Required Libraries:

I am using regular libraries like

- NUMPY
- PANDAS
- MATPLOTLIB
- SEABORN

  for Data Collection, EDA and Visualization etc.,

Apart from that I am using

- Ydata Profiling – for deep analysis
- DATETIME – to handle date columns
- SCIPY.STATS – for applying hypothesis

Data Collection:

A 'csv' file named 'home_insurance' with 256136 rows and 66 columns (given) was uploaded in python platform using panda's library.

Since the data is huge, I am using google colab instead of my local machine to avoid overload.

The data contains datatype of int, float and object.

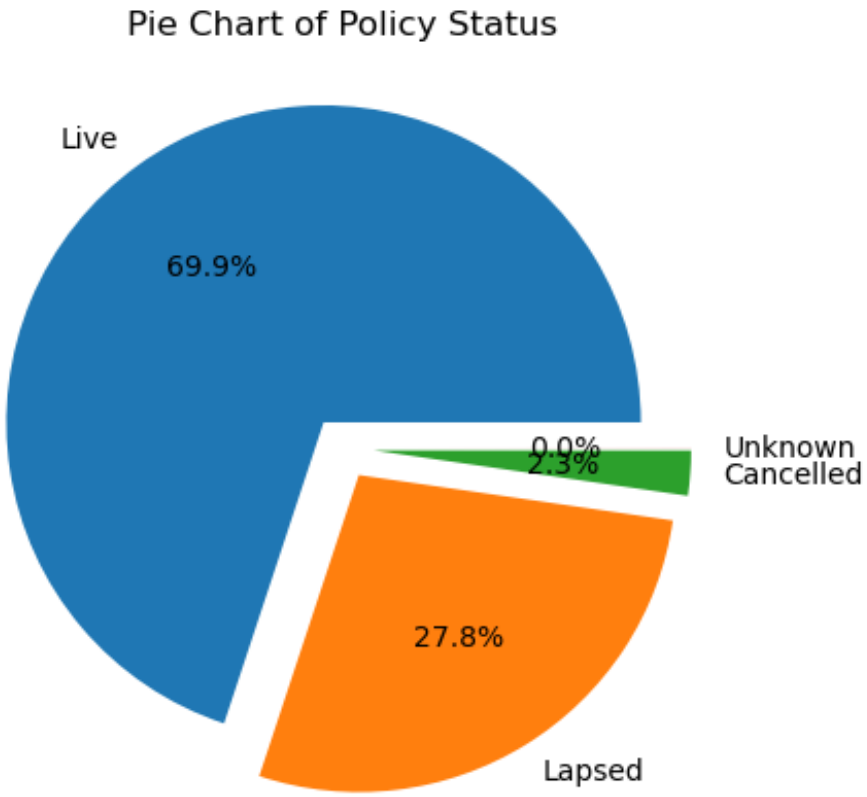
Since I did not get any description on categorical data, I am just mentioning those values according to the labels.

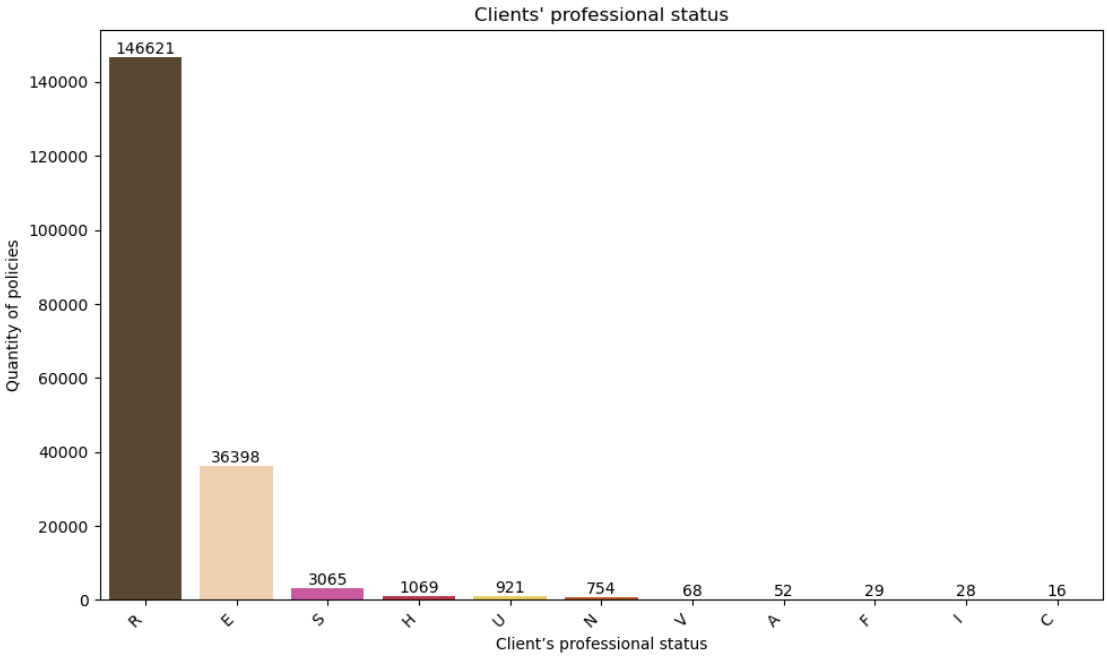
Data Cleaning:

Data contains many missed values for entire rows for columns and more than 60% of missing data in some columns

After cleaning the data (with the guidance of my mentor), we are left with 189021 rows and 59 columns.

Data Analysis:

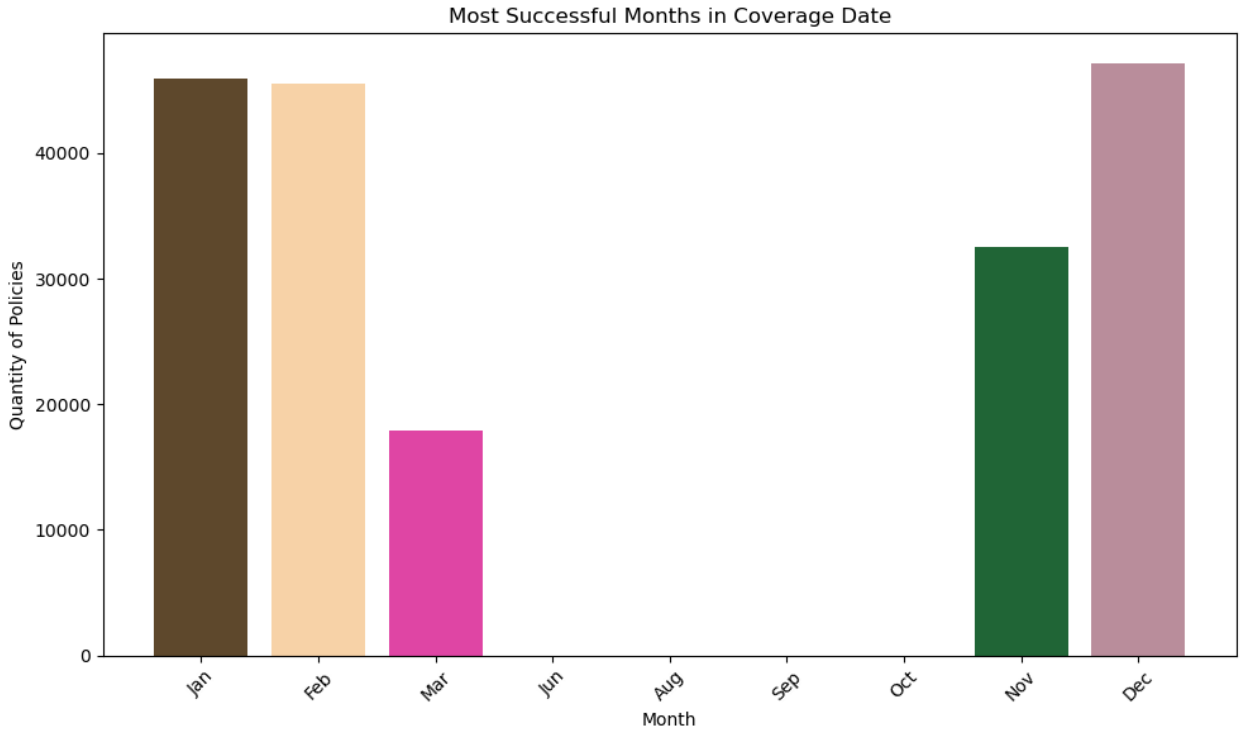




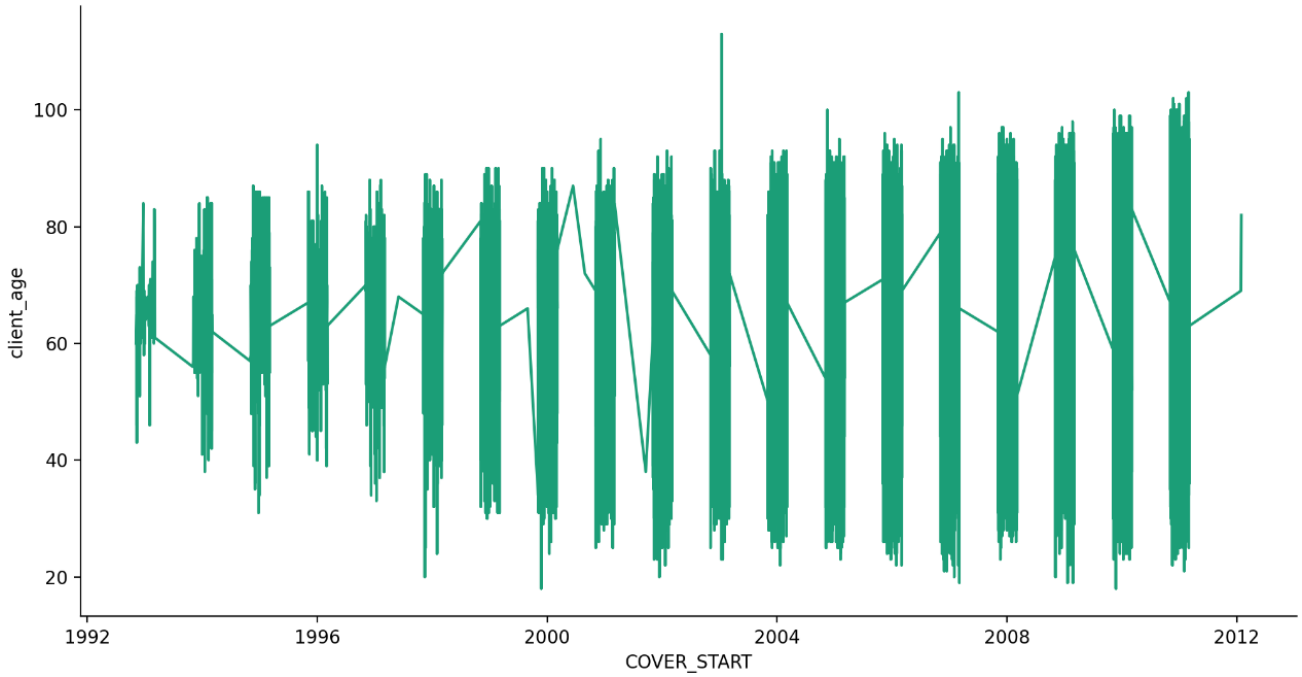
We can see that approx. (~70%) policies are live, with more than 10 lakhs dollars with R labelled professional clints possess most of it.

We can also see that beginning and ending of months are busy for company as new policies were registered in these months.

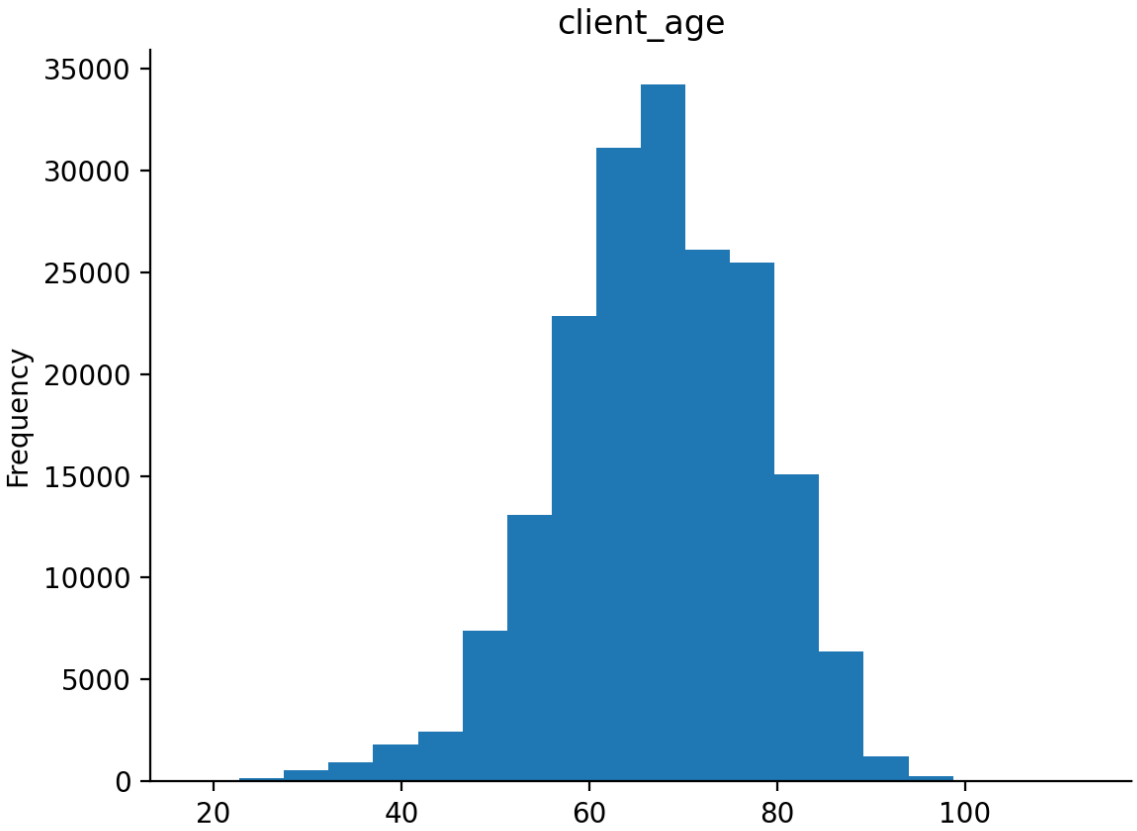
We can also see that beginning and ending of months are busy for company as new policies were registered in these months.



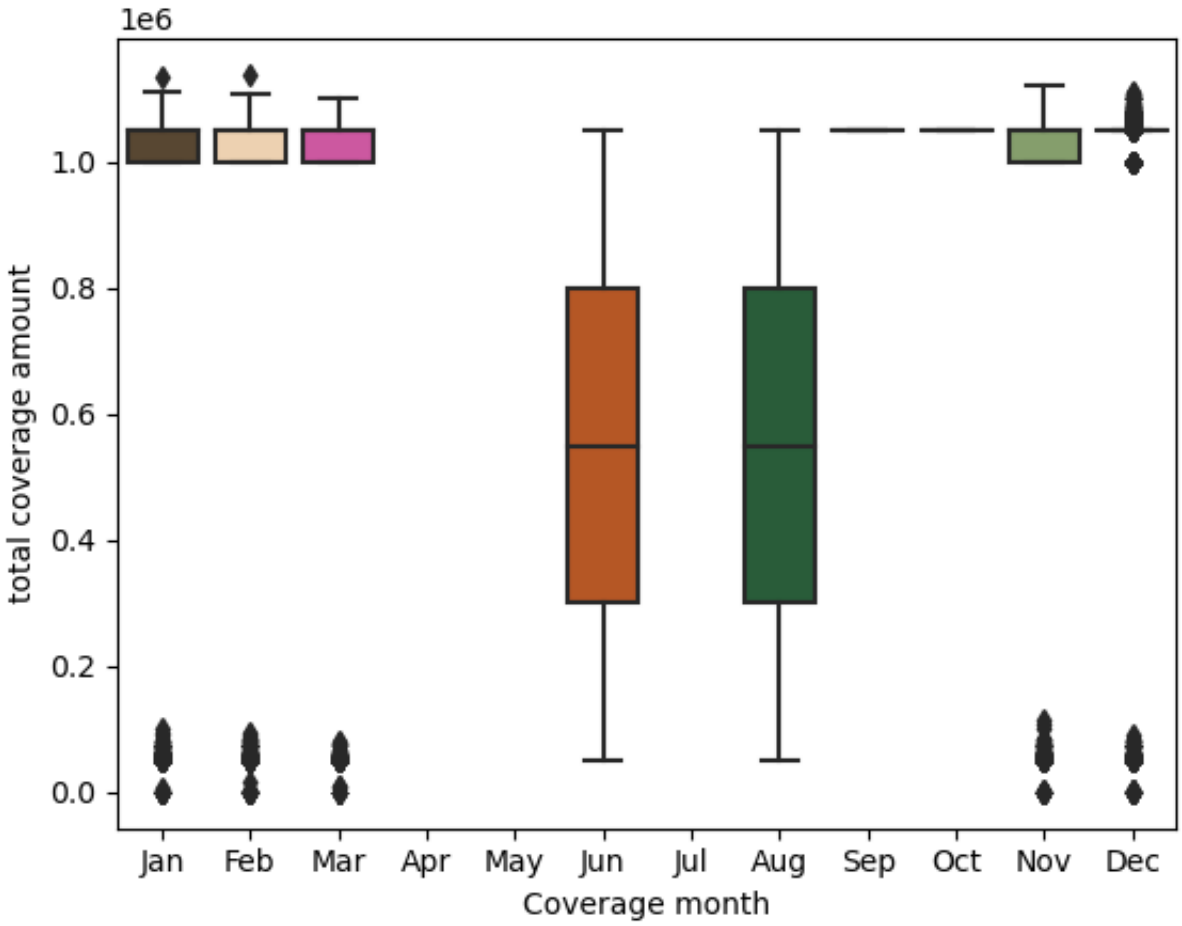
We can also see that beginning and ending of months are busy for company as new policies were registered in these months.



Between 2000 and 2004, we can see that most of the age groups are interested in taking new policies. Which I believe some significant event or intervention has impacted. So, I recommend to analyse the events occurred during that timeline.

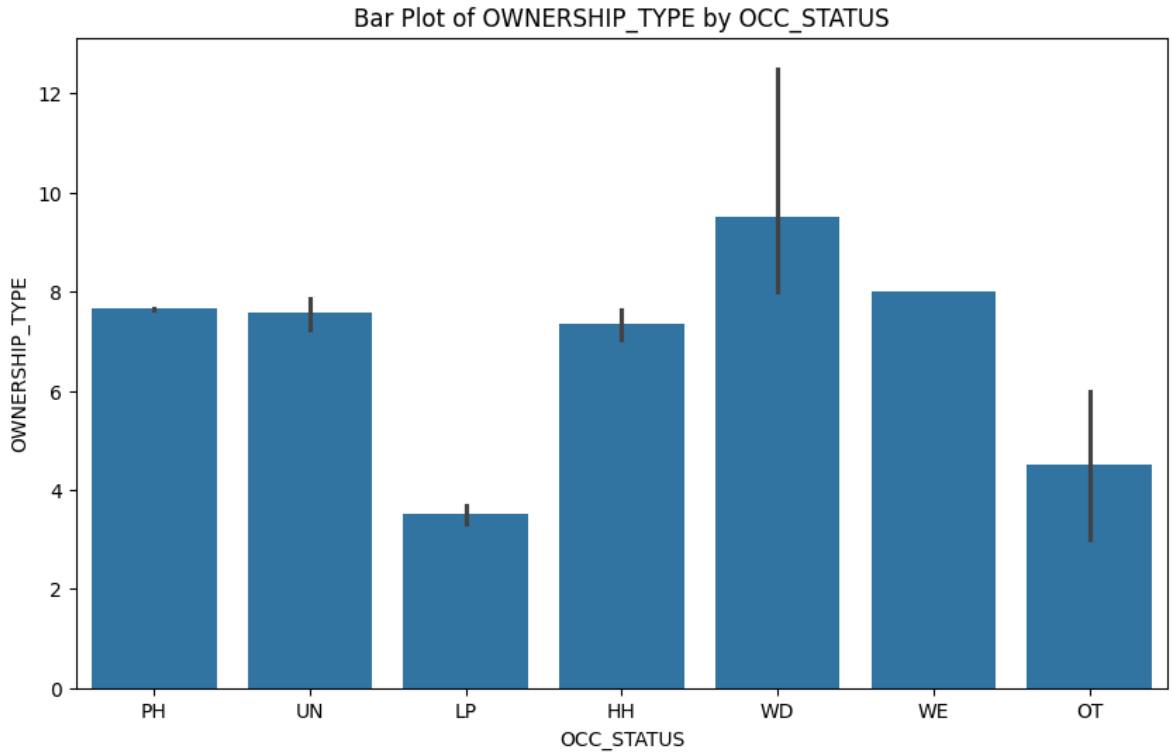


Clients from age group of 60-70 are showing much interest in policies. So much focus should be kept on certain age groups.

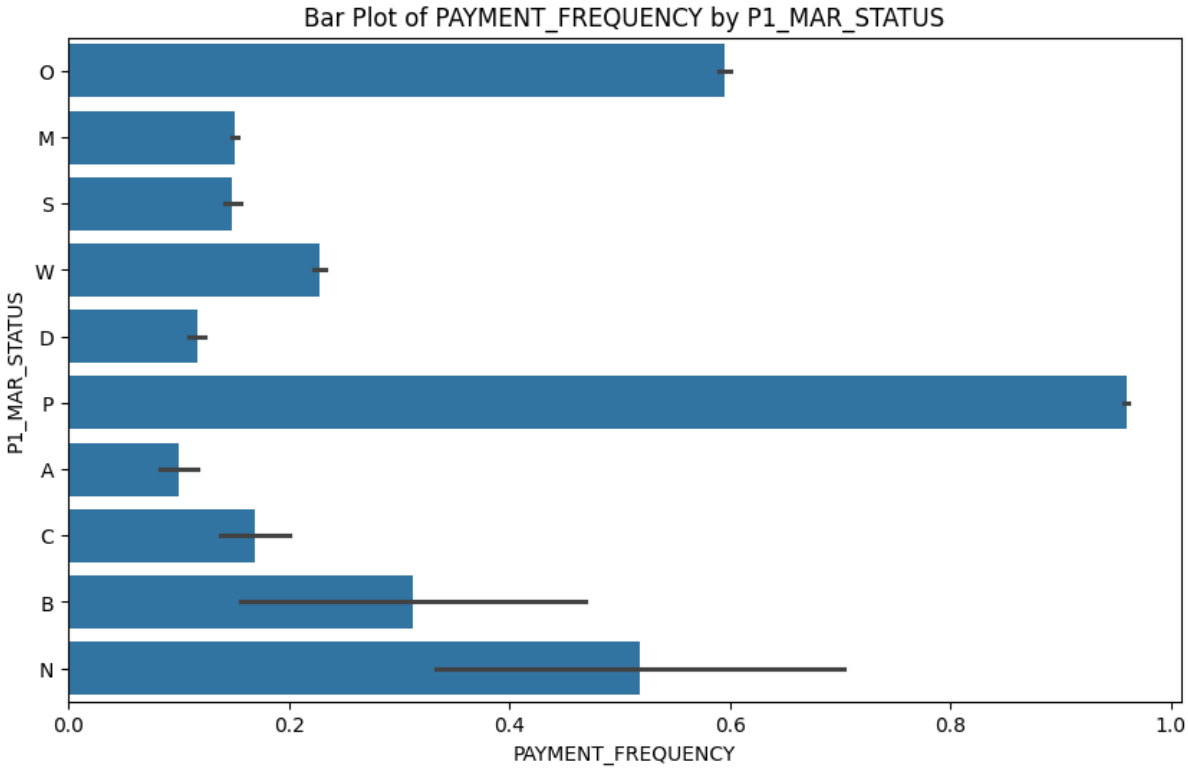


We see that effectively the months of January, February, March and November tend to yield the highest median returns. However, December does not have a high total coverage, even when this month has the highest number of policies.

On the other hand, June and August have a not so high coverage and finally the resting months are the least successful months. 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.

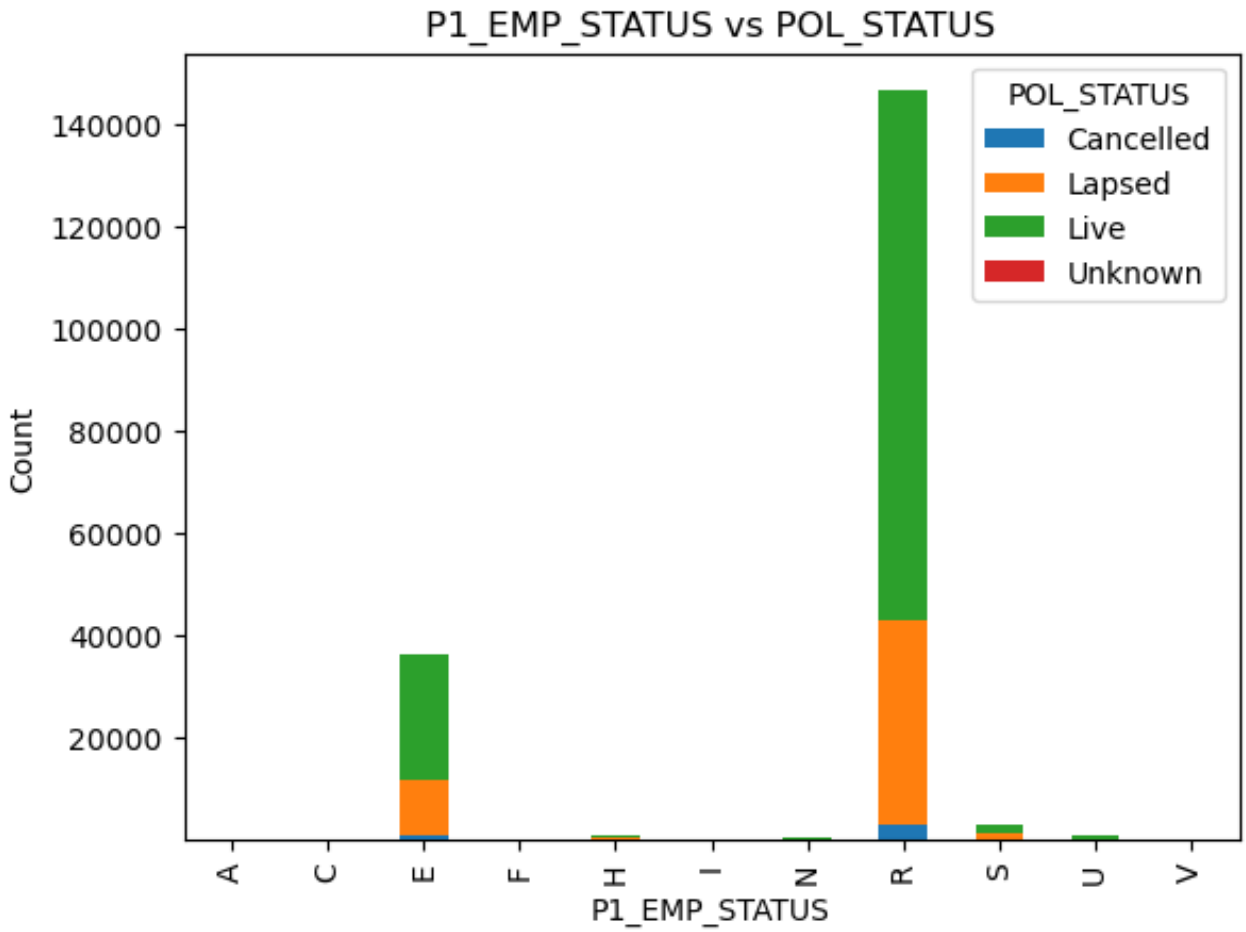


The persons WD labelled occupational status has high number of ownership type.



The clients with marital status of labelled P are paying premium much frequently. Where with marital status of labelled A have least.

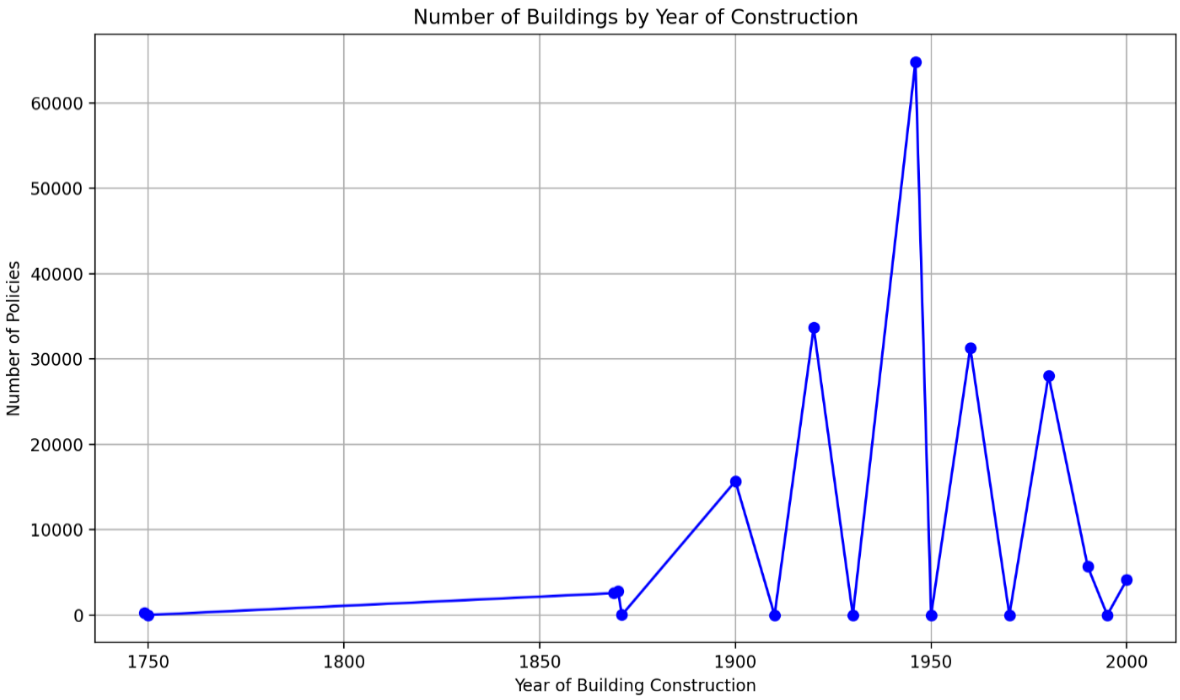
So, it would be wiser to give preference to the client with marital status of labelled P.



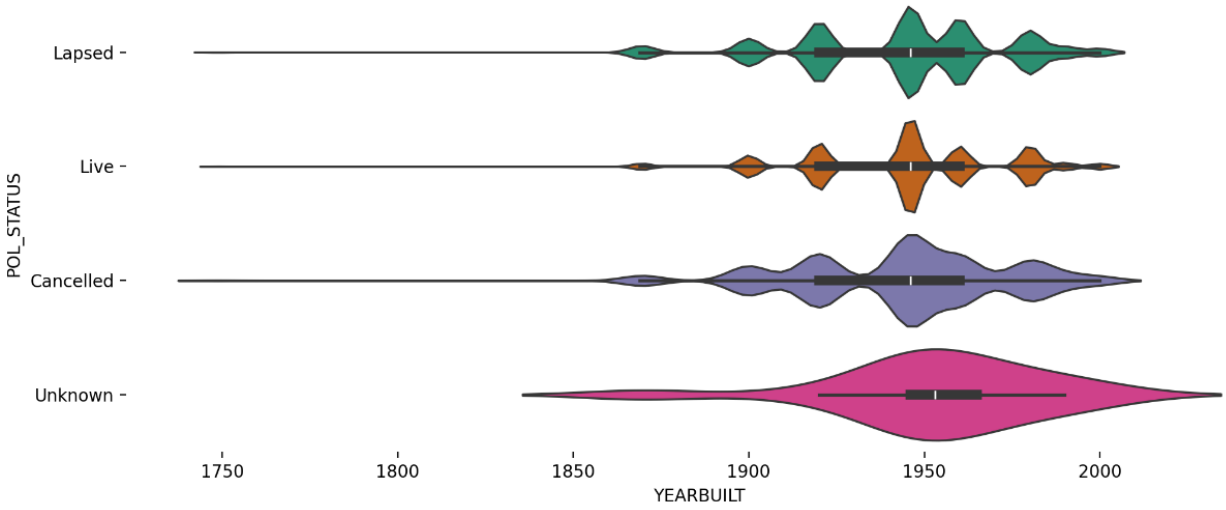
The clients with employment status R possess more Live policies followed by E.

So, marketing strategies must increase and also research must be conducted for remaining groups and provide encouraging offers according to their need.

For R and E category, I am suggesting for provide loyalty points and intensive's so that they will encourage others of their circle group to join the policy.



From the graph we can conclude that the majority of buildings have between 20 and 80 years of constructed approximately.



Live:

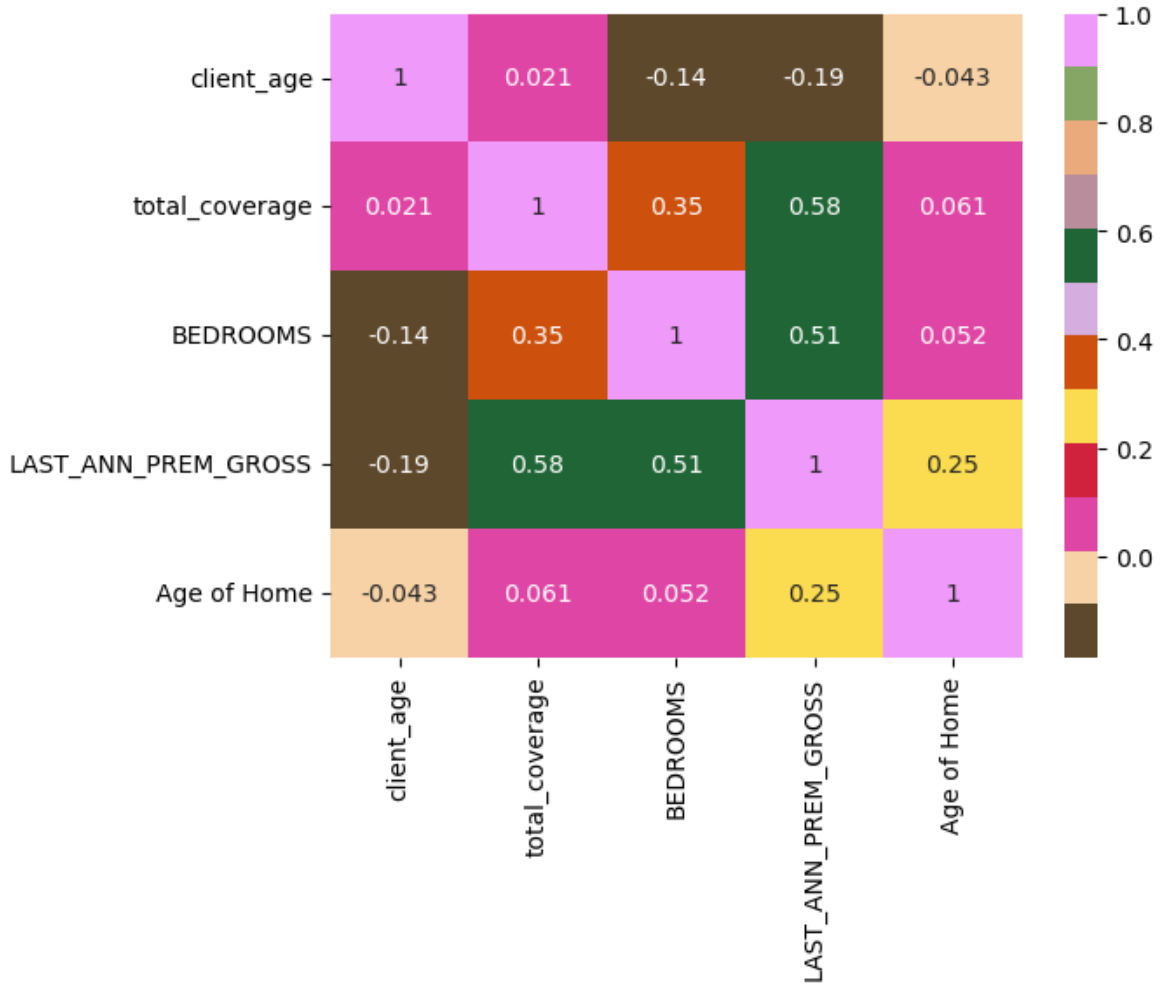
There is a noticeable concentration of data points around the year 1900, indicating a high number of buildings constructed during this period.

Lapsed, Cancelled, and Unknown:

These categories have wider distributions, suggesting more variability in the number of buildings constructed over time.

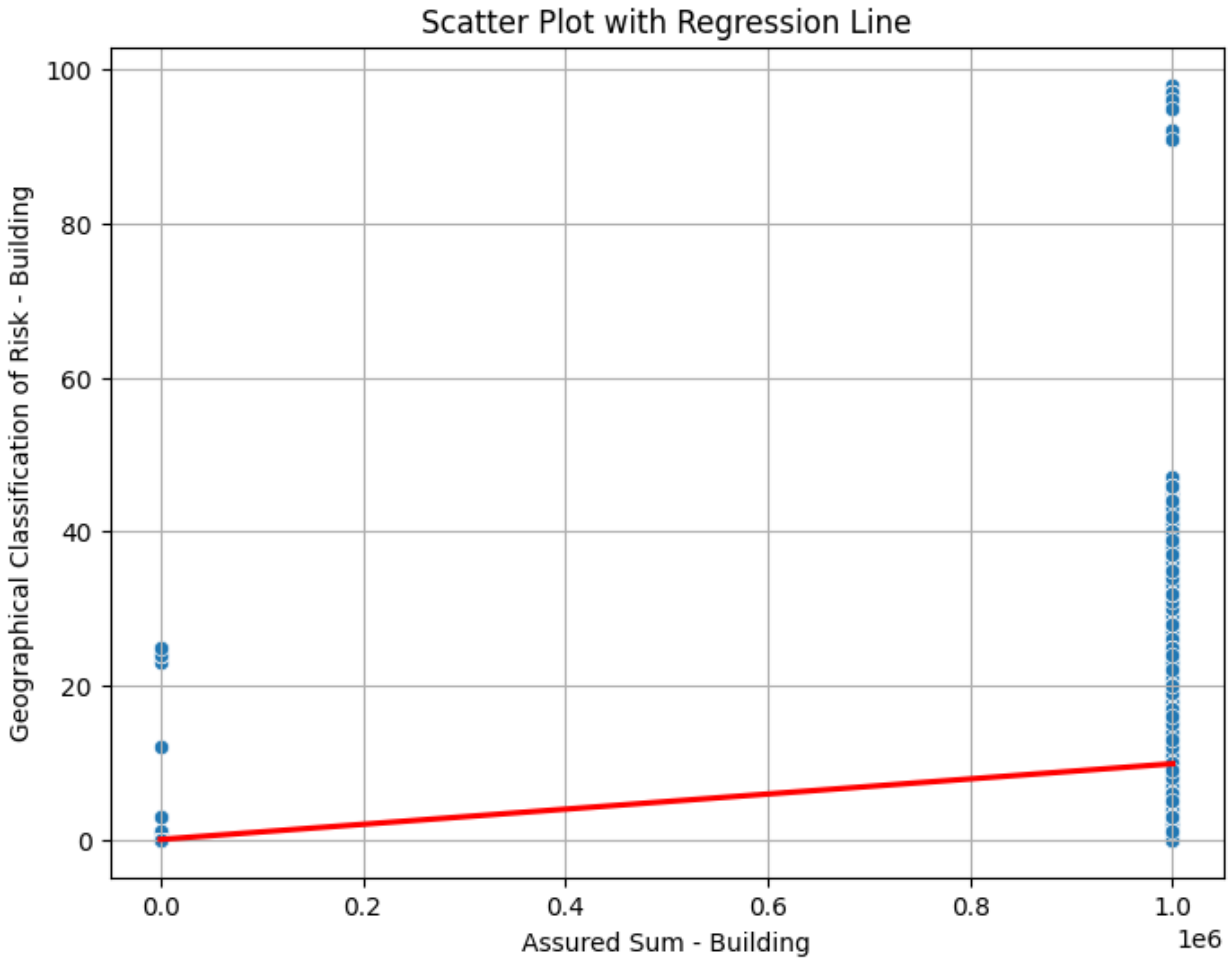
Possible Reasons for Peaks:

- Historical Events: Significant historical events, such as the Industrial Revolution and post-World War II economic boom, could have led to increased construction activity.
- Economic Conditions: Periods of economic prosperity often result in higher construction rates.
- Government Policies: Initiatives and policies promoting construction could also contribute to these peaks.



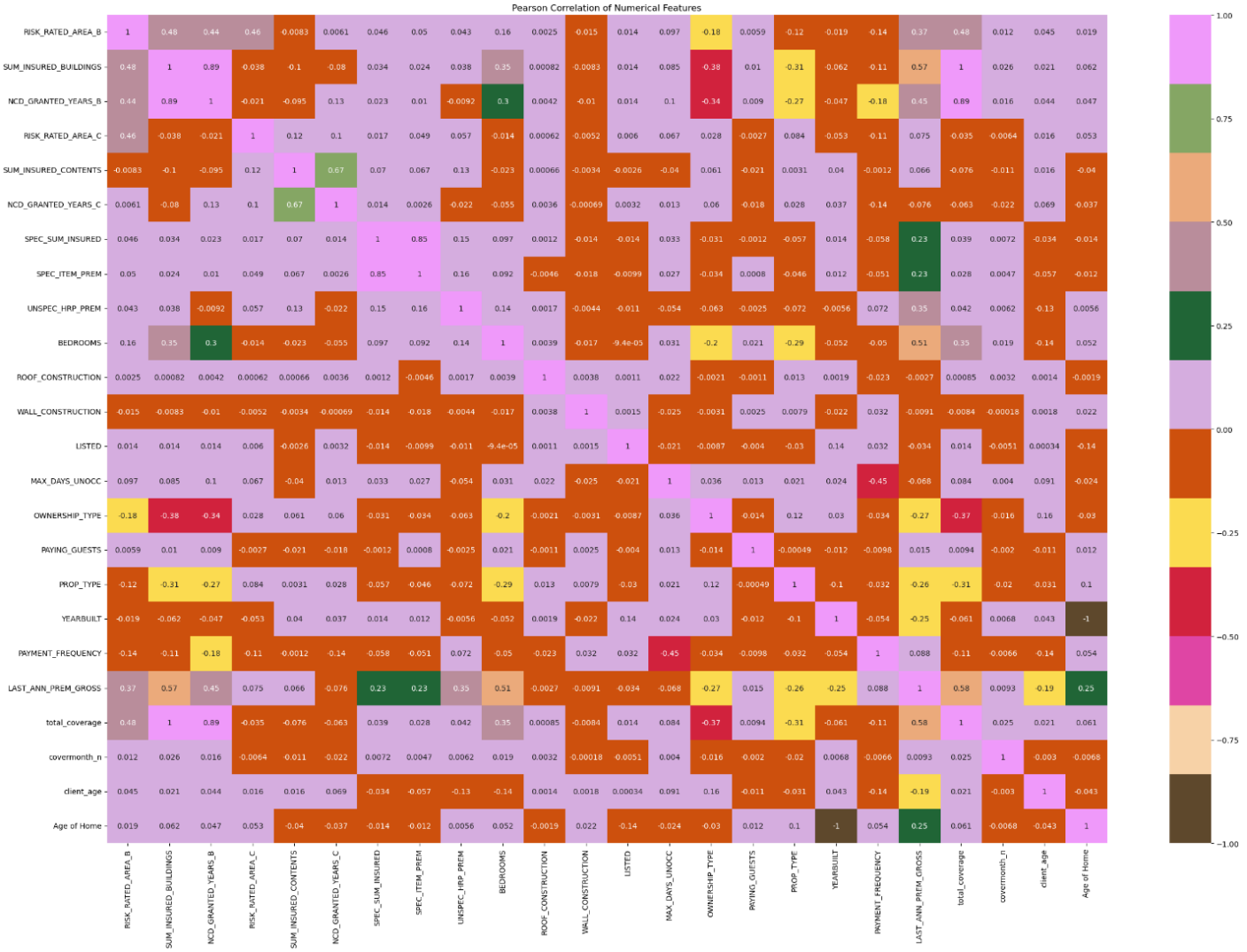
From the above heatmap, we can say that, The older buildings have more insurance coverage,

Lesser the client age, more chances of getting insured



The upward slope of the regression line suggests that higher insured sums are associated with higher geographical risk classifications.

Most data points are clustered near the origin, indicating that many buildings have lower insured sums and lower geographical risk classifications.



And finally, the above correlation matrix provides easy understanding of relationship among the features.