Foundations of Data Science
(UCS548)

Dashboard Submission

Health Insurance Survey

Submitted By:
Yatin Goyal
102003655
3COE26

Submitted To:
Dr. Sharad Saxena

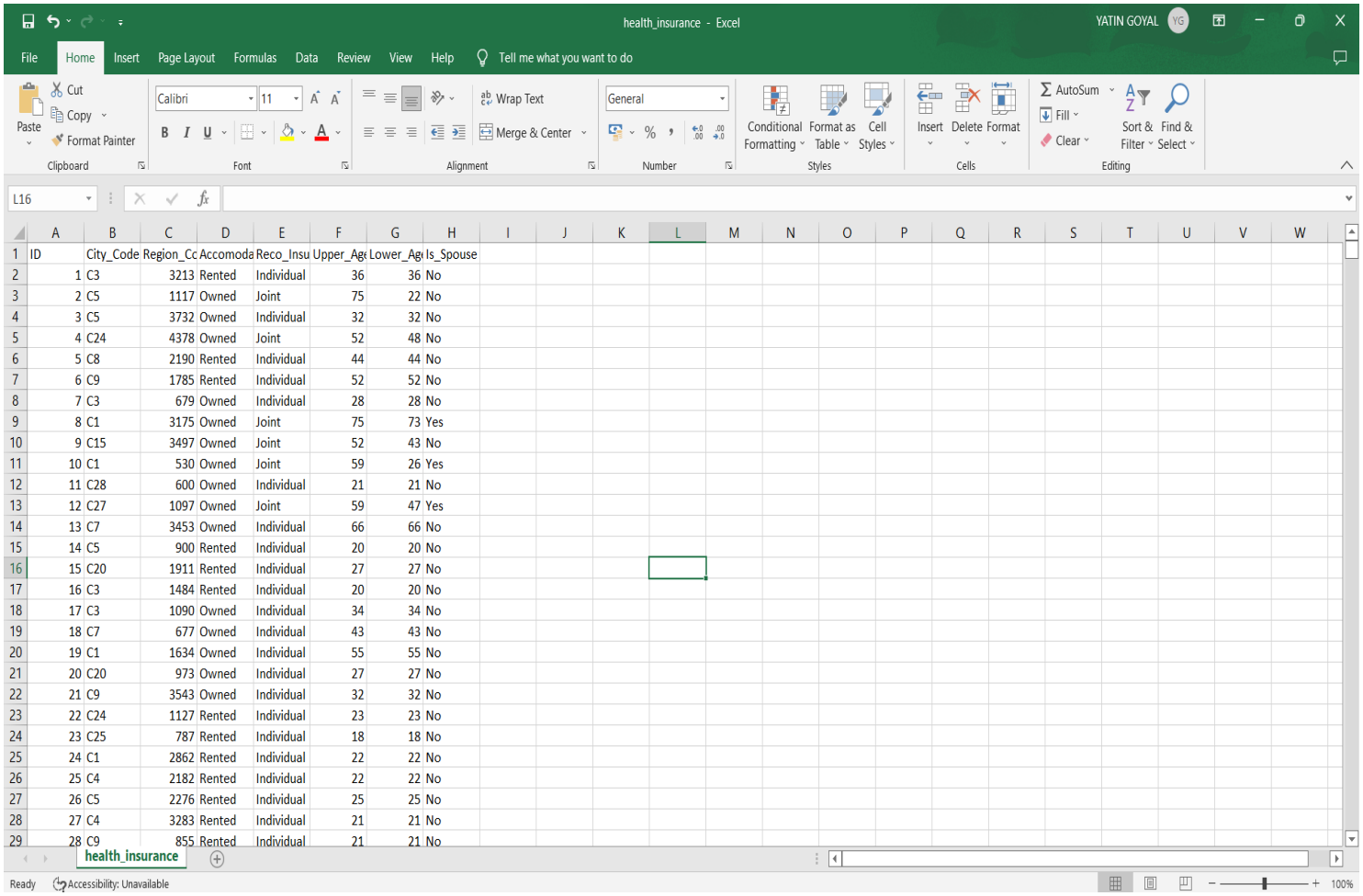
July 2022 – December 2022

About the dataset

FinMan Company is looking to leverage their client base by cross selling insurance products to existing customers. Insurance policies are offered to prospective and existing clients based on website landing and consumer election to fill out additional information forms. The project dataset is provided by Analytics Vidhya via Kaggle. Data includes demographic features, policy features (for current customers) and example positive classifications for ML model validation and interpretation.

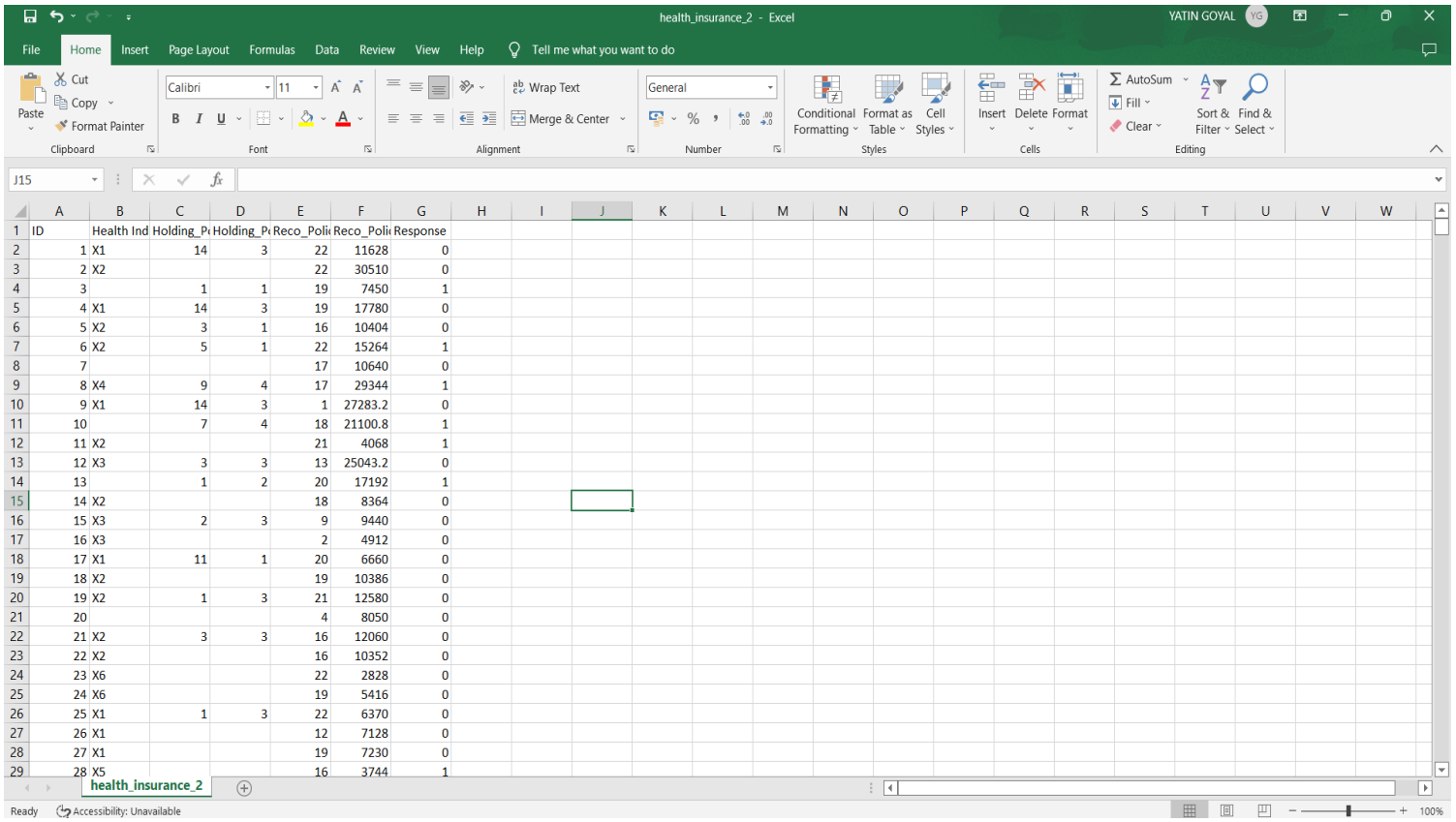
This dataset is divided into 2 multiple table:

Table 1:



ID	City_Code	Region_Cc	Accomoda	Reco_Insu	Upper_Ag	Lower_Ag	Is_Spouse
1	C3	3213	Rented	Individual	36	36	No
2	C5	1117	Owned	Joint	75	22	No
3	C5	3732	Owned	Individual	32	32	No
4	C24	4378	Owned	Joint	52	48	No
5	C8	2190	Rented	Individual	44	44	No
6	C9	1785	Rented	Individual	52	52	No
7	C3	679	Owned	Individual	28	28	No
8	C1	3175	Owned	Joint	75	73	Yes
9	C15	3497	Owned	Joint	52	43	No
10	C1	530	Owned	Joint	59	26	Yes
11	C28	600	Owned	Individual	21	21	No
12	C27	1097	Owned	Joint	59	47	Yes
13	C7	3453	Owned	Individual	66	66	No
14	C5	900	Rented	Individual	20	20	No
15	C20	1911	Rented	Individual	27	27	No
16	C3	1484	Rented	Individual	20	20	No
17	C3	1090	Owned	Individual	34	34	No
18	C7	677	Owned	Individual	43	43	No
19	C1	1634	Owned	Individual	55	55	No
20	C20	973	Owned	Individual	27	27	No
21	C9	3543	Owned	Individual	32	32	No
22	C24	1127	Rented	Individual	23	23	No
23	C25	787	Rented	Individual	18	18	No
24	C1	2862	Rented	Individual	22	22	No
25	C4	2182	Rented	Individual	22	22	No
26	C5	2276	Rented	Individual	25	25	No
27	C4	3283	Rented	Individual	21	21	No
28	C9	855	Rented	Individual	21	21	No

Table 2:

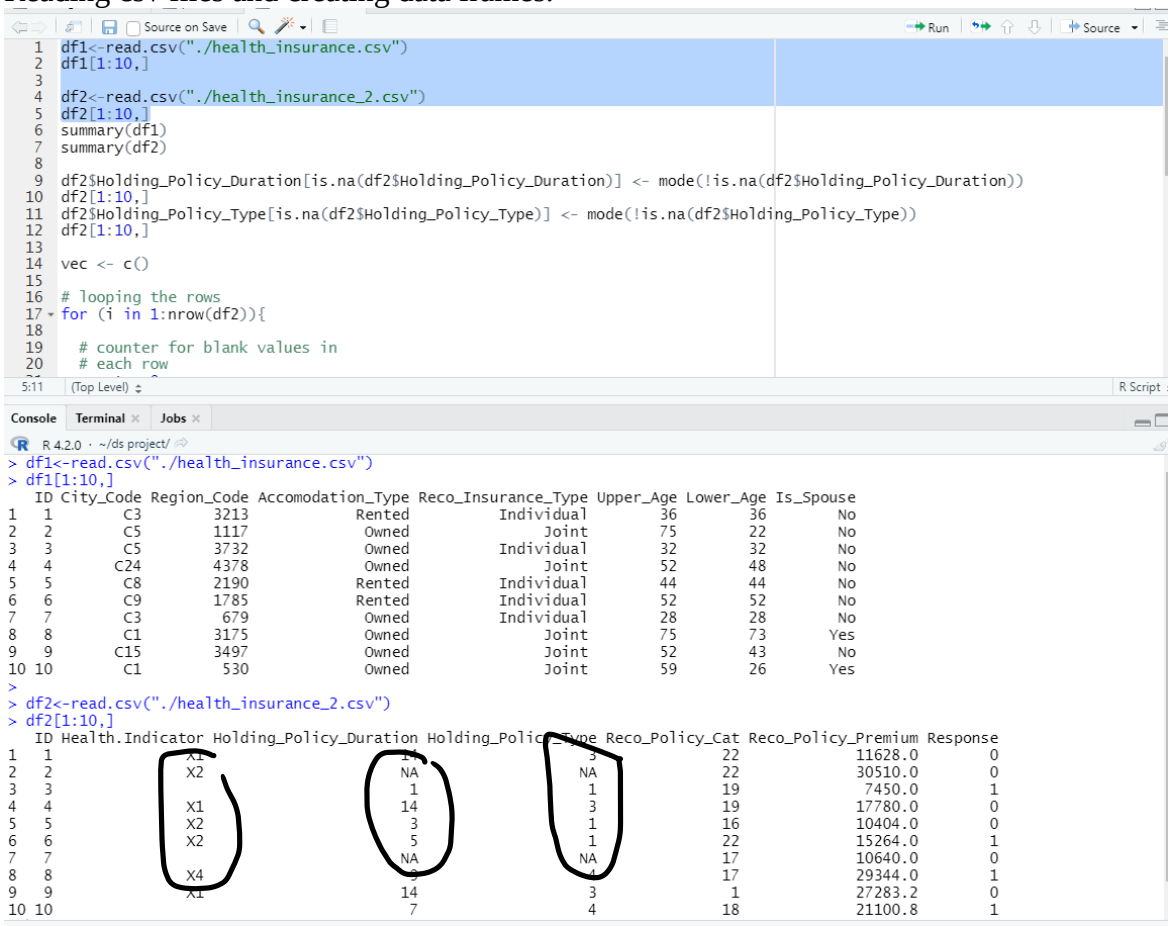


ID	Health Ind	Holding_P	Holding_P	Reco_Poli	Reco_Poli	Response
1	X1	14	3	22	11628	0
2	X2			22	30510	0
3		1	1	19	7450	1
4	X1	14	3	19	17780	0
5	X2	3	1	16	10404	0
6	X2	5	1	22	15264	1
7				17	10640	0
8	X4	9	4	17	29344	1
9	X1	14	3	1	27283.2	0
10		7	4	18	21100.8	1
11	X2			21	4068	1
12	X3	3	3	13	25043.2	0
13		1	2	20	17192	1
14	X2			18	8364	0
15	X3	2	3	9	9440	0
16	X3			2	4912	0
17	X1	11	1	20	6660	0
18	X2			19	10386	0
19	X2	1	3	21	12580	0
20				4	8050	0
21	X2	3	3	16	12060	0
22	X2			16	10352	0
23	X6			22	2828	0
24	X6			19	5416	0
25	X1	1	3	22	6370	0
26	X1			12	7128	0
27	X1			19	7230	0
28	X5			16	3744	1

Now, running queries in R to clean these tables.

Data Cleaning

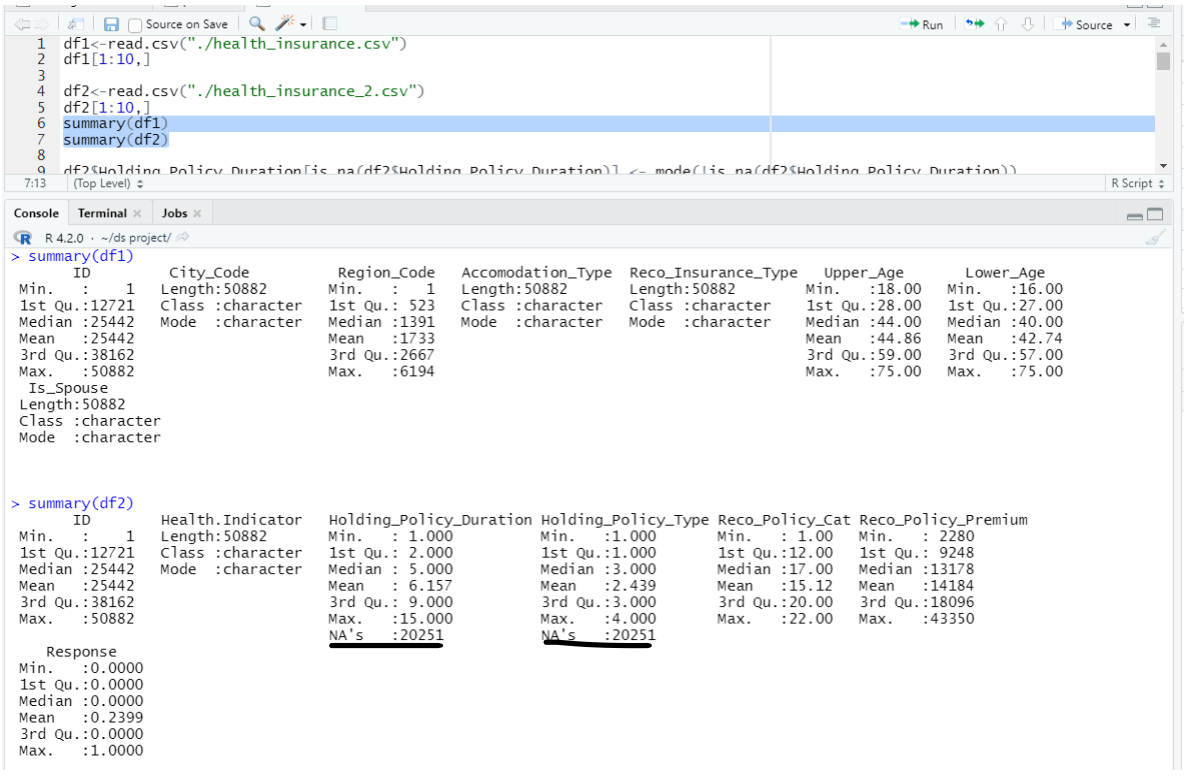
Reading csv files and creating data frames.



```
df1<-read.csv("./health_insurance.csv")
df1[1:10,]

df2<-read.csv("./health_insurance_2.csv")
df2[1:10,]
summary(df1)
summary(df2)

df2$Holding_Policy_Duration[is.na(df2$Holding_Policy_Duration)] <- mode(!is.na(df2$Holding_Policy_Duration))
df2[1:10,]
df2$Holding_Policy_Type[is.na(df2$Holding_Policy_Type)] <- mode(!is.na(df2$Holding_Policy_Type))
df2[1:10,]

vec <- c()

# looping the rows
for (i in 1:nrow(df2)){

  # counter for blank values in
  # each row
```

```
> df1<-read.csv("./health_insurance.csv")
> df1[1:10,]
   ID City_Code Region_Code Accomodation_Type Reco_Insurance_Type Upper_Age Lower_Age Is_Spouse
1   1        C3        3213            Rented          Individual        36        36        No
2   2        C5        1117             Owned               Joint        75        22        No
3   3        C5        3732             Owned          Individual        32        32        No
4   4       C24        4378             Owned               Joint        52        48        No
5   5        C8        2190            Rented          Individual        44        44        No
6   6        C9        1785            Rented          Individual        52        52        No
7   7        C3         679             Owned          Individual        28        28        No
8   8        C1        3175             Owned               Joint        75        73       Yes
9   9       C15        3497             Owned               Joint        52        43        No
10 10        C1         530             Owned               Joint        59        26       Yes
>
> df2<-read.csv("./health_insurance_2.csv")
> df2[1:10,]
   ID Health.Indicator Holding_Policy_Duration Holding_Policy_Type Reco_Policy_Cat Reco_Policy_Premium Response
1   1               X1                      1                   3             22            11628.0        0
2   2               X2                     NA                  NA             22            30510.0        0
3   3                                       1                   1             19             7450.0        1
4   4               X1                     14                   3             19            17780.0        0
5   5               X2                      3                   1             16            10404.0        0
6   6               X2                      5                   1             22            15264.0        1
7   7                                      NA                  NA             17            10640.0        0
8   8               X4                      3                   4             17            29344.0        1
9   9               X1                     14                   3              1            27283.2        0
10 10                                       7                   4             18            21100.8        1
```

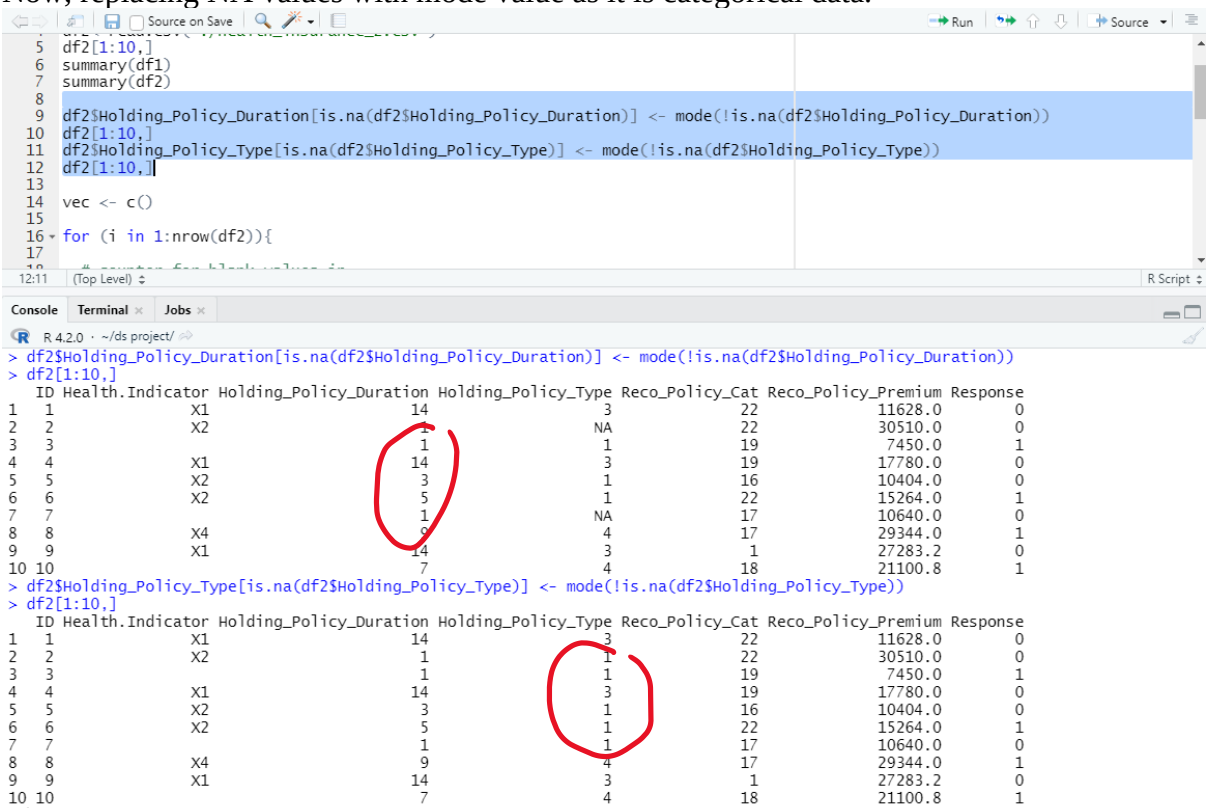
Checking NA and missing values in data.



```
df1<-read.csv("./health_insurance.csv")
df1[1:10,]

df2<-read.csv("./health_insurance_2.csv")
df2[1:10,]
summary(df1)
summary(df2)
```

```
> summary(df1)
       ID          City_Code          Region_Code   Accomodation_Type  Reco_Insurance_Type   Upper_Age       Lower_Age
 Min.   :    1   Length:50882       Min.   :   1   Length:50882       Length:50882        Min.   :18.00   Min.   :16.00
 1st Qu.:12721   Class :character   1st Qu.: 523   Class :character   Class :character    1st Qu.:28.00   1st Qu.:27.00
 Median :25442   Mode  :character   Median :1391   Mode  :character   Mode  :character    Median :44.00   Median :40.00
 Mean   :25442                      Mean   :1733                                          Mean   :44.86   Mean   :42.74
 3rd Qu.:38162                      3rd Qu.:2667                                          3rd Qu.:59.00   3rd Qu.:57.00
 Max.   :50882                      Max.   :6194                                          Max.   :75.00   Max.   :75.00
  Is_Spouse
 Length:50882
 Class :character
 Mode  :character



> summary(df2)
       ID        Health.Indicator   Holding_Policy_Duration Holding_Policy_Type Reco_Policy_Cat Reco_Policy_Premium
 Min.   :    1   Length:50882       Min.   : 1.000          Min.   :1.000       Min.   : 1.00   Min.   : 2280
 1st Qu.:12721   Class :character   1st Qu.: 2.000          1st Qu.:1.000       1st Qu.:12.00   1st Qu.: 9248
 Median :25442   Mode  :character   Median : 5.000          Median :3.000       Median :17.00   Median :13178
 Mean   :25442                      Mean   : 6.157          Mean   :2.439       Mean   :15.12   Mean   :14184
 3rd Qu.:38162                      3rd Qu.: 9.000          3rd Qu.:3.000       3rd Qu.:20.00   3rd Qu.:18096
 Max.   :50882                      Max.   :15.000          Max.   :4.000       Max.   :22.00   Max.   :43350
                                    NA's   :20251           NA's   :20251
    Response
 Min.   :0.0000
 1st Qu.:0.0000
 Median :0.0000
 Mean   :0.2399
 3rd Qu.:0.0000
 Max.   :1.0000
```

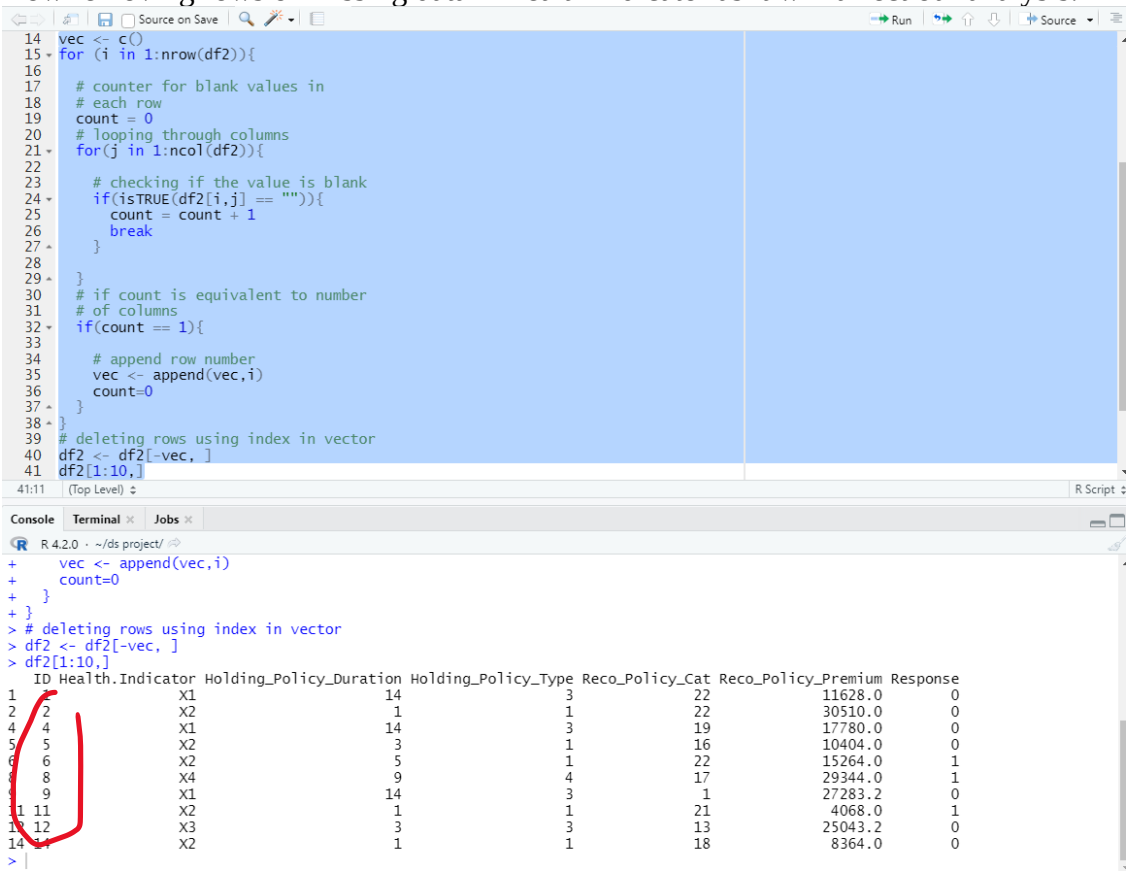
Now, replacing NA values with mode value as it is categorical data.



```
df2[1:10,]
summary(df1)
summary(df2)

df2$Holding_Policy_Duration[is.na(df2$Holding_Policy_Duration)] <- mode(!is.na(df2$Holding_Policy_Duration))
df2[1:10,]
df2$Holding_Policy_Type[is.na(df2$Holding_Policy_Type)] <- mode(!is.na(df2$Holding_Policy_Type))
df2[1:10,]

vec <- c()

for (i in 1:nrow(df2)){
```

```
> df2$Holding_Policy_Duration[is.na(df2$Holding_Policy_Duration)] <- mode(!is.na(df2$Holding_Policy_Duration))
> df2[1:10,]
   ID Health.Indicator Holding_Policy_Duration Holding_Policy_Type Reco_Policy_Cat Reco_Policy_Premium Response
1   1               X1                      14                   3              22             11628.0        0
2   2               X2                       1                  NA              22             30510.0        0
3   3                                        1                   1              19              7450.0        1
4   4               X1                      14                   3              19             17780.0        0
5   5               X2                       3                   1              16             10404.0        0
6   6               X2                       5                   1              22             15264.0        1
7   7                                        1                  NA              17             10640.0        0
8   8               X4                       9                   4              17             29344.0        1
9   9               X1                      14                   3               1             27283.2        0
10 10                                        7                   4              18             21100.8        1
> df2$Holding_Policy_Type[is.na(df2$Holding_Policy_Type)] <- mode(!is.na(df2$Holding_Policy_Type))
> df2[1:10,]
   ID Health.Indicator Holding_Policy_Duration Holding_Policy_Type Reco_Policy_Cat Reco_Policy_Premium Response
1   1               X1                      14                   3              22             11628.0        0
2   2               X2                       1                   1              22             30510.0        0
3   3                                        1                   1              19              7450.0        1
4   4               X1                      14                   3              19             17780.0        0
5   5               X2                       3                   1              16             10404.0        0
6   6               X2                       5                   1              22             15264.0        1
7   7                                        1                   1              17             10640.0        0
8   8               X4                       9                   4              17             29344.0        1
9   9               X1                      14                   3               1             27283.2        0
10 10                                        7                   4              18             21100.8        1
>
```

Now removing rows of missing data in health indicator as it will affect our analysis.



```
vec <- c()
for (i in 1:nrow(df2)){

  # counter for blank values in
  # each row
  count = 0
  # looping through columns
  for(j in 1:ncol(df2)){

    # checking if the value is blank
    if(isTRUE(df2[i,j] == "")){
      count = count + 1
      break
    }

  }
  # if count is equivalent to number
  # of columns
  if(count == 1){

    # append row number
    vec <- append(vec,i)
    count=0
  }
}
# deleting rows using index in vector
df2 <- df2[-vec, ]
df2[1:10,]
```

```
+     vec <- append(vec,i)
+     count=0
+   }
+ }
> # deleting rows using index in vector
> df2 <- df2[-vec, ]
> df2[1:10,]
   ID Health.Indicator Holding_Policy_Duration Holding_Policy_Type Reco_Policy_Cat Reco_Policy_Premium Response
1   1               X1                      14                   3              22             11628.0        0
2   2               X2                       1                   1              22             30510.0        0
4   4               X1                      14                   3              19             17780.0        0
5   5               X2                       3                   1              16             10404.0        0
6   6               X2                       5                   1              22             15264.0        1
8   8               X4                       9                   4              17             29344.0        1
9   9               X1                      14                   3               1             27283.2        0
11 11               X2                       1                   1              21              4068.0        1
12 12               X3                       3                   3              13             25043.2        0
14 14               X2                       1                   1              18              8364.0        0
>
```

After cleaning the data, we will merge the two dataframes.

Queries for merging dataset:

Merging all dataset into a single dataset by using inner_join() function as in one dataframe some rows were removed due to missing values.

```
43 #joining subsets
44 df= df1 %>% inner_join(df2,by="ID")
45 df[1:10,]
```

45:10 (Top Level) R Script

Console Terminal Jobs

R 4.2.0 · ~/ds project/

```
   ID City_Code Region_Code Accomodation_Type Reco_Insurance_Type Upper_Age Lower_Age Is_Spouse Health.Indicator
1   1        C3         3213            Rented          Individual        36        36        No               X1
2   2        C5         1117             Owned               Joint        75        22        No               X2
3   4       C24         4378             Owned               Joint        52        48        No               X1
4   5        C8         2190            Rented          Individual        44        44        No               X2
5   6        C9         1785            Rented          Individual        52        52        No               X2
6   8        C1         3175             Owned               Joint        75        73       Yes               X4
7   9       C15         3497             Owned               Joint        52        43        No               X1
8  11       C28          600             Owned          Individual        21        21        No               X2
9  12       C27         1097             Owned               Joint        59        47       Yes               X3
10 14        C5          900            Rented          Individual        20        20        No               X2
   Holding_Policy_Duration Holding_Policy_Type Reco_Policy_Cat Reco_Policy_Premium Response
1                       14                   3              22              11628.0        0
2                        1                   1              22              30510.0        0
3                       14                   3              19              17780.0        0
4                        3                   1              16              10404.0        0
5                        5                   1              22              15264.0        1
6                        9                   4              17              29344.0        1
7                       14                   3               1              27283.2        0
8                        1                   1              21               4068.0        1
9                        3                   3              13              25043.2        0
10                       1                   1              18               8364.0        0
>
```

Finally saving the dataset as rdata file named as insurance.rdata.

```
47 save(df,file = "C:/insurance.rdata")
```

47:37 (Top Level)

Console Terminal Jobs

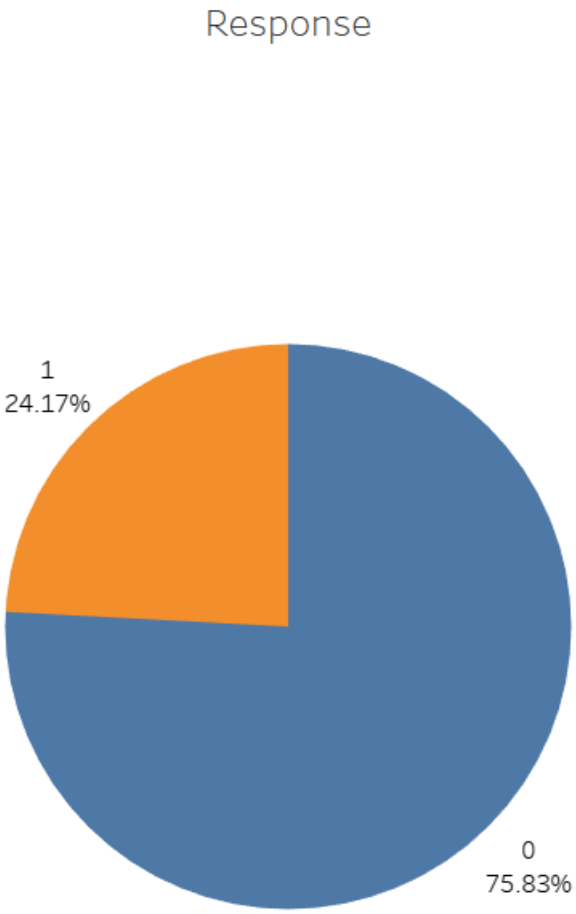
R 4.2.0 · ~/ds project/

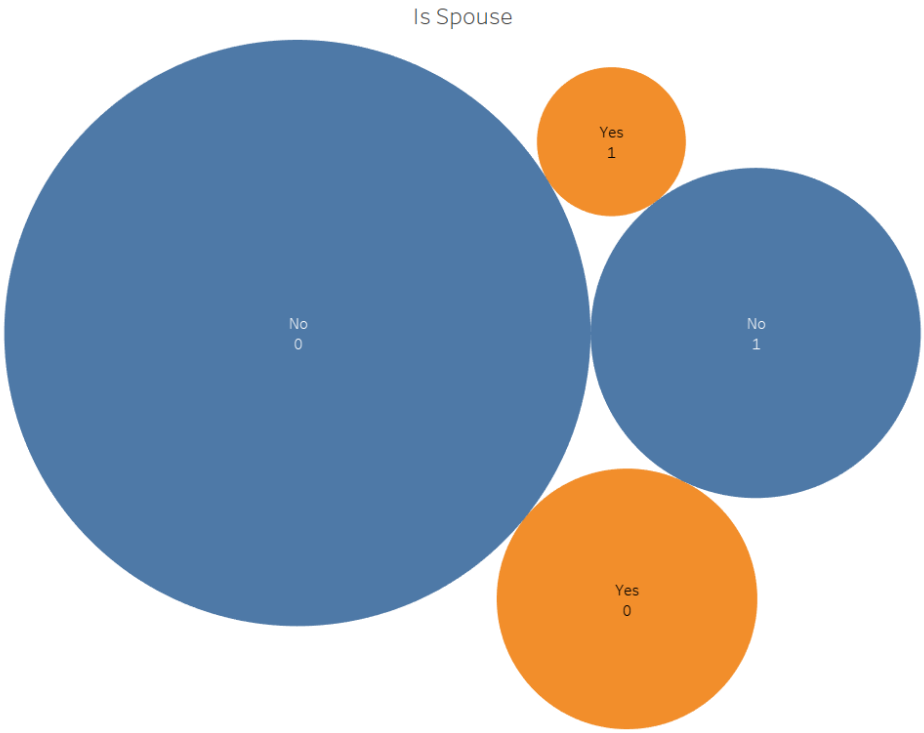
```
> save(df,file = "C:/insurance.rdata")
>
```

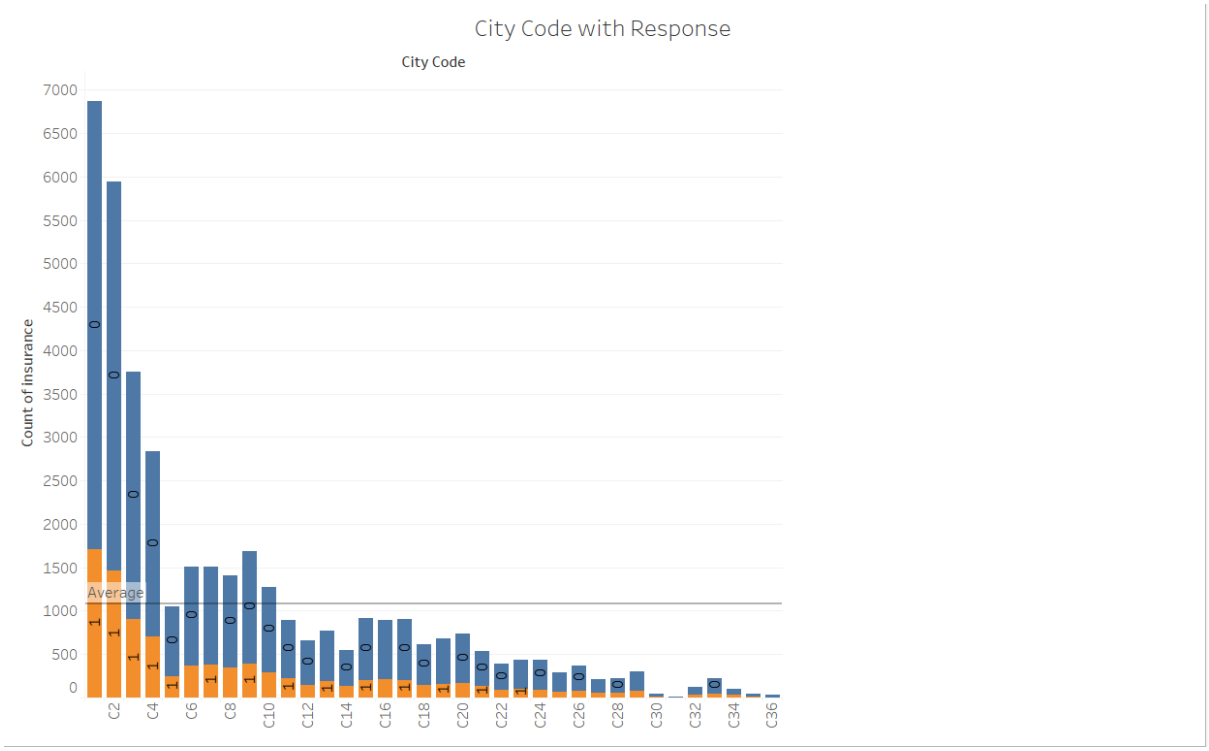
This dataset contains the following data:

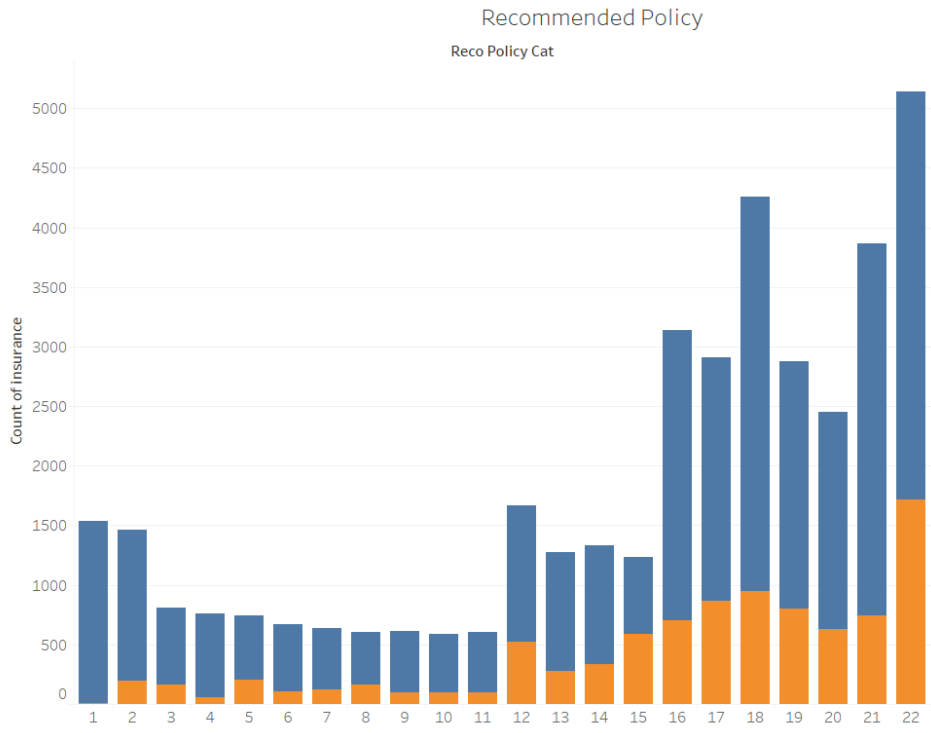
1. **ID** - Unique id.
2. **City_Code** - Code for the city of users.
3. **Region_Code** - Code for the region of the users.
4. **Accomodation_Type** - Customer owns/rents the house.
5. **Reco_Insurance_Type** - Joint or individual type for the recommended insurance.
6. **Upper_Age** - Maximium age of the customer.
7. **Lower_Age** - Minimium age of the customer.
8. **Is_Spouse** - If the customer is married or not.
9. **Health Indicator** - Encoded values for health of the customer.
10. **Holding_Policy_Duration** - Duration in year of holding policy.
11. **Holding_Policy_Type** - Type of holding policy.
12. **Reco_Policy_Cat** - Encoded values of recommended health insurance.
13. **Reco_Policy_Premium**- Annual premium (INR) for the recommended health insurance.
14. **Response**- Whether the client filled the form or not.

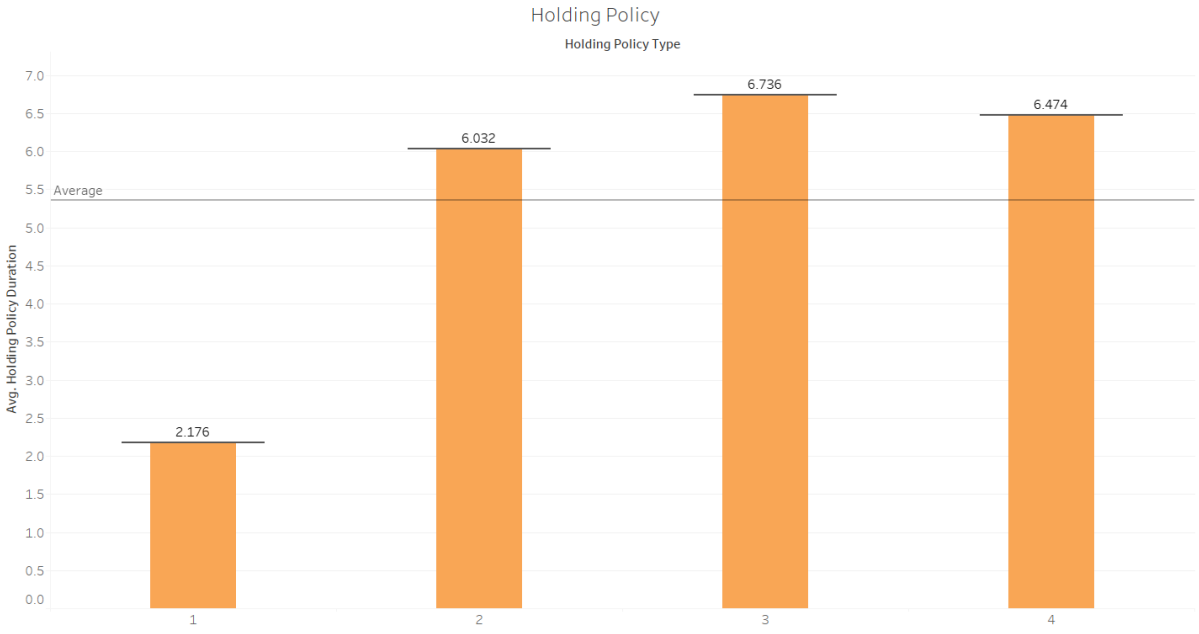
Tableau Queries



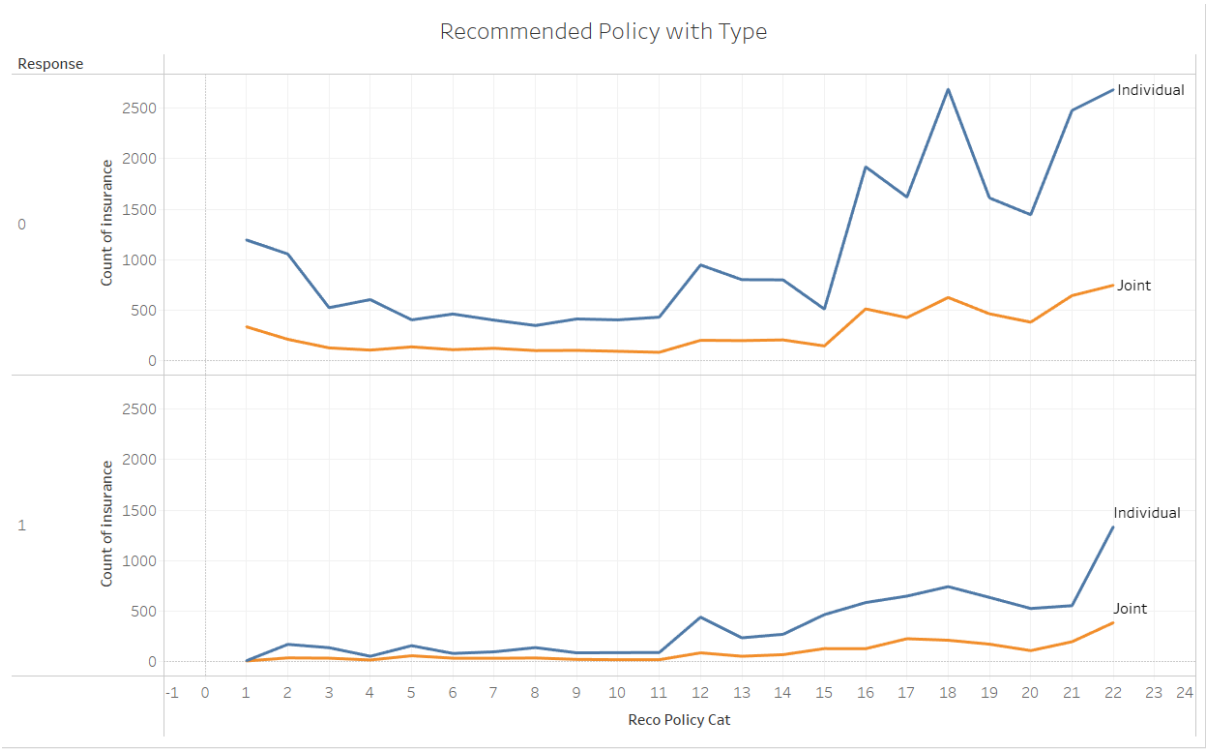

Is Spouse
Yes
1
No
0
No
1
Yes
0


City Code with Response
City Code
Count of insurance
7000
6500
6000
5500
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
Average
C2
C4
C6
C8
C10
C12
C14
C16
C18
C20
C22
C24
C26
C28
C30
C32
C34
C36

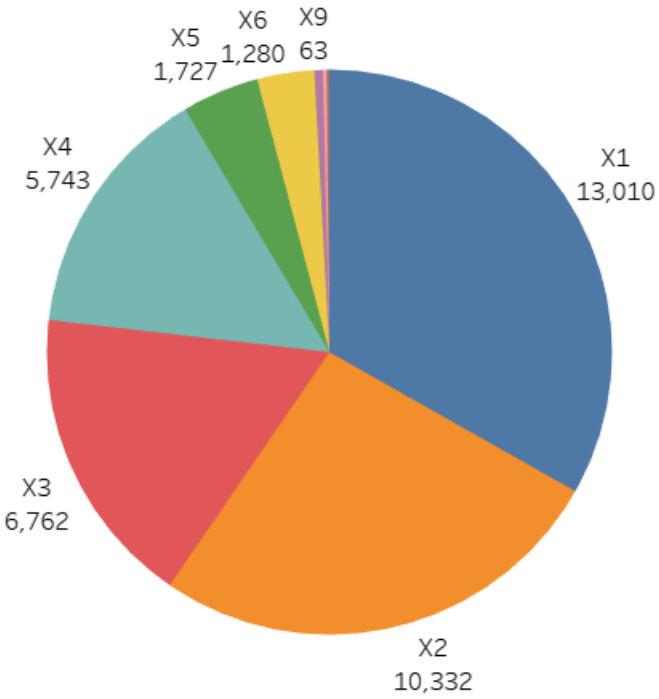

Recommended Policy
Reco Policy Cat
Count of Insurance
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22

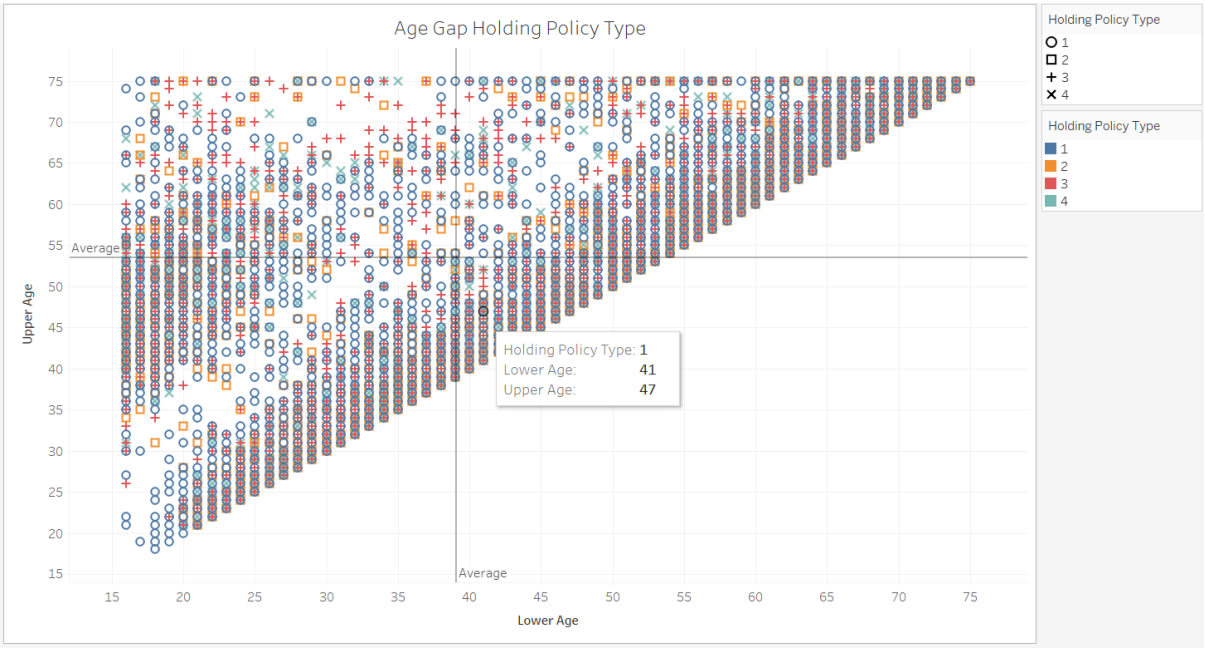

Holding Policy
Holding Policy Type
Avg. Holding Policy Duration
7.0
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Average
2.176
6.032
6.736
6.474
1
2
3
4


Recomended Policy with Premium
Reco Policy Cat
Reco Policy Premium
₹ 45.00K
₹ 40.00K
₹ 35.00K
₹ 30.00K
₹ 25.00K
₹ 20.00K
₹ 15.00K
₹ 10.00K
₹ 5.00K
₹ 0.00K
Average
Minimum
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22


Recommended Policy with Type
Response
0
1
Count of insurance
2500
2000
1500
1000
500
0
Individual
Joint
-1 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
Reco Policy Cat

Health indicator





Final Dashboard