

# lab-07-simpsons.Rmd

Asrar Almutairi

17 March 2021

## Packages

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

## Exercises

1.

```
?Whickham
```

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

2.

```
nrow(Whickham)
```

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

Your answer; There are 1,314 obs. As we Know every row is an obs

3.

```
names(Whickham)
```

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

Your answer:

There are 3 variables, “outcome”, “smoker”, “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
```

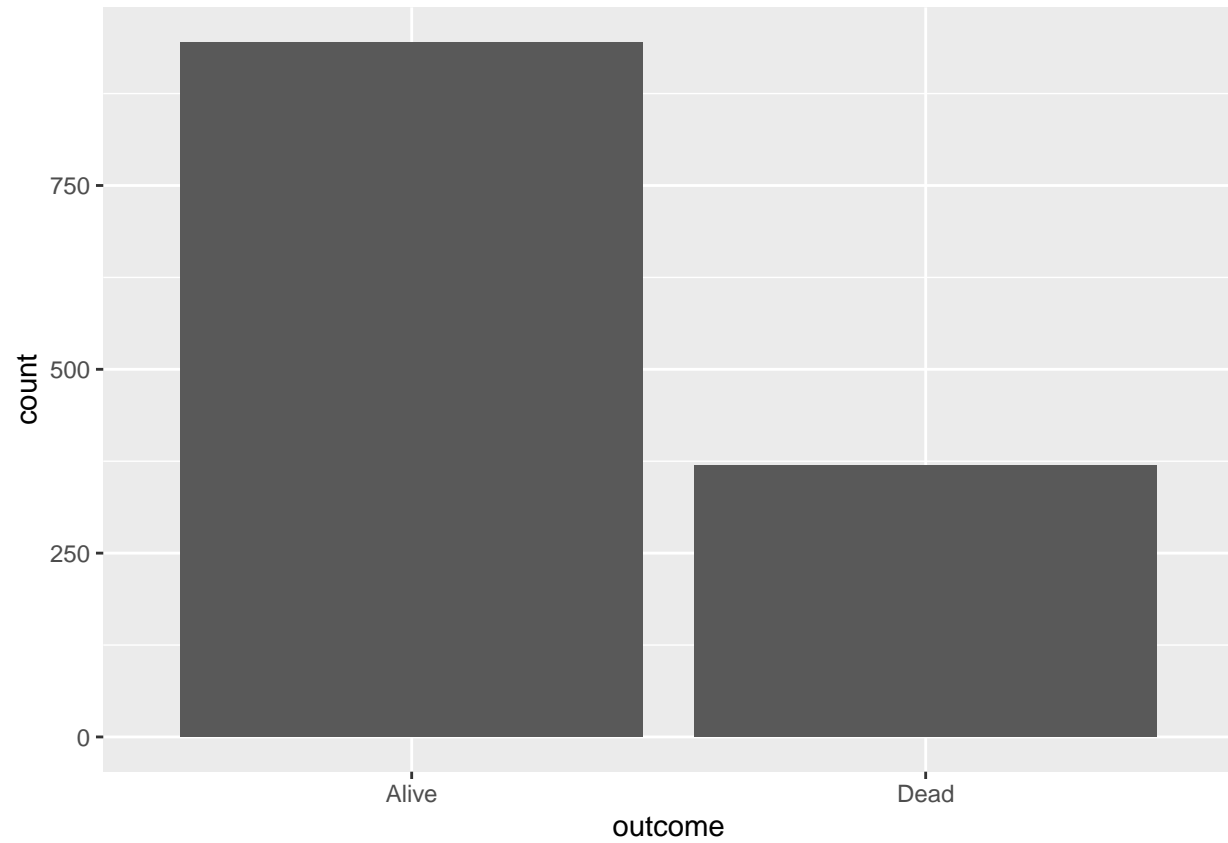
```
## [51] 44 74 46 35 77 57 42 81 19 63 78 83 82 70 58 41 65
```

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

categorical non-ordinal. Age is numerical continuous data.

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

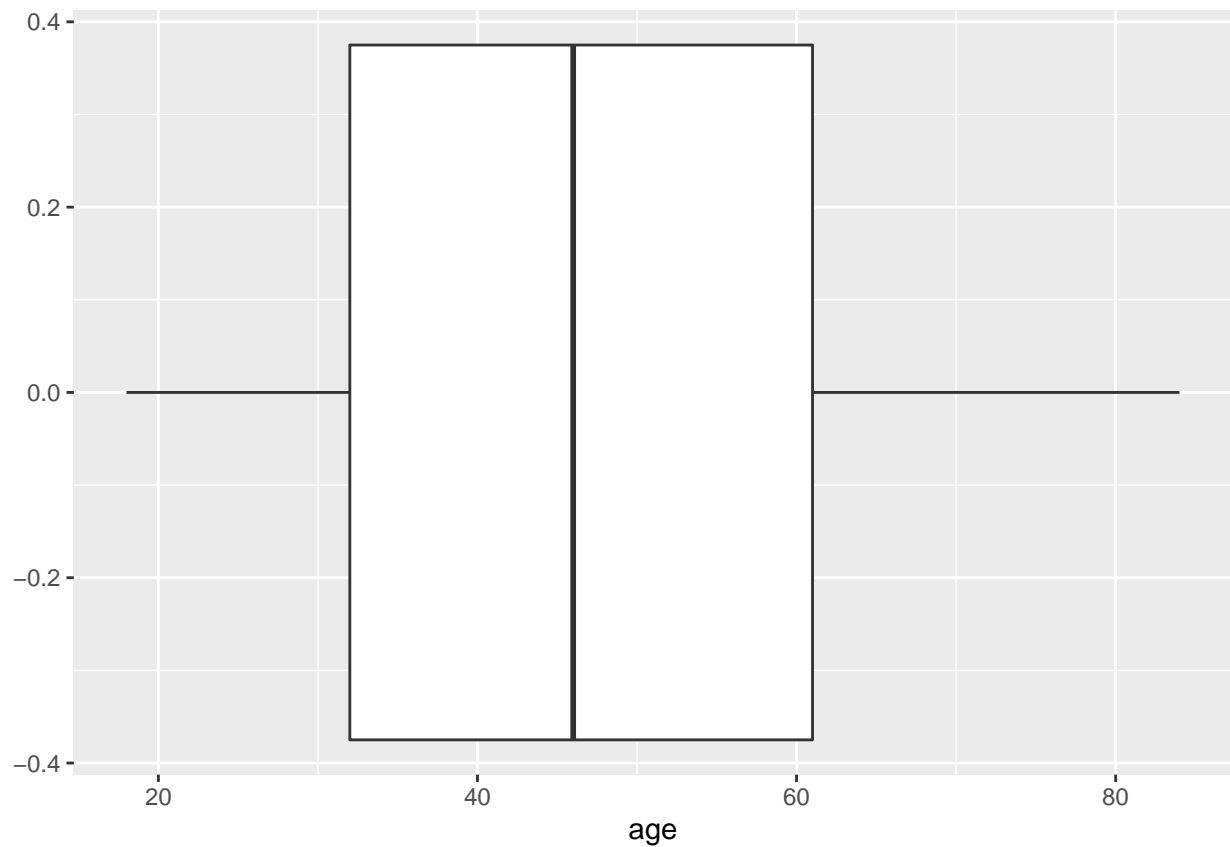
```
ggplot(Whickham, aes(x = outcome)) +  
  geom_bar()
```



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

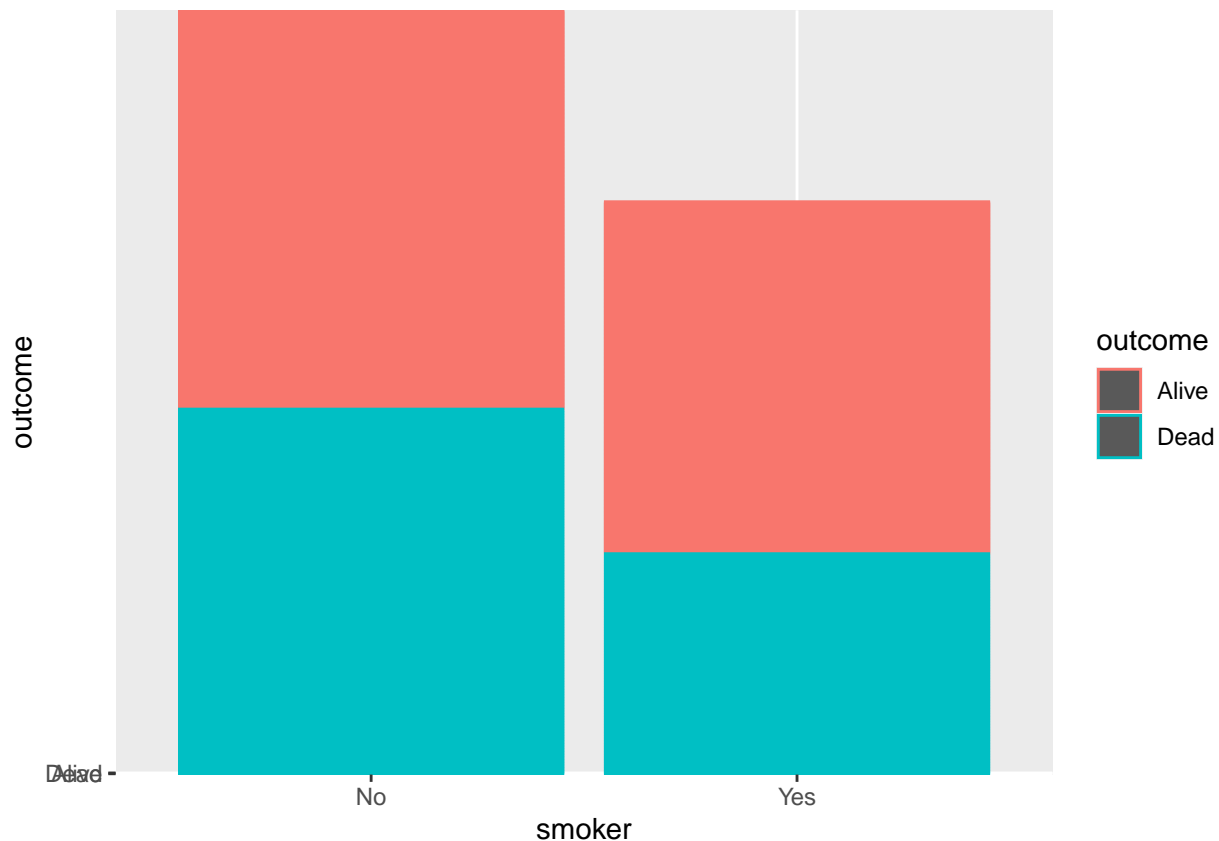


```
ggplot(Whickham, aes(x = age)) +  
  geom_boxplot()
```



4. Smoking is a bad habit if it continues. It can lead to death and increase the proportion.

```
ggplot(data=Whickham, aes(x=smoker, y=outcome, color=outcome)) + geom_bar(stat="identity")
```



Knit, commit, and push to github.

5. smoker no (732) : 31,4 (dead) » (68,6) alive smoker yes (582) : 23,8 (dead) » (76,2) alive

```
Whickham %>%
  count(smoker, outcome)
```

```
##   smoker outcome    n
## 1     No    Alive 502
## 2     No    Dead 230
## 3     Yes   Alive 443
## 4     Yes   Dead 139
```

```
502+230
```

```
## [1] 732
```

```
230/732
```

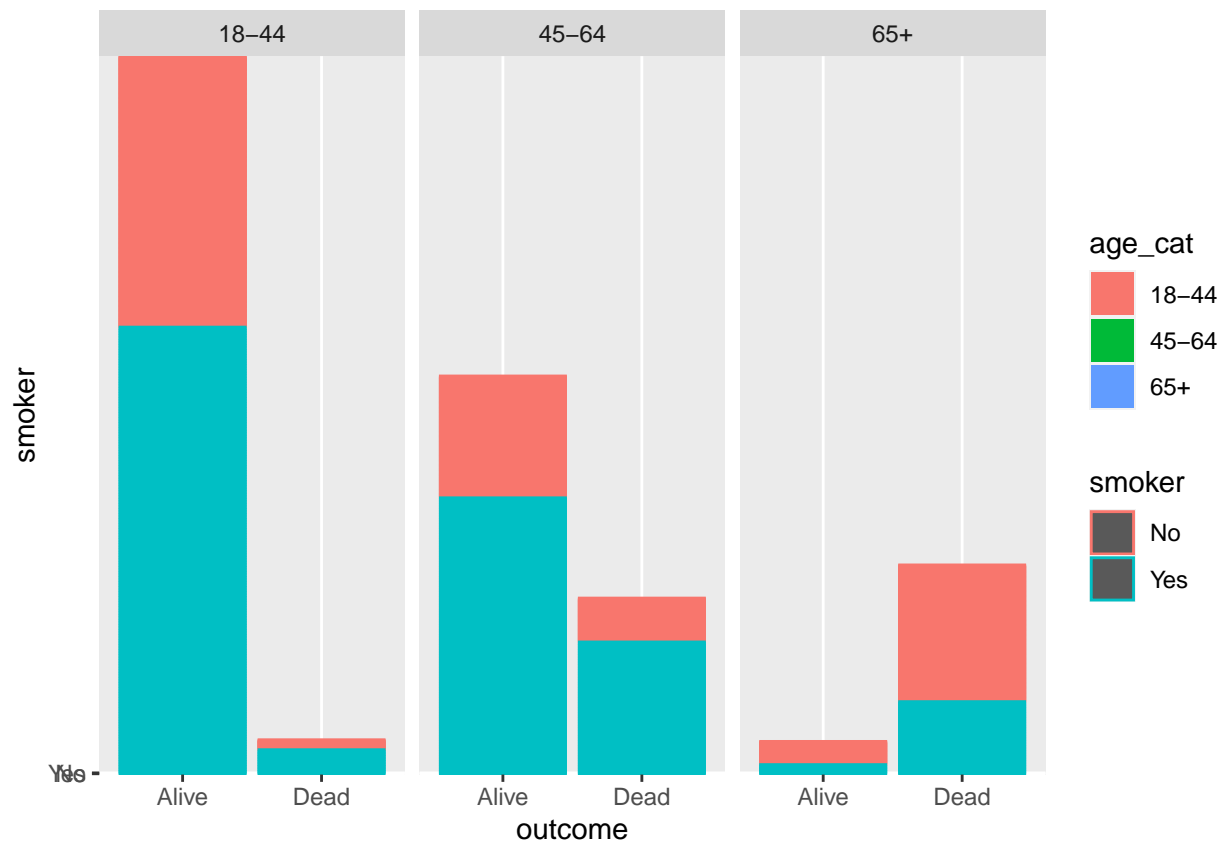
```
## [1] 0.3142077
```

6.

```
Whickham <- Whickham%>% mutate (age_cat = case_when (age <= 44 ~ "18-44", age > 44. & age <= 64 ~ "45-64", age > 64 ~ "65+"))
```

7. what changes > the category of the age it's appear to us and we see the most of dead people not smoker in age (65+).. but in age (45-64)and (18-44)the most dead people are smoker that is relationship between the smoking and helth not clearly but can say that your helth will be change to worst if you be smoker.

```
ggplot(data=Whickham, aes(x=outcome, y=smoker,color=smoker, fill=age_cat)) + geom_bar(stat="identity")
```



Knit, commit, and push to github.