Identify premium pricing attributes for home insurance using R

CRISP-DM (CRoss Industry Standard Process for Data Mining) The need to standardize the lessons learned into a common method became increasingly pressing as the 1990s went on.In 1996, Daimler, NCR, and OHRA, along with two of the leading tool providers of the time, SPSS and Teradata, formed a Special Interest Group (SIG). In less than a year, they were able to codify the CRISP-DM, or CRoss Industry Standard Process for Data Mining. Actually, CRISP-DM wasn't the first. SEMMA, which stands for "Sample, Explore, Modify, Model, Assess," was the SAS Institute's own version. However, within a year or two, many more practitioners were incorporating CRISP-DM into their strategies.

● CRISP-DM Methodology The CRISP-DM process or methodology of CRISP-DM is described in these six major steps

1. Business Understanding concentrates on comprehending the project's requirements and objectives from a business perspective. The analyst creates a preliminary plan and presents this information as a data mining problem.

2.Data Understanding The analyst proceeds with activities to become familiar with the data, identify issues with data quality, and discover initial insights into the data, beginning with the initial data collection. The analyst might also find interesting subsets during this phase to generate hypotheses about hidden information.

3. Data Preparation The data preparation phase covers all activities to construct the final dataset from the initial raw data

4. Modeling The analyst looks at, chooses, and uses the right modeling methods. because some methods, like neural nets, have specific requirements for the data's form. This could lead back to data preparation

5. Evaluation Based on the loss functions that were chosen, the analyst constructs and selects models that appear to be of high quality. The analyst then puts them through tests to make sure they can apply the models to new data. The analyst then confirms that the models adequately address all important business issues. The champion model selection is the end result.

Identify premium pricing attributes for home insurance using R

6. Deployment In most cases, this will entail incorporating a model code representation into an operating system. Mechanisms for scoring or classifying newly generated, unseen data are also included in this. The new data should be used by the mechanism to solve the original business problem. Importantly, all data preparation steps prior to modeling must be included in the code representation. This guarantees that the model will treat new raw data in the same way as it did when it was being developed.

● Characteristics of CRISP-DM

1.A number of characteristics, in my opinion, account for CRISP-DM's longevity in a field that is undergoing rapid change: It encourages data miners to concentrate on business objectives to guarantee that the project's outputs have a direct impact on the organization. Too frequently, analysts can lose sight of the ultimate business purpose of their analysis, turning it into a means to an end rather than an end in and of itself. The CRISP-DM method helps to keep the project's business objectives at the center of everything.

2.CRISP-DM offers an iterative approach, with frequent chances to compare the project's progress to its original goals.As a result, there is less chance that the project will reach its conclusion and reveal that the company's goals have not really been met. Additionally, it indicates that the project stakeholders are able to modify and adapt the objectives in light of new findings.

3.The CRISP-DM approach does not discriminate based on technology or issue. Any software you like can be used for your analysis and applied to any data mining issue. CRISP-DM will still provide you with a framework with enough structure to be useful, regardless of the nature of your data mining project

Steps we are performing to Identify premium pricing attributes for home insurance

DATA PREPROCESSING
EDA
Comparison chart of features we are interested in
Transform Data

Identify premium pricing attributes for home insurance using R

DATA PREPROCESSING

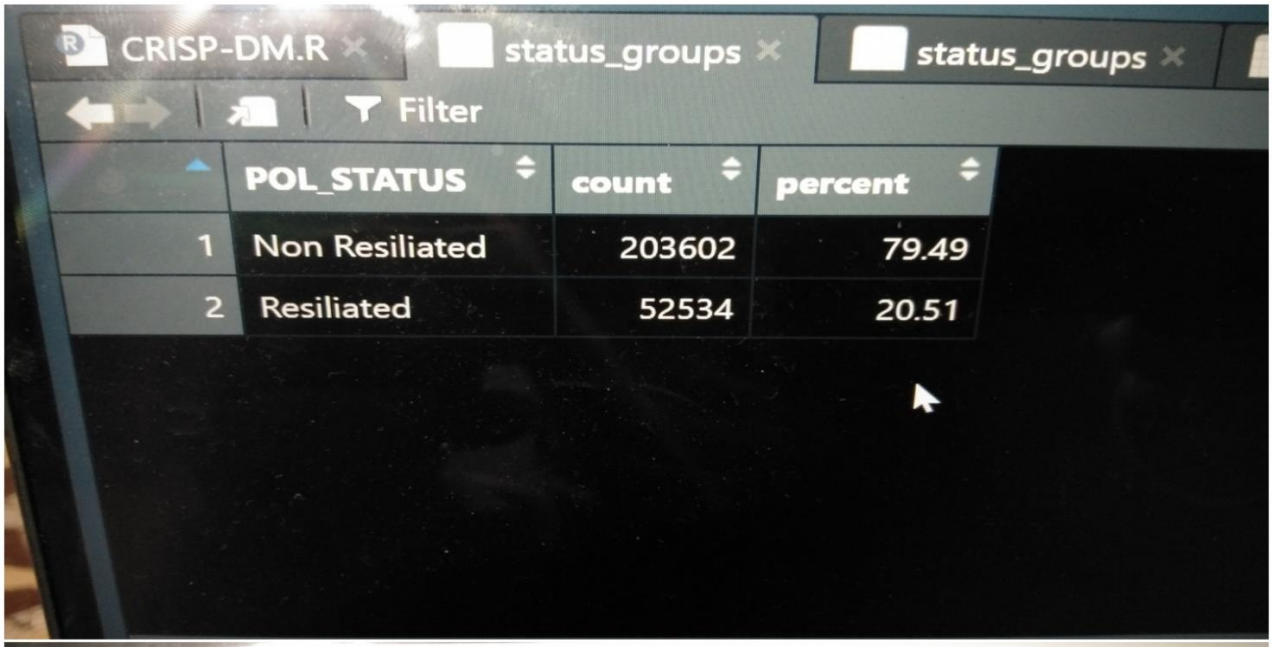
There are a total of **256,136 policies** with **66 features**. Most of the features have so much null values. I will attempt at cleaning this dataset to a form suitable for analysis.

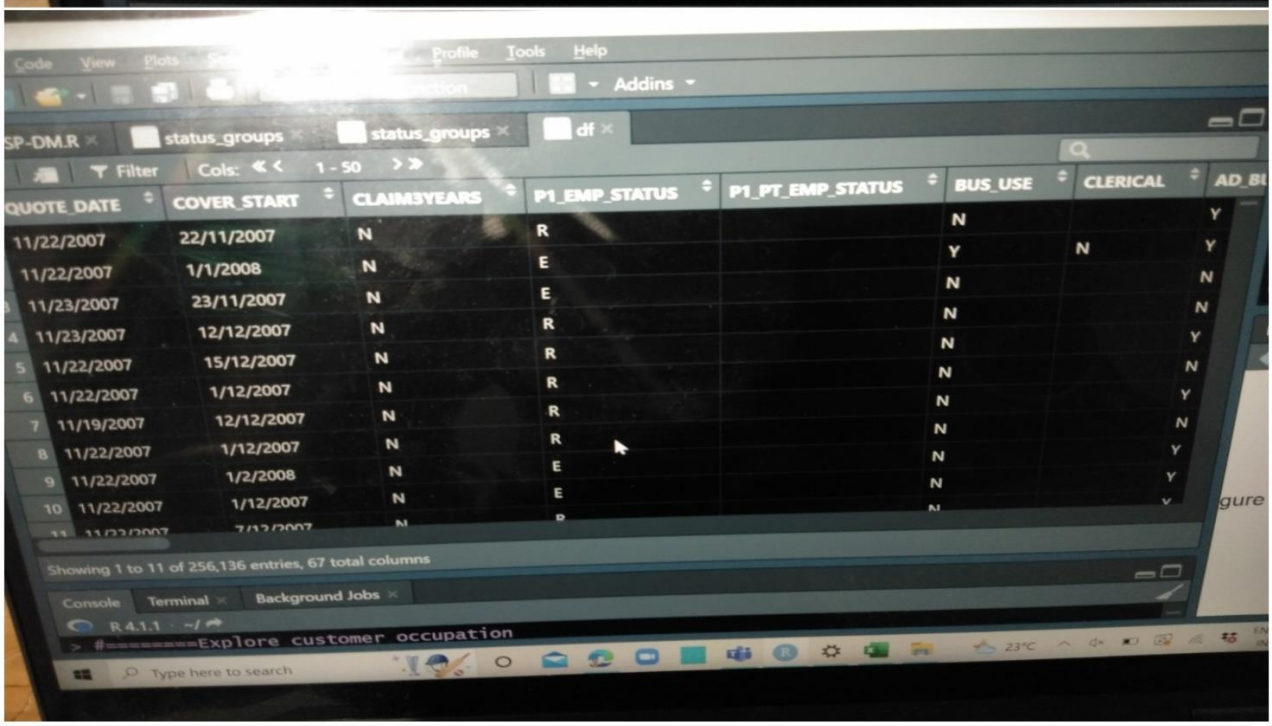
Data Wrangling

The data that was originally obtained was in the form of a Microsoft Office Access File (.accdb). This was converted manually into a CSV file (in Microsoft Office Excel) to arrive at an input that could be loaded into a R DataFrame effortlessly. In other words, this dataset is already relatively clean. I will however attempt at learning more about this features and performing appropriate wrangling steps to arrive at a form that is more suitable for analysis.

Identify premium pricing attributes for home insurance using R

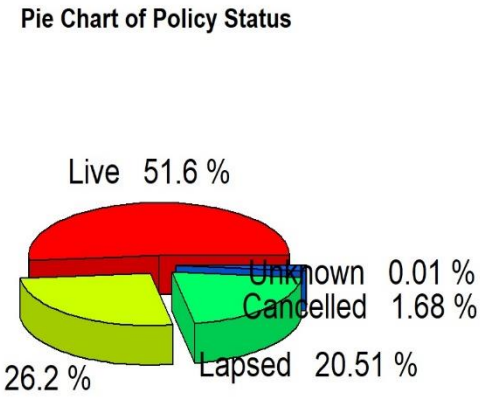
DataFrame





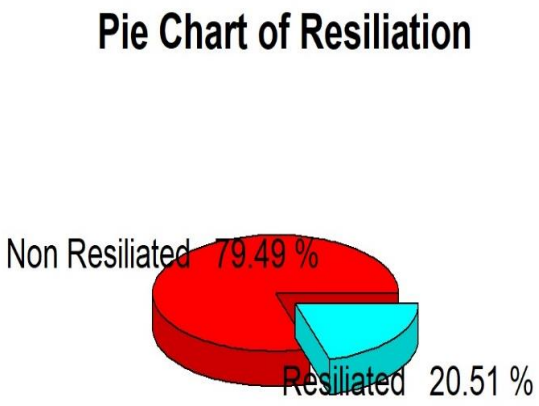
Identify premium pricing attributes for home insurance using R

Pie Chart of Policy Status



We can see that the majority of the policies have a status of live (**51.6%**).

Pie Chart of resiliation



Identify premium pricing attributes for home insurance using R

Exploratory Data Analysis

Policy best covered

I am curious to discover the most covered policy among the others. I will wrangle the data to find out adding a new column that show the total coverage of the policy

The Policy with ID **P098293** is the most covered policy in almost **1,138k** dollars.
The **P032217** come in a close second with a **1,132k** dollars. Policy **P028759** is third but this policy has significantly less Valuable Personal Property compared to the two first ones in the list and therefore, a much smaller total coverage.

Client's professional status

In this section, I will look at the client's professional status of the policies in the HI dataset.

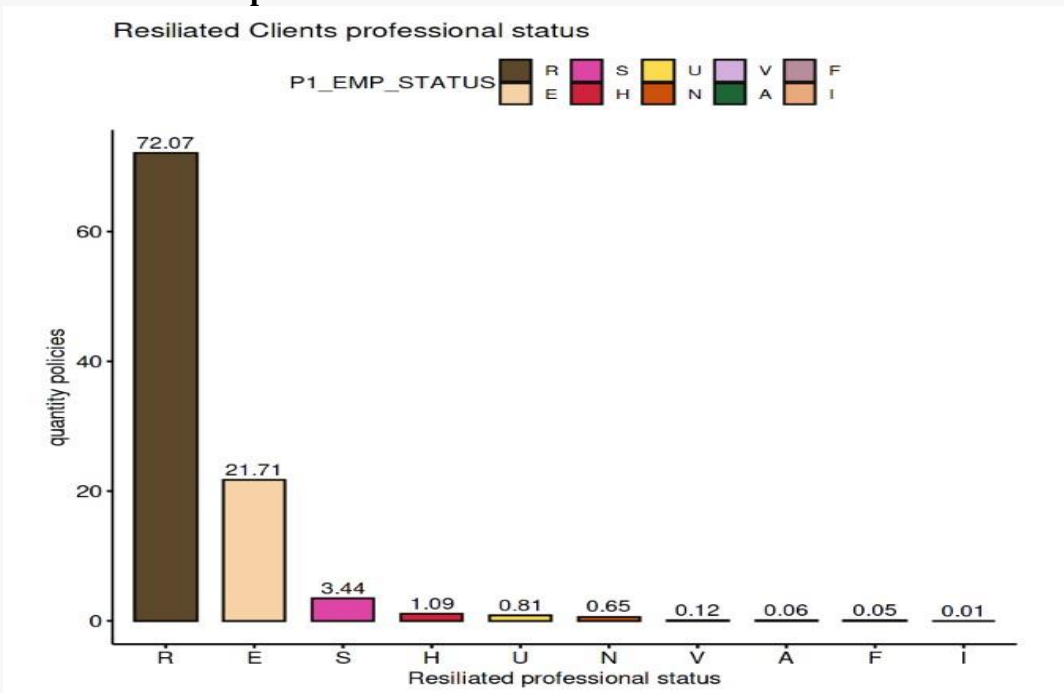
This was the information I have received from the research. Sorry is not completed, the main goal of this project is just get some insights, even when we don't have the complete information.

- R = Retired,
- E = Employed,
- N = Not Available,
- H = House person,
- S = Student ,
- U = Unemployed.

Anyways, there are over **11** professional status represented in the HI dataset (avoiding the null status). The **Retired** clients form the overwhelmingly majority. The **Employees** and **Students** come at a very distant second and third respectively.

Identify premium pricing attributes for home insurance using R

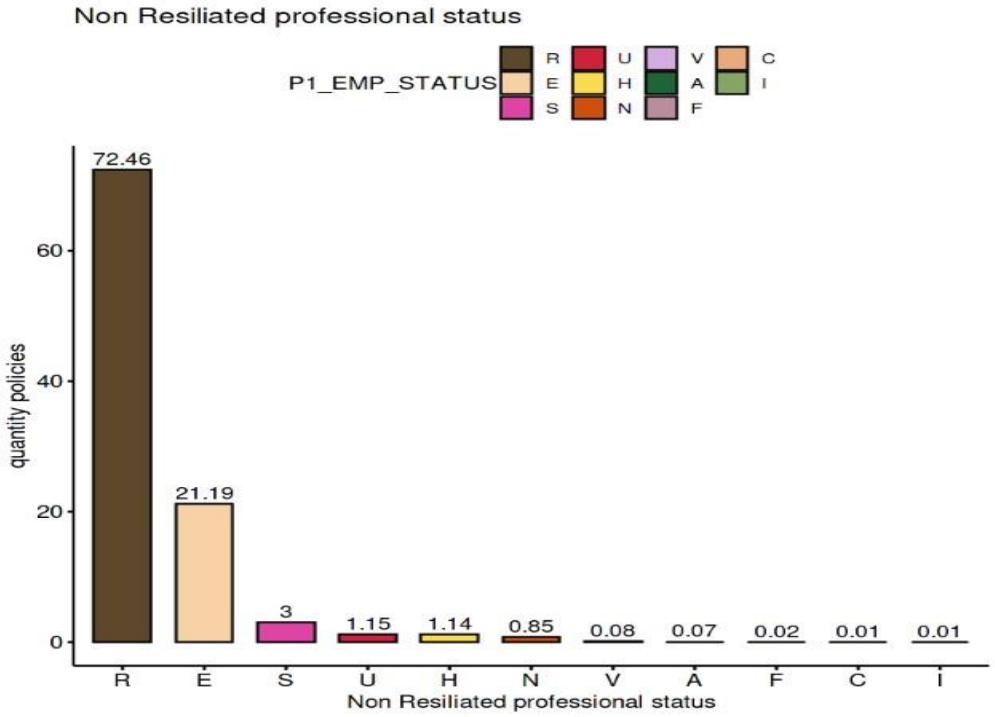
Resiliated Clients professional status

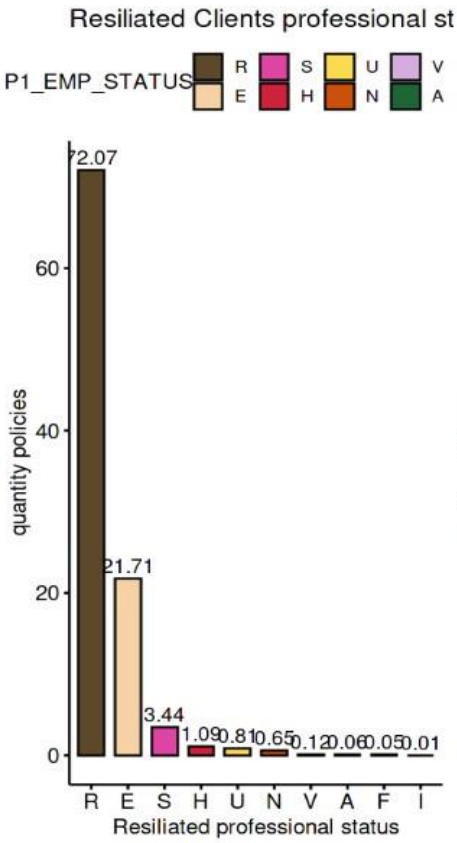


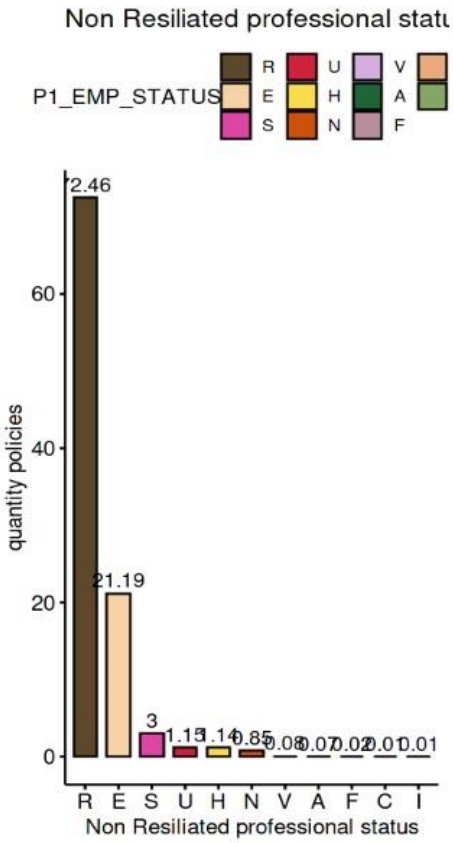
As mentioned earlier, **Retired** and **Employees** clients are the most commonly occurring professional status. **V**, **A**, **F**, **I**, **C** form the minority. I can imagine that this last statuses are two or the following:

1. The richest clients of the sample.
2. The clients living in a very uncomfortable zone/building.

Identify premium pricing attributes for home insurance using R

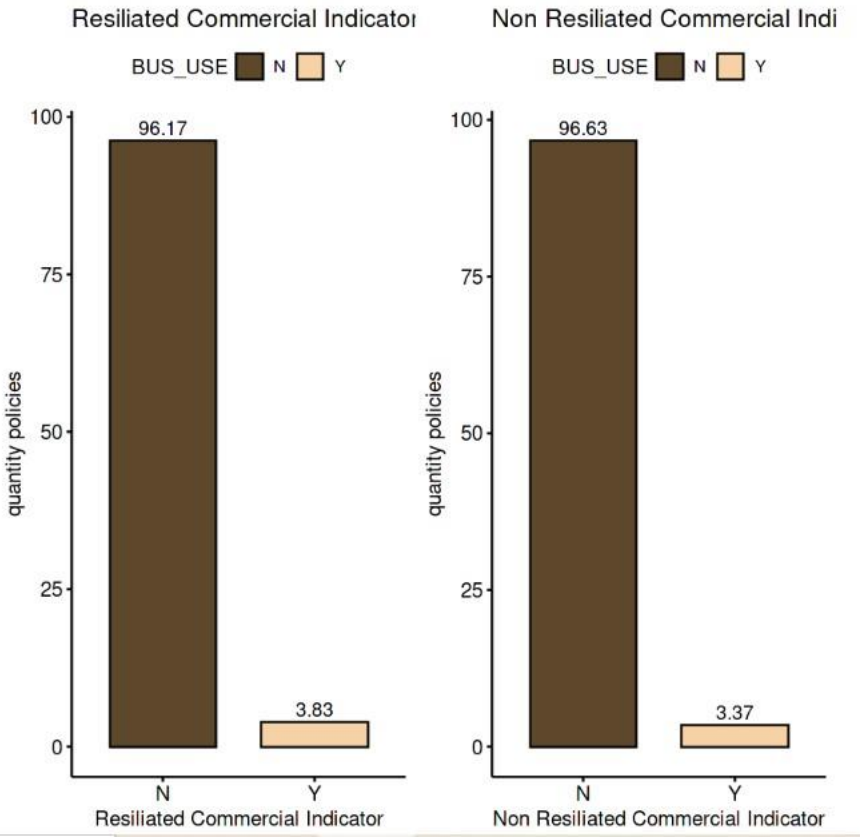




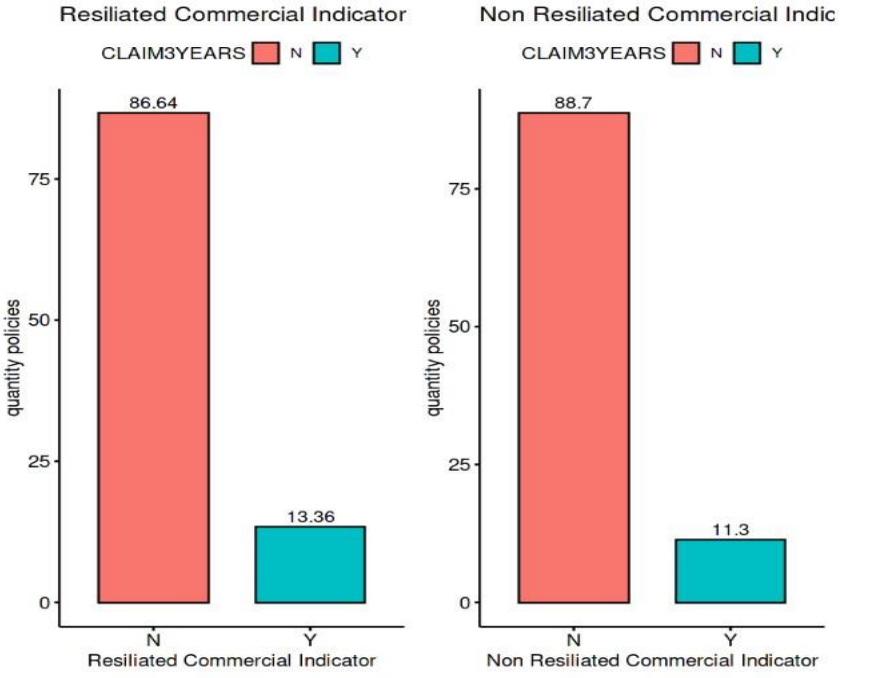


Identify premium pricing attributes for home insurance using R

Comparison chart of features we are interested in

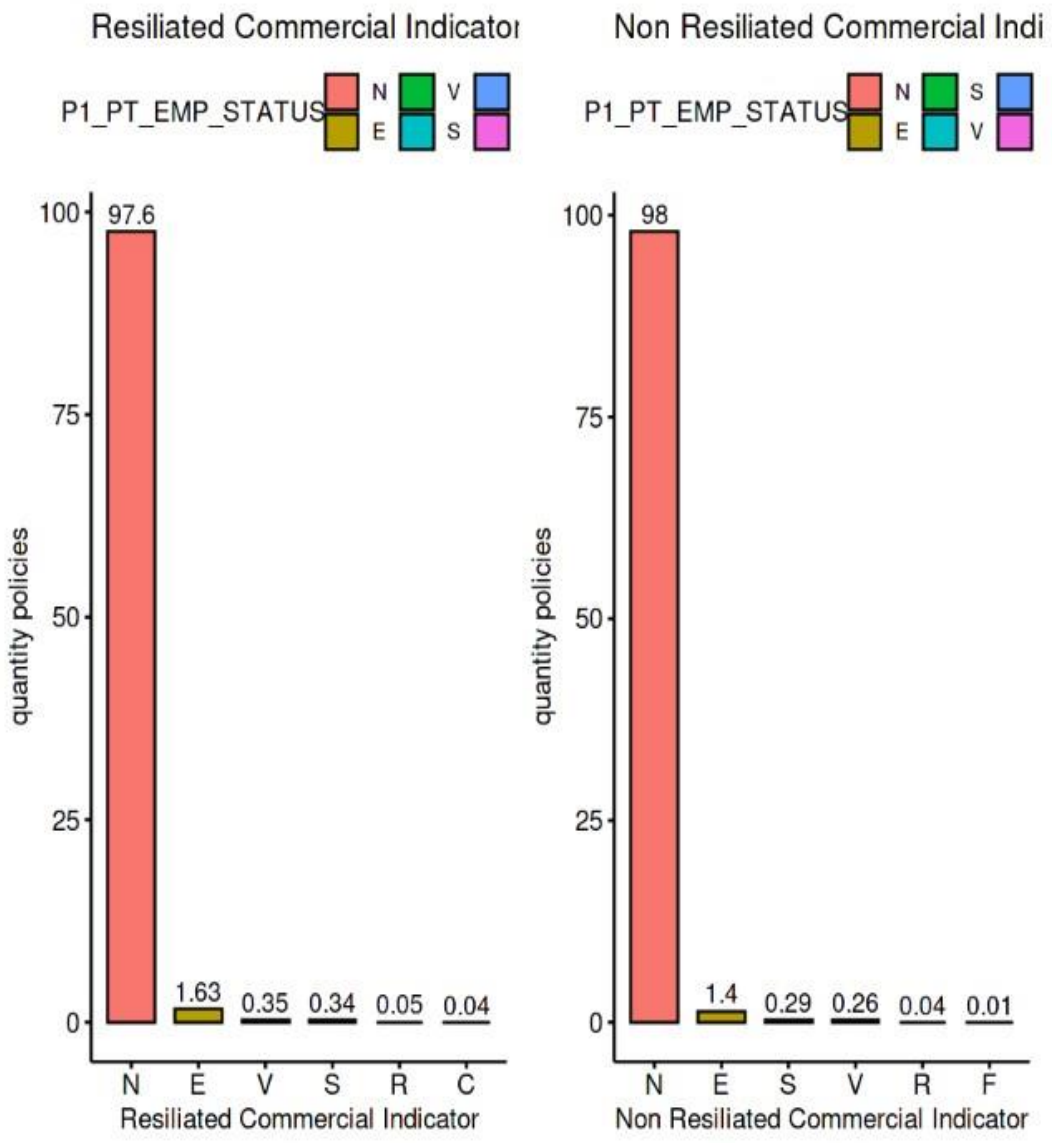


Bar chart for Resiliated and non-resiliated (CLAIM3YEARS) is given below



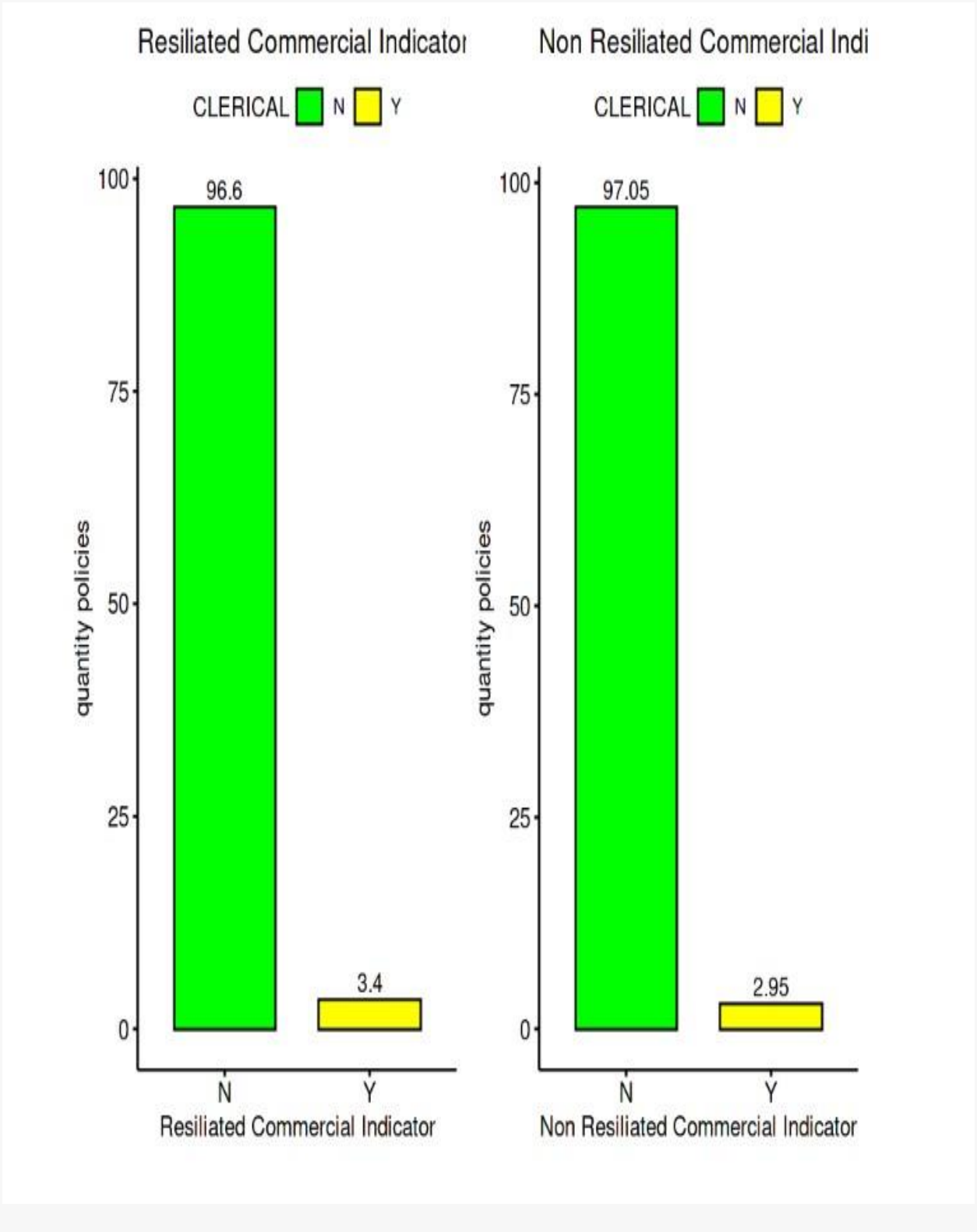
Identify premium pricing attributes for home insurance using R

Bar chart for Resiliated(P1_PT_EMP_STATUS) and non_resiliated (P1_PT_EMP_STATUS) is given below:



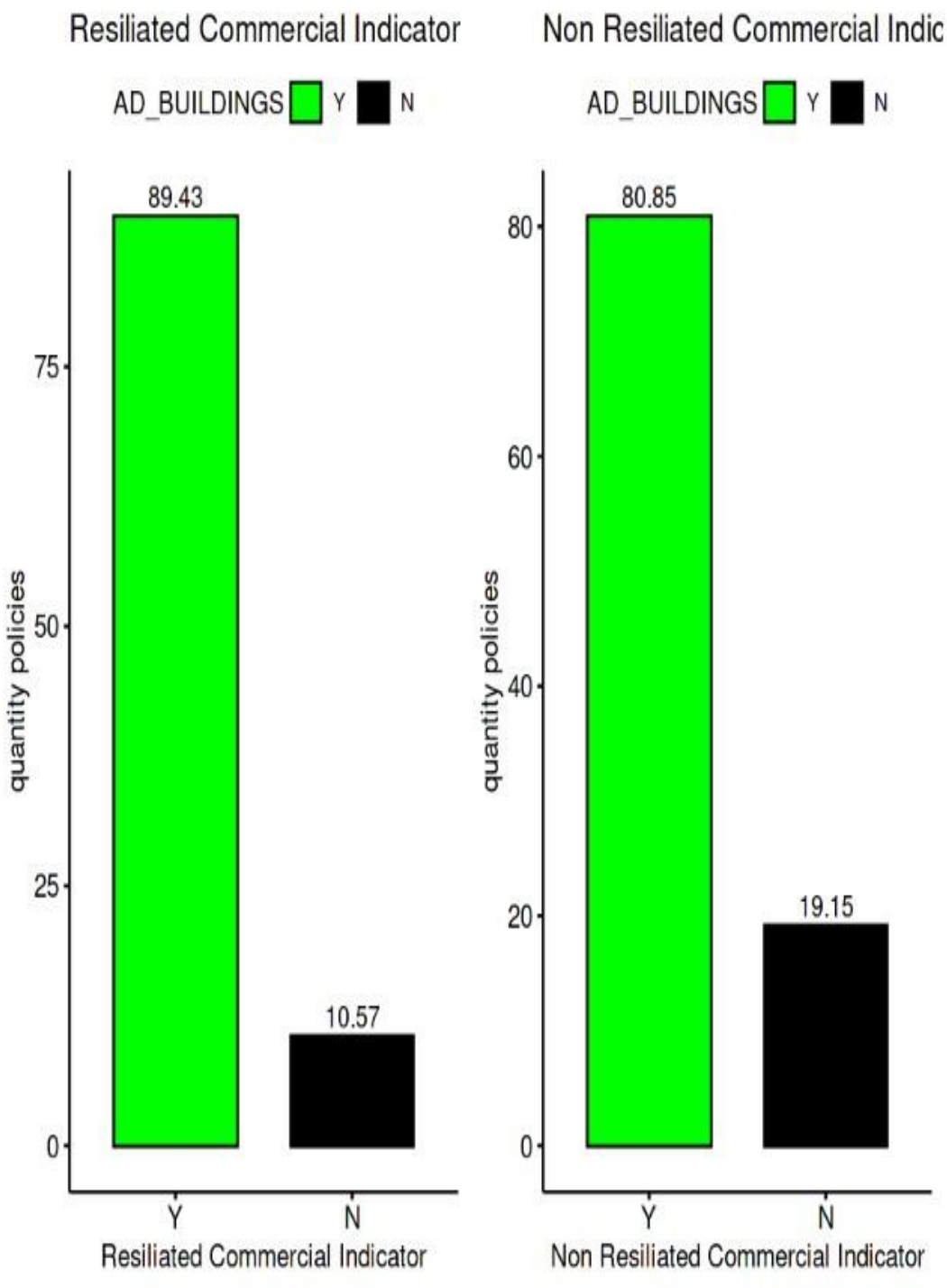
Identify premium pricing attributes for home insurance using R

Bar chart for Resiliated (CLERICAL)and non- Resiliated (CLERICAL) is given below:



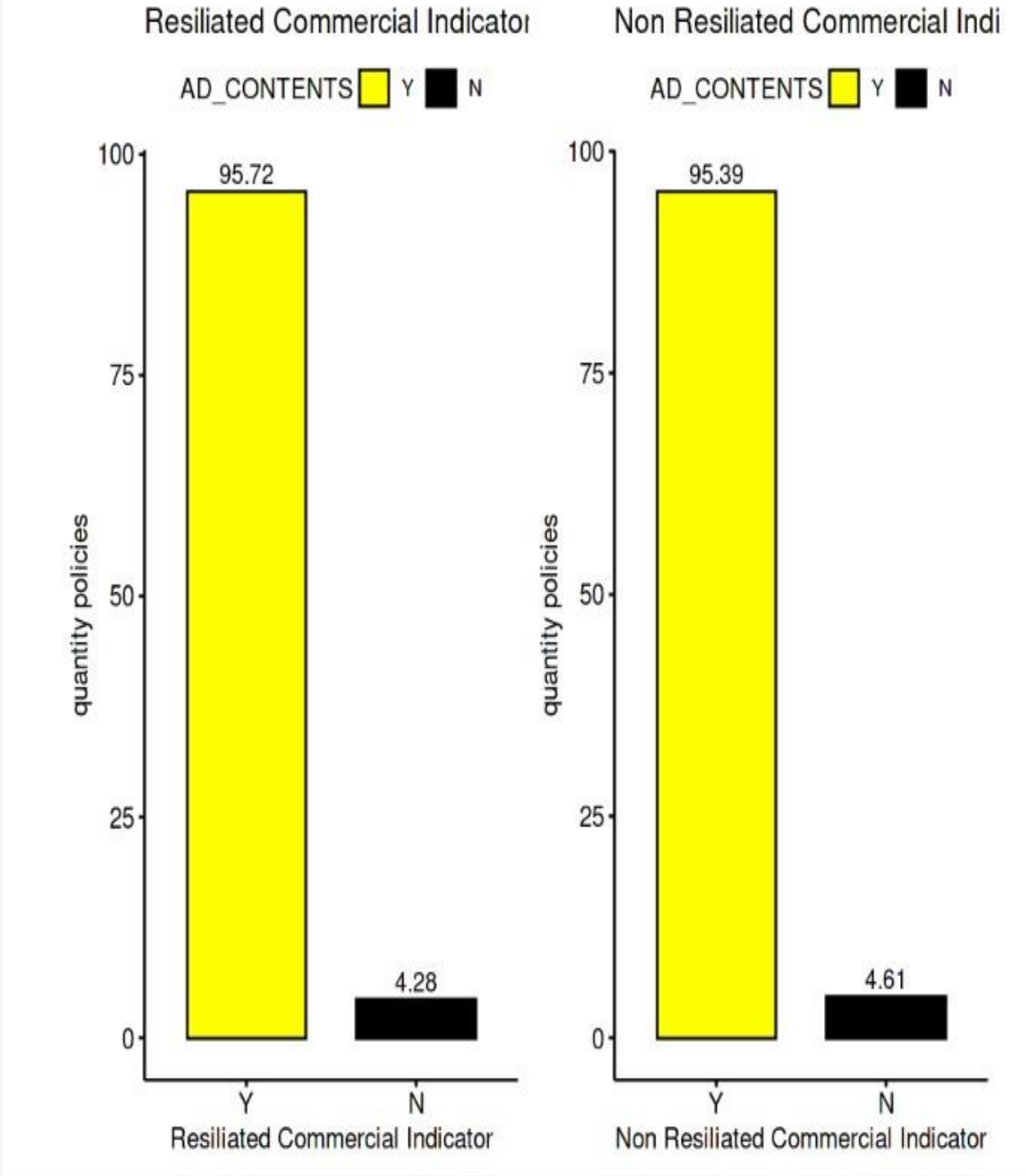
Identify premium pricing attributes for home insurance using R

Bar chart for Resiliated (AD_BUILDINGS) and Non-Resiliated (AD_BUILDINGS)is given belo w



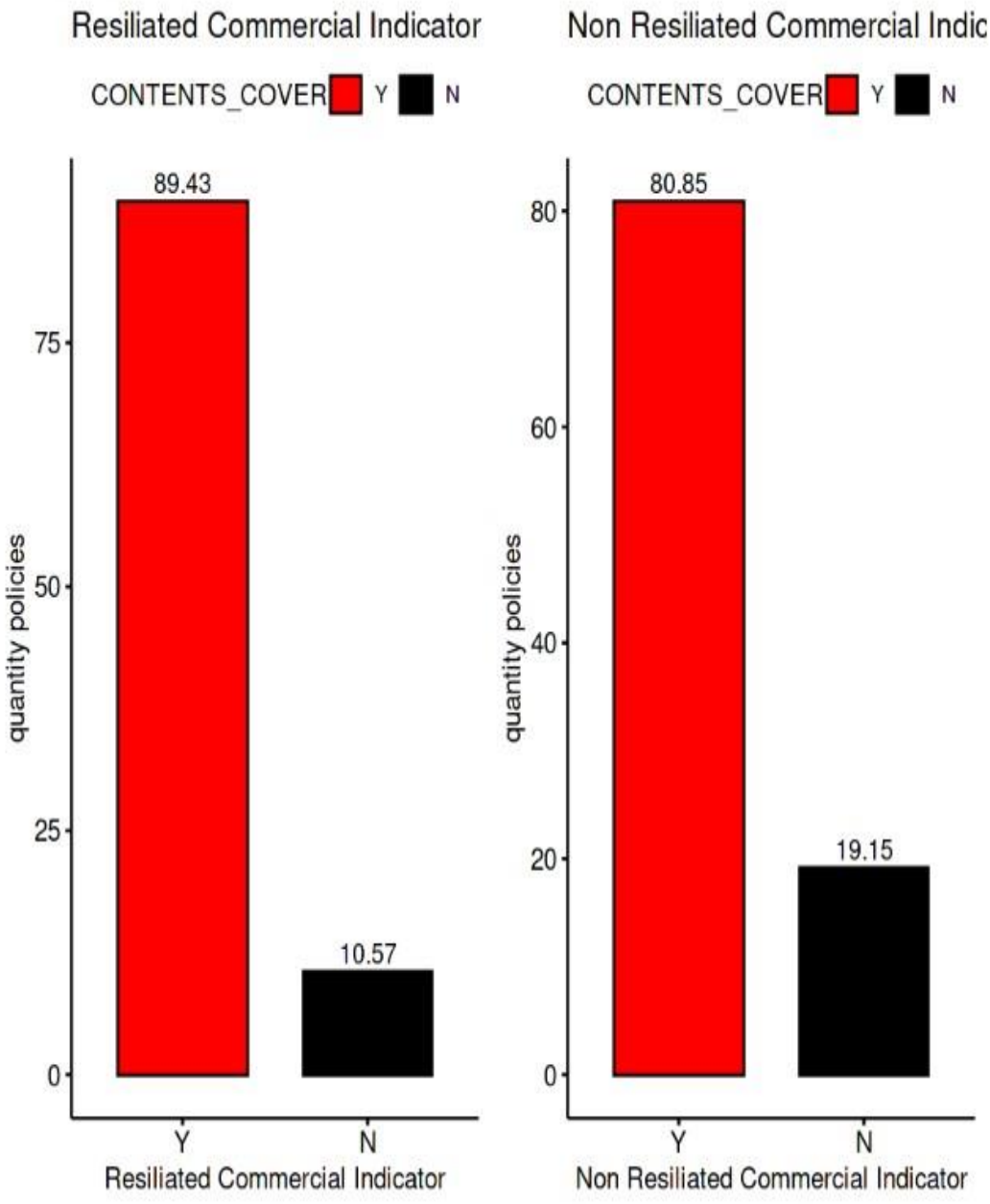
Identify premium pricing attributes for home insurance using R

Bar chart for Resiliated (AD_CONTENTS) and non_ Resiliated (AD_CONTENTS)



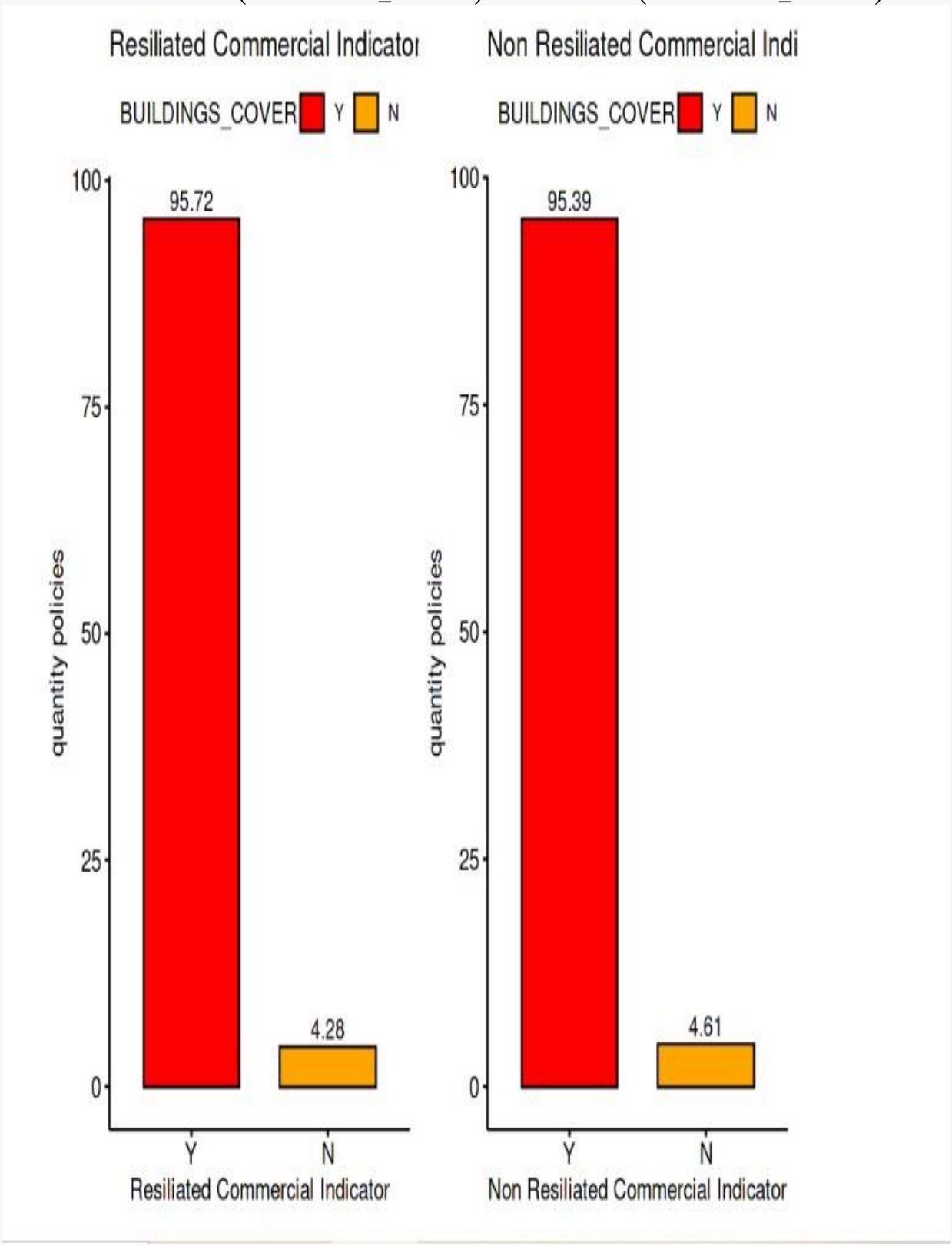
Identify premium pricing attributes for home insurance using R

Bar chart for Resiliated (CONTENTS_COVER) and non- Resiliated (CONTENTS_COVER



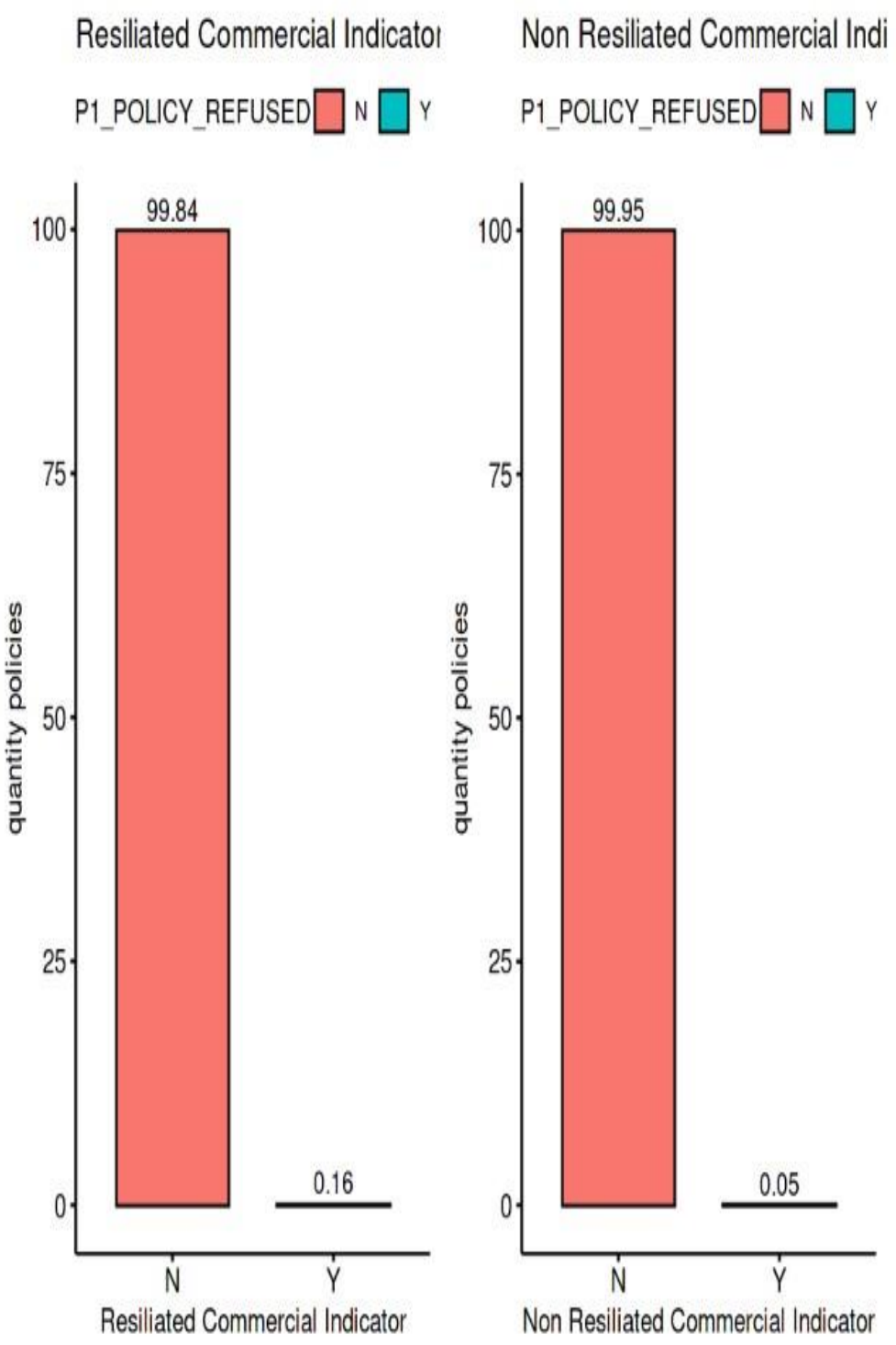
Identify premium pricing attributes for home insurance using R

Bar chart for Resiliated (BUILDINGS_COVER) non-Resiliated (BUILDINGS_COVER)



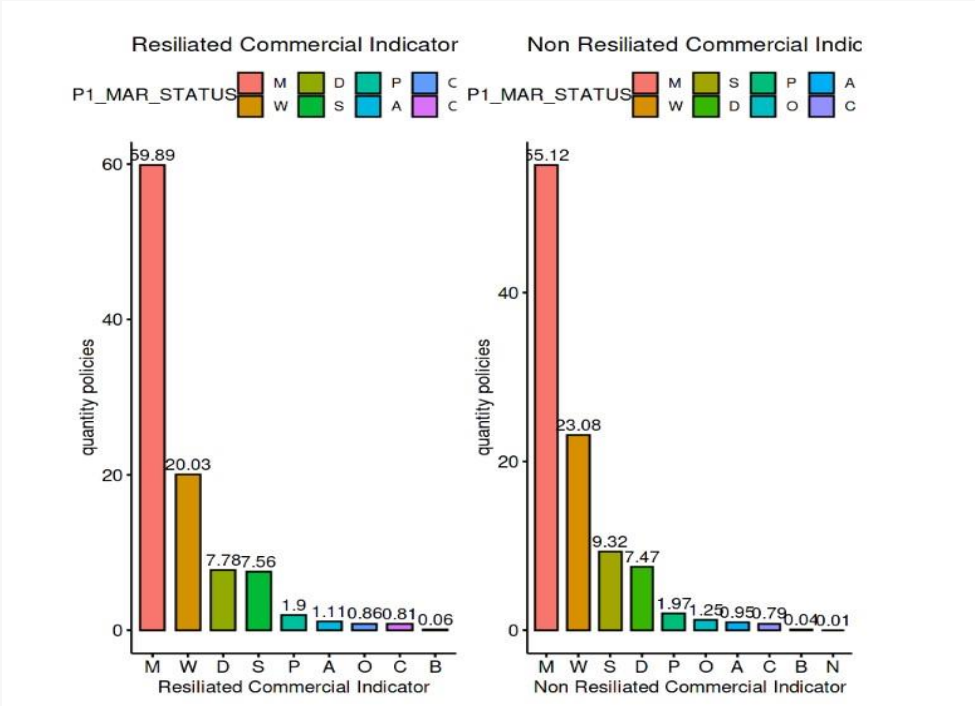
Identify premium pricing attributes for home insurance using R

Bar chart for Resiliated (P1_POLICY_REFUSED) and non- Resiliated (P1_POLICY_REFUSED)

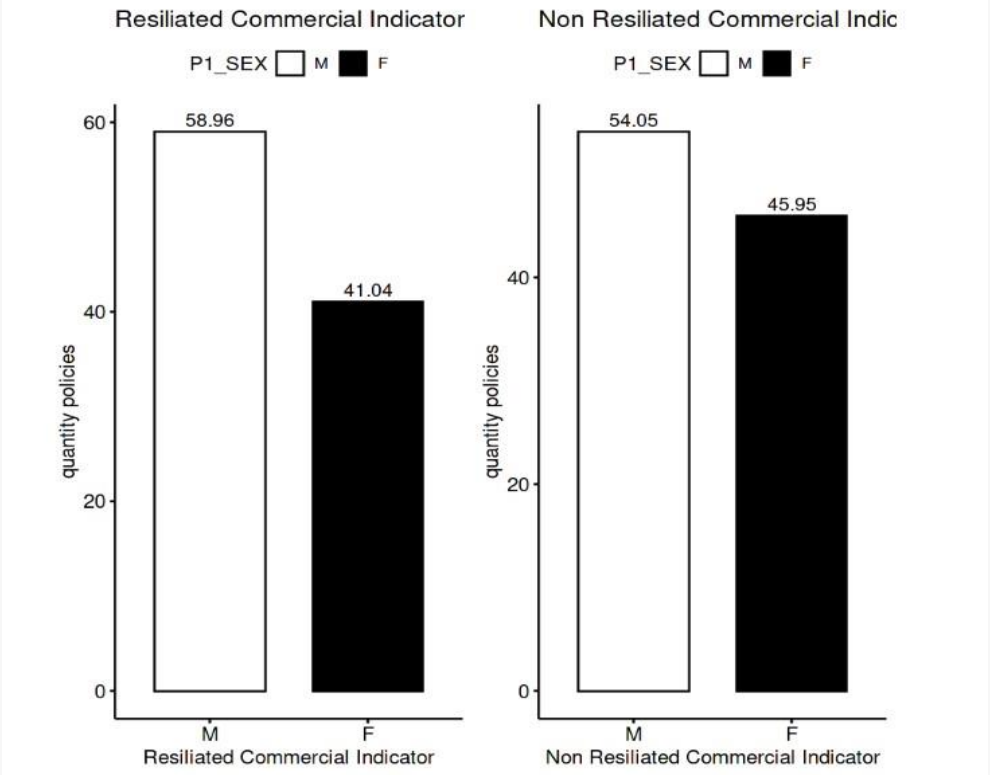


Identify premium pricing attributes for home insurance using R

Bar chart for Resiliated (P1_MAR_STATUS) and non- for Resiliated (P1_MAR_STATUS)

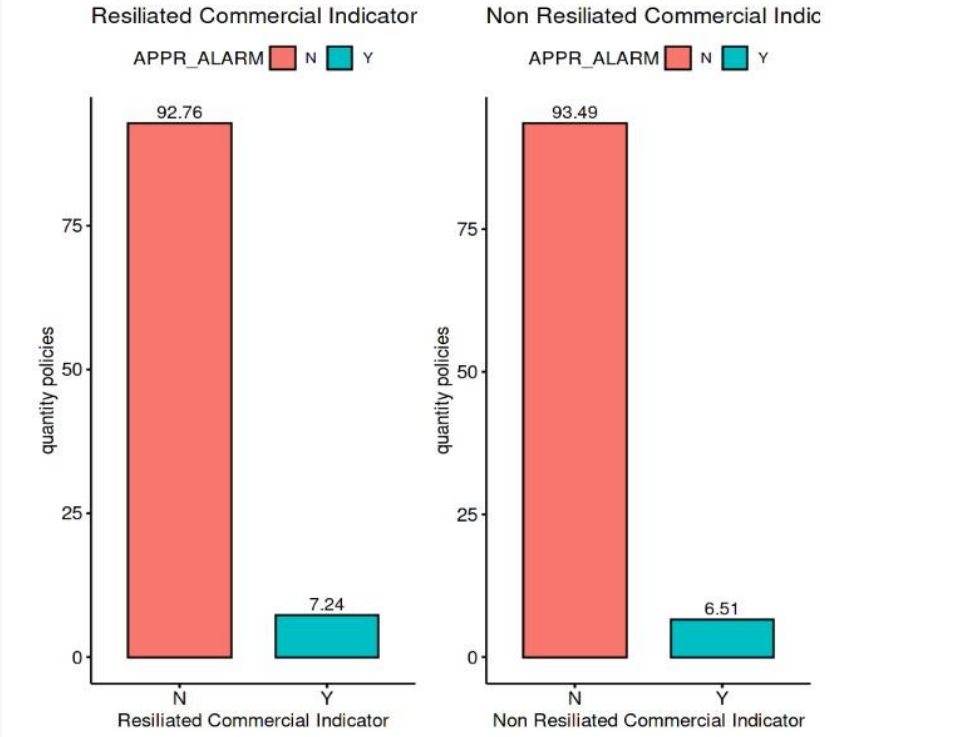


Bar chart for Resiliated (P1_SEX) and non-Resiliated (P1_SEX)

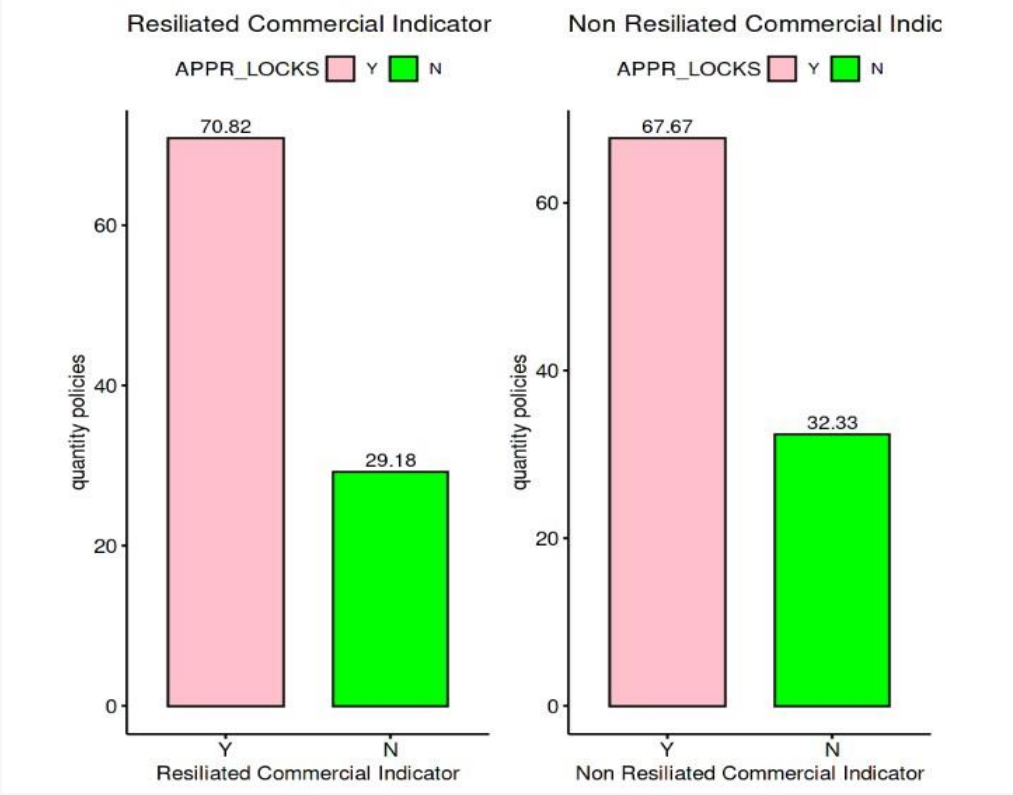


Identify premium pricing attributes for home insurance using R

Bar chart for Resiliated (APPR_ALARM) and non- Resiliated (APPR_ALARM)

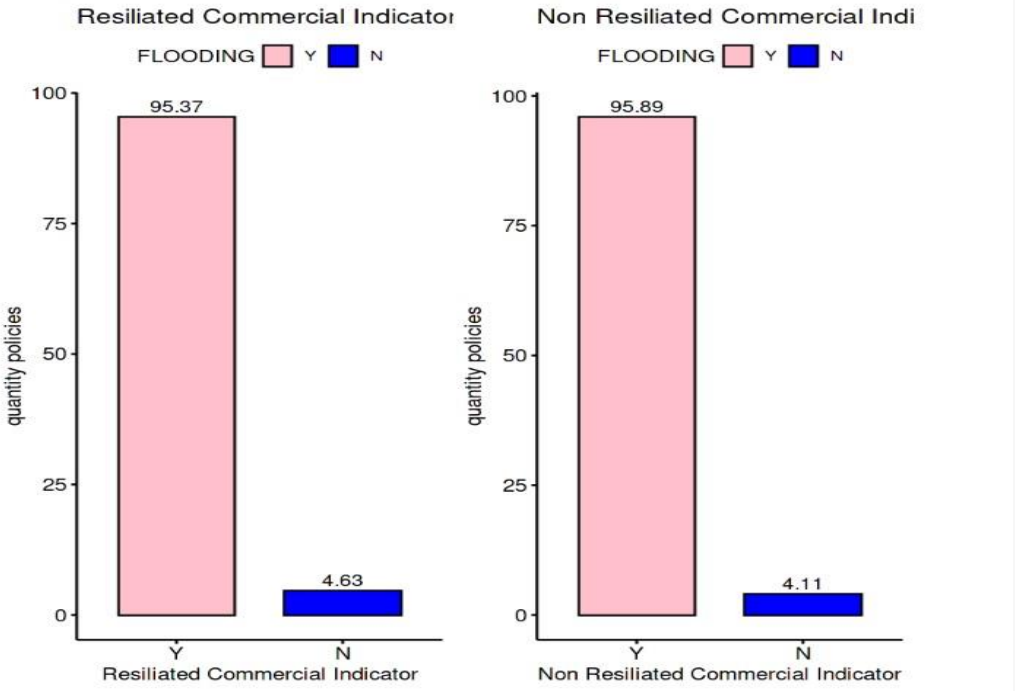


Bar chart for Resiliated (APPR_LOCKS) and non- Resiliated (APPR_LOCKS)

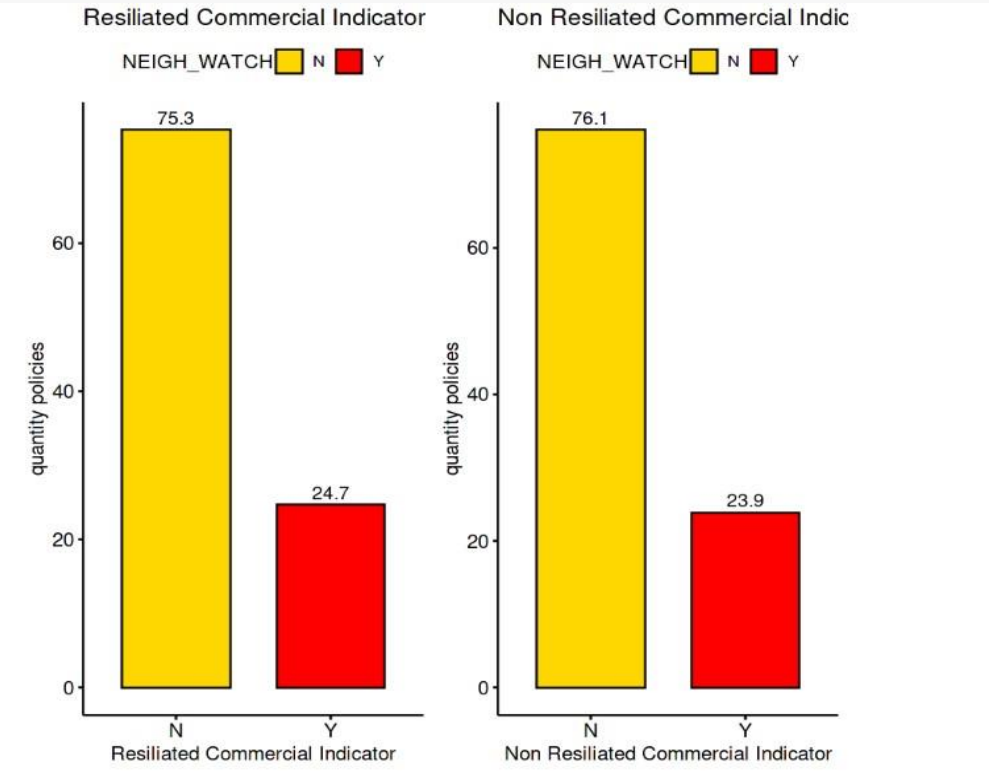


Identify premium pricing attributes for home insurance using R

Bar chart for Resiliated (FLOODING) and non_ Resiliated (FLOODING) is shown below

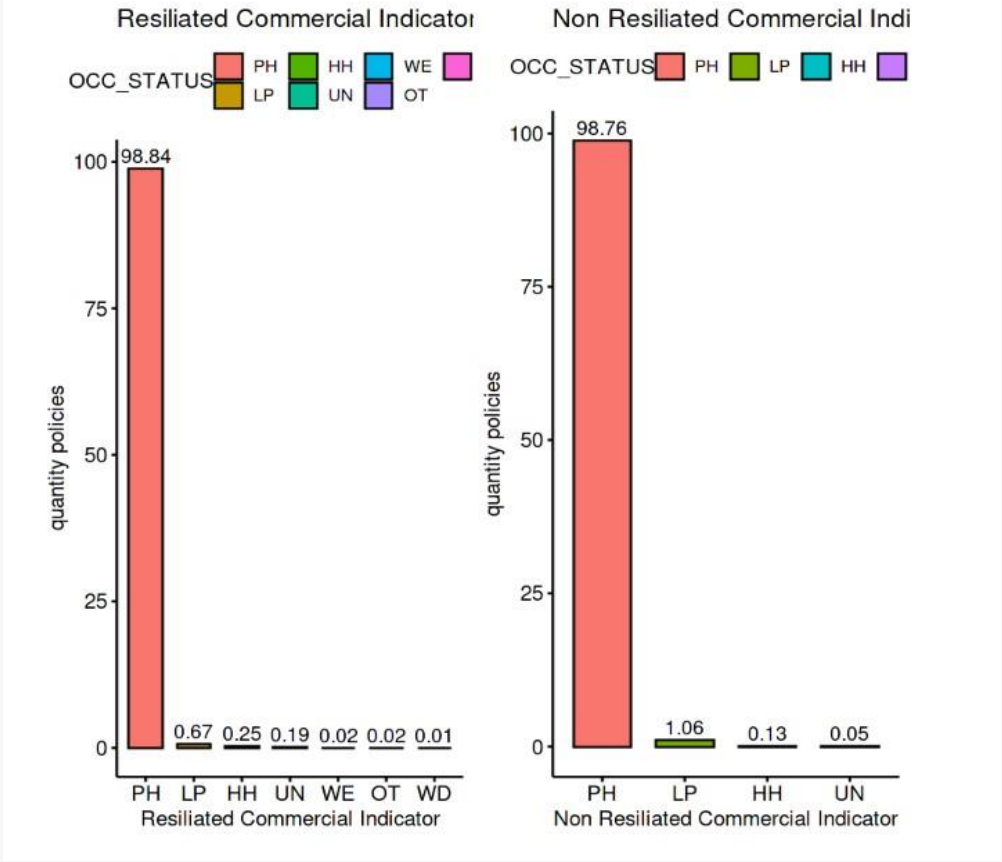


Bar chart for Resiliated (NEIGH_WATCH) and non_ Resiliated (NEIGH_WATCH) shown bel ow

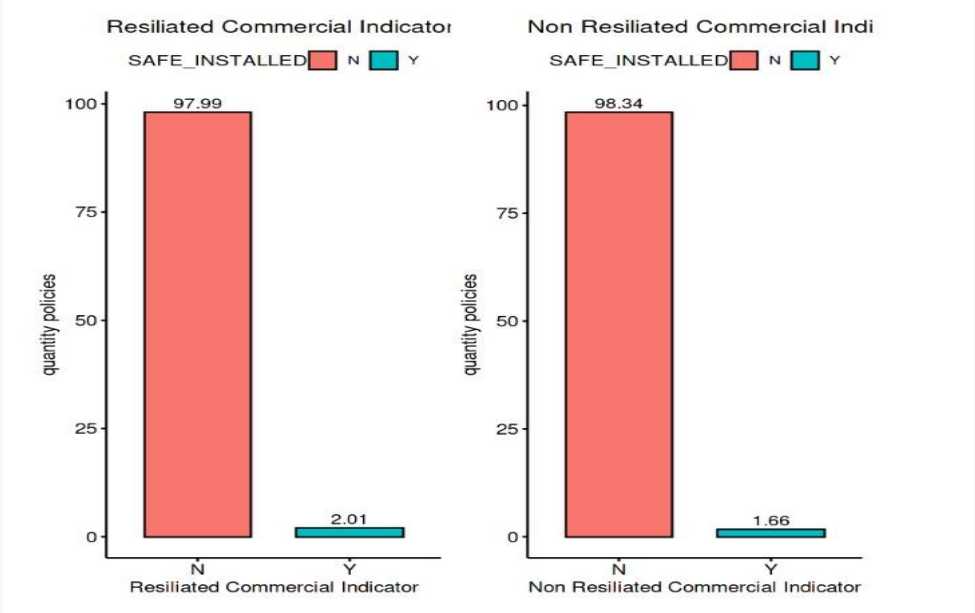


Identify premium pricing attributes for home insurance using R

Bar chart for Resiliated (OCC_STATUS) and non_ Resiliated (OCC_STATUS) shown below

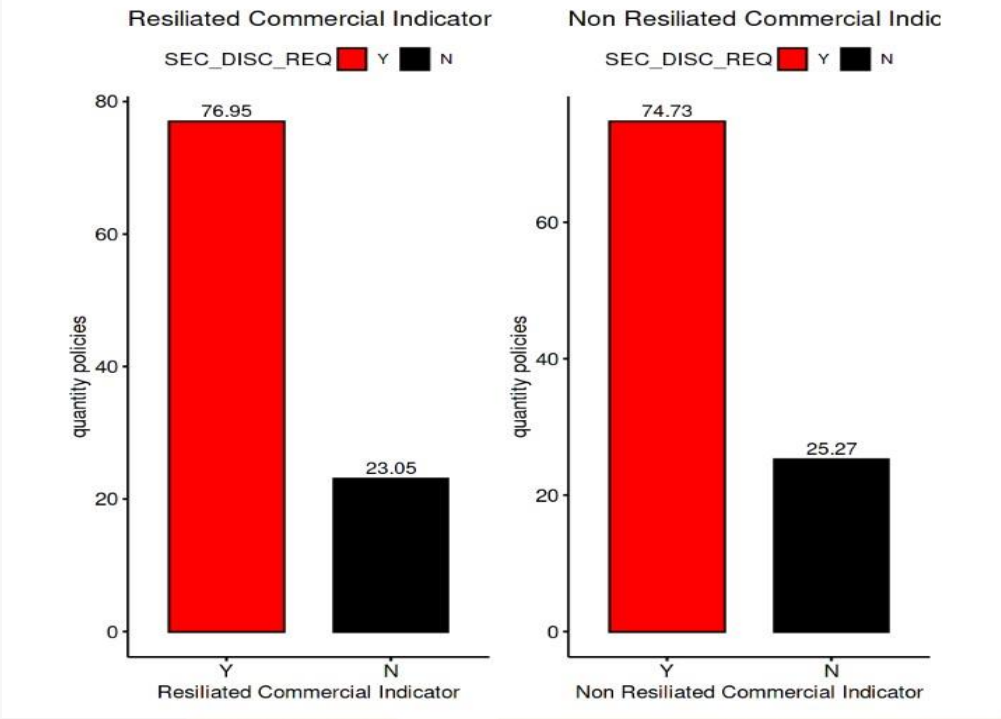


Bar chart for Resiliated (SAFE_INSTALLED) and non- Resiliated (SAFE_INSTALLED) is give -n below

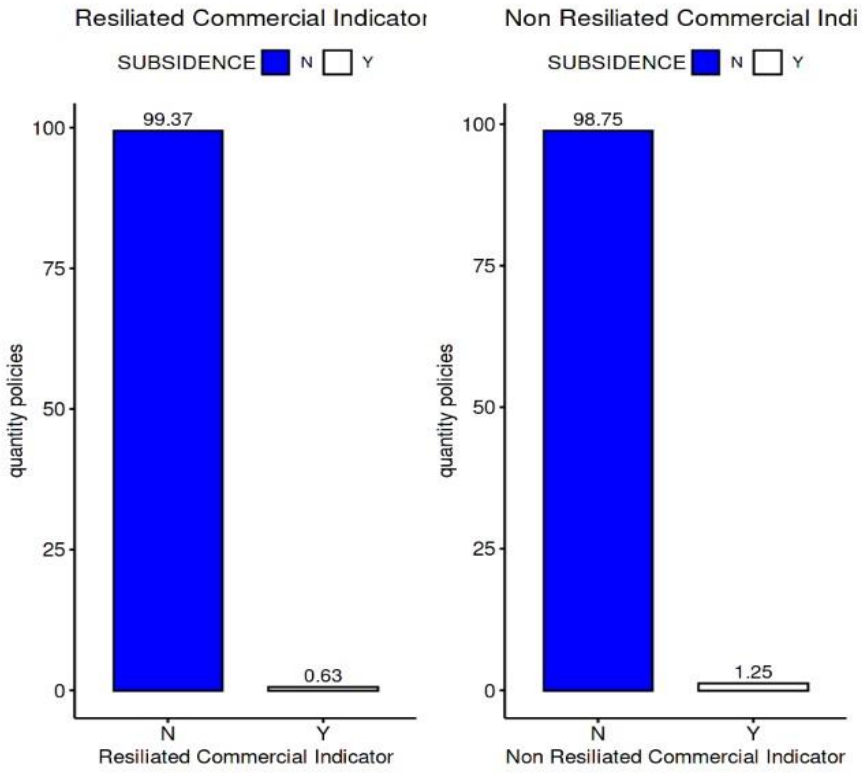


Identify premium pricing attributes for home insurance using R

Bar chart for Resiliated (SEC_DISC_REQ) and non- Resiliated (SEC_DISC_REQ)

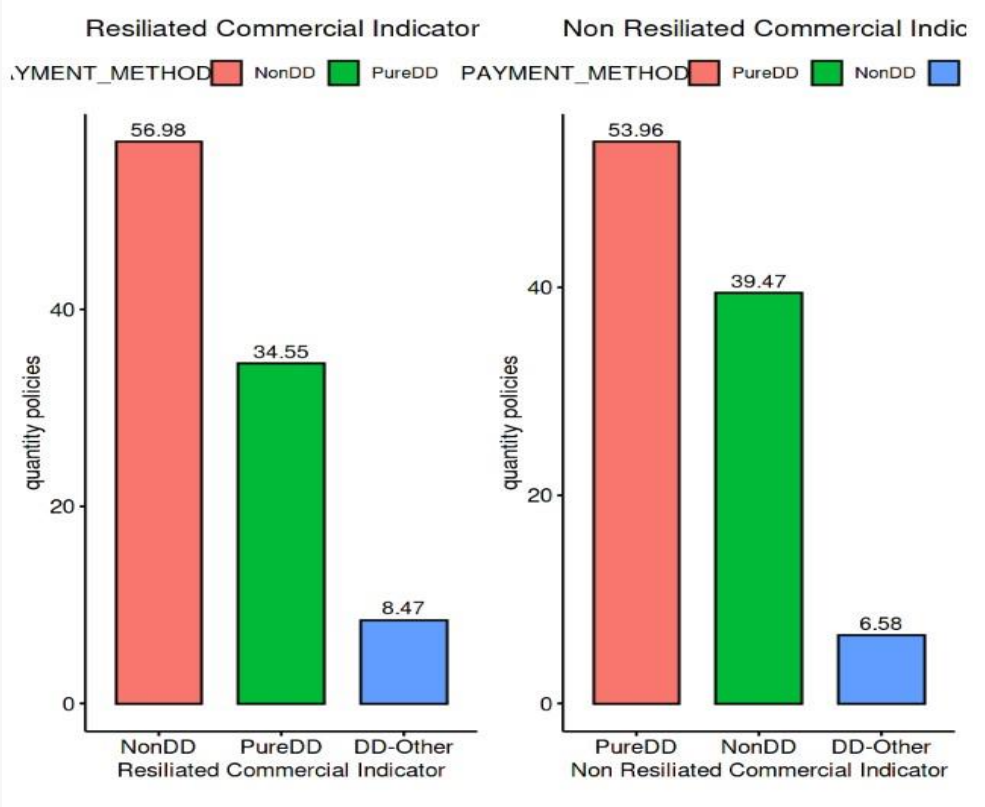


Bar chart for Resiliated (SUBSIDENCE) and non_ Resiliated (SUBSIDENCE) is given below

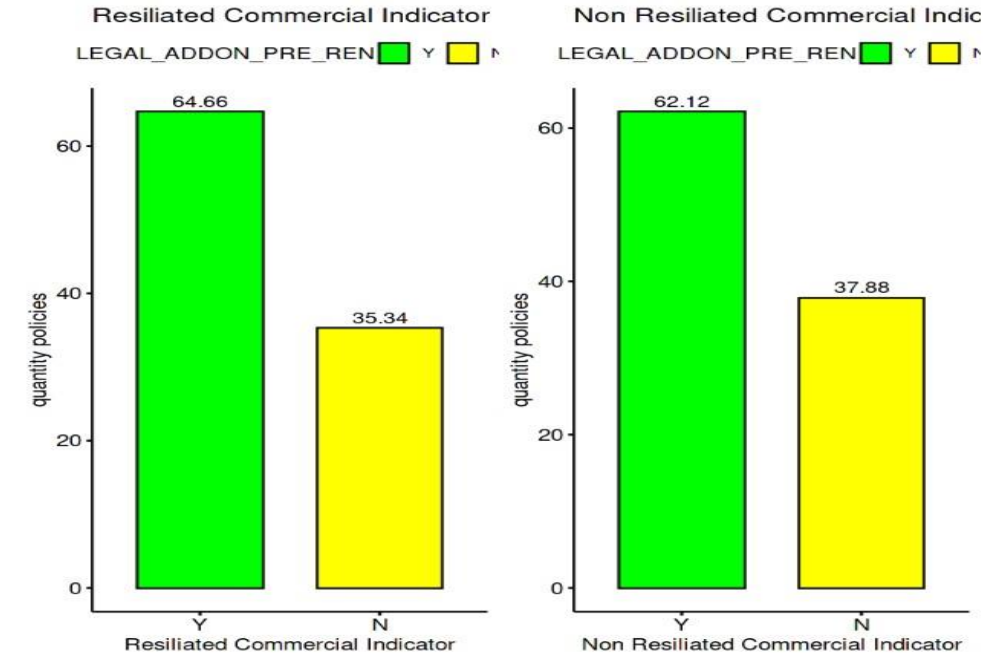


Identify premium pricing attributes for home insurance using R

Bar chart for Resiliated (PAYMENT_METHOD) and non_ Resiliated (PAYMENT_METHOD) is given below

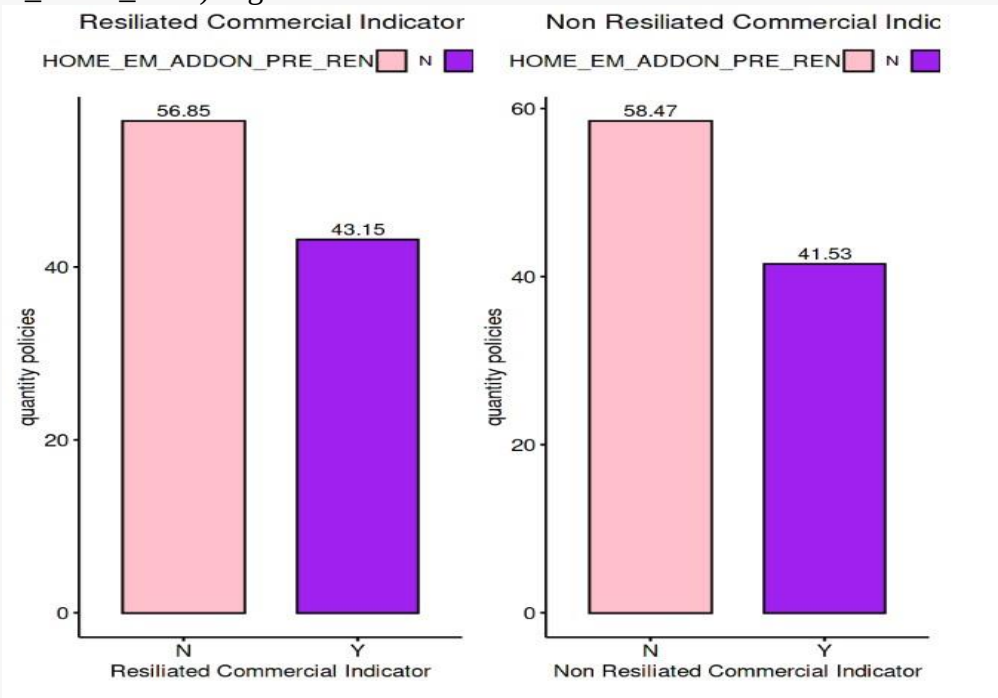


Bar chart for Resiliated (LEGAL_ADDON_PRE_REN) and non- Resiliated (LEGAL_ADDON_ PRE_REN)

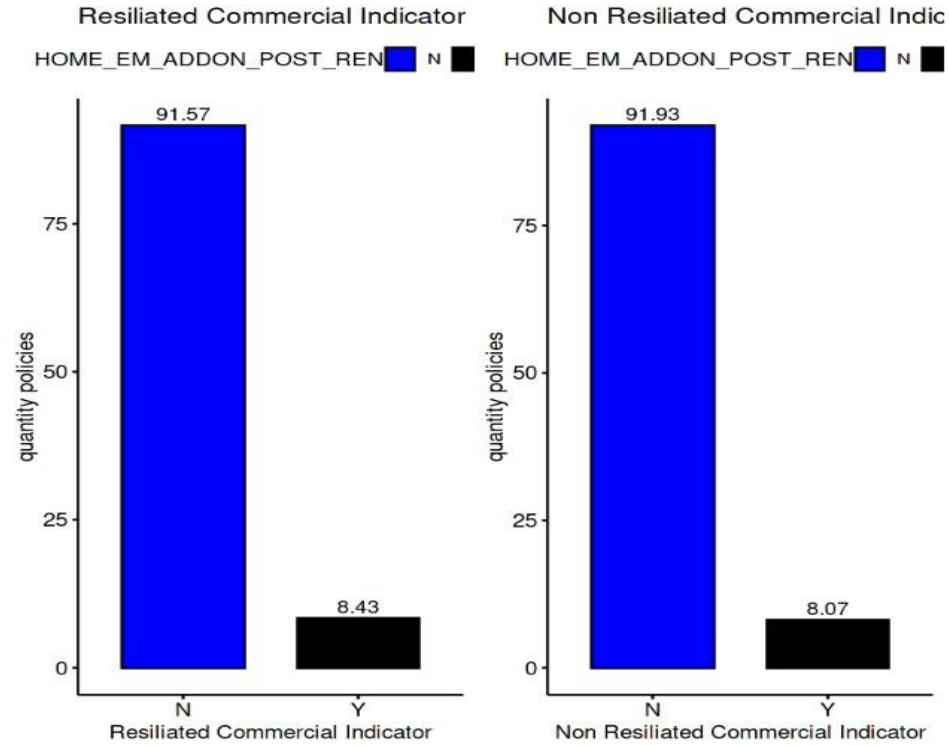


Identify premium pricing attributes for home insurance using R

Bar chart for Resiliated (LEGAL_ADDON_POST_REN) and non- Resiliated (LEGAL_ADDON_POST_REN) is given below:

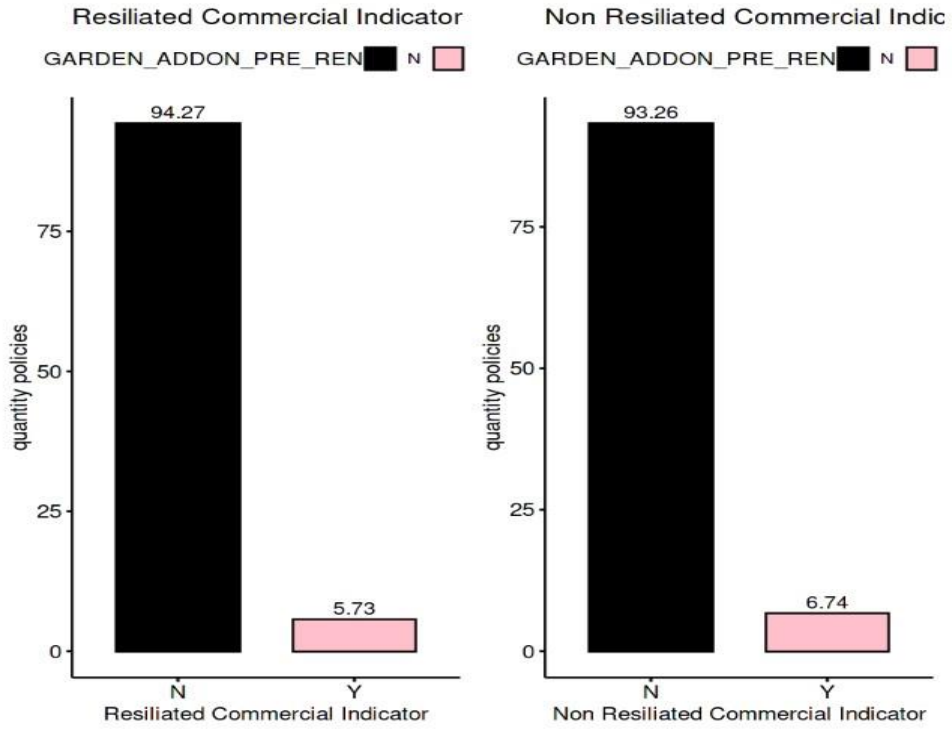


Bar chart for Resiliated (HOME_EM_ADDON_POST_REN) and non- Resiliated (HOME_EM_ADDON_POST_REN)

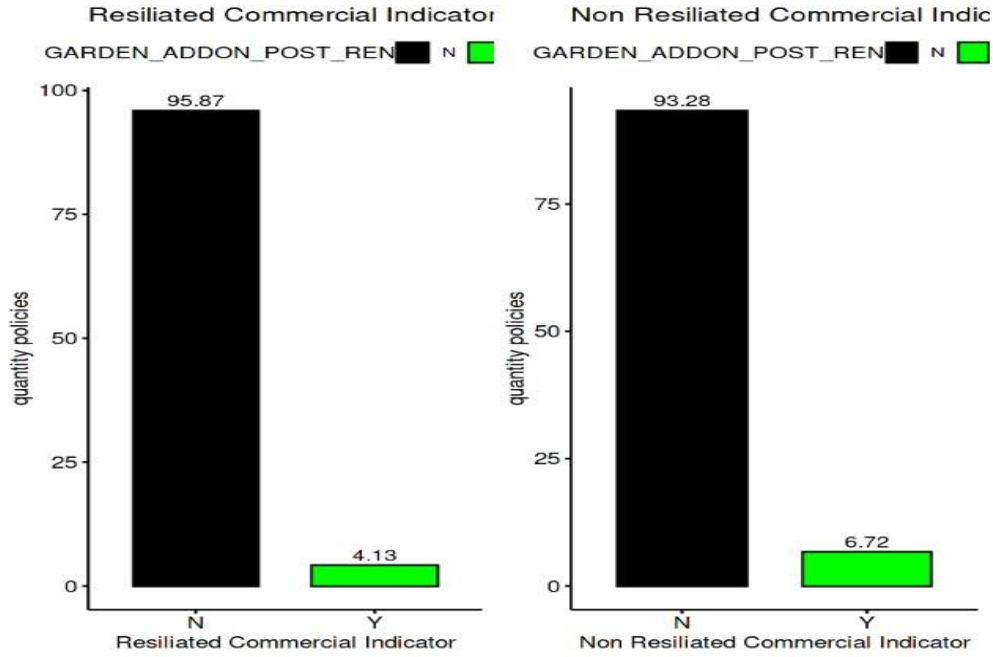


Identify premium pricing attributes for home insurance using R

Bar chart for Resiliated (GARDEN_ADDON_PRE_REN) and non- Resiliated (GARDEN_ADDON_PRE_REN) is given below

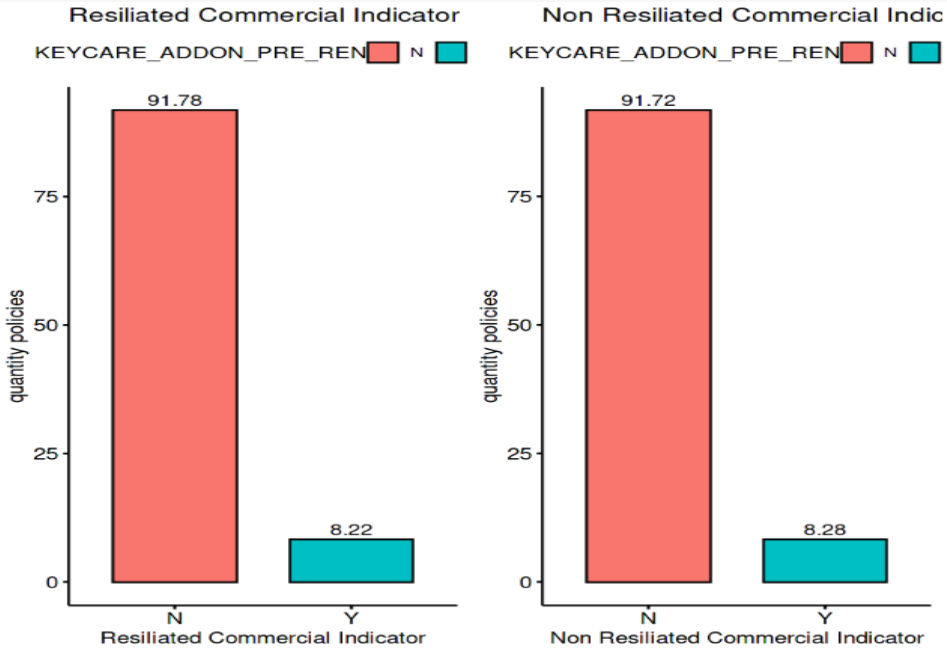


Bar chart for Resiliated (GARDEN_ADDON_POST_REN) and non for Resiliated (GARDEN_ADDON_POST_REN) is given below

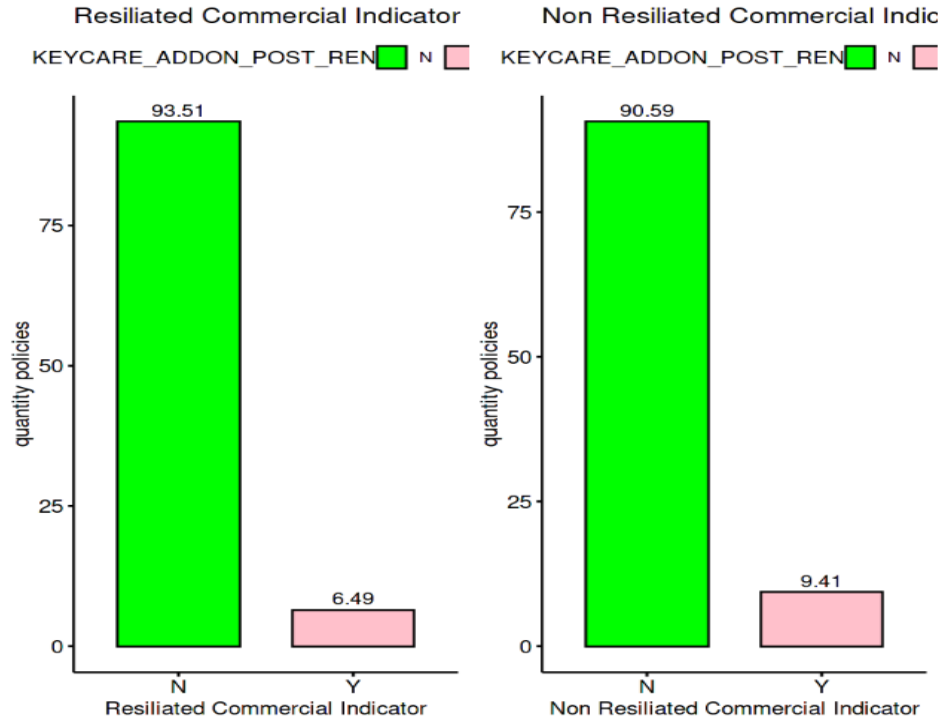


Identify premium pricing attributes for home insurance using R

Bar chart for Resiliated (KEYCARE_ADDON_PRE_REN) and non- Resiliated (KEYCARE_ADDON_PRE_REN) is given below

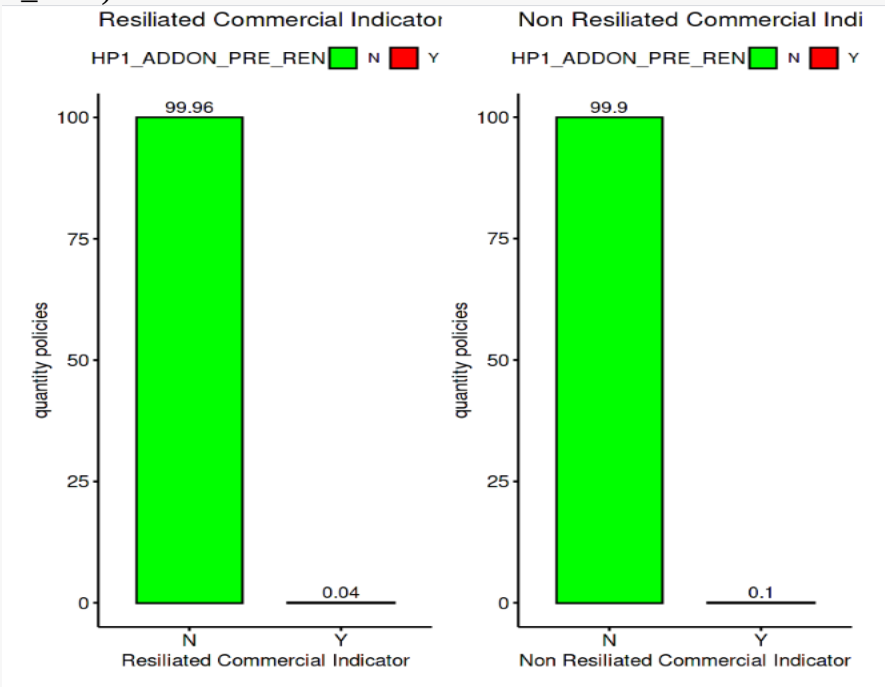


Bar chart for Resiliated (KEYCARE_ADDON_POST_REN) and non- Resiliated (KEYCARE_ADDON_POST_REN) is given below

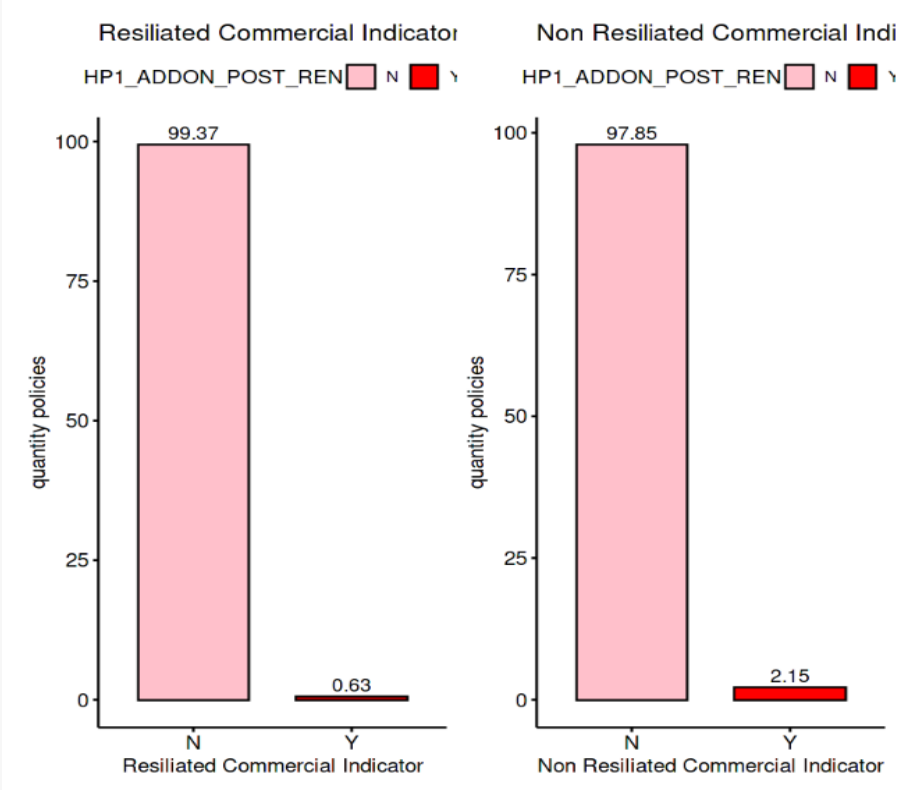


Identify premium pricing attributes for home insurance using R

Bar chart for Resiliated (HP1_ADDON_PRE_REN) and non- for Resiliated (HP1_ADDON_PRE_REN)

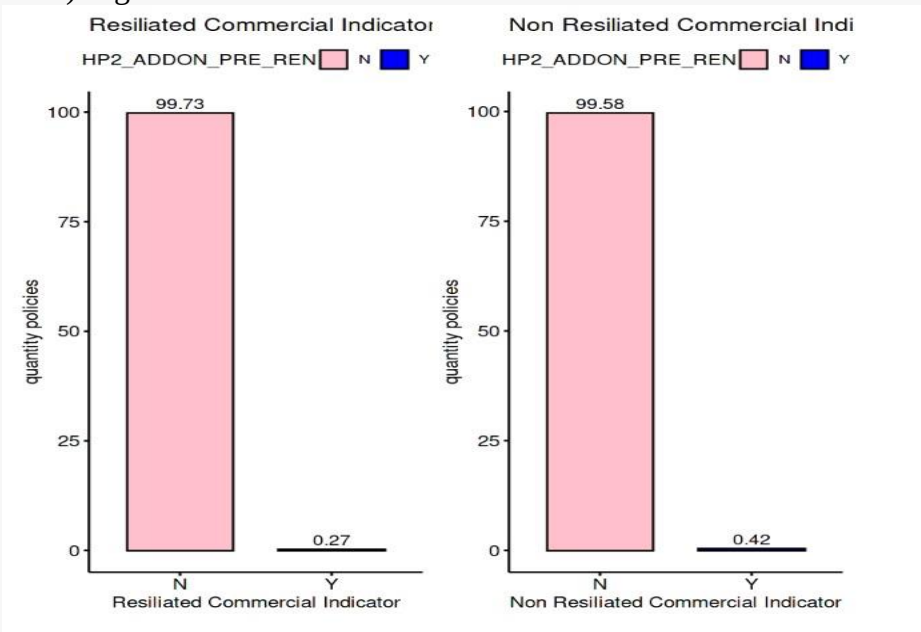


Bar chart for Resiliated (HP1_ADDON_POST_REN) and non- Resiliated (HP1_ADDON_POST_REN)

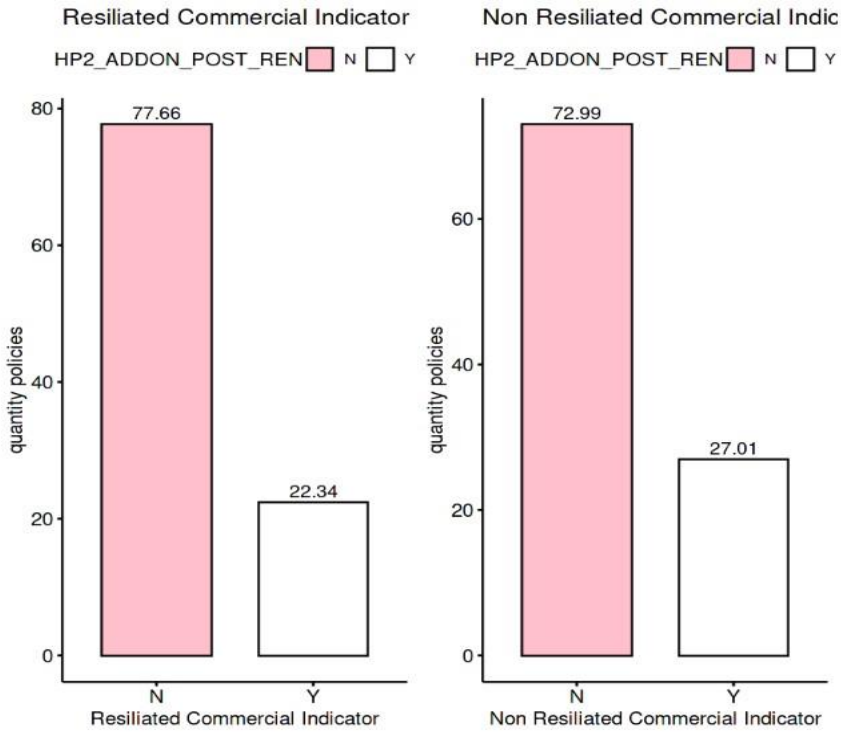


Identify premium pricing attributes for home insurance using R

Bar chart for Resiliated (HP2_ADDON_PRE_REN) And non-Resiliated (HP2_ADDON_PRE_REN) is given below

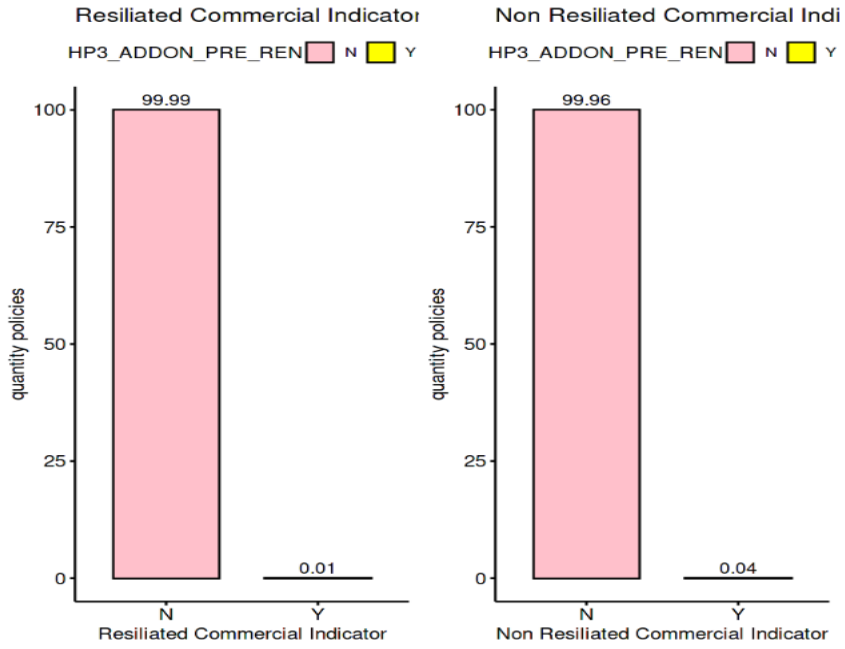


Bar chart for Resiliated (HP2_ADDON_POST_REN) and non- Resiliated (HP2_ADDON_POST_REN)

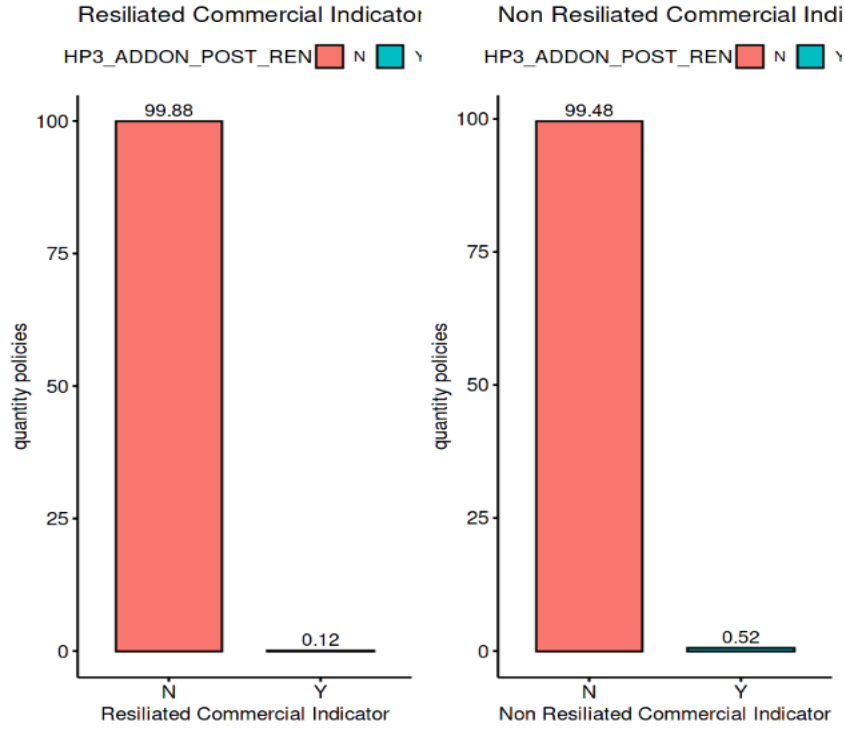


Identify premium pricing attributes for home insurance using R

Bar chart for Resiliated (HP3_ADDON_PRE_REN) and non- for Resiliated (HP3_ADDON_PRE_REN)

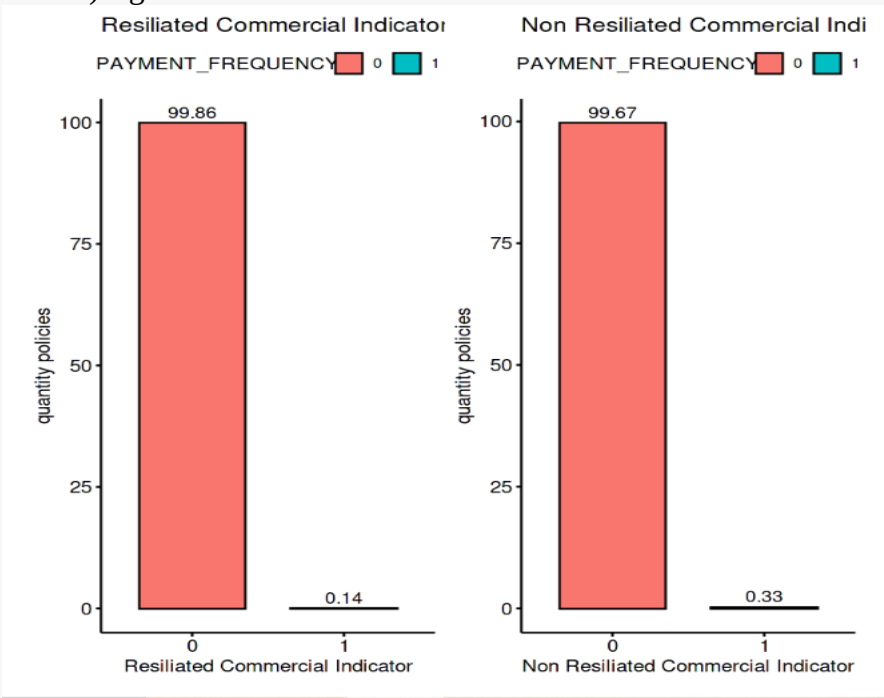


Bar chart for Resiliated (HP3_ADDON_POST_REN) and non- Resiliated (HP3_ADDON_POST_REN)

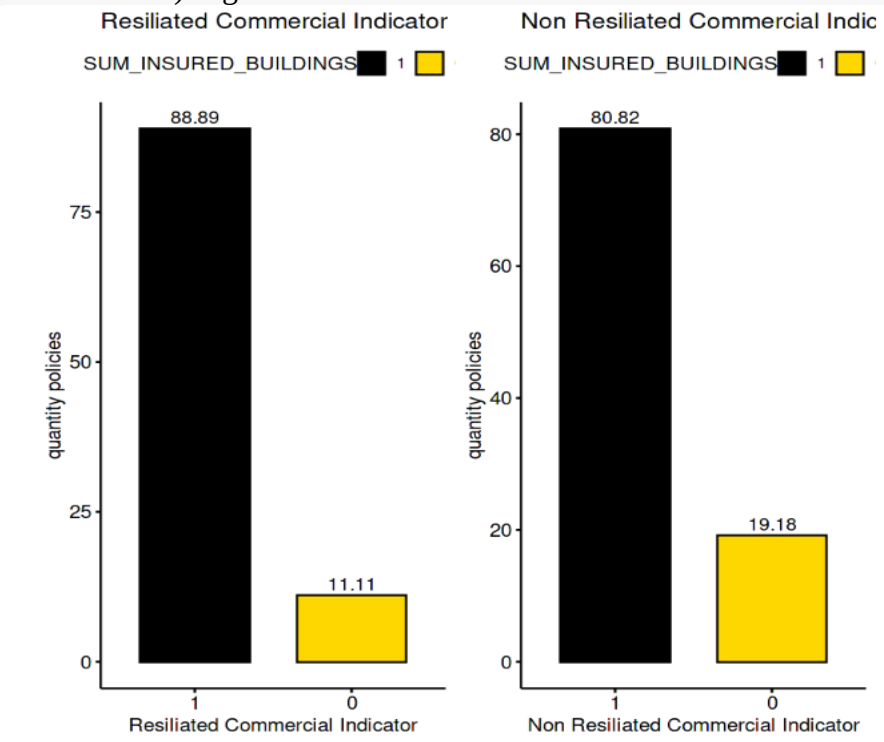


Identify premium pricing attributes for home insurance using R

Bar chart for Resiliated (PAYMENT_FREQUENCY) and non Resiliated (PAYMENT_FREQUENCY)is given below

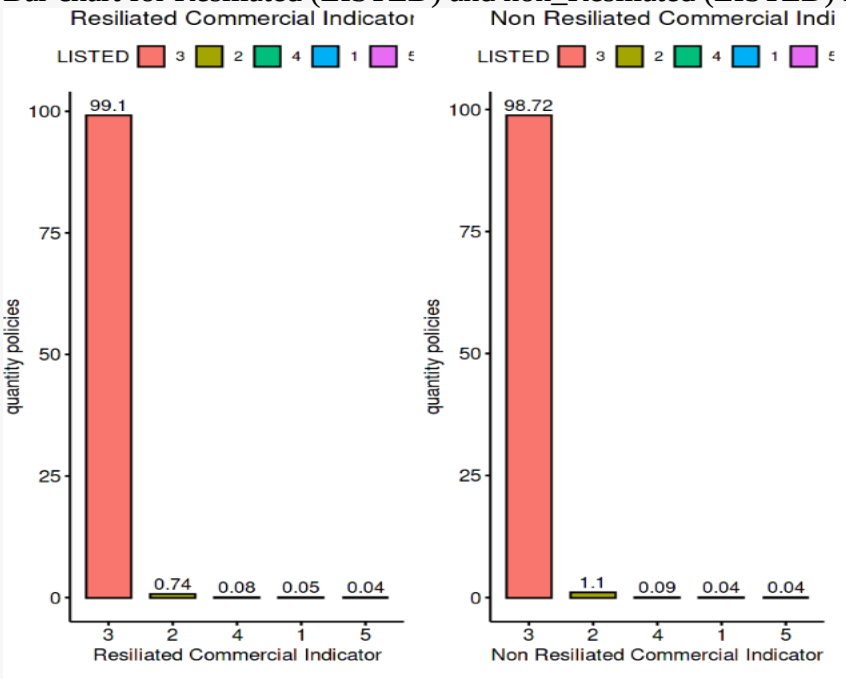


Bar chart for Resiliated (SUM_INSURED_BUILDINGS) and non Resiliated (SUM_INSURED_BUILDINGS) is given below



Identify premium pricing attributes for home insurance using R

Bar chart for Resiliated (LISTED) and non_Resiliated (LISTED) is given below



Correlations

As we can see in the **Geographical Classification of Risk - Building and Assured Sum - Building** plot, the Pearson Coefficient of the two aforementioned quantities is **-0.012** which suggests that there is not a tangible correlation. In other words, the buildings Geographical Classification of Risk and the Assured Sum are independent quantities. The points fall close to the line, which indicates that there is a strong negative relationship between the variables.

We can see that for those premium clients with Total for the previous year bonus (more than 0/null) are between 0 and 1,000.Now i will check the most popular and most successful months and day for quotation and cover start.

In the Most successful months in Quotation date plot
It appears that **January** is the most popular month when it comes to policies quotations. This is maybe for the salary bonus that employees get in christmas and the clients makes theirs plan come true making the quote.

Most successful months in Coverage date plot It seems that the beggining and ending months of the year have the highest number of policies. This can be attributed to the fact that people tend to dedicate their money in the summer to vacances when the kids are out of school, the parents are on vacation and therefore, the policies is more likely to be practically none.

Now i want to know if there are months that tend to be more successful than others. For this i will create a boxplot between the total coverage and the months.

Identify premium pricing attributes for home insurance using R

By using box plot analysis We see that effectively the months of **January**, **February**, **March** an d **November** tend to yield the highest median returns. However **December** does not have a high total coverage, even when this month has the highest number of policies. On the other hand **Jun e** and **August** have a not so high coverage and finally the resting months are the least successful months on the aforementioned metrics. Again, the success of the starting and ending months can be attributed to the fact that in summer the people tend to spend their money on vacation stuffs.

Number of buildings by year

We notice that there is a rise in the number of buildings constructed the **1870s** decade and there is a peak in the **1940s** decade with more than **60 thousands** policies. It can be concluded that the majority of buildings have between **20** and **80** years of constructed.

Relation between age of client (at the time that made the first payment) and the duration of the policy (for those who have a cancelled policy:

Surprisingly, the median age of the clients at the moment of the coverage payment is **66** years old and the peak is around **70's**. This makes sense with the the previous discovery that most clients are retired. Now, I want to know if there is a correlation between the age of the client and the policy duration

In the scattered graphic, the Pearson Coefficient between the client's age and the policy duration (in years) is **-0.074** which suggests that there is no tangible correlation. So we can say that this quantitative variables are independent quantities. The relationship is negative because, as one variable increases, the other variable decreases.

Machine Learning:

I will try to build the next models: Linear Logistic Regression, Decision tree, Random Forest , Extreme Boost, SVM and Neuron Network. As well as Error Matrix, ROC curves and Area Under the Precision-Recall (PR) for each model. My goal with this is to get the best model to predict which clients have more probabilities to remove/resiliate their policy.

Pearson Correlation

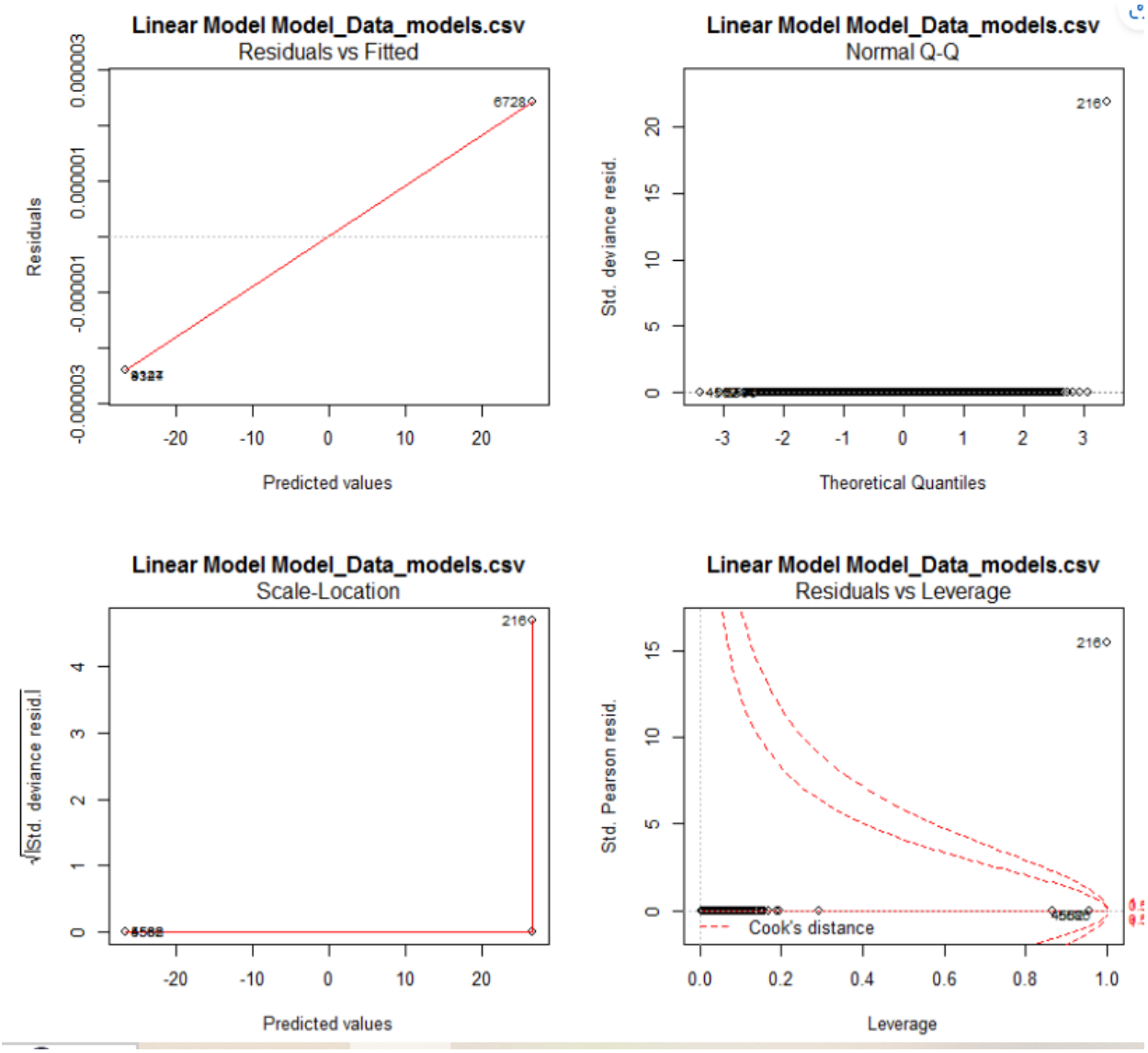
As we can see in this complete correlation analysis, there are some evidently relations like while the building has more bedroom the coverage increments (as well as the total coverage). However we can also see that there is weak inversely proportional relation between the age of the client

and the duration of the policy. This may be due to the fact that young people move a lot, either through studies, work, family, etc.

Models

Logistic Linear Regression

This is a logistic regression for the Resiliation indicator. This can be used to predict of a client will renew his policy given certains features.

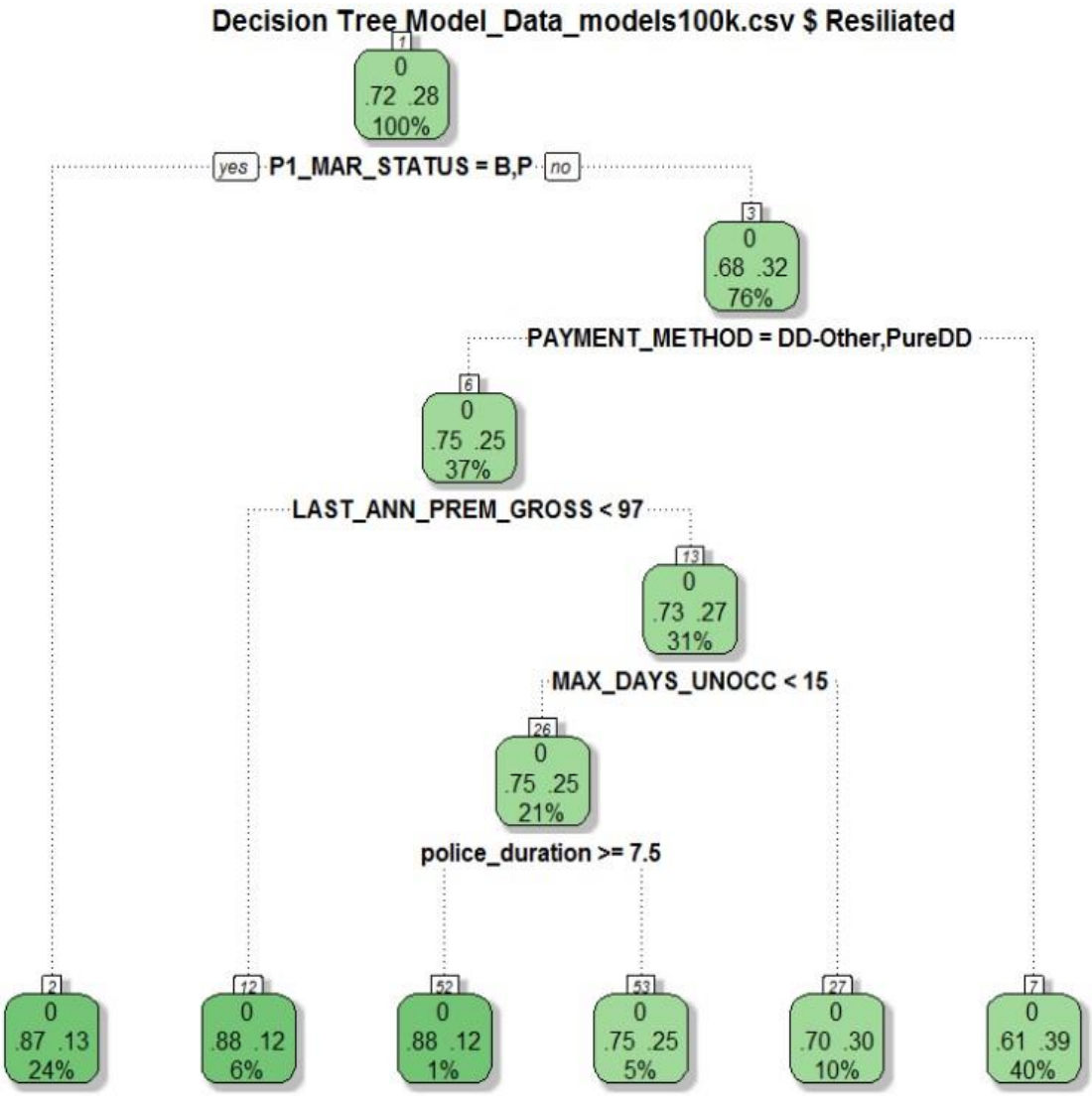


Identify premium pricing attributes for home insurance using R

We can predict if a policy will be cancelled/resiliated firstly asking if the marriage status is **B** or **P**, if yes then there is 24% probability that that client give up with his policy. On the other hand the purest subset of this decision tree is the client that doesn't have a payment method of **DD-Other** or **PureDD** which give us **40%** of probability that this client will cancel the policy. However this is not enough to be considered, so this model can be safely dropped since we did not get a strong subset that allow us to make a decision (based on this model).

Naïve Bayes: Accuracy given by naïve bayes model 65.7%

Decision Tree: Accuracy given by **Decision Tree** Model is 68.37%



Identify premium pricing attributes for home insurance using R

Random Forest: Accuracy given by **Random Forest** Model is 70.38%

XGBOOST: Accuracy given by **XGBOOST** Model is 70.52%

KNN Model: Accuracy given by **KNN** Model is 66.33%

Testing

For the testing i will choose the **Decision Tree** since it was the best predictive model to know those clients who have more probabilities to dismiss their policy.

As we saw earlier, the clients that doesn't have a marital status of **B** or **P** have already **76%** probabilities of resiliate their policy and for this subset those who have a **Non-DD (Non-Direct Debit)** method of payment have a 40% of resiliate the policy too. This means that the clients that don't have a financial engagement with the company are more likely to cancel their policy (since they are free to go when they want to).

Solution Proposed

Already explored all the data, models and discovered some insights, I suggest to this Insurance company to offer extra privileges if the clients choose the **Direct-Debit** payment method, since the clients with financial engagement are the ones more likely to stay in the insurance company and therefore are the most loyal to it. I also suggest that it will be better to do this "extra privileges" marketing in the **first and last trimester of the year**, since these are the months with the best earnings for the company and last but not least I suggest to aim this marketing specially to the **Retired** people since they are the ones who buy the most insurance in the history of the company.