REPORT ON HOME INSUARNCE DATASET

1.Understanding data

Understanding the column names which is in dataset

1. **QUOTE_DATE:** Day where the quotation was made
2. **COVER_START:** Beginning of the cover payment
3. **CLAIM3YEARS:** 3 last years loss
4. **P1_EMP_STATUS:** Client's professional status
5. **P1_PT_EMP_STATUS:** Client's part-time professional status
6. **BUS_USE:** Commercial use indicator
7. **CLERICAL:** Administration office usage indicator
8. **AD_BUILDINGS:** Building coverage - Self damage
9. **RISK_RATED_AREA_B:** Geographical Classification of Risk - Building
10. **SUM_INSURED_BUILDINGS:** Assured Sum - Building
11. **NCD_GRANTED_YEARS_B:** Bonus Malus - Building
12. **AD_CONTENTS:** Coverage of personal items - Self Damage
13. **RISK_RATED_AREA_C:** Geographical Classification of Risk - Personal Objects
14. **SUM_INSURED_CONTENTS:** Assured Sum - Personal Items
15. **NCD_GRANTED_YEARS_C:** Malus Bonus - Personal Items
16. **CONTENTS_COVER:** Coverage - Personal Objects indicator
17. **BUILDINGS_COVER:** Cover - Building indicator
18. **SPEC_SUM_INSURED:** Assured Sum - Valuable Personal Property
19. **SPEC_ITEM_PREM:** Premium - Personal valuable items
20. **UNSPEC_HRP_PREM:** Unknown
21. **P1_DOB:** Date of birth of the client
22. **P1_MAR_STATUS:** Marital status of the client

23. **P1_POLICY_REFUSED:** Policy Emission Denial Indicator
24. **P1_SEX:** customer sex
25. **APPR_ALARM:** Appropriate alarm
26. **APPR_LOCKS:** Appropriate lock
27. **BEDROOMS:** Number of bedrooms
28. **ROOF_CONSTRUCTION:** Code of the type of construction of the roof
29. **WALL_CONSTRUCTION:** Code of the type of wall construction
30. **FLOODING:** House susceptible to floods
31. **LISTED:** National Heritage Building
32. **MAX_DAYS_UNOCC:** Number of days unoccupied
33. **NEIGH_WATCH:** Vigils of proximity present
34. **OCC_STATUS:** Occupancy status
35. **OWNERSHIP_TYPE:** Type of membership
36. **PAYING_GUESTS:** Presence of paying guests
37. **PROP_TYPE:** Type of property
38. **SAFE_INSTALLED:** Safe installs
39. **SEC_DISC_REQ:** Reduction of premium for security
40. **SUBSIDENCE:** Subsidence indicator (relative downwards motion of the surface)
41. **YEARBUILT:** Year of construction
42. **CAMPAIGN_DESC:** Description of the marketing campaign
43. **PAYMENT_METHOD:** Method of payment
44. **PAYMENT_FREQUENCY:** Frequency of payment
45. **LEGAL_ADDON_PRE_REN:** Option "Legal Fees" included before 1st renewal
46. **LEGAL_ADDON_POST_REN:** Option "Legal Fees" included after 1st renewal
47. **HOME_EM_ADDON_PRE_REN:** "Emergencies" option included before 1st renewal
48. **HOME_EM_ADDON_POST_REN:** Option "Emergencies" included after 1st renewal
49. **GARDEN_ADDON_PRE_REN:** Option "Gardens" included before 1st renewal
50. **GARDEN_ADDON_POST_REN:** Option "Gardens" included after 1st renewal

51. **KEYCARE_ADDON_PRE_REN:** Option "Replacement of keys" included before 1st renewal
52. **KEYCARE_ADDON_POST_REN:** Option "Replacement of keys" included after 1st renewal
53. **HP1_ADDON_PRE_REN:** Option "HP1" included before 1st renewal
54. **HP1_ADDON_POST_REN:** Option "HP1" included after 1st renewal
55. **HP2_ADDON_PRE_REN:** Option "HP2" included before 1st renewal
56. **HP2_ADDON_POST_REN:** Option "HP2" included afterrenewal
57. **HP3_ADDON_PRE_REN:** Option "HP3" included before 1st renewal
58. **HP3_ADDON_POST_REN:** Option "HP3" included after renewal
59. **MTA_FLAG:** Mid-Term Adjustment indicator
60. **MTA_FAP:** Bonus up to date of Adjustment
61. **MTA_APRP:** Adjustment of the premium for Mid-Term Adjustmen
62. **MTA_DATE:** Date of Mid-Term Adjustment
63. **LAST_ANN_PREM_GROSS:** Premium - Total for the previous year
64. **POL_STATUS:** Policy status
65. **Police:** Policy number

So there are 65 columns in dataset.

Total rows in dataset is 256136

2. 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

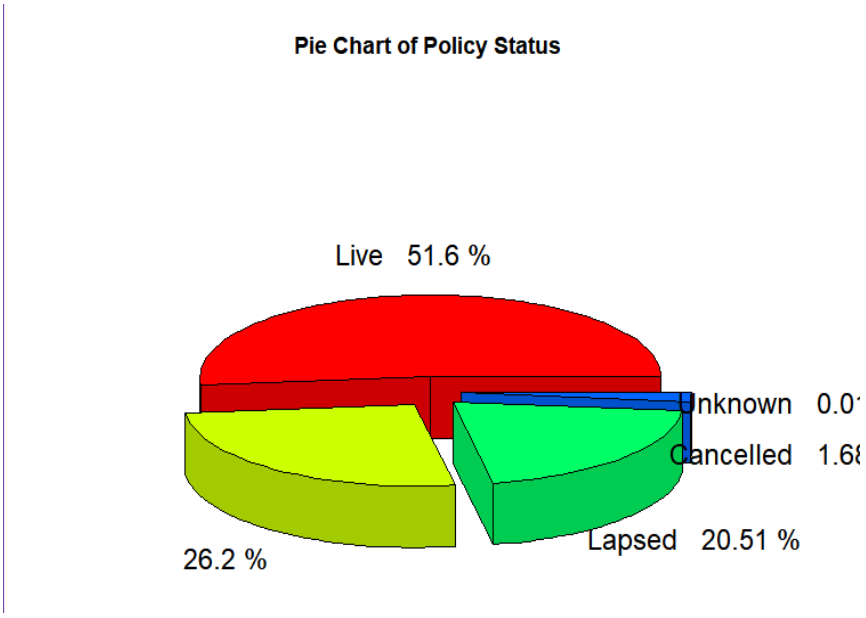
(in Microsoft Office Excel) to arrive at an input that could be loaded into a R DataFrame effortlessly. In other words, this dataset is already relatively clean. I will however attempt at learning more about this features and performing appropriate wrangling steps to arrive at a form that is more suitable for analysis.

1. Using R created a pie chart of policy status.

home_insurance_sagar_satpute.R × status_groups ×

Filter

	POL_STATUS	count	percent
1	Live	132160	51.60
2		67115	26.20
3	Lapsed	52534	20.51
4	Cancelled	4311	1.68
5	Unknown	16	0.01



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

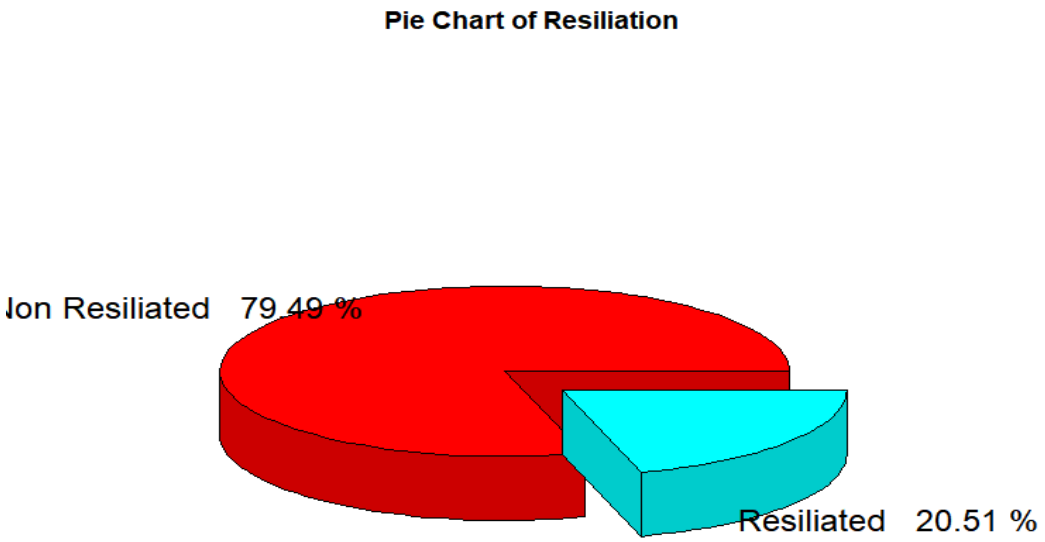
2. Pie chart of resiliation

After working on policy status I decided to get data of resiliation.

home_insurance_sagar_satpute.R status_gr

Filter

	POL_STATUS	count	percent
1	Non Resiliated	203602	79.49
2	Resiliated	52534	20.51



Here we can see that 20.5% of the clients return to the policy after cancelled.

3.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

This are the top 8 of total coverage of the policy which is i added in the dataset

home_insurance_sagar_satpute.R* × policy_coverd ×

Filter

	Police	SUM_INSURED_BUILDINGS	SUM_INSURED_CONTENTS	SPEC_SUM_INSURED	total_coverage
1	P098293	1e+06	1e+05	38080	1138080
2	P032217	1e+06	1e+05	32800	1132800
3	P028759	1e+06	1e+05	20000	1120000
4	P159900	1e+06	1e+05	11480	1111480
5	P103729	1e+06	1e+05	11160	1111160
6	P079048	1e+06	1e+05	10000	1110000
7	P245704	1e+06	1e+05	9100	1109100
8	P025046	1e+06	1e+05	8880	1108880

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

4. Client professional status.

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

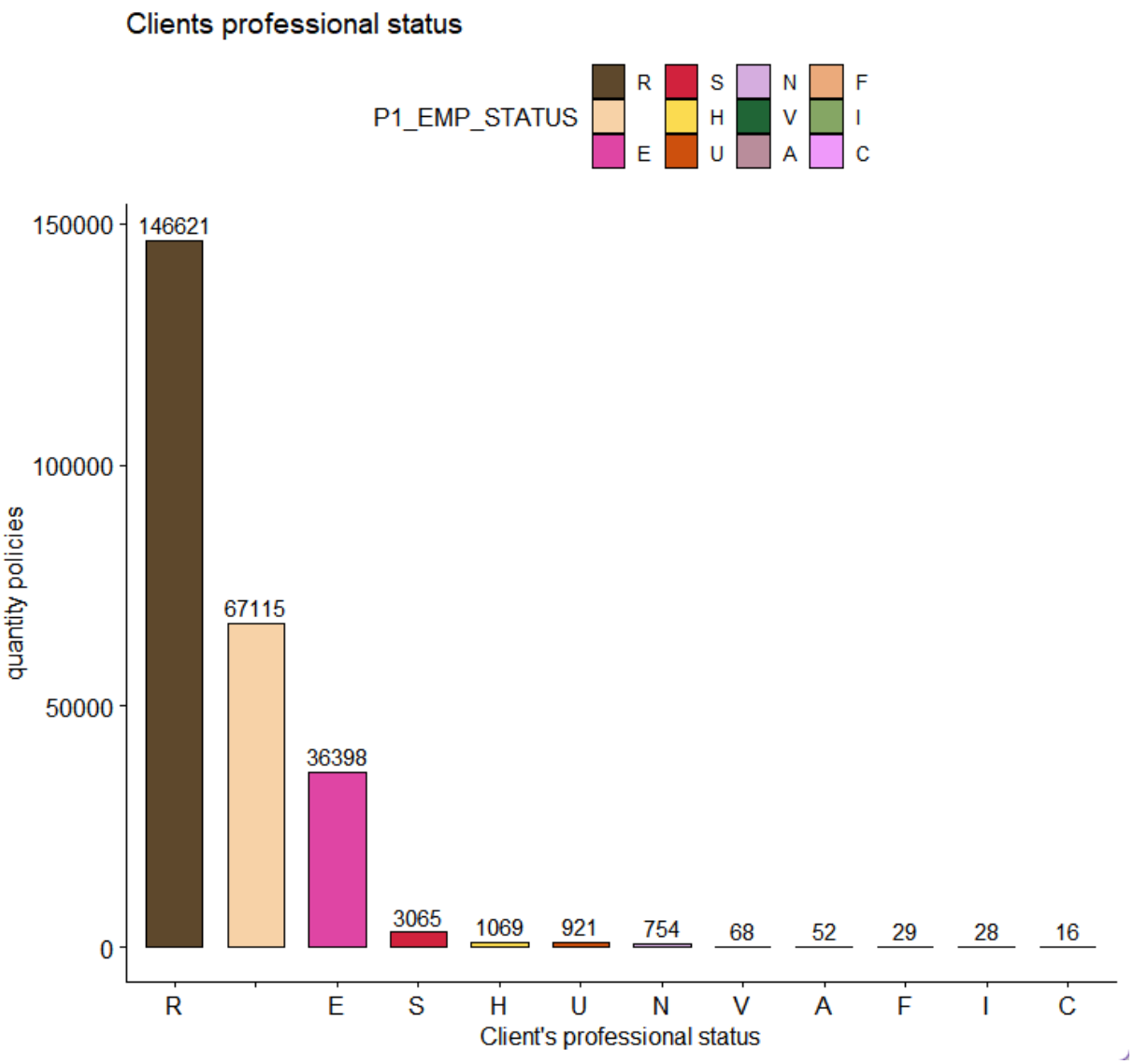
home_insurance_sagar_satpute.R*

Filter

	P1_EMP_STATUS	count
1	R	146621
2		67115
3	E	36398
4	S	3065
5	H	1069
6	U	921
7	N	754
8	V	68
9	A	52
10	F	29
11	I	28
12	C	16

1. R = Retired,
2. E = Employed,
3. N = Not Available,
4. H = House person,
5. S = Student ,
6. U = Unemployed.

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

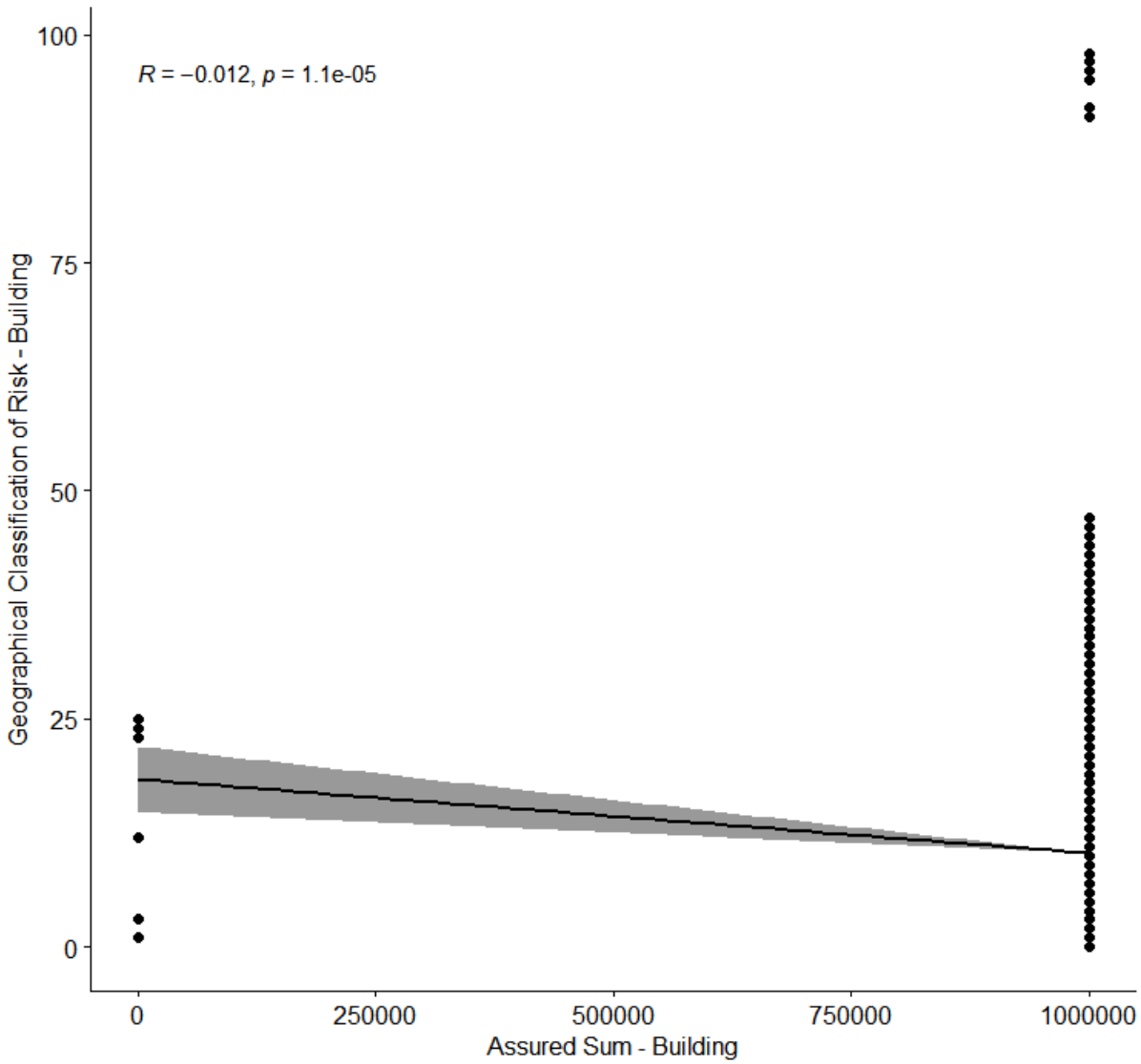


As mentioned earlier, Retired and Employees clients are the most commonly occurring professional status. V, A, F, I, C form the minority. I can imagine that this last statuses are two or the following:

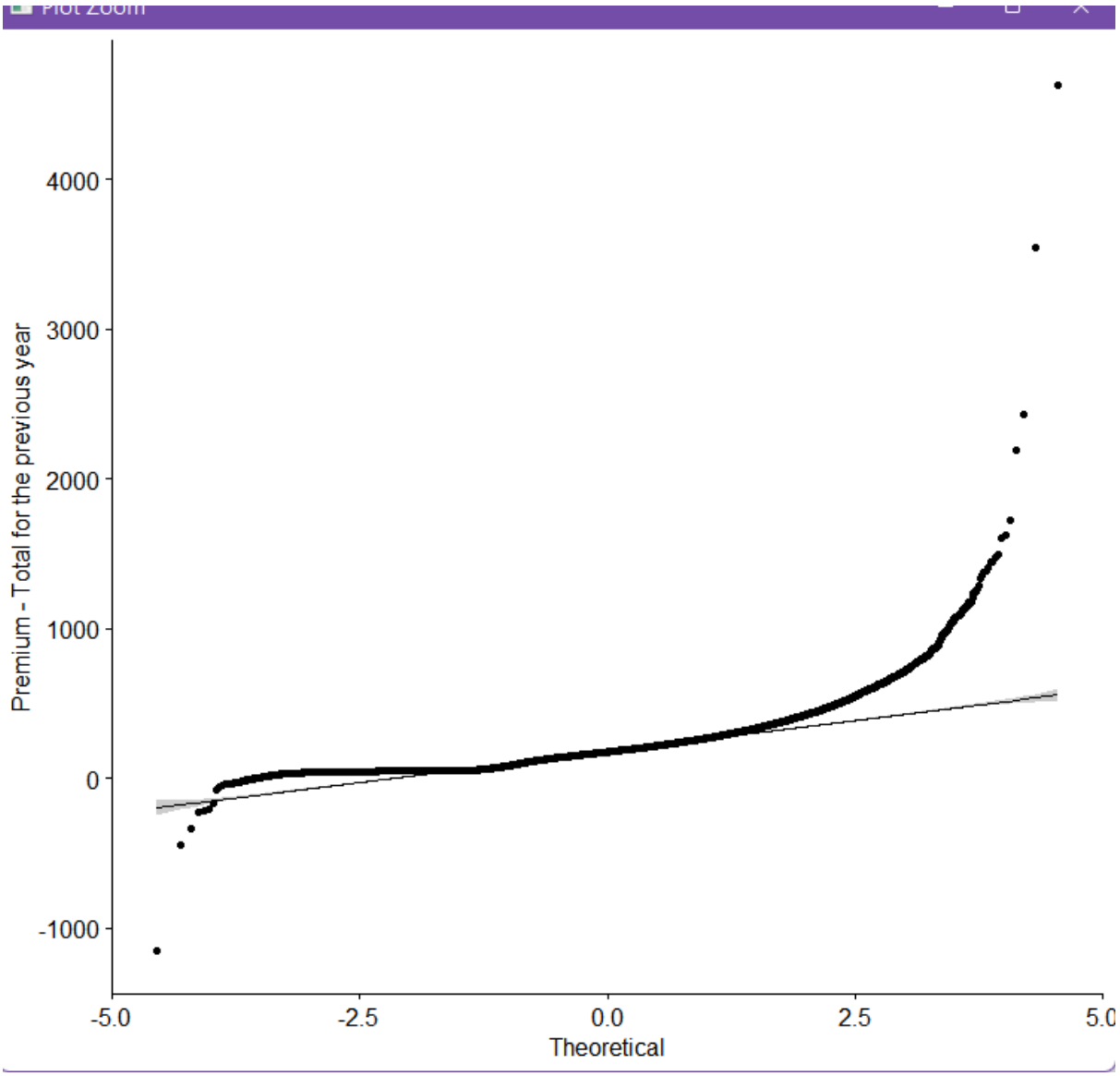
1. ***The richest clients of the sample.***
2. ***The clients living in a very uncomfortable zone/building***

5.Correlation

Geographical Classification of Risk - Building and Assured Sum - Building



As we can see in the 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.



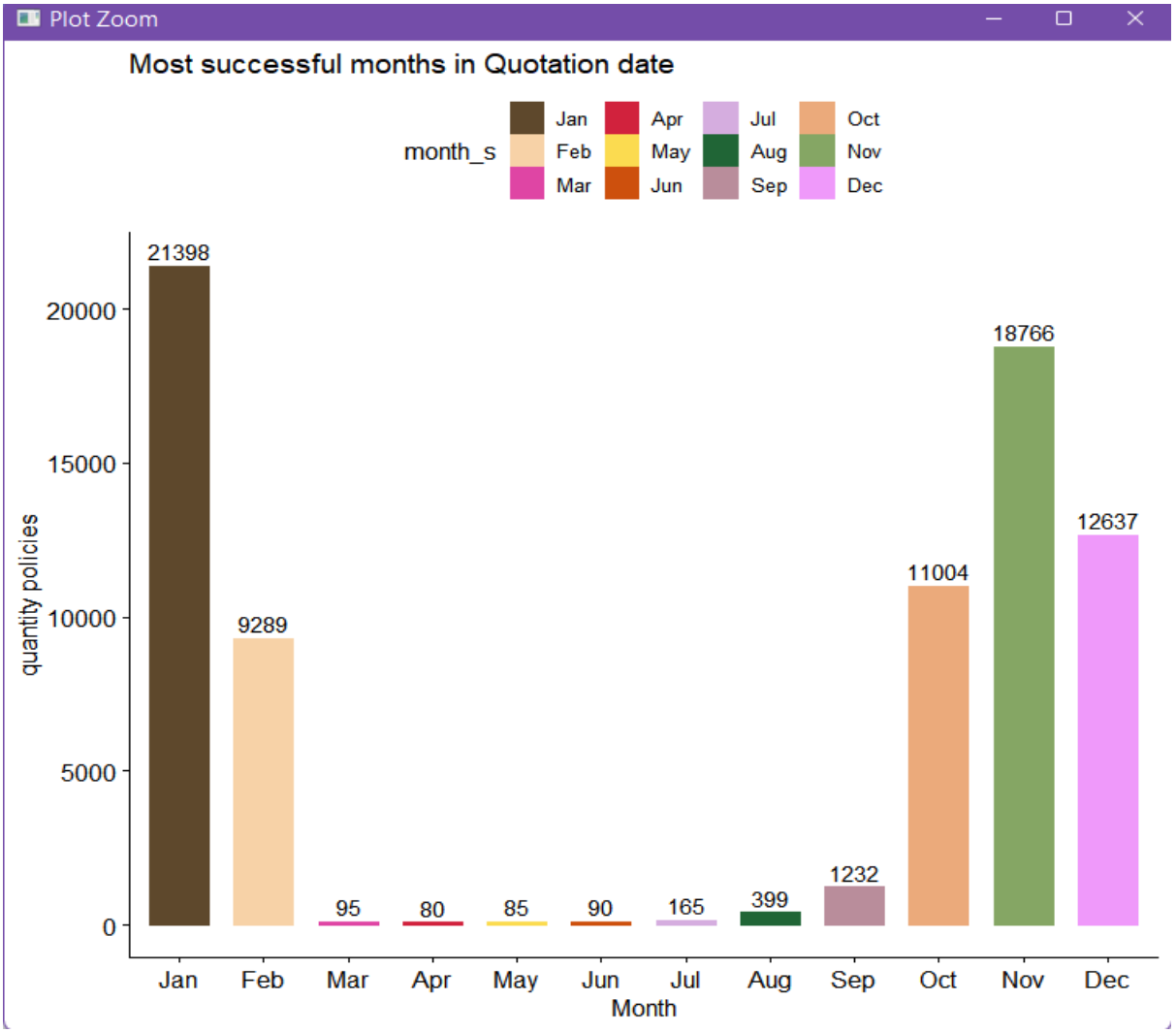
We can see that for those premium clients with Total for the previous year bonus (more than 0/null) are between 0 and 1,000.

6.most popular and most successful months and day for quotation start.

home_insurance_sagar_satpute.R* quot

Filter

	month_n	count	month_s
1	1	21398	Jan
2	2	9289	Feb
3	3	95	Mar
4	4	80	Apr
5	5	85	May
6	6	90	Jun
7	7	165	Jul
8	8	399	Aug
9	9	1232	Sep
10	10	11004	Oct
11	11	18766	Nov
12	12	12637	Dec



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

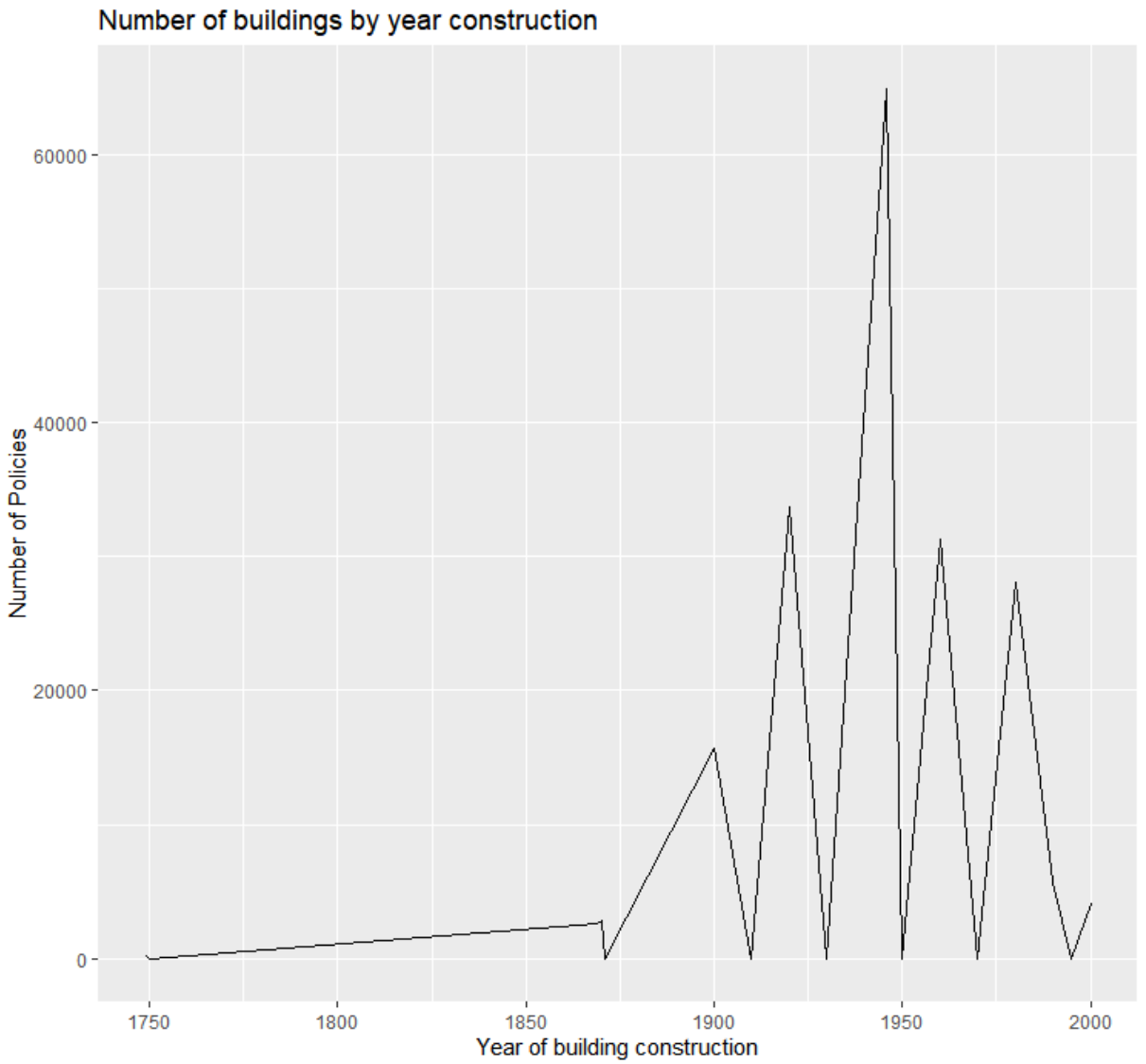
We see that effectively the months of January, February, March and November tend to yield the highest median returns. However December does not have a high total coverage, even when this month has the highest number of policies. On the other hand June and August have a not so high coverage and finally the resting months are the least successful months 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 stuff

7.Number of buildings by year

year_built × home_insurance_

Filter

	YEARBUILT	count
1	1749	222
2	1750	6
3	1869	2564
4	1870	2835
5	1871	11
6	1900	15687
7	1910	1
8	1920	33679
9	1930	4
10	1946	64831
11	1950	4
12	1960	31287
13	1970	2
14	1980	28057
15	1990	5692
16	1995	3
17	2000	4136



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.

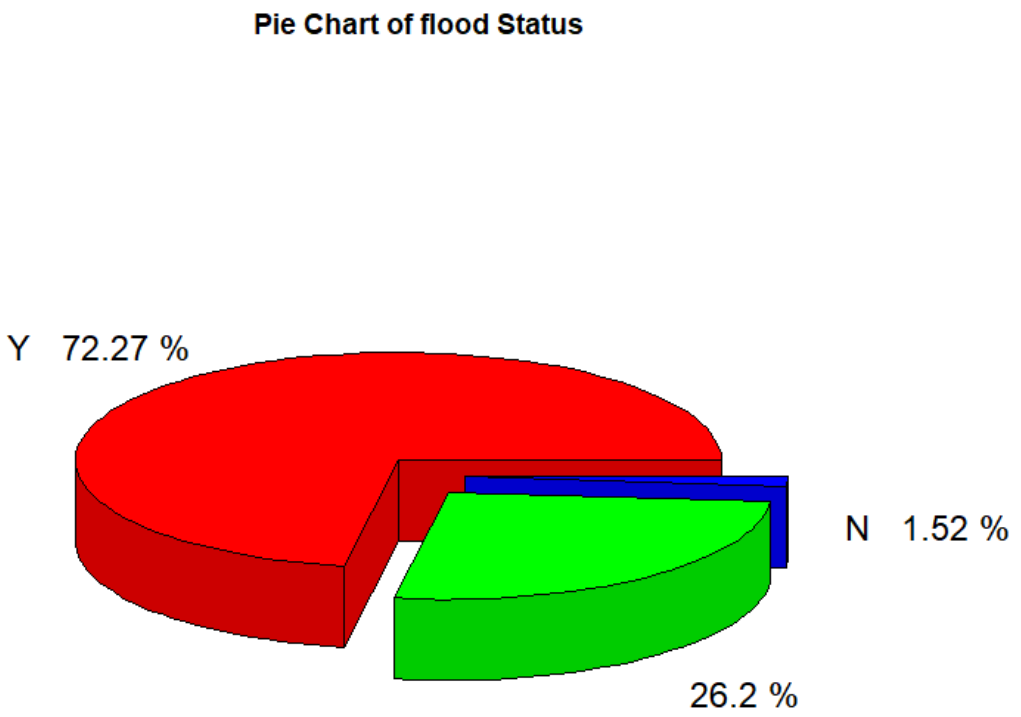
It appears that January is the most popular month when it comes to policies quotations. This is maybe for the salary bonus that employees get in christmas and the clients makes their plan come true making the quote.

8.HOUSE SUSCEPTIBLE TO FLOODS

ling_and_subsetting_sagar_... home_insurance.Rmd

Filter

	FLOODING	count	percent
1	Y	185121	72.27
2		67115	26.20
3	N	3900	1.52



72.27% POLICY IS SUSCEPTIBLE TO FLOODS WHERE 26.2% POLICY'S DATA IS NOT AVAILABLE

9.PAYING GUESTS NOT = TO 0

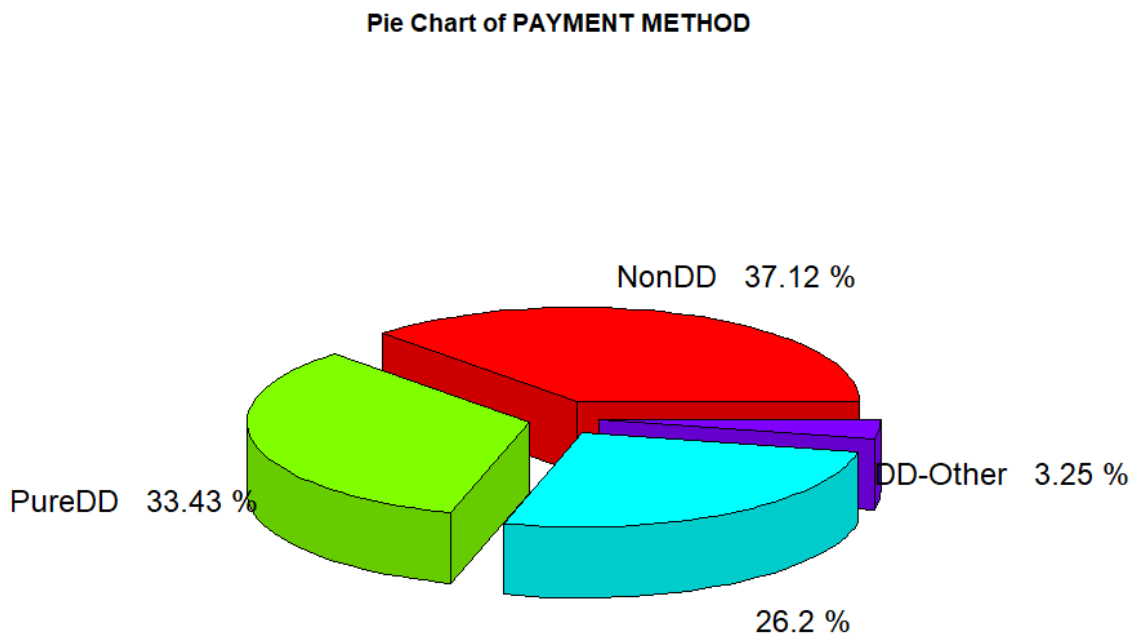
After subseting a data i find out 207 policies paying guests are not equals to 0

10.PAYMENT METHOD

ling_and_subsetting_sagar_... home_insurance.Rmd

Filter

	PAYMENT_METHOD	count	percent
1	NonDD	95065	37.12
2	PureDD	85619	33.43
3		67115	26.20
4	DD-Other	8337	3.25



37.12% of policies payment method is not DD.but 33.43% policies payment method is 33.34% is by DD.and 26.2% policies payment method is unknown

11.OWENERSHIP TYPE

year_built flooding FL_TOP

Filter

	OWNERSHIP_TYPE	Police
1	3	P000002
2	3	P000014
3	3	P000018
4	3	P000024
5	3	P000025
6	3	P000031
7	3	P000035
8	3	P000036
9	3	P000040
10	3	P000041
11	3	P000048
12	3	P000054
13	3	P000055

Showing 1 to 14 of 27,388 entries, 2 total columns

There are total 27388 policies where ownership type is "3"

Conclusion :

Inference means a conclusion reached on the basis of evidence and reasoning.
Inference is using observation and background to reach a logical conclusion.lets see an example of a inference using my mentorship
so my project name is identify primium pricing attributes of insurance industry
so month is a premium attribute beacuse January is the most popular month when it comes to policies quotations.
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 on the aforementioned metrics.

second is employee status Retired and Employees clients are the most commonly occurring professional status

3rd is build year of a building 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

Suggestion

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.

Thank you,

Written by

Sagar Satpute

Data science batch 3

Rise WPU