# 4most Exercise 2:

Stephen Brownsey

## Introduction

These are paired annuities and as such there is only one expiry, this expiry will be the same for both the male and female from each group of data. The plots in this section will be used to visualise the data data from a couple of different perspectives, each of which will be coloured by sex to demonstrate the slight differences between genders. From the describe we can see that females are slightly younger on average and there are less female deaths.

The next section will cover some EDA into the dataset and main summarised points: * Men are slightly older on average than women * The death rate of males is significantly higher than women * As age increases the proportion of women outliving men will also increase

```
#Load the data data
data = readxl::read_xlsx("data/MockDataSet2.xlsx")
```

```
## New names:
## * `` -> `...1`
```

```
head(data)
```

```
## # A tibble: 6 x 6
##    ...1 EntryAgeM EntryAgeF DeathTimeM DeathTimeF AnnuityExpiredM
##   <dbl>     <dbl>     <dbl>      <dbl>      <dbl>           <dbl>
## 1     1      60.7      62.1          0          0            5.01
## 2     2      69.1      68.4          0          0            5.01
## 3     3      66.2      65.0          0          0            1.67
## 4     4      58.1      60.0          0          0            1.67
## 5     5      70.6      65.1          0          0            3.17
## 6     6      67.9      64.9          0          0            3.25
```

```
describe(data)
```

```
## data
##
##  6  Variables      14889  Observations
## --------------------------------------------------------------------------------
## ...1
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##    14889        0    14889        1     7445     4963    745.4   1489.8
##      .25      .50      .75      .90      .95
##   3723.0   7445.0  11167.0  13400.2  14144.6
##
```

```
## lowest :     1     2     3     4     5, highest: 14885 14886 14887 14888 14889
## ---------------------------------------------------------------------------
## EntryAgeM
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##    14889        0     7406        1    68.08    6.843    58.05    60.61
##      .25      .50      .75      .90      .95
##    64.75    67.90    72.11    75.45    77.71
##
## lowest :   0.0657  22.6545  22.7050  23.9890  25.7911
## highest:  93.5000  94.9045  95.6776  96.8415 104.8826
## ---------------------------------------------------------------------------
## EntryAgeF
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##    14889        0     8065        1     65.2    7.823    53.00    56.30
##      .25      .50      .75      .90      .95
##    61.15    65.53    69.65    73.49    75.93
##
## lowest :  0.2676  0.4906  1.1326 24.3498 26.5301
## highest: 91.6600 92.4413 92.6162 92.7527 93.7664
## ---------------------------------------------------------------------------
## DeathTimeM
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##    14889        0      857    0.282   0.2564   0.4775   0.0000   0.0000
##      .25      .50      .75      .90      .95
##   0.0000   0.0000   0.0000   0.2131   2.5598
##
## lowest : 0.0000 0.0121 0.0149 0.0176 0.0286, highest: 4.9671 4.9754 4.9781 4.9808 5.0055
## ---------------------------------------------------------------------------
## DeathTimeF
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##    14889        0      386    0.111  0.09824   0.1914        0        0
##      .25      .50      .75      .90      .95
##        0        0        0        0        0
##
## lowest : 0.0000 0.0505 0.0560 0.0724 0.0751, highest: 4.9644 4.9781 4.9836 4.9863 5.0027
## ---------------------------------------------------------------------------
## AnnuityExpiredM
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##    14889        0     1096    0.666    4.454   0.8922    1.863    2.979
##      .25      .50      .75      .90      .95
##    4.332    5.005    5.005    5.005    5.005
##
## lowest : 0.0246 0.0301 0.0602 0.0630 0.0739, highest: 4.9961 4.9988 5.0000 5.0027 5.0055
## ---------------------------------------------------------------------------
```

```r
#Combine the data:
data_fixer = function(data, col1, col2, col3, sex){
  data %>%
  select(col1, col2, col3) %>%
  mutate("sex" = sex) %>%
  rename("EntryAge" = col1,"DeathTime" = col2, "AnnuityExpired" =  col3) %>%
  mutate("death" = if_else(DeathTime > 0, 1 ,0)) %>%
  mutate("time" = if_else(death == 1, DeathTime , AnnuityExpired)) %>%
  mutate("time" = floor(as.numeric(time * 365.25))) %>%
```

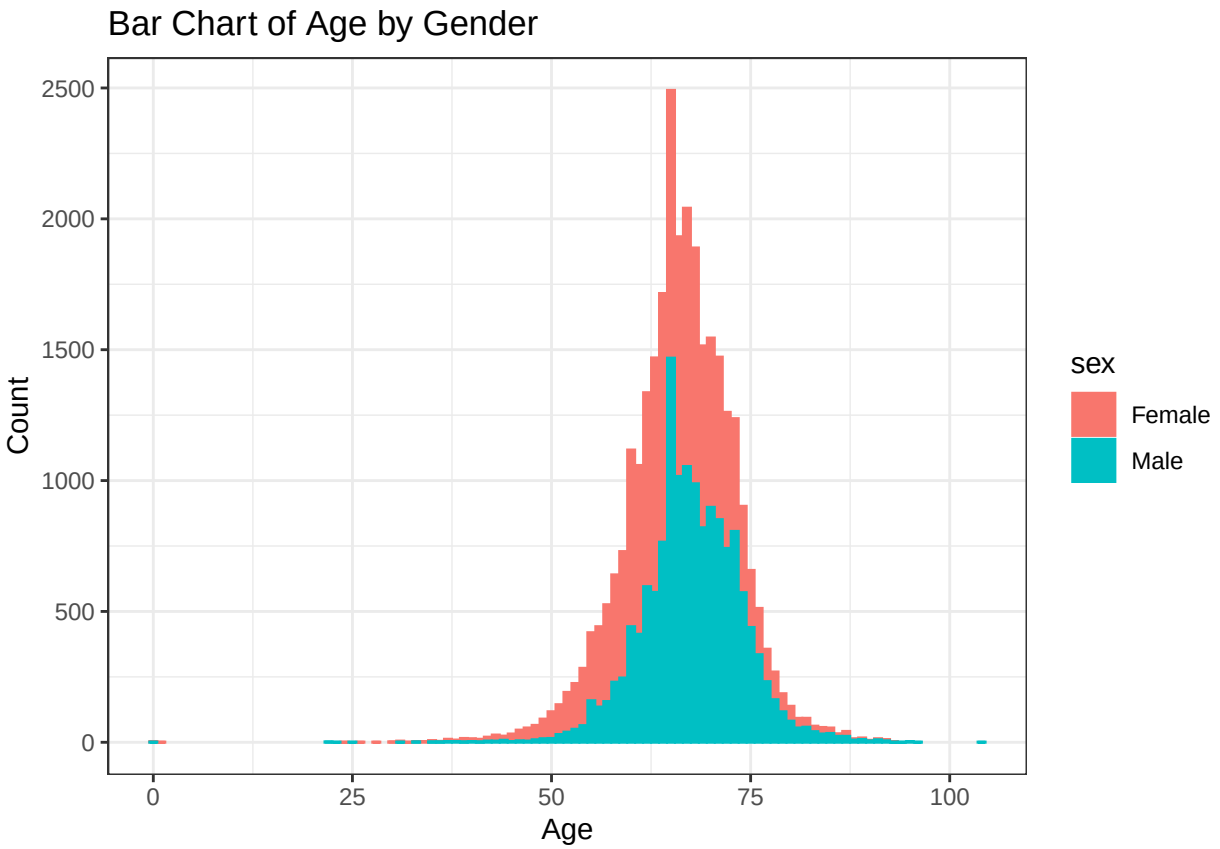
```
  select(sex, EntryAge, death, time)
}

data <- rbind(data_fixer(data,"EntryAgeM", "DeathTimeM", "AnnuityExpiredM", "Male"),
data_fixer(data,"EntryAgeF", "DeathTimeF", "AnnuityExpiredM", "Female"))
```

```
## Note: Using an external vector in selections is ambiguous.
## i Use ‘all_of(col1)‘ instead of ‘col1‘ to silence this message.
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.
## Note: Using an external vector in selections is ambiguous.
## i Use ‘all_of(col2)‘ instead of ‘col2‘ to silence this message.
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.
## Note: Using an external vector in selections is ambiguous.
## i Use ‘all_of(col3)‘ instead of ‘col3‘ to silence this message.
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.
```

```
data %>%
  group_by(sex) %>%
  summarise(deaths = sum(death))
```

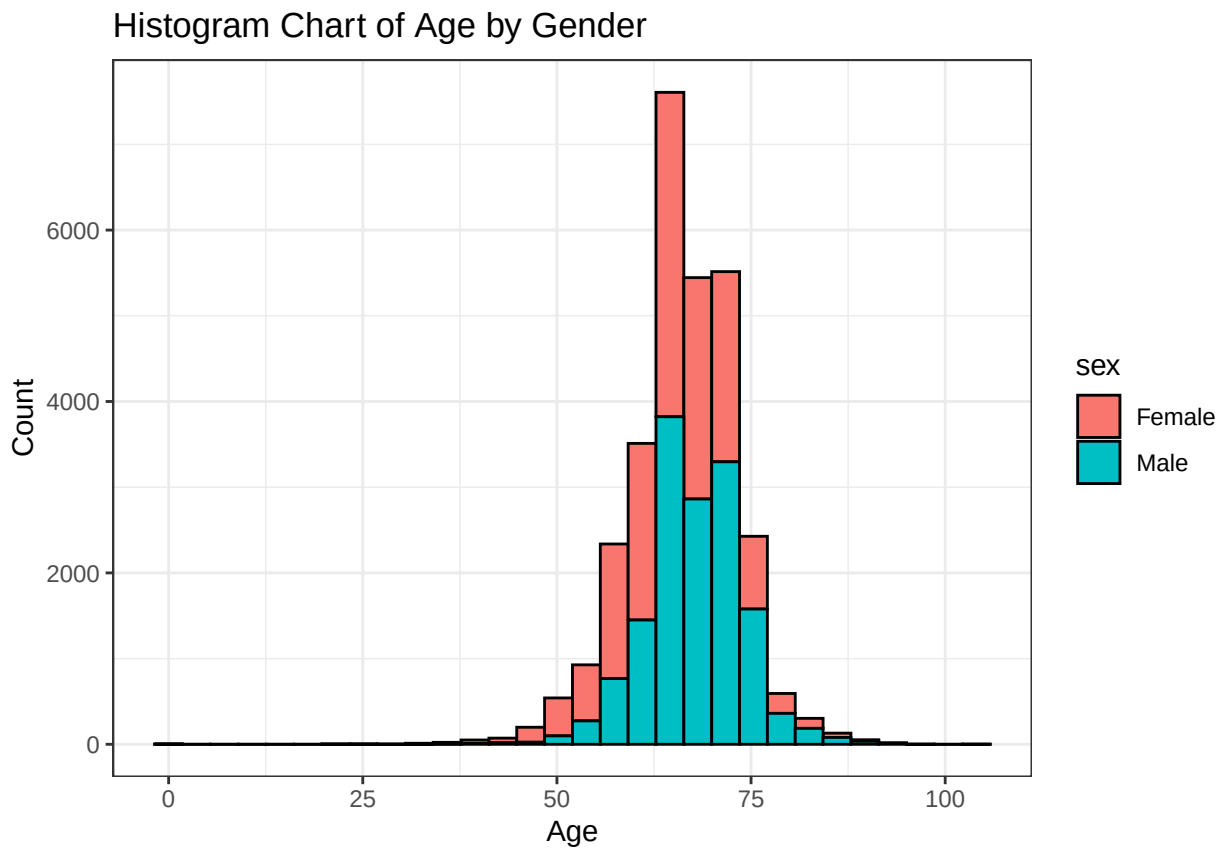
```
## # A tibble: 2 x 2
##   sex    deaths
##   <chr>   <dbl>
## 1 Female    572
## 2 Male     1554
```

```
data %>%
  ggplot() +
  geom_bar(aes(x = floor(EntryAge), colour = sex, fill = sex), position = "stack") +
  theme_bw() +
  labs(y = "Count",
       x = "Age", title = "Bar Chart of Age by Gender")
```

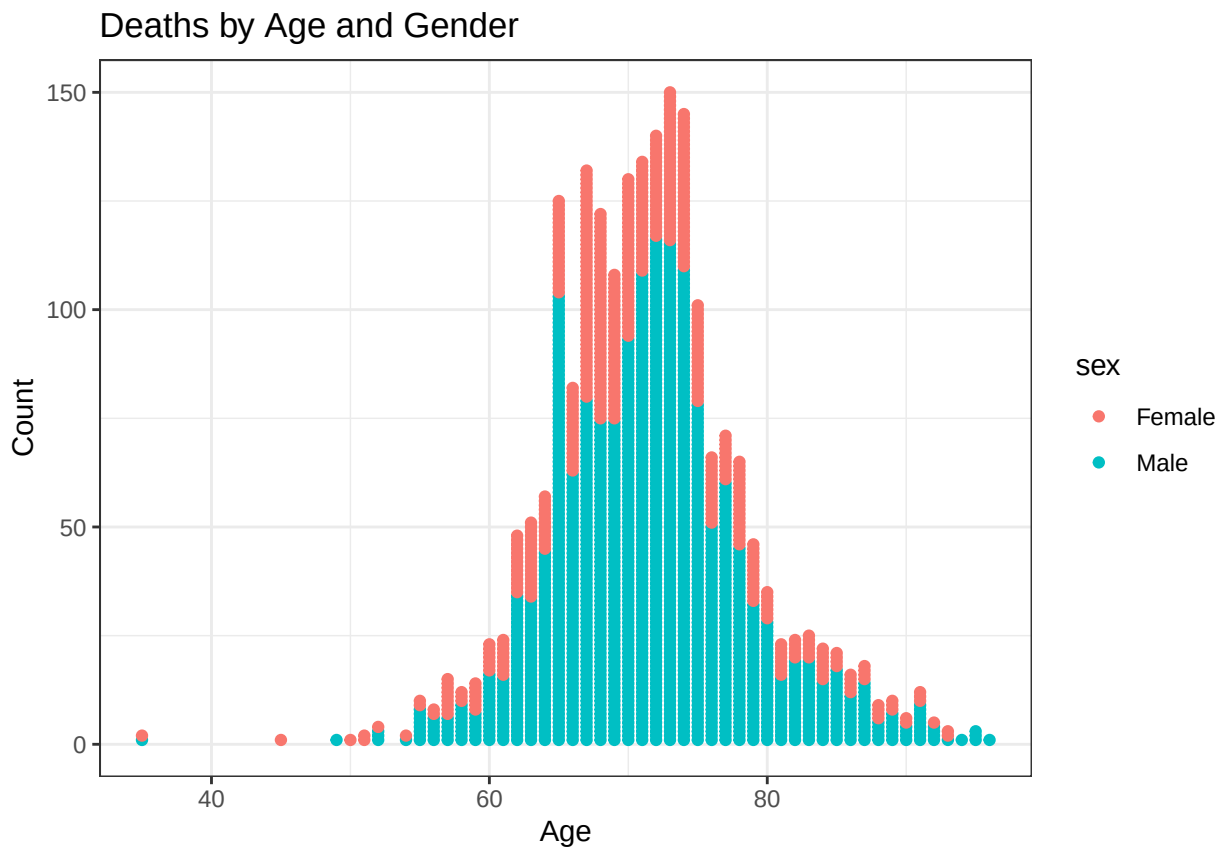



```
#Histogram demonstrating the distribution of the ages by gender.
#As arguably histogram may be easier to read than bar but same idea
data %>%
  ggplot() +
  geom_histogram(aes(floor(EntryAge), fill = sex), col = "black") +
  theme_bw()  +
  labs(y = "Count",
       x = "Age", title = "Histogram Chart of Age by Gender")
```

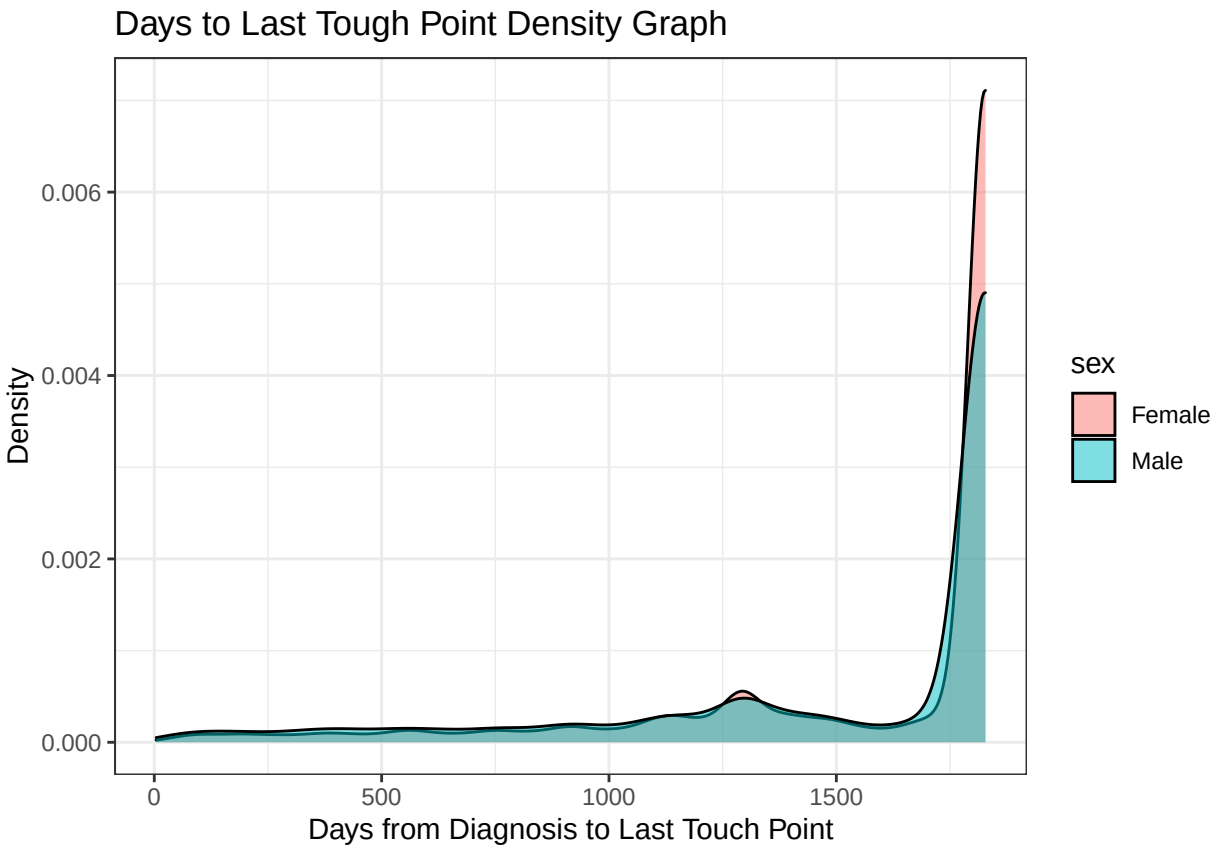
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
#Data points showing Days till last follow up against subject ID, coloured by Sex
#Position stack allows a geom_point to act similar to a bar
data %>%
  filter(death == 1) %>%
  ggplot() +
  geom_point(aes(x = floor(EntryAge), y = death, colour = sex, fill = sex),
             position = "stack") +
  theme_bw() +
  labs(y = "Count",
       x = "Age", title = "Deaths by Age and Gender")
```




```
#Density plot to show distributions of of both Male and Females, shows males taper
# off due to increased deaths towards the end and density of females at greater
# ages increases
data %>%
  ggplot() +
  geom_density(aes(time, fill = sex), alpha = 0.5, col = "black") +
  theme_bw() +
  ggtitle("Density showing Time till Last Touch Point Distribution by Gender") +
  labs(y = "Density", x = "Days from Diagnosis to Last Touch Point",
       title = "Days to Last Tough Point Density Graph")
```




### Question 2b):

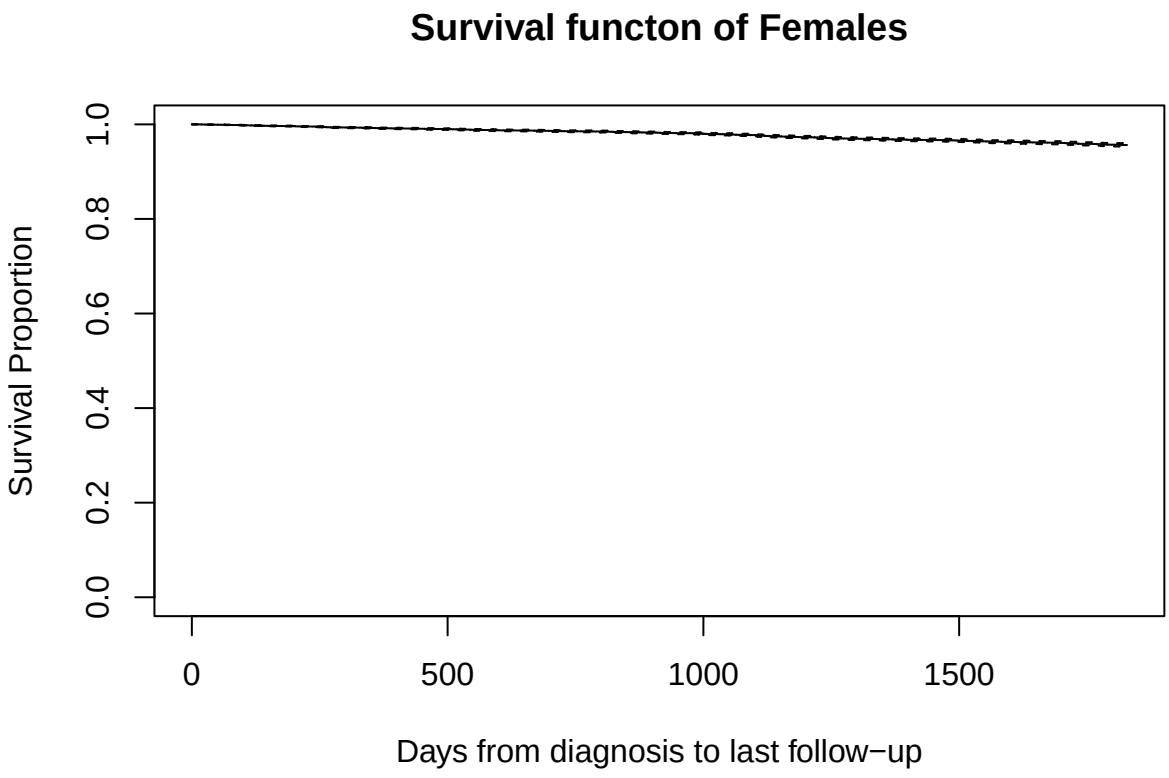
The Kaplan-Meier Estimator is a non-parametric statistic used to estimate the survival function from lifetime data. The estimator, which allows for ties, can be written as:

$$\hat{s}(t) = \prod_{i=y_i \leq t} \left(1 - \frac{d_i}{n_i}\right)^{\delta(i)}$$

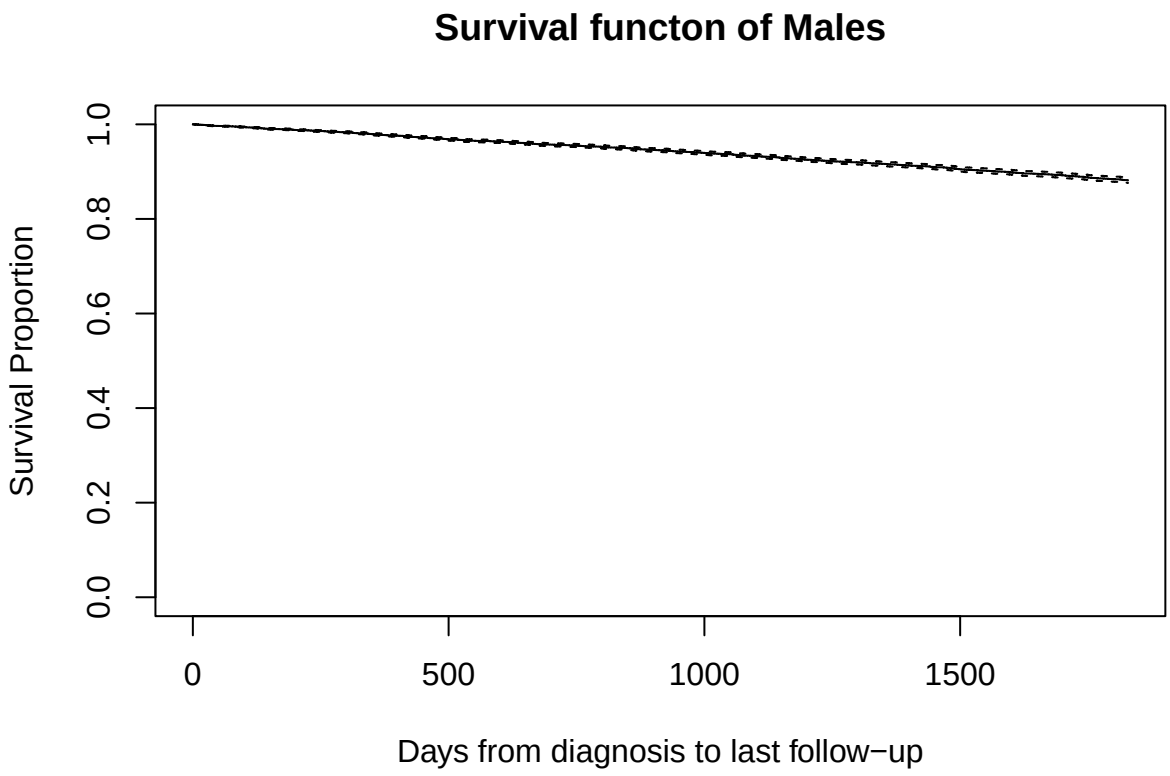
Where $n_i$ denotes the number of subjects at beginning of time point $I_i$, $d_i$ denotes the number of deaths that occur in $I_i$, $T_i$ denotes the time of death for observation i, $C_i$ denotes the censoring and $\delta_i$ is the indicator function that both $T_i$ and $C_i$ occur $\delta_i = 1(T_i, C_i)$

The plot, as seen below, is a series of decreasing horizontal steps which, with a sufficiently large sample size approaches the true survival rate of the population.

```
km_fit_female = survfit(Surv(time, death) ~ 1, data = data %>% filter(sex == "Female"), conf.type = "no
plot(km_fit_female, main="Survival functon of Females",
     xlab = "Days from diagnosis to last follow-up", ylab = "Survival Proportion")
```



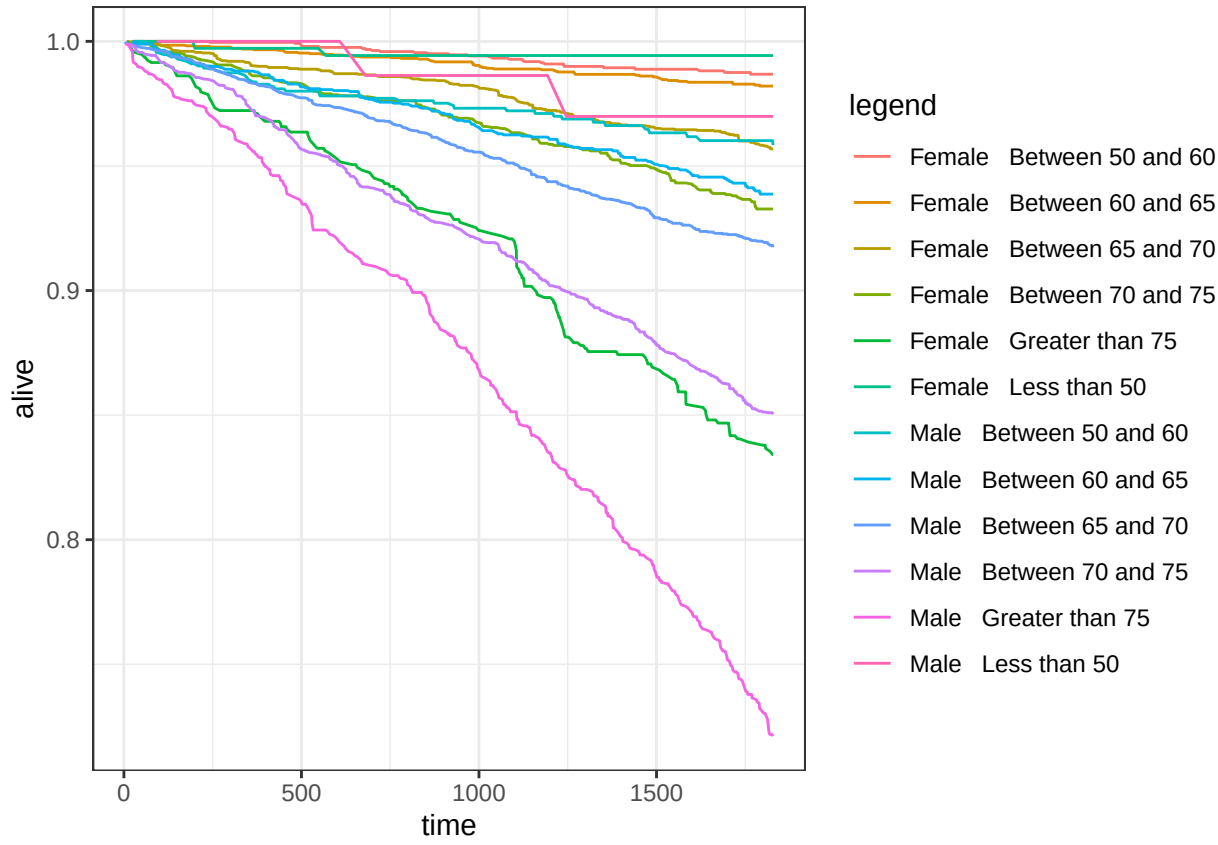

```
km_fit_male = survfit(Surv(time, death) ~ 1, data = data %>% filter(sex == "Male"))
plot(km_fit_male, main="Survival functon of Males",
     xlab = "Days from diagnosis to last follow-up", ylab = "Survival Proportion")
```




Let's create a nicer image to allow comparison between Age and Gender on the same graph: This looks at the different age bands as per the banding defined but this can be updated as and how desired to get views for more specific bands if required.

```
#Age banding
data = data %>%
  mutate(age_band = if_else(EntryAge < 50, "Less than 50",
                    if_else((EntryAge >= 50 & EntryAge < 60), "Between 50 and 60",
                    if_else((EntryAge >= 60 & EntryAge < 65), "Between 60 and 65",
                    if_else((EntryAge >= 65 & EntryAge < 70), "Between 65 and 70",
                    if_else((EntryAge >= 70 & EntryAge < 75), "Between 70 and 75",
                    if_else((EntryAge >= 75), "Greater than 75",
                            "Invalid Age")))))))

write_csv(data, "data/km_data.csv")

# Loop through and create the graph data for each combination of age and sex
# Extract the time and surv properties of the model and use these for the line graph
for (i in sort(unique(data$age_band))){
  for (j in sort(unique(data$sex))){
    data2 = data %>%
      filter(age_band == i) %>%
      filter(sex == j)

    fit = survfit(Surv(time, death) ~ 1, data = data2, conf.type = "none")
```

```
    if(!exists("graph_data")){
      graph_data = tibble(fit$time, fit$surv) %>%
  rename("time"="fit$time", "alive" = "fit$surv") %>%
  mutate(legend = paste(j, " " , i) )

    }
    else{

      graph_data1 = tibble(fit$time, fit$surv) %>%
  rename("time"="fit$time", "alive" = "fit$surv") %>%
  mutate(legend = paste(j, " " , i) )

      graph_data = rbind(graph_data, graph_data1)

    }

  }

}

#Plotting the graph
graph_data %>%
  ggplot(aes(x = time, y = alive, col = legend)) +
  geom_line() +
  theme_bw() +
  labs()
```




```
# Note from this graph, there is no real steep drop off observed
```

It is possible to test whether the two groupings come from the same population of data and this can help distinguish at a statistically significant level that there is a difference.

The hypothesis test undertaken will test for a difference between two survival curves using the G-rho family of tests. The idea being that if the null hypothesis is true then there is no difference between the two survival curves and therefore the two datasets contain survival data which originates from the same underlying population. On the contrary, if the alternative hypothesis is true, then there is statistically significant evidence to suggest that since the two datasets have different survival curves then they come from different underlying populations. The test can be formulated as followed:

Null Hypothesis: $H0$ : There is no difference between the two survival curves. Alternative Hypothesis: $H1$ : There is a difference between the two survival curves. Conclusion: Since T $<$ 0.05, there is sufficient evidence to reject the null hypothesis and accept the alternative hypothesis - there is a difference between the survival curves generated between male and females. We could band this down further to prove it was for each grouping of data but not really any point as they will all be significant, although less so as it will account for some of the age bias as males are slightly older than females in the population.

```
survdiff(Surv(time, death) ~ sex, data = data)
```

```
## Call:
## survdiff(formula = Surv(time, death) ~ sex, data = data)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## sex=Female 14889      572     1084       242       493
```

```
## sex=Male   14889     1554     1042       251       493
##
##  Chisq= 493  on 1 degrees of freedom, p= <2e-16
```

```
survdiff(Surv(time, death) ~ sex, data = data %>%
           filter(age_band == "Greater than 75"))
```

```
## Call:
## survdiff(formula = Surv(time, death) ~ sex, data = data %>% filter(age_band ==
##     "Greater than 75"))
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## sex=Female   942      146      217      23.1      36.7
## sex=Male    1682      437      366      13.6      36.7
##
##  Chisq= 36.7  on 1 degrees of freedom, p= 1e-09
```

## Question 2c): Gompertz solution

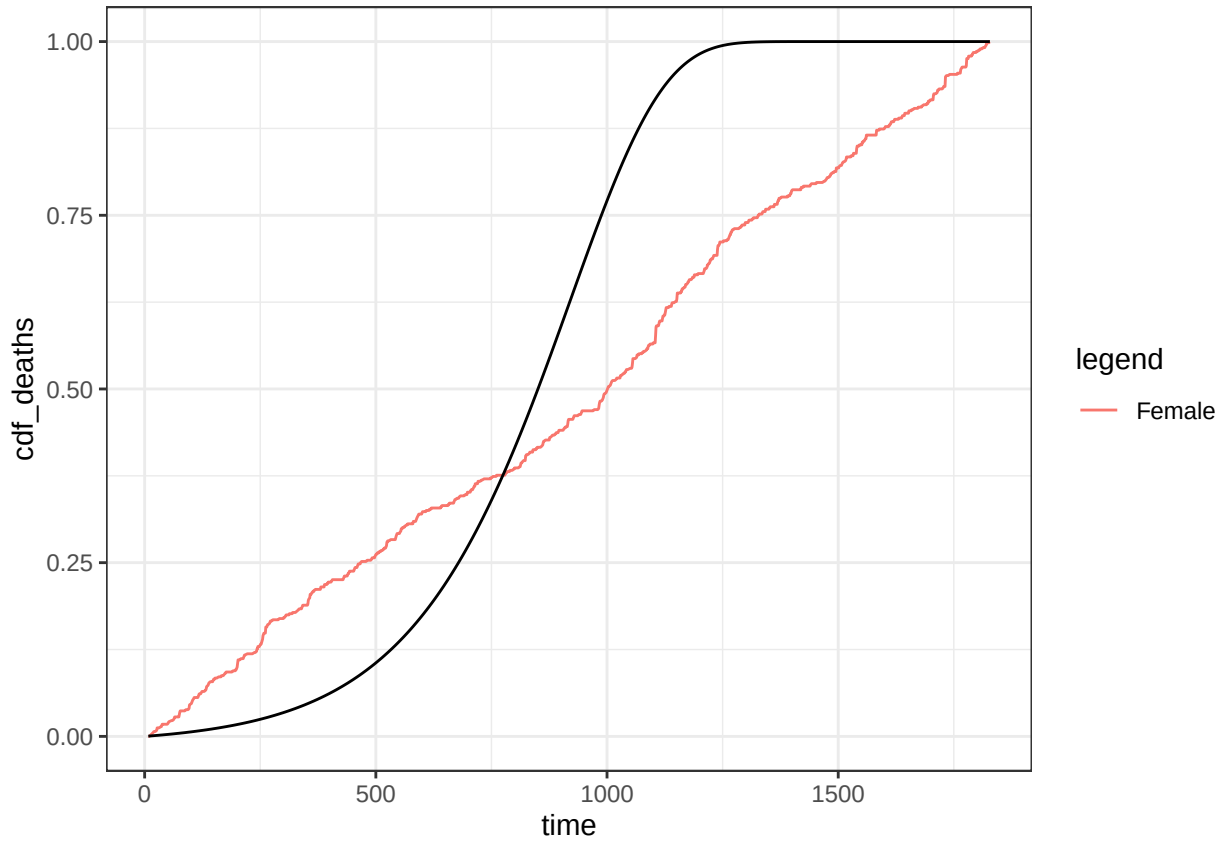
Wasn't sure exactly what do do with this question as CDF is ascending but the km curve will be descending. Wasn't sure what M/B would refer to exactly so I implemented the CDF from Wikipedia as it was slightly simpler and enabled just two parameters to be defined $\eta$ and b referring to the shape and scale parameters respectively.

This formula for the Gompertz CDF can be given by $1 - e^{(-\eta(e^{bx-1}))}$

I did two lots of graphs one using purely the deaths which followed a somewhat straight line. But would go up to a CDF of 1 and one which had a denomiator of the people left in the study, both of these were straighter in respect to time than the Gompertz CDF function which would have a sharp drop-off.

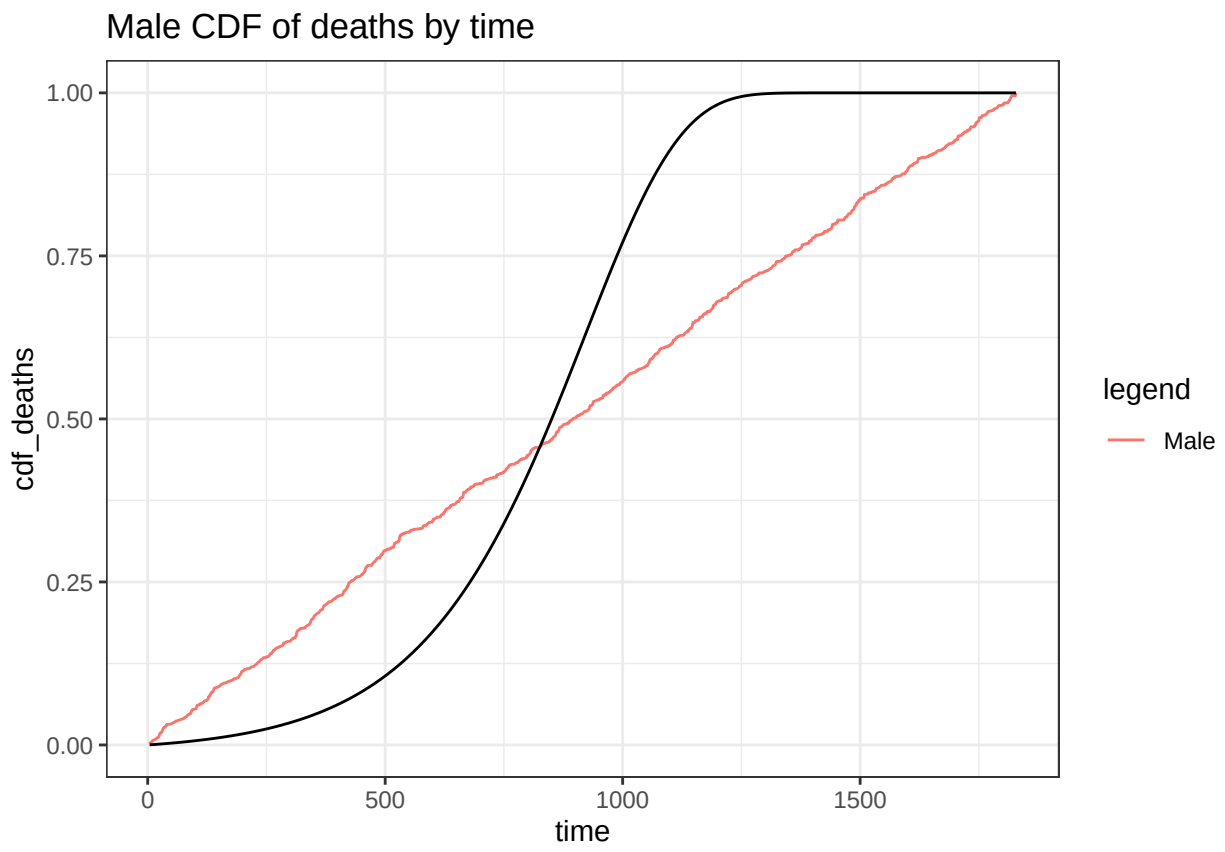
```
#gompertz function
calc_gompertz <- function(x, eta, b){
  (1 - (exp(-eta*(exp(b*x) - 1))))
}

#Just doing it for males and females
fit_f = survfit(Surv(time, death) ~ 1, data = data %>% filter(sex == "Female"), conf.type = "none")
fit_m = survfit(Surv(time, death) ~ 1, data = data %>% filter(sex == "Male"), conf.type = "none")

#Create the df required for the graphing
create_deaths_df <- function(fit, legend){
  totals <- array()
  j = 0
  for (i in fit$n.event){
    j= i +j
    totals = append(totals, j)
  }
  totals = totals[!is.na(totals)]#First element was na - removing
  total = max(totals)

  df = tibble(fit$n.event, fit$time, fit$n.risk,totals) %>%
    rename(n_event = "fit$n.event", time =  "fit$time", risk = "fit$n.risk")
```

```
  df %>%
    mutate(legend = legend) %>%
    mutate(in_study = risk - totals) %>%
    mutate(cdf = totals/in_study) %>%
    mutate(cdf_deaths = totals/total) %>%
    #Had a little play with the values to see which looked somewhat respectable
    mutate(gompertz = calc_gompertz(time, 0.01, 0.005)) %>%
    mutate(gompertz_total = calc_gompertz(time, 0.003, 0.0025))

}

gompertz_df_f = create_deaths_df(fit_f, "Female" )
gompertz_df_m = create_deaths_df(fit_m, "Male")

#Some plots with gompertz added
gompertz_df_f %>%
  ggplot() +
  geom_line(aes(x = time, y = cdf_deaths, col = legend)) +
  geom_line(aes(x= time, y= gompertz)) +
  theme_bw() +
  labs("Female CDF of deaths by time")
```

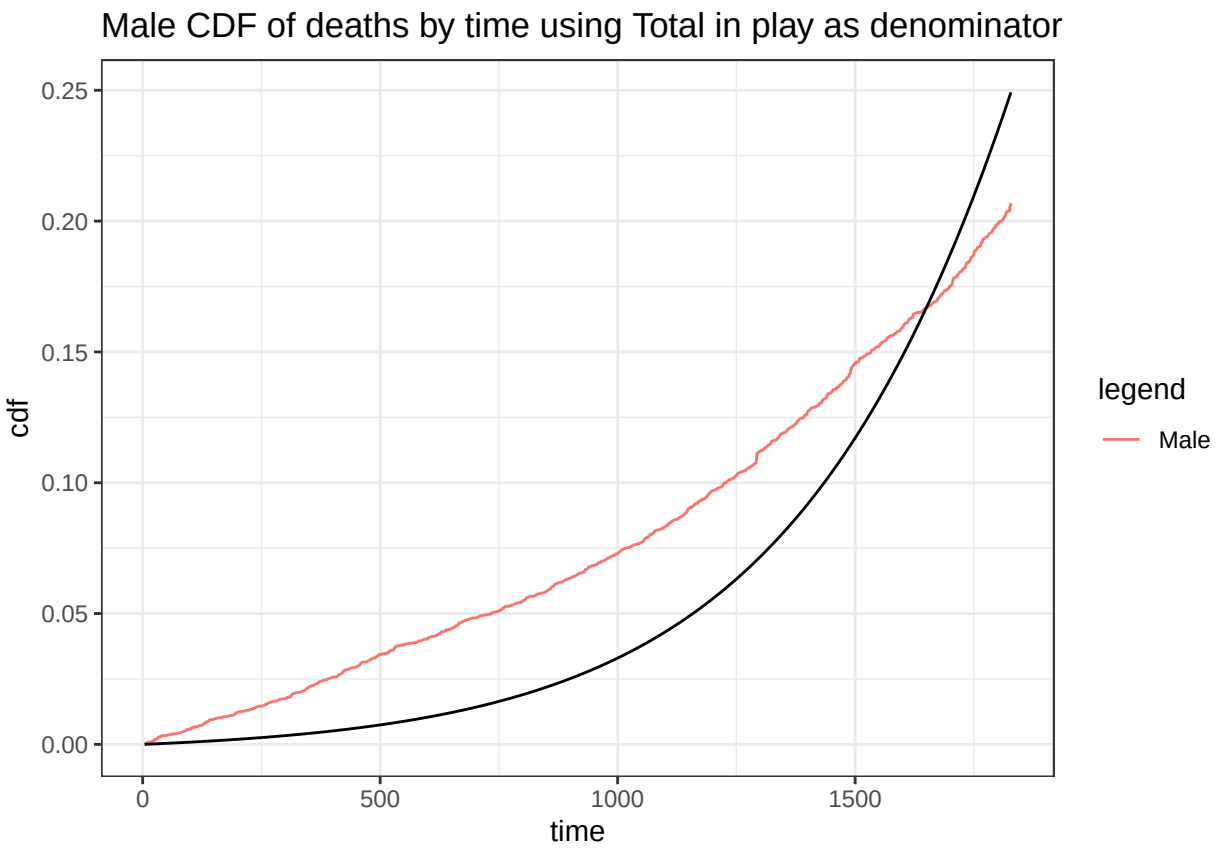



```
gompertz_df_m %>%
  ggplot() +
  geom_line(aes(x = time, y = cdf_deaths, col = legend)) +
```

```
  geom_line(aes(x= time, y= gompertz)) +
  theme_bw() +
  labs(title = "Male CDF of deaths by time")
```

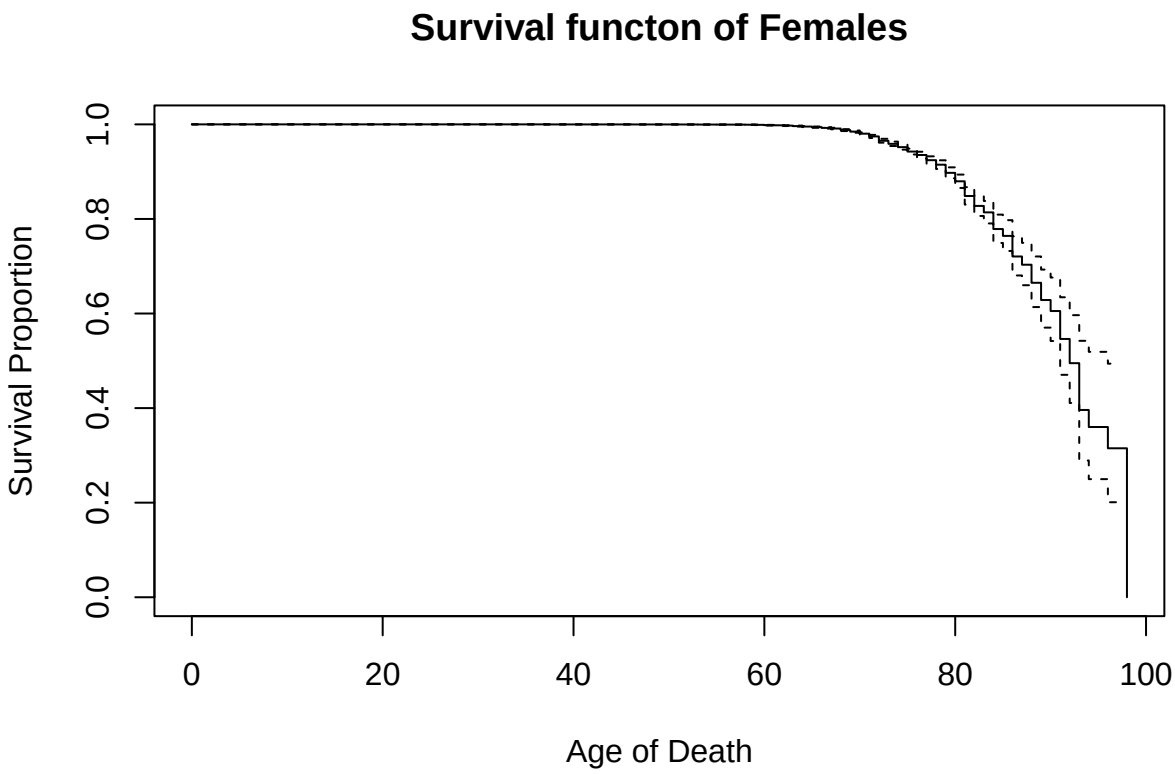



```
gompertz_df_m %>%
  ggplot() +
  geom_line(aes(x = time, y = cdf, col = legend)) +
  geom_line(aes(x= time, y= gompertz_total)) +
  theme_bw() +
  labs(title = "Male CDF of deaths by time using Total in play as denominator")
```

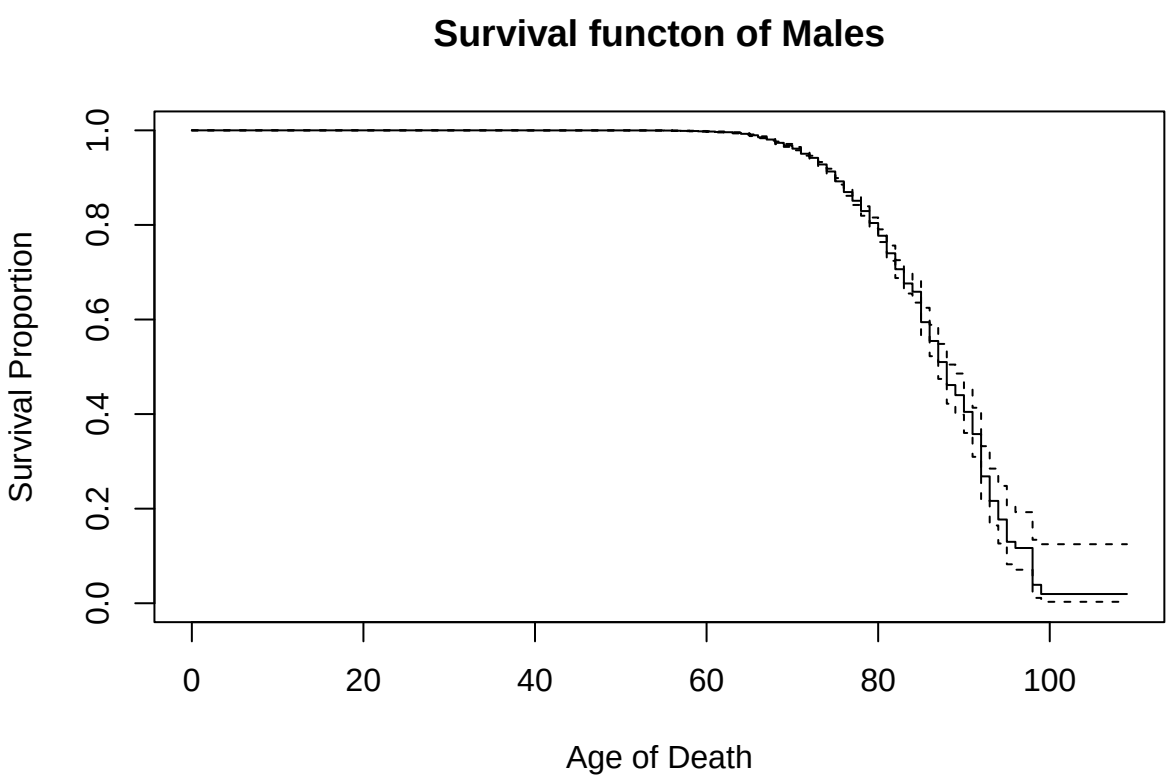



Thoughts

```
data <- data %>%
  mutate(age = floor(EntryAge + (time/365.25)))
km_fit_female = survfit(Surv(age, death) ~ 1, data = data %>% filter(sex == "Female"), conf.type = "none
plot(km_fit_female, main="Survival functon of Females",
     xlab = "Age of Death", ylab = "Survival Proportion")
```




```
km_fit_male = survfit(Surv(age, death) ~ 1, data = data %>% filter(sex == "Male"))
plot(km_fit_male, main="Survival functon of Males",
     xlab = "Age of Death", ylab = "Survival Proportion")
```


Survival functon of Males
Survival Proportion
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
Age of Death