lab-07-simpsons.Rmd

shahad abdulah al- mutairi

17 March 2021

Packages

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

Exercises

1.

```
?Whickham
```

Your answer: The data is observational as the description states that is based on age, smoking, and mortality, which are all observable events and not produced via experiments.

2.

```
nrow(Whickham)
```

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

Your answer;There ane 1,314 observations. As we know every row is an observation.

3.

```
names(Whickham)
```

```
## [1] "outcome" "smoker"  "age"
```

Your answer: There are 3 variables, “outcome” “smoker”, and “age”

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

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

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

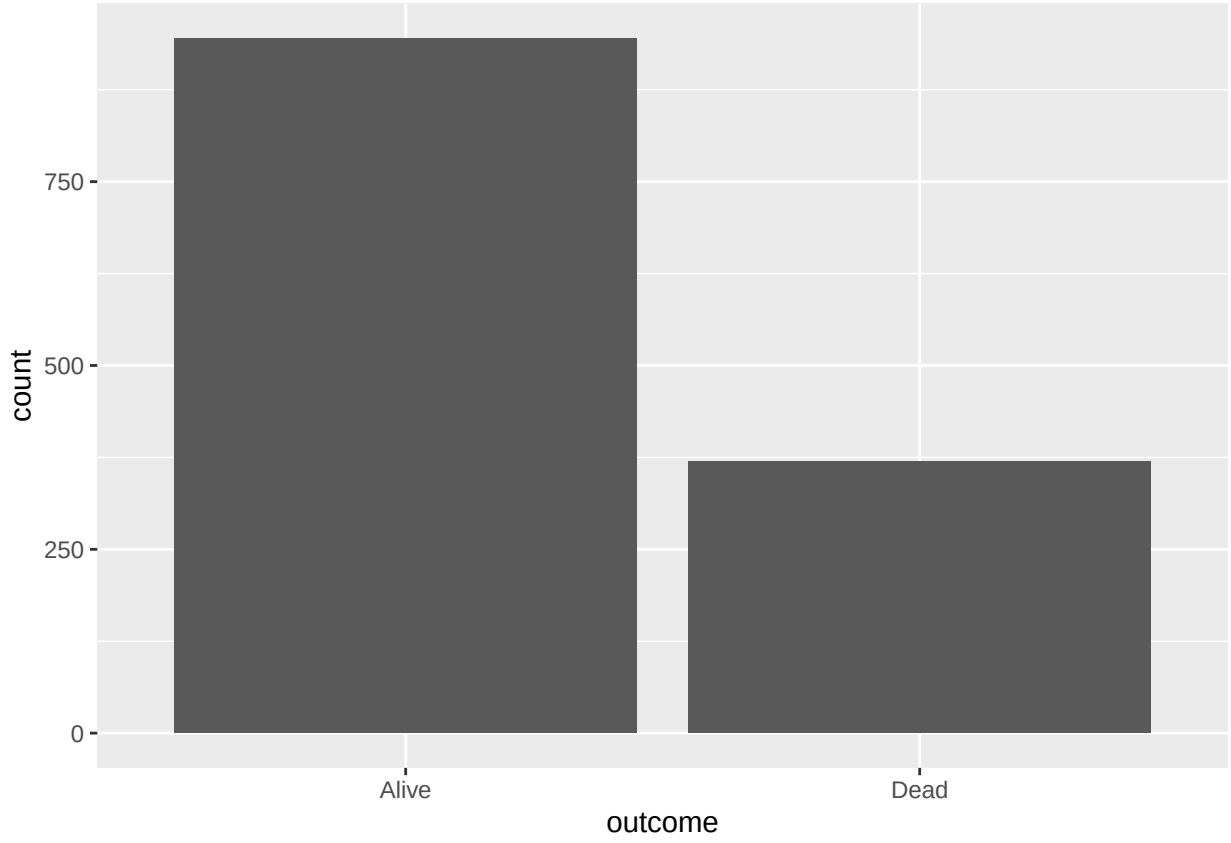
```
## [1] Yes No
## Levels: No Yes
```

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

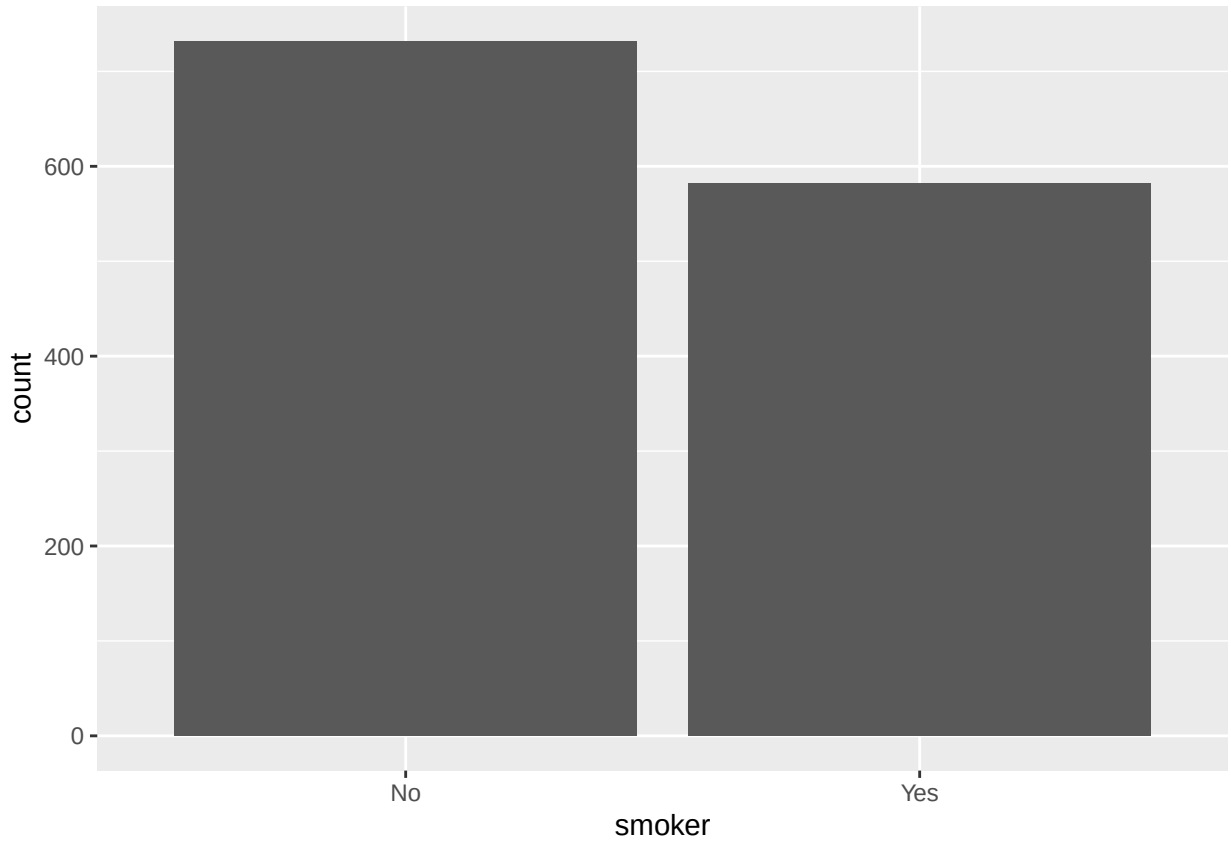
```
##  [1] 23 18 71 67 64 38 45 76 28 27 34 20 72 48 66 30 33 68 61 43 47 22 39 80 59
## [26] 56 62 51 32 60 37 36 50 55 73 52 25 53 31 54 69 79 75 21 29 24 26 49 84 40
## [51] 44 74 46 35 77 57 42 81 19 63 78 83 82 70 58 41 65
```

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

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

 geom_bar()
```

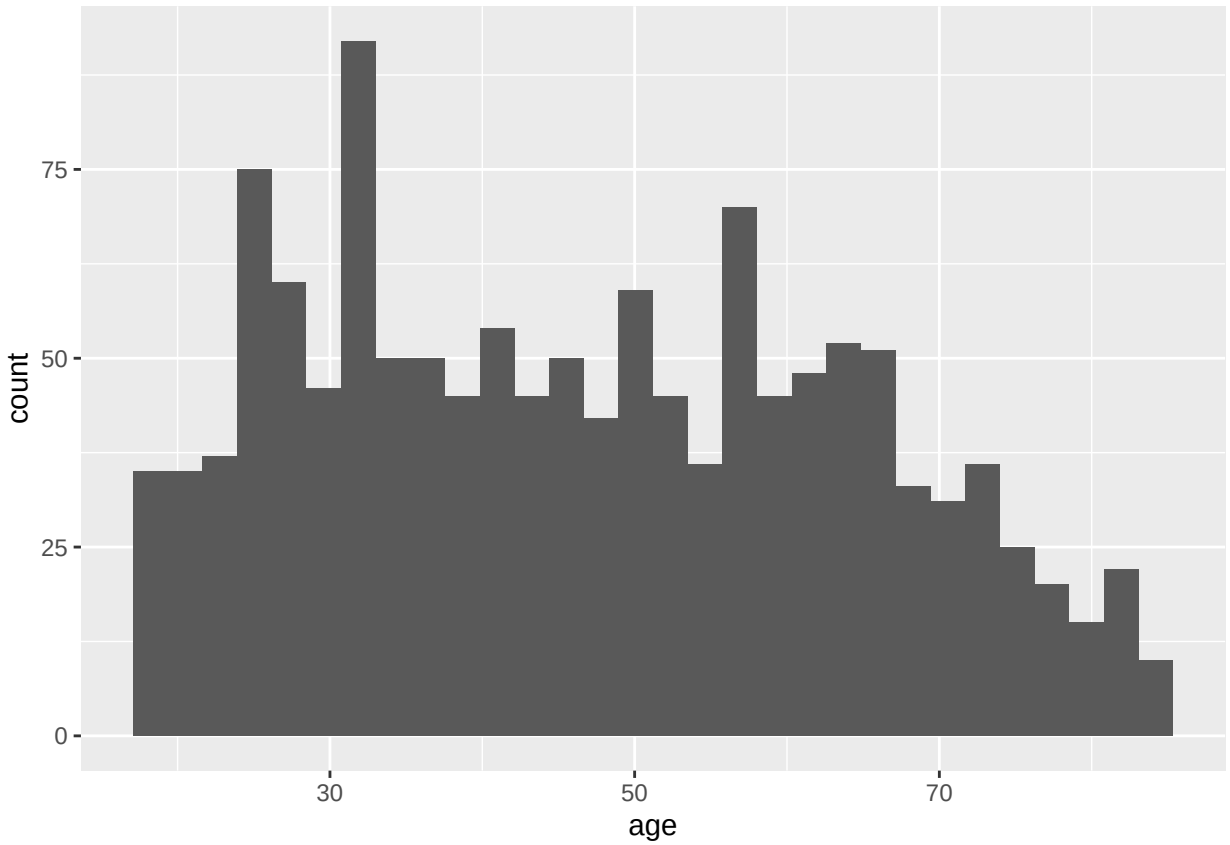



```
ggplot(Whickham, aes(x = smoker)) +

 geom_bar()
```




```
ggplot(Whickham, aes(x = age)) +

 geom_histogram()
```

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



4 Before plotting the correlation between smoking status and health outcomes.I would like to assume there is a strong correlation between the two.based on the well-known impact of smoking on the health.

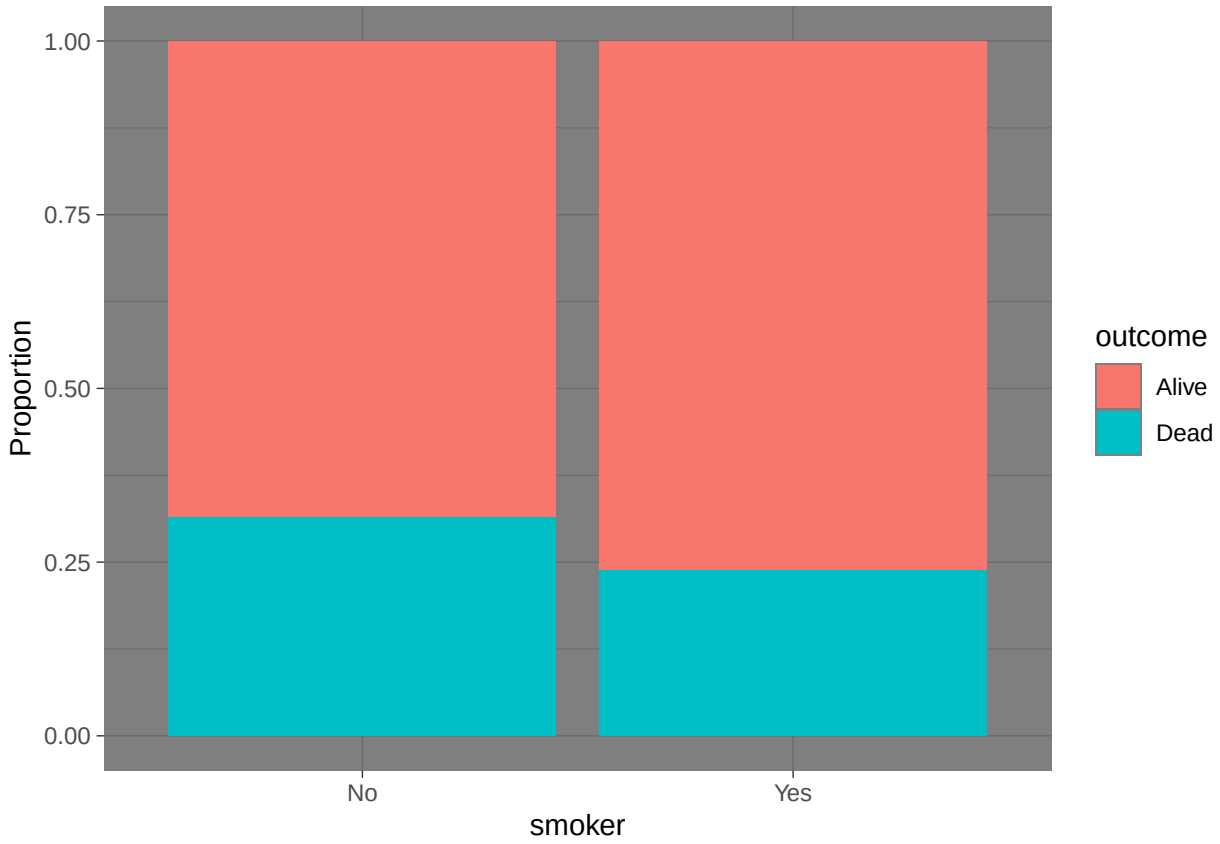
Knit, commit, and push to github.

5.

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

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

```
ggplot(Whickham, aes(x=smoker, fill=outcome)) +
  geom_bar(position = "fill") + labs(y="Proportion") +
  theme_dark()
```



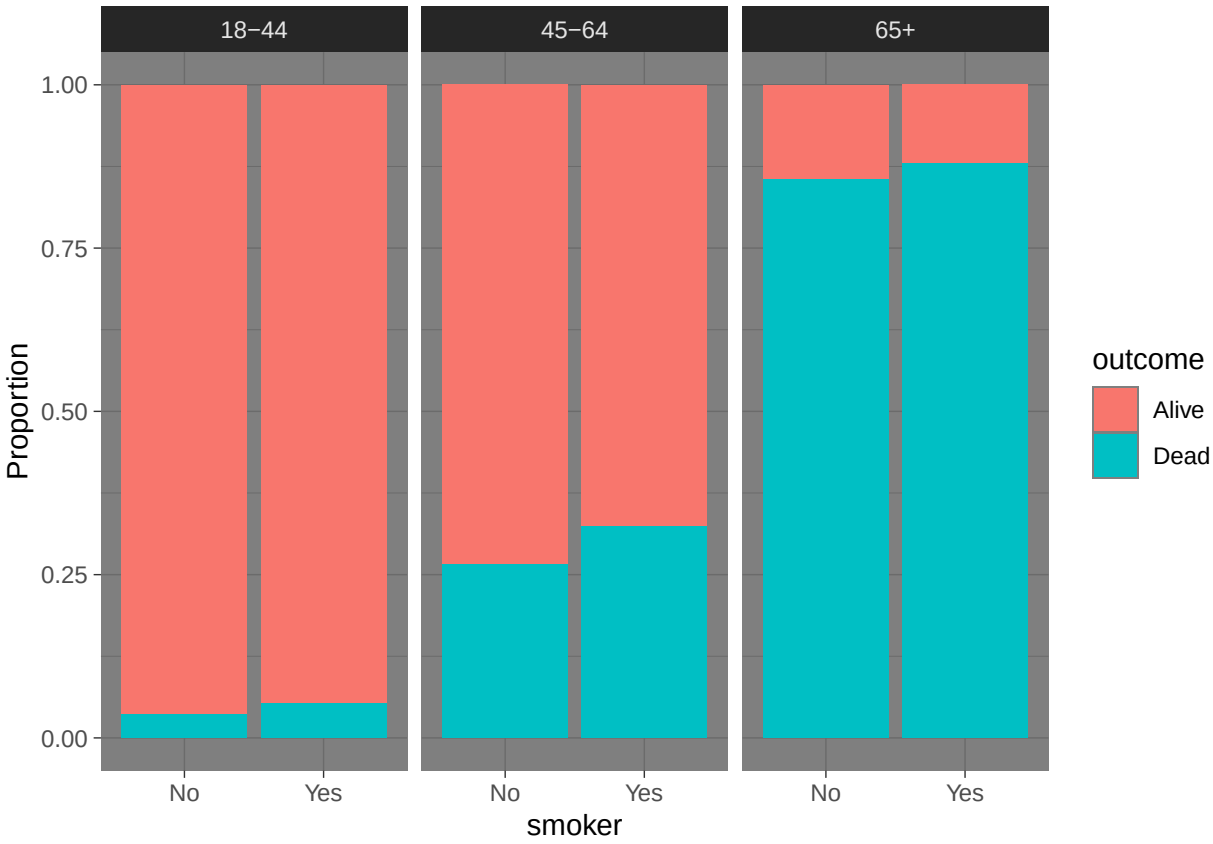
Insights: First, what the data shows is unrealistic.but the data never lies, and we should always belive it.the data shows that people who smokes are more likely to live longer than the people who dont.we need to investigate the data further.

6.

```
Whickham <- Whickham%>% mutate (age_cat = case_when (age <= 44 ~ "18-44", age > 44. & age <= 64 ~ "45-64
```

7.

```
ggplot(Whickham, aes(x=smoker, fill=outcome)) + geom_bar(position = "fill") +
  labs(y="Proportion") + facet_grid(. ~ age_cat) + theme_dark()
```



Observations:

by distributing the data by age group we find people who smoke are more likely to die sooner than the people who don't - Smoking has a negitive impact on humen health. - The impact of smoking does'nt greatly effect the human health until the late 60's.

Knit, commit, and push to github.