lab-07-simpsons.Rmd

Ahmed adnan

17 March 2021

Packages

```
library(tidyverse)
library(mosaicData)
```

Exercises

1.

```
glimpse(Whickham)
```

```
## Rows: 1,314
## Columns: 3
## $ outcome <fct> Alive, Alive, Dead, Alive, Alive, Alive, Alive, Dead, Alive, A~
## $ smoker  <fct> Yes, Yes, Yes, No, No, Yes, Yes, No, No, No, No, Yes, No, Yes,~
## $ age     <int> 23, 18, 71, 67, 64, 38, 45, 76, 28, 27, 28, 34, 20, 72, 48, 45~
```

Your answer: The data is experimental because specific information about the people was collected.

2.

```
nrow(Whickham)
```

```
## [1] 1314
```

Your answer;There are 1,134 rows/observation in this dataset.

3.

```
ncol(Whickham)
```

```
## [1] 3
```

Your answer:there are 3 variables/columns in this dataset.

```
unique(Whickham$outcome)
```

```
## [1] Alive Dead
## Levels: Alive Dead
```

```
unique(Whickham$smoker)
```

```
## [1] Yes No
## Levels: No Yes
```

```
unique(Whickham$age)
```

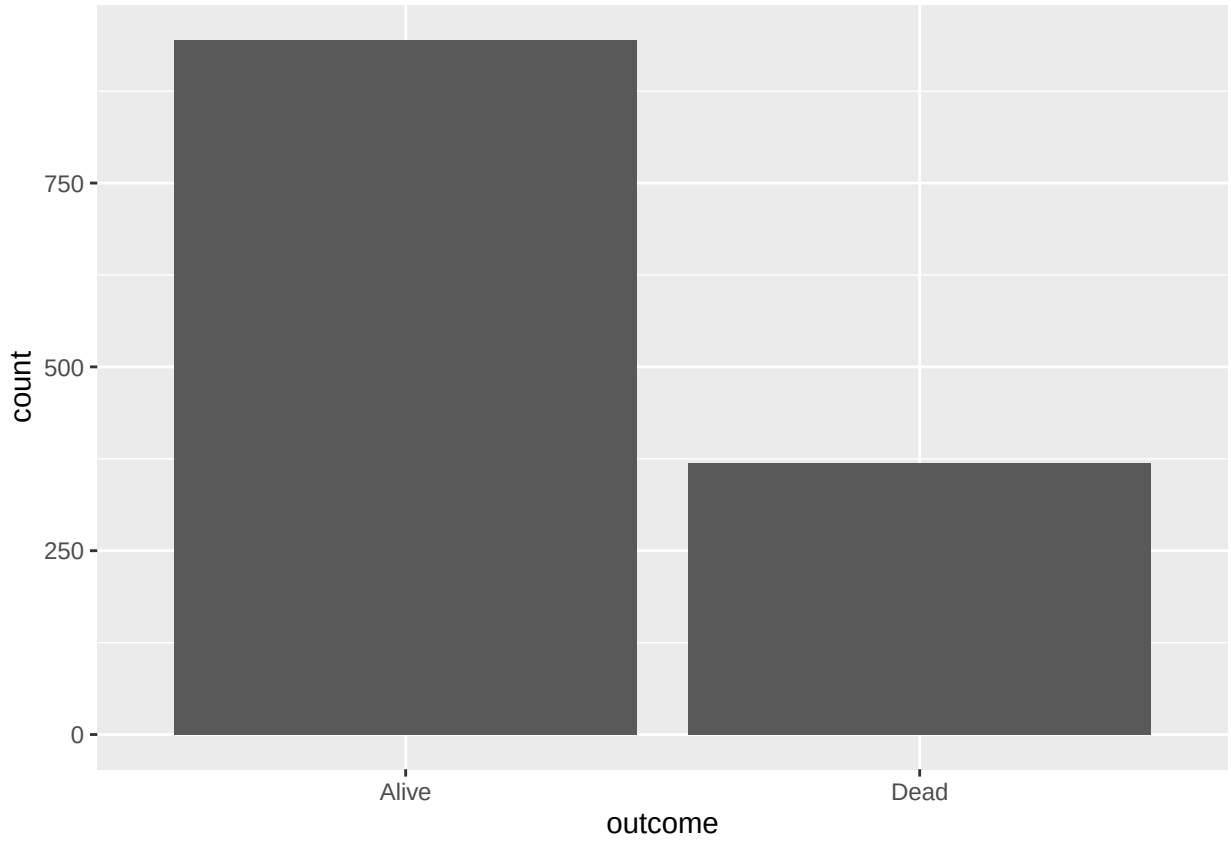
```
##  [1] 23 18 71 67 64 38 45 76 28 27 34 20 72 48 66 30 33 68 61 43 47 22 39 80 59
## [26] 56 62 51 32 60 37 36 50 55 73 52 25 53 31 54 69 79 75 21 29 24 26 49 84 40
```

```
## [51] 44 74 46 35 77 57 42 81 19 63 78 83 82 70 58 41 65
```

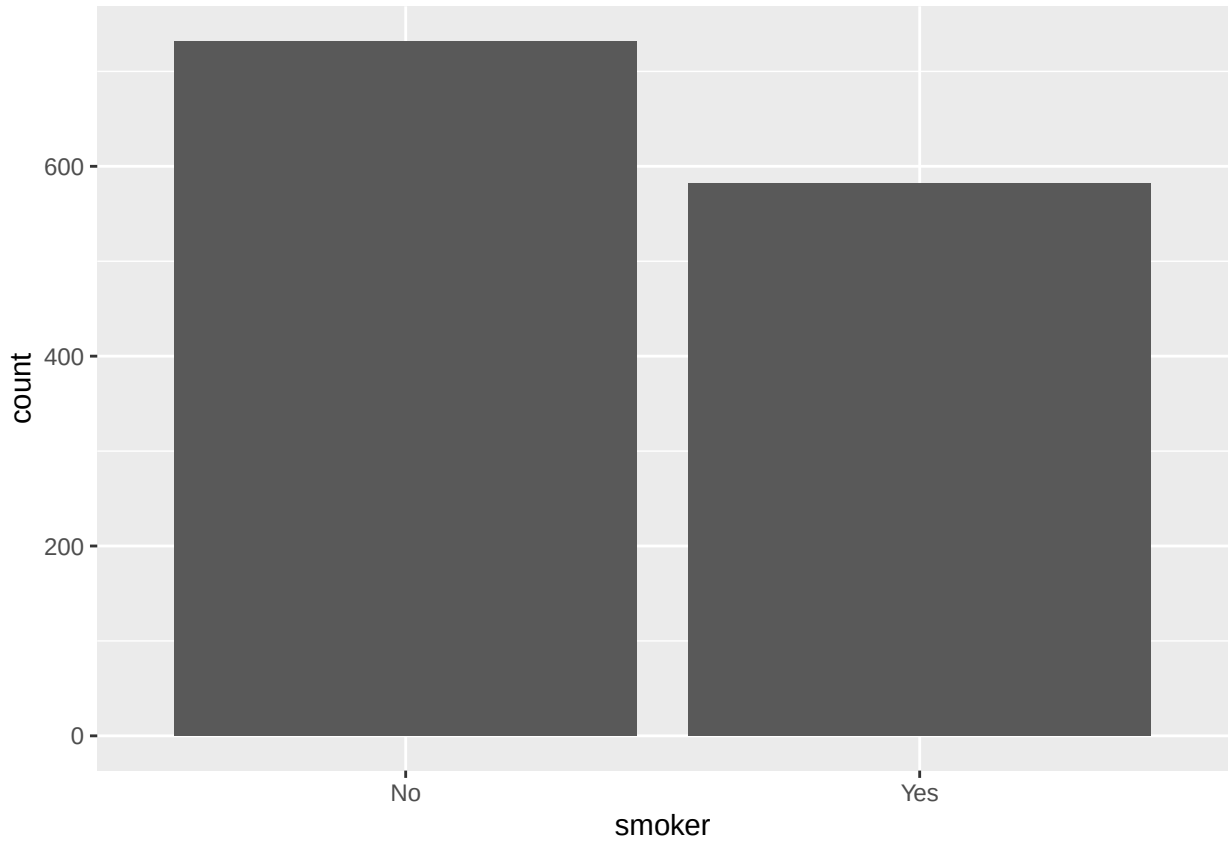
Your answer: Using the 'unique()' function on the 3 variables, we could see that "outcome" varible only takes Alive or Dead value,which makes it categorical non-ordinal. "smoker" variable only takes Yes or No, which also makes it categorical non-ordinal. Age is numerical continuous data.

One of the best ways to visualize categorical data is through the use of bar charts.

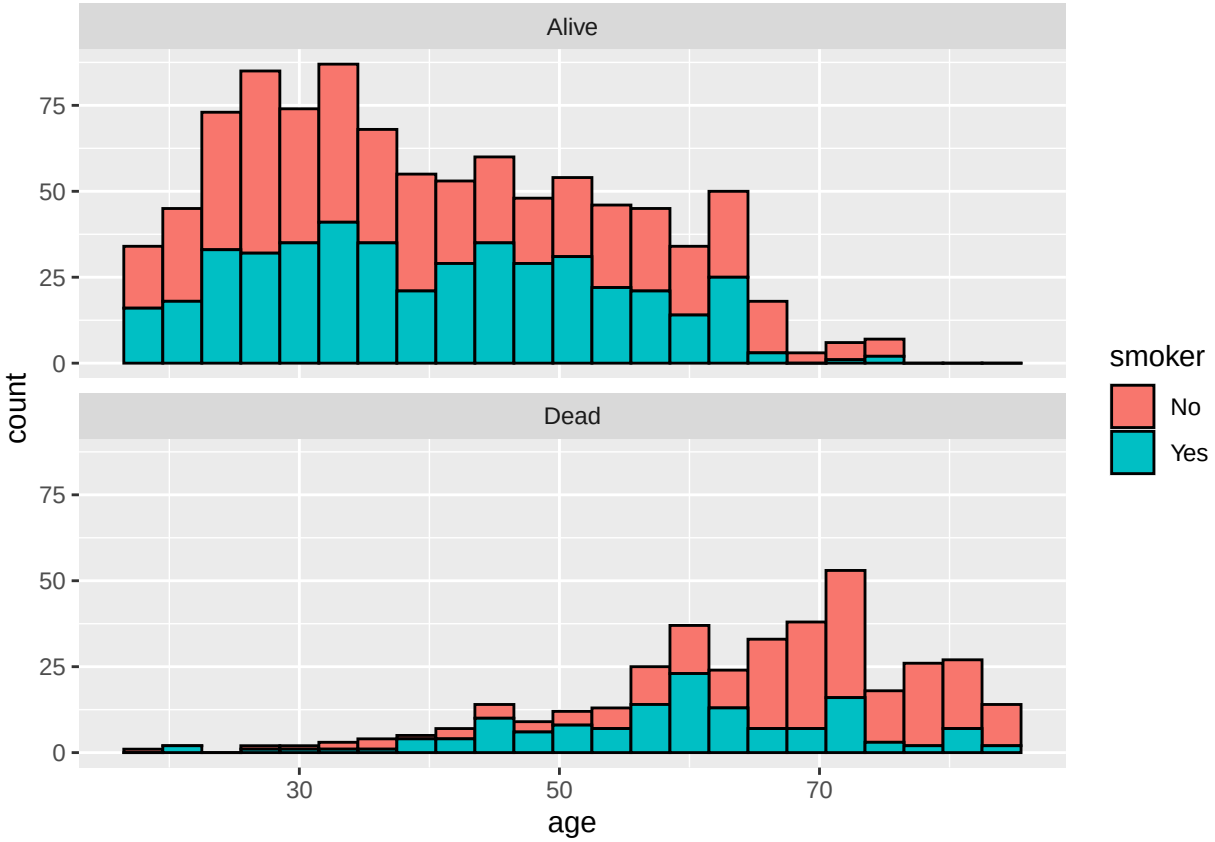
```
ggplot(Whickham, aes(x= outcome)) +
  geom_bar()
```




```
ggplot(Whickham, aes(x= smoker)) +
  geom_bar()
```




```
ggplot(Whickham, aes(x= age, fill = smoker)) +
  geom_histogram(binwidth =3, color = "black") +
  facet_wrap(~ outcome, nrow = 2)
```

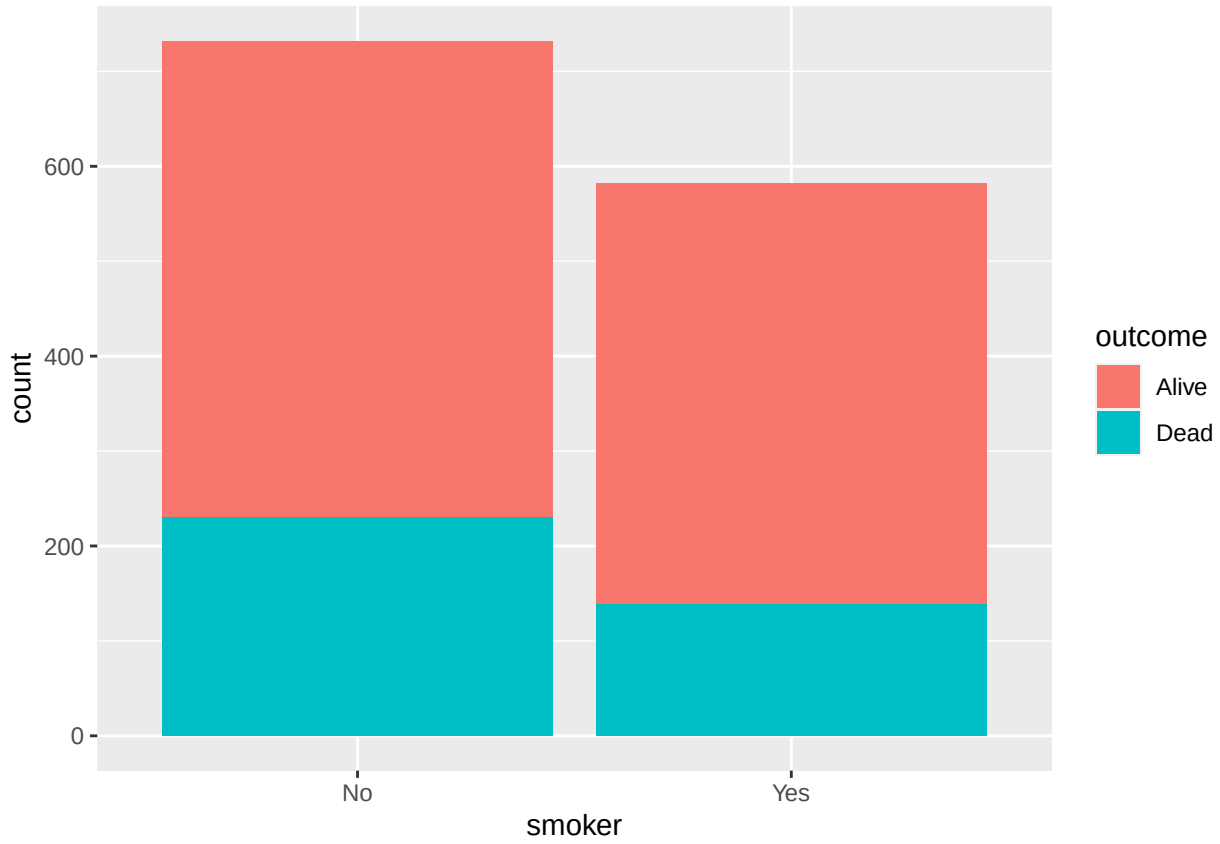


4. I expect that there will be more deaths of smoker than there will be of non-smokers..

Knit, commit, and push to github.

5. Below is the visualization depicting the relationship between smoking status and health outcome.

```
ggplot(Whickham, aes(x = smoker, fill = outcome)) +
  geom_bar()
```




```
Whickham %>%
  count(smoker, outcome) %>%
  group_by(smoker) %>%
  mutate(outcome_perc = n / sum(n)) %>%
  filter(outcome == "Dead")
```

```
## # A tibble: 2 x 4
## # Groups:   smoker [2]
##   smoker outcome     n outcome_perc
##   <fct>  <fct>   <int>        <dbl>
## 1 No     Dead      230        0.314
## 2 Yes    Dead      139        0.239
```

31.4% of smokers had died by the follow-up and 23.9% of non-smoker had dide by the follow-up after the same period . Itappears that, contrary to what I expected, smokers survived the 20 years follow-up more than non-smokers..

6.
7.

Knit, commit, and push to github.