Data Cleaning

Helitha Dharmadasa - z5451805

2024-02-23

```
library(tidyverse)
```

```
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.
```

```
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 O8FN60R4KXIS  T20               2001        54 F          100000 NS
## 2 K0JK2XD81ZNI  SPWL              2001        54 M         1000000 NS
## 3 AH3A98MHTO8H  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 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>
```

```
summary(superlife_df)
```

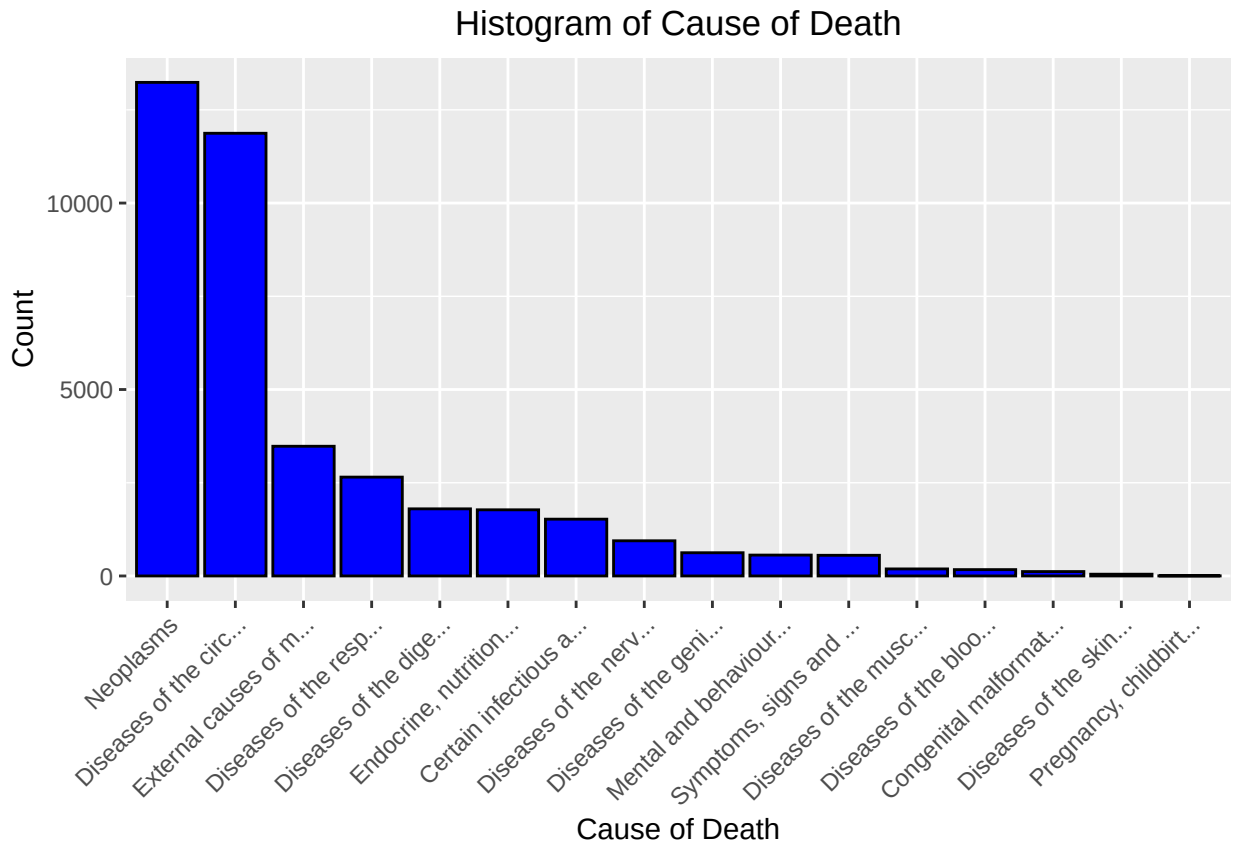
```
##  Policy.number      Policy.type          Issue.year     Issue.age
##  Length:978582      Length:978582      Min.   :2001   Min.   :26.0
##  Class :character   Class :character   1st Qu.:2009   1st Qu.:36.0
##  Mode  :character   Mode  :character   Median :2015   Median :44.0
##                                        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
```

Misc Plots for Initial Analysis

```
plot_df <- superlife_df %>%
    filter(Cause.of.Death.Description != "NA") %>%
    mutate(Cause.of.Death.Description = ifelse(nchar(Cause.of.Death.Description) >
        20, paste0(str_sub(Cause.of.Death.Description, end = 20), "..."), Cause.of.Death.Description)) %
    group_by(Cause.of.Death.Description) %>%
    summarise(count = n())

ggplot(plot_df, aes(x = reorder(Cause.of.Death.Description, desc(count)), y = count)) +
    geom_col(fill = "blue", color = "black") + labs(title = "Histogram of Cause of Death",
    x = "Cause of Death", y = "Count") + theme(axis.text.x = element_text(angle = 45,
    hjust = 1), plot.title = element_text(hjust = 0.5))
```

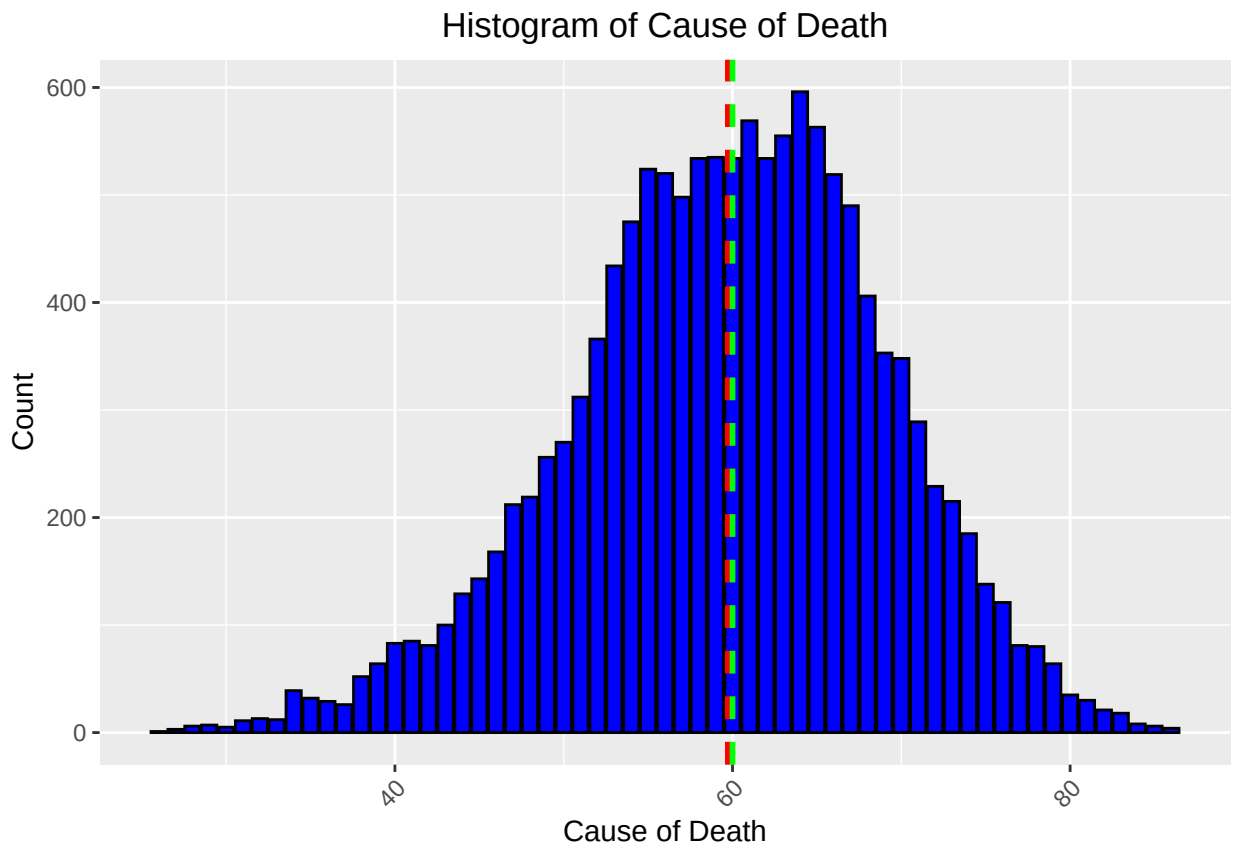



```
neoplasm_df <- superlife_df %>%
    filter(Cause.of.Death.Description == "Neoplasms") %>%
    group_by(Age.at.Death) %>%
    summarise(count = n())

mean <- weighted.mean(neoplasm_df$Age.at.Death, neoplasm_df$count)
median <- median(rep(neoplasm_df$Age.at.Death, times = neoplasm_df$count))

hist <- ggplot(neoplasm_df, aes(x = Age.at.Death, y = count)) + geom_col(fill = "blue",
    color = "black") + labs(title = "Histogram of Cause of Death", x = "Cause of Death",
    y = "Count") + theme(axis.text.x = element_text(angle = 45, hjust = 1), plot.title = element_text(h

hist + geom_vline(xintercept = mean, color = "red", linetype = "dashed", size = 1) +
    geom_vline(xintercept = median, color = "green", linetype = "dashed", size = 1)
```



Generate and write Neoplasm loading based on cancer death rates

```
neoplasm_df <- neoplasm_df %>%
    filter(Age.at.Death >= 50) %>%
    mutate(Weight = count/sum(count)) %>%
    select(Age.at.Death, Weight)

write_csv(neoplasm_df, "../Data/Processed Data/Neoplasm_Mortality_Loading.csv")
```

```
sum(neoplasm_df$Weight)
```

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

```
neoplasm_df <- superlife_df %>%
    filter(Cause.of.Death.Description == "Neoplasms") %>%
    group_by(Sex, Age.at.Death) %>%
    summarise(count = n())
```

```
## `summarise()` has grouped output by 'Sex'. You can override using the `.groups`
## argument.
```

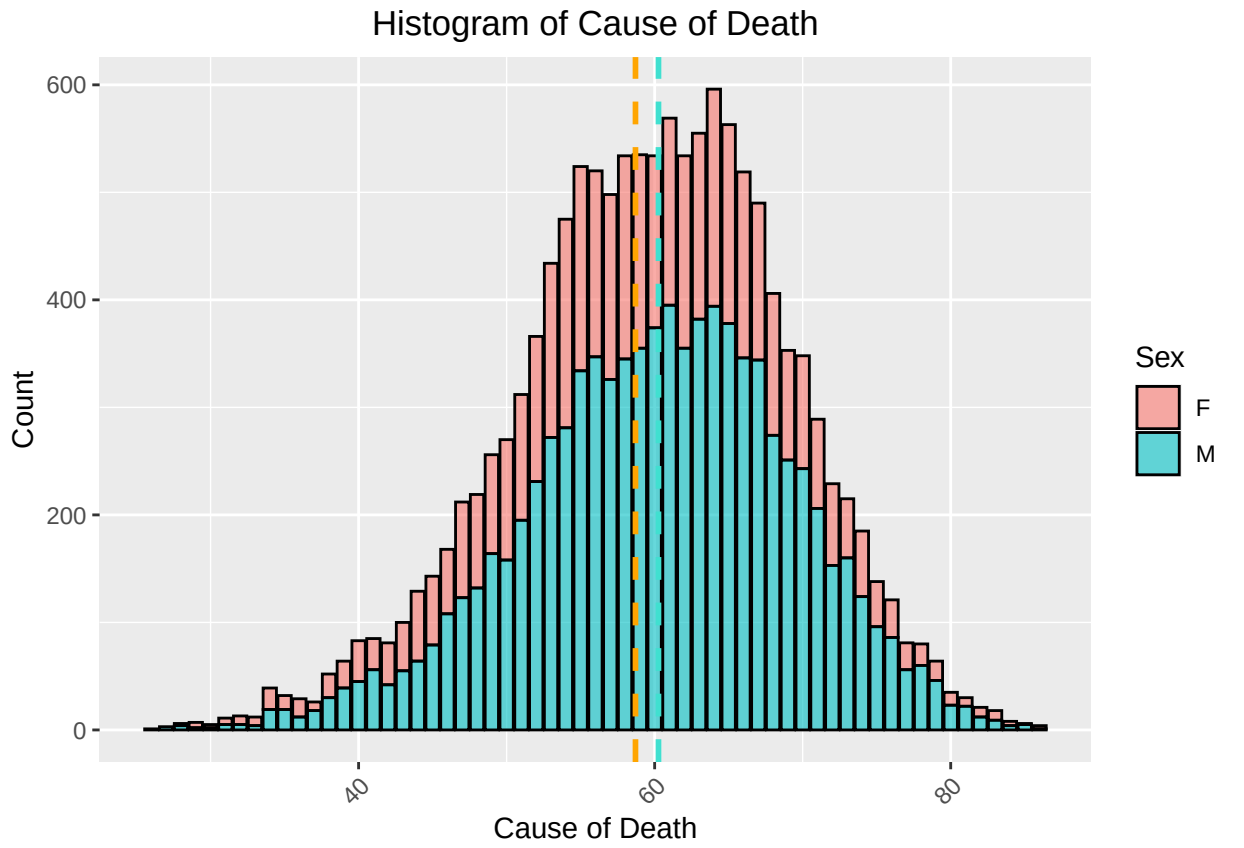
```r
means <- neoplasm_df %>%
    group_by(Sex) %>%
    summarise(mean = weighted.mean(Age.at.Death, count))

mean_f <- means %>%
    filter(Sex == "F") %>%
    pull(mean)


mean_m <- means %>%
    filter(Sex == "M") %>%
    pull(mean)

hist <- ggplot(neoplasm_df, aes(x = Age.at.Death, y = count, fill = Sex)) + geom_col(color = "black",
    alpha = 0.6) + labs(title = "Histogram of Cause of Death", x = "Cause of Death",
    y = "Count") + theme(axis.text.x = element_text(angle = 45, hjust = 1), plot.title = element_text(h

hist + geom_vline(xintercept = mean_f, color = "orange", linetype = "dashed", size = 1) +
    geom_vline(xintercept = mean_m, color = "turquoise", linetype = "dashed", size = 1)
```




```r
# Smoking rate in inforce data
smokers <- superlife_df %>%
    filter(Smoker.Status == "S") %>%
    nrow()
```

```r
smokers/nrow(superlife_df)
```

```
## [1] 0.06309129
```