Around Simpson's Paradox

Benjamin Cathelineau

21/11/2021

Introduction

The first study was carried in 1977, the second study in 1995.

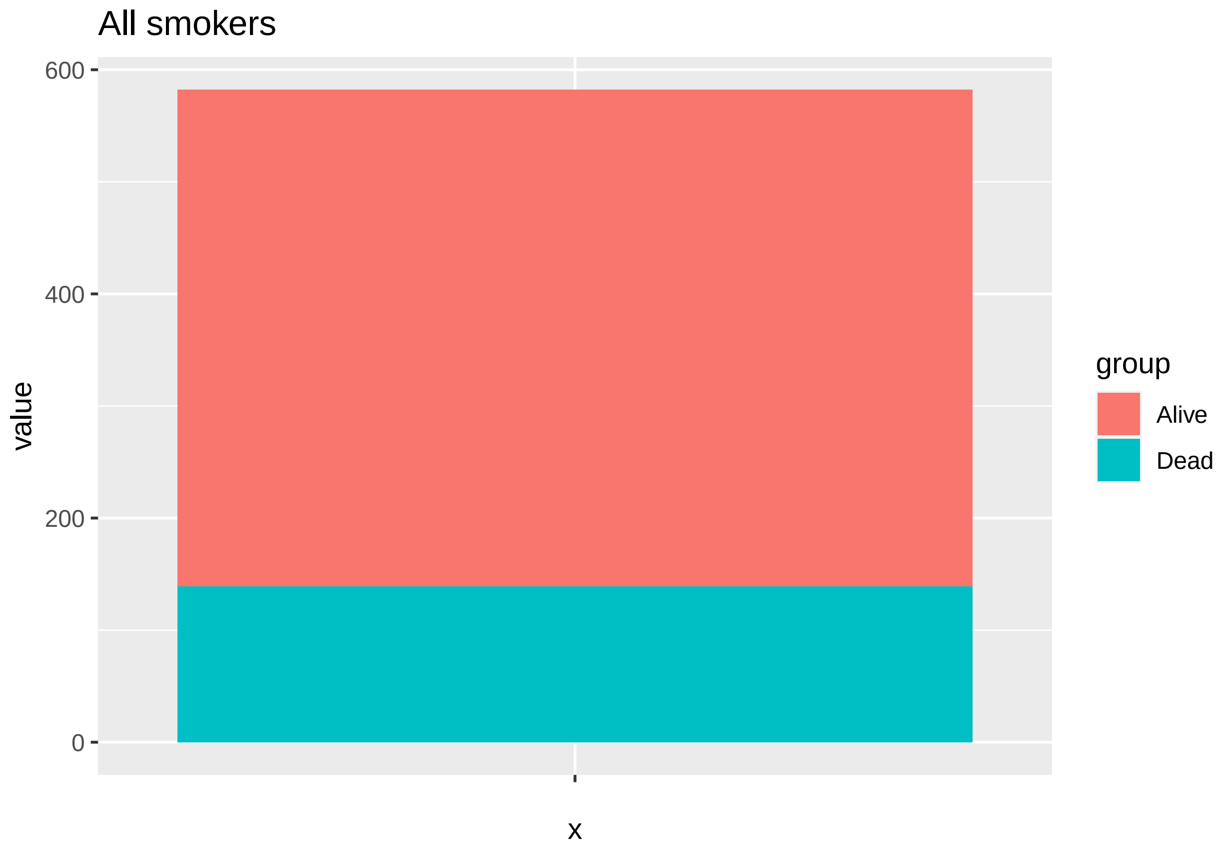
Question 1

TODO pie chart

```
df = read.csv("Subject6_smoking.csv")
compute_mortality_rate <- function(smoker_arg,df_arg, title_arg){
  nb_alive= df_arg %>% filter(Status == "Alive" & Smoker== smoker_arg) %>% nrow()
  nb_dead= df_arg %>% filter(Status == "Dead" & Smoker== smoker_arg) %>% nrow()


  df <- data.frame(
  group = c("Alive", "Dead"),
  value = c(nb_alive,nb_dead)
  )

  bp<- ggplot(df, aes(x="", y=value, fill=group))+
  geom_bar(width = 1, stat = "identity") + ggtitle(title_arg)
  print(bp)


  nb_dead / (nb_alive+nb_dead) # divide the number of dead by the total, the total being the addition o
}
```

We declare a function so that we can' compute both rates (for smokers and non smokers), without repeating our code

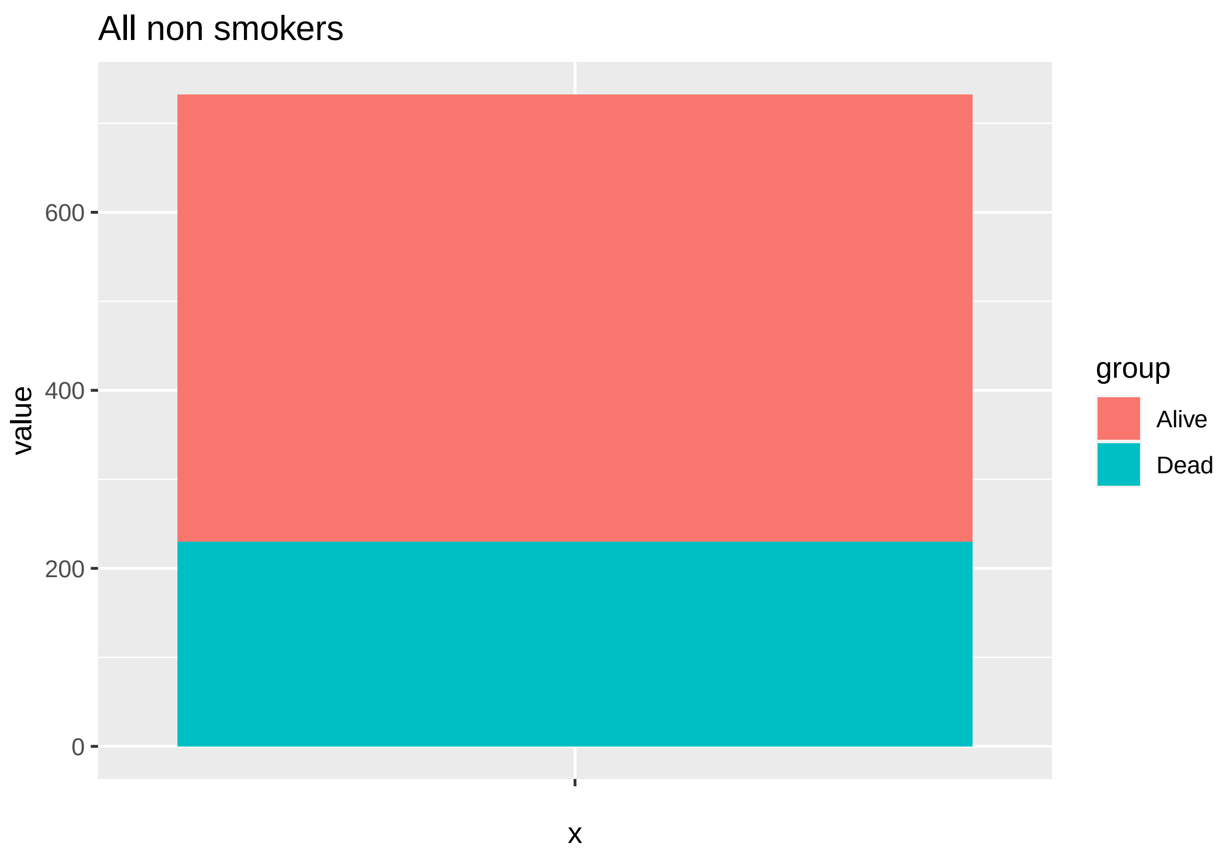
```
compute_mortality_rate(smoker_arg = "Yes", df_arg =df,title_arg = "All smokers" )
```




```
## [1] 0.2388316
```

The rate for the smoker group

```
compute_mortality_rate(smoker_arg = "No", df_arg =df, title_arg = "All non smokers")
```




```
## [1] 0.3142077
```

The rate for the non smoking group

The mortality rate is significantly higher for the group that is not smoking. In other words, in with this data, a woman who smoked in 1977 is less likely to have died in 1995 than a woman who did not smoke in 1977.

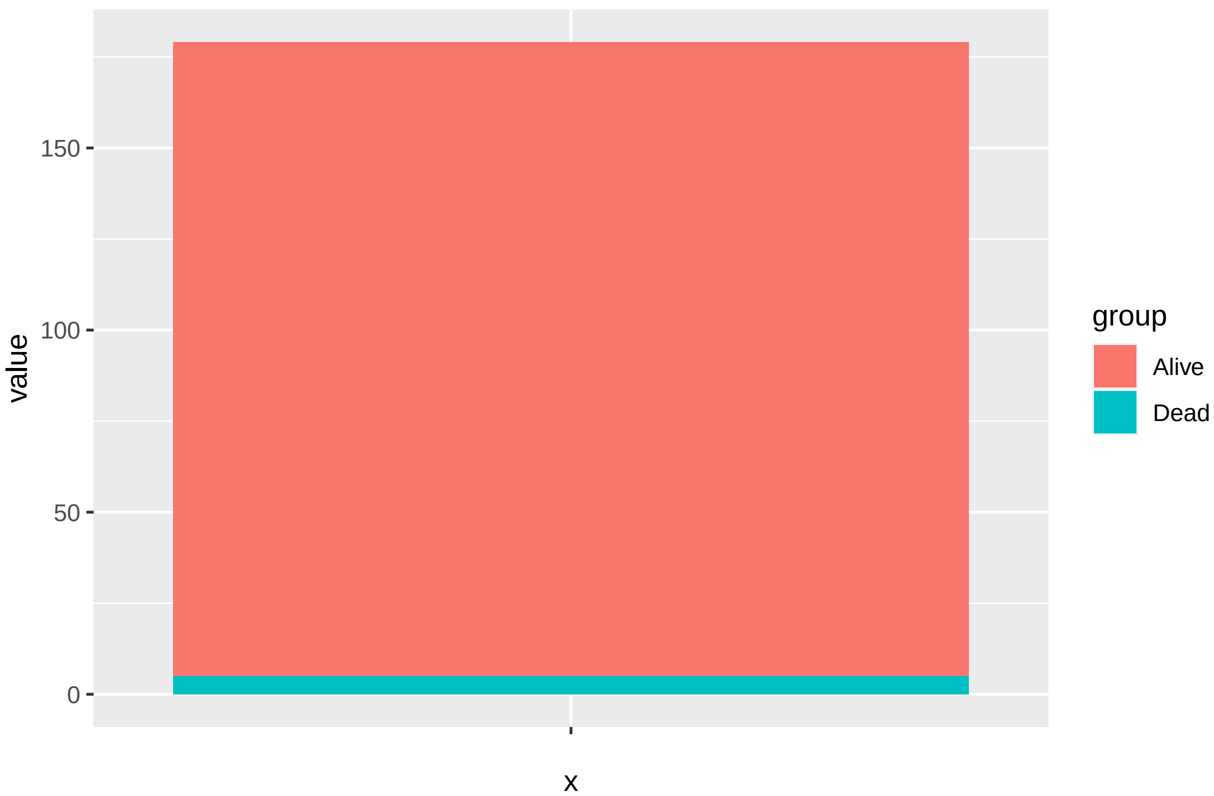
Of course, this is very surprising because it is now known that smoking cigarette increases the risk of death, trough various mechanisms, such as increased risk of cancer and cardiovascular disease. For more details, consult the relevant wikipedia article.

Question 2

We will use the recommended age grouping.

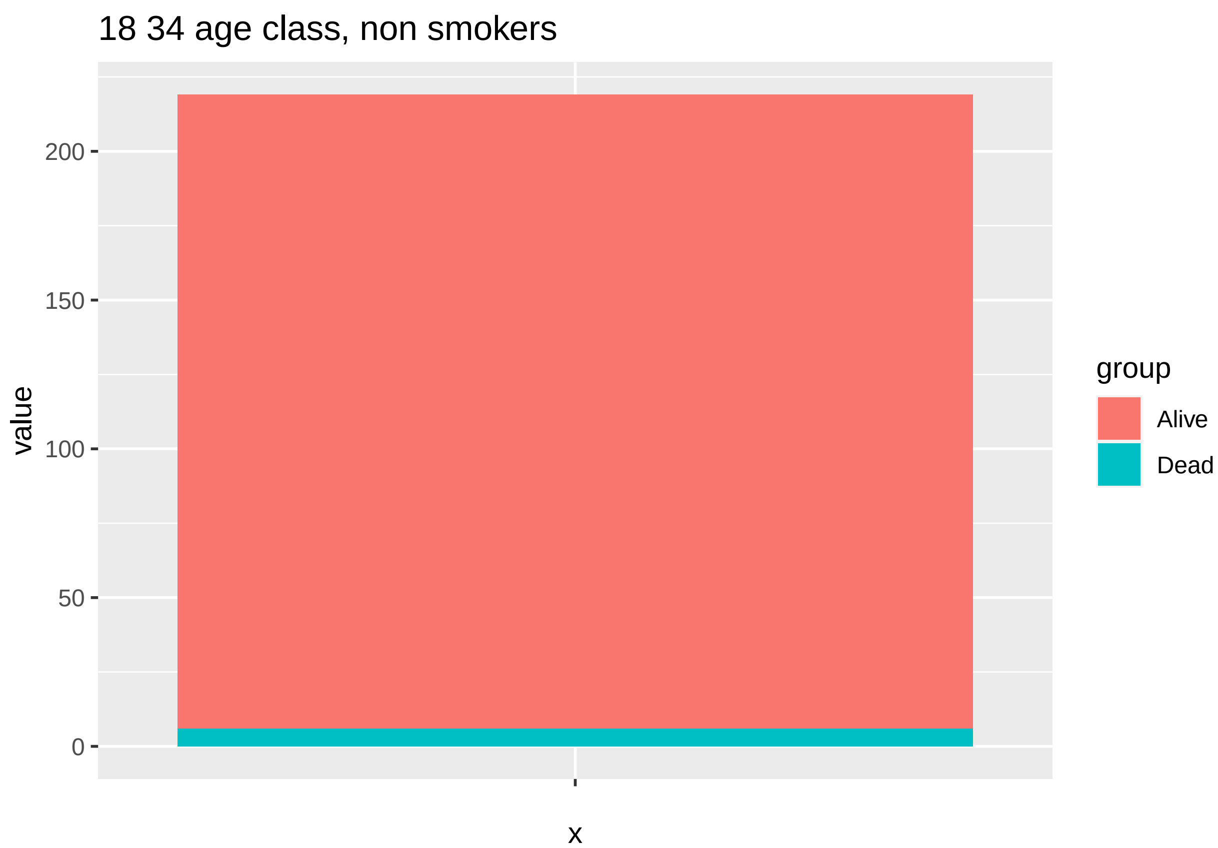
```
class1834 = df %>% filter(Age >= 18 & Age < 34)
class3454 = df %>% filter(Age >= 34 & Age < 54)
class5464 = df %>% filter(Age >= 54 & Age < 64)
class64 = df %>% filter(Age >= 64)
compute_mortality_rate(smoker_arg = "Yes", df_arg = class1834,title_arg = "18 34 age class, smokers")
```

18 34 age class, smokers



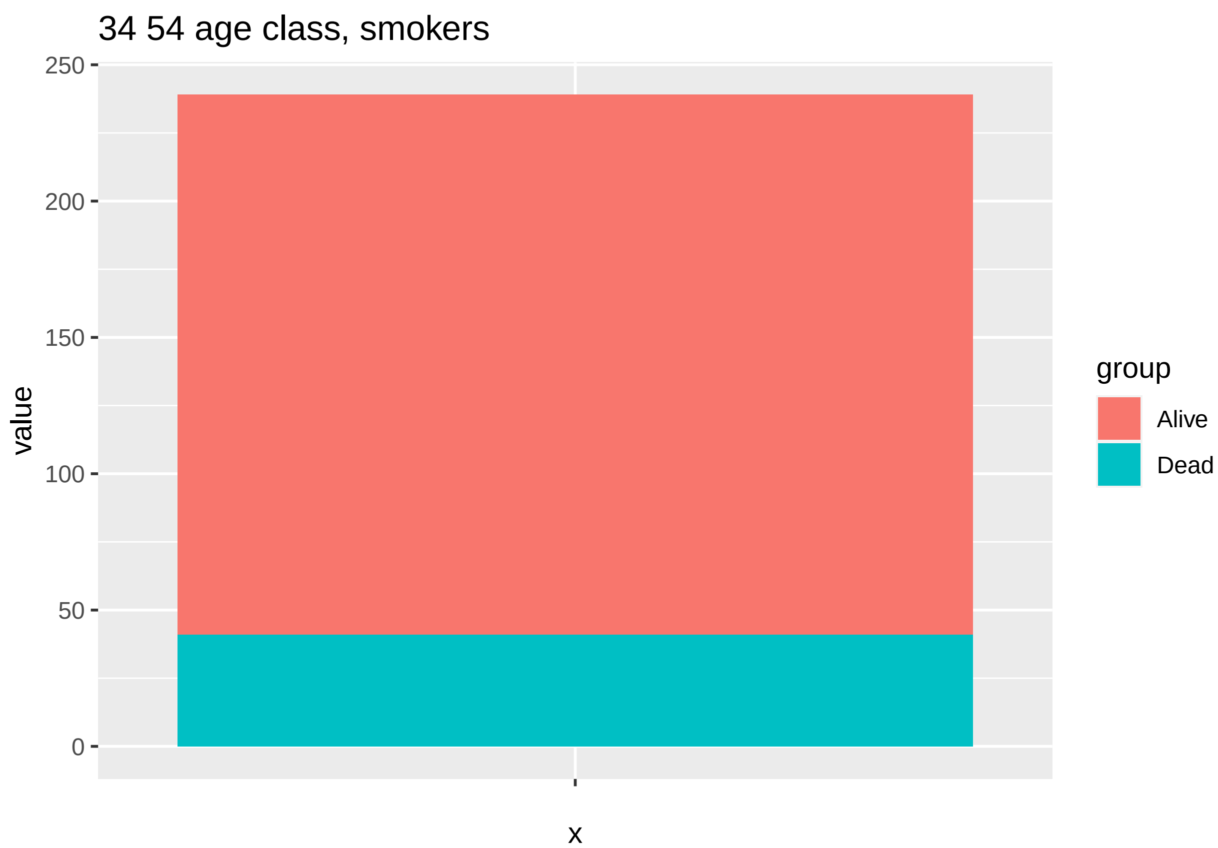

```
## [1] 0.02793296
```

```
compute_mortality_rate(smoker_arg = "No", df_arg = class1834,title_arg = "18 34 age class, non smokers")
```



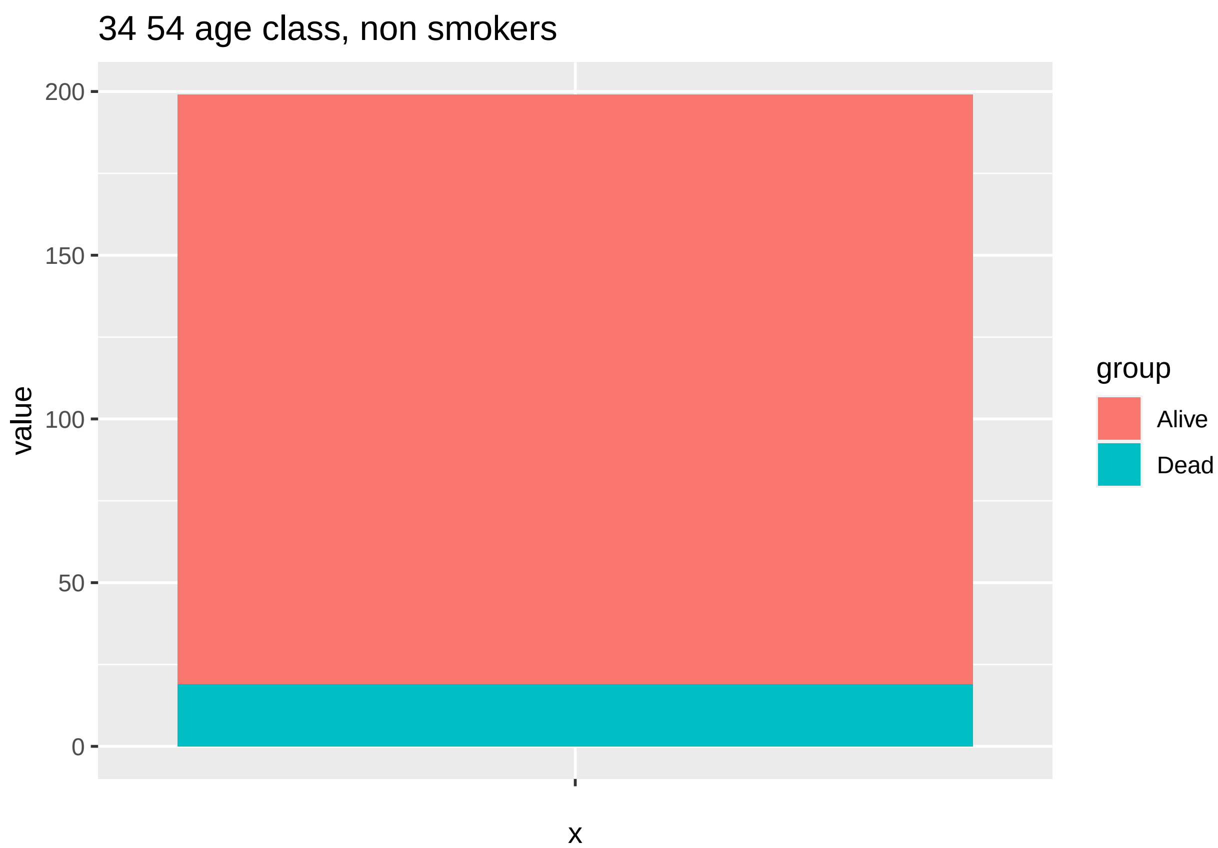

```
## [1] 0.02739726
```

```
compute_mortality_rate(smoker_arg = "Yes", df_arg = class3454,title_arg = "34 54 age class, smokers")
```



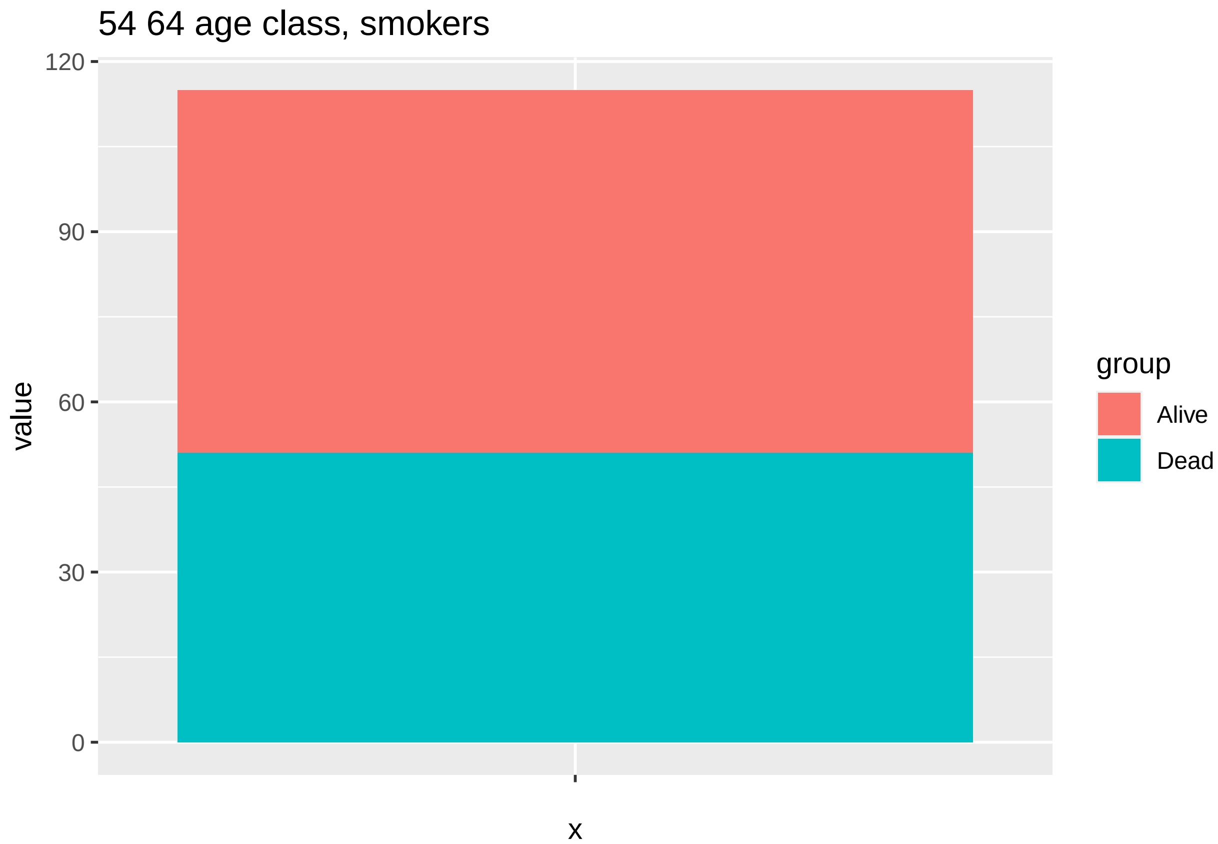

```
## [1] 0.1715481
```

```
compute_mortality_rate(smoker_arg = "No", df_arg = class3454,title_arg = "34 54 age class, non smokers")
```



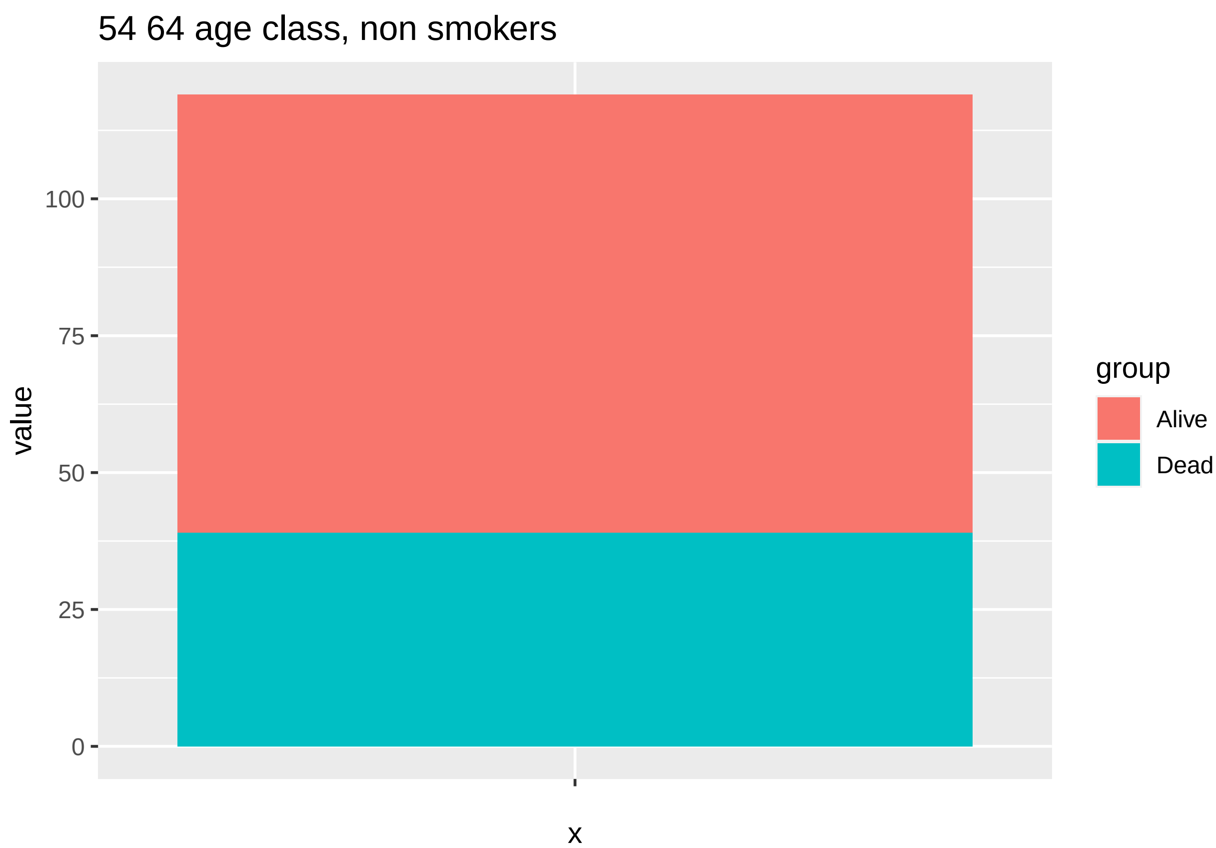

```
## [1] 0.09547739
```

```
compute_mortality_rate(smoker_arg = "Yes", df_arg = class5464,title_arg = "54 64 age class, smokers")
```



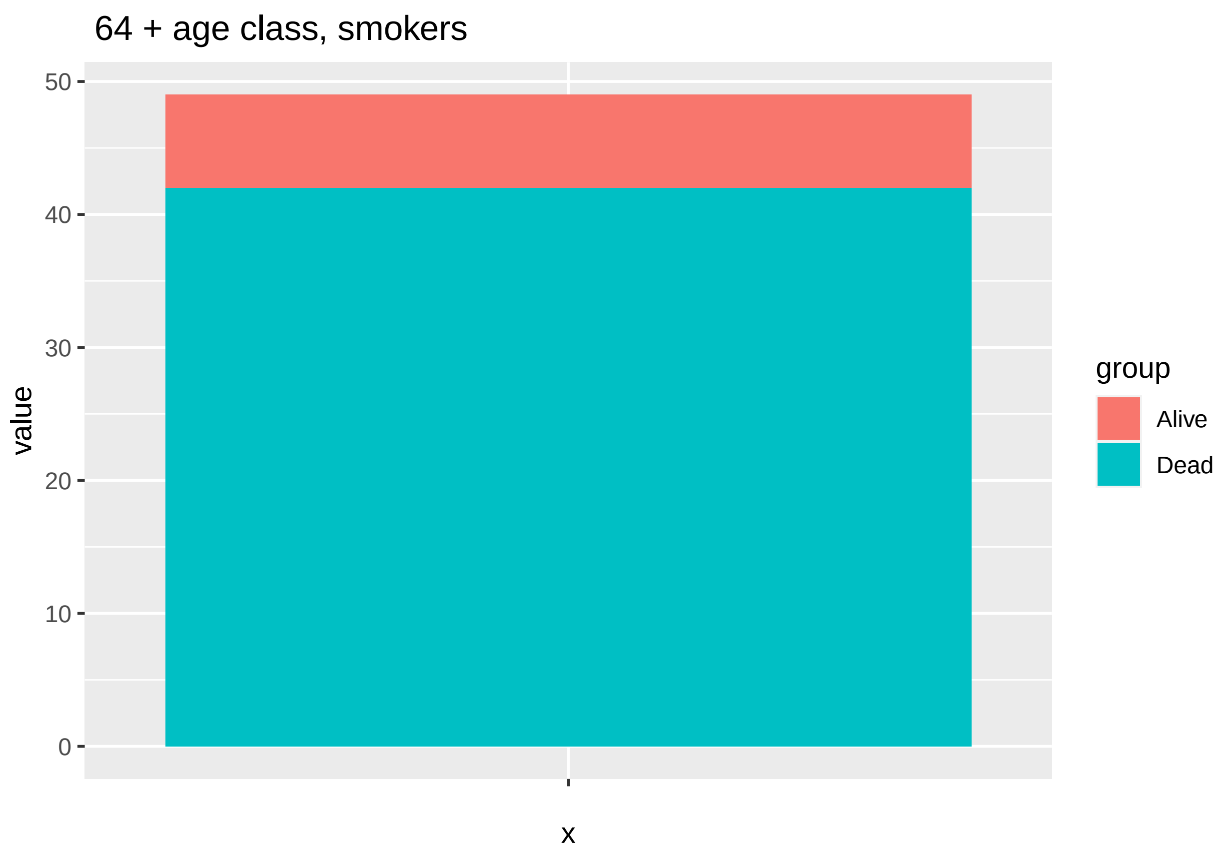

```
## [1] 0.4434783
```

```
compute_mortality_rate(smoker_arg = "No", df_arg = class5464,title_arg = "54 64 age class, non smokers")
```



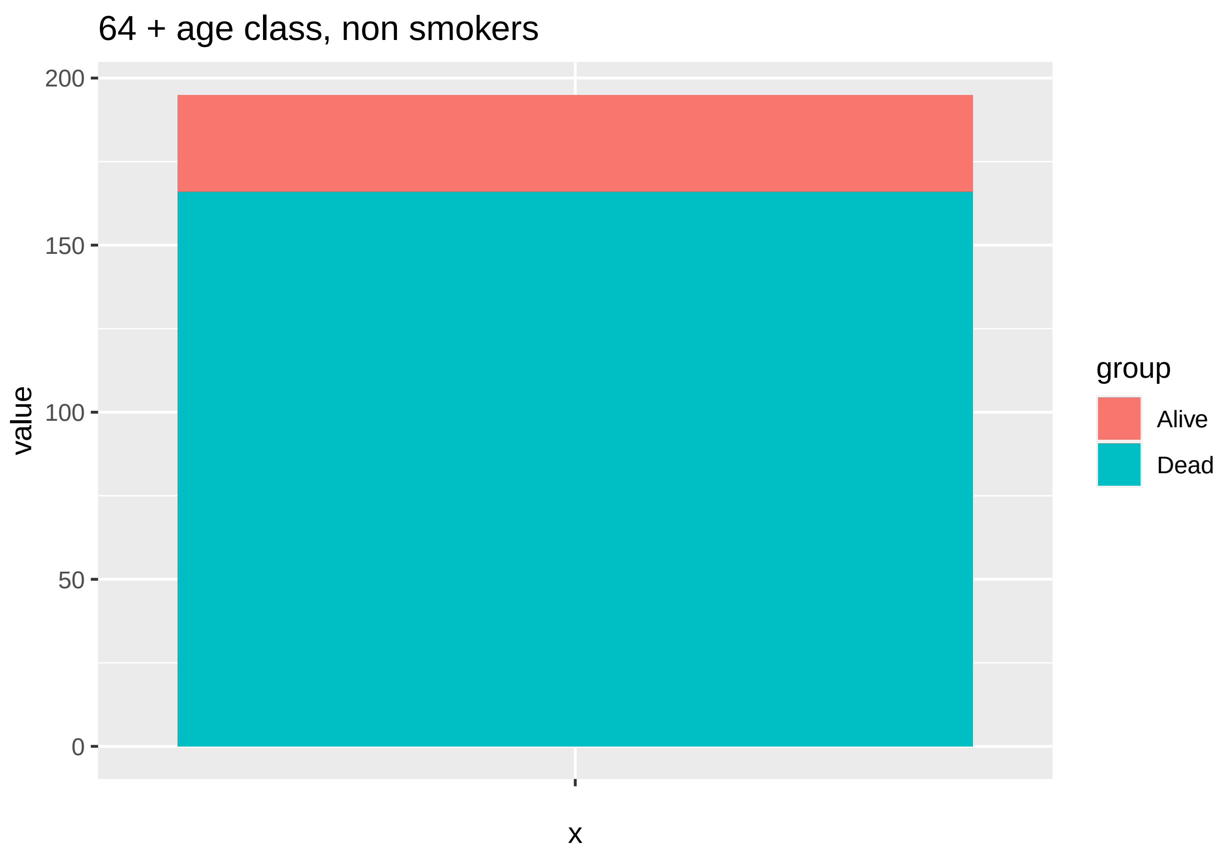

```
## [1] 0.3277311
```

```
compute_mortality_rate(smoker_arg = "Yes", df_arg = class64,title_arg = " 64 + age class, smokers")
```




```
## [1] 0.8571429
```

```
compute_mortality_rate(smoker_arg = "No", df_arg = class64,title_arg = "64 + age class, non smokers")
```




```
## [1] 0.8512821
```

This is very surprising, because, as we saw in question 1, the mortality rate was higher for *non smoker*. But now, after organizing the data in age classes, for every single class, the mortality is higher for the *smoker* group. relevant youtube video # Conclusion