Health Insurance Data analysis ⍰ ⍰

Predict Health Insurance Owners' who will be interested in Vehicle Insurance

About Dataset

Our client is an Insurance company that has provided Health Insurance to its customers now they need your help in building a model to predict whether the policyholders (customers) from past year will also be interested in Vehicle Insurance provided by the company.

An insurance policy is an arrangement by which a company undertakes to provide a guarantee of compensation for specified loss, damage, illness, or death in return for the payment of a specified premium. A premium is a sum of money that the customer needs to pay regularly to an insurance company for this guarantee.

For example, you may pay a premium of Rs. 5000 each year for a health insurance cover of Rs. 200,000/- so that if, God forbid, you fall ill and need to be hospitalised in that year, the insurance provider company will bear the cost of hospitalisation etc. for upto Rs. 200,000. Now if you are wondering how can company bear such high hospitalisation cost when it charges a premium of only Rs. 5000/-, that is where the concept of probabilities comes in picture. For example, like you, there may be 100 customers who would be paying a premium of Rs. 5000 every year, but only a few of them (say 2-3) would get hospitalised that year and not everyone. This way everyone shares the risk of everyone else.

Just like medical insurance, there is vehicle insurance where every year customer needs to pay a premium of certain amount to insurance provider company so that in case of unfortunate accident by the vehicle, the insurance provider company will provide a compensation (called 'sum assured') to the customer.

Building a model to predict whether a customer would be interested in Vehicle Insurance is extremely helpful for the company because it can then accordingly plan its communication strategy to reach out to those customers and optimise its business model and revenue.

Now, in order to predict, whether the customer would be interested in Vehicle insurance, you have information about demographics (gender, age, region code type), Vehicles (Vehicle Age, Damage), Policy (Premium, sourcing channel) etc.

Select a real-world dataset

```
!pip install jovian opendatasets --upgrade --quiet
```

```
# Change this
dataset_url = 'https://www.kaggle.com/datasets/anmolkumar/health-insurance-cross-sell-p
```

```
import opendatasets as od
od.download(dataset_url)
```

```
Skipping, found downloaded files in ".\health-insurance-cross-sell-prediction" (use
force=True to force download)
```

```
project_name = 'health-insurance-data-analysis'
```

```
!pip install jovian --upgrade -q
```

```
import pandas as pd
import numpy as np
```

```
# Change this
data_dir = './health-insurance-cross-sell-prediction'
```

```
import os
os.listdir(data_dir)
```

```
['sample_submission.csv', 'test.csv', 'train.csv']
```

```
import jovian
```

```
jovian.commit(project=project_name)
```

```
[jovian] Updating notebook "shashi-tron/health-insurance-data-analysis" on
https://jovian.com/
[jovian] Committed successfully! https://jovian.com/shashi-tron/health-insurance-data-
analysis
```

```
'https://jovian.com/shashi-tron/health-insurance-data-analysis'
```

Data preparation & cleaning

Now that we have the data imported, will prepare the data for further analysis by pre-anaylsing cleaning the data set.

First we load all base csv files

```
insurance_raw_df = pd.read_csv(data_dir + '/train.csv')
```

```
insurance_raw_df[:5]
```

	id	Gender	Age	Driving_License	Region_Code	Previously_Insured	Vehicle_Age	Vehicle_Damage	Annual_Premium
0	1	Male	44	1	28.0	0	> 2 Years	Yes	40454.0
1	2	Male	76	1	3.0	0	1-2 Year	No	33536.0
2	3	Male	47	1	28.0	0	> 2 Years	Yes	38294.0
3	4	Male	21	1	11.0	1	< 1 Year	No	28619.0
4	5	Female	29	1	41.0	1	< 1 Year	No	27496.0

#data is already clean and in case of further anylsis requried operation is perfomwed to manipulate the data

```
insurance_raw_df.head(2)
```

	id	Gender	Age	Driving_License	Region_Code	Previously_Insured	Vehicle_Age	Vehicle_Damage	Annual_Premium
0	1	Male	44	1	28.0	0	> 2 Years	Yes	40454.0
1	2	Male	76	1	3.0	0	1-2 Year	No	33536.0

Checking the info of all columns to know the datatypes and Non-Null counts

```
##
insurance_raw_df.info()
```

```
<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
```

```
## finding any null value
insurance_raw_df.isnull()
```

	id	Gender	Age	Driving_License	Region_Code	Previously_Insured	Vehicle_Age	Vehicle_Damage	Annua
0	False	False	False	False	False	False	False	False	
1	False	False	False	False	False	False	False	False	
2	False	False	False	False	False	False	False	False	
3	False	False	False	False	False	False	False	False	
4	False	False	False	False	False	False	False	False	
...	...	...	...	...	...	...	...	...	
381104	False	False	False	False	False	False	False	False	
381105	False	False	False	False	False	False	False	False	
381106	False	False	False	False	False	False	False	False	
381107	False	False	False	False	False	False	False	False	

	id	Gender	Age	Driving_License	Region_Code	Previously_Insured	Vehicle_Age	Vehicle_Damage	Annua
381108	False	False	False	False	False	False	False	False	

381109 rows × 12 columns

-To find the missing values in the rows we use `isnull().sum()` funcion and sorted it in Descending order by using `sort_values(ascending=False)**

Finding of null values

```
insurance_raw_df.isnull().sum()
```

```
id                      0
Gender                  0
Age                     0
Driving_License         0
Region_Code             0
Previously_Insured      0
Vehicle_Age             0
Vehicle_Damage          0
Annual_Premium          0
Policy_Sales_Channel    0
Vintage                 0
Response                0
dtype: int64
```

Above data set does not contain any null value so data clean.in case found null value it would calculated with mean respective column filled value with it. ## command to calculate mean value example -- insurance_raw_df[age].fillna(insurance_raw_df[age].mean(),inplace = True)

```
insurance_raw_df.describe()
```

	id	Age	Driving_License	Region_Code	Previously_Insured	Annual_Premium	Policy
count	381109.000000	381109.000000	381109.000000	381109.000000	381109.000000	381109.000000	3
mean	190555.000000	38.822584	0.997869	26.388807	0.458210	30564.389581	
std	110016.836208	15.511611	0.046110	13.229888	0.498251	17213.155057	
min	1.000000	20.000000	0.000000	0.000000	0.000000	2630.000000	
25%	95278.000000	25.000000	1.000000	15.000000	0.000000	24405.000000	
50%	190555.000000	36.000000	1.000000	28.000000	0.000000	31669.000000	
75%	285832.000000	49.000000	1.000000	35.000000	1.000000	39400.000000	
max	381109.000000	85.000000	1.000000	52.000000	1.000000	540165.000000	

Above data table no error values all min and maximum value are valid

```
import jovian
```

```
 jovian.commit()
```

```
[jovian] Updating notebook "shashi-tron/health-insurance-data-analysis" on
https://jovian.com/
[jovian] Committed successfully! https://jovian.com/shashi-tron/health-insurance-data-
analysis
```

```
'https://jovian.com/shashi-tron/health-insurance-data-analysis'
```

Exploratory Analysis and Visualization

Exploratory data analysis is a way to better understand data. The numerical data will be visualized in graphs.

Let's begin by importing `matplotlib.pyplot and seaborn`. Some graph visualization options are set as well.

```
import seaborn as sns
import matplotlib
import matplotlib.pyplot as plt
%matplotlib inline

sns.set_style('darkgrid')
matplotlib.rcParams['font.size'] = 14
matplotlib.rcParams['figure.figsize'] = (9, 5)
matplotlib.rcParams['figure.facecolor'] = '#00000000'
```

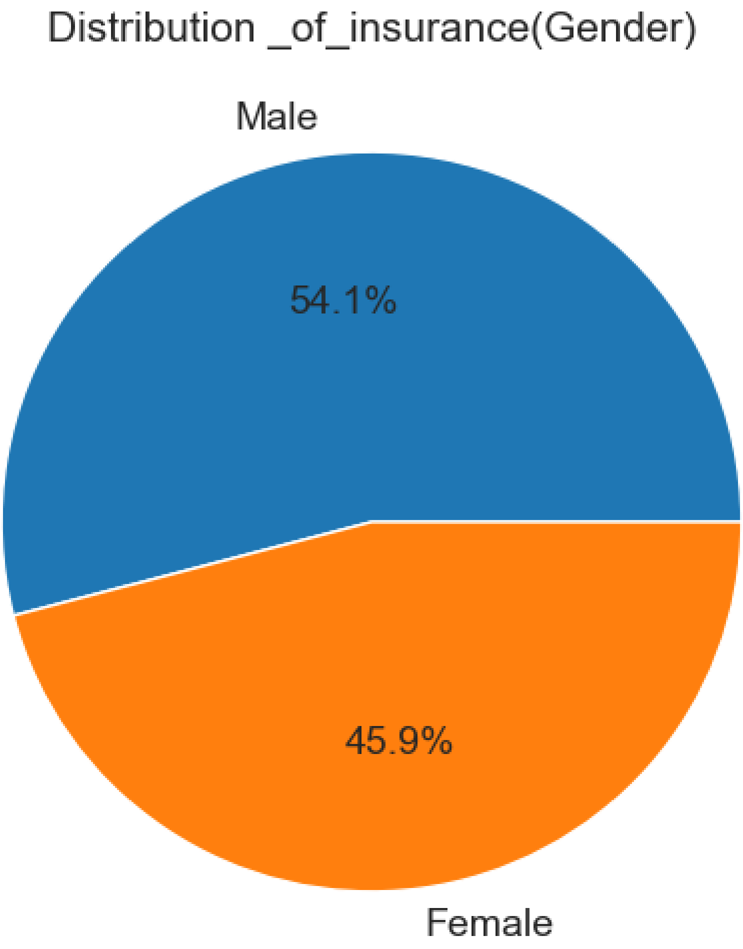
Lets First check the customers based on gender
which gives us insight which need customer need to be targeted first for cross -sell of insurance as combo product while purchasing vehicle insurance , which helps us to make informed decisions in business

```
insurance_by_sex =insurance_raw_df['Gender'].value_counts()
```

```
insurance_by_sex
```

```
Male      206089
Female    175020
Name: Gender, dtype: int64
```

```
plt.figure(figsize=(12,6))
plt.title('Distribution _of_insurance(Gender)',pad= 7, fontsize = 15)
plt.pie(insurance_by_sex,labels=insurance_by_sex.index,autopct='%1.1f%%');
```



above pie chart clearly shown male 54.1% and female 45.9% as customer base who bought the Vehicle insurance

Insight
pie chart gives clear view, which section customers need to be focused based on gender but there are missing elements which need to be taken considration targetd sales.

> people of what age need to be targeted which narrow down to quality customers.

> Customer with vechile damge histroy

```
import jovian
```

```
jovian.commit()
```

```
[jovian] Updating notebook "shashi-tron/health-insurance-data-analysis" on
https://jovian.com/
[jovian] Committed successfully! https://jovian.com/shashi-tron/health-insurance-data-
analysis
```

```
'https://jovian.com/shashi-tron/health-insurance-data-analysis'
```

4.Upon initial inspection of the data, we can start thinking of some questions about it that we would want to answer.

Here above data set holds customer information customers having car insurance. Here some of question can be formed which help cross health insurance targeted customers.

```
### Following question help us to make informed decision
```

1.what is the relationship between age, gender, and response to health insurance cross-sell offers? Can we identify specific age and gender groups that are more likely to be interested in purchasing health insurance as a cross-sell product?

2.Does vehicle damage have any impact on the likelihood of a customer purchasing health insurance as a cross-sell product? Are customers who have experienced vehicle damage more or less likely to be interested in purchasing health insurance?

3.What is the most effective policy sales channel for marketing health insurance cross-sell products to car insurance customers? Are there specific channels that are more effective for reaching certain demographic groups or types of customers?

4.what is the impact of the annual premium on the likelihood of purchasing health insurance as a cross-sell product? Are customers with higher or lower annual premiums more likely to respond to health insurance cross-sell offers?

Q1.what is the relationship between age, gender, and response to health insurance cross-sell offers? Can we identify specific age and gender groups that are more likely to be interested in purchasing health insurance as a cross-sell product?

Here above question we gone identify group customers likely to purchase Health insurance along car insurance.**

```
# we have grouped columns
sort_insurance_df = insurance_raw_df[['Gender','Age','Vehicle_Damage','Vehicle_Age','Re
```

```
sort_insurance_df
```

	Gender	Age	Vehicle_Damage	Vehicle_Age	Response	Policy_Sales_Channel	Annual_Premium
0	Male	44	Yes	> 2 Years	1	26.0	40454.0
1	Male	30	Yes	1-2 Year	1	157.0	2630.0
2	Male	36	Yes	1-2 Year	1	26.0	2630.0
3	Male	24	Yes	< 1 Year	0	152.0	33480.0
4	Male	49	Yes	1-2 Year	0	26.0	29180.0
...	...	...	...	...	...	...	...
381104	Female	48	Yes	1-2 Year	0	13.0	38427.0

	Gender	Age	Vehicle_Damage	Vehicle_Age	Response	Policy_Sales_Channel	Annual_Premium
381105	Female	25	No	< 1 Year	0	152.0	28658.0
381106	Female	70	No	1-2 Year	0	26.0	41991.0
381107	Female	26	No	< 1 Year	0	152.0	29552.0
381108	Female	35	No	1-2 Year	0	151.0	35150.0

381109 rows × 7 columns

Transforming data into numerical data type

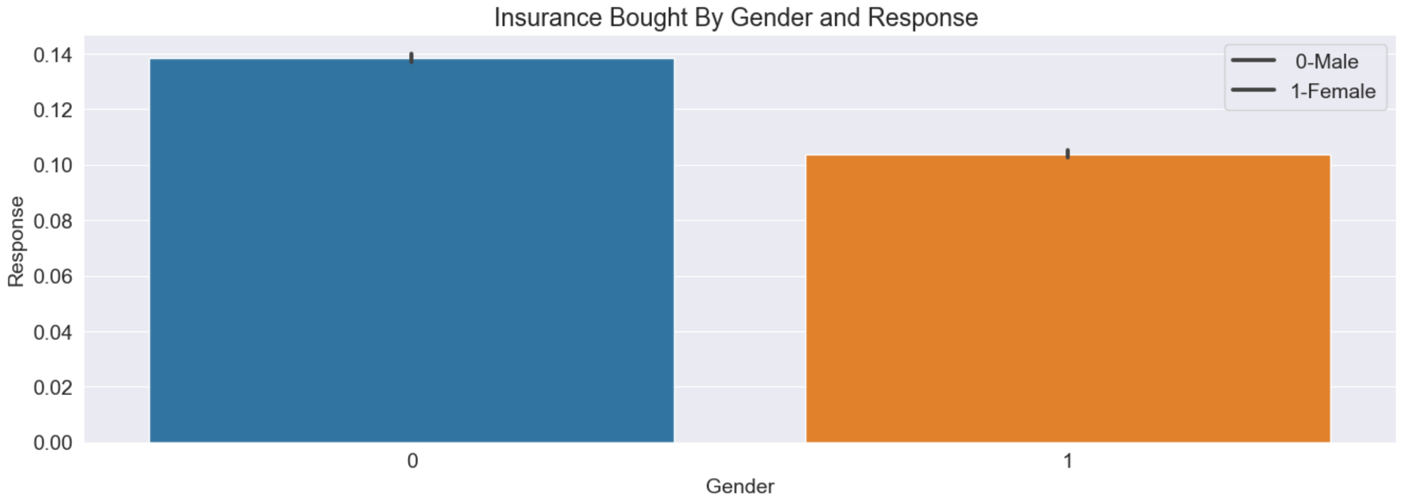
```
## Now we will cnovert object type data to integer values which helps us to make inform
sort_insurance_df['Gender'] = sort_insurance_df['Gender'].map({'Male': 0, 'Female': 1})
sort_insurance_df['Vehicle_Age'] =  sort_insurance_df['Vehicle_Age'].map({'< 1 Year': 0
sort_insurance_df['Vehicle_Damage'] = sort_insurance_df['Vehicle_Damage'].map({'Yes':1,
```

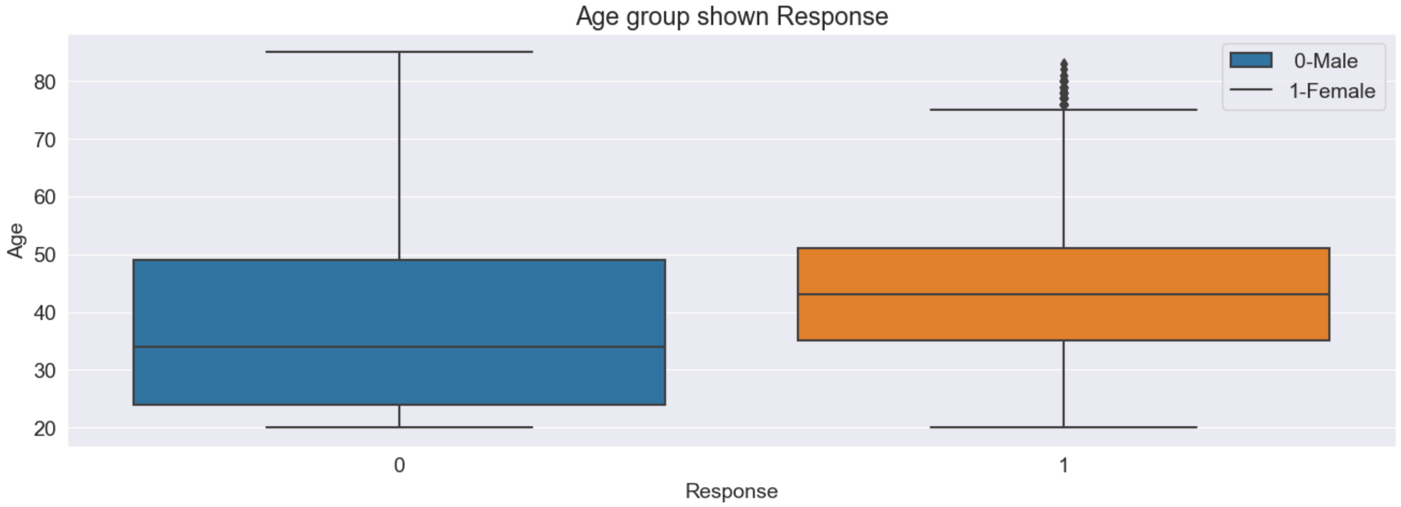
```
sort_insurance_df
```

	Gender	Age	Vehicle_Damage	Vehicle_Age	Response	Policy_Sales_Channel	Annual_Premium
0	0	44	1	2	1	26.0	40454.0
1	0	30	1	1	1	157.0	2630.0
2	0	36	1	1	1	26.0	2630.0
3	0	24	1	0	0	152.0	33480.0
4	0	49	1	1	0	26.0	29180.0
...	...	...	...	...	...	...	...
381104	1	48	1	1	0	13.0	38427.0
381105	1	25	0	0	0	152.0	28658.0
381106	1	70	0	1	0	26.0	41991.0
381107	1	26	0	0	0	152.0	29552.0
381108	1	35	0	1	0	151.0	35150.0

381109 rows × 7 columns

```
#perfoming explortitary data anylsis by visual
plt.figure(figsize=(16,5))
plt.title('Insurance Bought By Gender and Response ')
sns.barplot(x='Gender', y='Response', data=sort_insurance_df)
plt.legend([' 0-Male' ,'1-Female', ])
plt.show()
plt.figure(figsize=(16,5))
plt.title('Age group shown Response')
sns.boxplot(x='Response', y='Age', data=sort_insurance_df)
plt.legend([' 0-Male' ,'1-Female'])
plt.show()
```





>> in above plot can seen result shown represntes more "men" are instered as per response consideration which is represnted in '0'
>> in term of response age group above 50 age apperas more intersetd compare other age groups

insight

>> response among gender 'male' that comapre to 'female' , Sell of health insurance with combo offer can be given Customer above age of 50.

>> By giving combo offers loyal customer will be created so purchase of Vehicle insurance with health insurance.

2.Does vehicle damage have any impact on the likelihood of a customer purchasing health insurance as a cross-sell product? Are customers who have experienced vehicle damage more or less likely to be interested in purchasing health insurance?

Lets plot graph where customer who have vehicle damage are Responded

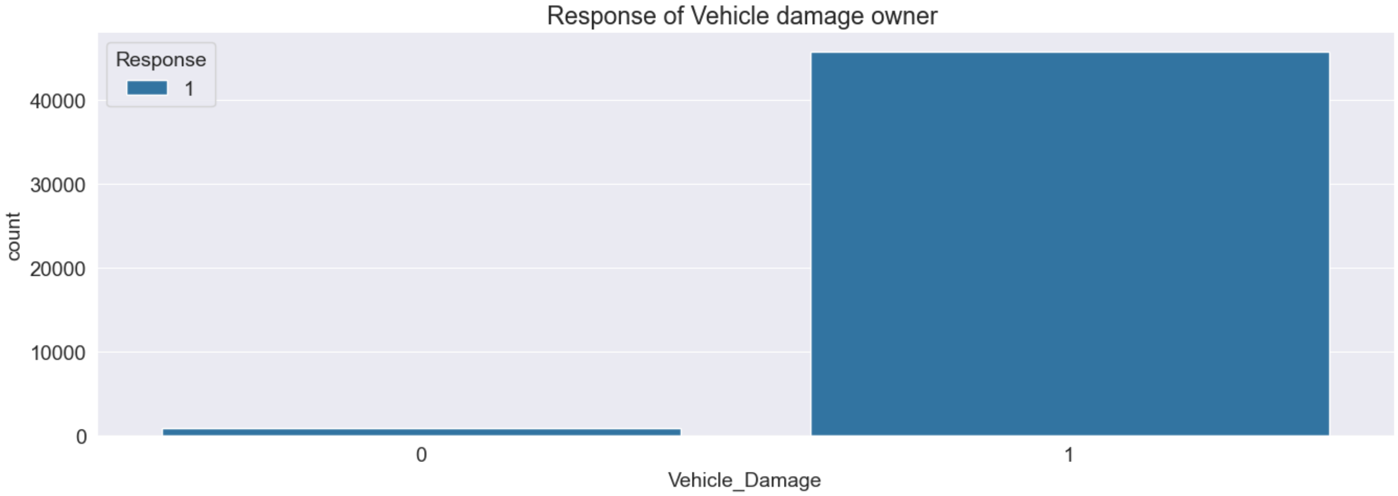
```python
#filtering the responses by cutomer response
interested_in_health_insurance= sort_insurance_df[sort_insurance_df['Response'] == 1]
```

```python
interested_in_health_insurance
```

	Gender	Age	Vehicle_Damage	Vehicle_Age	Response	Policy_Sales_Channel	Annual_Premium
0	0	44	1	2	1	26.0	40454.0
1	0	30	1	1	1	157.0	2630.0
2	0	36	1	1	1	26.0	2630.0
6	0	58	1	1	1	26.0	42819.0
9	0	73	1	2	1	122.0	50041.0
...	...	...	...	...	...	...	...
381041	1	40	1	1	1	124.0	33934.0
381050	1	43	1	1	1	16.0	2630.0
381059	1	38	1	1	1	26.0	24955.0
381061	1	31	1	1	1	124.0	59028.0
381084	1	34	1	1	1	26.0	32083.0

46710 rows × 7 columns

```python
#Perfoming Visulization
plt.figure(figsize=(16,5))
plt.title('Response of Vehicle damage owner ')
sns.countplot(x='Vehicle_Damage', hue='Response', data=interested_in_health_insurance)
plt.show()
```



yes it appears to be Vehicle owners with vehicle damage as more response , this would potential customers to cross-sell offer

insight

> > customer with vechile damge indeed purchase Vehicle insurance as compulsary as cross -sell product health insurance would be bought by customers.

> > offering combo offer and long term insurance with better discounts gives good scope customer purchasing insurance products.

```
import jovian
```

```
jovian.commit(project=project_name)
```

```
[jovian] Updating notebook "shashi-tron/health-insurance-data-analysis" on
https://jovian.com/
[jovian] Committed successfully! https://jovian.com/shashi-tron/health-insurance-data-
analysis
```

```
'https://jovian.com/shashi-tron/health-insurance-data-analysis'
```

3.What is the most effective policy sales channel for marketing health insurance cross-sell products to car insurance customers? Are there specific channels that are more effective for reaching certain demographic groups or types of customers?

Here we gone take consideration specfice columns which would help us find more effective channel cross-sell products likely to be bought.

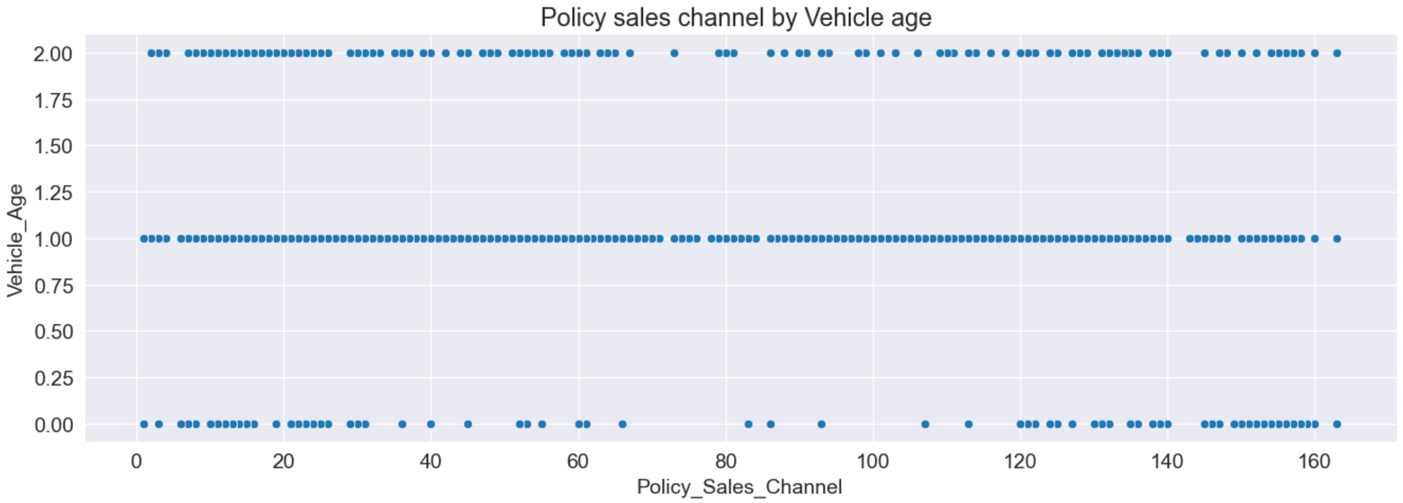
```
channel_df = sort_insurance_df[['Vehicle_Damage','Policy_Sales_Channel','Vehicle_Age']]
```

```
channel_df.head(5)
```

	Vehicle_Damage	Policy_Sales_Channel	Vehicle_Age
0	1	163.0	2
1	1	163.0	1

	Vehicle_Damage	Policy_Sales_Channel	Vehicle_Age
2	1	163.0	1
3	1	163.0	1
4	1	163.0	0

```
plt.figure(figsize=(16,5))
sns.scatterplot(x= channel_df.Policy_Sales_Channel, y= channel_df.Vehicle_Age);
plt.title('Policy sales channel by Vehicle age ');
```

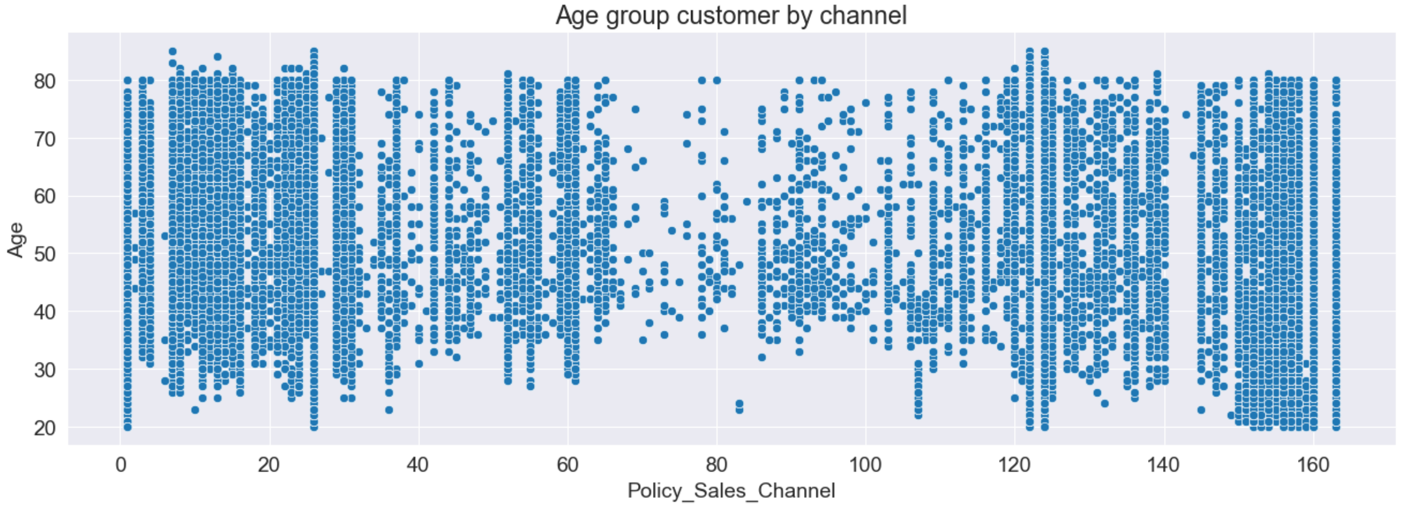


above scatterplot shows conncetration of policy sales and vehicle Age more conncentation of policy appears in range of (0-20) and (120-160) , which are more effective in terms of policy can be sold .

insight

> Here Customer conncentration in limited number of channel which policy bought , there is requirement of marketing and promoting insurance product rest of policy channel for effective sales of insurance product.

```
plt.figure(figsize=(16,5))
plt.title('Age group customer by channel')
sns.scatterplot(x='Policy_Sales_Channel', y='Age', data=sort_insurance_df)
plt.show()
```



Here above scatterplot can be seen , clusters more density are good policy channel which are rangeing between (120 - 160) and (0-20)

insight

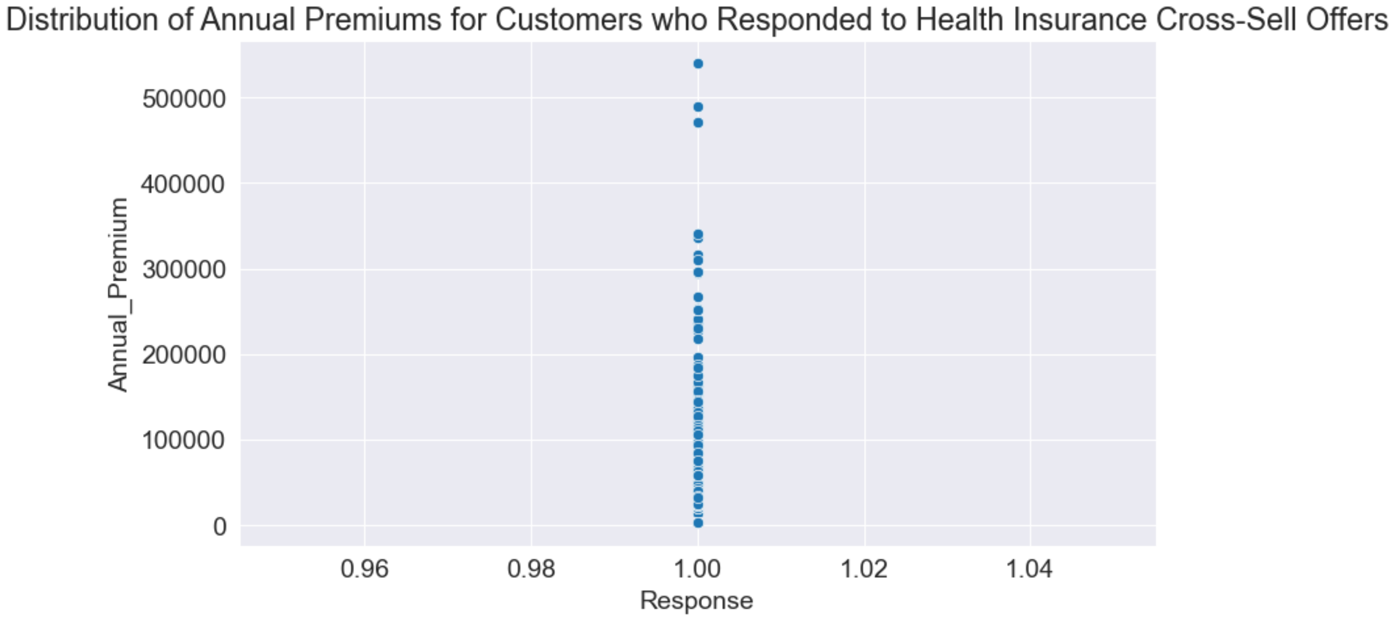
> The concentartion of elderly population lies in policy channel shown scatterplot , but elderly population is scatterd in rest of channel.

> So , it appears there is not much need campgin scatterd area (40-120), focus on the areas which already elderly customer exist.

4. What is the impact of the annual premium on the likelihood of purchasing health insurance as a cross-sell product? Are customers with higher or lower annual premiums more likely to respond to health insurance cross-sell offers?

```
# step 1 filter customer who responded
cross_sell_data =sort_insurance_df[sort_insurance_df['Response'] == 1]
```

```
#visulaztion
sns.scatterplot(y = cross_sell_data['Annual_Premium'], x = cross_sell_data['Response'])
plt.title('Distribution of Annual Premiums for Customers who Responded to Health Insura
plt.show()
```



Annual premium insurance with low premium are responded more that of compare to higher premium insurance , which are ranging from below 2,00,000

insight

> So, customer responded are more which includes approx 90% customer base .

> launch of cross sell - health insurance appears effective since most customer base appears intersed in purchase of product

```
mean_annual_premium = cross_sell_data['Annual_Premium'].mean()
print(f"Mean Annual Premium for Customers who Responded to Health Insurance Cross-Sell
```

```
Mean Annual Premium for Customers who Responded to Health Insurance Cross-Sell Offers:
31604.09
```

```
import jovian
```

```
jovian.commit(project=project_name)
```

```
[jovian] Updating notebook "shashi-tron/health-insurance-data-analysis" on
https://jovian.com/
[jovian] Committed successfully! https://jovian.com/shashi-tron/health-insurance-data-
analysis
```

```
'https://jovian.com/shashi-tron/health-insurance-data-analysis'
```

Inferences and Conclusion

According to this analysis the cross sell of Health insurance can be done through spefice policy_channel and based on gender and Vehicle_damages.

Based on anylsis people vehicle damge tends buy health insurance.

According gender spefice 'male' can be seen targetd customer sell health insurance

```
import jovian
```

```
jovian.commit()
```

```
[jovian] Updating notebook "shashi-tron/health-insurance-data-analysis" on
https://jovian.com/
[jovian] Committed successfully! https://jovian.com/shashi-tron/health-insurance-data-
analysis
```

```
'https://jovian.com/shashi-tron/health-insurance-data-analysis'
```

References

https://matplotlib.org/

https://pandas.pydata.org/

https://www.kaggle.com/

Youtube videos :-

Jovian - Build an Exploratory Data Analysis Project from Scratch with Python, Numpy, and Pandas

Datasets Source
'https://www.kaggle.com/datasets/anmolkumar/health-insurance-cross-sell-prediction'

Future work:
In the future, machine learning techniques could be applied to the health insurance cross sell prediction to undersatnd behavior of customers.

```
import jovian
```

```
jovian.commit()
```