Project for R Course Data Science and Application Advanced Diploma Metro College

Ana Clara Tupinambá Freitas, oriented by Professor Hamid Rajaee

5/24/2021

Methodology

- Business understanding;
- EDA:
  - Univariate analysis; and
  - Bivariate analysis.

Introduction

What is Churn?

Churn is a measurement of the percentage of accounts that cancel or choose not to renew their subscriptions. A high churn rate can negatively impact Monthly Recurring Revenue (MRR) and can also indicate dissatisfaction with a product or service.

$Churn = \frac{CustomersLostInaPperiod}{CustomersAttheBeginningOfaPeriod}$

This project will treat churn related to patients of a diverse set of practices clinics.

Source: https://www.productplan.com/glossary/churn/

Methodology

This project will perform EDA and make presumptions about the data since contact to subject matter experts was not possible at this moment.

The goal is the creation of a model to predict churn of patients.

Loading Data

First Look at Data:

Shape of data:

```
## [1] 25000    13
```

Features:

```
##  [1] "practice_id"                    "patient_id"
##  [3] "gender"                         "age"
##  [5] "zip"                            "primary_insurance_company_id"
##  [7] "secondary_insurance_company_id" "patient_referral"
##  [9] "other_referral"                 "FirstVisit"
## [11] "LastVisit"                      "DaysLastVisit"
## [13] "Chrun"
```

Data Structure: We can see that the only numeric features, at this moment, are: patient_id(categorical) and age.

```
## 'data.frame':    25000 obs. of  13 variables:
##  $ practice_id                   : chr  "D17435" "D17435" "D17435" "D17435" ...
##  $ patient_id                    : int  806553553 806553536 806553528 806553525 806553524 806553517 8
##  $ gender                        : chr  "Female" "Female" "Male" "Male" ...
##  $ age                           : int  47 74 76 57 53 74 69 73 47 22 ...
##  $ zip                           : chr  "V8P5H7" "V9Z1C5" "V8L2P7" "V9B2W3" ...
##  $ primary_insurance_company_id  : chr  "4136" "4273" "3816" "3816" ...
##  $ secondary_insurance_company_id: chr  "3386" NA NA "1947" ...
##  $ patient_referral              : chr  NA NA NA NA ...
##  $ other_referral                : chr  "Choboter David" "Groff Tera" "Culligan Peter" "Thom David"
##  $ FirstVisit                    : chr  "1/28/2016" "2/16/2016" "2/24/2016" "1/31/2017" ...
##  $ LastVisit                     : chr  "6/29/2017" "12/31/2019" "11/27/2019" "6/28/2017" ...
##  $ DaysLastVisit                 : chr  "7/3/1902" "1/0/1900" "2/3/1900" "7/4/1902" ...
##  $ Chrun                         : chr  "YES" "NO" "NO" "YES" ...
```

```
## Rows: 25,000
## Columns: 13
## $ practice_id                    <chr> "D17435", "D17435", "D17435", "D17435",~
## $ patient_id                     <int> 806553553, 806553536, 806553528, 806553~
## $ gender                         <chr> "Female", "Female", "Male", "Male", "Fe~
## $ age                            <int> 47, 74, 76, 57, 53, 74, 69, 73, 47, 22,~
## $ zip                            <chr> "V8P5H7", "V9Z1C5", "V8L2P7", "V9B2W3",~
## $ primary_insurance_company_id   <chr> "4136", "4273", "3816", "3816", "1769",~
## $ secondary_insurance_company_id <chr> "3386", NA, NA, "1947", NA, NA, NA, NA,~
## $ patient_referral               <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ other_referral                 <chr> "Choboter David", "Groff Tera", "Cullig~
## $ FirstVisit                     <chr> "1/28/2016", "2/16/2016", "2/24/2016", ~
## $ LastVisit                      <chr> "6/29/2017", "12/31/2019", "11/27/2019"~
## $ DaysLastVisit                  <chr> "7/3/1902", "1/0/1900", "2/3/1900", "7/~
## $ Chrun                          <chr> "YES", "NO", "NO", "YES", "YES", "NO", ~
```

First 3 observations:

```
##   practice_id patient_id gender age    zip primary_insurance_company_id
## 1      D17435  806553553 Female  47 V8P5H7                         4136
## 2      D17435  806553536 Female  74 V9Z1C5                         4273
## 3      D17435  806553528   Male  76 V8L2P7                         3816
##   secondary_insurance_company_id patient_referral other_referral FirstVisit
## 1                           3386             <NA> Choboter David  1/28/2016
## 2                           <NA>             <NA>     Groff Tera  2/16/2016
## 3                           <NA>             <NA> Culligan Peter  2/24/2016
##    LastVisit DaysLastVisit Chrun
## 1  6/29/2017      7/3/1902   YES
## 2 12/31/2019      1/0/1900    NO
## 3 11/27/2019      2/3/1900    NO
```

Last 3 observations:

```
##       practice_id patient_id gender age     zip primary_insurance_company_id
## 24998      D24402       9889   Male  26 L7E 2P5                         <NA>
## 24999      D24402       9886 Female  15 L7E 0A4                           20
## 25000      D24402       9885   Male   6 L9W 2W7                           78
##       secondary_insurance_company_id patient_referral other_referral FirstVisit
## 24998                           <NA>             <NA>           <NA>   1/8/2018
## 24999                             91             <NA>           <NA>   2/5/2018
## 25000                           <NA>             <NA>           <NA>  1/17/2018
##       LastVisit DaysLastVisit Chrun
## 24998  3/7/2018           664    NO
## 24999 9/11/2019           111    NO
## 25000 8/14/2019           139    NO
```

Summary:

```
##  practice_id         patient_id           gender               age       
##  Length:25000       Min.   :        3   Length:25000       Min.   :  1.0  
##  Class :character   1st Qu.:    12014   Class :character   1st Qu.: 19.0  
##  Mode  :character   Median :    25638   Mode  :character   Median : 33.0  
##                     Mean   :  4778068                      Mean   : 38.6  
##                     3rd Qu.:   152315                      3rd Qu.: 55.0  
##                     Max.   :806553553                      Max.   :120.0  
##      zip            primary_insurance_company_id secondary_insurance_company_id
##  Length:25000       Length:25000                 Length:25000                  
##  Class :character   Class :character             Class :character              
##  Mode  :character   Mode  :character             Mode  :character              
##                                                                                
##                                                                                
##                                                                                
##  patient_referral   other_referral      FirstVisit         LastVisit        
##  Length:25000       Length:25000       Length:25000       Length:25000      
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
##  DaysLastVisit         Chrun          
##  Length:25000       Length:25000      
##  Class :character   Class :character  
##  Mode  :character   Mode  :character  
##                                       
##                                       
##                                       
```

```
## [1] 0
```

Range, minimum and maximum values of numeric features of data frame (data) :

```
## [1] "The range of < patient_id > is:  806553550 and its minimum and maximum values are: 3 & 806553553
## [2] "The range of < age > is:  119 and its minimum and maximum values are: 1 & 120"
```

Is there NAs in data frame? What's its percentage in each feature?

We can see that there are many more NAs in procedure description than in procedure code, both will be merged to create a new combined feature.

```
##                              Names Total_of_NAs Prop.NAs
## 1                      practice_id            0      0 %
## 2                       patient_id            0      0 %
## 3                           gender          279      1 %
## 4                              age            0      0 %
## 5                              zip          323      1 %
## 6    primary_insurance_company_id        14495     58 %
## 7  secondary_insurance_company_id        23493     94 %
## 8                 patient_referral        21834     87 %
## 9                   other_referral        21767     87 %
## 10                      FirstVisit            0      0 %
## 11                       LastVisit            0      0 %
## 12                  DaysLastVisit            0      0 %
## 13                           Chrun            0      0 %
```

There is a greater number of NAs for:

- primary_insurance_company_id;
- secondary_insurance_company_id;
- patient_referral; and
- other_referral.

These features will be converted to binary:

- If there is a value:Yes;
- If don't: No.

And renamed:

- primary_insurance_company_id: primary_insurance;
- secondary_insurance_company_id: secondary_insurance

```
##                  Names Total_of_NAs Prop.NAs
## 1          practice_id            0      0 %
## 2           patient_id            0      0 %
## 3               gender          279      1 %
## 4                  age            0      0 %
## 5                  zip          323      1 %
## 6    primary_insurance            0      0 %
## 7  secondary_insurance            0      0 %
## 8     patient_referral            0      0 %
## 9       other_referral            0      0 %
## 10          FirstVisit            0      0 %
## 11           LastVisit            0      0 %
## 12       DaysLastVisit            0      0 %
## 13               Chrun            0      0 %
```

Since 97.76% of data will be preserved and features don't have more NAs with more than 1% participation within group, NAs will be dropped.

Features

What practice_id presents ?

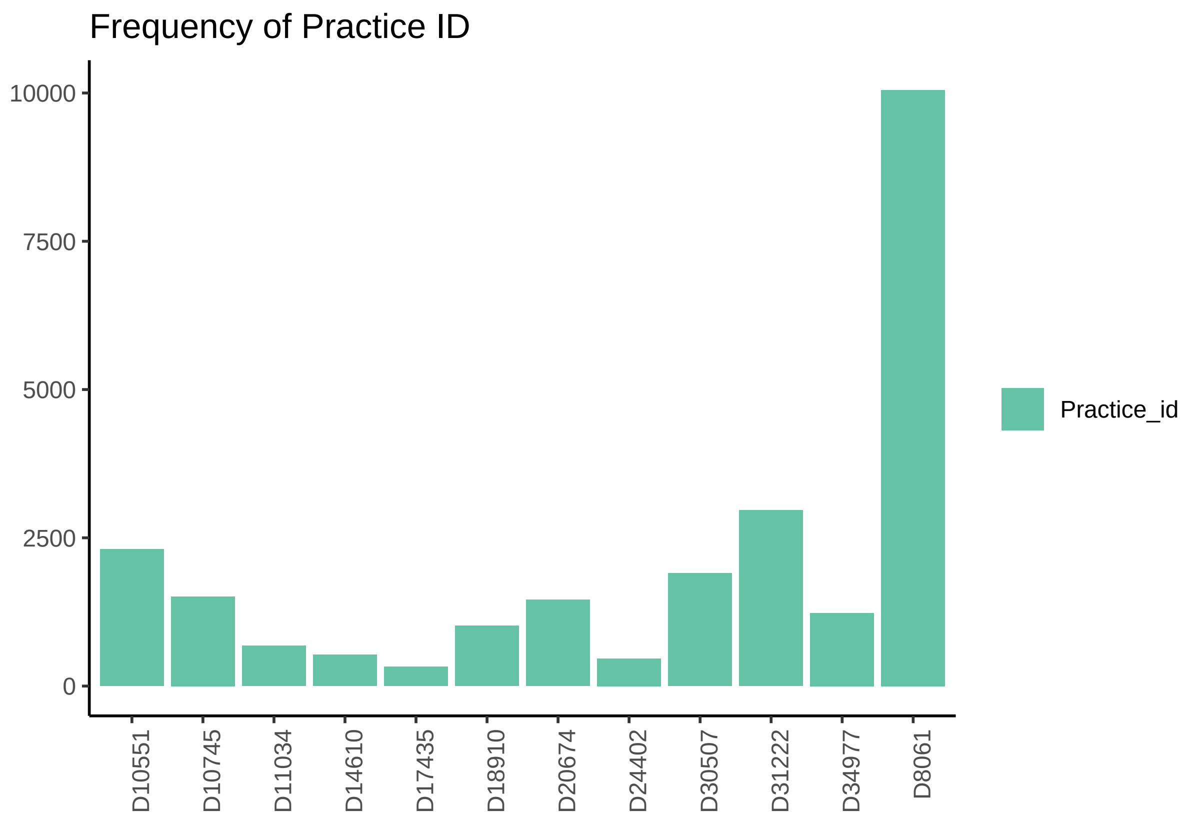
First 3 and last 3 observations:

```
## [1] "D17435" "D17435" "D17435"
```

```
## [1] "D24402" "D24402" "D24402"
```

Table 1: Frequency

practice_id	Freq
D10551	2310
D10745	1510
D11034	678
D14610	526
D17435	323
D18910	1020
D20674	1455
D24402	466
D30507	1903
D31222	2966
D34977	1232
D8061	10050



What patient_id presents ?

First 3 and last 3 observations:

```
## [1] 806553553 806553536 806553528
```

```
## [1] 9889 9886 9885
```

Is patient_id unique?

First observations:

```
## [1] 806553553 806553536 806553528 806553525 806553524 806553517
## 22739 Levels: 3 5 6 7 8 12 14 15 16 17 19 20 21 23 26 27 29 30 32 33 36 42 ... 806553553
```

Table 2: Frequency

patient_id	Freq
3	1
5	2
6	1
7	2
8	1
12	1

```
## [1] "There are  1700  duplicated values in patient_id."
```

A further look at the first observations, gives us some insights that patient id may be indeed insurance holder, since there are at least different genders, ages and zip for the same patient id.

```
## # A tibble: 10 x 13
## # Groups:   patient_id [8]
##    practice_id patient_id gender   age zip    primary_insuran~ secondary_insura~
##    <fct>       <fct>      <chr>  <int> <chr>  <chr>            <chr>
##  1 D11034      3          F         78 K2J2Z3 YES              NO
##  2 D18910      5          Female    63 M6L1T8 YES              NO
##  3 D11034      5          M         78 K2L2K8 YES              NO
##  4 D18910      6          Male      39 M6L1T8 NO               NO
##  5 D18910      7          Male      64 M6L1T8 YES              YES
##  6 D11034      7          M         67 K0A1A0 NO               NO
##  7 D18910      8          Male      41 M6L1T8 NO               NO
##  8 D11034      12         F         73 K2L2K8 YES              NO
##  9 D18910      14         Male      59 M6S2J4 YES              NO
## 10 D18910      15         Male      30 M6S2J4 NO               NO
## # ... with 6 more variables: patient_referral <chr>, other_referral <chr>,
## #   FirstVisit <chr>, LastVisit <chr>, DaysLastVisit <chr>, Chrun <chr>
```

What gender presents ?

First 3 and last 3 observations:

```
## [1] "Female" "Female" "Male"
```

```
## [1] "Male"   "Female" "Male"
```

What are the unique values?

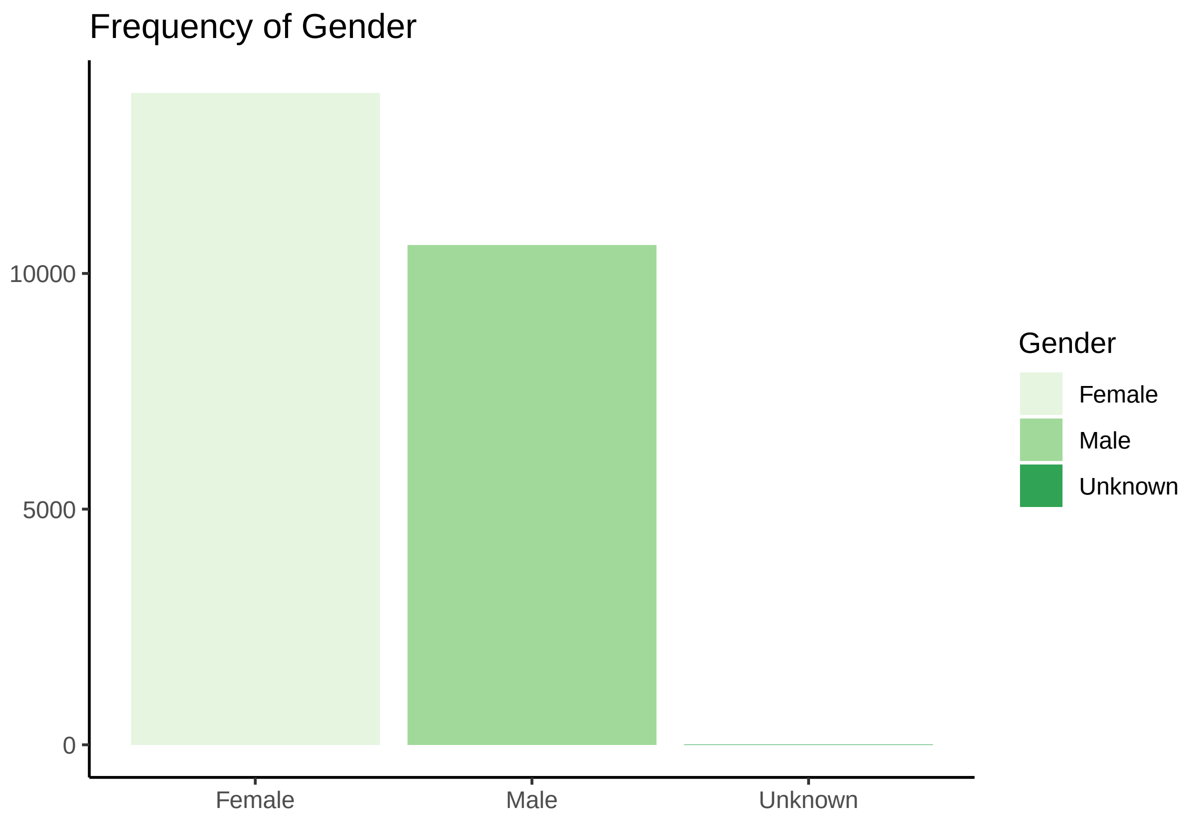
We see that there are different values that can be made to one:

```
## [1] "Before transformation: Female, Male, F, M, U"
```

```
## [1] "After transformation: Female, Male, Unknown"
```

Table 3: Frequency

gender	Freq
Female	13832
Male	10601
Unknown	6



We see that the number of observations of unknown gender are very, small. These observations will be dropped.

What age presents ?

First 3 and last 3 observations:

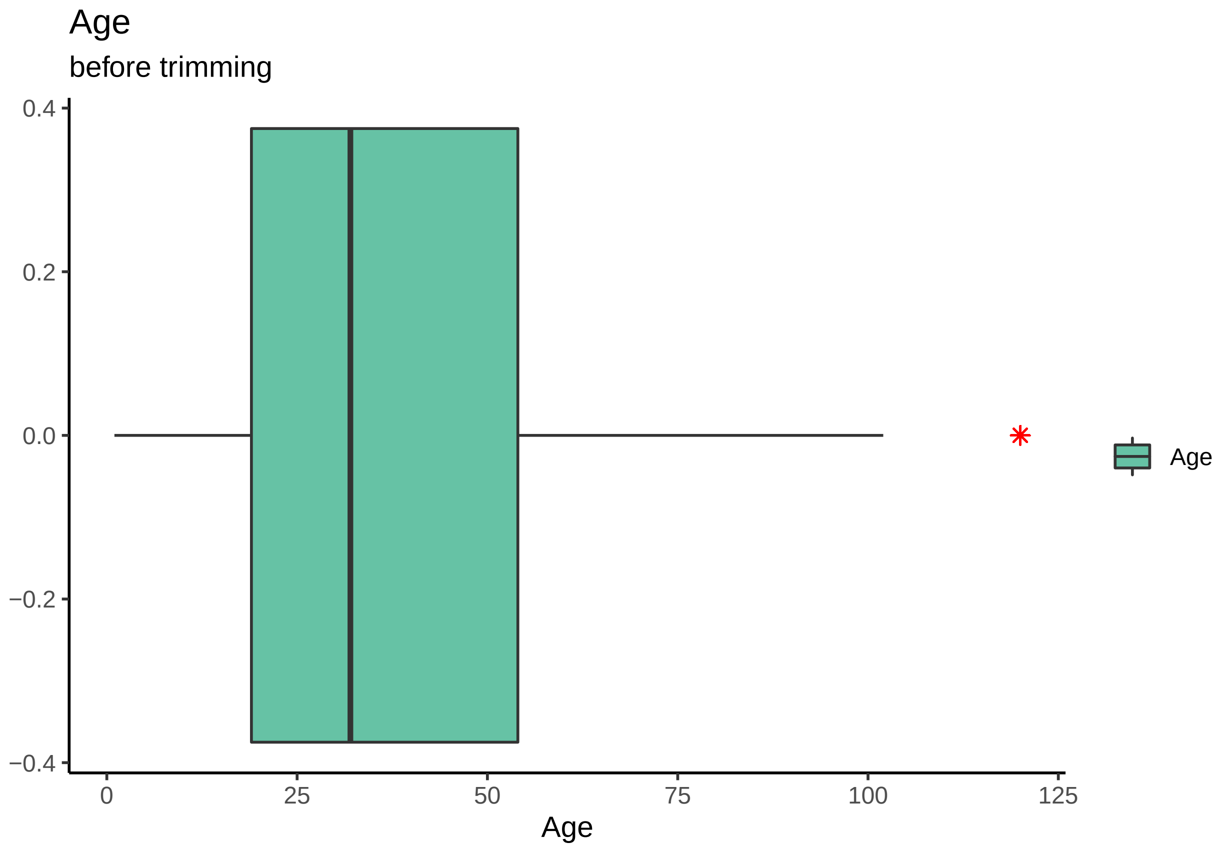
```
## [1] 47 74 76
```

```
## [1] 26 15  6
```

Summary:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   19.00   32.00   37.89   54.00  120.00
```

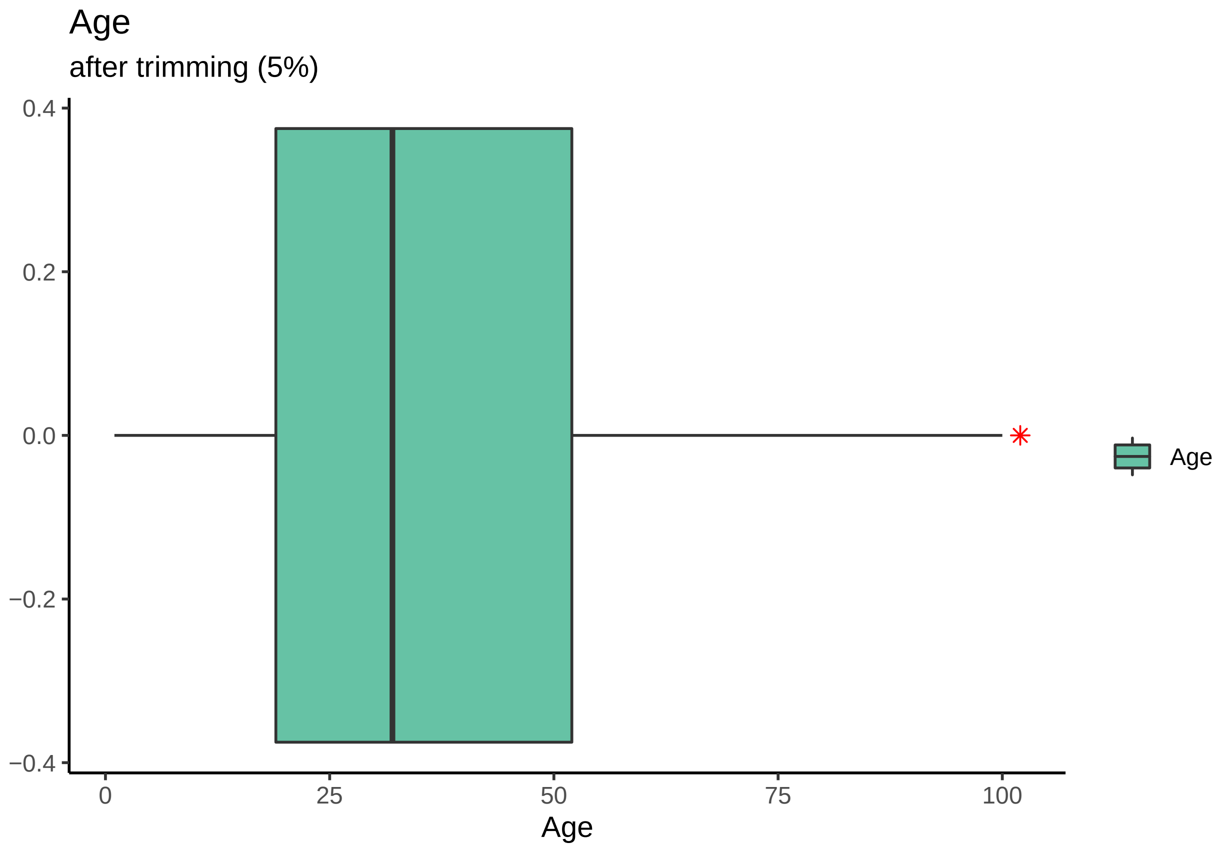
The mode of age is 13 with 676 observations. The total of observations is 24433. We can see that there is an outlier:



We see a lot of 120 values in age. These will be replaced by the 5% trimmed mean of age.

Table 4: Last observations of Frequency

age	Freq
97	2
98	4
100	3
102	1
120	516
Sum	24433



What zip presents ?

First 3 and last 3 observations:

```
## [1] "V8P5H7" "V9Z1C5" "V8L2P7"
```

```
## [1] "L7E 2P5" "L7E 0A4" "L9W 2W7"
```

We can see there is some observations with "=" for zip. If there is another row with zip information zip will be replaced. If, not, the row will be dropped:

Table 5: First observations of Frequency

zip	Freq
=	2
0K1H0	2
10024	1
1060	1
12962	1
13655	5

```r
#What rows have zip "="
data[which(data$zip=="="),] #patient_id: 10709,22681
```

```
## # A tibble: 2 x 13
## # Groups:   patient_id [2]
##   practice_id patient_id gender   age zip   primary_insurance secondary_insuran~
##   <fct>       <fct>      <fct>  <int> <chr> <chr>             <chr>
## 1 D31222      10709      Female    32 =     YES               NO
## 2 D10551      22681      Female    47 =     NO                NO
## # ... with 6 more variables: patient_referral <chr>, other_referral <chr>,
## #   FirstVisit <chr>, LastVisit <chr>, DaysLastVisit <chr>, Chrun <chr>
```

```r
data[which(data$patient_id==10709),] #just one row
```

```
## # A tibble: 1 x 13
## # Groups:   patient_id [1]
##   practice_id patient_id gender   age zip   primary_insurance secondary_insuran~
##   <fct>       <fct>      <fct>  <int> <chr> <chr>             <chr>
## 1 D31222      10709      Female    32 =     YES               NO
## # ... with 6 more variables: patient_referral <chr>, other_referral <chr>,
## #   FirstVisit <chr>, LastVisit <chr>, DaysLastVisit <chr>, Chrun <chr>
```

```r
data <- data[-which(data$patient_id==10709),] #dropping row

data[which(data$patient_id==22681),] #2 rows, second one with zip: "L5B3M1"
```

```
## # A tibble: 2 x 13
## # Groups:   patient_id [1]
##   practice_id patient_id gender   age zip    primary_insurance secondary_insura~
##   <fct>       <fct>      <fct>  <int> <chr>  <chr>             <chr>
## 1 D10551      22681      Female    47 =      NO                NO
## 2 D8061       22681      Female    31 L5B3M1 NO                NO
## # ... with 6 more variables: patient_referral <chr>, other_referral <chr>,
## #   FirstVisit <chr>, LastVisit <chr>, DaysLastVisit <chr>, Chrun <chr>
```

```r
data[which(data$zip=="="),"zip"] <- "L5B3M1" #replacing with second value
```

What primary_insurance presents ?

First 3 and last 3 observations:

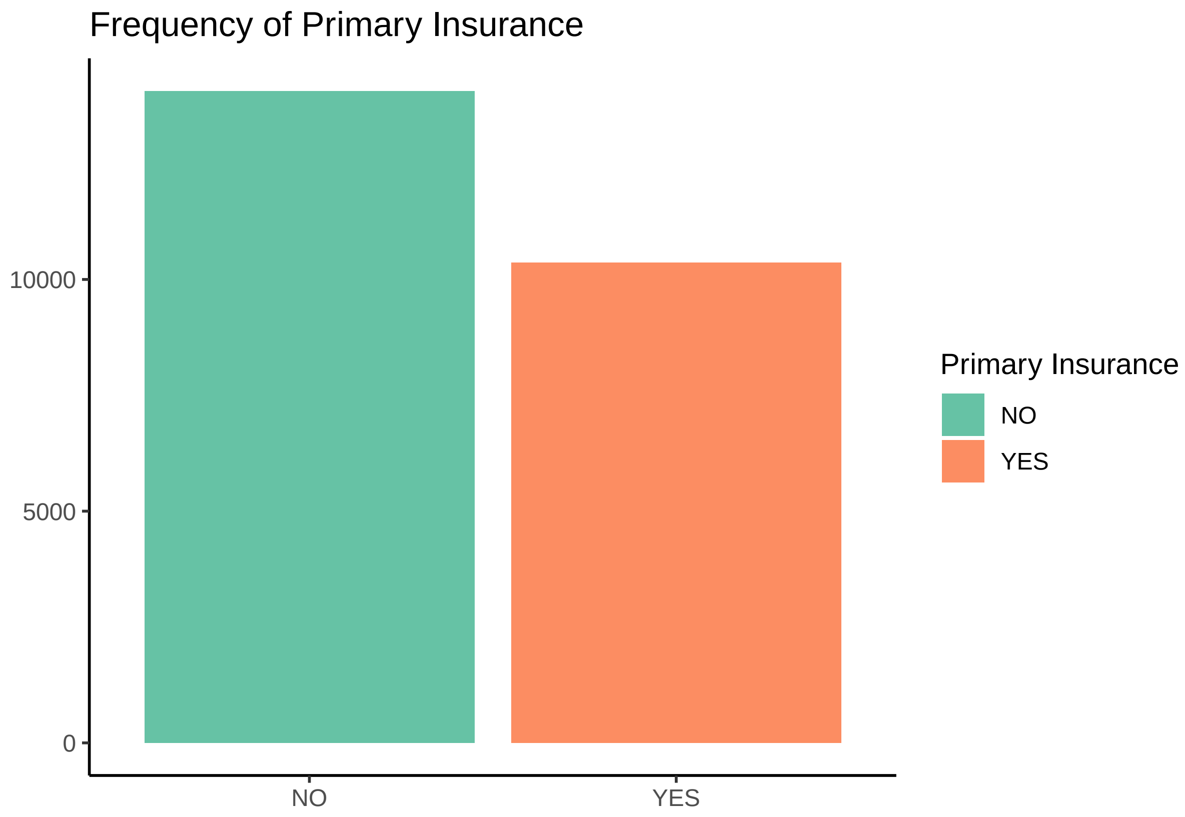
```
## [1] "YES" "YES" "YES"
```

```
## [1] "NO"  "YES" "YES"
```

We can see the majority of observations have at insurance(primary).

Table 6: Frequency

primary_insurance	Freq
NO	14068
YES	10364



What secondary_insurance presents ?

First 3 and last 3 observations:

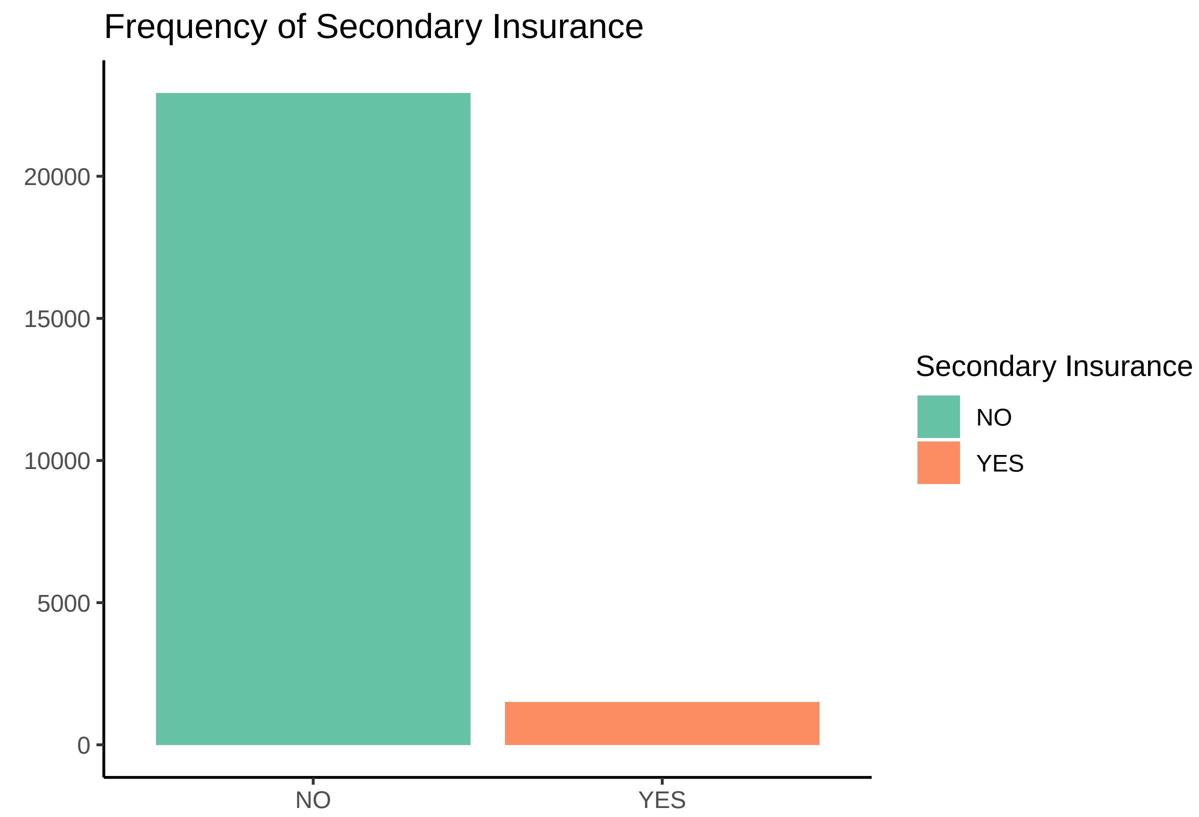
```
## [1] "YES" "NO"  "NO"
```

```
## [1] "NO"  "YES" "NO"
```

We can see the majority of observations have don't have a secondary insurance.

Table 7: Frequency

secondary_insurance	Freq
NO	22934
YES	1498



What other_referral presents ?

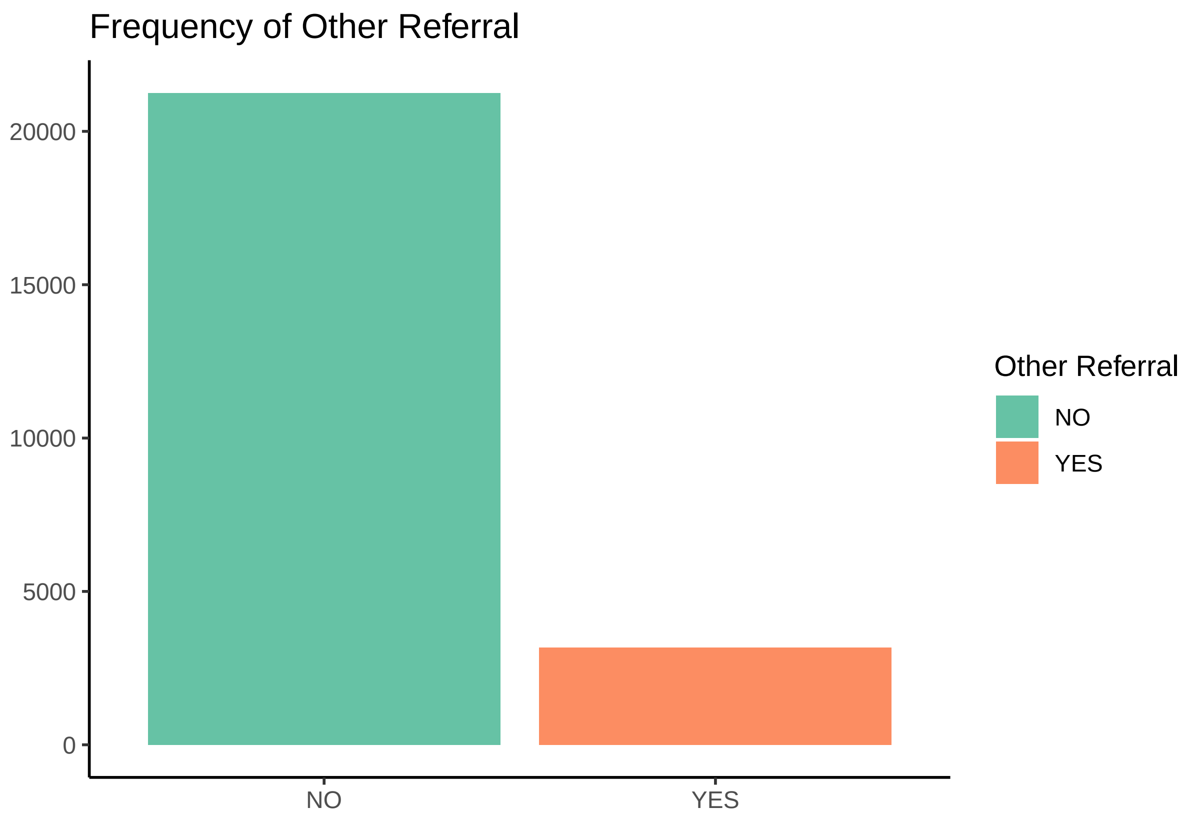
First 3 and last 3 observations:

```
## [1] "YES" "YES" "YES"
```

```
## [1] "NO" "NO" "NO"
```

Table 8: Frequency

other_referral	Freq
NO	21256
YES	3176



What FirstVisit presents ?

First 3 and last 3 observations:

```
## [1] "1/28/2016" "2/16/2016" "2/24/2016"
```

```
## [1] "1/8/2018"  "2/5/2018"  "1/17/2018"
```

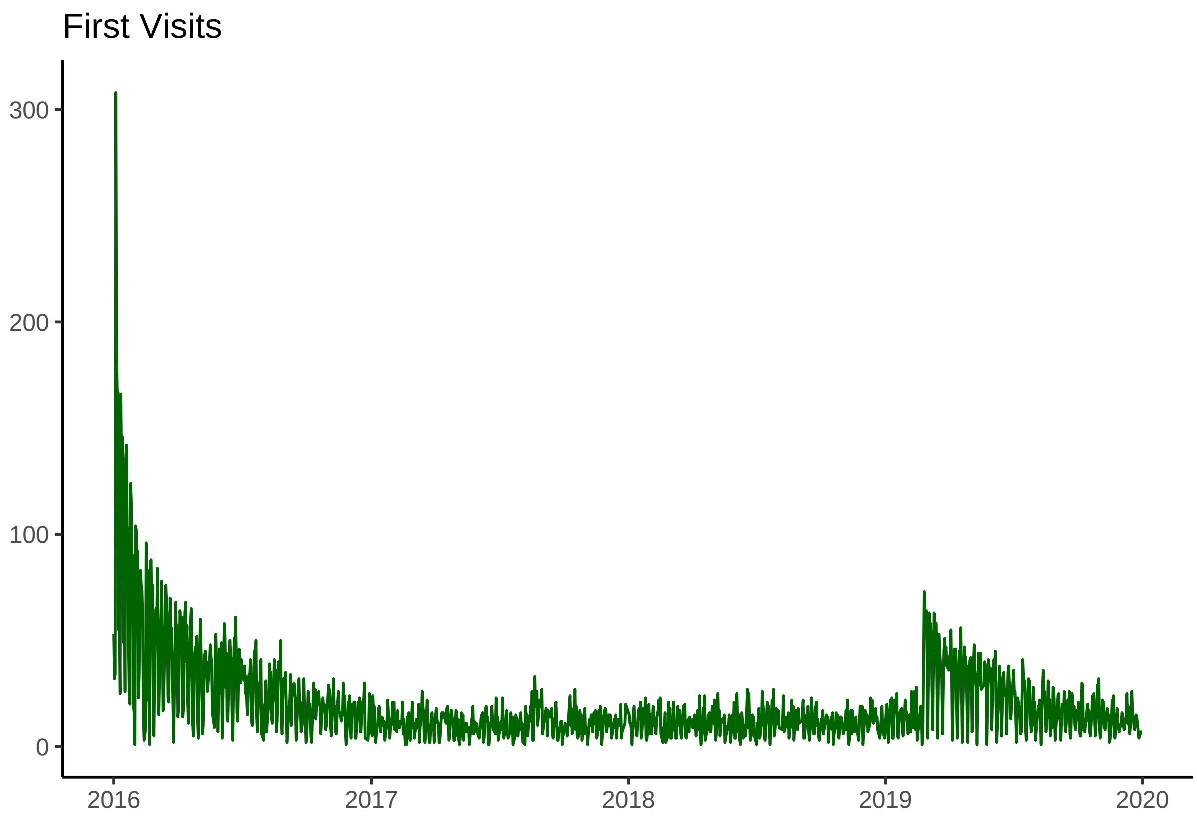
We can see a peak of First visits at the beginning of data collection.

```
## [1] "2016-01-28" "2016-02-16" "2016-02-24"
```

```
## [1] "2018-01-08" "2018-02-05" "2018-01-17"
```

Table 9: First observations of Frequency

FirstVisit	Freq
2016-01-01	53
2016-01-02	32
2016-01-03	34
2016-01-04	308
2016-01-05	186
2016-01-06	164



What LastVisit presents ?

First 3 and last 3 observations:

```
## [1] "6/29/2017"  "12/31/2019" "11/27/2019"
```

```
## [1] "3/7/2018"  "9/11/2019" "8/14/2019"
```

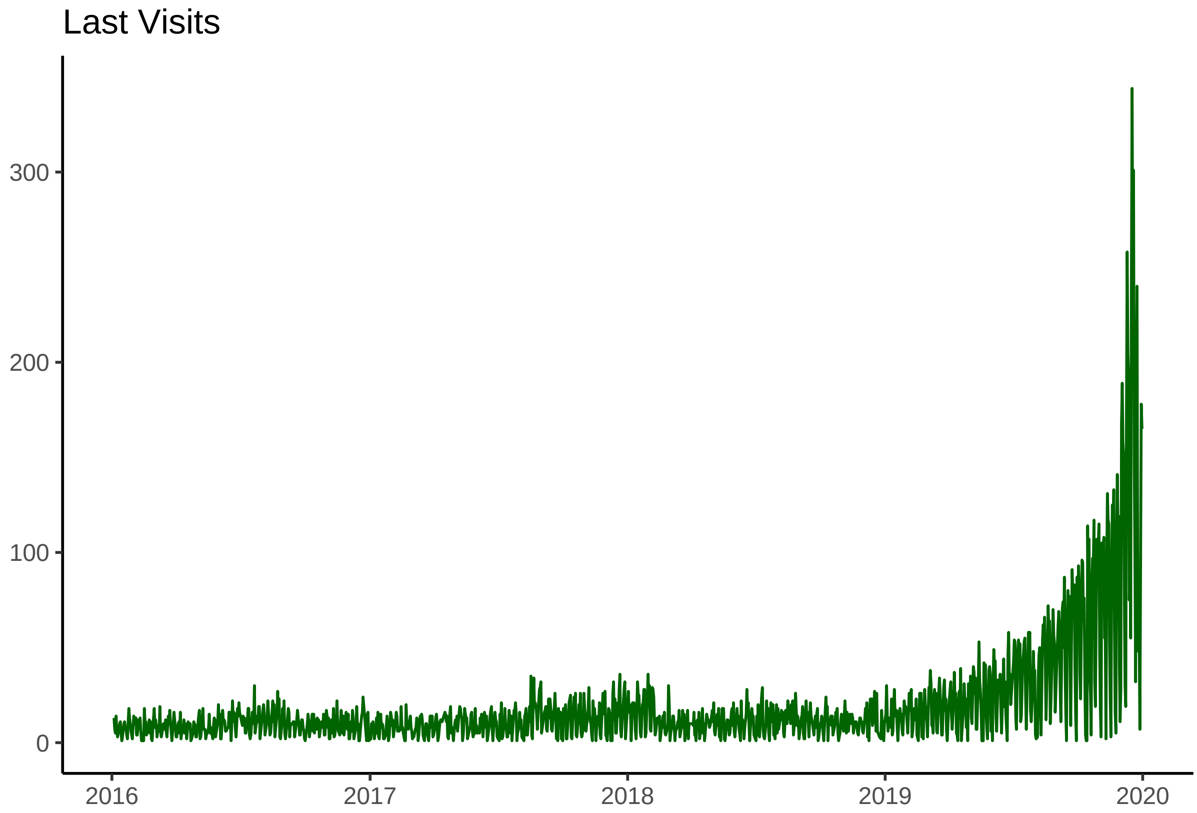
We can see a peak of Last visits at the end of data collection.

```
## [1] "2017-06-29" "2019-12-31" "2019-11-27"
```

```
## [1] "2018-03-07" "2019-09-11" "2019-08-14"
```

Table 10: First observations of Frequency

LastVisit	Freq
2016-01-04	13
2016-01-05	7
2016-01-06	5
2016-01-07	14
2016-01-08	6
2016-01-09	3



What DaysLastVisit presents ?

First 3 and last 3 observations:

```
## [1] "7/3/1902" "1/0/1900" "2/3/1900"
```

```
## [1] "664" "111" "139"
```

We can see that there are 2 formats: date and numeric. Probably due to how data was export. Since the data set posses First Visit and Last Visit dates, this feature will be replaced by a calculated one:

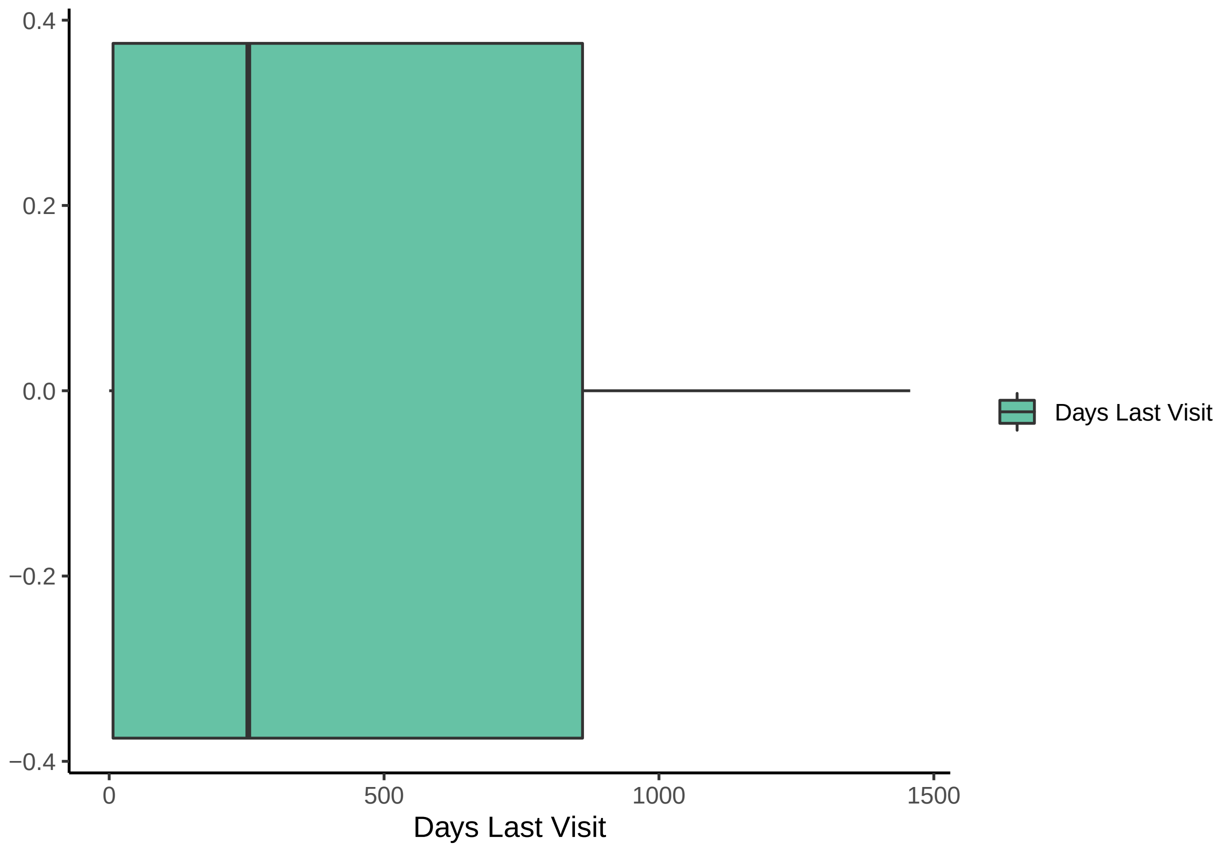
$DaysLastVisit = LastVisit - FirstVisit$

Summary:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     0.0     7.0   253.0   461.3   861.0  1457.0
```

The mode of DaysLastVisit is 0 with 5577 observations. With the total of observations being 24432.

This values correspond to 23% of observations. These observations will not be dropped before consultation with a subject matter expert.



What churn presents ?

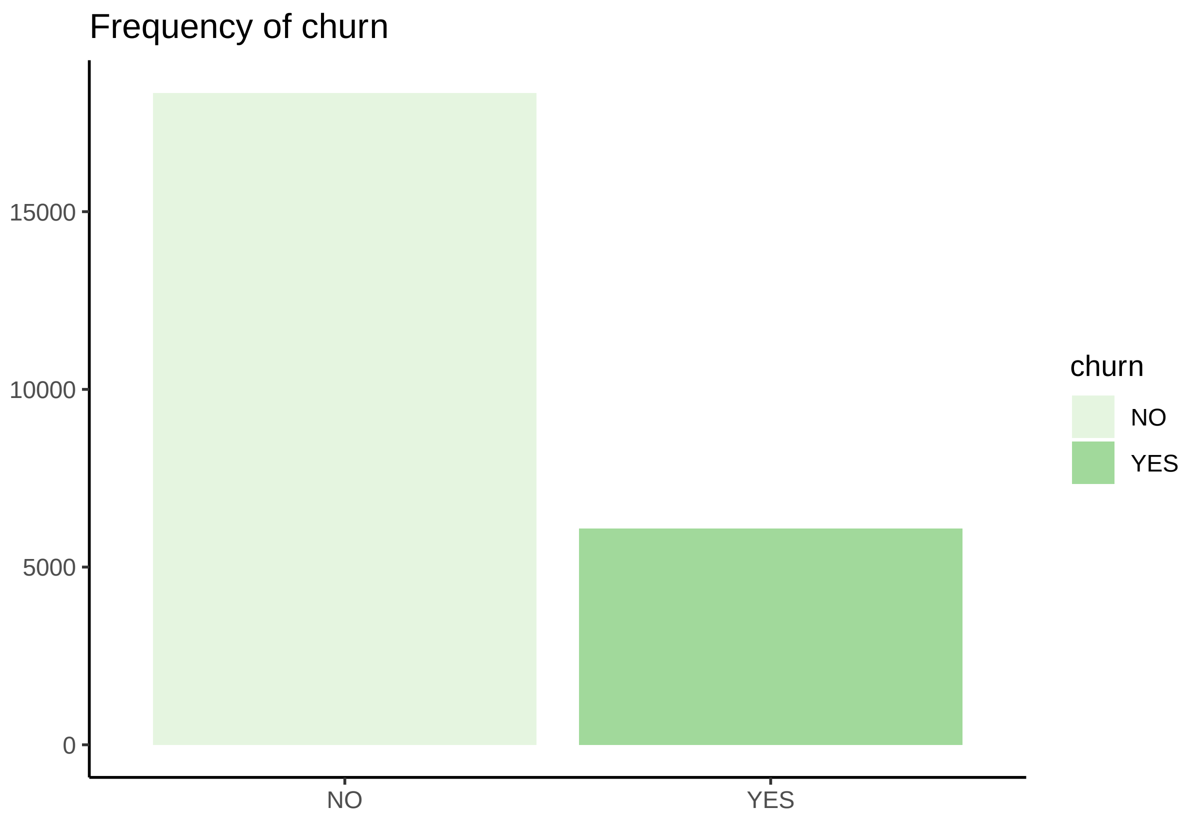
First 3 and last 3 observations:

```
## [1] "YES" "NO"  "NO"
```

```
## [1] "NO" "NO" "NO"
```

Table 11: Frequency

Var1	Freq
NO	18343
YES	6089



Segmenting by age groups

The age groups will be divided as following :

- 0-18;
- 19-44;
- 45-64;
- 65-84; and
- 85 and over.

Data source: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/Age-and-Gender

```
# There is not any age minor to 0 or that is NA
sum(is.na(data$age)|data$age<0)
```

```
## [1] 0
```

After inclusion of feature age_group:

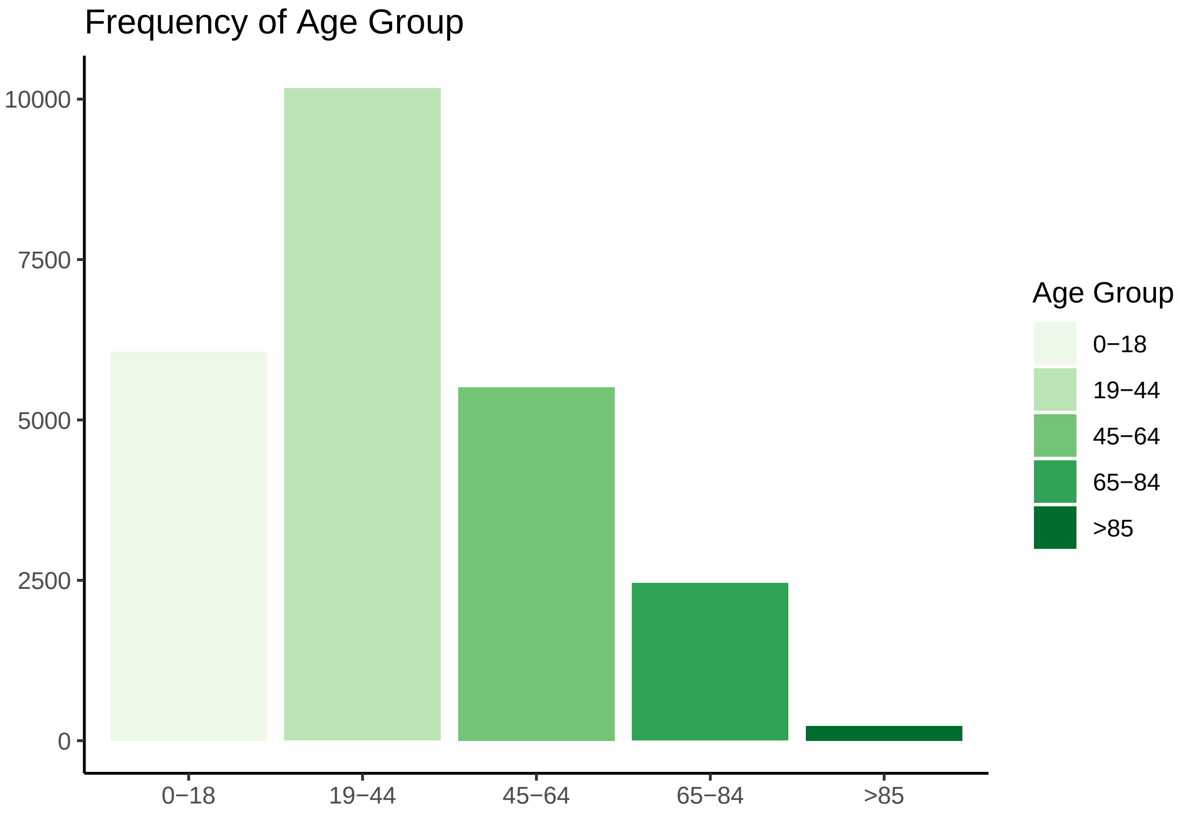
Table 12: First observations

practice_id	patient_id	gender	age	age_group	zip
D17435	806553553	Female	47	45-64	V8P5H7
D17435	806553536	Female	74	65-84	V9Z1C5
D17435	806553528	Male	76	65-84	V8L2P7
D17435	806553525	Male	57	45-64	V9B2W3
D17435	806553524	Female	53	45-64	V9B2W3
D17435	806553517	Female	74	65-84	V8S2N3

We can see that the most frequent age group is 19-44.

Table 13: Frequency

Age.Group	Freq
0-18	6064
19-44	10169
45-64	5511
65-84	2456
>85	232



Creating an ID to identifying patient and renaming patient_id to main_insurance_holder:

```
#Is there a duplicated value in the data frame?
sum(duplicated(data))
```

```
## [1] 0
```

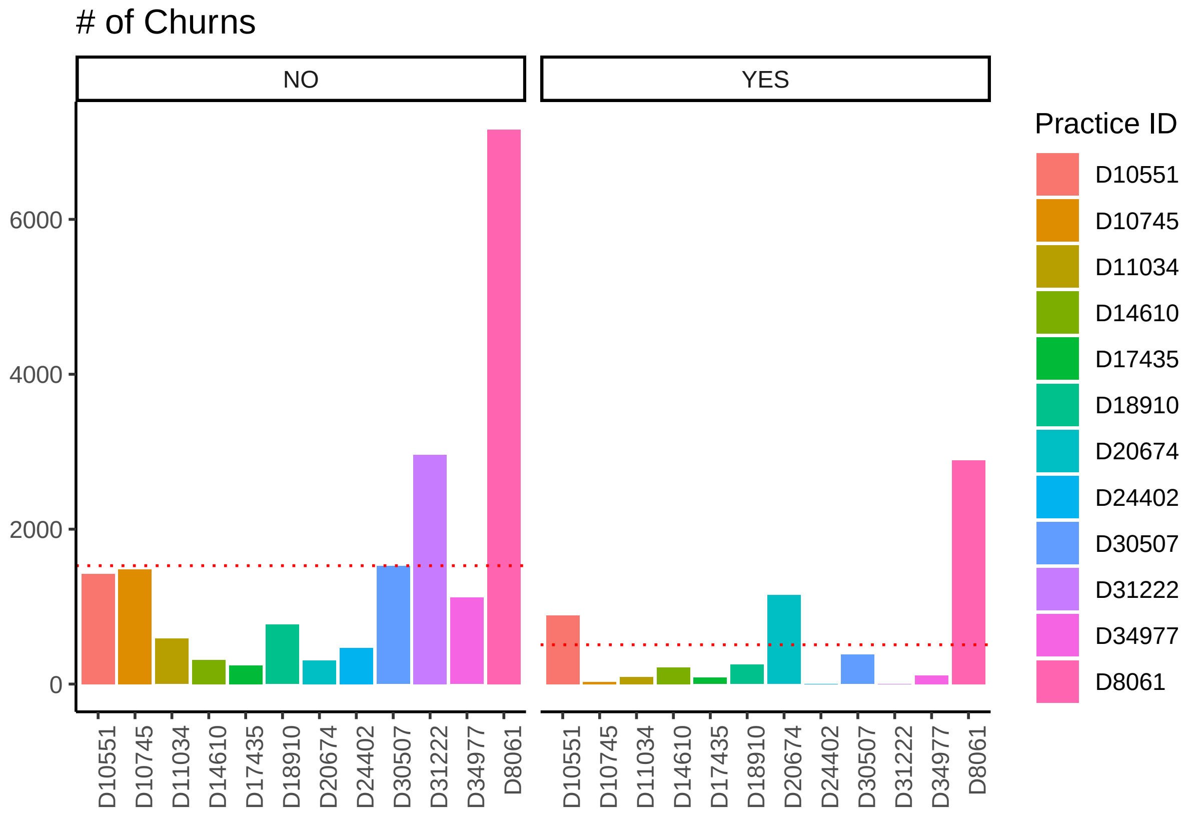
Table 14: Head of data

patient_id	practice_id	main_insurance_holder	gender	age	age_group	zip	primary_insurance
p_1	D17435	806553553	Female	47	45-64	V8P5H7	YES
p_2	D17435	806553536	Female	74	65-84	V9Z1C5	YES
p_3	D17435	806553528	Male	76	65-84	V8L2P7	YES
p_4	D17435	806553525	Male	57	45-64	V9B2W3	YES
p_5	D17435	806553524	Female	53	45-64	V9B2W3	YES
p_6	D17435	806553517	Female	74	65-84	V8S2N3	NO

Which practice have the most and least churn numbers? Is there a relationship between these features?

We can see that D8061 leads both groups.

	NO	YES
D10551	1422	888
D10745	1483	27
D11034	586	92
D14610	311	215
D17435	240	83
D18910	768	252
D20674	306	1149
D24402	466	0
D30507	1522	381
D31222	2962	3
D34977	1116	110
D8061	7161	2889



Test of Independence(Chi-Square) - Practice vs Churn (0.05 significance level)

```
##
##  Pearson's Chi-squared test
##
## data:  t
## X-squared = 4448, df = 11, p-value < 2.2e-16
```

Assumptions:

1. N, the total frequency, should be reasonably large, say greater than 50;
2. The sample observations should be independent. No individual item should be included twice or more in the sample;
3. No expected frequencies should be small. Preferably each expected frequency should be larger than 10 but in any case not less than 5.

Null hypothesis: Practice is independent of the Churn

If condition of chi-square are satisfied and p-value is less than significant level (5%) reject null hypothesis: There is a relation ship between them at 5% significant level.

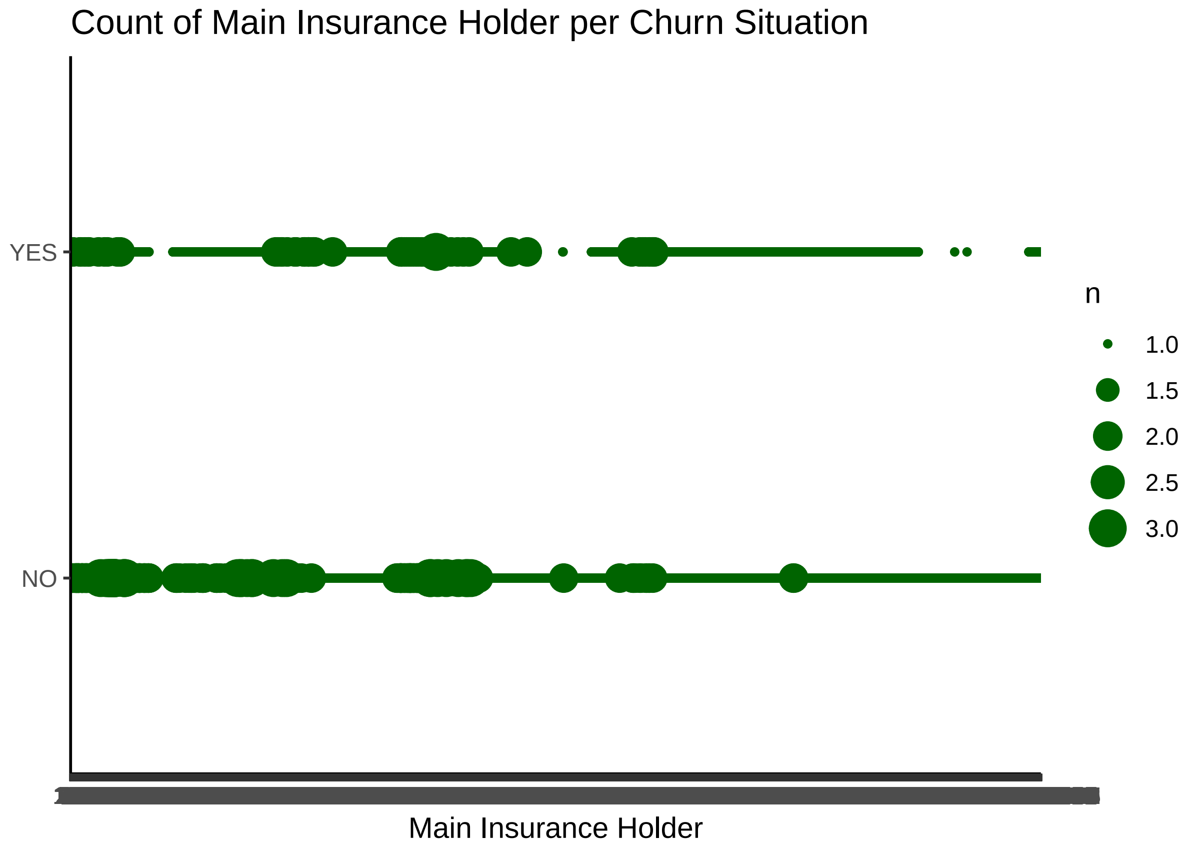
We can see that There is a relationship between Practice and Churn, since pvalue is less than 0.

Is there a relationship between main_insurance_holder and churn features?

We can see that appears that there are main insurance holder presents in both churn and non-churn situation:

Table 16: 2-way table(First observations)

	NO	YES
3	1	0
5	1	1
6	0	1
7	1	1
8	0	1
12	0	1



Test of Independence(Chi-Square) - Main Insurance Holder vs Churn (0.05 significance level)

```
## Warning in chisq.test(t): Chi-squared approximation may be incorrect

##
##  Pearson's Chi-squared test
##
## data:  t
## X-squared = NaN, df = 22738, p-value = NA
```

Assumptions:

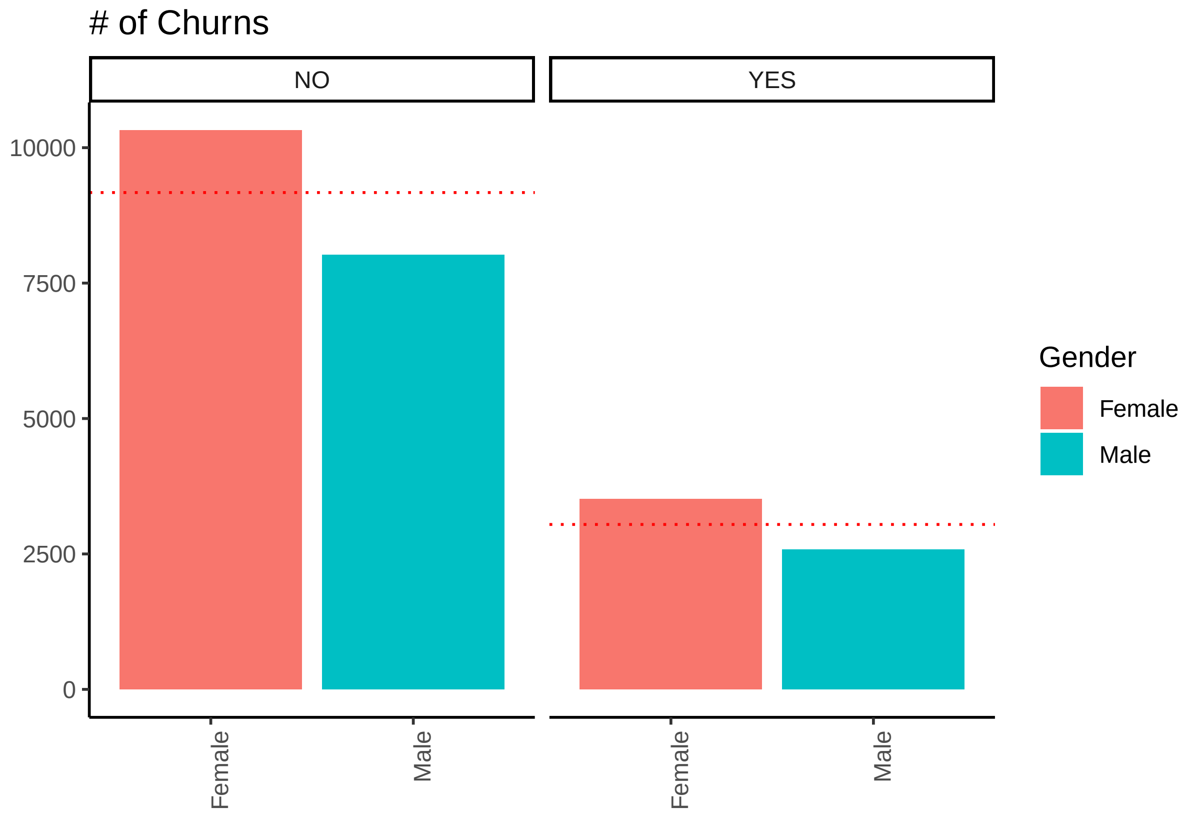
1. N, the total frequency, should be reasonably large, say greater than 50;
2. The sample observations should be independent. No individual item should be included twice or more in the sample;
3. No expected frequencies should be small. Preferably each expected frequency should be larger than 10 but in any case not less than 5.

As seen before the maximum number of observations per Main Insurance Holder is 3 and most of the are at the minimum of 1, so the assumptions are not satisfied.

Which gender have the most and least churn numbers? Is there a relationship between these features?

We can see that females leads both categories.

	NO	YES
Female	10319	3512
Male	8024	2577



Test of Independence(Chi-Square) - Gender vs Churn (0.05 significance level)

To satisfy assumptions and perform chi-square test, observations of unknows will be dropped.

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  t
## X-squared = 3.7056, df = 1, p-value = 0.05423
```

Assumptions:

1. N, the total frequency, should be reasonably large, say greater than 50;
2. The sample observations should be independent. No individual item should be included twice or more in the sample;
3. No expected frequencies should be small. Preferably each expected frequency should be larger than 10 but in any case not less than 5.

   Null hypothesis: Gender is independent of the Churn

   If condition of chi-square are satisfied and p-value is less than significant level (5%) reject null hypothesis: There is a relation ship between them at 5% significant level.

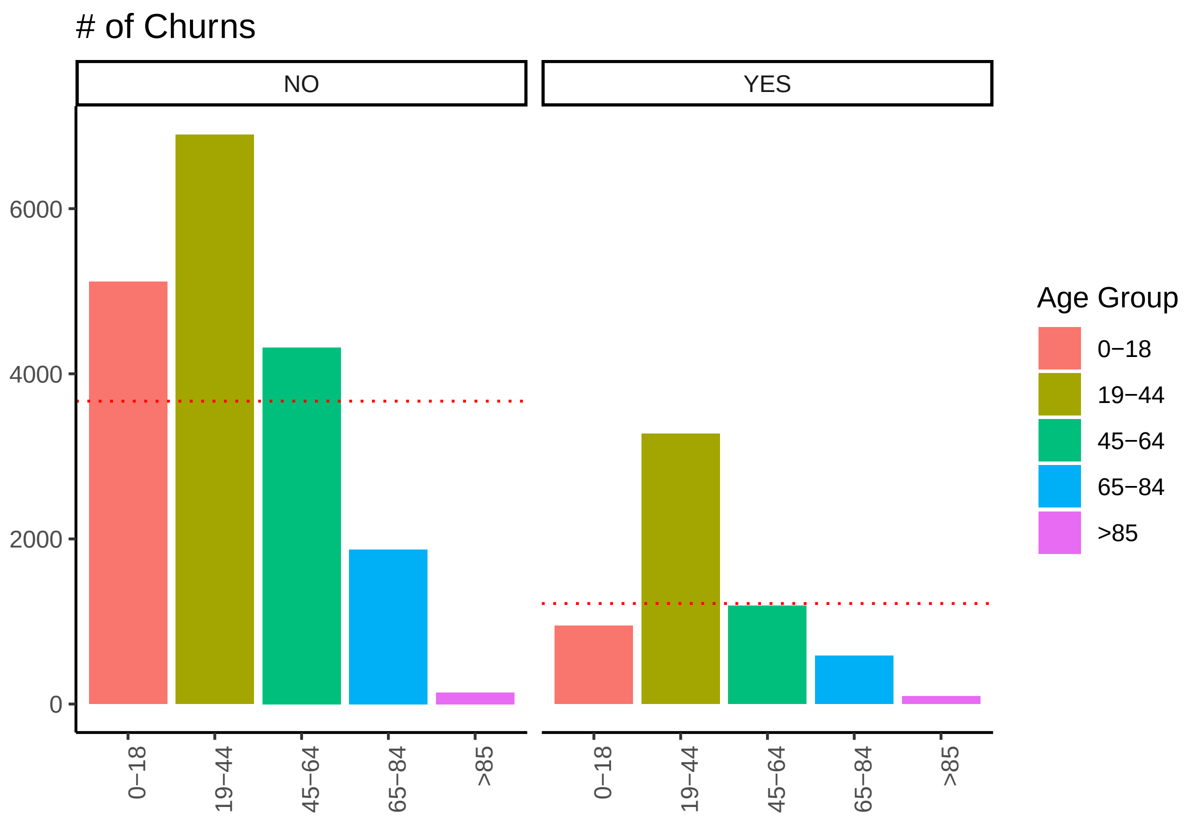
We can see that There is NOT a relationship between Gender and Churn, since pvalue is 0.0542298621449605.

Which age group have the most and least churn numbers? Is there a relationship between these features?

We can see that 19-44 age group leads in both categories.

	NO	YES
0-18	5115	949
19-44	6897	3272
45-64	4319	1192
65-84	1872	584
>85	140	92

```
## [1] "NO"  "YES"
```



Test of Independence(Chi-Square) - Age Group vs Churn (0.05 significance level)

```
##
##  Pearson's Chi-squared test
##
## data:  t
## X-squared = 625.19, df = 4, p-value < 2.2e-16
```

Assumptions:

1. N, the total frequency, should be reasonably large, say greater than 50;
2. The sample observations should be independent. No individual item should be included twice or more in the sample;
3. No expected frequencies should be small. Preferably each expected frequency should be larger than 10 but in any case not less than 5.

   Null hypothesis: Age Group is independent of the Churn

   If condition of chi-square are satisfied and p-value is less than significant level (5%) reject null hypothesis: There is a relation ship between them at 5% significant level.

We can see that There is a relationship between Practice and Churn, since pvalue is 5.47096905941567e-134.

Which zip(location) have the most and least churn numbers? Is there a relationship between these features?

Table 19: First observations of 2-way table

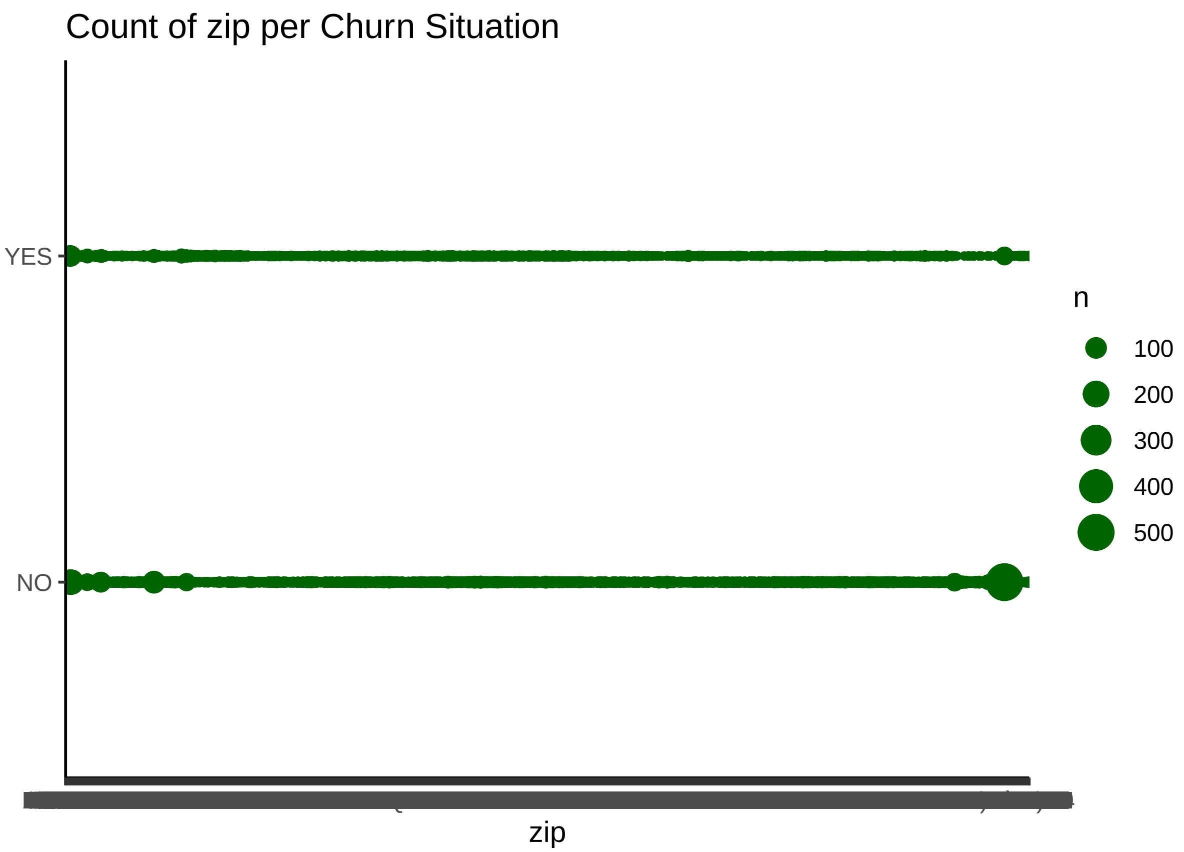
	NO	YES
0K1H0	2	0
10024	1	0
1060	1	0
12962	1	0
13655	3	2
14001	1	0
14094	3	0
14132	1	0
14150	1	0
14209	1	0

Table 20: 5 zip with least churn numbers

zip	Churn	Freq
V0J2N0	NO	533
B0K1H0	NO	169
B0K2A0	NO	169
B0K1S0	NO	157
K6H5R7	NO	118

Table 21: 5 zip with most churn numbers

zip	Churn	Freq
B0K1H0	YES	101
B0K2A0	YES	83
B0K1S0	YES	75
B0K1X0	YES	75
V0J2N0	YES	59



Test of Independence(Chi-Square) - zip vs Churn (0.05 significance level)

Assumptions:

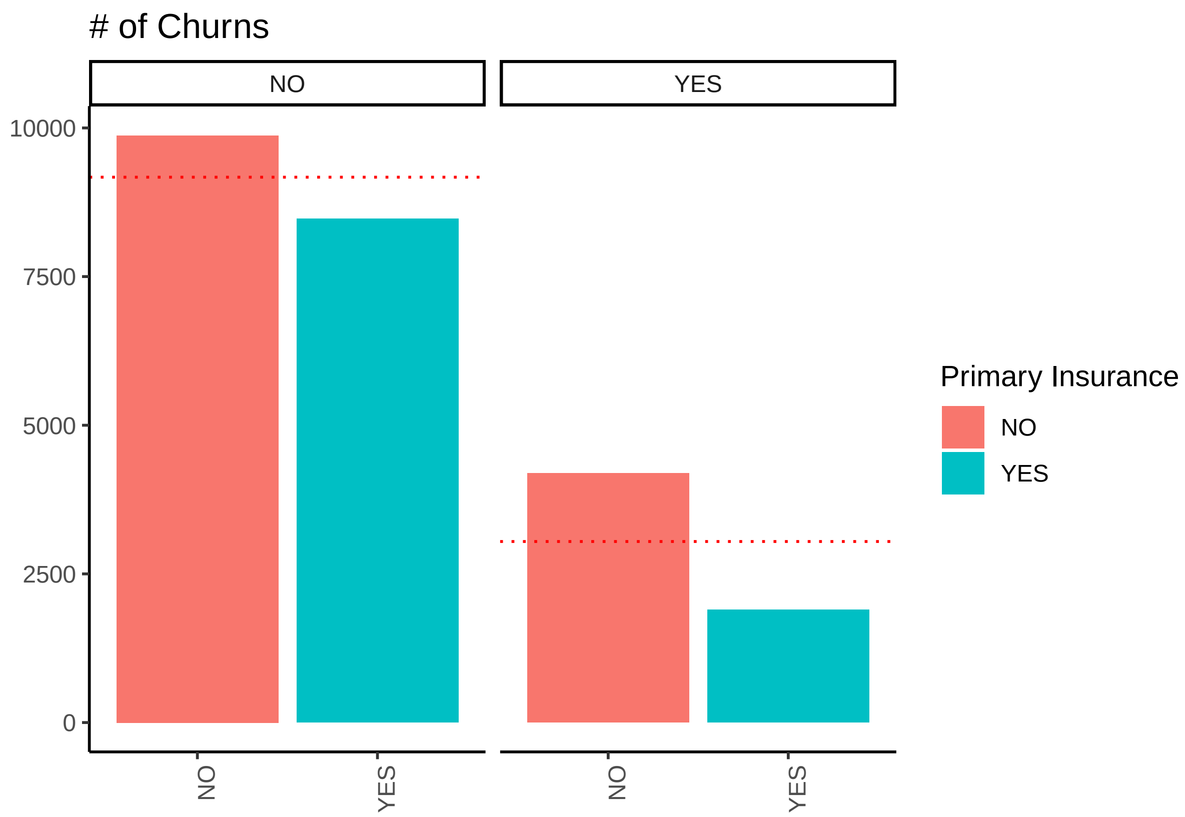
1. N, the total frequency, should be reasonably large, say greater than 50;
2. The sample observations should be independent. No individual item should be included twice or more in the sample;
3. No expected frequencies should be small. Preferably each expected frequency should be larger than 10 but in any case not less than 5.

We can see that there are many observations with frequency of value 0, not satisfying chi-square assumptions.

How having insurance impact on churn numbers? Is there a relationship between these features?

We can see that not having insurance leads both groups.

	NO	YES
NO	9873	4195
YES	8470	1894



Test of Independence(Chi-Square) - Primary Insurance vs Churn (0.05 significance level)

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  t
## X-squared = 424.46, df = 1, p-value < 2.2e-16
```

Assumptions:

1. N, the total frequency, should be reasonably large, say greater than 50;
2. The sample observations should be independent. No individual item should be included twice or more in the sample;
3. No expected frequencies should be small. Preferably each expected frequency should be larger than 10 but in any case not less than 5.

   Null hypothesis: Primary Insurance is independent of the Churn

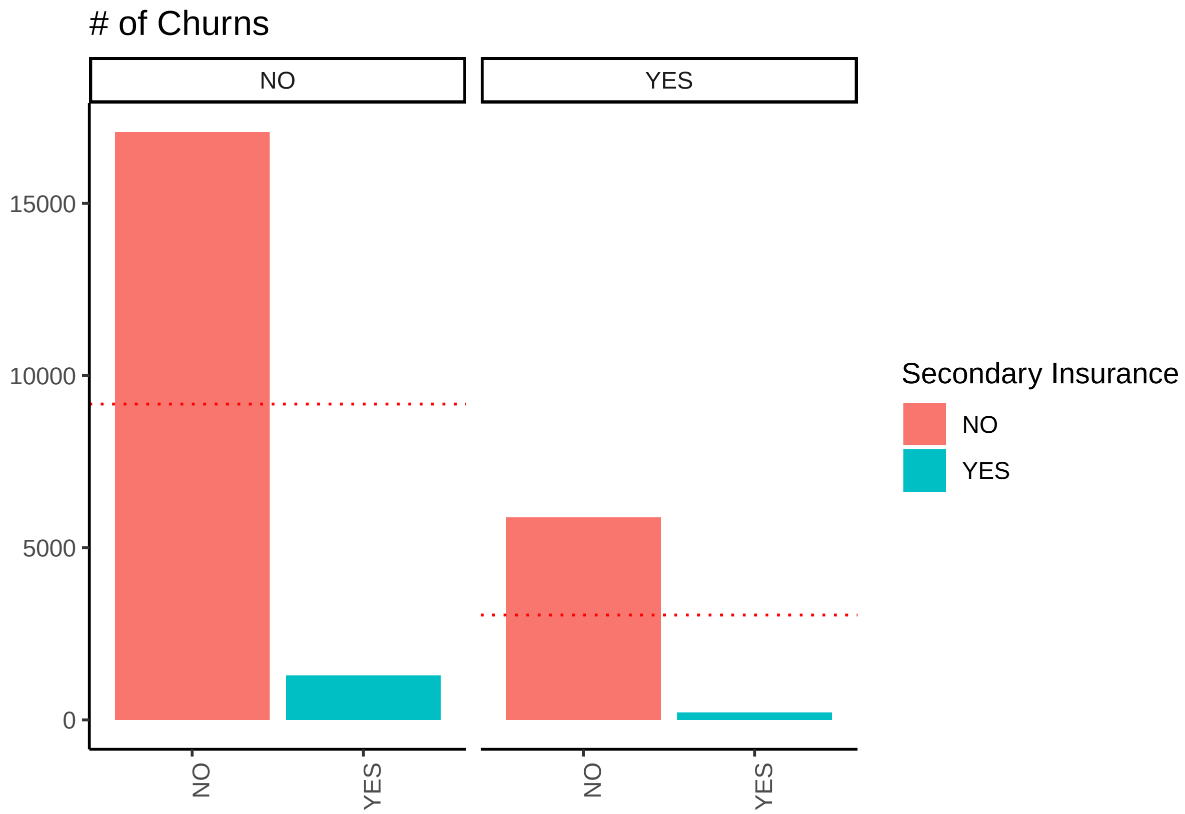
   If condition of chi-square are satisfied and p-value is less than significant level (5%) reject null hypothesis: There is a relation ship between them at 5% significant level.

We can see that **There is a relationship between Primary Insurance and Churn, since pvalue is 2.61452143266085e-94.**

How having a secondary insurance impact on churn numbers? Is there a relationship between these features?

We can see that not having a secondary insurance leads in both categories.

	NO	YES
NO	17057	5877
YES	1286	212



Test of Independence(Chi-Square) - Secondary Insurance vs Churn (0.05 significance level)

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  t
## X-squared = 98.317, df = 1, p-value < 2.2e-16
```

Assumptions:

1. N, the total frequency, should be reasonably large, say greater than 50;
2. The sample observations should be independent. No individual item should be included twice or more in the sample;
3. No expected frequencies should be small. Preferably each expected frequency should be larger than 10 but in any case not less than 5.

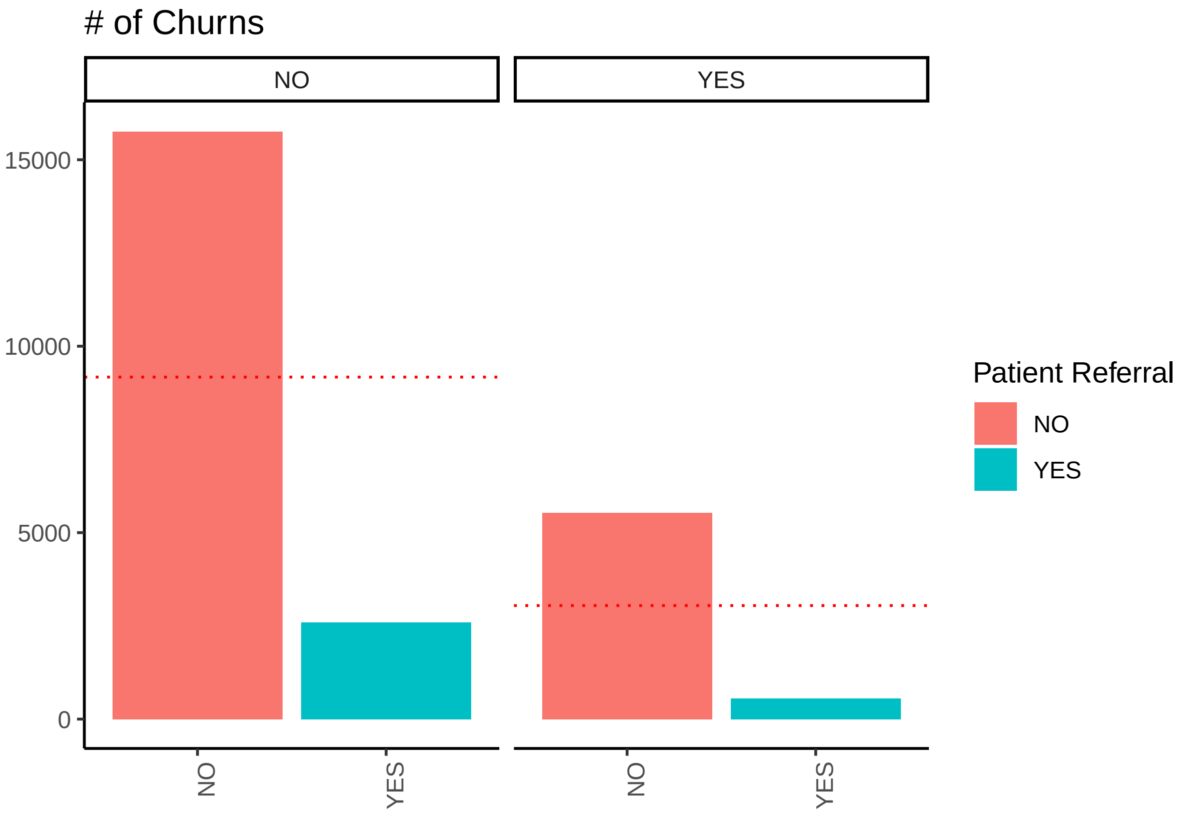
> **Null hypothesis: Secondary Insurance is independent of the Churn**
>
> If condition of chi-square are satisfied and p-value is less than significant level (5%) reject null hypothesis: There is a relation ship between them at 5% significant level.

We can see that **There is a relationship between Secondary Insurance and Churn, since pvalue is 3.56415392187723e-23.**

How having been referred impact on churn numbers? Is there a relationship between these features?

We can see that not being referred leads in both categories.

	NO	YES
NO	15753	5531
YES	2590	558



Test of Independence(Chi-Square) - Patient Referral vs Churn (0.05 significance level)

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
```

```
## data:  t
## X-squared = 99.584, df = 1, p-value < 2.2e-16
```

Assumptions:

1. N, the total frequency, should be reasonably large, say greater than 50;
2. The sample observations should be independent. No individual item should be included twice or more in the sample;
3. No expected frequencies should be small. Preferably each expected frequency should be larger than 10 but in any case not less than 5.

   Null hypothesis: Patient Referral is independent of the Churn

   If condition of chi-square are satisfied and p-value is less than significant level (5%) reject null hypothesis: There is a relation ship between them at 5% significant level.

We can see that **There is a relationship between Patient Referral and Churn, since pvalue is 1.88028195323435e-23.**

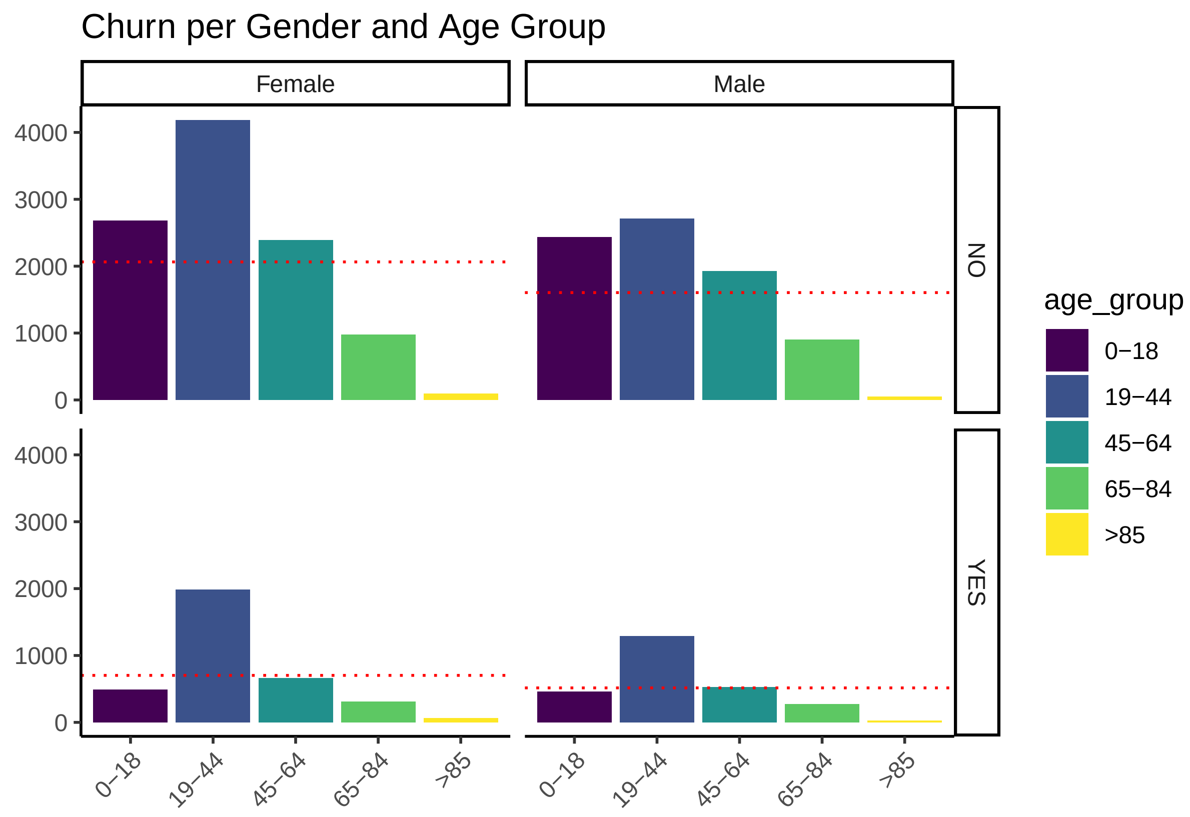
How Gender and Age Groups impacts on churn numbers? Is there a relationship between these features?

```
## ‘summarise()‘ has grouped output by ’gender’, ’age_group’. You can override using the ‘.groups‘ argu
```

```
## ‘summarise()‘ has grouped output by ’churn’. You can override using the ‘.groups‘ argument.
```

Table 25: Churn per Gender and Age Group

gender	age_group	churn	Freq
Female	0-18	NO	2681
Female	0-18	YES	490
Female	19-44	NO	4184
Female	19-44	YES	1984
Female	45-64	NO	2391
Female	45-64	YES	663
Female	65-84	NO	972
Female	65-84	YES	313
Female	>85	NO	91
Female	>85	YES	62
Male	0-18	NO	2434
Male	0-18	YES	459
Male	19-44	NO	2713
Male	19-44	YES	1288
Male	45-64	NO	1928
Male	45-64	YES	529
Male	65-84	NO	900
Male	65-84	YES	271
Male	>85	NO	49
Male	>85	YES	30



```
## mapping: yintercept = ~mean(Freq)
## geom_hline: na.rm = FALSE
## stat_identity: na.rm = FALSE
## position_identity
```

Test of Independence(Asymptotic General Independence Test) - Gender and Age Group vs Churn (0.05 significance level)

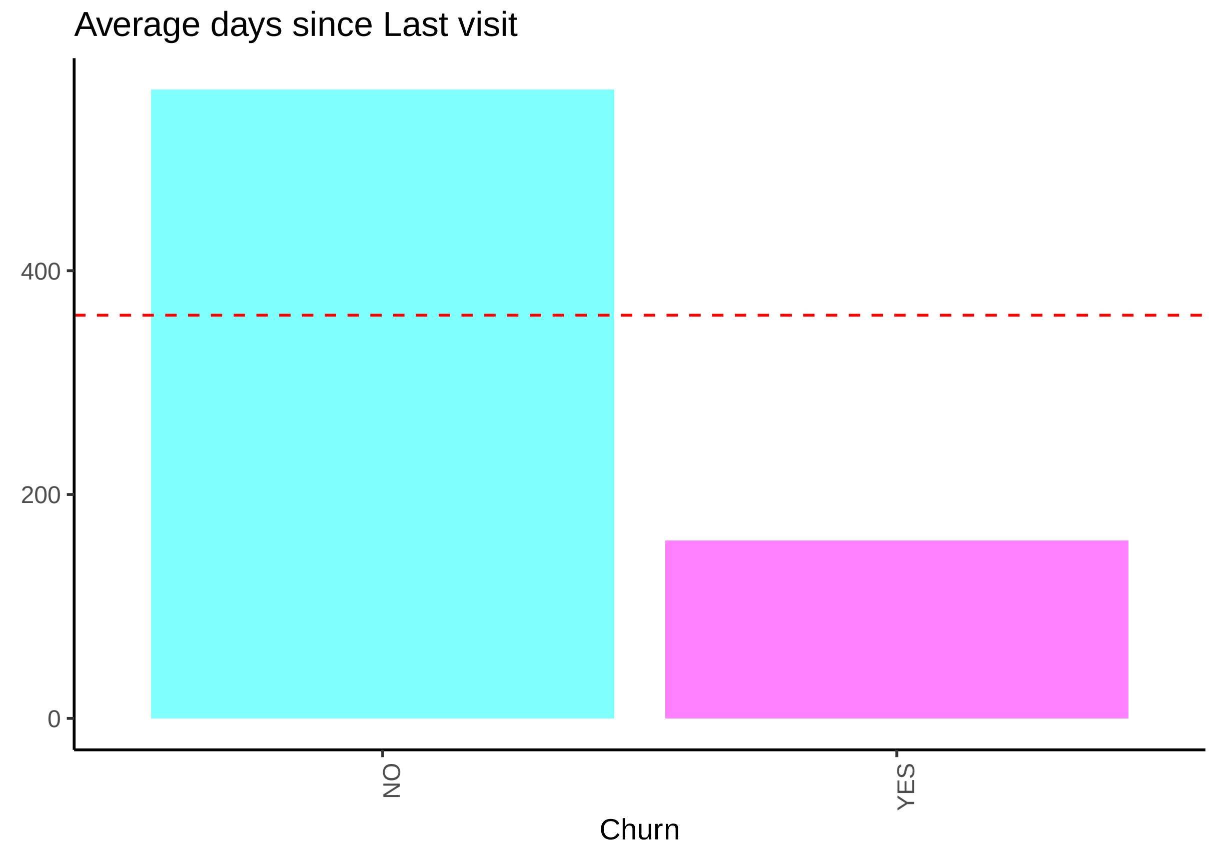
We can see that Churn is dependent of Gender and Age Group.

```
##
##  Asymptotic General Independence Test
##
## data:  churn by
##   age_group (0-18 < 19-44 < 45-64 < 65-84 < >85)
##   stratified by gender
## Z = -6.5867, p-value = 4.497e-11
## alternative hypothesis: two.sided
```

How much time since last visit impacts on churn numbers? Is there a relationship between these features?

Table 26: Mean of Days Since Last Visit

Churn	Days Last Visit
NO	561.7486
YES	158.6408



Test of Independence(Chi-Square) - Patient Referral vs Churn (0.05 significance level)

```
##
##  Welch Two Sample t-test
##
## data:  DaysLastVisit by churn
## t = 86.896, df = 23089, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group NO and group YES is not equal to 0
## 95 percent confidence interval:
##  394.0152 412.2004
## sample estimates:
##  mean in group NO mean in group YES
##          561.7486          158.6408
```

Assumptions:

1. The first assumption made regarding t-tests concerns the scale of measurement. The assumption for a t-test is that the scale of measurement applied to the data collected follows a continuous or ordinal scale, such as the scores for an IQ test;

2. The second assumption made is that of a simple random sample, that the data is collected from a representative, randomly selected portion of the total population;

3. The third assumption is the data, when plotted, results in a normal distribution, bell-shaped distribution curve. When a normal distribution is assumed, one can specify a level of probability (alpha level, level of significance, p) as a criterion for acceptance. In most cases, a 5% value can be assumed;

4. The fourth assumption is a reasonably large sample size is used. A larger sample size means the distribution of results should approach a normal bell-shaped curve;

5. The final assumption is homogeneity of variance. Homogeneous, or equal, variance exists when the standard deviations of samples are approximately equal.

   If condition of t-test are satisfied and p-value is less than significant level (5%) reject null hypothesis: true difference in means between group NO and group YES is equal to 0

We can see that the **True difference in means between group NO and group YES IS EQUAL to 0, since pvalue is virtually 0.**

Data source: https://www.investopedia.com/ask/answers/073115/what-assumptions-are-made-when-conducting-ttest.asp

Is there any association between Days Last Visit, and Age?

Test of Independence - (0.05 significance level)

Assumptions:

1.level of measurement[1];

2.related pairs[2];

3.absence of outliers[3]; and

4.linearity[4].

Source: https://www.statisticssolutions.com/pearson-correlation-assumptions/

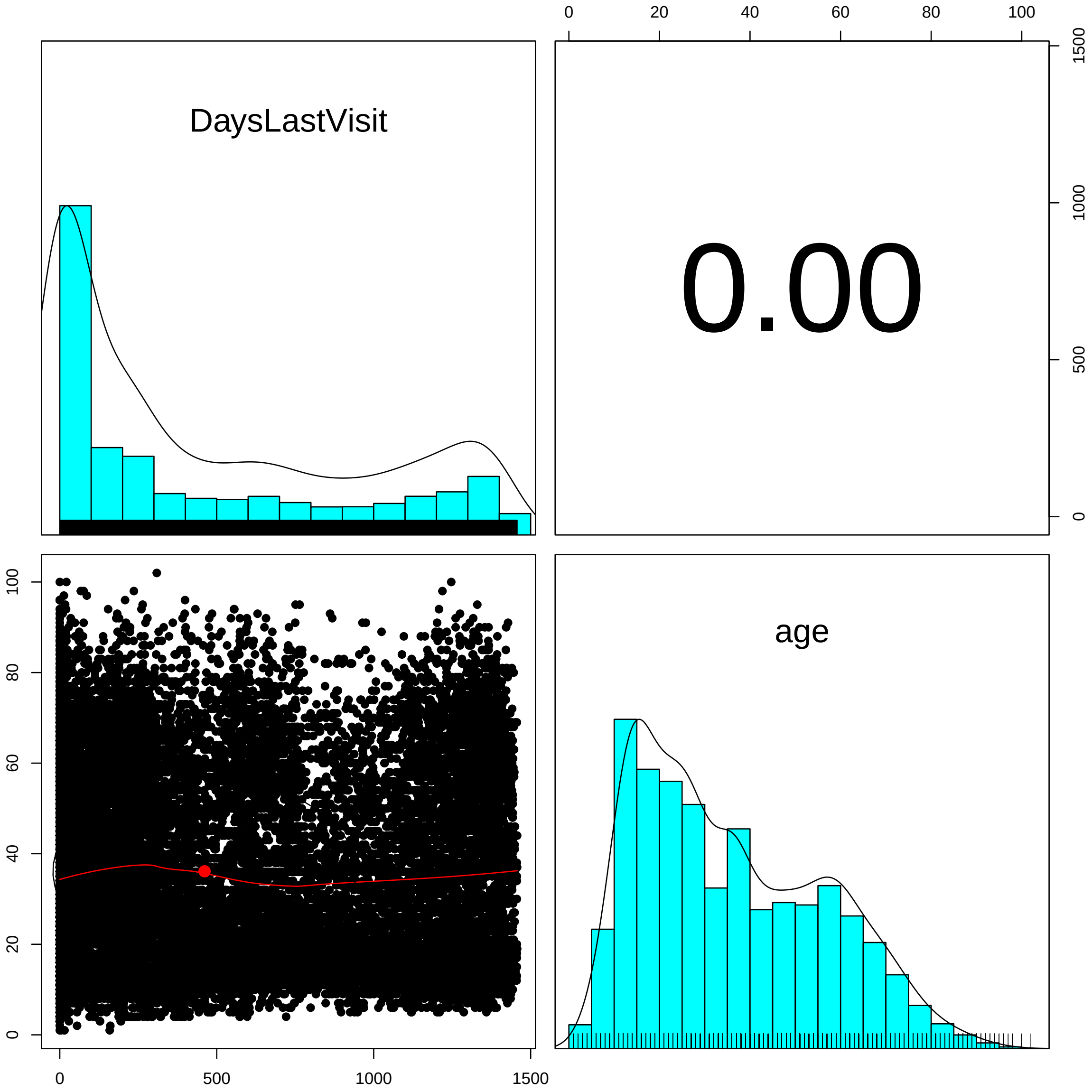
We can see that due to assumptions not met, results may not be reliable.

[1] Level of measurement refers to each variable. For a Pearson correlation, each variable should be continuous. If one or both of the variables are ordinal in measurement, then a Spearman correlation could be conducted instead.

[2] Related pairs refers to the pairs of variables. Each participant or observation should have a pair of values. So if the correlation was between weight and height, then each observation used should have both a weight and a height value.

[3] Absence of outliers refers to not having outliers in either variable. Having an outlier can skew the results of the correlation by pulling the line of best fit formed by the correlation too far in one direction or another. Typically, an outlier is defined as a value that is 3.29 standard deviations from the mean, or a standardized value of less than ±3.29.

[4] Linearity refers to the shape of the values formed by the scatterplot. For linearity, a “straight line” relationship between the variable should be formed. If a line were to be drawn between all the dots going from left to right, the line should be straight and not curved.



Interpretation:

r < 0.25 No relationship

0.25 < r < 0.5 Weak relationship

0.5 < r < 0.75 Moderate relationship

r > 0.75 Strong relationship

Results: there's no linear correlation between Age and Days Last Visit.

Conclusion

1. Almost all features have relationship with the target.

Recommendations

1. Perform targeted marketing related to each of the the feature, e.g: age_group, gender.