Data Cleaning

Helitha Dharmadasa - z5451805

2024-02-23

```r
library(tidyverse)
```

Read Data

```r
superlife_df <- read_csv("../Data/Processed Data/CLEANED_2024-srcsc-superlife-inforce-dataset.csv")
```

```
## Rows: 978582 Columns: 18
## -- Column specification ----------------------------------------------------------
## Delimiter: ","
## chr (9): Policy.number, Policy.type, Sex, Smoker.Status, Underwriting.Class,...
## dbl (9): Issue.year, Issue.age, Face.amount, Region, Death.indicator, Year.o...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```r
head(superlife_df)
```

```
## # A tibble: 6 x 18
##   Policy.number Policy.type Issue.year Issue.age Sex   Face.amount Smoker.Status
##   <chr>         <chr>            <dbl>     <dbl> <chr>       <dbl> <chr>
## 1 08FN60R4KXIS  T20               2001        54 F          100000 NS
## 2 K0JK2XD81ZNI  SPWL              2001        54 M         1000000 NS
## 3 AH3A98MHT08H  T20               2001        27 F           50000 NS
## 4 C9QPJMIH8H9Y  T20               2001        55 F         2000000 NS
## 5 2C1HL2XQ0WME  T20               2001        39 F          250000 NS
## 6 LKW7MA7BPAV1  SPWL              2001        41 M         2000000 NS
## # i 11 more variables: Underwriting.Class <chr>, Urban.vs.Rural <chr>,
## #   Region <dbl>, Distribution.Channel <chr>, Death.indicator <dbl>,
## #   Year.of.Death <dbl>, Lapse.Indicator <dbl>, Year.of.Lapse <dbl>,
## #   Cause.of.Death <chr>, Age.at.Death <dbl>, Cause.of.Death.Description <chr>
```

```r
summary(superlife_df)
```

```
##  Policy.number       Policy.type          Issue.year     Issue.age
##  Length:978582      Length:978582      Min.   :2001   Min.   :26.0
##  Class :character   Class :character   1st Qu.:2009   1st Qu.:36.0
##  Mode  :character   Mode  :character   Median :2015   Median :44.0
##                                         Mean   :2014   Mean   :44.1  
##                                         3rd Qu.:2020   3rd Qu.:52.0  
##                                         Max.   :2023   Max.   :65.0  
##                                                                      
##      Sex             Face.amount      Smoker.Status      Underwriting.Class
##  Length:978582      Min.   :  50000   Length:978582      Length:978582     
##  Class  :character  1st Qu.: 100000   Class  :character  Class  :character 
##  Mode   :character  Median : 500000   Mode   :character  Mode   :character 
##                     Mean   : 665574                                        
##                     3rd Qu.:1000000                                        
##                     Max.   :2000000                                        
##                                                                            
##  Urban.vs.Rural         Region      Distribution.Channel Death.indicator
##  Length:978582      Min.   :1.000   Length:978582        Min.   :1      
##  Class  :character  1st Qu.:1.000   Class  :character    1st Qu.:1      
##  Mode   :character  Median :2.000   Mode   :character    Median :1      
##                     Mean   :2.748                        Mean   :1      
##                     3rd Qu.:4.000                        3rd Qu.:1      
##                     Max.   :6.000                        Max.   :1      
##                                                          NA's   :938206 
##  Year.of.Death    Lapse.Indicator  Year.of.Lapse    Cause.of.Death    
##  Min.   :2001     Min.   :1        Min.   :2001     Length:978582     
##  1st Qu.:2015     1st Qu.:1        1st Qu.:2017     Class  :character 
##  Median :2019     Median :1        Median :2021     Mode   :character 
##  Mean   :2018     Mean   :1        Mean   :2019                       
##  3rd Qu.:2021     3rd Qu.:1        3rd Qu.:2022                       
##  Max.   :2023     Max.   :1        Max.   :2023                       
##  NA's   :938206   NA's   :867693   NA's   :867693                     
##   Age.at.Death    Cause.of.Death.Description
##  Min.   :26.0     Length:978582             
##  1st Qu.:52.0     Class  :character         
##  Median :59.0     Mode   :character         
##  Mean   :58.6                               
##  3rd Qu.:66.0                               
##  Max.   :87.0                               
##  NA's   :938206
```

```r
max_year <- max(superlife_df$Issue.year)

superlife_df <- superlife_df %>%
    filter(is.na(Lapse.Indicator)) %>%
    mutate(Max.age = coalesce(Age.at.Death, max_year - Issue.year + Issue.age))

max_obs <- nrow(superlife_df)

superlife_df
```

```
## # A tibble: 867,693 x 19
##    Policy.number Policy.type Issue.year Issue.age Sex   Face.amount
##    <chr>         <chr>            <dbl>     <dbl> <chr>       <dbl>
##  1 K0JK2XD81ZNI  SPWL              2001        54 M         1000000
##  2 LKW7MA7BPAV1  SPWL              2001        41 M         2000000
##  3 MWUNTLGLE8NR  SPWL              2001        37 F          100000
##  4 BJJ1U7SIJUCS  SPWL              2001        48 F         1000000
##  5 JTFR6CAODMLQ  T20               2001        46 M           50000
##  6 CHBTT2PBPQYC  SPWL              2001        50 M         1000000
##  7 K3H8WN6O2QMJ  SPWL              2001        50 M          100000
##  8 HSITVHDV2XTJ  T20               2001        48 F          250000
##  9 KN7X1NLMWUIN  T20               2001        52 M         1000000
## 10 ISEEQXTXIIV4  SPWL              2001        42 F         2000000
## # i 867,683 more rows
## # i 13 more variables: Smoker.Status <chr>, Underwriting.Class <chr>,
## #   Urban.vs.Rural <chr>, Region <dbl>, Distribution.Channel <chr>,
## #   Death.indicator <dbl>, Year.of.Death <dbl>, Lapse.Indicator <dbl>,
## #   Year.of.Lapse <dbl>, Cause.of.Death <chr>, Age.at.Death <dbl>,
## #   Cause.of.Death.Description <chr>, Max.age <dbl>
```

Calculate Inforce Mortality

```
# Calculate mortality rate of inforce dataset
mortality_df <- superlife_df %>%
    select(Max.age) %>%
    rowwise() %>%
    mutate(Age = list(seq(1, Max.age))) %>%
    unnest(c(Age)) %>%
    group_by(Age) %>%
    summarise(lx = n()) %>%
    mutate(mortality_rate = 1 - ifelse(is.na(lead(lx)), 0, (lead(lx)/lx)))

mortality_df
```

```
## # A tibble: 87 x 3
##      Age     lx mortality_rate
##    <int>  <int>          <dbl>
##  1     1 867693              0
##  2     2 867693              0
##  3     3 867693              0
##  4     4 867693              0
##  5     5 867693              0
##  6     6 867693              0
##  7     7 867693              0
##  8     8 867693              0
##  9     9 867693              0
## 10    10 867693              0
## # i 77 more rows
```

```
write_csv(mortality_df, "../Data/Processed Data/Superlife-inforce-mortality-table.csv")
```