

Data Cleaning

Helitha Dharmadasa - z5451805

2024-02-23

```
library(tidyverse)
```

Read Data

```
superlife_df <- read_csv("../Data/Case Study Data/2024-srcsc-superlife-inforce-dataset.csv",  
  skip = 3)
```

```
## Rows: 978582 Columns: 16  
## -- Column specification -----  
## Delimiter: ","  
## chr (9): Policy.number, Policy.type, Sex, Smoker.Status, Underwriting.Class,...  
## dbl (7): 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.
```

```
cause_of_death_map <- read_csv("../Data/External Data/superlife_inforce-causes_of_death.csv")
```

```
## Rows: 17 Columns: 2  
## -- Column specification -----  
## Delimiter: ","  
## chr (2): Unique.Cause.of.Death, Description  
##  
## 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.
```

```
cause_of_death_map
```

```
## # A tibble: 17 x 2  
##   Unique.Cause.of.Death Description  
##   <chr>                <chr>  
## 1 A00-B99              Certain infectious and parasitic diseases  
## 2 C00-D48              Neoplasms  
## 3 D50-D89              Diseases of the blood and blood-forming organs and cer~  
## 4 E00-E88              Endocrine, nutritional and metabolic diseases  
## 5 F01-F99              Mental and behavioural disorders  
## 6 G00-G98              Diseases of the nervous system
```

```
## 7 I00-I99 Diseases of the circulatory system
## 8 J00-J98 Diseases of the respiratory system
## 9 K00-K92 Diseases of the digestive system
## 10 L00-L98 Diseases of the skin and subcutaneous tissue
## 11 M00-M99 Diseases of the musculoskeletal system and connective ~
## 12 N00-N98 Diseases of the genitourinary system
## 13 <NA> <NA>
## 14 O00-O99 Pregnancy, childbirth and the puerperium
## 15 Q00-Q99 Congenital malformations, deformations and chromosomal~
## 16 R00-R99 Symptoms, signs and abnormal clinical and laboratory f~
## 17 V01-Y89 External causes of morbidity and mortality
```

```
head(superlife_df)
```

```
## # A tibble: 6 x 16
##   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 2C1HL2XQOWME T20             2001      39 F     250000 NS
## 6 LKW7MA7BPAV1 SPWL             2001      41 M     2000000 NS
## # i 9 more variables: Underwriting.Class <chr>, Urban.vs.Rural <chr>,
## #   Region <dbl>, Distribution.Channel <chr>, Death.indicator <dbl>,
## #   Year.of.Death <dbl>, Lapse.Indicator <chr>, Year.of.Lapse <dbl>,
## #   Cause.of.Death <chr>
```

```
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
##                                     3rd Qu.:4.000
##                                     Max.   :6.000
##                                     Max.   :1
```

```
##                                     NA's      :938206
## Year.of.Death      Lapse.Indicator      Year.of.Lapse      Cause.of.Death
## Min.      :2001      Length:978582      Min.      :2001      Length:978582
## 1st Qu.:2015      Class :character      1st Qu.:2017      Class :character
## Median :2019      Mode  :character      Median :2021      Mode  :character
## Mean   :2018                                     Mean   :2019
## 3rd Qu.:2021                                     3rd Qu.:2022
## Max.   :2023                                     Max.   :2023
## NA's   :938206                                     NA's   :867693
```

Clean & Transform Data

```
superlife_df <- superlife_df %>%
  mutate(Lapse.Indicator = ifelse(Lapse.Indicator == "Y", 1, Lapse.Indicator),
         Age.at.Death = Year.of.Death - Issue.year + Issue.age)
```

```
# Join Cause of death desc. with main dataset
```

```
superlife_df <- left_join(superlife_df, cause_of_death_map, by = c(Cause.of.Death = "Unique.Cause.of.Death"))
```

```
superlife_df <- superlife_df %>%
  rename(Cause.of.Death.Description = "Description")
```

```
# Write cleaned data
```

```
write_csv(superlife_df, "../Data/Processed Data/CLEANED_2024-srcsc-superlife-inforce-dataset.csv")
```