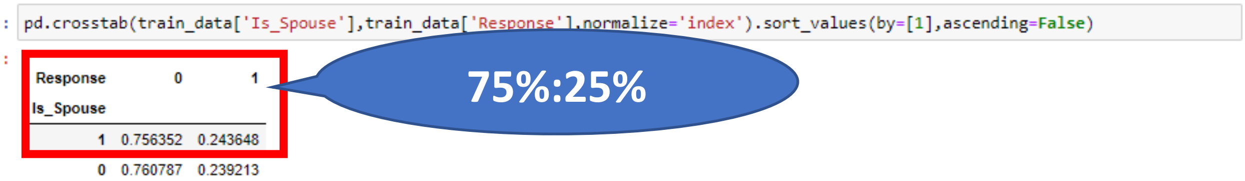
Health Insurance Lead Prediction

Your Client FinMan is a financial services company that provides various financial services like loan, investment funds, insurance etc. to its customers. FinMan wishes to cross-sell health insurance to the existing customers who may or may not hold insurance policies with the company. The company recommend health insurance to it's customers based on their profile once these customers land on the website. Customers might browse the recommended health insurance policy and consequently fill up a form to apply. When these customers fill-up the form, their Response towards the policy is considered positive and they are classified as a lead.

Once these leads are acquired, the sales advisors approach them to convert and thus the company can sell proposed health insurance to these leads in a more efficient manner.

Now the company needs your help in building a model to predict whether the person will be interested in their proposed Health plan/policy

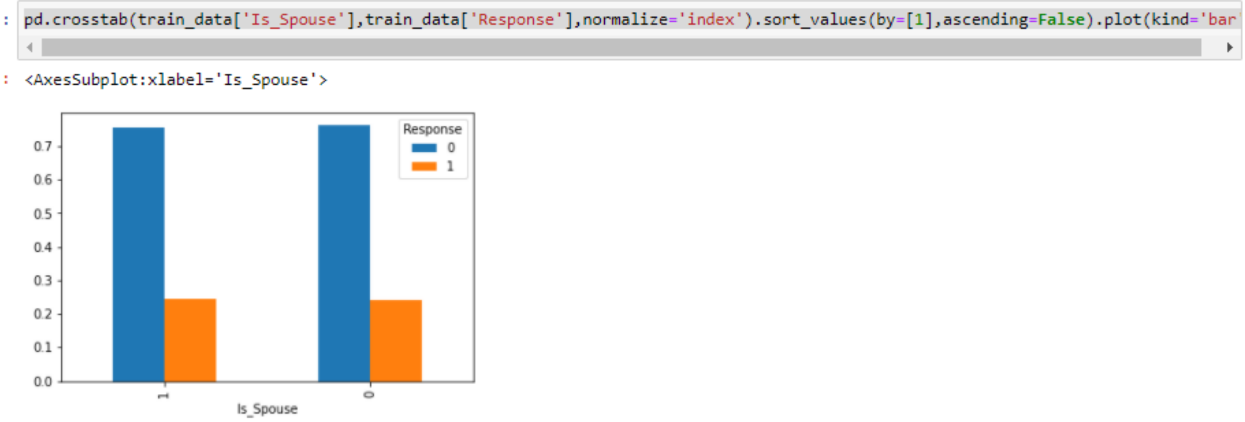
Some Meaningful Insights



```
pd.crosstab(train_data['Is_Spouse'],train_data['Response'],normalize='index').sort_values(by=[1],ascending=False)
```

```
Response         0         1
Is_Spouse
        1  0.756352  0.243648
        0  0.760787  0.239213
```

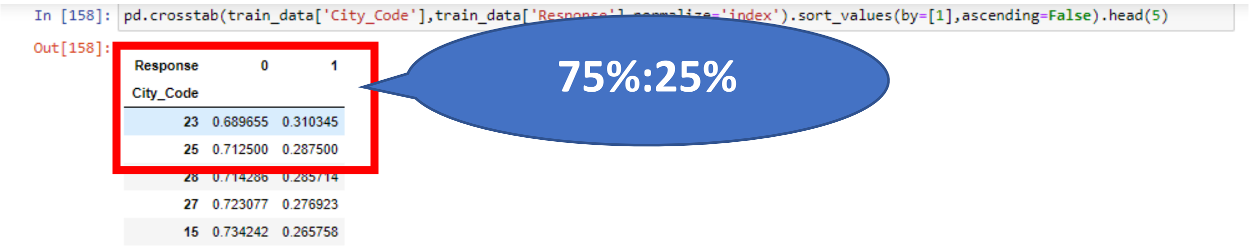
75%:25%



```
pd.crosstab(train_data['Is_Spouse'],train_data['Response'],normalize='index').sort_values(by=[1],ascending=False).plot(kind='bar
```

```
<AxesSubplot:xlabel='Is_Spouse'>
```

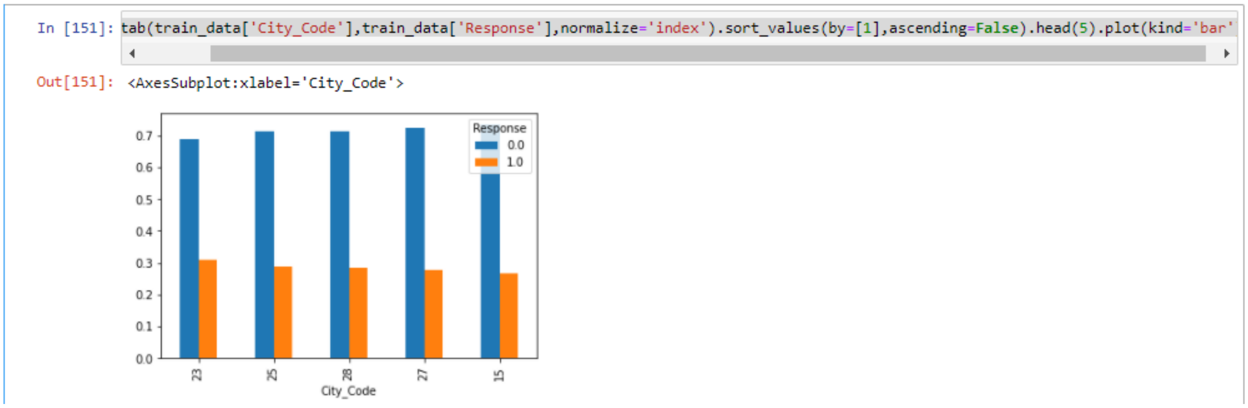
Is Spouse with respect to Response variable has distribution of 75:25 %.. Likelihood of positive response while married approximately equals to while not married. Hence this does not explain the target variable



```
In [158]: pd.crosstab(train_data['City_Code'],train_data['Response'],normalize='index').sort_values(by=[1],ascending=False).head(5)
Out[158]:
```

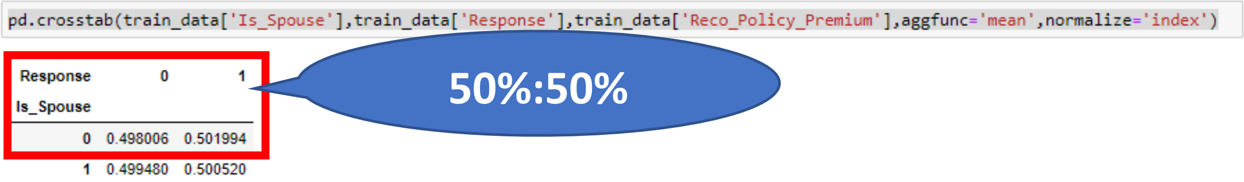
Response	0	1
City_Code		
23	0.689655	0.310345
25	0.712500	0.287500
28	0.714286	0.285714
27	0.723077	0.276923
15	0.734242	0.265758

75%:25%



```
In [151]: tab(train_data['City_Code'],train_data['Response'],normalize='index').sort_values(by=[1],ascending=False).head(5).plot(kind='bar'
Out[151]: <AxesSubplot:xlabel='City_Code'>
```

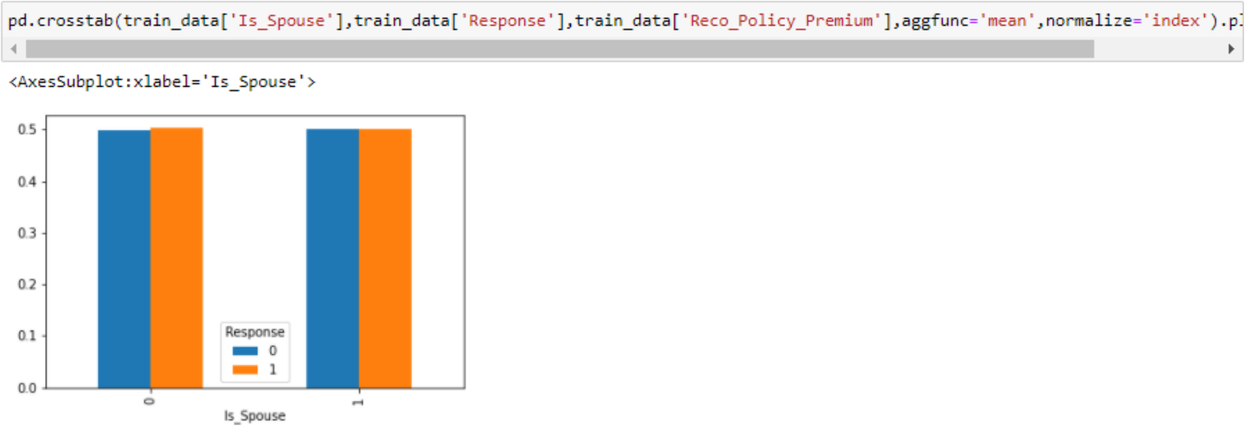
City Code with respect to Response variable has distribution of 75:25 %.



```
pd.crosstab(train_data['Is_Spouse'],train_data['Response'],train_data['Reco_Policy_Premium'],aggfunc='mean',normalize='index')
```

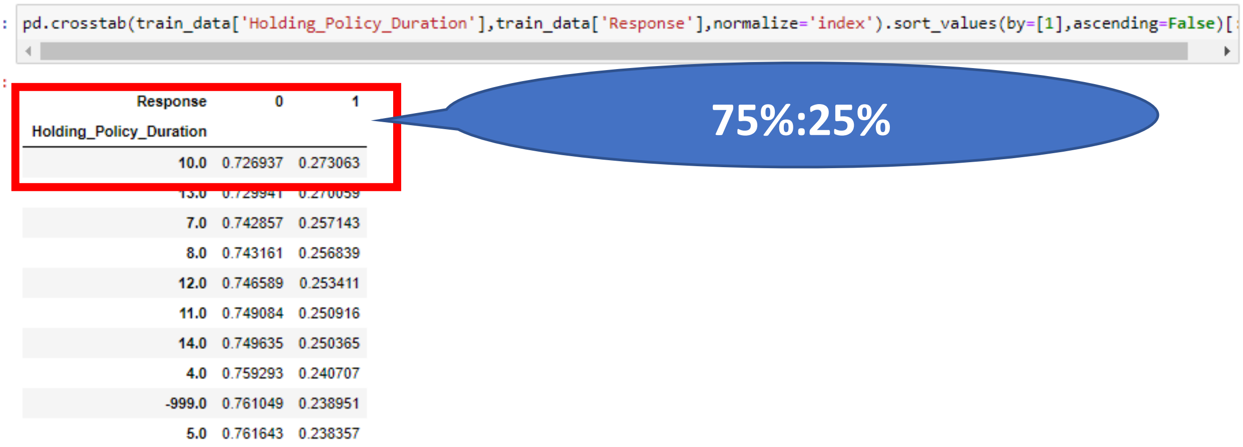
Response	0	1
Is_Spouse		
0	0.498006	0.501994
1	0.499480	0.500520

50%:50%



```
pd.crosstab(train_data['Is_Spouse'],train_data['Response'],train_data['Reco_Policy_Premium'],aggfunc='mean',normalize='index').pl
<AxesSubplot:xlabel='Is_Spouse'>
```

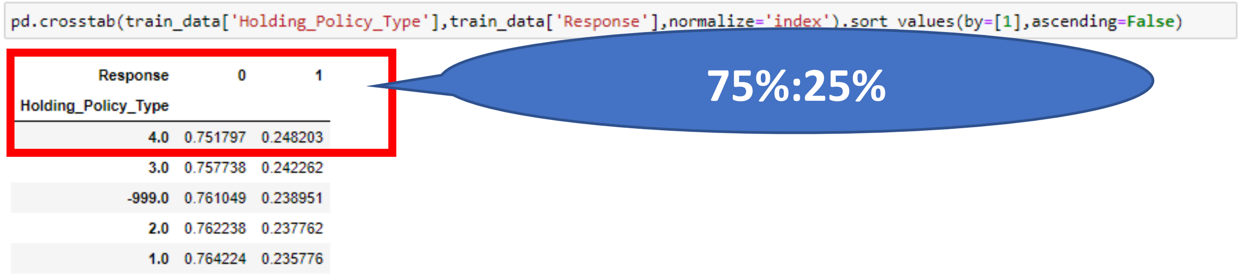
Average Premium of Is Spouse category per Response type has distribution of 50%:50%. Average Premium per spouse category does not explains target varable, as average premium for all the combinations is .50(Approx).



```
: pd.crosstab(train_data['Holding_Policy_Duration'],train_data['Response'],normalize='index').sort_values(by=[1],ascending=False)[:
```

Response	0	1
Holding_Policy_Duration		
10.0	0.726937	0.273063
13.0	0.729941	0.270059
7.0	0.742857	0.257143
8.0	0.743161	0.256839
12.0	0.746589	0.253411
11.0	0.749084	0.250916
14.0	0.749635	0.250365
4.0	0.759293	0.240707
-999.0	0.761049	0.238951
5.0	0.761643	0.238357

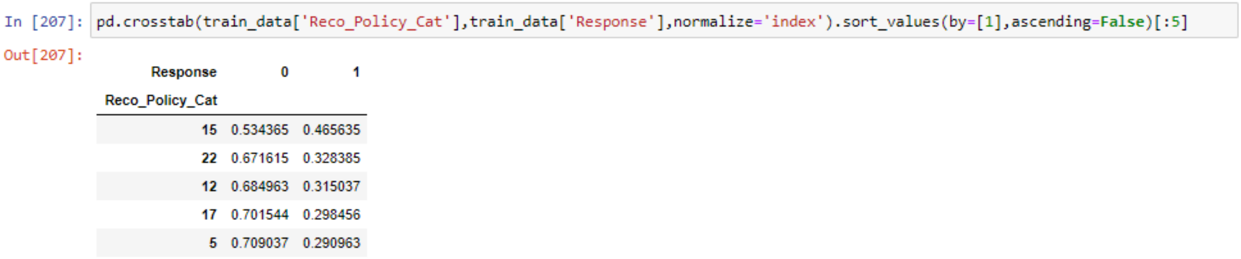
Distribution of Holding Policy Duration to Response type is 75%:25%(Approx). Which states that there is approximately 25% probability for positive response type.



```
pd.crosstab(train_data['Holding_Policy_Type'],train_data['Response'],normalize='index').sort_values(by=[1],ascending=False)
```

Response	0	1
Holding_Policy_Type		
4.0	0.751797	0.248203
3.0	0.757738	0.242262
-999.0	0.761049	0.238951
2.0	0.762238	0.237762
1.0	0.764224	0.235776

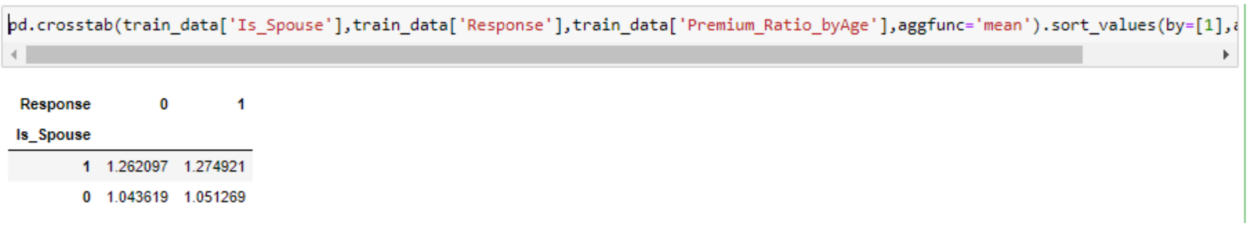
Distribution per Response type for Policy Type is 75% :25% (Approx)



```
In [207]: pd.crosstab(train_data['Reco_Policy_Cat'],train_data['Response'],normalize='index').sort_values(by=[1],ascending=False)[:5]
Out[207]:
```

Response	0	1
Reco_Policy_Cat		
15	0.534365	0.465635
22	0.671615	0.328385
12	0.684963	0.315037
17	0.701544	0.298456
5	0.709037	0.290963

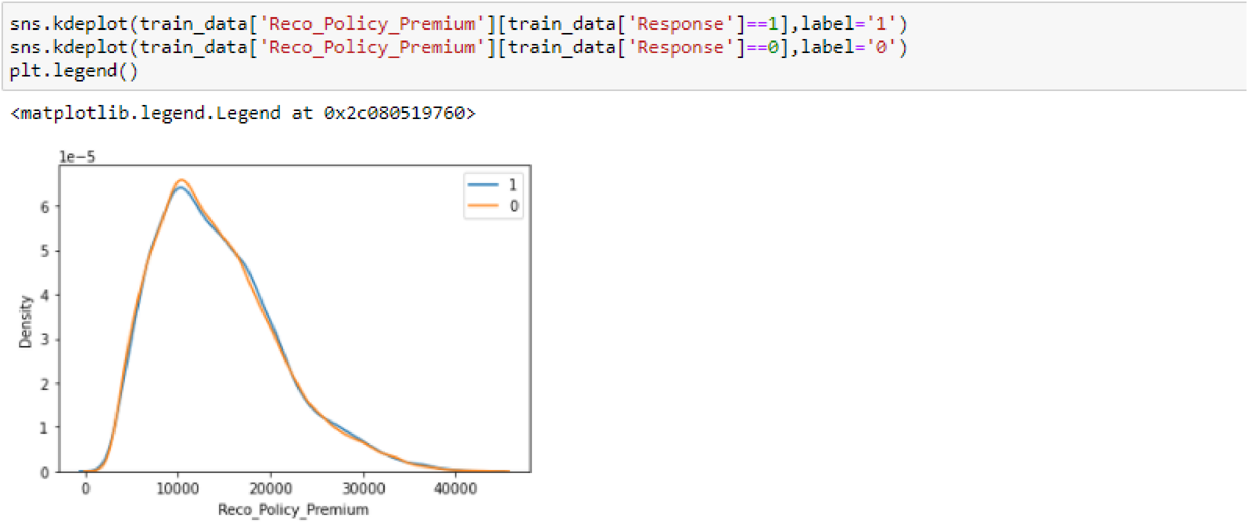
The distribution of Policy Category to Response type shows variation, hence we can say policy category is a potential important variable.



```
pd.crosstab(train_data['Is_Spouse'],train_data['Response'],train_data['Premium_Ratio_byAge'],aggfunc='mean').sort_values(by=[1],a
```

Response	0	1
Is_Spouse		
1	1.262097	1.274921
0	1.043619	1.051269

We notice that within same spouse category it is hard to state difference for each response type, distribution is approximately same. But when we compare the categories of Spouse type , we can see the change in distribution. Which means for non married , the upper age has to spend less in comparison to married. But we can not distinguish for response categories for each spouse type.



```
sns.kdeplot(train_data['Reco_Policy_Premium'][train_data['Response']==1],label='1')
sns.kdeplot(train_data['Reco_Policy_Premium'][train_data['Response']==0],label='0')
plt.legend()
```

```
<matplotlib.legend.Legend at 0x2c080519760>
```

We see the premium distribution for response type is almost same, there is no difference in distribution for positive and negative response. Hence it does not help in explaining the Response, both category has equal likelihood.