# Report on Spain Consumer Life Insurance

In this report, we outline the process of analysis, show patterns found, and share the important findings.

## Approach

**Data Description:** The data is on life-risk insurance policies in force on December 31, 2009, of consumers in Spain for a company. This is 76102 consumers' policies and has 6 primary variables. There are no duplicate records. All IDs are unique, interpreted as no consumer has 2 policies.

```
RangeIndex: 76102 entries, 0 to 76101
Data columns (total 15 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   ID                  76102 non-null  int64
 1   Gender              76102 non-null  object
 2   Birth_Date          76102 non-null  datetime64[ns]
 3   Effecitive_Date     76102 non-null  datetime64[ns]
 4   Capital             76102 non-null  float64
 5   Renewal_Date        76102 non-null  datetime64[ns]
 6   Age                 76102 non-null  float64
```

**Objective of Study:** The first objective of the study is to find general patterns and insights in the data. The second objective is to be able to predict the preferable Capital to offer a consumer for a policy based on features available in the data.

**Assumption:** The data generated is an outcome of the decision-making of the consumer.

**Approach:** To achieve the objectives of the study we first do exploratory data analysis using univariate and bivariate analysis. Then we run a few machine learning models to probe further. We start with the primary variables in the data and construct new variables (as defined below) based on the objective of the study and as and when the iterative process of analysis suggests so.

**Defined Variables:** Based on the objective of the study and taking clues from the process of exploratory data analysis we construct a few variables listed below.

*PolicyStartAge:* The variable is the age of the consumer at the time when the policy was issued (effective date). We chose this defined variable for analysis over the primary variable Age (highly correlated) because we are interested in description and prediction at the time of policy issue rather than risk evaluation at any given point in time where Age will play a more relevant role.

*PolicyMonth:* This is the calendar month derived from the policy effective date which is when the policy is issued. This is to study any seasonality in policy issuance.

*BirthMonth:* This is the calendar month derived from the birthday date. This is to study any seasonality based on when in the year the birthday of the consumer falls.

*PolicyQuarter:* This is the calendar (not financial) quarter of the year derived from the variable PolicyMonth. This is to study if grouping monthly data into quarters can show a clearer pattern.

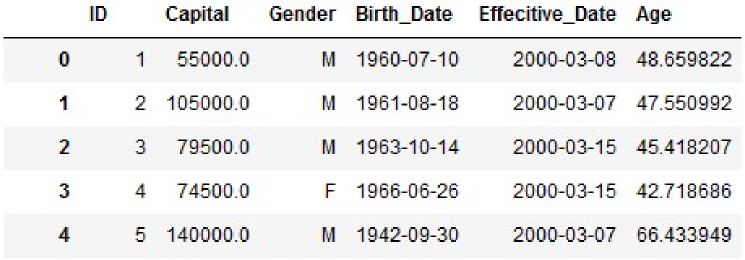
*BirthQuarter:* This is the calendar (not financial) quarter of the year derived from the variable BirthMonth. This is to study if clubbing monthly data into quarters can show a clearer pattern.

*PBQuarter:* This is the difference of months between the PolicyMonth and BirthMonth that is after how many months the policy got issued after the last birthday of the consumer. This is to study if there is a presence of birthday sentiment in the decision-making of the consumer. This is because we suspect that birthdays, like some other events e.g. festivals, bringing family and friends closer can invoke the importance of life nudging consumers to opt for life insurance.

## Exploratory Data Analysis

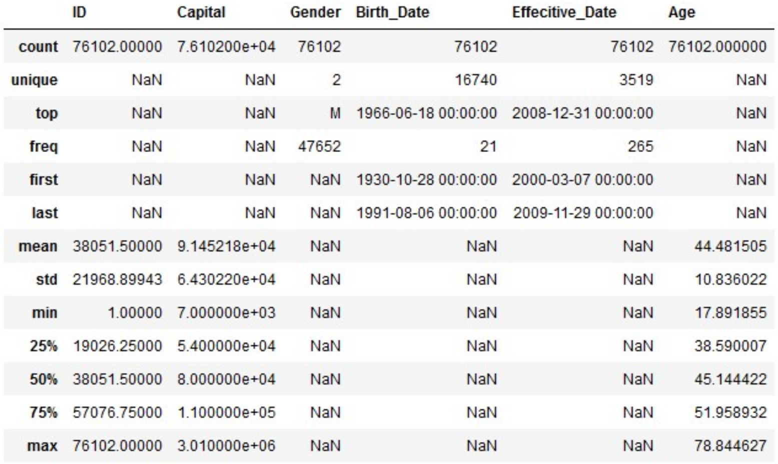
We started with considering primary variables for our study and constructed any derived variable we saw a need for.

*A look at sample data with primary variables.*



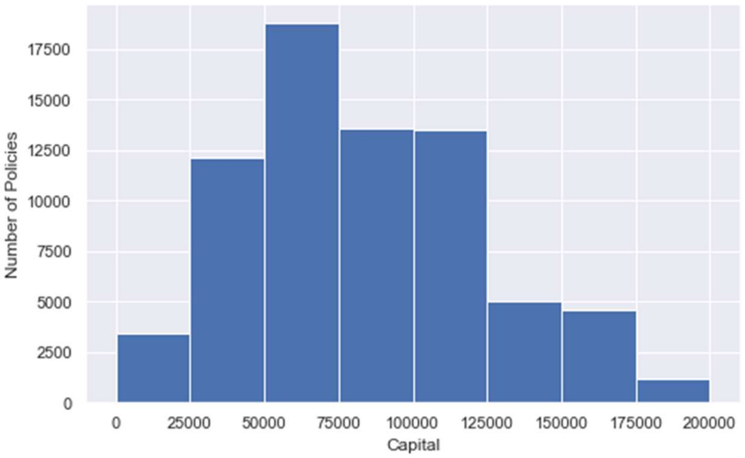
|  | ID | Capital | Gender | Birth_Date | Effecitive_Date | Age |
|---|---|---|---|---|---|---|
| **0** | 1 | 55000.0 | M | 1960-07-10 | 2000-03-08 | 48.659822 |
| **1** | 2 | 105000.0 | M | 1961-08-18 | 2000-03-07 | 47.550992 |
| **2** | 3 | 79500.0 | M | 1963-10-14 | 2000-03-15 | 45.418207 |
| **3** | 4 | 74500.0 | F | 1966-06-26 | 2000-03-15 | 42.718686 |
| **4** | 5 | 140000.0 | M | 1942-09-30 | 2000-03-07 | 66.433949 |

*Summary statistics of the primary variables are as below.*

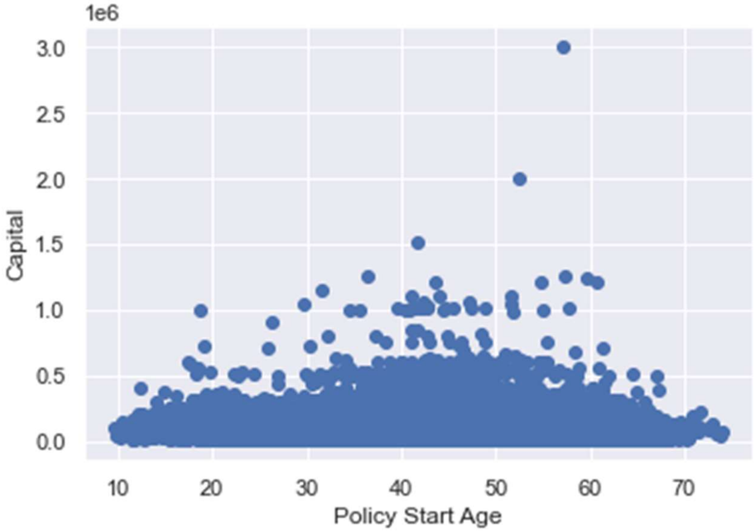


|  | ID | Capital | Gender | Birth_Date | Effecitive_Date | Age |
|---|---|---|---|---|---|---|
| **count** | 76102.00000 | 7.610200e+04 | 76102 | 76102 | 76102 | 76102.000000 |
| **unique** | NaN | NaN | 2 | 16740 | 3519 | NaN |
| **top** | NaN | NaN | M | 1966-06-18 00:00:00 | 2008-12-31 00:00:00 | NaN |
| **freq** | NaN | NaN | 47652 | 21 | 265 | NaN |
| **first** | NaN | NaN | NaN | 1930-10-28 00:00:00 | 2000-03-07 00:00:00 | NaN |
| **last** | NaN | NaN | NaN | 1991-08-06 00:00:00 | 2009-11-29 00:00:00 | NaN |
| **mean** | 38051.50000 | 9.145218e+04 | NaN | NaN | NaN | 44.481505 |
| **std** | 21968.89943 | 6.430220e+04 | NaN | NaN | NaN | 10.836022 |
| **min** | 1.00000 | 7.000000e+03 | NaN | NaN | NaN | 17.891855 |
| **25%** | 19026.25000 | 5.400000e+04 | NaN | NaN | NaN | 38.590007 |
| **50%** | 38051.50000 | 8.000000e+04 | NaN | NaN | NaN | 45.144422 |
| **75%** | 57076.75000 | 1.100000e+05 | NaN | NaN | NaN | 51.958932 |
| **max** | 76102.00000 | 3.010000e+06 | NaN | NaN | NaN | 78.844627 |

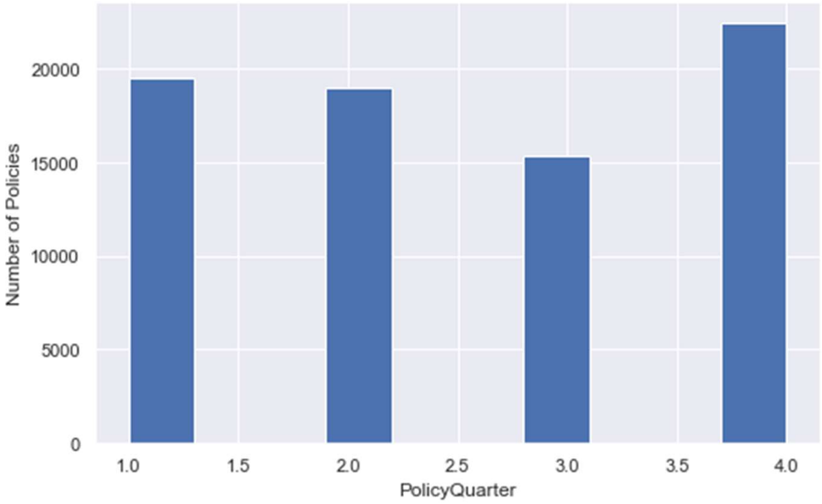
*Capital Frequency:* Most of the policies are in the range of 25k-125k with the range that has the highest number of policies being 50k-75k



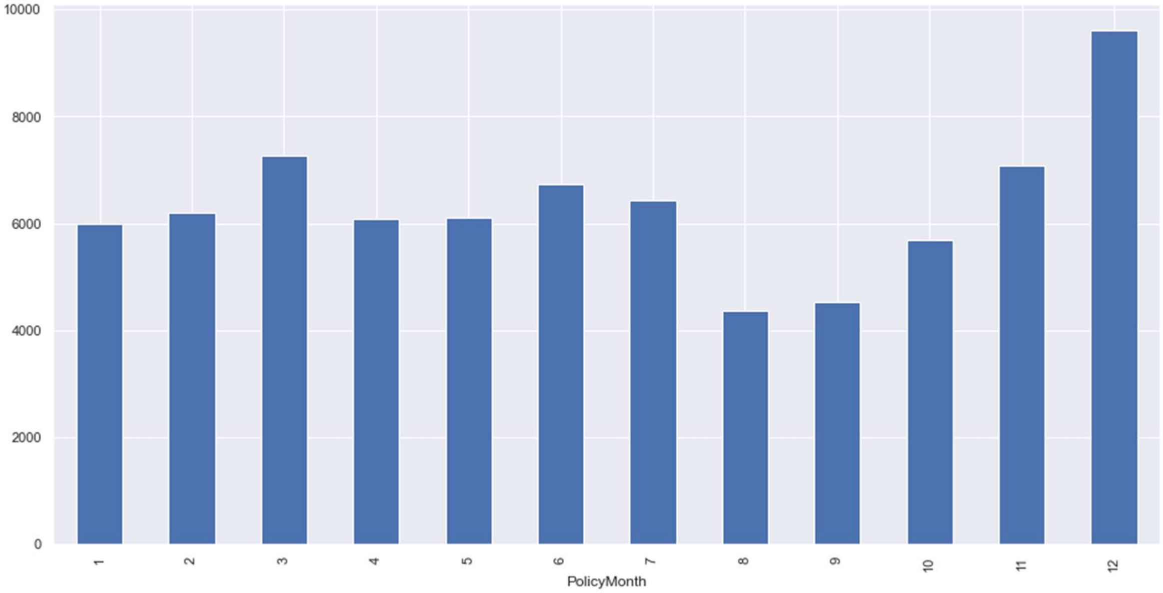
*Capital and Policy Start Age joint scatter plot:*



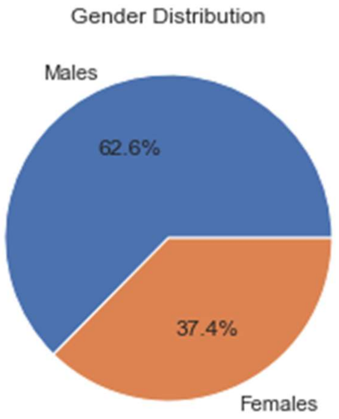
*Policy frequency quarter-wise:*



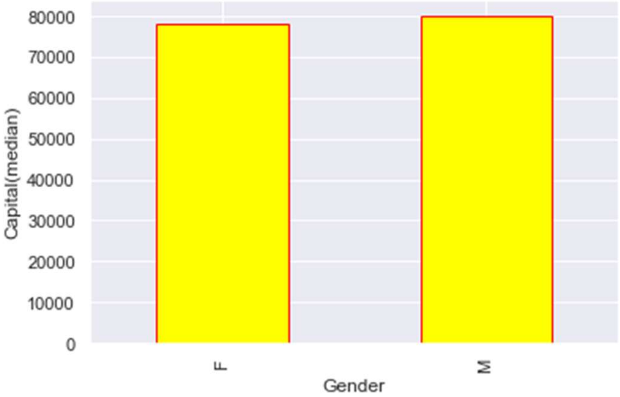
*Policy frequency month-wise:*



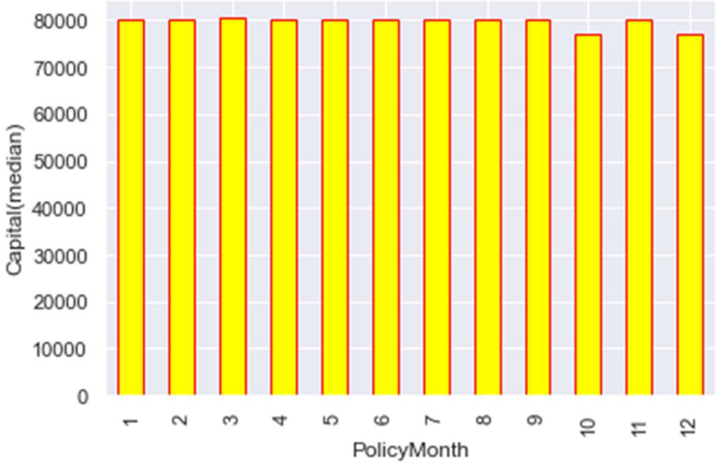
*Proportion of policies for two genders:*



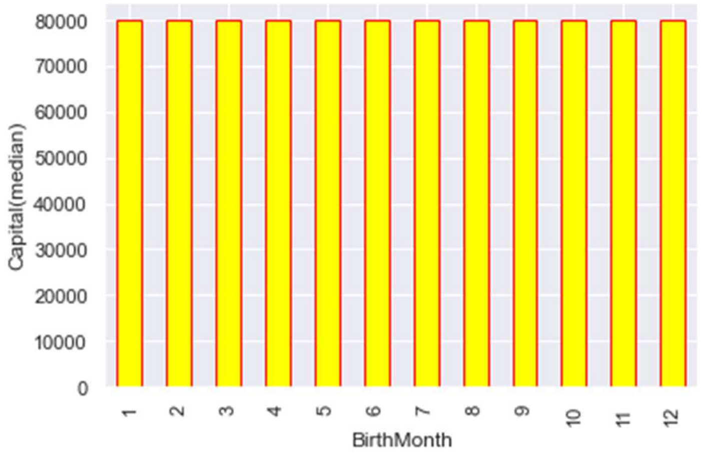
*Median Capital for two genders:* We observe that the median capital for males is slightly more than that of females.



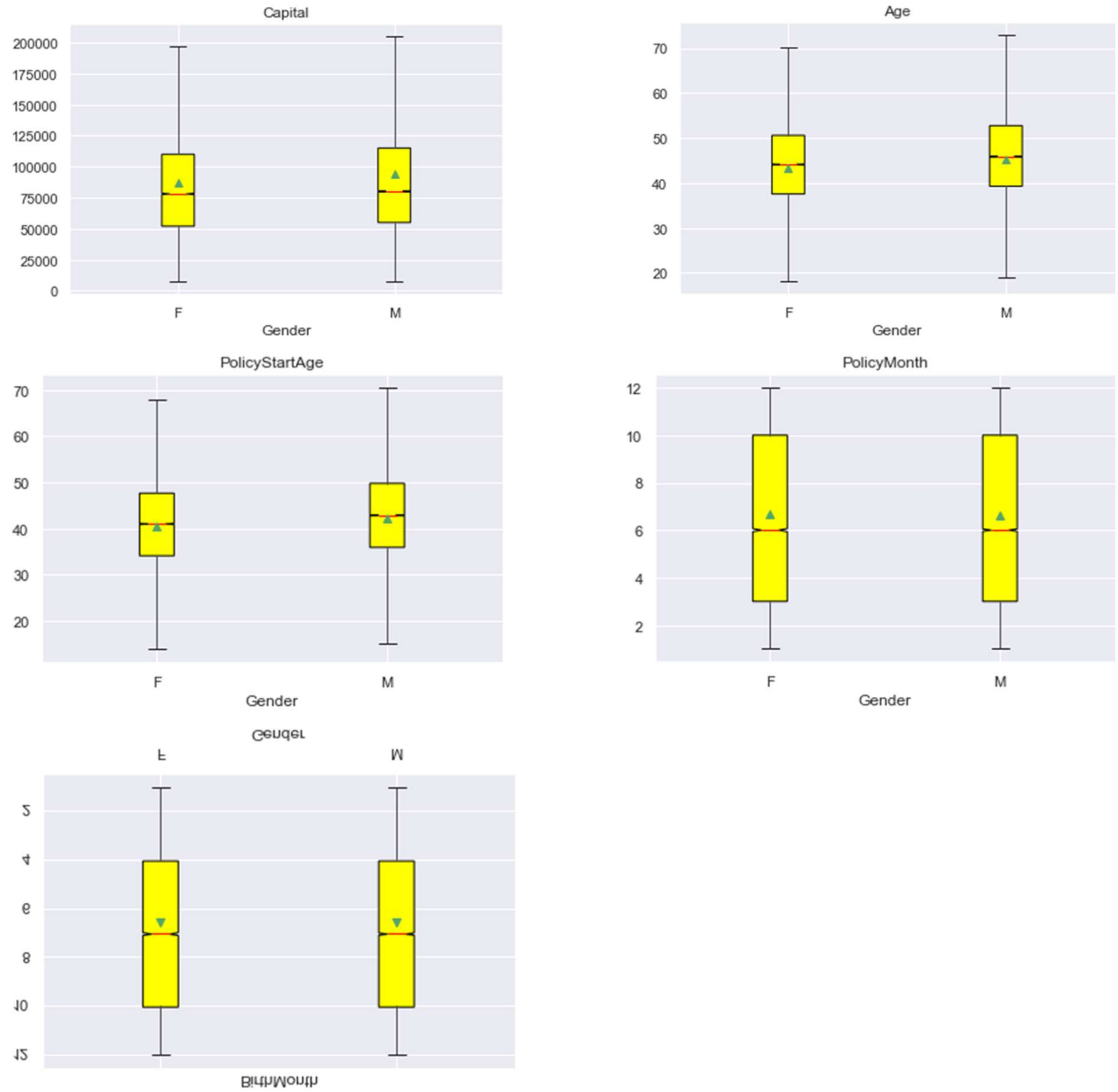
*Median Capital for each policy month:* We observe that there appears to be some seasonality in the last quarter based on when the policy was issued.



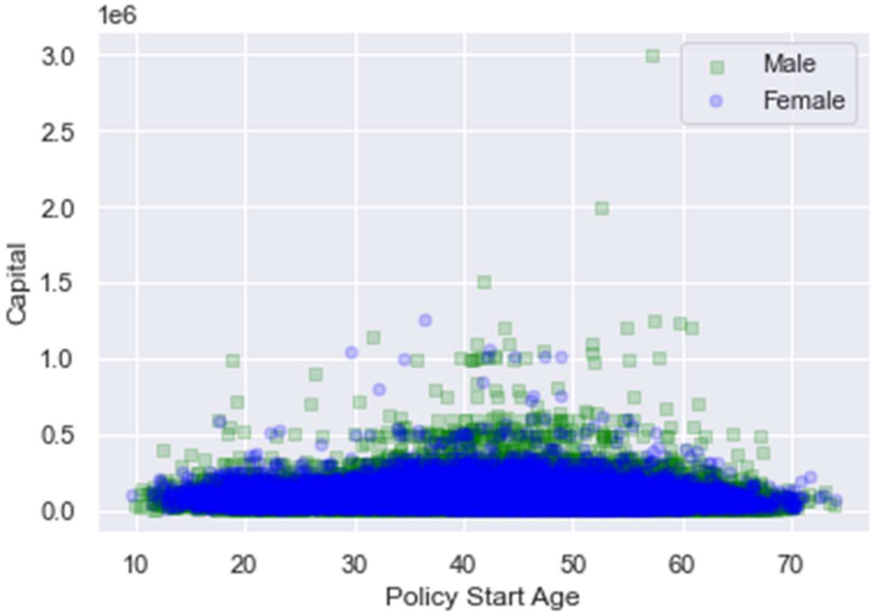
*Median Capital for each birth month:* We observe that there is no seasonality based on in which month the consumer is born.



*Gender Summary Statistics Boxplots:* We partition the dataset based on gender to see if these have different distributions of variables and need to be analyzed separately.

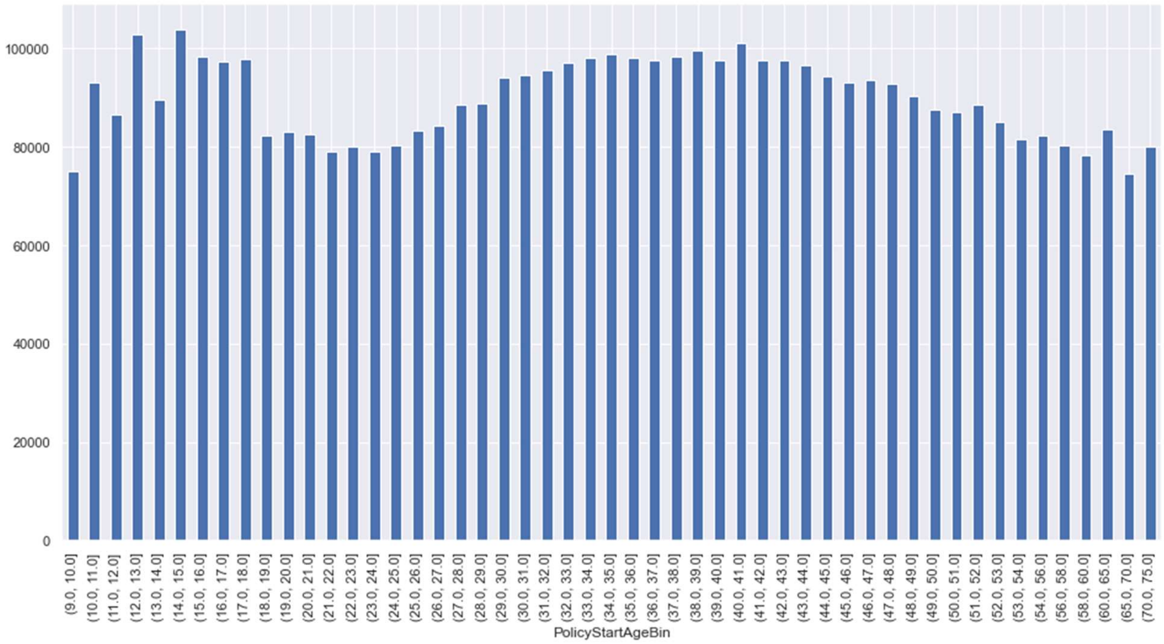


We further look at the two important variables Capital and Policy Start Age. The joint distribution as observed in the scatter plot below shows that these are similar.



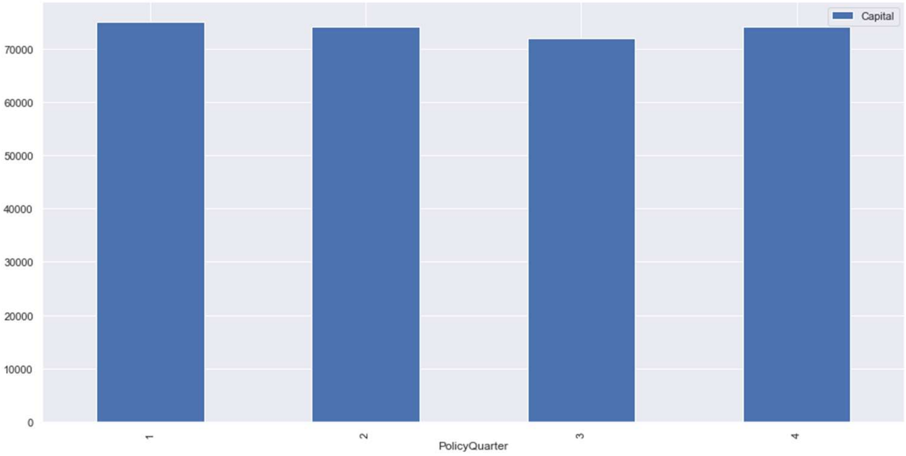
Based on the above charts we conclude that male and female data can be analyzed together as variables in these datasets appears to have similar relationships among them.

*Capital and Policy Start Age:* Probing the joint distribution of Capital with Policy Start Age further finds that the data can be partitioned into 3 data sets with the Policy Start Age cutoffs as 24 and 40. We will name these as data1: Students, data2: Young Professionals, data3: Mature Professionals. The nomenclature is based on what majority (not all) of the population is thought to be in that cohort.

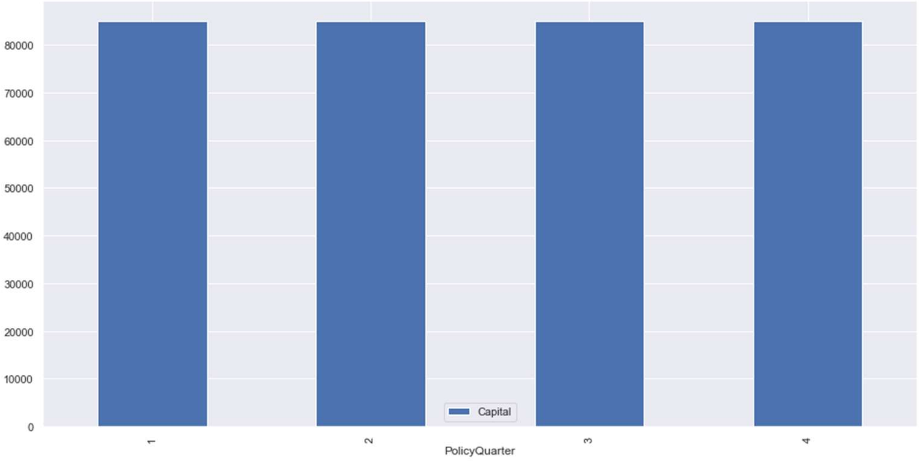


*Median Capital and Policy Quarter for three segments:* The Students and Mature Professionals show some quarterly seasonality but Young Professionals remain at the same amount of policy Capital

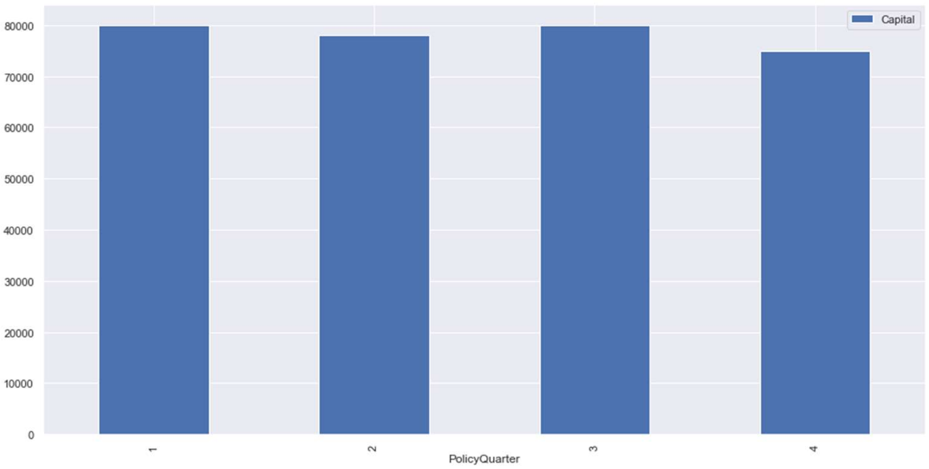
Students



Young Professionals

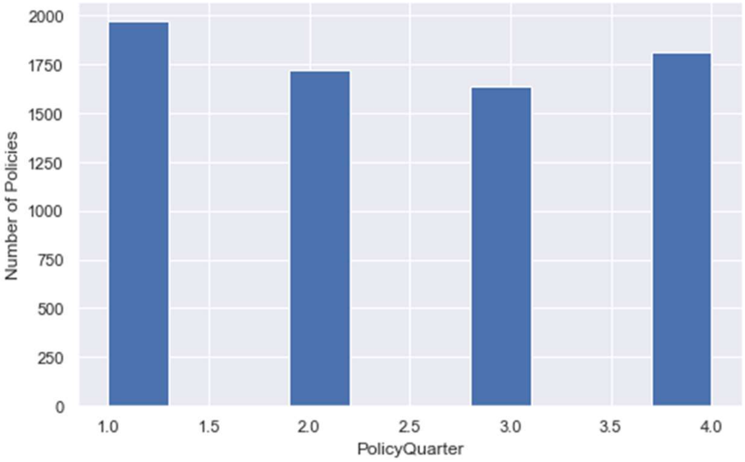


Mature Professionals

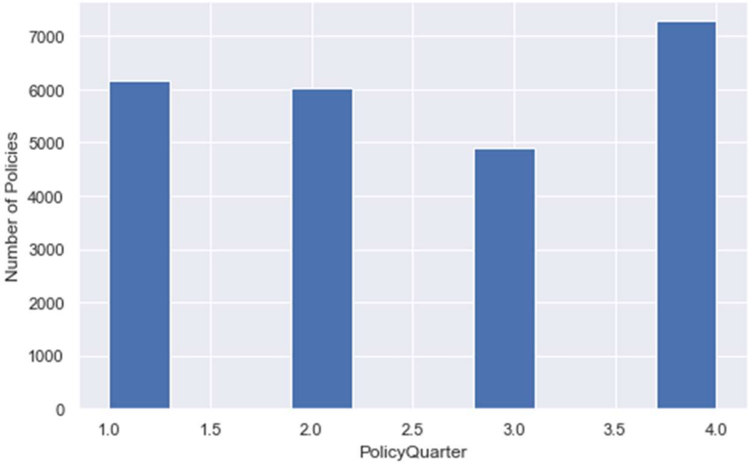


*Policy frequency for three segments quarter-wise:* The summer observes the minimum number of policies in all the segments. Professionals have a maximum number of policies issued in the last quarter while Students do in the first quarter.

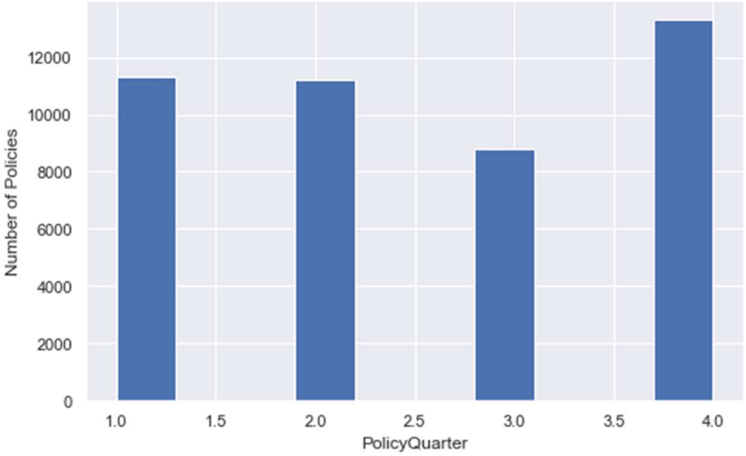
Students



Young Professionals

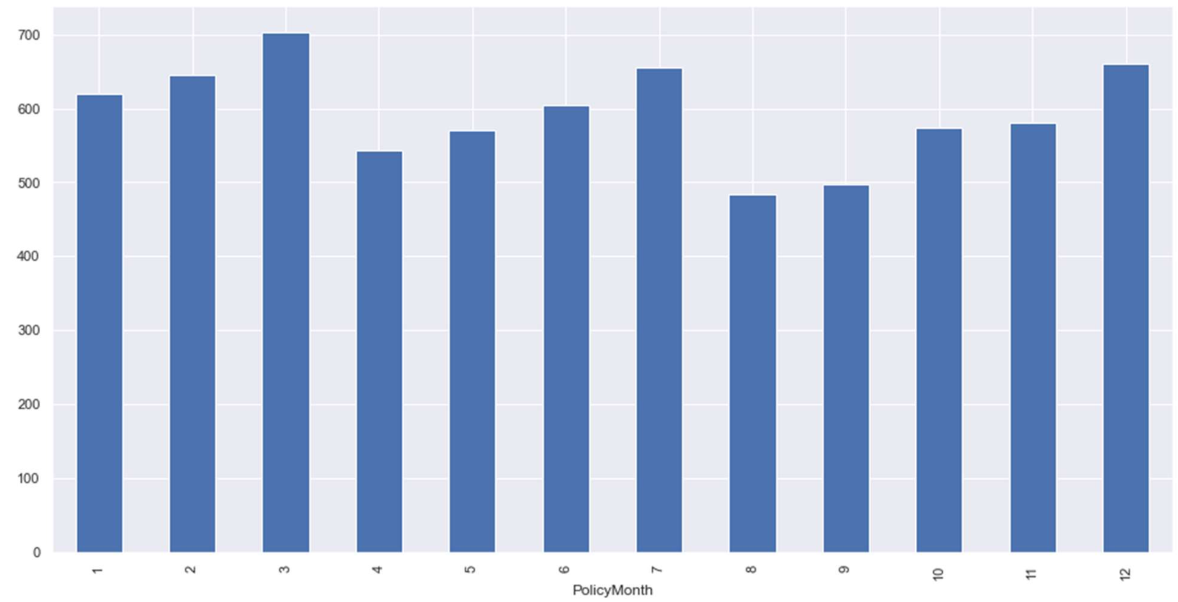


Mature Professionals

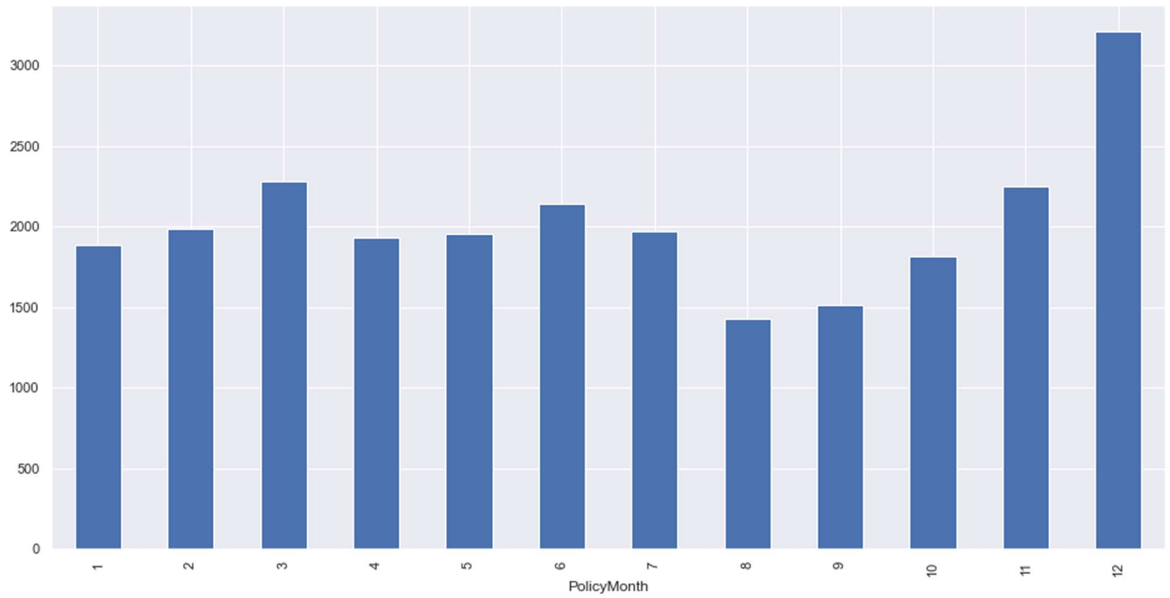


*Policy frequency for three segments month-wise:* When we zoom in on the quarterly seasonality, we observe that there are two periods of trend for all the segments and one period can have a trend or may remain roughly flat: 1) from Jan that peaks in March (financial year end and close to Easter) 2) dipping in April then increasing and peaking in June/July or remaining flat 3) dipping in August then increasing and peaking in December (holiday and Christmas).

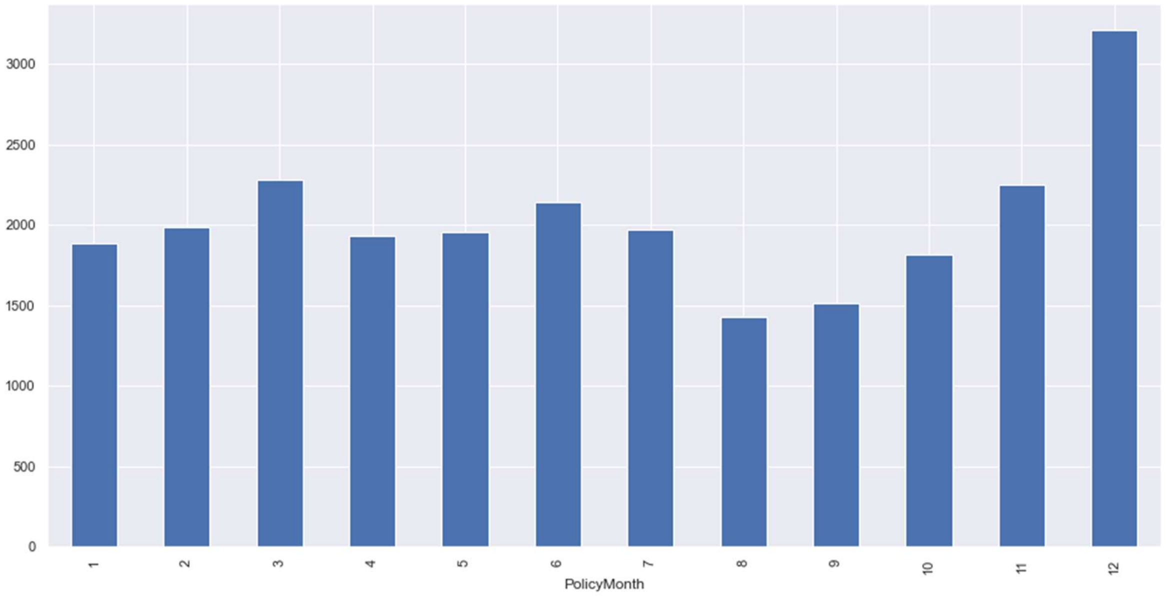
Students



Young Professionals

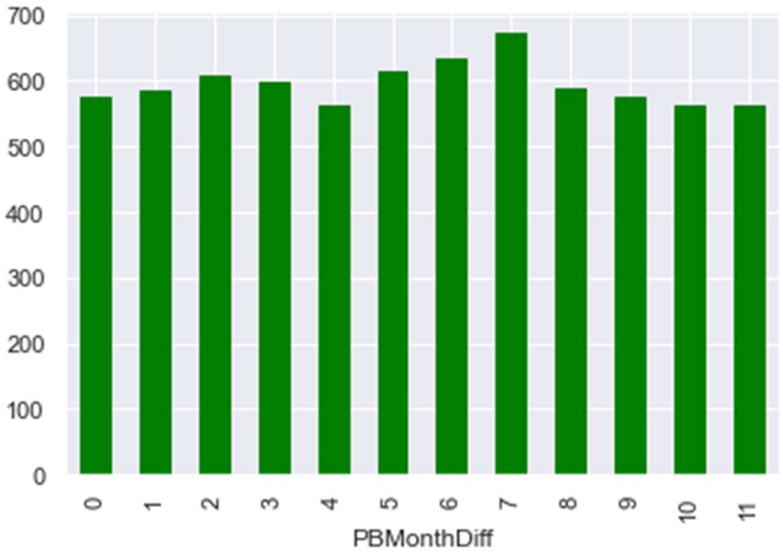


Mature Professionals

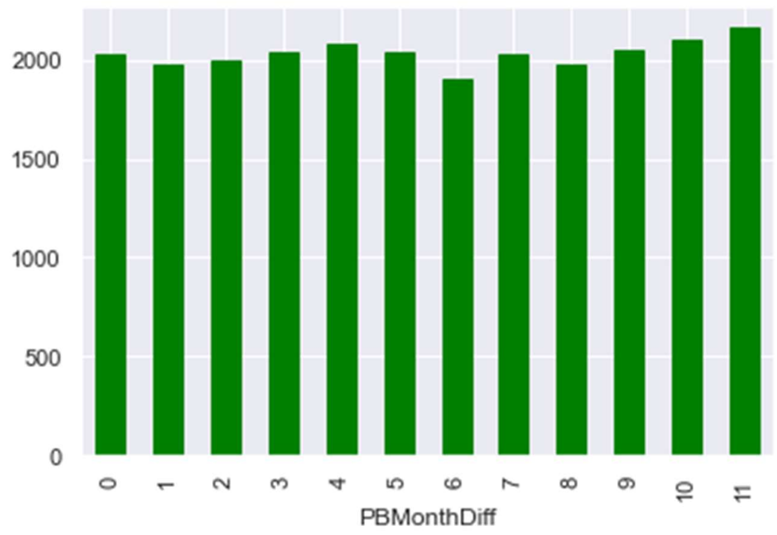


*Birthday Sentiment:* We examine the three datasets for the presence of birthday sentiment. The variable PBMonthDiff is defined as the difference of months between the birthday month and the policy issue month. The Student segment shows the increment in policy around half a year passed the birth month. The Young Professionals have some increase in policy issuance in anticipation of birthdays. The Mature Professionals have had some increase in policy issuance past birthday month.

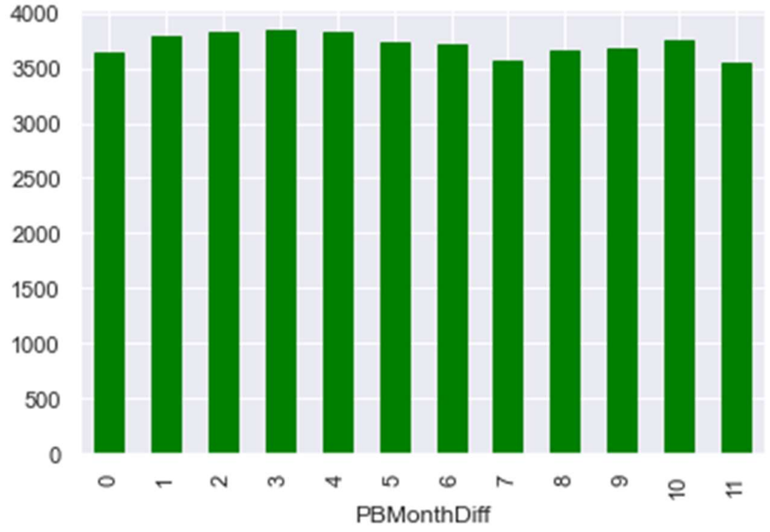
Students:



Young Professionals

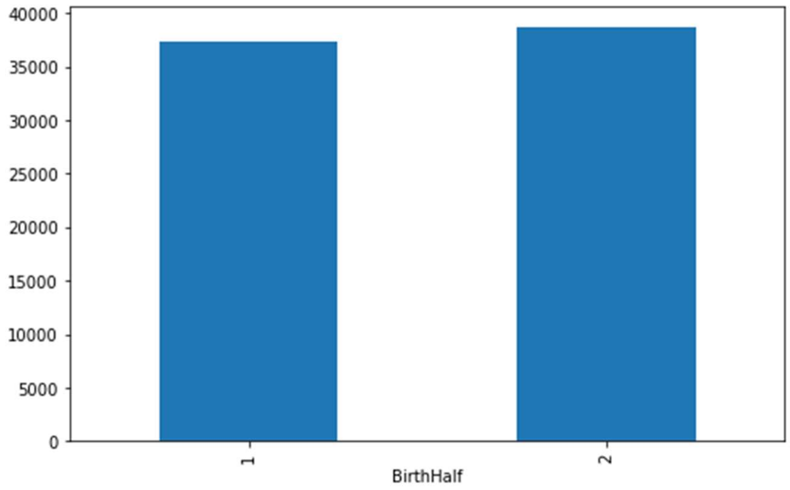


Mature Professionals

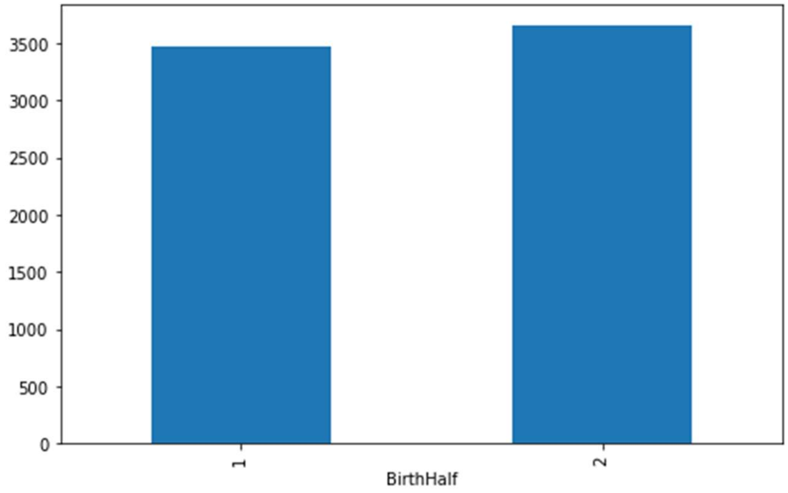


*Policy frequency for three segments based on BirthHalf:*

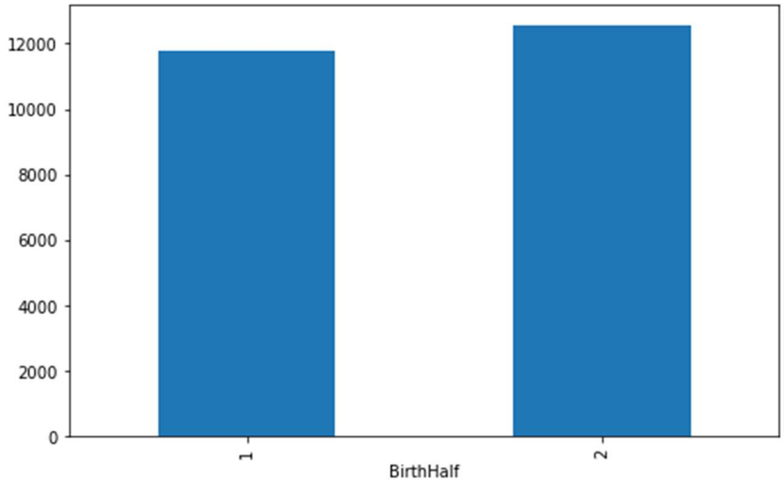
Whole dataset



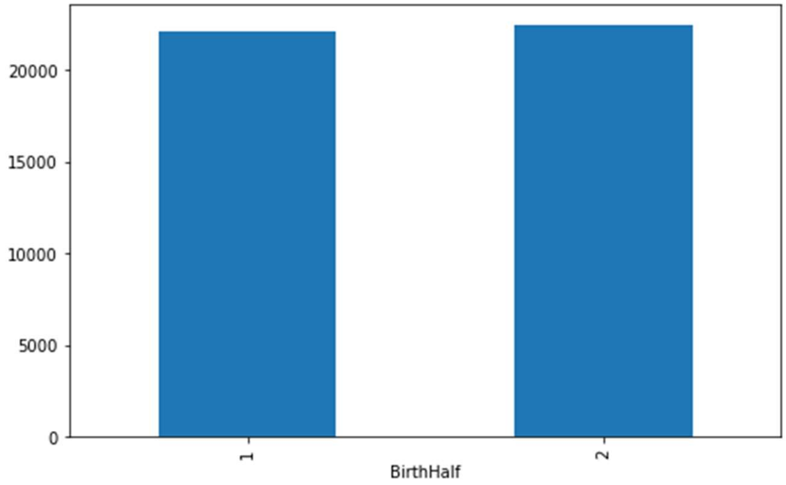
Students:



Young Professionals:



Mature Professionals



## Machine Learning Model(s)

So far, we have observed some patterns using exploratory data analysis of one variable at a time and two variables jointly. This helped us to understand when consumers prefer to get the policy issued based on available information. Now we are interested in understanding if the consumer decides to take a policy then what is their preferred Capital amount. We employ machine learning models for predicting Capital, separately for all three segments. We will examine the linear regression. Then we run non-linear models to see if those can improve results. If non-linear models show better results, then we will examine them in detail.

*Linear Regression (LR):* We follow a comprehensive procedure to prepare the data for running a linear regression model and extracting important metrics to evaluate results. The features are 'Gender', 'PolicyStartAge', 'PolicyQuarter', 'BirthQuarter', and 'PBQuarter', and the target variable is ‘Capital’. The important findings from the results for each segment are presented below.

*LR on Students:* Based on the results of the linear regression on the segment Students the significant variable (based on p-value) to predict Capital is Policy Start Age. On average, as the Policy Start Age increases by one year the Capital falls by 1152 euros.

*LR on Young Professionals:* Based on the results of the linear regression on the segment Young Professionals the significant variables to predict Capital are Gender and Policy Start Age. Males have 2990 more euros than females, on average as the Policy Start Age increases by one year the Capital increases by 676 euros.

*LR on Mature Professionals:* Based on the results of the linear regression on the segment Mature Professionals the significant variable that has the power to predict Capital are Gender, Policy Start Age, Policy Quarter 4, Birth Quarter 2, and Birth Quarter 1. Males prefer higher Capital over females, as Policy Start Age increases by one year the Capital falls by 844 euros, and the consumers who were issued policy in the 4th quarter have lesser Capital by 3513 euros compared to those who got issued in the 3rd quarter, the consumers who are born in 2nd quarter have 1594 euros less Capital compared to those who are born in 3rd quarter and the consumers who are born in 1st quarter have 1222 euros less Capital compared to those who are born in 3rd quarter.

# Conclusion

We also ran a k-mean clustering algorithm to understand if there are compact clusters in terms of two variables, we felt were important. The results, the location of the centroid in optimal clustering, do not show the presence of dense clusters.

We further run non-linear models like Polynomial Regression to examine interaction variables if it improves results, that is, increase predictability and explanation of Capital based on more variables; Decision Tree Regression to partition data into smaller homogenous groups that can improve predictability and explanation; and Neural Network to see if it can enhance predictability given it can approximate structural model underlying the data compared to linear regressions which are reduced form approaches.

All the non-linear regression approaches applied to the three segments do not further improve results. This is interpreted based on different metrics of goodness of fit. Hence, we do not further probe into those models.

Based on the approach and methodology used so far to analyze this data we have the following major findings:

1. Two-thirds of the consumers are males and one-third are females. The relationship among variables for males and females is similar.
2. Males have somewhat higher Capital compared to females.
3. A range of Capital with the maximum number of policies is 50-75k euros. Most of the policies are in the range of 25-125k euros.
4. The consumers can be divided into three population segments (Students, Young Professionals, and Mature Professionals) based on age as they show qualitatively different behaviors.
5. There is seasonality in terms of when consumers prefer to buy policies: The last quarter of the calendar year which is Autumn (due to high sales in December) is the highest and the third quarter which is Summer (due to low sales in August and September) is the lowest for all three segments.
6. There are two trends present in all three segments in terms of when consumer purchases policy: one starts in Jan and peaks in March (financial year end and Easter), and another starts in August and peaks in December (Christmas and holiday period).
7. There is some presence of seasonality based on which half of the year the birthday of the consumer falls into. For all the segments, more consumers are born in the second half of the year.
8. There appears no strong presence of birthday sentiment nor there is the presence of a common seasonality w.r.t. birthday.
9. Based on results from multiple machine learning models, the data is too limited to be able to explain and predict Capital sufficiently.
10. For all segments, Policy Start Age is a significant variable in determining Capital.
11. For all Professionals, Gender is a significant variable in determining Capital. The reason it is not important for Students might be that their decision is made by their parents.

12. For Mature Professionals, the largest segment, birth, and policy month also matters in determining Capital.

In further work on this data, one can examine the results of machine learning models if quarterly variables are replaced with monthly variables. It can be a useful exercise to probe further into polynomial regression to examine if any interaction variables are significant despite poor goodness of fit. Based on the new variables one can go back and iterate the models to see if these can enhance results.