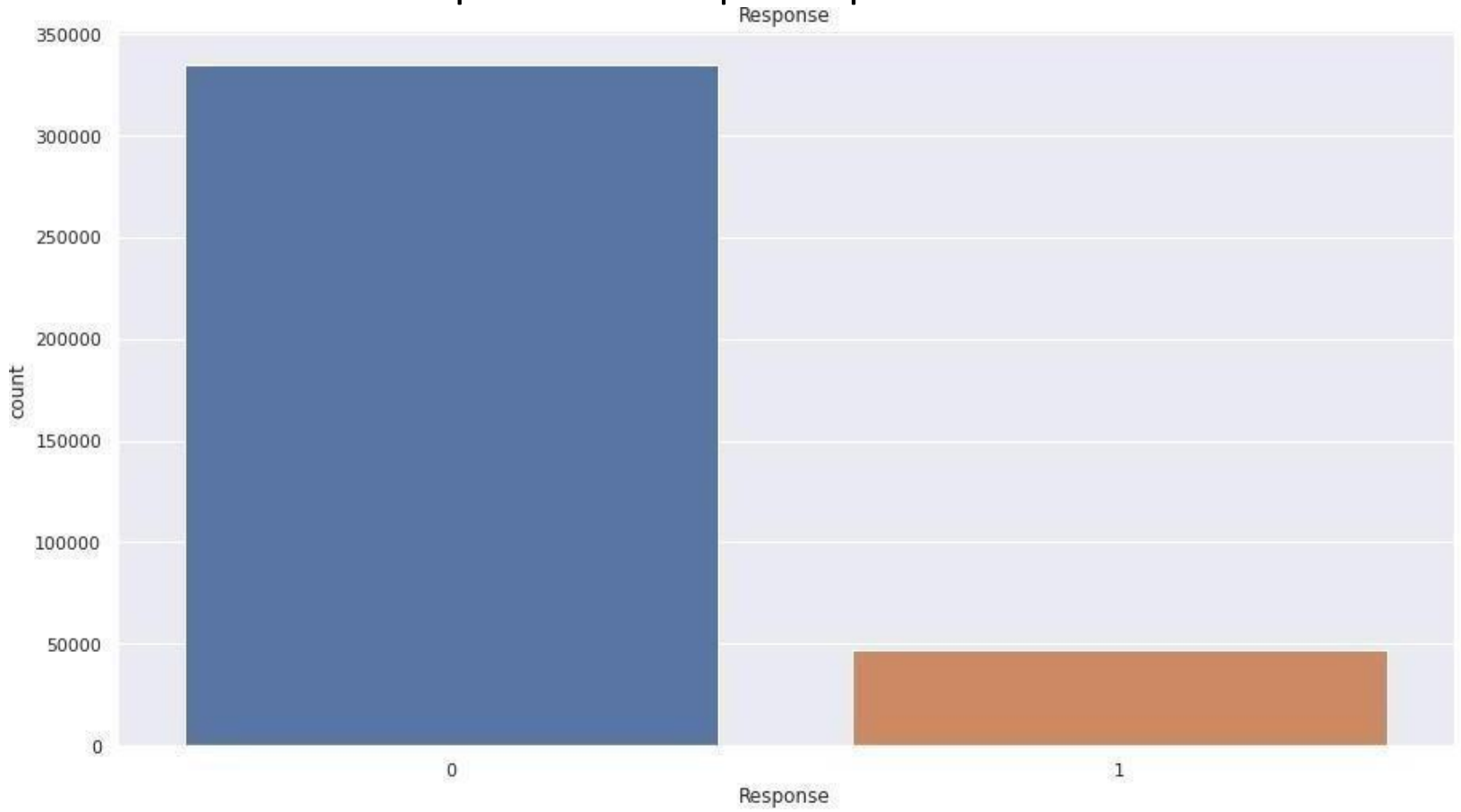
	id	Gender	Age	Driving_License	Region_Code	Previously_Insured	Vehicle_Age	Vehicle_Damage	Annual_Premium	Policy_Sales_Channel	Vintage	Response
0	1	Male	44	1	28.0	0	> 2 Years	Yes	40454.0	26.0	217	1
1	2	Male	76	1	3.0	0	1-2 Year	No	33536.0	26.0	183	0
2	3	Male	47	1	28.0	0	> 2 Years	Yes	38294.0	26.0	27	1
3	4	Male	21	1	11.0	1	< 1 Year	No	28619.0	152.0	203	0
4	5	Female	29	1	41.0	1	< 1 Year	No	27496.0	152.0	39	0

Exploratory data analysis

- Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns,to spot anomalies,to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.
- EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task.

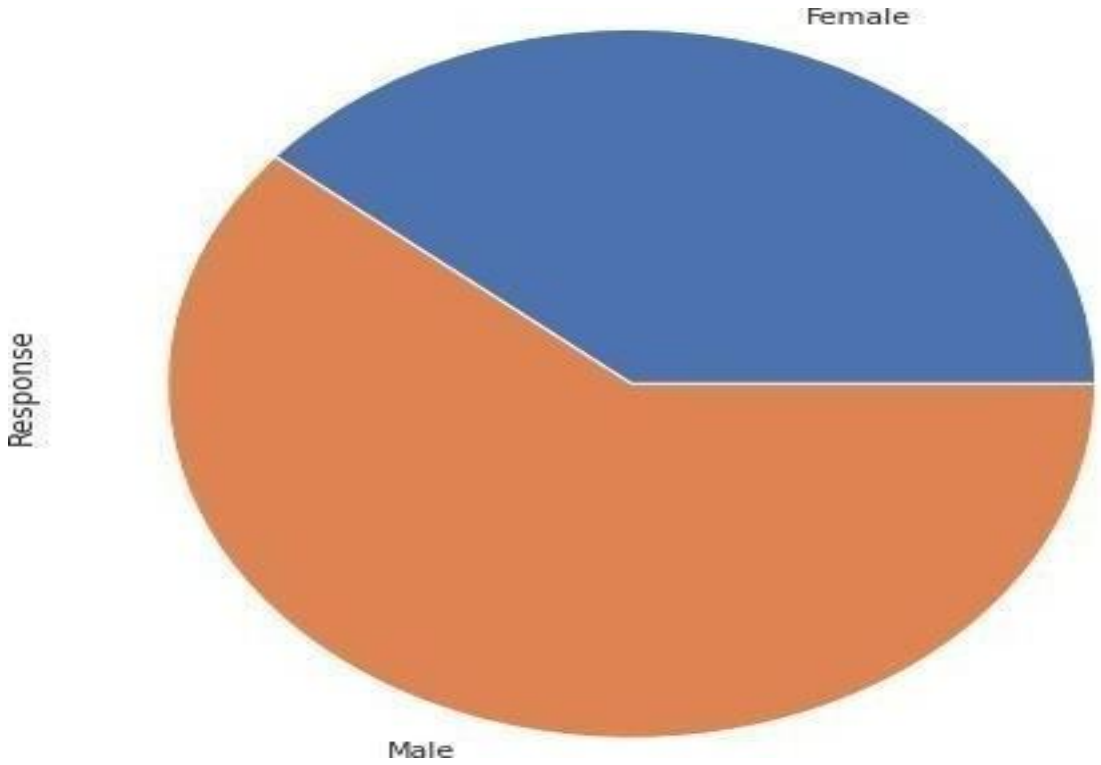
Distribution of Responses by Customers

No Response : 0 | Response: 1



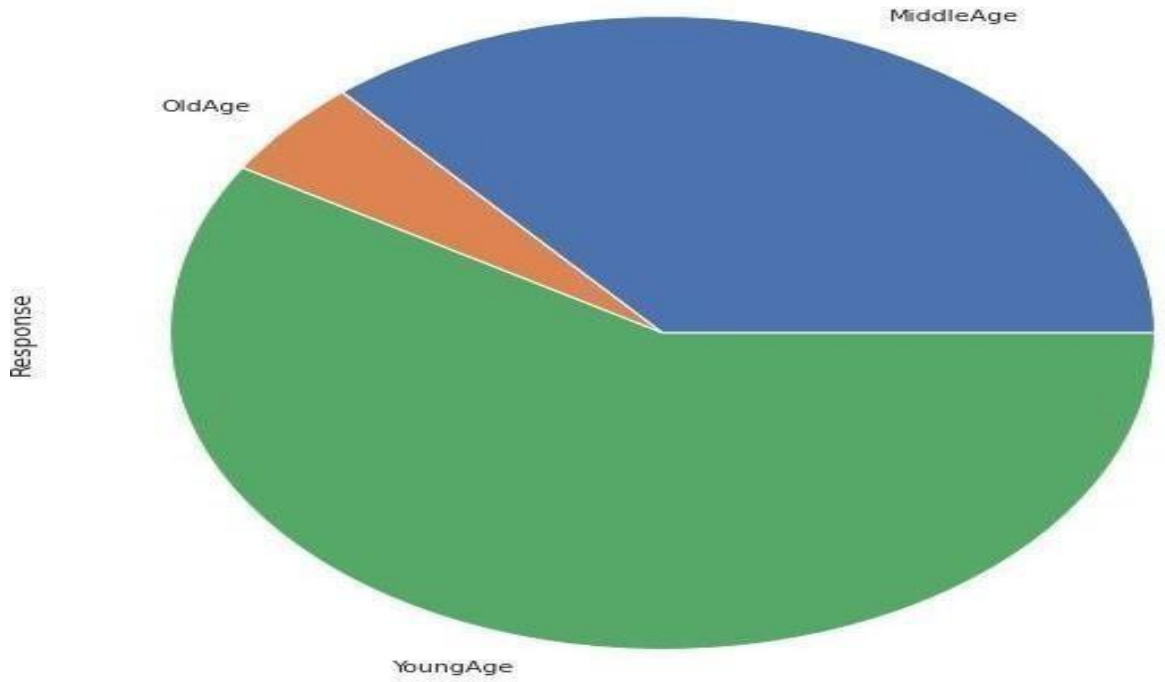
There are more number of people who don't prefer to purchase vehicle insurance

Distribution of number of responses from Gender



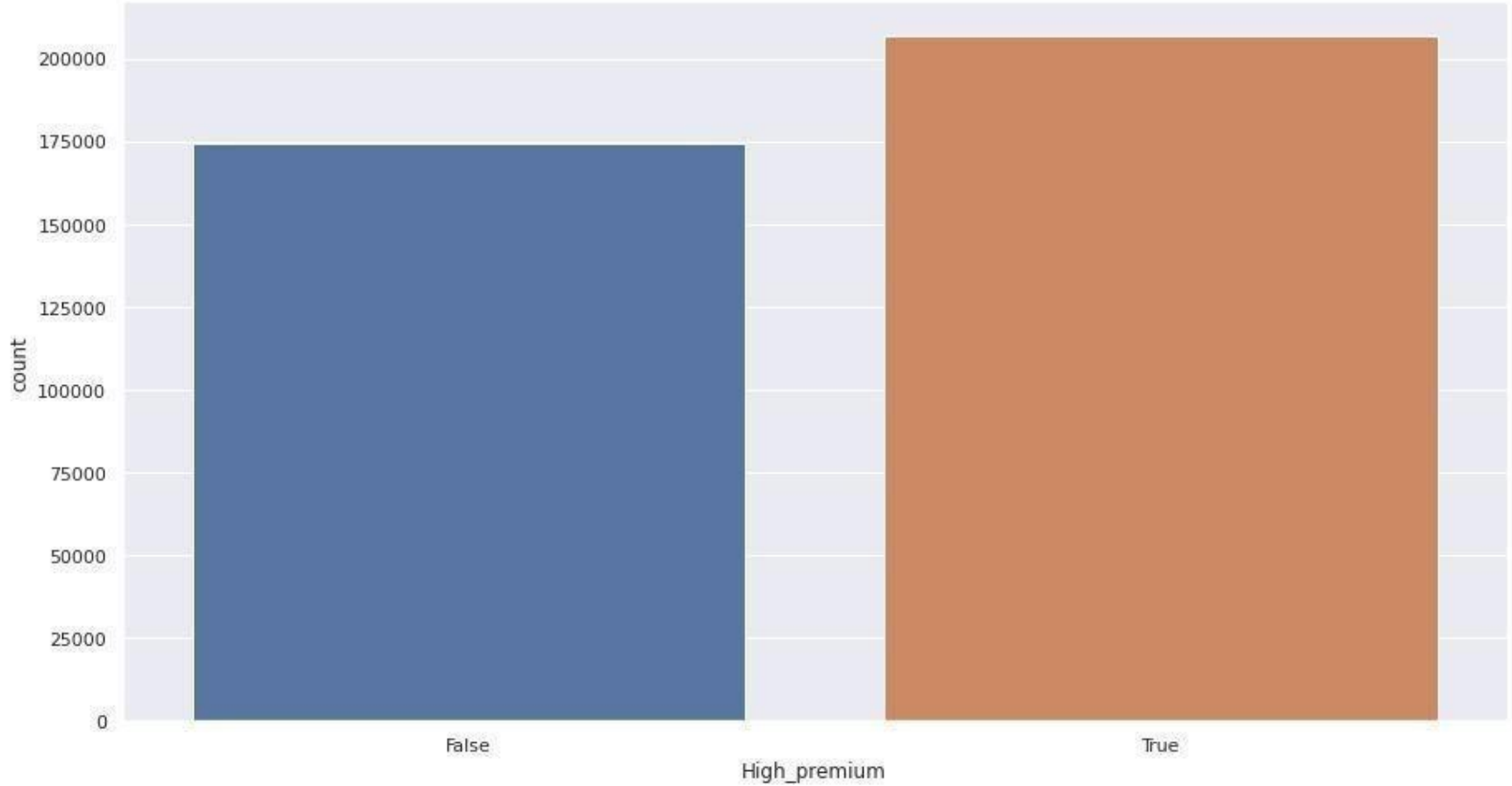
From the Pie charts we can conclude Males were more interested in subscribing to Vehicle Insurance

Distribution of customers by Age Group



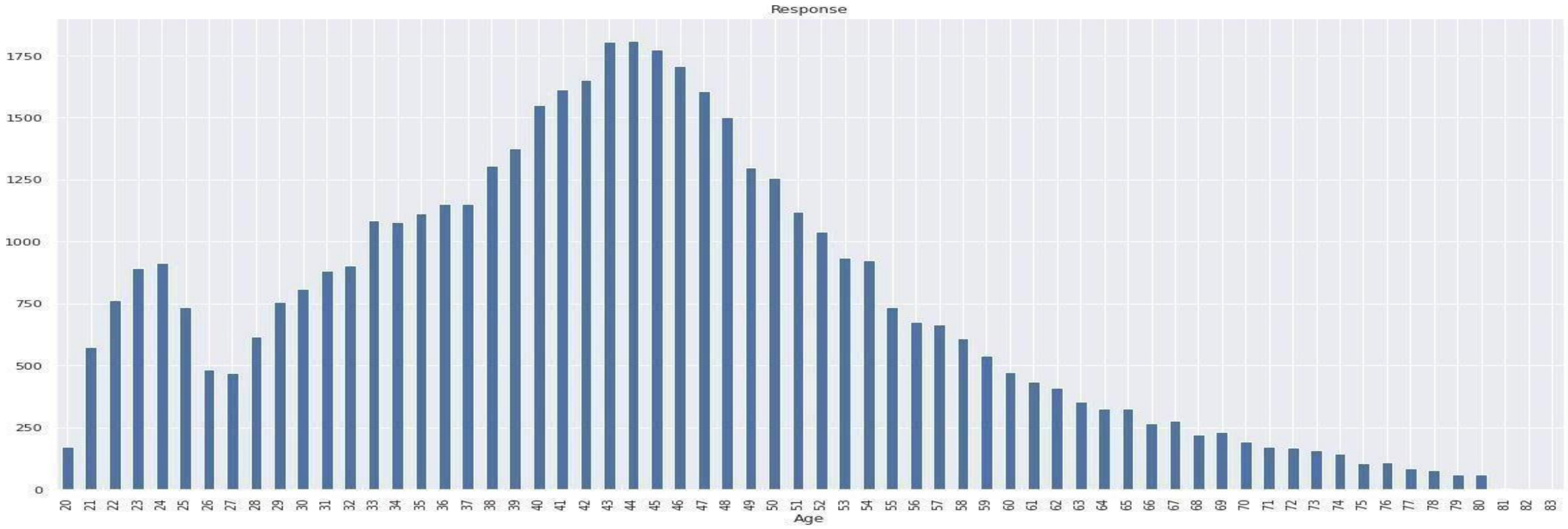
This age group plot has customers who are prospects for Vehicle Insurance. From the plot it is evident that there is a remarkable customer activity regarding Vehicle Insurance from Young Age group as they are more conscious about Vehicle Insurance because of their reckless driving

Distribution of Type of subscriptions Customers prefer



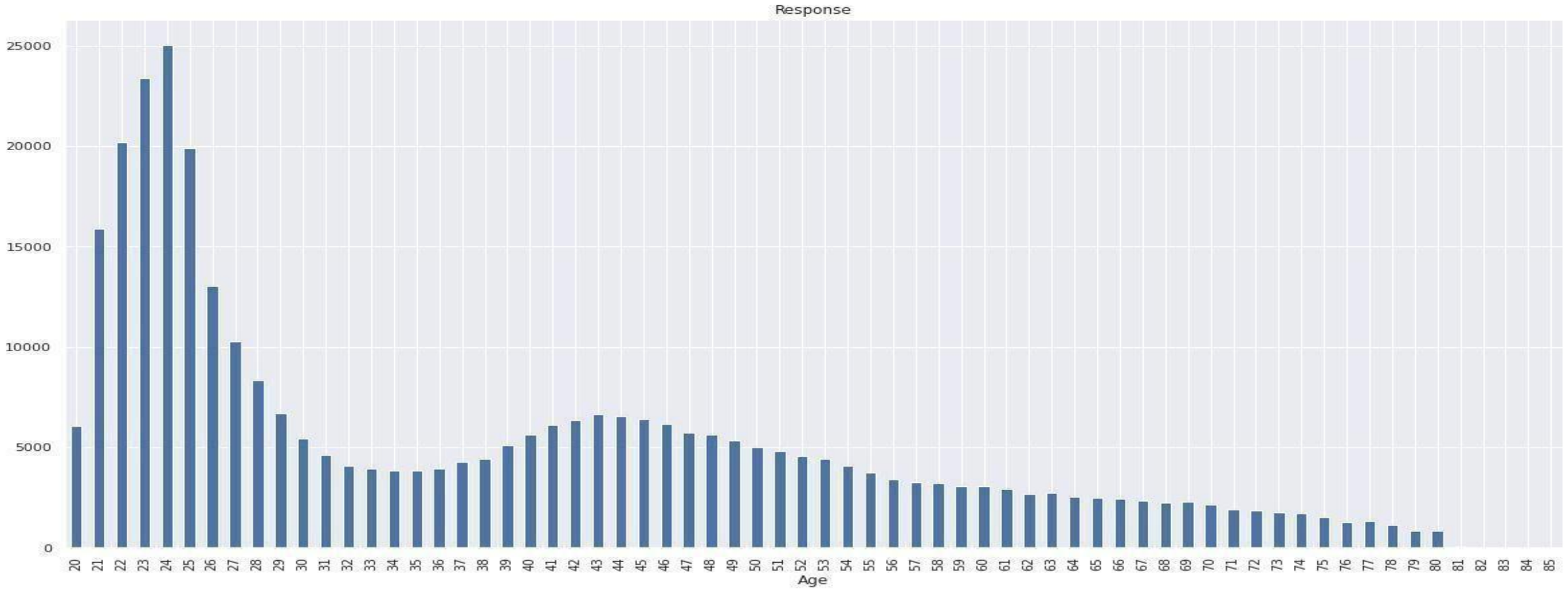
From the above plot it is evident that people prefer high end premiums than compared low end.

Customers of different age groups who opted for Vehicle Insurance



From the plot it is evident that there is a remarkable response regarding Vehicle Insurance from the Age groups 40 to 48

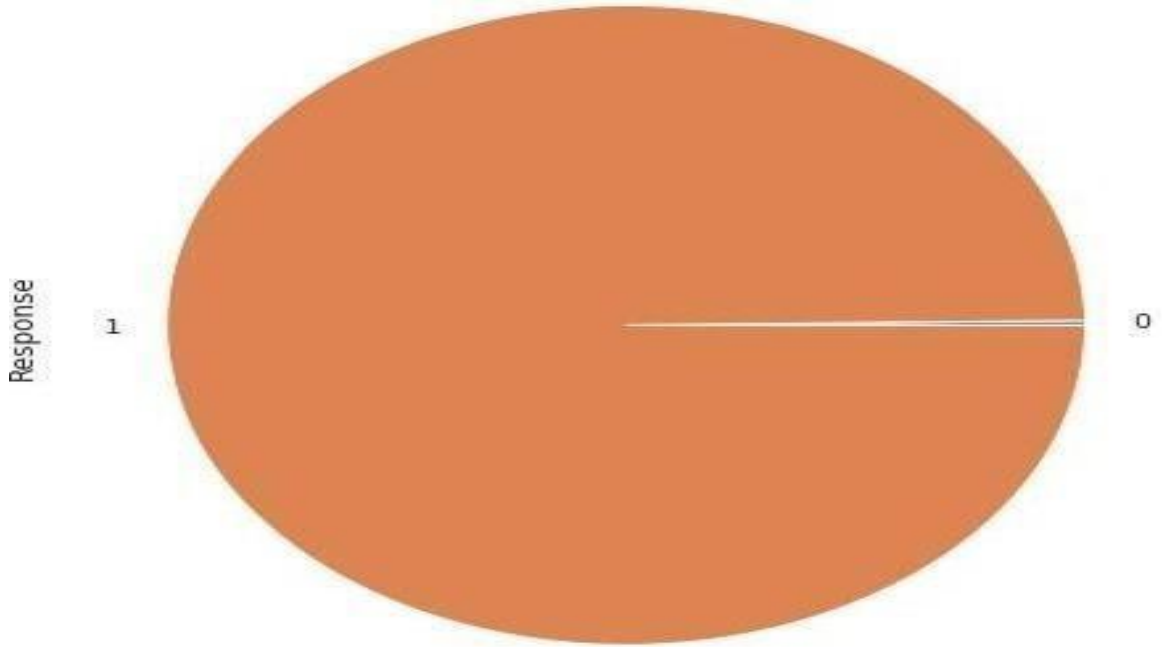
Customers of different age groups who did not opt for Vehicle Insurance



From the plot it is evident that there is a negative responses regarding Vehicle Insurance from the Age groups 21 to 25

Distribution of Responses with respect to Driving License

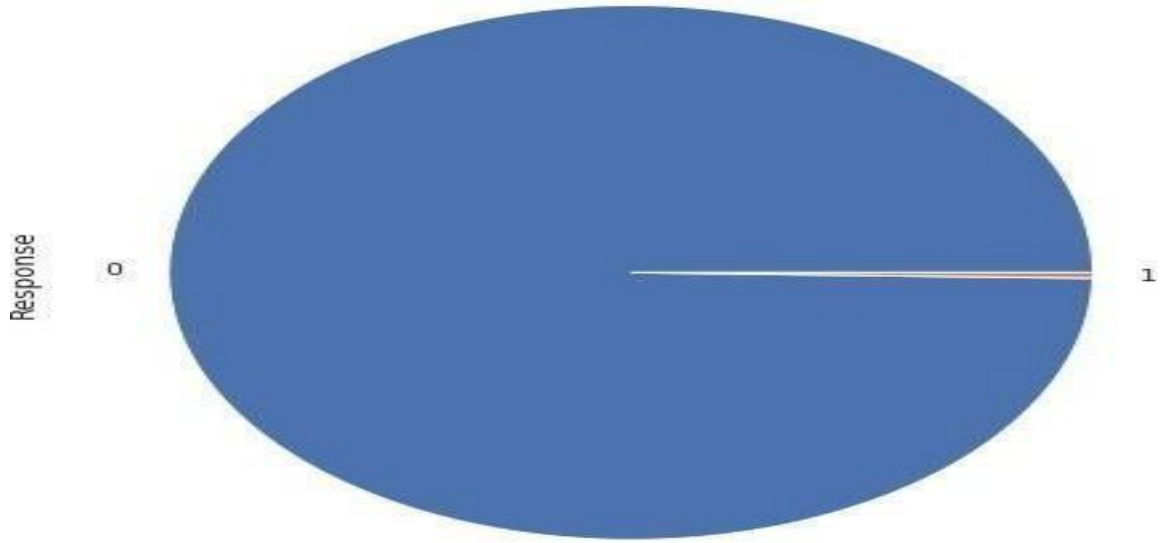
No Driving License : 0 | Driving License: 1



People who posses driving license are more likely to get their vehicle insured than that of the people with no driving license

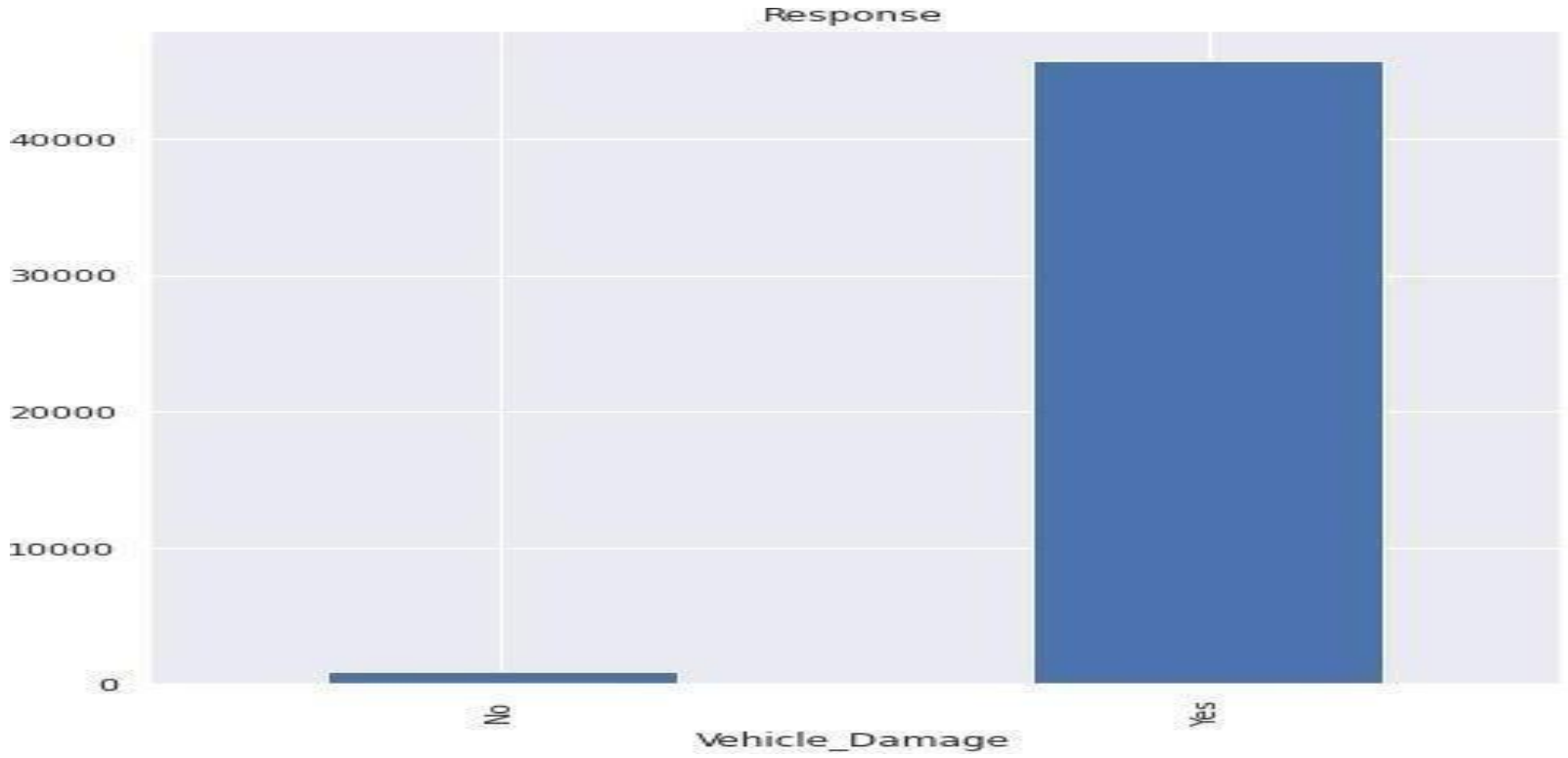
Distribution of Responses with respect to Customers who were Previously Insured

Not Previously Insured : 0 | Previously Insured: 1



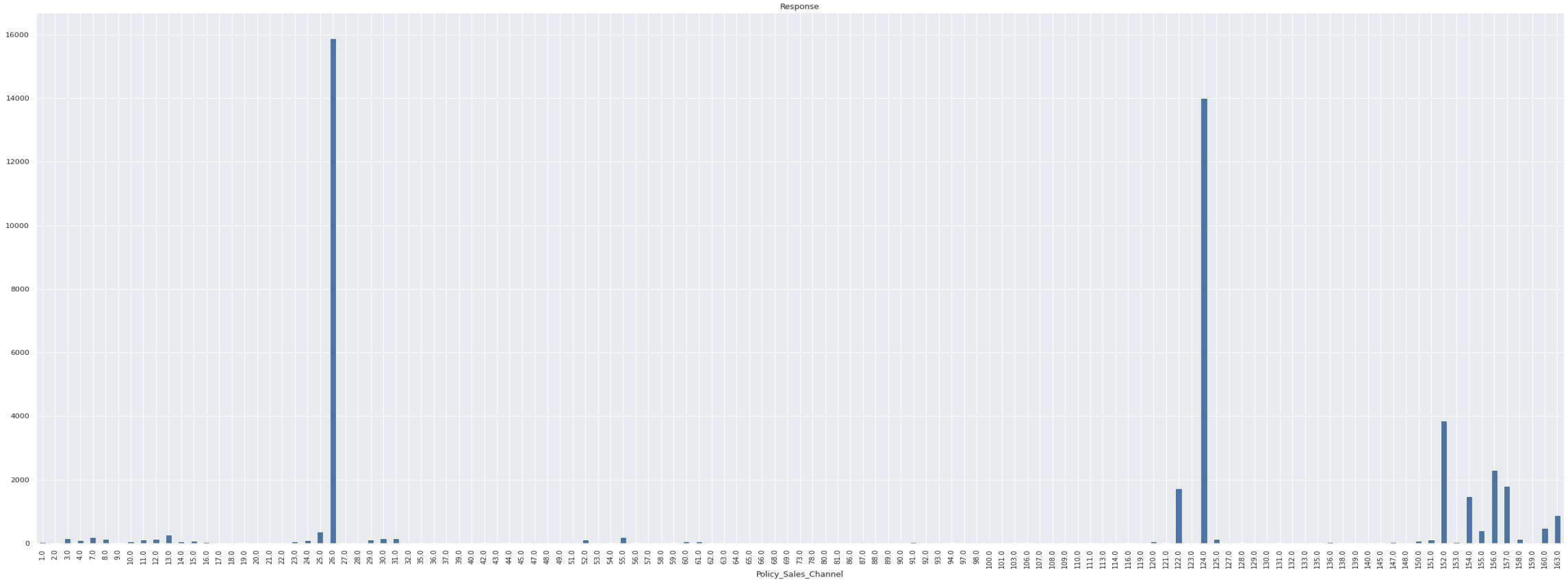
This Pie Chart is about the distribution of number of responses by people who are previously insured and not insured. It suggests that people who are not previously insured almost always opt in for vehicle Insurance.

Distribution of responses of Customers who possess a damaged vehicle



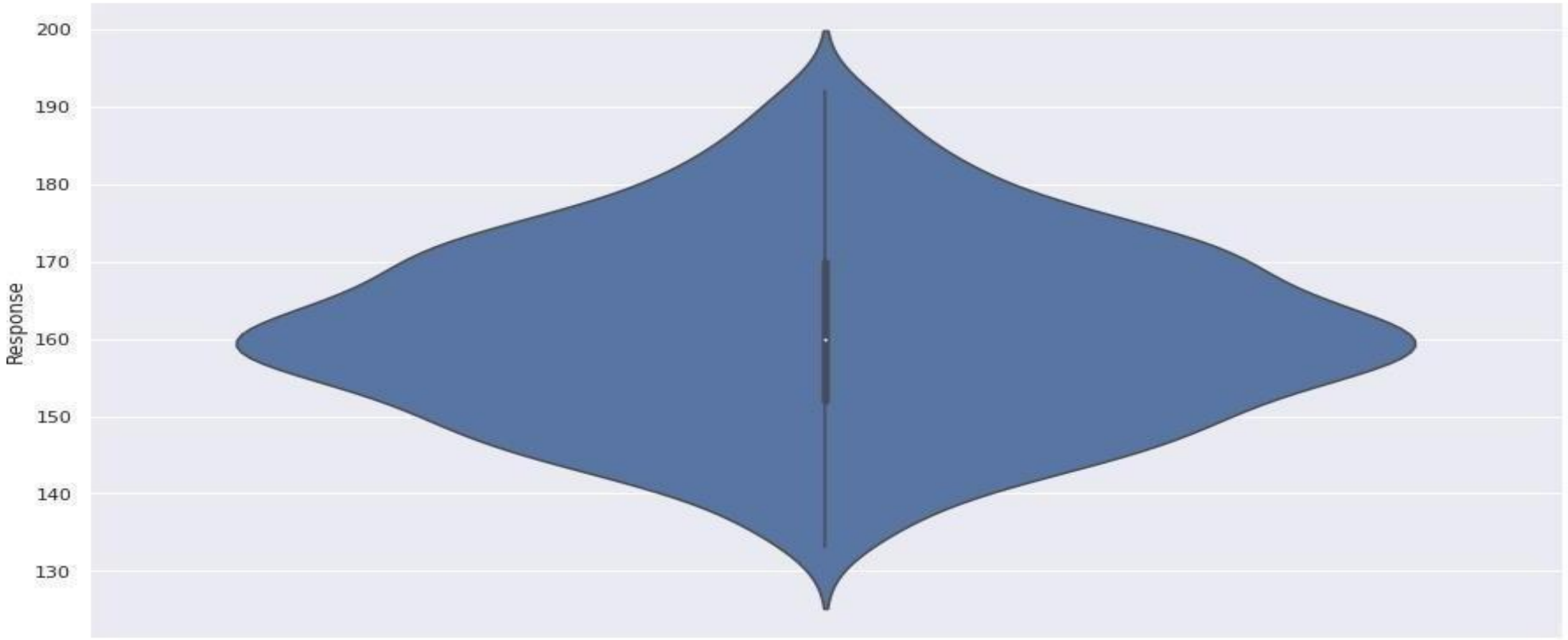
From this plot, we can say that the people who have a damaged vehicle are more inclined to get a vehicle insurance. Hence we reject the null hypothesis here.

Distribution of Responses from different Policy Channels



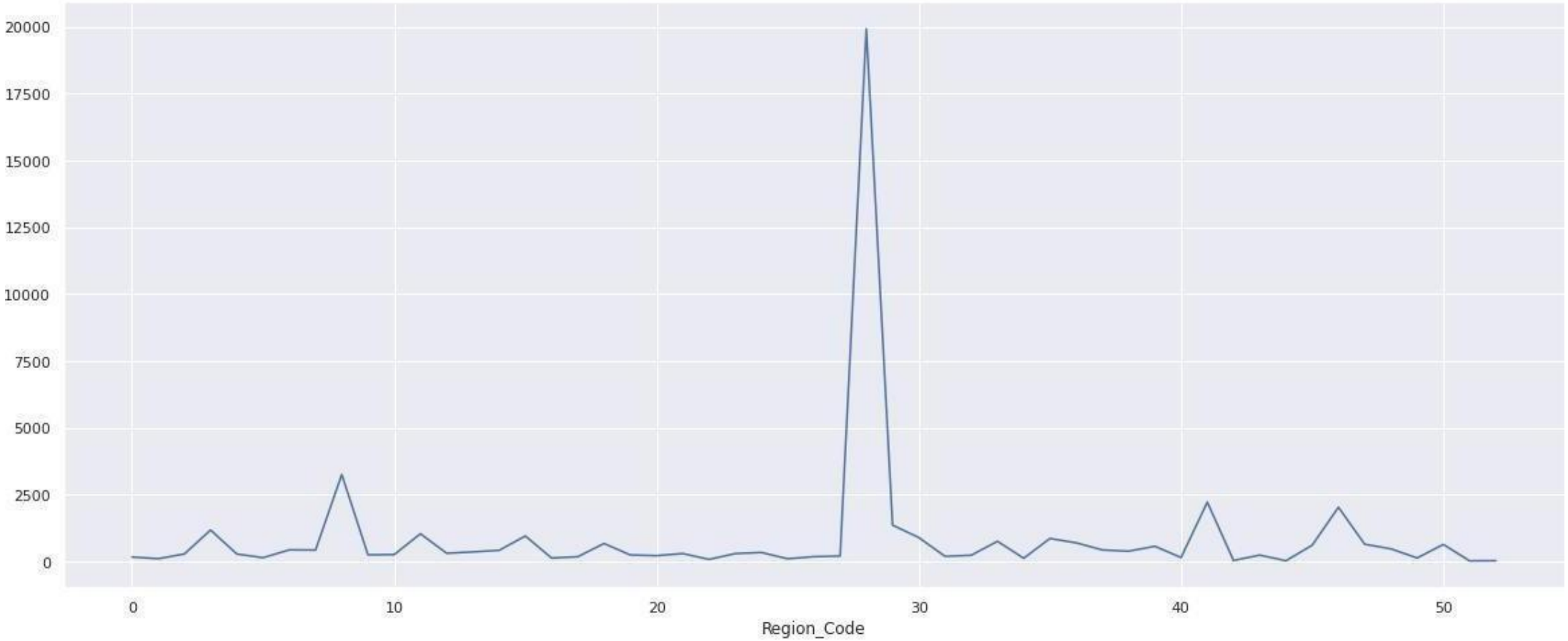
From this line plots, it is clearly evident that the substanstial number of responses are from channels 26 and 124

Customers with different tenures for insurance



This violinplot suggests that the poisitive responses are high from the customers who are loyal for 150 or more days and it drops after 180 days being a customer

Customers from different Region Codes

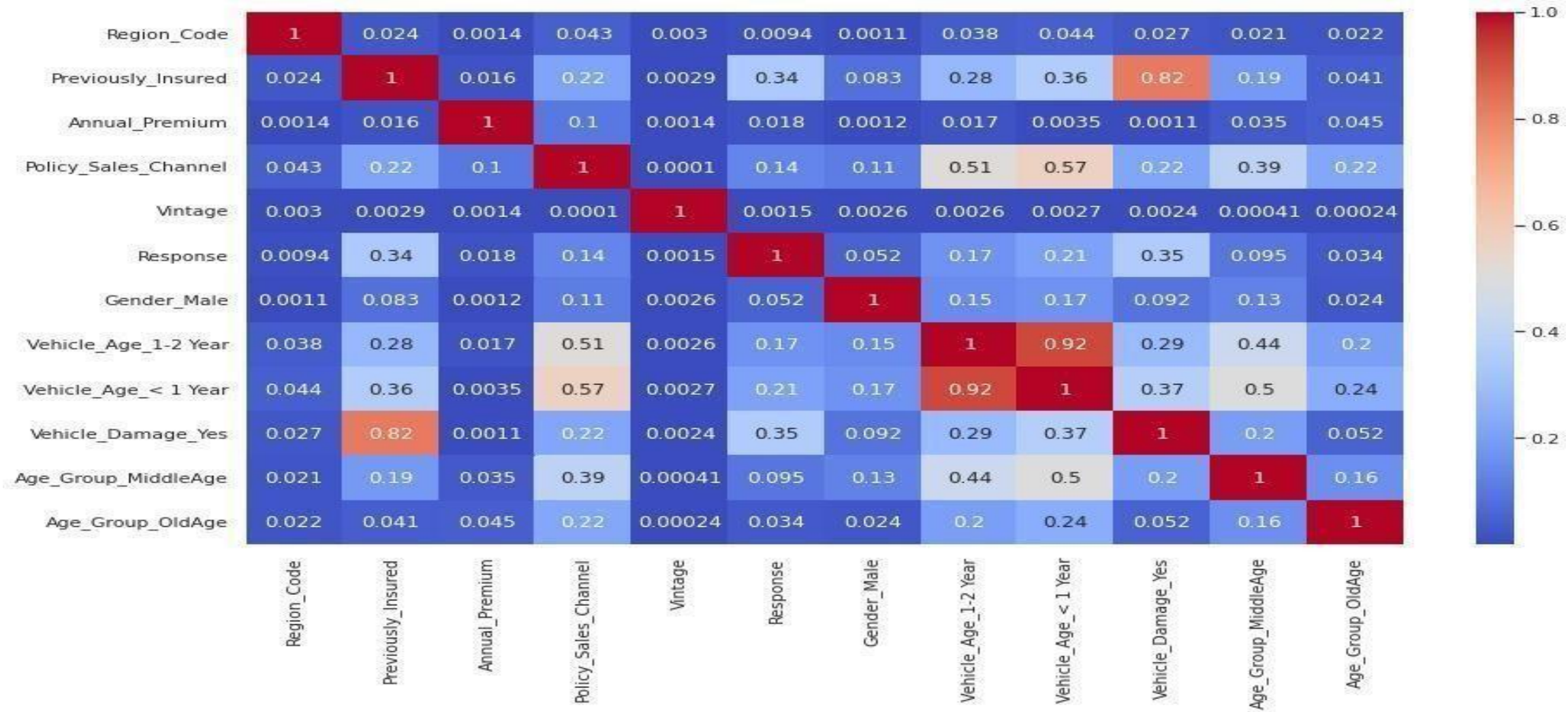


This Line plot shows the positive responses from the customers across all the regions.There is substantial response from the Area code 28 followed by codes 8 and 46

Model preparation

1. Calculating multicollinearity through VIF and filtering our data.
2. Plotting the correlation heatmap of all the columns to visualize the collinearity after dropping columns using VIF.
3. All of the values are of desired data type of modelling i.e, Int64 and Float64
4. There aren’t any null values across the dataset.

Correlation Heat Map



Models used

- Logistic Regression
- Random Forest Classifier
- XGBoost Classifier
- Naive Bayes Classifier

Evaluation of models

Logistic Regression

	precision	recall	f1-score	support
0	0.60	0.95	0.73	60873
1	0.97	0.71	0.82	133624
accuracy			0.78	194497
macro avg	0.78	0.83	0.78	194497
weighted avg	0.85	0.78	0.79	194497

Random Forest

	precision	recall	f1-score	support
0	0.87	1.00	0.93	87345
1	1.00	0.89	0.94	112374
accuracy			0.93	199719
macro avg	0.93	0.94	0.93	199719
weighted avg	0.94	0.93	0.93	199719

XGBoost

	precision	recall	f1-score	support
0	0.65	0.93	0.76	67693
1	0.95	0.73	0.83	126804
accuracy			0.80	194497
macro avg	0.80	0.83	0.79	194497
weighted avg	0.84	0.80	0.80	194497

Naive Bayes Classifier

	precision	recall	f1-score	support
0	0.81	0.67	0.73	233134
1	0.72	0.84	0.78	232877
accuracy			0.76	466011
macro avg	0.76	0.76	0.75	466011
weighted avg	0.76	0.76	0.75	466011

Test set report

	precision	recall	f1-score	support
0	0.81	0.67	0.73	99731
1	0.72	0.84	0.78	99988
accuracy			0.76	199719
macro avg	0.77	0.76	0.76	199719
weighted avg	0.77	0.76	0.76	199719

Challenges faced

1. Pre-processing the data was one of the challenges we faced which includes removing highly correlated variables from the data so as to not hinder the performance of our regression model.

2. Exploring all the columns and calculating VIF for multicollinearity was challenging because it might decrease the models performance.

3. Selecting the appropriate models to maximize the accuracy of our predictions was one of the challenges faced.

Conclusion

- We trained our model on logistic regression and other models .

- Out of all models used , with the Random Forest classification model we were able to get the F1-score of 0.93.

- The model which performed poorly was Naive Bayes Classification model with r2-score of 0.73. Given the size of data and the amount of irrelevance in the data , the above score is good.