Sunrise Life Insurance

Jun 2023

By Huy Pham
LinkedIn: https://www.linkedin.com/in/huypham040100/
Email: qhuy1508@gmail.com

Background

In 2023, Sunrise Life Insurance is looking to anticipate which customers are likely to lapse on their policies. By identifying such customers, the company intends to proactively reach out to them with some customer care program in a bid to retain them.

To achieve this, Sunrise Life Insurance will need to employ predictive analytics techniques that can analyze vast amounts of data on customer and other relevant factors. Once the predictive model has been developed, company will use it to generate a list of customers who are at high risk of lapsing. The company will then reach out to these customers with tailored offers that are designed to meet their unique needs and preferences.

The goal is to leverage data insights to improve customer satisfaction, reduce lapse, and increase customer loyalty.

Data Dictionary

1. POLICY_OWNER_ID: Policy owner ID (unique)
2. POLICY_TERM_YEARS: Life Assured period (i.e. period of the policy during which protection is applied)
3. POLICY_OWNER_AGE_AT_ISSUE : Age of members at policy issuance
4. TOTAL_LA: Total of Life Assured (i.e. individuals protected by the policy – excluding Policy owners)
5. ANNUAL_FEES: Annual insurance fee
6. POLICY_OWNER_ANNUAL_INCOME : Annual income of policy owner
7. POLICY_OWNER_OCCUPATION_ID: Occupation of policy owner
8. POLICY_OWNER_MARITAL_STATUS: Marriage status (Divorced, Married, Single, Widowed) of policy owner
9. POLICY_OWNER_GENDER: Gender of policy owner
10. PRODUCT_TYPE: Name of insurance product
11. PAYMENT_MODE: Payment mode of the policy (paid annually, semi-annually, monthly, quarterly, single-premium)
12. AGENT_CODE: Code of the agent who serves the policy owner
13. LAPSE_STATUS (Label): CANCELLED (i.e. Lapsed), INFORCE

Questions

1. **SQL:** We have below table:

- Table [Customer_Master] would contain data about Customer
- Table [Policy_Master] would contain data about Policies (One Customer can have multiple policies)

Table Name	Column Name	Data Type	Example	Description
Customer_Master	POLICY_OWNER_ID **(PRIMARY KEY)**	Int	7526648	Customer ClientNumber - Unique identifier for each Customer
Customer_Master	CUSTOMER_CITY	Nvarchar	Ho Chi Minh	Customer Address
Customer_Master	CUSTOMER_AGE	Int	35	Customer Current Age
Customer_Master	CUSTOMER_NAME	Nvarchar	Eric Barker	Customer Name

Table Name	Column Name	Data Type	Example	Description
Policy_Master	POLICY_NUMBER **(PRIMARY KEY)**	Int	28365874	Policy Number - Unique identifier for each Policy
Policy_Master	POLICY_OWNER_ID	Nvarchar(8)	07526642	Customer ClientNumber - Unique identifier for each Customer
Policy_Master	POLICY_OWNER_GENDER	Nvarchar(1)	F	Gender of Customer (Policy Owner) when born
Policy_Master	POLICY_ISSUED_DATE	Date	2020-01-02	Issued Date of Policy

SQL

The Marketing department is requiring an extraction of a customer list that includes their gender information. To fulfill this requirement, please provide a SQL script that selects the POLICY_OWNER_ID and corresponding Age, Name, Gender from the database. It is important that the output is accurate and efficiently executed

```
SELECT
    p.POLICY_OWNER_ID,
    c.CUSTOMER_AGE,
    c.CUSTOMER_NAME,
    POLICY_OWNER_GENDER
FROM  Customer_Master as c
INNER JOIN Policy_Master as p ON c.POLICY_OWNER_ID = p.POLICY_OWNER_ID
```

Data Wrangling

1. Imputing the missing values. Describe how and explain your reasoning

Missing values:

1. POLICY_OWNER_ANNUAL_INCOME (**14.72%**)
2. POLICY_OWNER_OCCUPATION_ID (**0.37%**)

Types of Missing Data:

1. Missing completely at random (MCAR)
2. Missing at random (MAR)
3. Missing not at random (MNAR)

My approaches:

1. Used **MICE** imputation for ANNUAL_INCOME
2. Ignored nulls in Occupation_ID

```
Data columns (total 13 columns):
 #   Column                      Non-Null Count
---  ------                      --------------
 0   POLICY_OWNER_ID             8807 non-null
 1   POLICY_TERM_YEARS           8807 non-null
 2   ANNUAL_FEES                 8807 non-null
 3   POLICY_OWNER_MARITAL_STATUS 8807 non-null
 4   POLICY_OWNER_GENDER         8807 non-null
 5   POLICY_OWNER_ANNUAL_INCOME  7510 non-null
 6   POLICY_OWNER_OCCUPATION_ID  8774 non-null
 7   PRODUCT_TYPE                8807 non-null
 8   POLICY_OWNER_AGE_AT_ISSUE   8807 non-null
 9   TOTAL_LA                    8807 non-null
 10  PAYMENT_MODE                8807 non-null
 11  AGENT_CODE                  8807 non-null
 12  LAPSE_STATUS                8807 non-null
```

Describe how and explain your reasoning

Hypothesis 1:

Missingness in Annual Income is a MNAR.

My approach:

Used Chi-square test to examine

Results of Chi-square:

- **p-value** between annual income and owner age: **0.41** (> 0.05 – alpha)
- **p-valu**e annual income and annual fees: **0.07** (still > 0.05 – alpha)

Conclusion:

In this case, missingness does not belong to type 3 (MNAR). So we can try missingness imputation technique.

Other assumptions were made in my Google Colab to find out why missingness existed.

```
ANNUAL_FEES                 99999.99     100000.00    100000.02    \
POLICY_OWNER_ANNUAL_INCOME
False                              0          2271            2
True                               1           456            5

ANNUAL_FEES                 100000.03    100000.06    100000.08    \
POLICY_OWNER_ANNUAL_INCOME
False                              0             4            3
True                               1             0            0

ANNUAL_FEES                 100000.09    100000.10    100000.11    \
POLICY_OWNER_ANNUAL_INCOME
False                             11             6            6
True                               1             0            0

ANNUAL_FEES                 100000.12    ...  2010000.00   2500000.00   \
POLICY_OWNER_ANNUAL_INCOME               ...
False                              1     ...           1            1
True                               3     ...           0            0

ANNUAL_FEES                 2999999.98   3000000.00   3400000.00   \
POLICY_OWNER_ANNUAL_INCOME
False                              1             2            1
True                               0             0            0

ANNUAL_FEES                 4013256.00   4999999.98   5250042.00   \
POLICY_OWNER_ANNUAL_INCOME
False                              1             1            1
True                               0             0            0

ANNUAL_FEES                 10000000.00  10100000.00
POLICY_OWNER_ANNUAL_INCOME
False                              1             0
True                               0             1

[2 rows x 2732 columns]
The p-value result is: 0.07260012711404985
The missingness in 'POLICY_OWNER_ANNUAL_INCOME' is not MNAR.
```

Result after using MICE imputation

Train set impuatation

	POLICY_OWNER_ID	POLICY_TERM_YEARS	ANNUAL_FEES	POLICY_OWNER_MARITAL_STATUS	POLICY_OWNER_GENDER	POLICY_OWNER_ANNUAL_INCOME	POLICY_OWNER_OCCUPATION_ID	PRODUCT_TYPE
9	A04004	27	100000.000000	M	M	nan	1.000000	TYPE-B
13	A03206	52	100000.000000	M	F	nan	4.000000	TYPE-B
30	A08266	17	149988.230000	M	M	nan	6.000000	TYPE-A
32	A04438	27	100000.000000	M	M	nan	6.000000	TYPE-B
37	A04491	38	100000.000000	M	M	nan	1.000000	TYPE-B
49	A04978	12	200001.000000	M	M	nan	1.000000	TYPE-B
54	A02877	70	100000.000000	M	M	nan	1.000000	TYPE-B
55	A05379	12	150000.000000	M	F	nan	1.000000	TYPE-B
56	A00104	17	100000.000000	M	M	nan	1.000000	TYPE-B
62	A00757	12	119999.880000	M	F	nan	4.000000	TYPE-B
80	A03896	30	100000.000000	S	M	nan	6.000000	TYPE-B
81	A02920	31	300000.000000	S	M	nan	6.000000	TYPE-B
92	A00037	64	100000.020000	M	M	nan	1.000000	TYPE-B
93	A04831	34	200000.000000	M	M	nan	6.000000	TYPE-B
98	A05645	101	113369.660000	M	M	nan	1.000000	TYPE-A

	POLICY_OWNER_ID	POLICY_TERM_YEARS	ANNUAL_FEES	POLICY_OWNER_MARITAL_STATUS	POLICY_OWNER_GENDER	POLICY_OWNER_ANNUAL_INCOME	POLICY_OWNER_OCCUPATION_ID	PRODUCT_TYPE
0	A05948	12.000000	100000.000000	M	M	399996.000000	1.000000	TYPE-B
1	A05490	17.000000	100514.230000	M	F	549996.000000	1.000000	TYPE-A
2	A09094	12.000000	100000.800000	M	M	1299996.000000	1.000000	TYPE-B
3	A05341	12.000000	100000.990000	M	M	500004.000000	1.000000	TYPE-B
4	A08168	12.000000	200001.600000	M	M	499992.000000	2.000000	TYPE-B
5	A05541	12.000000	100000.800000	M	M	298992.000000	1.000000	TYPE-B
6	A00001	29.000000	113125.000000	M	M	25200000.000000	1.000000	TYPE-B
7	A09223	12.000000	144001.080000	M	F	999996.000000	2.000000	TYPE-B
8	A00350	17.000000	180000.000000	M	M	750000.000000	1.000000	TYPE-B
9	A04004	27.000000	100000.000000	M	M	812746.483509	1.000000	TYPE-B
10	A06340	12.000000	100000.000000	M	M	1299996.000000	2.000000	TYPE-B
11	A03595	42.000000	100000.000000	M	M	11127360.000000	1.000000	TYPE-B
12	A02587	22.000000	100000.000000	M	M	800004.000000	1.000000	TYPE-B
13	A03206	52.000000	100000.000000	M	F	812746.763627	4.000000	TYPE-B
14	A08723	98.000000	104045.930000	M	M	1099992.000000	1.000000	TYPE-A

Test set imputation

	POLICY_OWNER_ID	POLICY_TERM_YEARS	ANNUAL_FEES	POLICY_OWNER_MARITAL_STATUS	POLICY_OWNER_GENDER	POLICY_OWNER_ANNUAL_INCOME	POLICY_OWNER_OCCUPATION_ID	PRODUCT_TYPE
0	A01146	9	100000.000000	M	M	nan	1	TYPE-B
1	A05334	12	100000.000000	M	M	nan	1	TYPE-B
2	A05616	12	150000.000000	M	F	nan	1	TYPE-B
3	A04403	12	500000.000000	M	M	nan	1	TYPE-B
4	A05470	12	100534.490000	M	M	nan	6	TYPE-A
5	A00484	12	100000.000000	M	F	nan	4	TYPE-B
6	A02123	12	100000.000000	M	M	nan	1	TYPE-B
7	A04400	12	100000.800000	M	M	nan	1	TYPE-B
8	A01719	12	100000.000000	M	M	nan	1	TYPE-B
9	A05425	12	100002.000000	M	M	nan	6	TYPE-B
10	A05063	12	200000.000000	M	M	nan	1	TYPE-B
11	A04927	12	100000.000000	M	M	nan	1	TYPE-B
12	A05017	12	100000.000000	M	F	nan	1	TYPE-B
13	A01559	12	100000.000000	M	M	nan	1	TYPE-B
14	A05515	12	150000.000000	M	F	nan	1	TYPE-B

	POLICY_OWNER_ID	POLICY_TERM_YEARS	ANNUAL_FEES	POLICY_OWNER_MARITAL_STATUS	POLICY_OWNER_GENDER	POLICY_OWNER_ANNUAL_INCOME	POLICY_OWNER_OCCUPATION_ID	PRODUCT_TYPE
0	A01579	9.000000	149999.880000	M	M	549996.000000	1.000000	TYPE-B
1	A01732	9.000000	100000.000000	M	M	399996.000000	5.000000	TYPE-B
2	A01370	9.000000	100000.000000	M	M	999996.000000	1.000000	TYPE-B
3	A01512	9.000000	100000.000000	S	M	540000.000000	1.000000	TYPE-B
4	A03680	9.000000	100000.000000	M	M	309996.000000	1.000000	TYPE-B
5	A01473	9.000000	200000.280000	S	M	450000.000000	1.000000	TYPE-B
6	A01513	9.000000	100000.000000	M	M	500004.000000	1.000000	TYPE-B
7	A01254	9.000000	100000.800000	M	M	300000.000000	1.000000	TYPE-B
8	A04964	12.000000	100200.720000	M	M	500004.000000	1.000000	TYPE-B
9	A02389	12.000000	150000.000000	M	M	349992.000000	5.000000	TYPE-B

Data Wrangling

2. Standardize data as there are inconsistencies in some columns that manually inputted

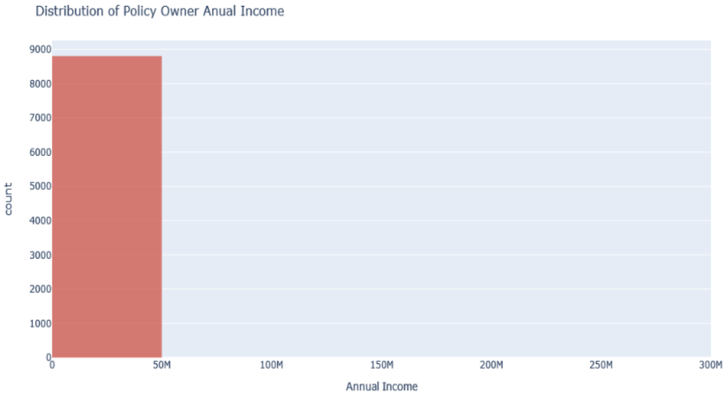
Inconsistencies in Owner_gender in train set

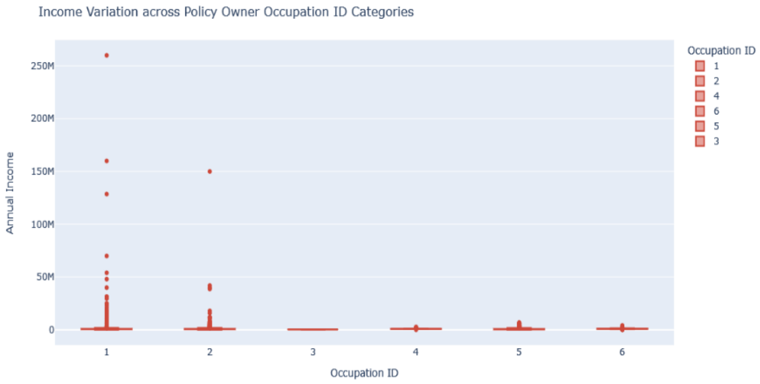
M	5753
F	1901
male	642
MaLe	308
FEmale	203

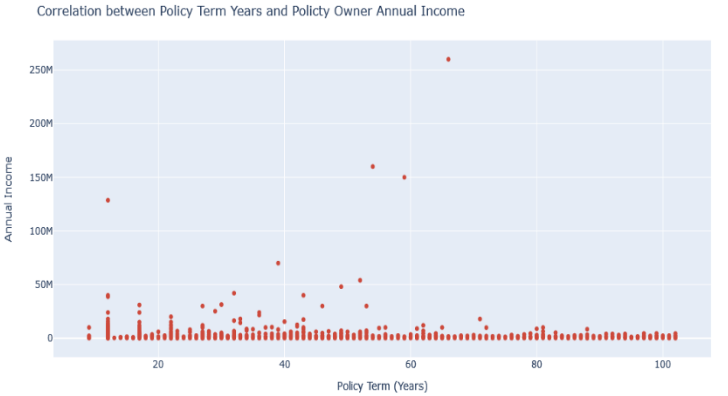
M	6703
F	2104

EDA

1. Income





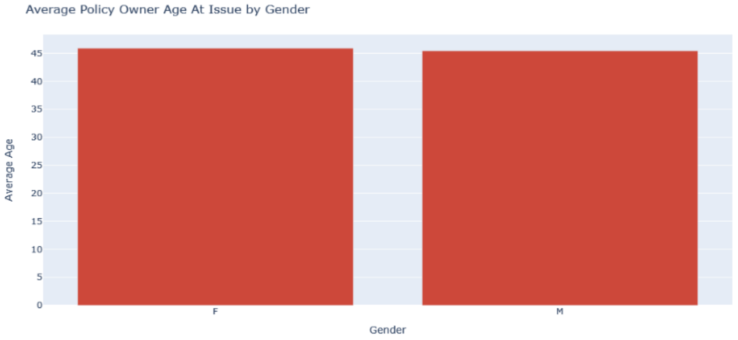


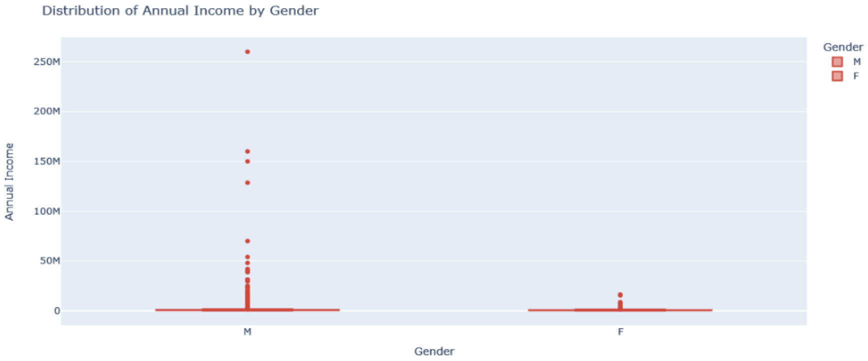
Observations:

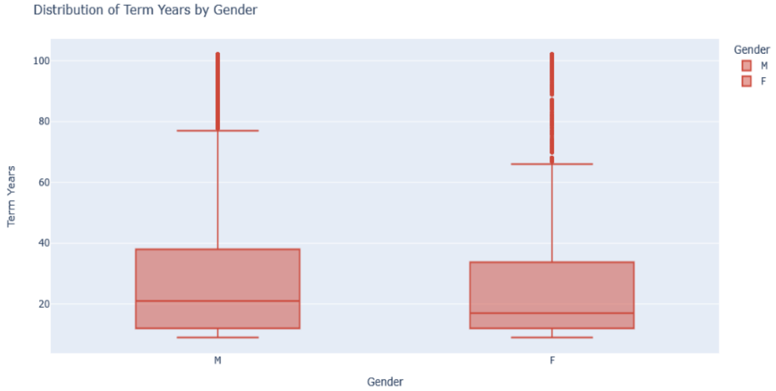
1. **8800** owners have income from **0-50M**
2. The highest annual income is **260M**
3. **Occupation_ID 1** has the highest income
4. No correlation between term years and annual income

EDA

2. Gender





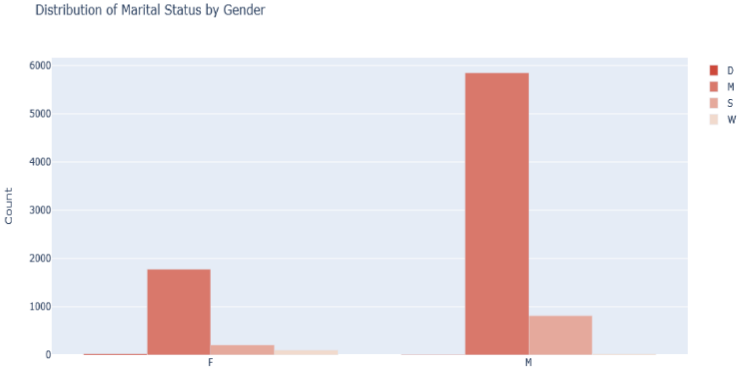


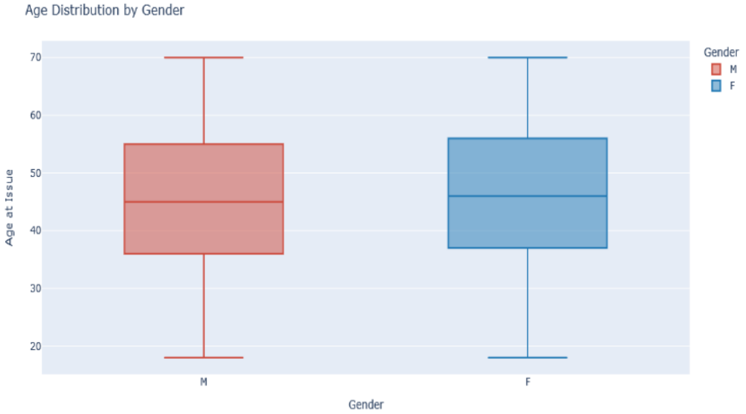
To conclude:

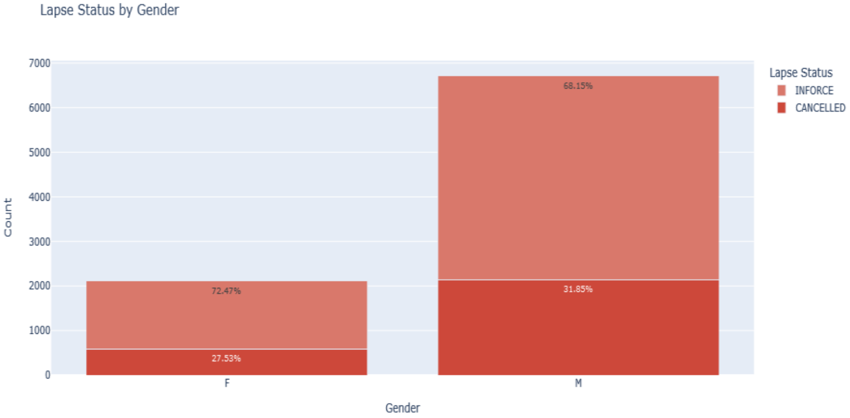
1. There is no different in average age of genders.
2. Male has higher annual income than Female.
3. Male also has more term-year than Female.
4. The highest income (260M) is from Male group.

EDA

2. Gender (Continued)





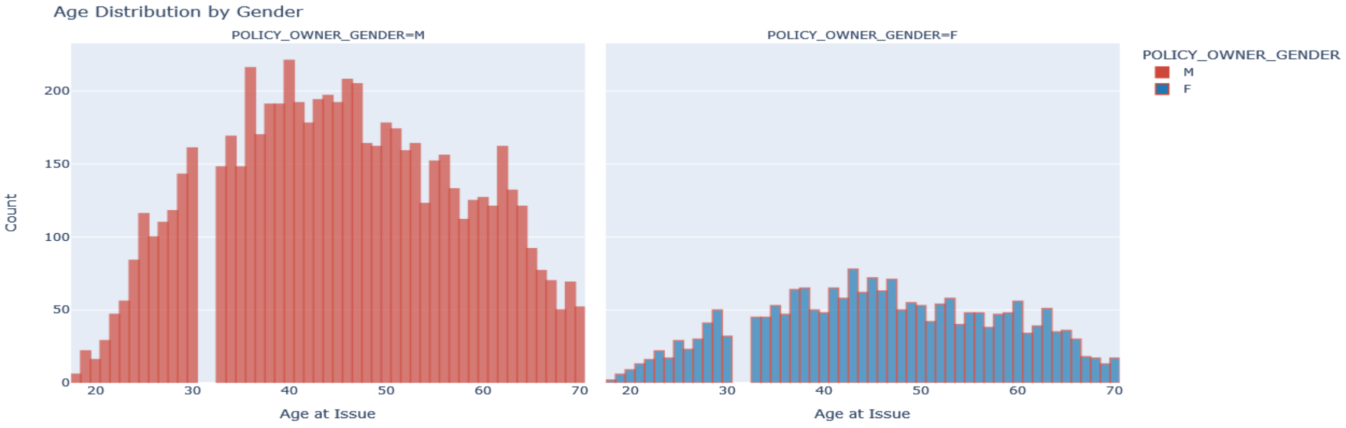


To conclude:

1. Most of owners are married
2. In lapse status, the percentage of cancelled in Male is 31.85%, while Female is 27.35%

EDA

2. Gender (Continued)



Observations about Gender, it is obvious that male seems to use insurance more than Female:

1. The average age between two genders are not much different
2. In term of Annual Income, Male has more Salary than Female
3. This also leads to Male will have more Annual Fees than Female
4. Male also longer term years than Female (the median of Male is 21 5.5. wheares Female is 17)
5. In Marital Status, we have more married male than female
6. For lapse status, Male with Inforce is higher than Female Inforce

Insight

Insight:

A survey shows that women play an important role in life insurance decisions, but they are more likely to buy life insurance for their husbands - the main breadwinner of the family and sometimes for their children first. (Link in References)

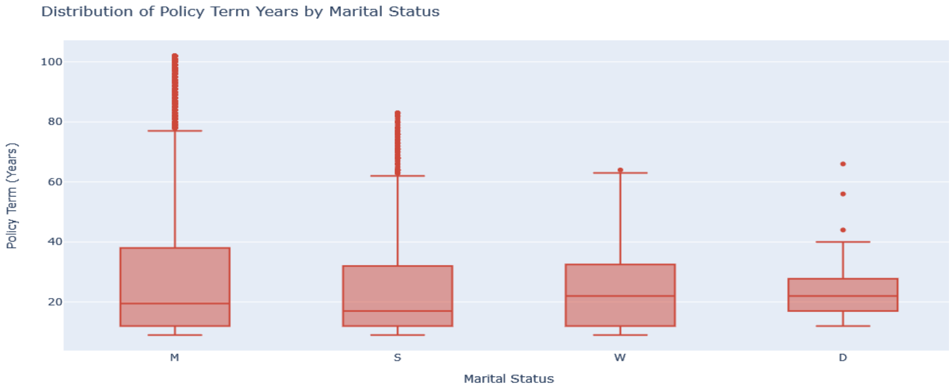
Action:

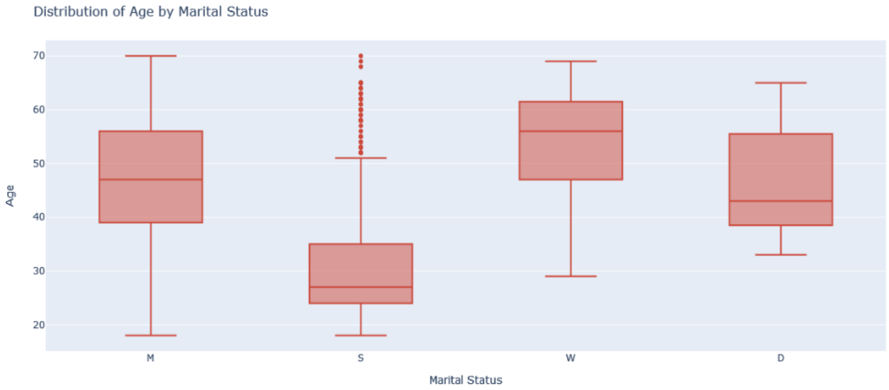
1. **Conduct marketing strategies** (heart touching campagins) to target each gender specifically. This can involve offering different insurance products or packages that align with their income levels and financial goals.
2. **Focus on exploiting the elderly group** in women using life insurance, this can be done by:

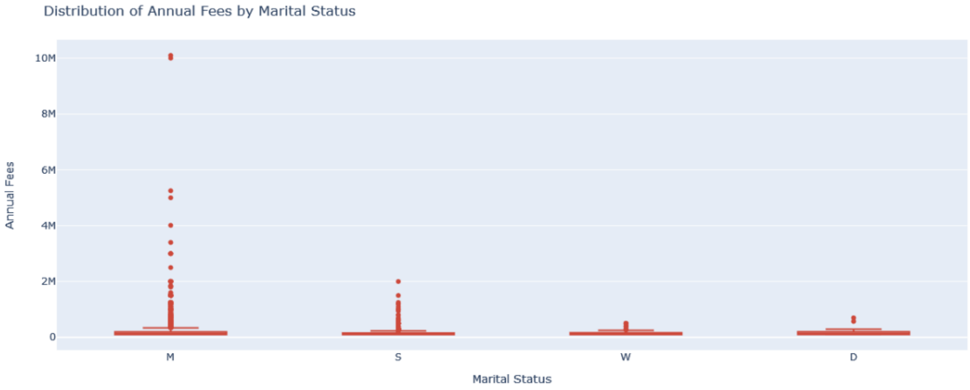
- Diversify insurance distribution channels for Female
- Change people's perception of life insurance (The perception that life insurance is only for the breadwinner is changing and tends to recognize the intangible economic value of mothers (housewives, housekeepers, parenting...)

EDA

3. Mariage status







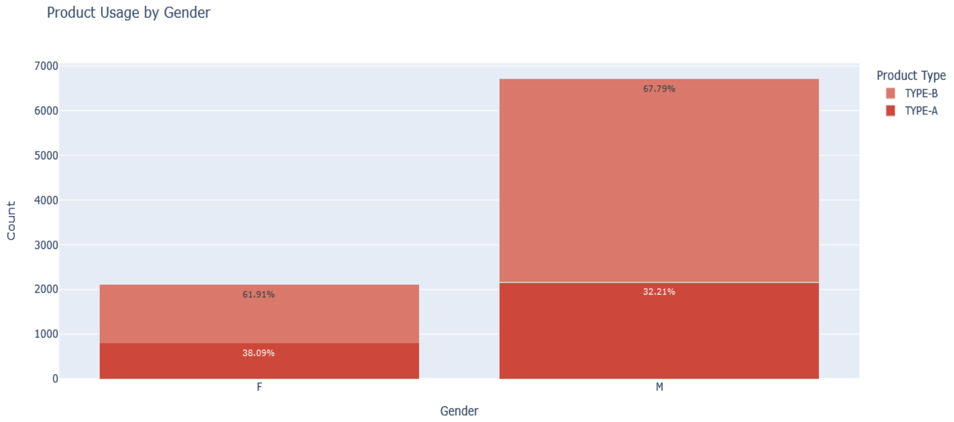
To conclude:

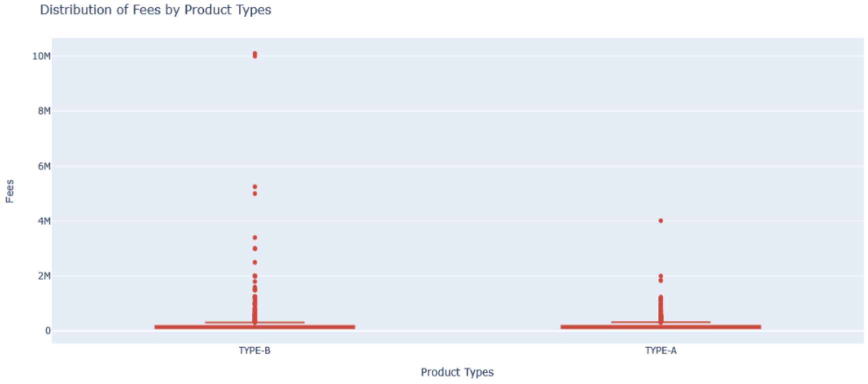
1. Married status has longer term-year than others
2. Married status generated more fees (highest at 10M/year)
3. The average age:

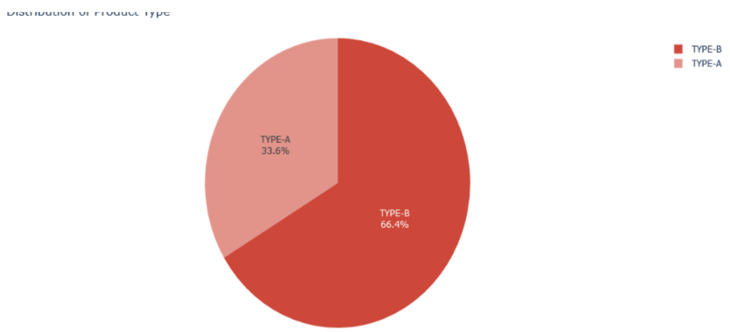
- Married: 47
- Single: 27
- Widowed: 56
- Divorced: 43

EDA

4. Product type





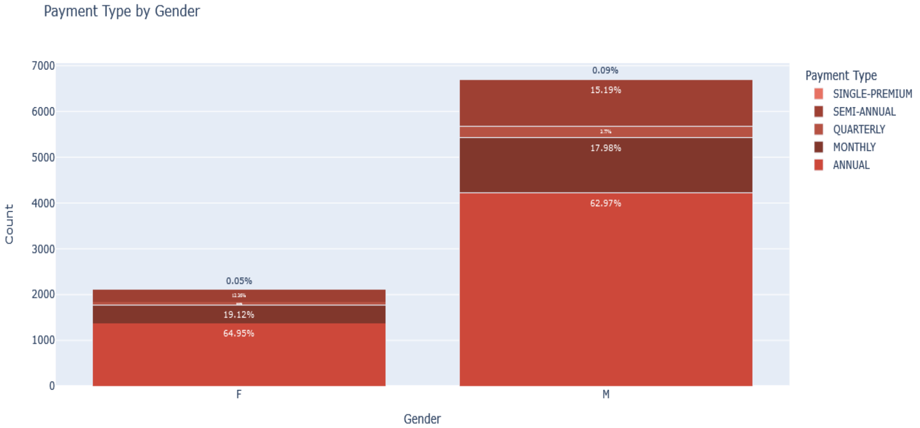


Observations:

1. Owners prefer Product Type B more than Type A
2. In **male** gender, the percentage of using Product Type B is **~68%** while Type A only has **32%**
3. In **female** gender, the percentage of using Product Type B is **~62%** while Type A only has **38%**.
4. Product Type B generated more fees than type A

EDA

5. Payment Mode



Observations:

1. Owners prefer annual payment the most
2. There is no significant difference between payment modes of genders

Statistical Analysis

a. Perform a hypothesis test to determine if there is a significant difference in

POLICY_OWNER_ANNUAL_INCOME between customers' gender? (significance level = 0.05)

Situation (Null hypothesis):

There is a significant difference in POLICY_OWNER_ANNUAL_INCOME between customers' gender (Null hypothesis: alpha = 0.05)

Tasks:

- Filter the data by gender
- Extract the annual income for each gender
- Perform two-sample t-test

Approach:

- Check if the p-value is less than the significance level

Result:

- t_stat: 4.777274088084046
- p_value: 0.0000018056030932618221

Action: p_value < alpha (< 0.05)

- Reject the null hypothesis, which means There is a significant difference in POLICY_OWNER_ANNUAL_INCOME between genders.

Statistical Analysis (continued a)

Try one more Hypothesis Testing with Mann-Whitney-U approach:
Because this approach works better with non-parametric (phi tham số), compares two sample means that come from the same population, and used to test whether two sample means are equal or not.

Situation (Null hypothesis):

There is a significant difference in POLICY_OWNER_ANNUAL_INCOME between customers' gender (Null hypothesis: alpha = 0.05)

Tasks:

- Perform Mann-Whitney U test

Approach:

- Check if the p-value is less than the significance level

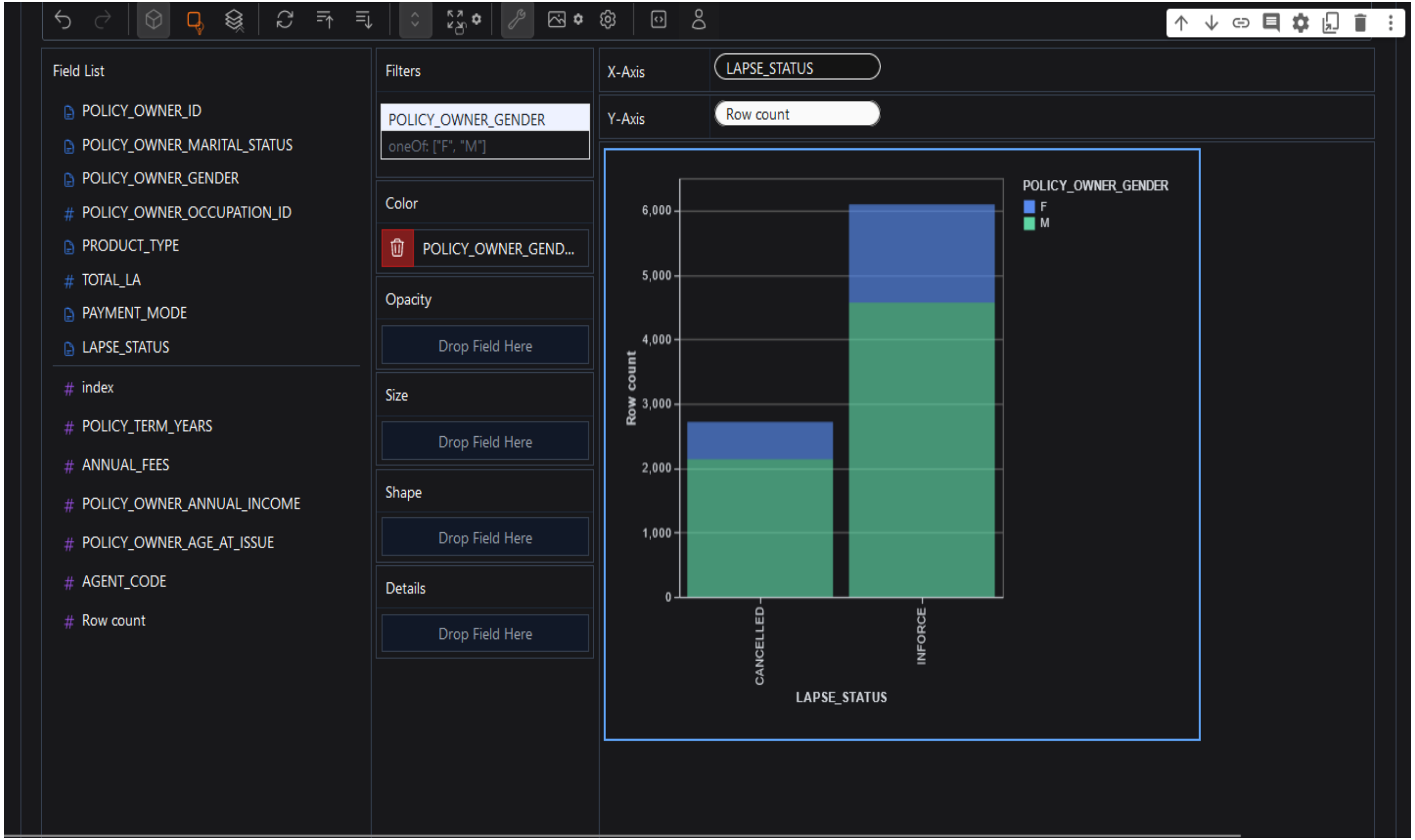
Result:

- t_stat: 8606129.0
- p_value: 5.089154882272343e-53

Action: p_value < alpha (< 0.05)

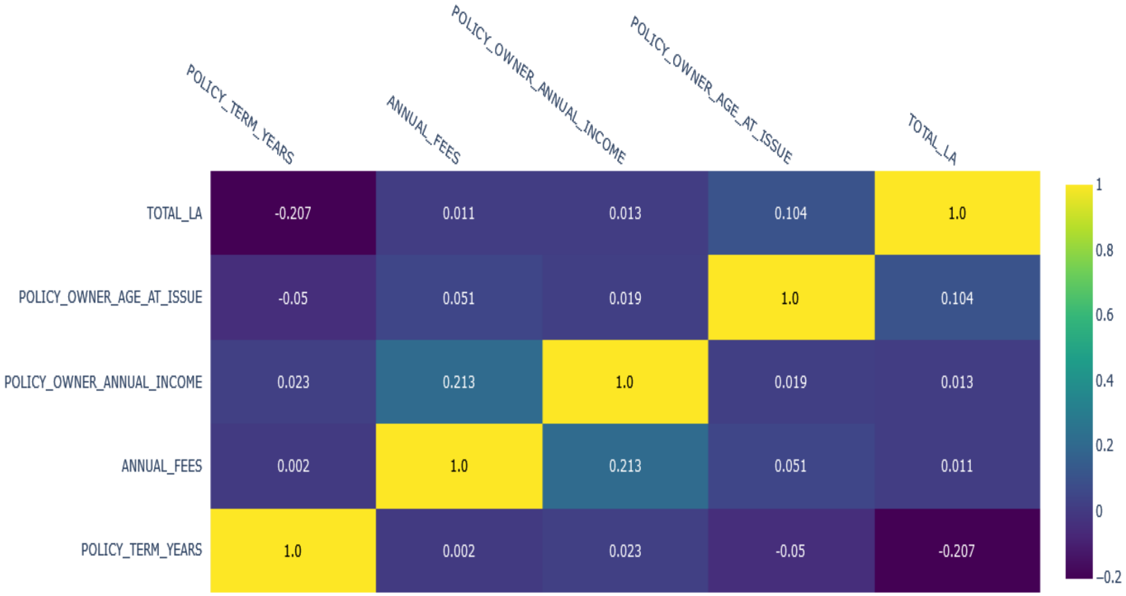
- Reject the null hypothesis, which means There is a significant difference in POLICY_OWNER_ANNUAL_INCOME between genders.

Data visualization with Pygwalker



Advanced Analytics: Building a classification model to predict a customer who is likely to lapse based on the provided dataset

Selecting the appropriate features into your model. Describe my approaches



Features selection:

By plotting the **correlation heatmap** with *Pearson* method, we can see:

- annual income is highly correlated to annue fees.
- Term-year has negative correlation with total lives assured.
- Age at issue and annual fees also have high correlation compared to other

By using techniques **Feature Importance, RFE, RFECV**
I decided to select:

- **Term-year**
- **Gender**
- **Annual Income**
- **Age**
- **Total LA**

Feature selection with business domain knowledge

To know who is likely to lapse in insurance, we will have to find out what the factors that motivate owners to purchase insurance are. Based on some scientific research that I have studied, I believe these factors (features) are also important in our models.

Tasmin, J. & Md. Mahiuddin, S. (2018) has proven that 03 factors are demographics (including: **age, gender, marital status**, number of dependents, education, **income**)
Nguyễn Thị Ánh Xuân (2004) has found out 05 factors that affect the buying trend of the group of customers who have not bought insurance insurance: the level of support of parents, the level of support of spouses, benefitsmorale, investment benefits and ultimately protective interests.
Li, 2008; Ćurak & partners, 2013; Shiferaw, 2017: **Gender** shows a positive correlation with the decision to participate in insurance insurance, whereby women tend to be more interested in and participate in insurance contracts than men.

Decision to buy life insurance and policy value

Demographic factor groups:

Gender
Marital status
Dependent lives (lives assured)
Income
Age

Social factor groups:

Influence from relatives
Business investment (real estate)
Quality of insurance

Research model of factors influencing the decision to buy insurance

Build models for classification

Overall, we will determine what tasks needed to build good performance models

Important tasks (Preprocessing):

- Label Encoder
- Upsampling data (imbalanced data)
- **Features selection (task required)**
- Scale data (Min-max scaler)

After selecting the right features and scaling data, this is our final data look-like for building classification models.

ANNUAL_FEES	POLICY_OWNER_MARITAL_STATUS	POLICY_OWNER_GENDER	POLICY_OWNER_ANNUAL_INCOME	POLICY_OWNER_OCCUPATION_ID	PRODUCT_TYPE	POLICY_OWNER_AGE_AT_ISSUE	TOTAL_L
2.226120e-04	0.333333	1.0	0.001692	0.2	0.0	0.673077	1.00000
1.000000e-09	0.333333	1.0	0.001885	0.0	1.0	0.346154	0.66666
2.369205e-03	0.333333	0.0	0.001115	0.0	0.0	0.365385	0.33333
2.007927e-03	0.333333	1.0	0.001115	0.0	0.0	0.403846	0.33333
3.089850e-04	0.333333	1.0	0.023039	0.0	0.0	0.500000	0.66666

Modeling

Here are the models that I choose to predict classification:
XGBClassifier, AdaBoostClassifier, CatBoostClassifier, Random Forest Classifier, Decision Tree Classifier.

Tasks:

- Fit models
- Predict
- Evaluation (accuracy)
- Combine y_predicts from models (majority voting, probabilities by averaging)
- Print result

Model	Accuracy
XGBoostClassifier	82.97%
AdaBoostClassifier	60.42%
CatBoostClassifier	86.80%
RandomForestClassifier	83.40%
DecisionTreeClassifier	66.38%

Models performance

Since we don't have true labels in test set, we cannot perform evaluations like: Confusion matrix, f1-score.. So I choose a method that combines y_predict from those models to get a better result

Apply a combination by majority voting

	POLICY_OWNER_ID	LAPSE_STATUS
0	A01579	0
1	A01732	0
2	A01370	1
3	A01091	1
4	A01311	1
5	A01512	1
6	A04153	1
7	A03680	1
8	A01146	1
9	A01473	0
10	A01513	1
11	A01254	0
12	A04964	1
13	A04625	0
14	A02418	1
15	A02389	0
16	A07512	0
17	A02010	1
18	A06816	1

In conclusion

After knowing which owners would likely to lapse, we should also have some actionable decisions

It is important to know, we just predicted based on demographic factors, whereas social factors also play a key role in determining the lapse status of our owners. We should prepare some actions for different type of owners.

Action:

- **Conduct marketing strategies** (heart touching campagins) to target each gender specifically. This can involve offering different insurance products or packages that align with their income levels and financial goals.
- **Focus on exploiting the elderly group** in women using life insurance, this can be done by:

  - Diversify insurance distribution channels for Female

  - Change people's perception of life insurance (The perception that life insurance is only for the breadwinner is changing and tends to recognize the intangible economic value of mothers (housewives, housekeepers, parenting...)

- **Strengthen consultancy activities** on the benefits of insurance insurance and guide the process and procedures for implementing insurance contracts, transparent public activities, diversify product packages suitable to the income of each customer group.
- **Expand distribution channels** (Bancassurance) with many banks, because this distribution model is considered as the dominant market channel
- **In the case of low-income pensioners** (from pensions), research to develop co-interactive models, support with participation from insurance enterprises, non-profit organizations in and overseas.

References

- Các yếu tố ảnh hưởng đến ý định mua bảo hiểm nhân thọ của người dân: trường hợp nghiên cứu tại quận Gò Vấp, Thành phố Hồ Chí Minh của Phạm Xuân Giang, Nguyễn Thị Thu Hằng (Tạp chí Khoa học và Công nghệ, Số 54, 2021).
- Nguyễn T. B. Minh và cộng sự. HCMCOUJS-Kinh tế và Quản trị Kinh doanh, 16(2), 156-169, Determinants affecting the buying decision in life insurance of individual customers in Ho Chi Minh City.
- Phân tích quyết định tham gia bảo hiểm nhân thọ của người hưu trí tại thành phố Cần Thơ (Analysis of the life insurance purchase decision of retirees in Can Tho city) Huy, H. T., Dung, N. T. K., Uyên, N. M., & Liêm, N. T. (2020). Phân tích quyết định tham gia bảo hiểm nhân thọ của người hưu trí tại thành phố Cần Thơ.
- Do, A. H., & Pham, M. H. (2019). Nhân tố ảnh hưởng đến quyết định tham gia BHNT của người dân tại Thành phố Quảng Ngãi [Factors affecting the decision to participate in life insurance of people in Ho Chi Minh City. Quang Ngai]. Retrieved May 12, 2020, from Tạp chí Tài chính Onine website: http://tapchitaichinh.vn/nghien-cuu-trao-doi/nhan-to- Nguyễn T. B. Minh và cộng sự. HCMCOUJS-Kinh tế và Quản trị Kinh doanh, 16(2), 156-169 169 anh-huong-den-quyet-dinh-tham-gia-bao-hiem-nhan-tho-cua-nguoi-dan-tai-tp-quang-ngai302400.html.
- Nguyen, T. T., & Nguyen, N. V. (2015). Các nhân tố ảnh hưởng đến quyết định mua BHNT của công ty BHNT Prudential Việt Nam [Factors affecting the decision to buy life insurance of Prudential Vietnam life insurance company]. Tạp chí Khoa học Công nghệ - Thủy sản, 1, 185-192
- Nguyễn Hữu Quỳnh Như / Tạp chí Khoa học và Công nghệ Đại học Duy Tân 4(47) (2021) 102-108

The end
Thank you