August 2022

Internship Final Report

Identify Premium Pricing Attribute for Home Insurance.

Presented by

Yash Agrawal

2021183

Table of Contents

Phases of Internship

Phase 1

Data Understanding / Wrangling

In this phase I learned about the data and the different type of variables in the data . The Columns who ever either empty or are not with enough information were dropped off completely

Phase 2

Cleaning the Data & Subsettiing It.

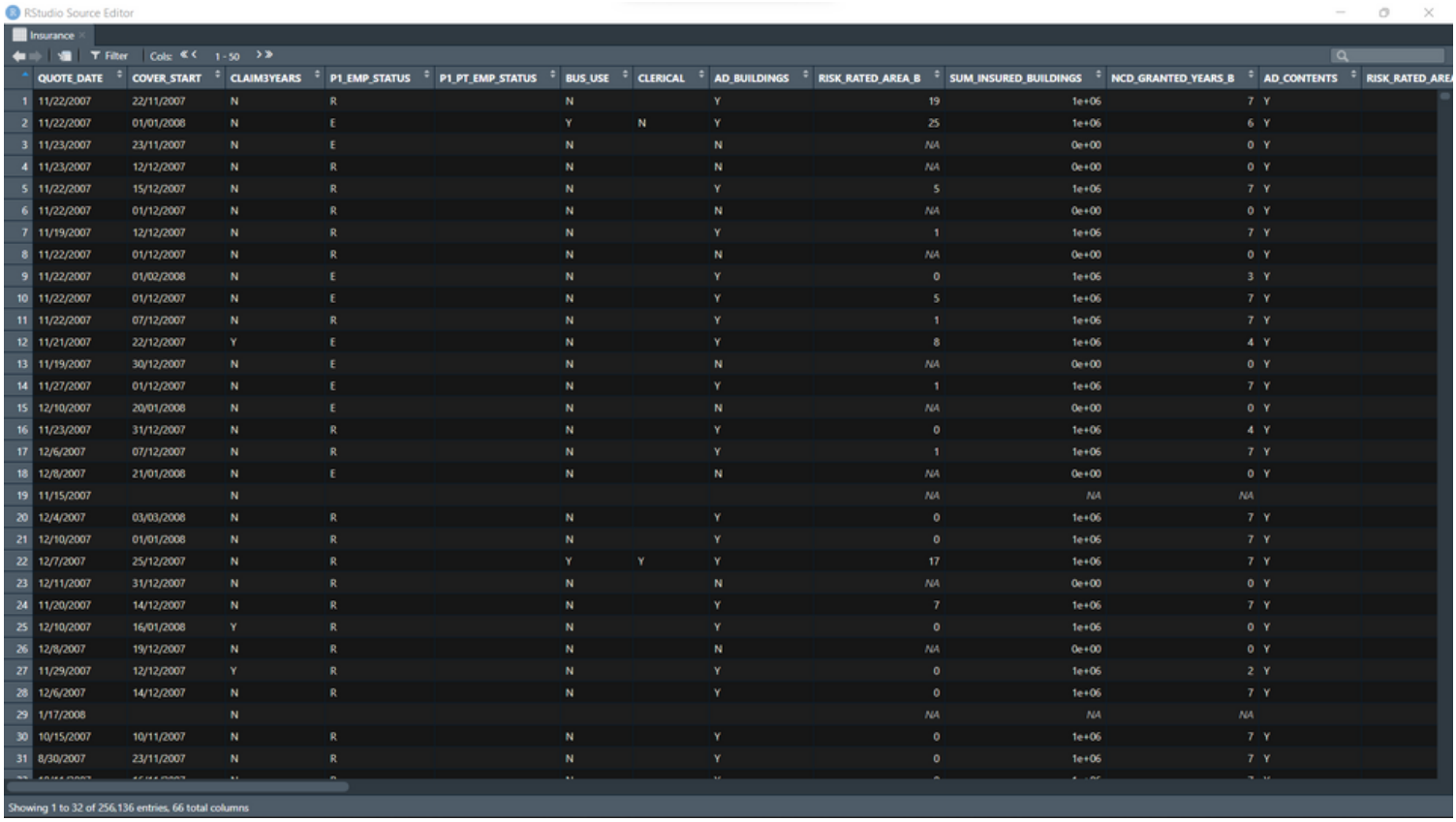
After the Phase 1 I got some idea on which columns to work on . The NA Values were either replaced with Median or dropped off . After cleaning the NA Values we subset the data into new Data frame for next phase.

Phase 3

Data Visualization

In this Phase I created some eye-catching Graphs and plots using the new data frame from the previous phase to Identify two things :-
1) Characteristics of Customers who are likely to make a default.
2) Features that drive up the Premium Prices of a Property.

Data Processing

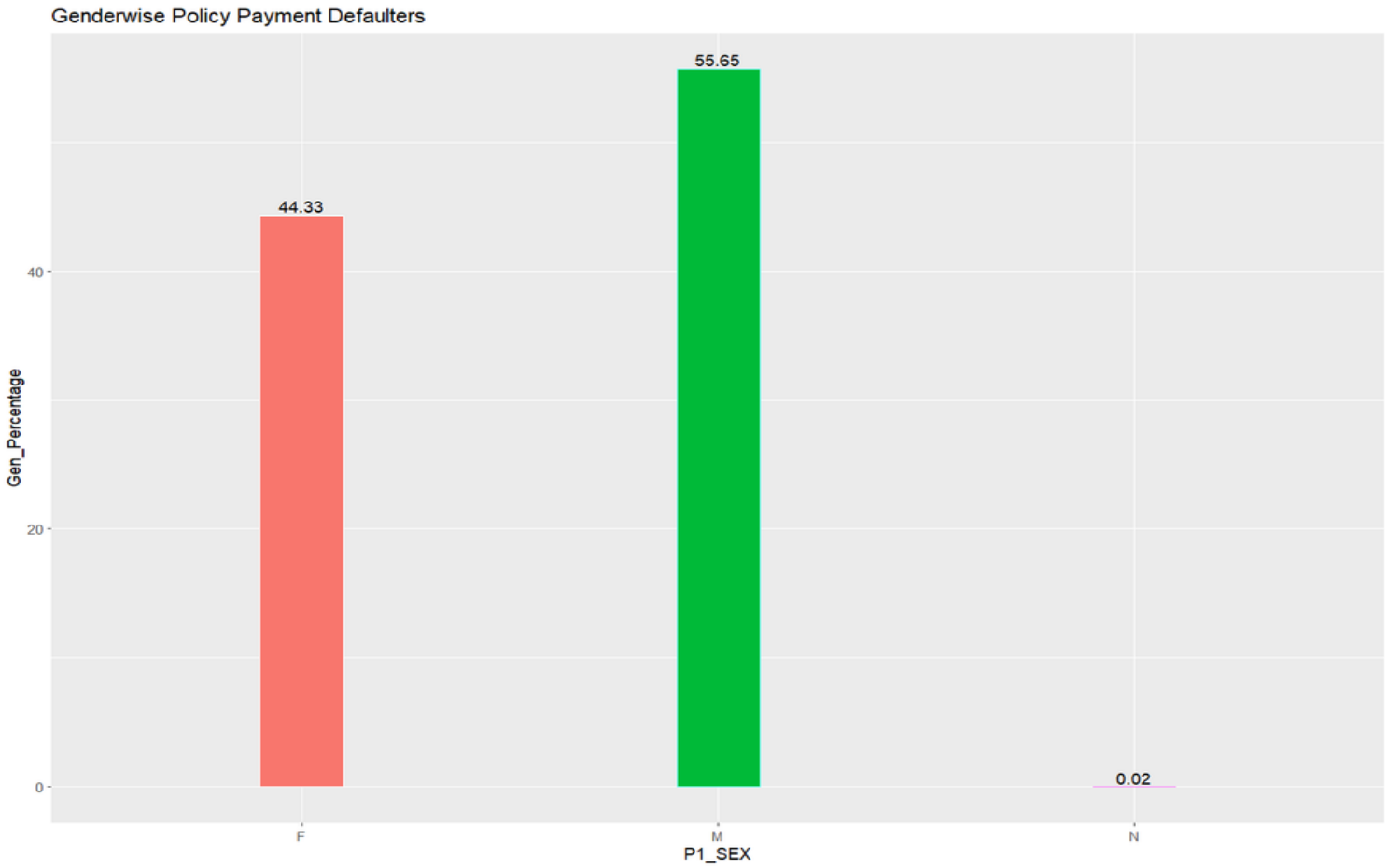


	QUOTE_DATE	COVER_START	CLAIM3YEARS	P1_EMP_STATUS	P1_PT_EMP_STATUS	BUS_USE	CLERICAL	AD_BUILDINGS	RISK_RATED_AREA_B	SUM_INSURED_BUILDINGS	NCD_GRANTED_YEARS_B	AD_CONTENTS	RISK_RATED_AREA
1	11/22/2007	22/11/2007	N	R		N		Y	19	1e+06	7	Y	
2	11/22/2007	01/01/2008	N	E		Y	N	Y	25	1e+06	6	Y	
3	11/23/2007	23/11/2007	N	E		N		N	NA	0e+00	0	Y	
4	11/23/2007	12/12/2007	N	R		N		N	NA	0e+00	0	Y	
5	11/22/2007	15/12/2007	N	R		N		Y	5	1e+06	7	Y	
6	11/22/2007	01/12/2007	N	R		N		N	NA	0e+00	0	Y	
7	11/19/2007	12/12/2007	N	R		N		Y	1	1e+06	7	Y	
8	11/22/2007	01/12/2007	N	R		N		N	NA	0e+00	0	Y	
9	11/22/2007	01/02/2008	N	E		N		Y	0	1e+06	3	Y	
10	11/22/2007	01/12/2007	N	E		N		Y	5	1e+06	7	Y	
11	11/22/2007	07/12/2007	N	R		N		Y	1	1e+06	7	Y	
12	11/21/2007	22/12/2007	Y	E		N		Y	8	1e+06	4	Y	
13	11/19/2007	30/12/2007	N	E		N		N	NA	0e+00	0	Y	
14	11/27/2007	01/12/2007	N	E		N		Y	1	1e+06	7	Y	
15	12/10/2007	20/01/2008	N	E		N		N	NA	0e+00	0	Y	
16	11/23/2007	31/12/2007	N	R		N		Y	0	1e+06	4	Y	
17	12/6/2007	07/12/2007	N	R		N		Y	1	1e+06	7	Y	
18	12/8/2007	21/01/2008	N	E		N		N	NA	0e+00	0	Y	
19	11/15/2007		N						NA	NA	NA		
20	12/4/2007	03/03/2008	N	R		N		Y	0	1e+06	7	Y	
21	12/10/2007	01/01/2008	N	R		N		Y	0	1e+06	7	Y	
22	12/7/2007	25/12/2007	N	R		Y	Y	Y	17	1e+06	7	Y	
23	12/11/2007	31/12/2007	N	R		N		N	NA	0e+00	0	Y	
24	11/20/2007	14/12/2007	N	R		N		Y	7	1e+06	7	Y	
25	12/10/2007	16/01/2008	Y	R		N		Y	0	1e+06	0	Y	
26	12/8/2007	19/12/2007	N	R		N		N	NA	0e+00	0	Y	
27	11/29/2007	12/12/2007	Y	R		N		Y	0	1e+06	2	Y	
28	12/6/2007	14/12/2007	N	R		N		Y	0	1e+06	7	Y	
29	1/17/2008		N						NA	NA	NA		
30	10/15/2007	10/11/2007	N	R		N		Y	0	1e+06	7	Y	
31	8/30/2007	23/11/2007	N	R		N		Y	0	1e+06	7	Y	

Initially the data contained 32 rows and total of 256136 columns including the empty rows and columns and NA values. After we performed **EDA** on the data , i.e. I dropped the ('i', 'CLERICAL', etc.) columns which were empty and no use , Replaced the Na values with median in some columns ('Risk_Rated_Area_A & C') and filtered out the Na Values . After doing all these things , the data did not contained any empty columns or rows or Na Values . The data was now more clean and precise and contained the data which could have been used in the next phase . The data was divided into smaller subsets in the next phase because the data needed to be divided on the basis of the task.

Characteristics of Defaulters

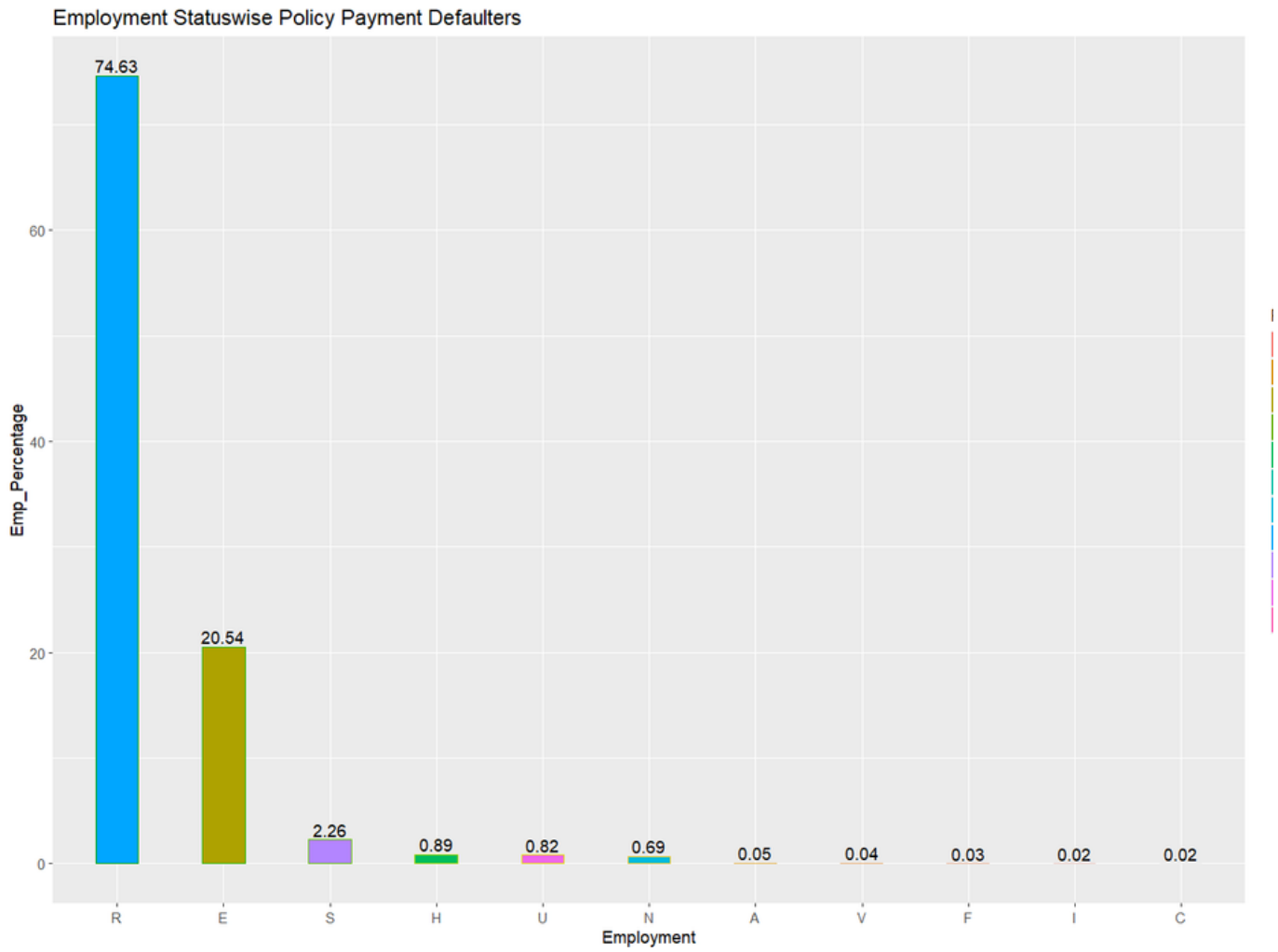
1) Checking the Defaulters Gender wise



In the above chart i have plotted the gender column who's payment Frequency is 0 and As we can see almost 11 % of more Male are more likely to make an Default than Female .

Characteristics of Defaulters

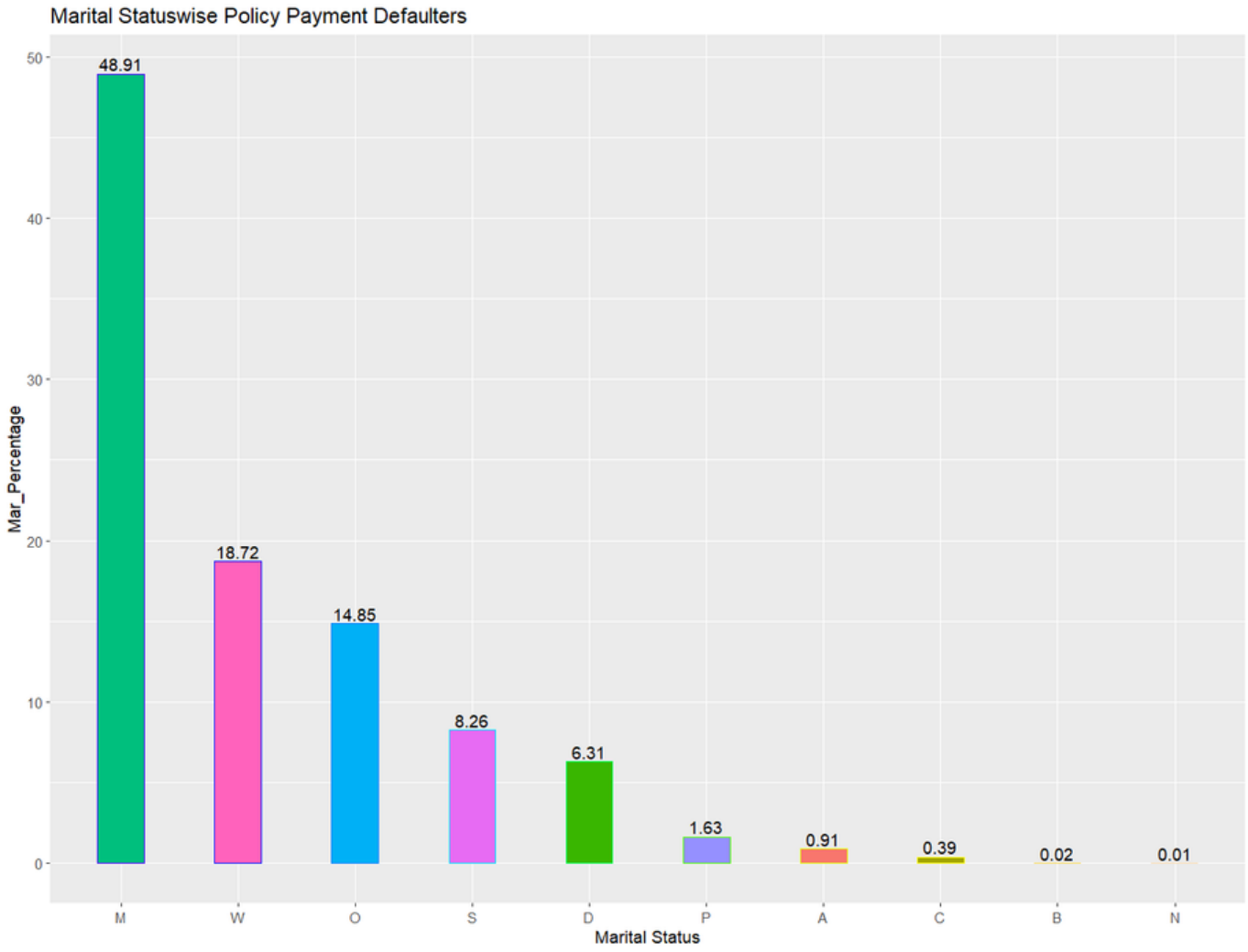
2) Checking the Defaulters Employment wise



In the above chart i have plotted the Employment column who's payment Frequency is 0 and As we can see the Retired customers are more likely to make an Default .

Characteristics of Defaulters

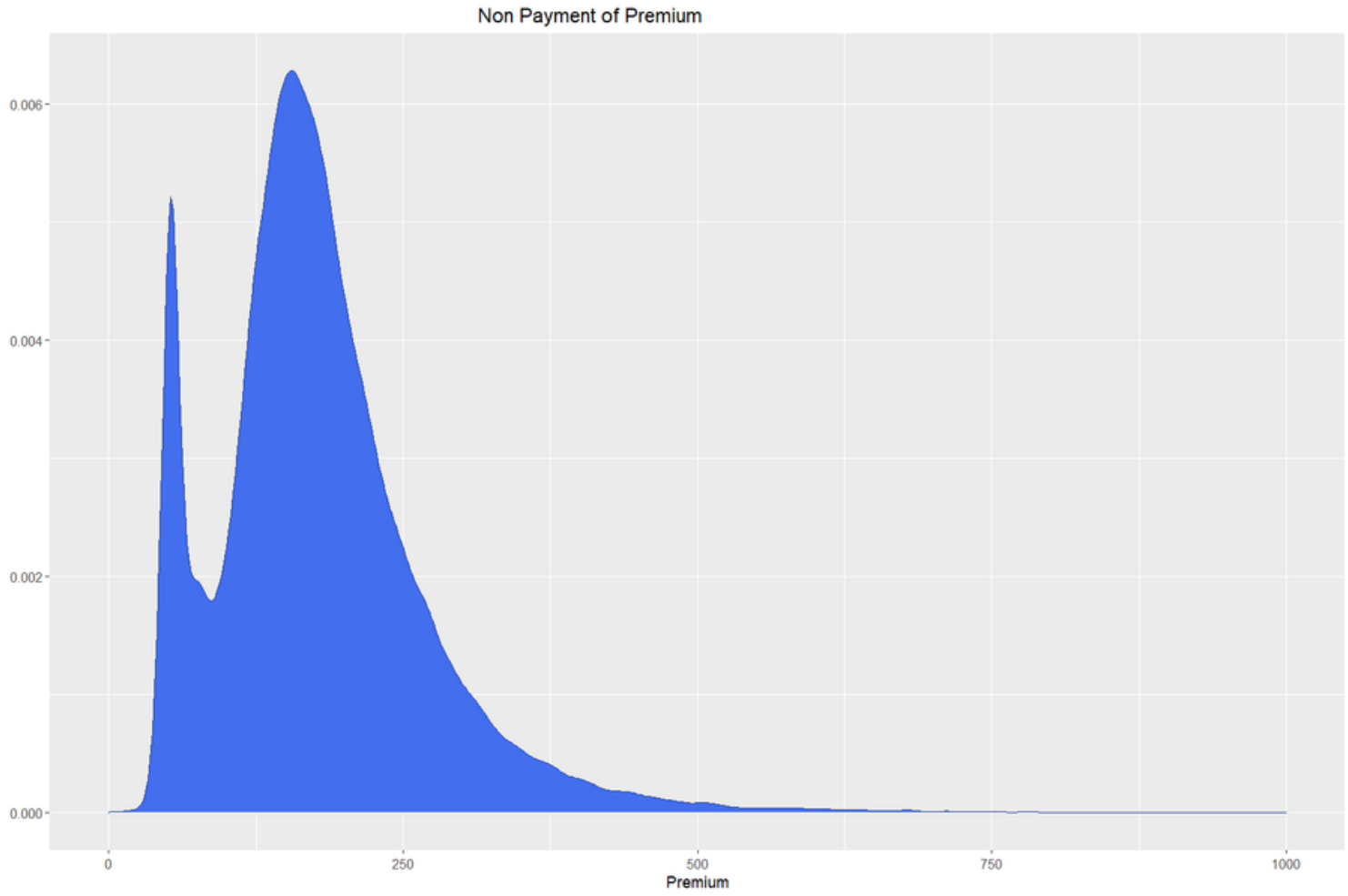
3) Checking the Defaulters Marital status wise



In the above chart i have plotted the Marital_Status column who's payment Frequency is 0 and As we can see the Married customers are more likely to make an Default.

Characteristics of Defaulters

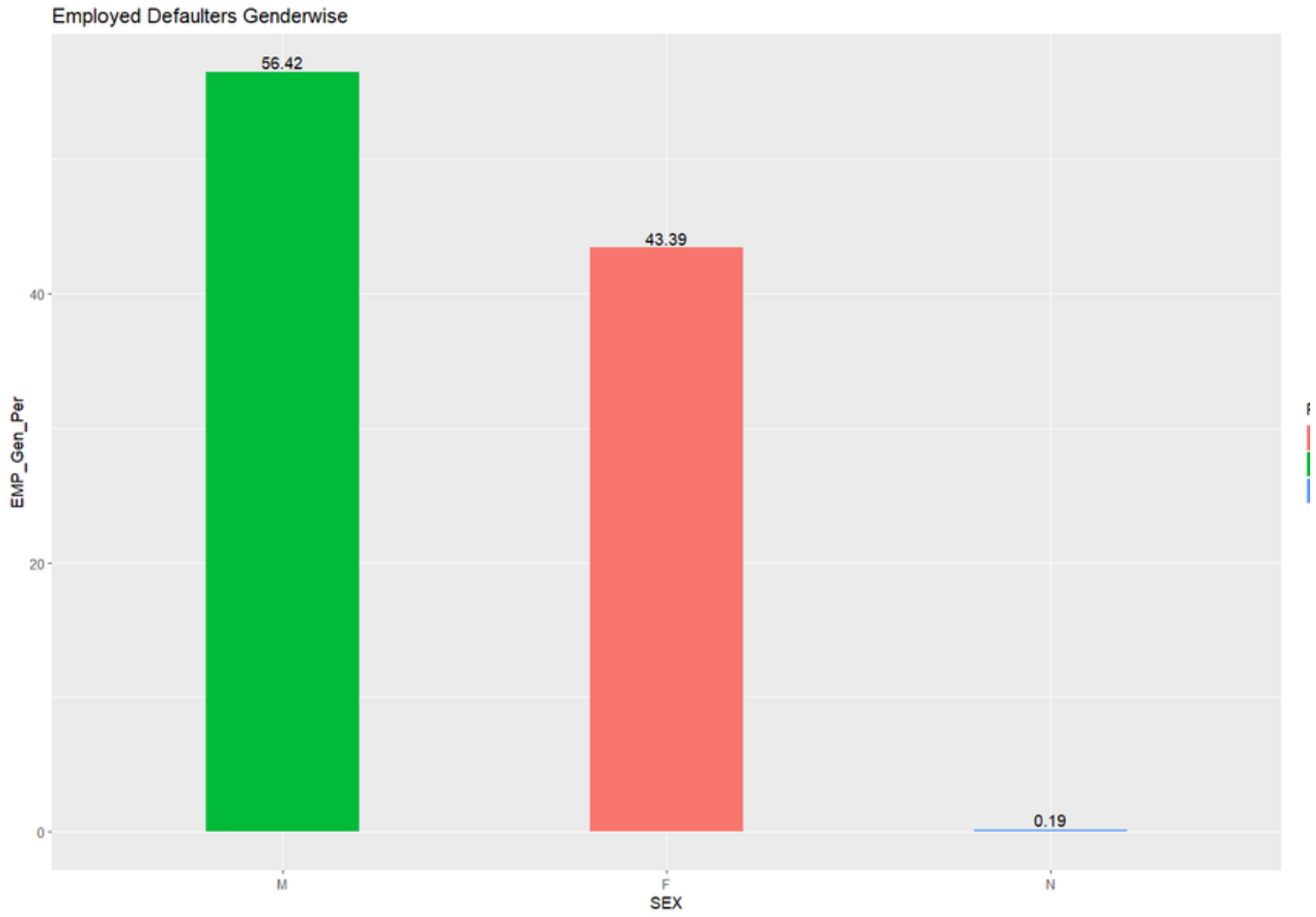
4) Checking the Frequency of Non-Payment of Premium



In the above chart i have plotted the Premium column who's payment Frequency is 0 and As we can see that the customers with premium amount above 250 dollars are more likely to make a default.

Characteristics of Defaulters

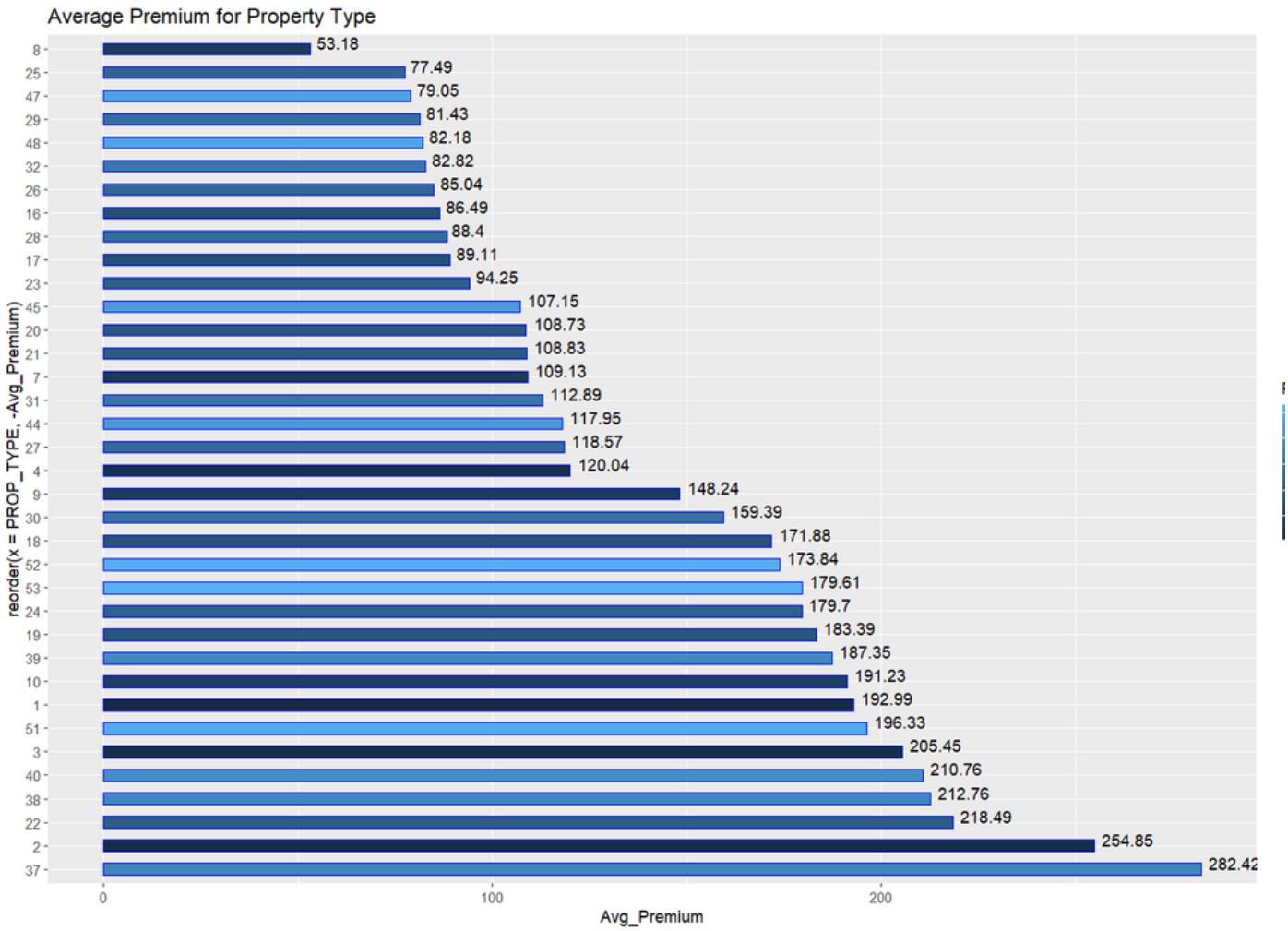
5) Checking the Employed Defaulters Gender wise.



In the above chart i have plotted the Gender Column who's payment Frequency is 0 and Employment Status as Employed & As we can see that the Employed Customers who are Male are more likely to make an Default.

Features of Properties

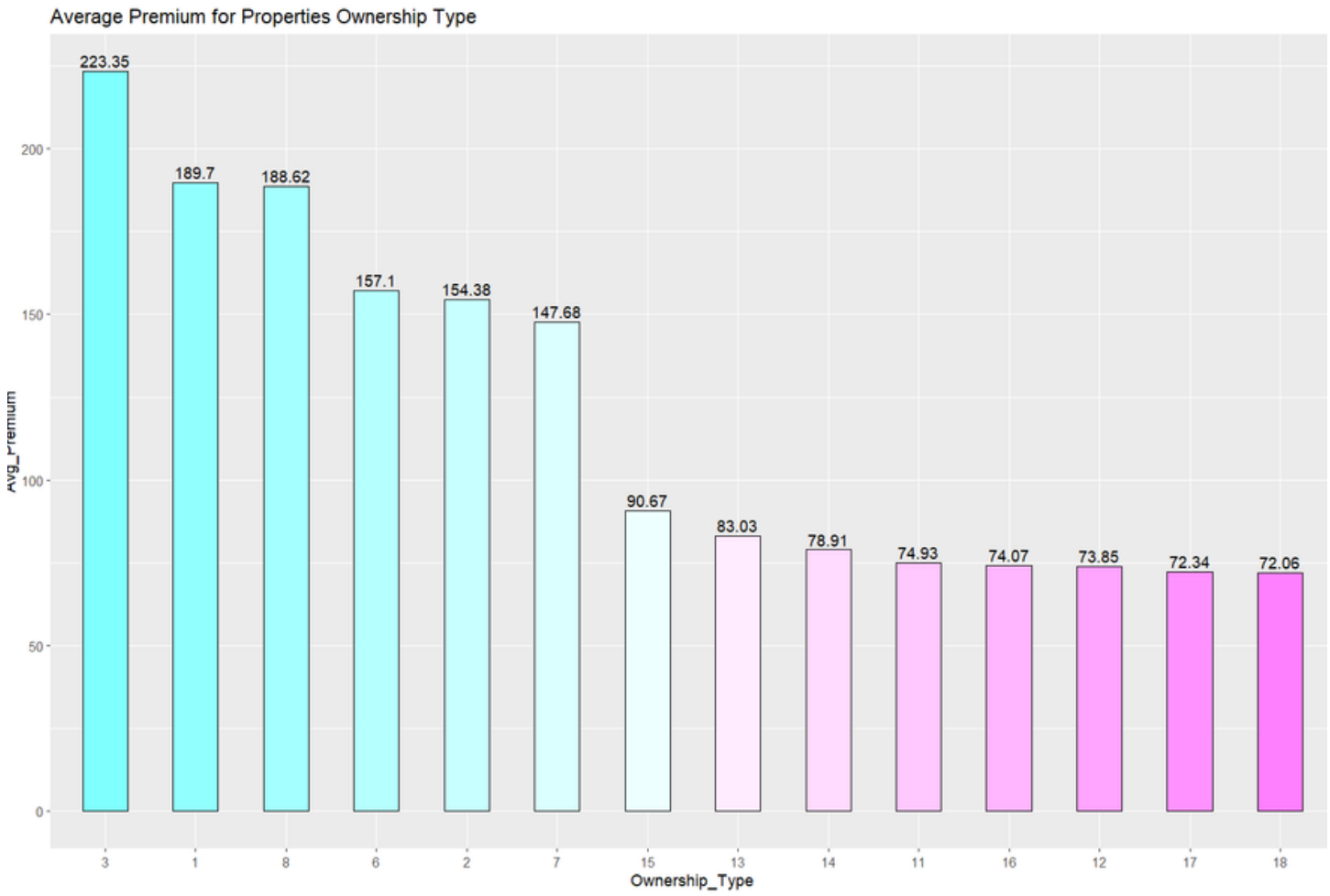
1) Checking the Premium for Property Type



As we can see that the Premium for the Property Type "37 " is the Highest , which is almost 282 dollars.

Features of Properties

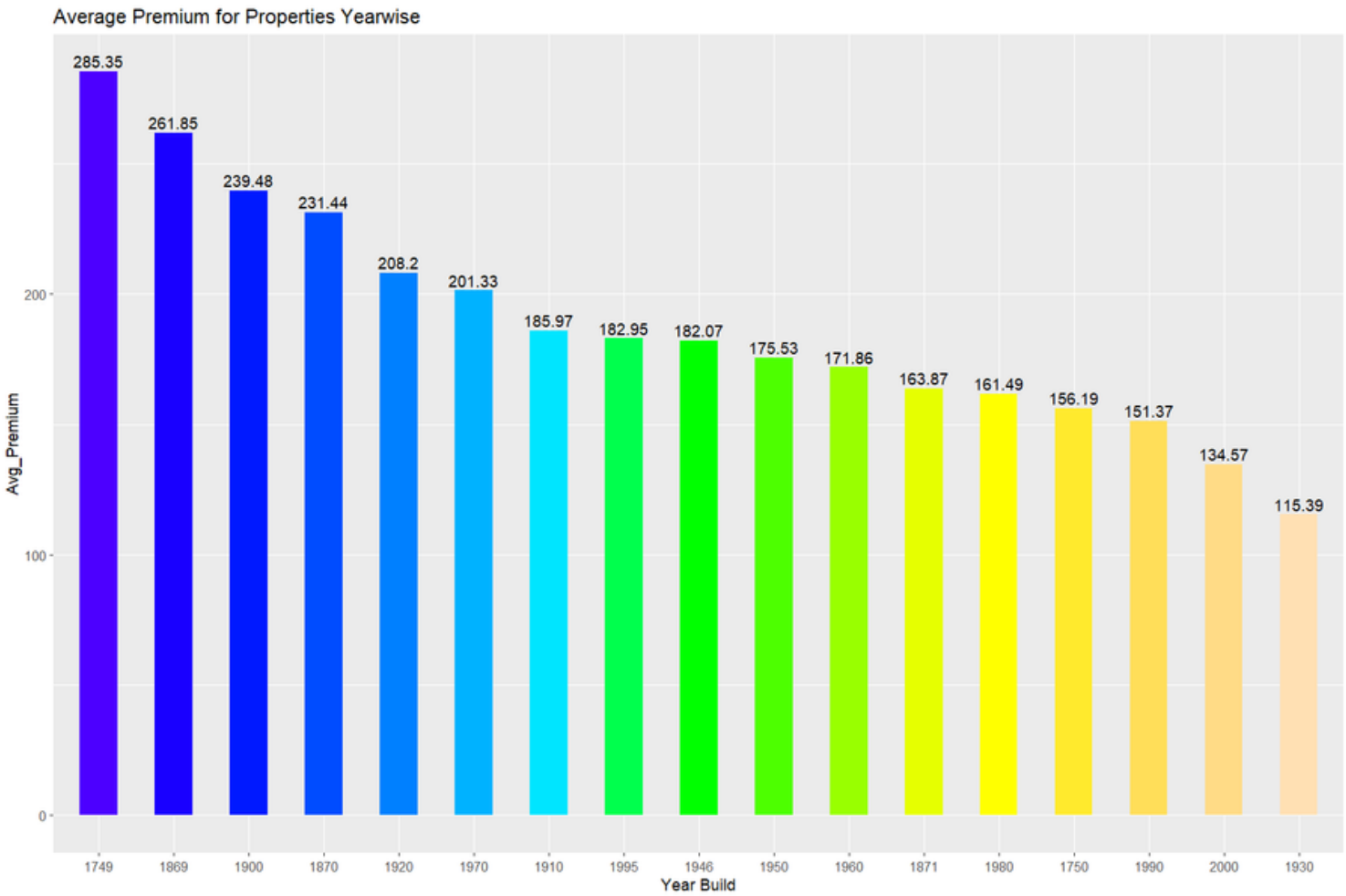
2) Checking the Premium for Ownership Type



As we can see that the Premium for the Ownership Type "3 " is the Highest , which is almost 223 dollars.

Features of Properties

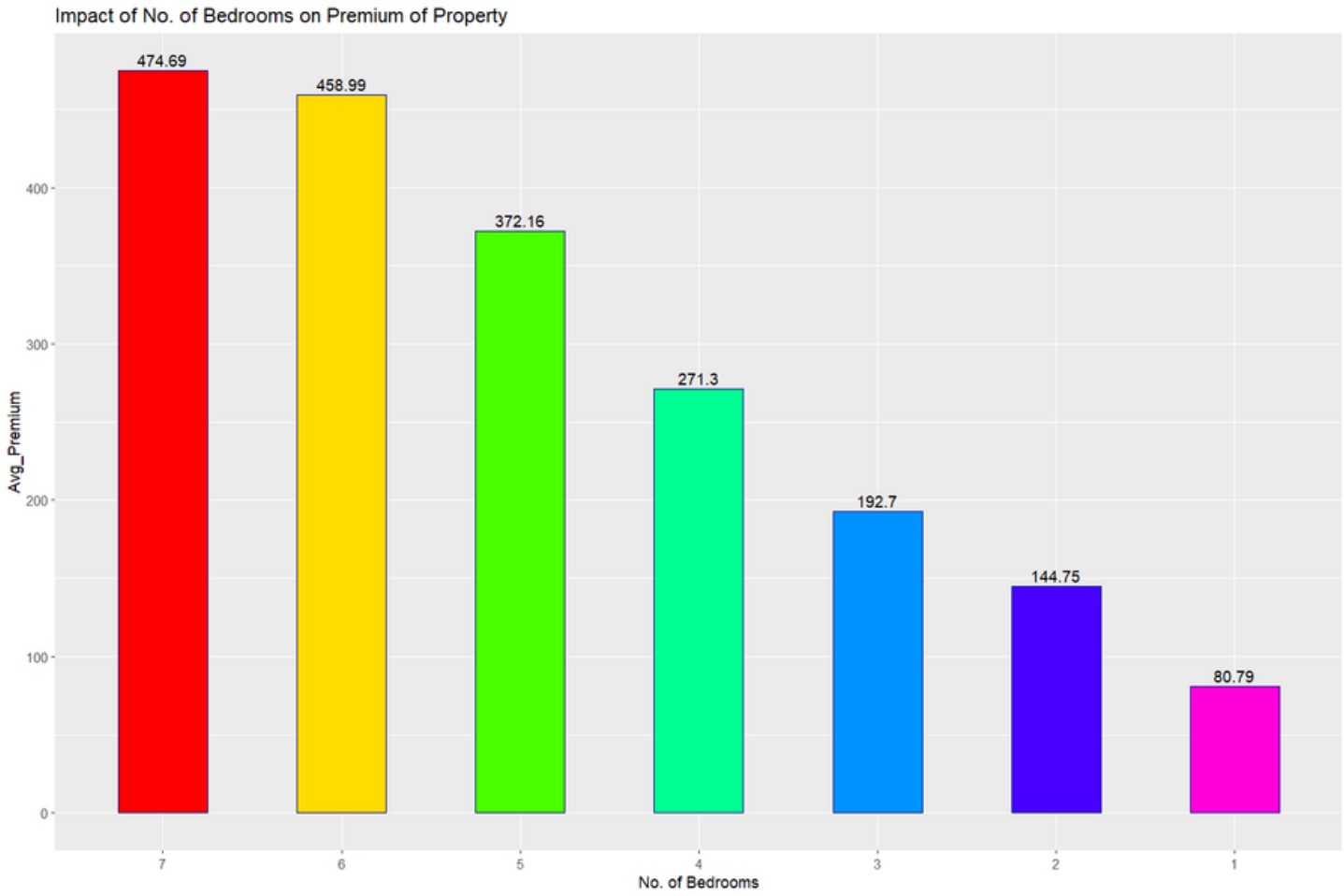
3) Checking the Premium for Year Build.



As we can see that the Premium for the Properties which were build during the 'Year 1749' is the Highest , which is almost 285 dollars.

Features of Properties

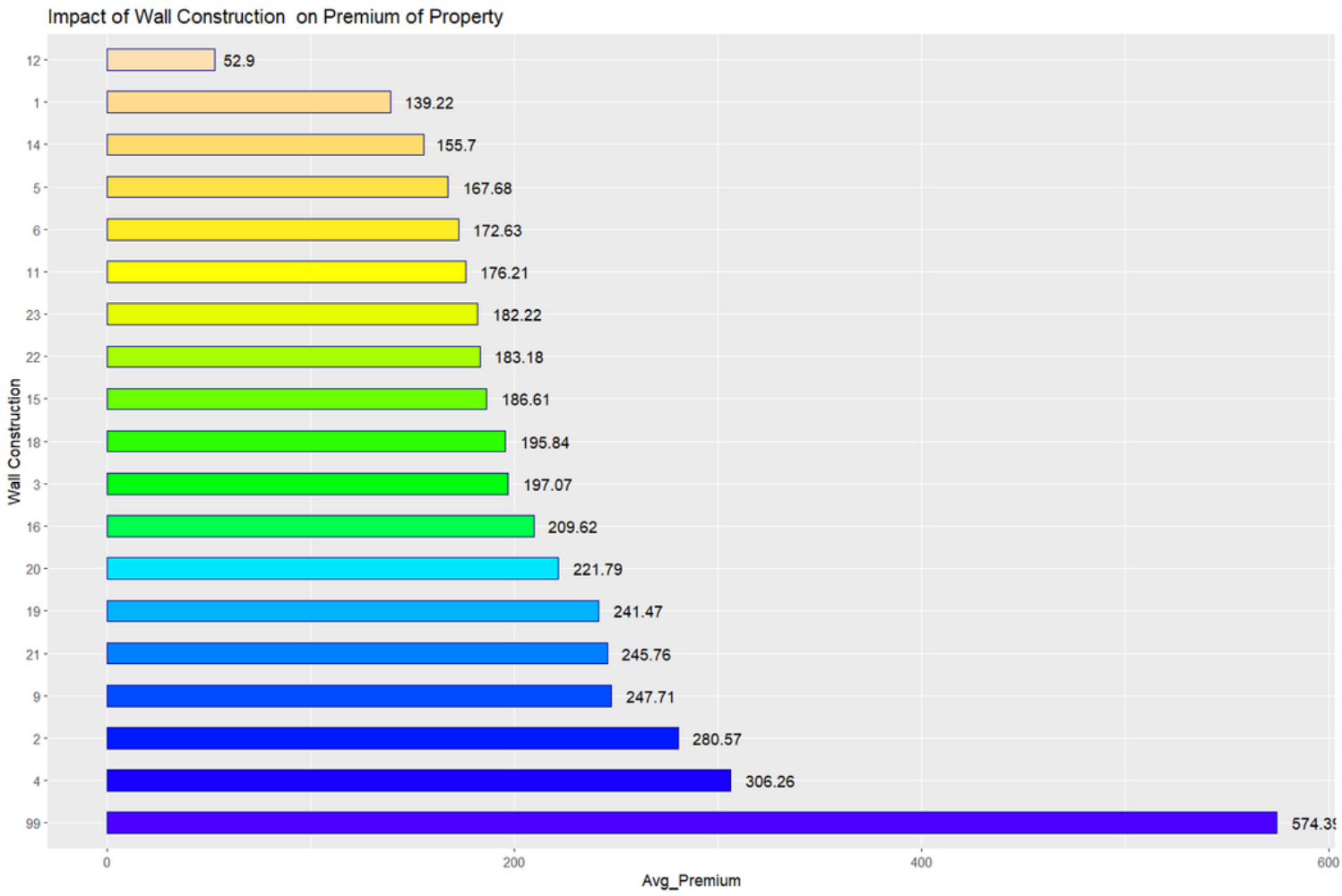
4) Checking the Premium for No. of Bedrooms



As we can see that the Premium for the Properties with 5 bedrooms is the Highest , which is almost 478 dollars.

Features of Properties

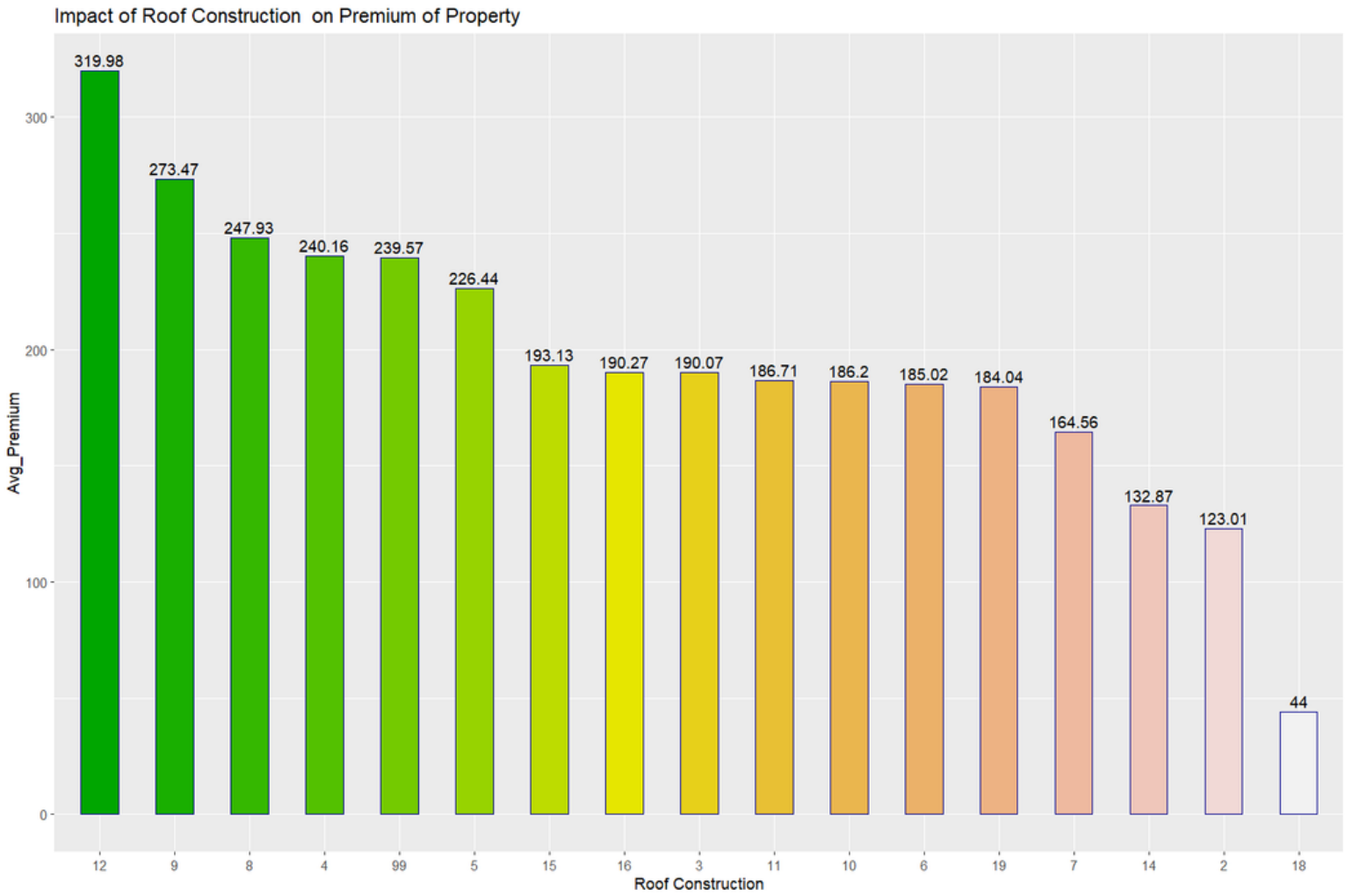
5) Checking the Premium for Wall Construction



As we can see that the Premium for the Properties with Wall Construction "99" is the Highest , which is almost 574 dollars.

Features of Properties

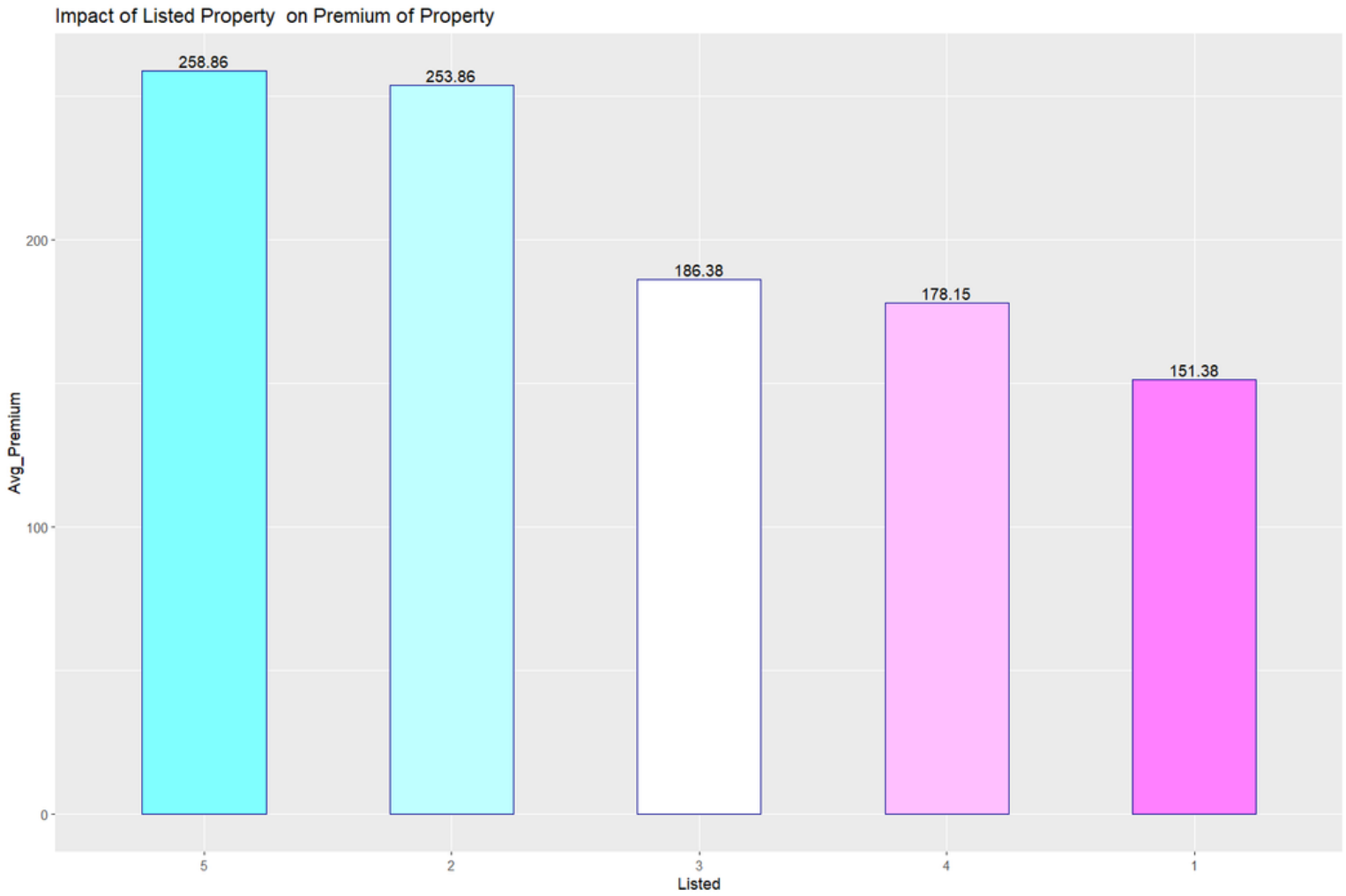
6) Checking the Premium for Roof Construction



As we can see that the Premium for the Properties with Roof Construction "12" is the Highest , which is almost 320 dollars.

Features of Properties

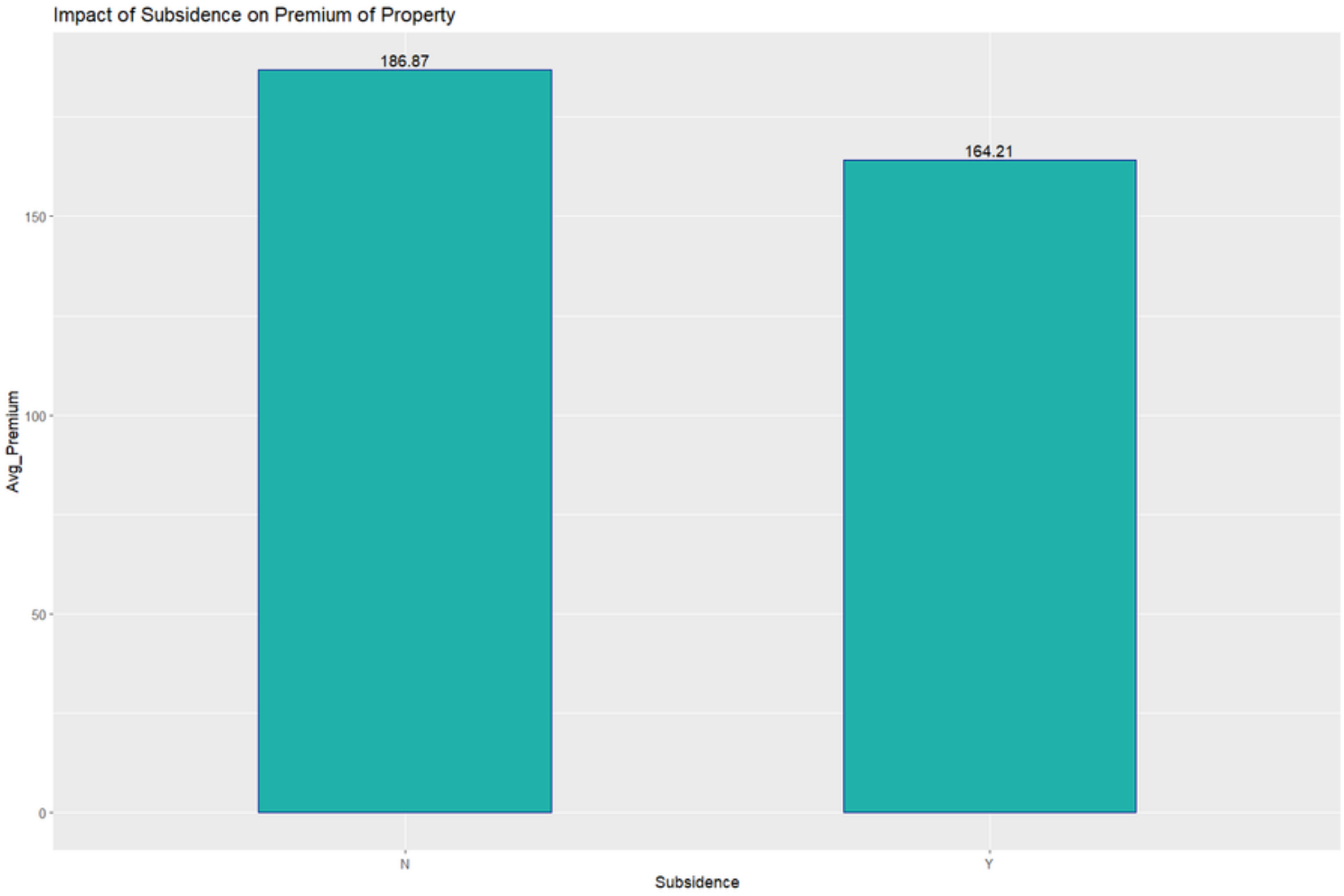
7) Checking the Premium for Listed Property



As we can see that the Premium for the Properties with Listing as "5" is the Highest , which is almost 258 dollars.

Features of Properties

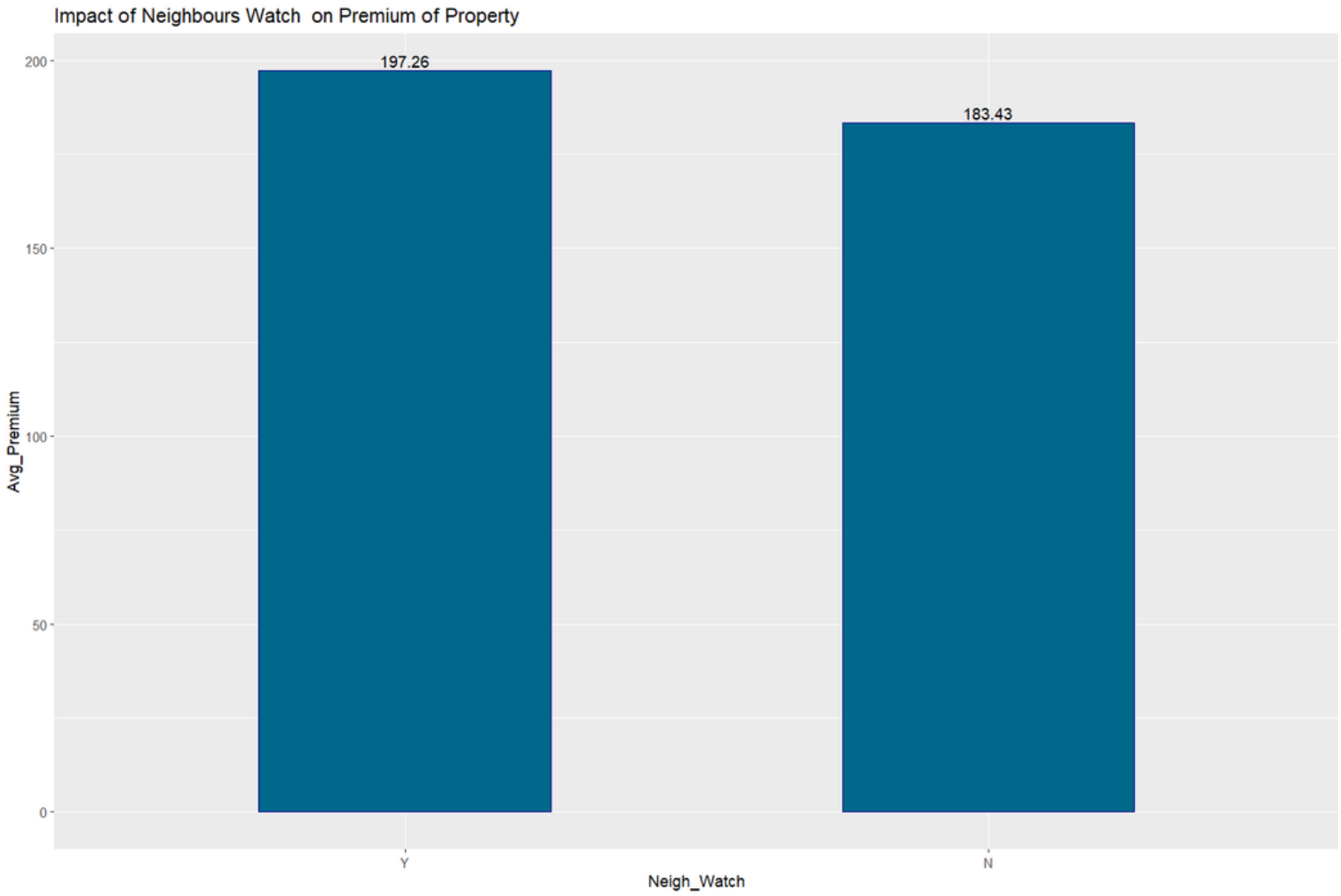
8) Checking the Premium for Property with Subsidence



As we can see that the Premium for the Properties with no Subsidence is the Highest , which is almost 187 dollars.

Features of Properties

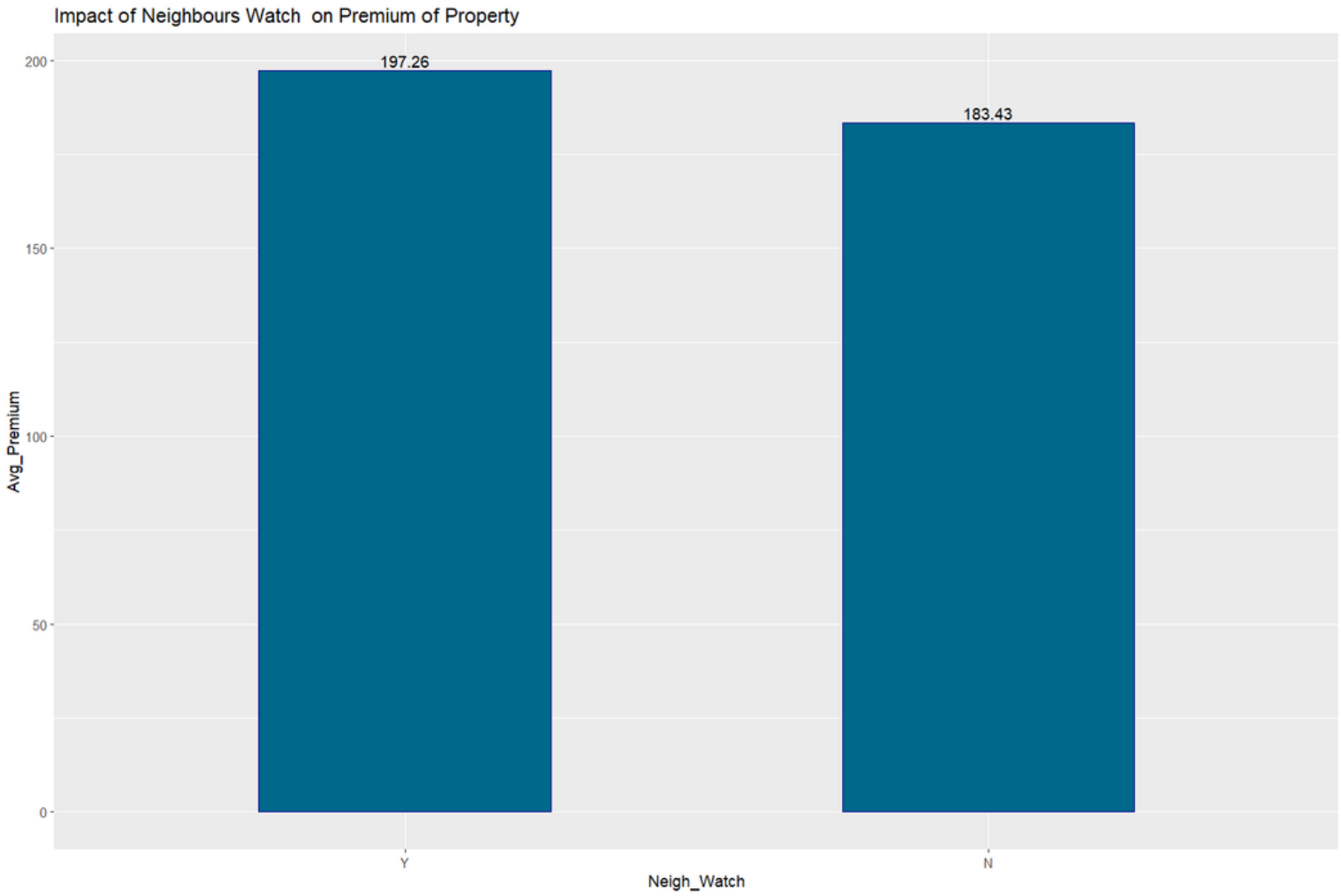
9) Checking the Premium for Property with Neigh_Watch



As we can see that the Premium for the Properties with Neighbours Watch is the Highest , which is almost 197 dollars.

Features of Properties

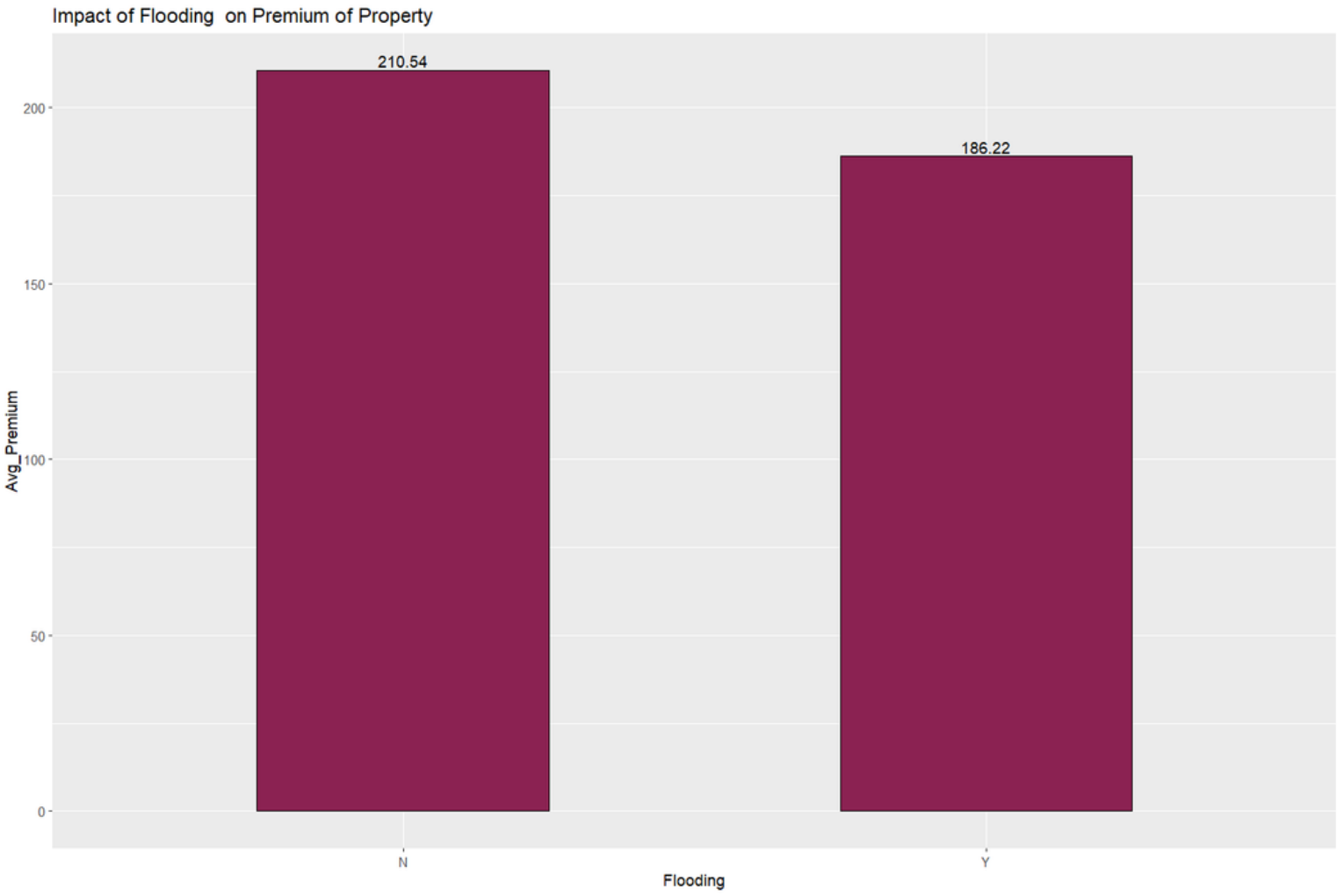
10) Checking the Premium for Property with Building Cover



As we can see that the Premium for the Properties with Neighbours Watch is the Highest , which is almost 197 dollars.

Features of Properties

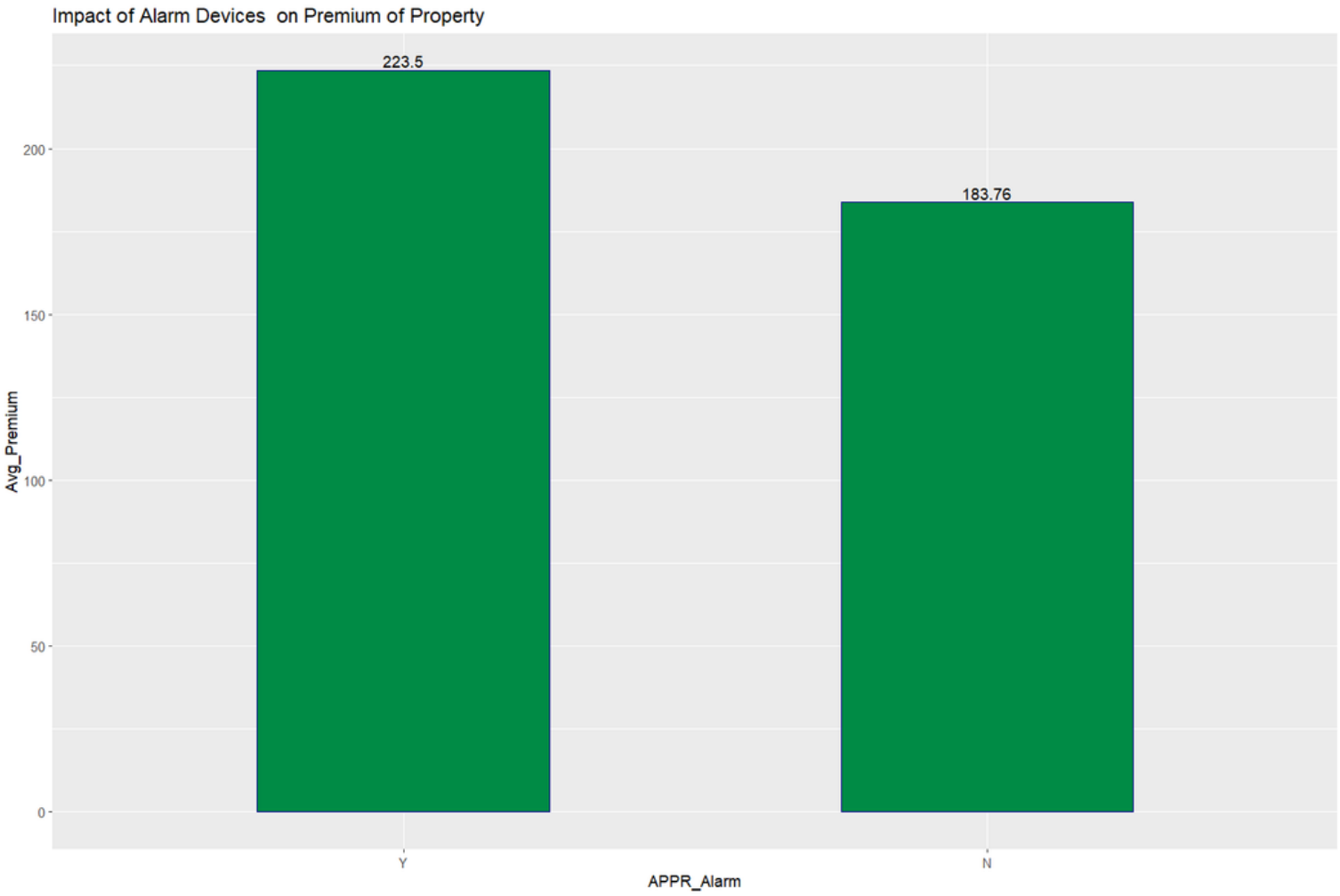
11) Checking the Premium for Property In Flooding Zone



As we can see that the Premium for the Properties which is not in Flooding Zone is the Highest , which is almost 210 dollars.

Features of Properties

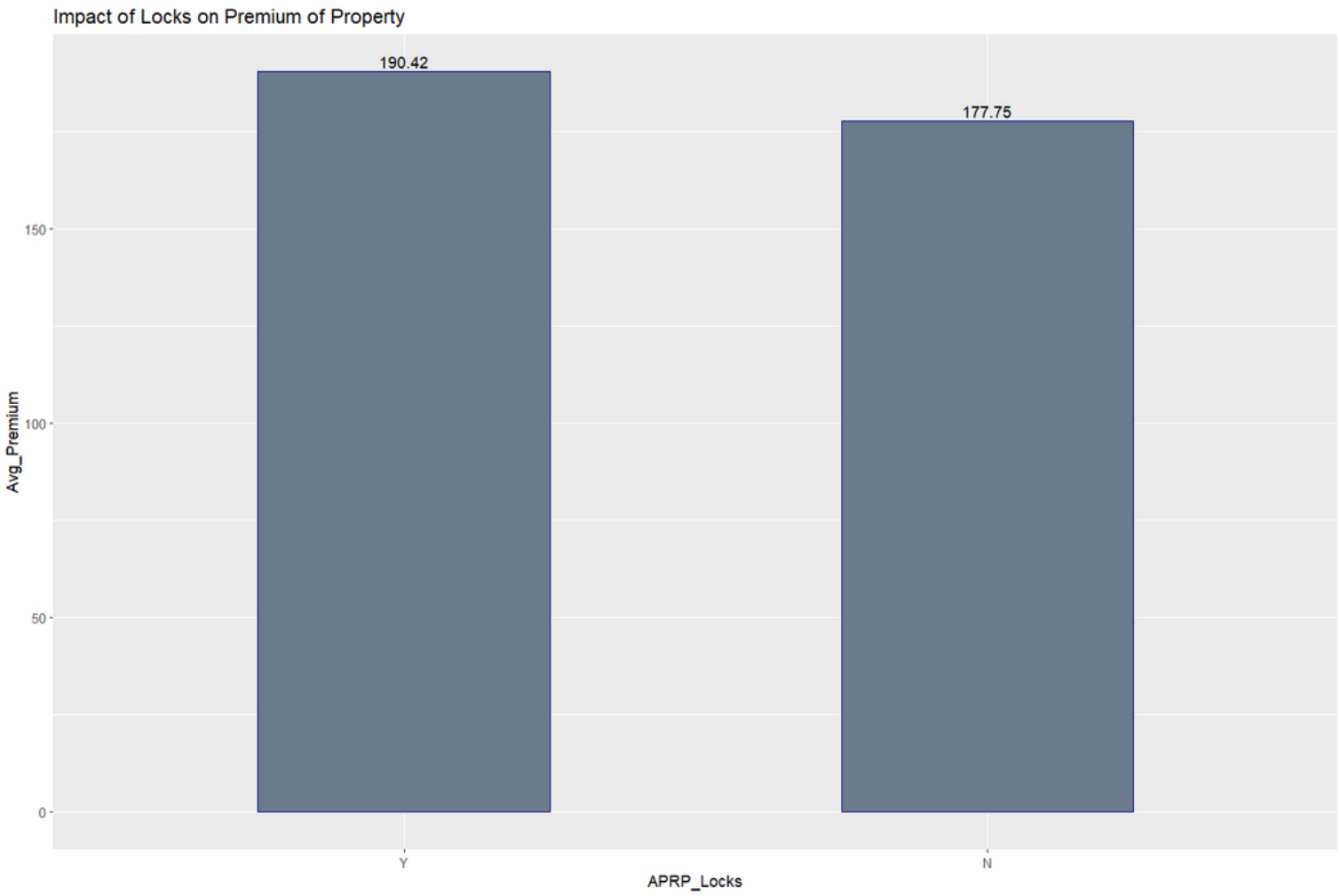
12) Checking the Premium for Property with Alarm Devices



As we can see that the Premium for the Properties with Alarm devices Installed is the Highest , which is almost 224 dollars.

Features of Properties

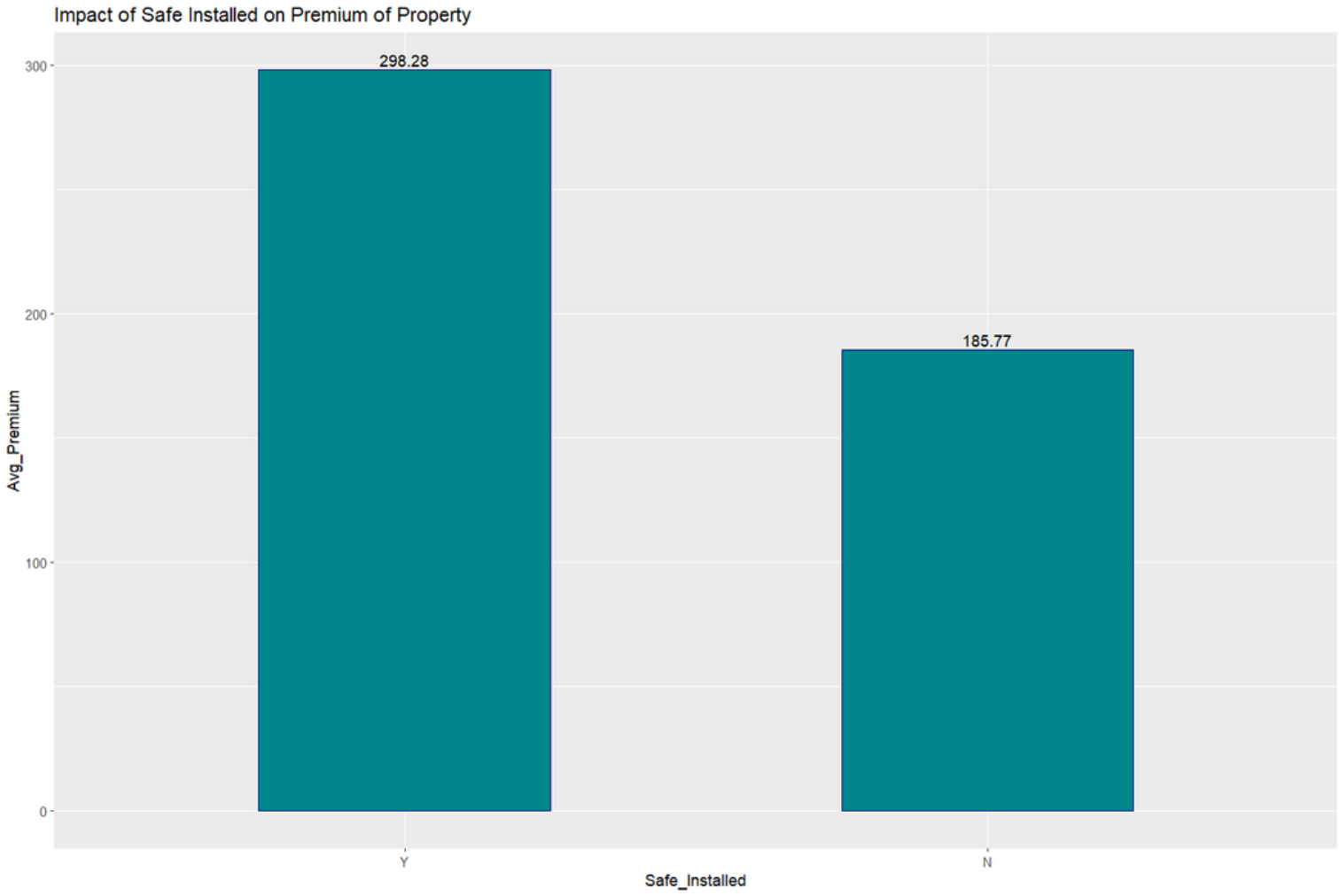
13) Checking the Premium for Property With Locks



As we can see that the Premium for the Properties with Locks installed is the Highest , which is almost 190 dollars.

Features of Properties

14) Checking the Premium for Property With Safe Installed



As we can see that the Premium for the Properties with Safe installed is the Highest , which is almost 298 dollars.

Conclusion

In this Internship we analyzed many variables to know their impact on the Home and Personnel Insurance Policies with the help of R-tools and R-Studio Software, Which really came in handy. These Variables were further plotted in form of Bar chart and Area chart for better understanding the analysis . The main objective of this Internship was to find the 1) Characteristics of Consumers who are likely to make an default & 2) Features of the Properties that drive up the Premium prices. So lets finally conclude the Report with answers we got from the Analysis , Those are :-

1) Characteristics of Customers who are more Likely to Default are :-
Male , Employed Male , Retired Customer , Married Customer .

2) Features of the Property that drive up Premium Prices are :-
Alarms , Safe and locks Installed & No Subsidence and No Flooding Zone & No. of Bedrooms and No. of times a property is listed & the buildings build before 1750

Important Links

1) Link for the Graphs :-

https://drive.google.com/drive/folders/1N2dD50YbHWSeT4wSoyYI-OYhG-8UMMfU?usp=sharing

2) Link for CRISP-DM Report :-

https://drive.google.com/drive/folders/1N2dD50YbHWSeT4wSoyYI-OYhG-8UMMfU?usp=sharing

3) R Internship File :-

https://drive.google.com/drive/folders/1N2dD50YbHWSeT4wSoyYI-OYhG-8UMMfU?usp=sharing

THANK YOU

Yash Agrawal