

# “COMPSCIX 415.2 Homework 6”

*Santosh Kanutala*

*7/13/2018*

## Contents

Github location . . . . .	2
Exercise 1 . . . . .	2
Question: 1 . . . . .	2
Question: 2 . . . . .	2
Question: 3 . . . . .	2
Question: 4 . . . . .	3
Exercise 2 . . . . .	5
Question: 1 . . . . .	5
Question: 2 . . . . .	5
Question: 3 . . . . .	5
Question: 4 . . . . .	5
Question: 5 . . . . .	5
Question: 6 . . . . .	5
Question: 7 . . . . .	5
Question: 8 . . . . .	5

## Github location

My homework assignments can be found at <https://github.com/santumagic/compscix-415-2assignments.git>

## Exercise 1

### Question: 1

```
# Load the required packages
library(tidyverse)
library(mdsr)
library(mosaicData)
# glimpse the given dataset
glimpse(Whickham)
```

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

### Answer:

Below are the three variables from the Whickham dataset.

- outcome
- smoker
- age

### Question: 2

### Answer:

There are 1314 observations. Each observation represents a person. The data set indicates if the individual is a smoker or not, current age an individual, and if the individual is alive or dead.

### Question: 3

```
library(mosaicData)
library(tidyverse)
Whickham %>% count( smoker , outcome )
```

```
## # A tibble: 4 x 3
##   smoker outcome     n
##   <fct>   <fct> <int>
## 1 No     Alive    502
## 2 No     Dead     230
## 3 Yes    Alive    443
## 4 Yes    Dead     139
```

### Answer:

By looking at the above table, it is difficult to conclude anything, so I calculate the proportions first by using the below code.

```
Whickham_proportions <- Whickham %>% group_by(smoker,outcome) %>%
  summarize(n = n()) %>%
  mutate ( prop = n/sum(n))
Whickham_proportions
```

```
## # A tibble: 4 x 4
## # Groups:   smoker [2]
##   smoker outcome      n prop
##   <fct>   <fct>   <int> <dbl>
## 1 No     Alive     502 0.686
## 2 No     Dead      230 0.314
## 3 Yes    Alive     443 0.761
## 4 Yes    Dead      139 0.239
```

By looking at the above proportions it is observed that, 31.4 % of non smokers are dead and 23.8 % of smokers are dead which means more healthy people are dead. So there must be other reasons for deaths or data might be wrong. So the data doesn't make any sense.

#### Question: 4

```
# creating the age groups column
Whickham_factor <- Whickham %>% mutate (category =
                                     factor (
case_when(age <= 44 ~ "age <= 44", age > 44 & age <= 64 ~ "age > 44 & age <= 64",
          age > 64 ~ "age > 64")))

head(Whickham_factor) # display the top rows of the result dataset
```

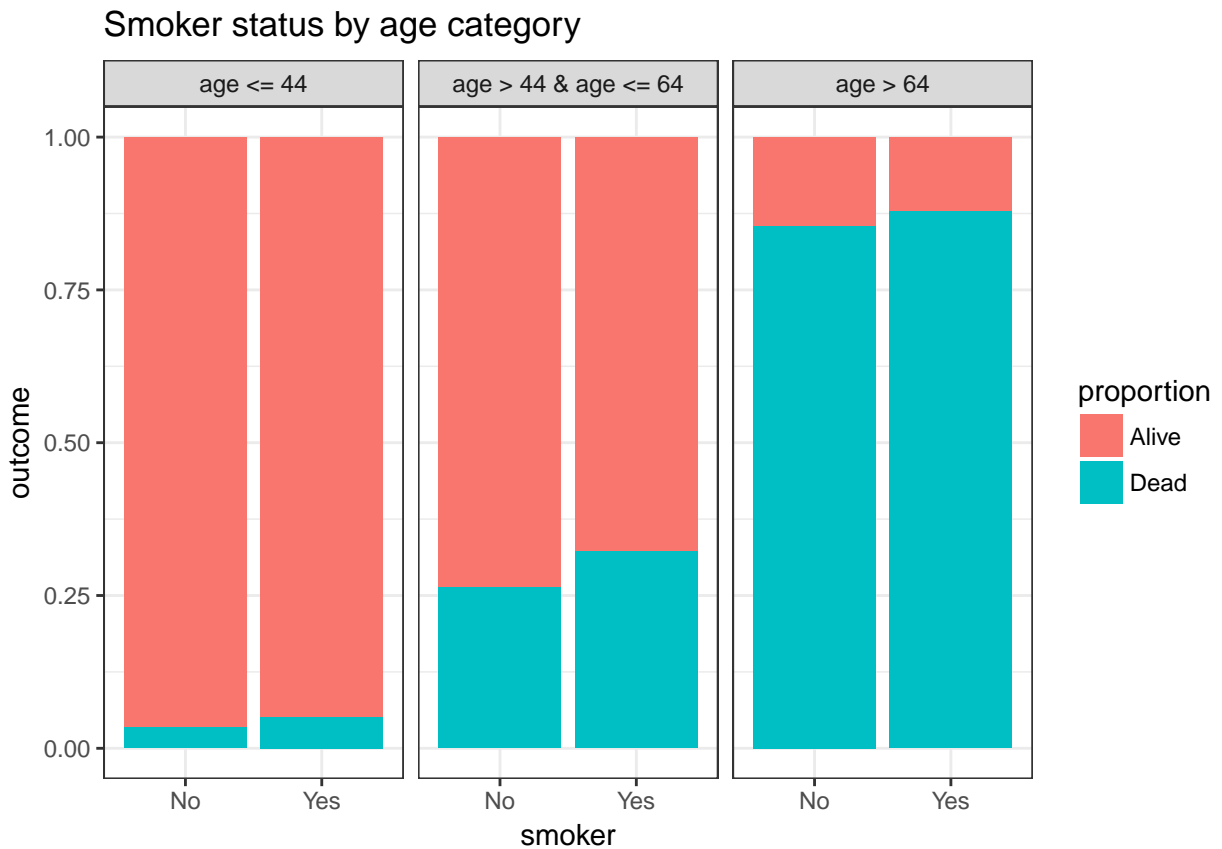
```
##   outcome smoker age      category
## 1  Alive   Yes  23   age <= 44
## 2  Alive   Yes  18   age <= 44
## 3  Dead   Yes  71   age > 64
## 4  Alive   No  67   age > 64
## 5  Alive   No  64 age > 44 & age <= 64
## 6  Alive   Yes  38   age <= 44
```

```
# reorganize the data by grouping, summarise the data and finding the proportions
Whickham_cat <- Whickham_factor %>%
group_by(category,smoker,outcome) %>% summarise( n = n()) %>%
mutate (proportion = n/sum(n))

head(Whickham_cat) # display the top rows of the result dataset
```

```
## # A tibble: 6 x 5
## # Groups:   category, smoker [3]
##   category      smoker outcome      n proportion
##   <fct>         <fct>   <fct>   <int>     <dbl>
## 1 age <= 44     No     Alive     327     0.965
## 2 age <= 44     No     Dead       12     0.0354
## 3 age <= 44     Yes    Alive     270     0.947
## 4 age <= 44     Yes    Dead       15     0.0526
## 5 age > 44 & age <= 64 No     Alive     147     0.735
## 6 age > 44 & age <= 64 No     Dead       53     0.265
```

```
# create the visualization with the above result set and facet on age categories
Whickham_cat %>%
ggplot() +
  geom_bar(aes(x = smoker, y = proportion, fill = outcome, label = round(proportion,2)),
    stat = 'identity', position = 'fill') +
  labs(x = 'smoker', y = 'outcome', fill = 'proportion',
    title = 'Smoker status by age category') +
  facet_grid(~ category) +
  theme_bw()
```



**Answer:**

From the above dataset it is observed that until the age of 44, non-smokers have only a 1% advantage compared to smokers, but this gap increases dramatically between the ages 44 & 64 where 6% more non-smokers are alive than smokers. Beyond 64 year of age, the difference drops to 3% with non-smokers still being alive more often than smokers.

## Exercise 2

Question: 1

Question: 2

Question: 3

Question: 4

Question: 5

Question: 6

Question: 7

Question: 8