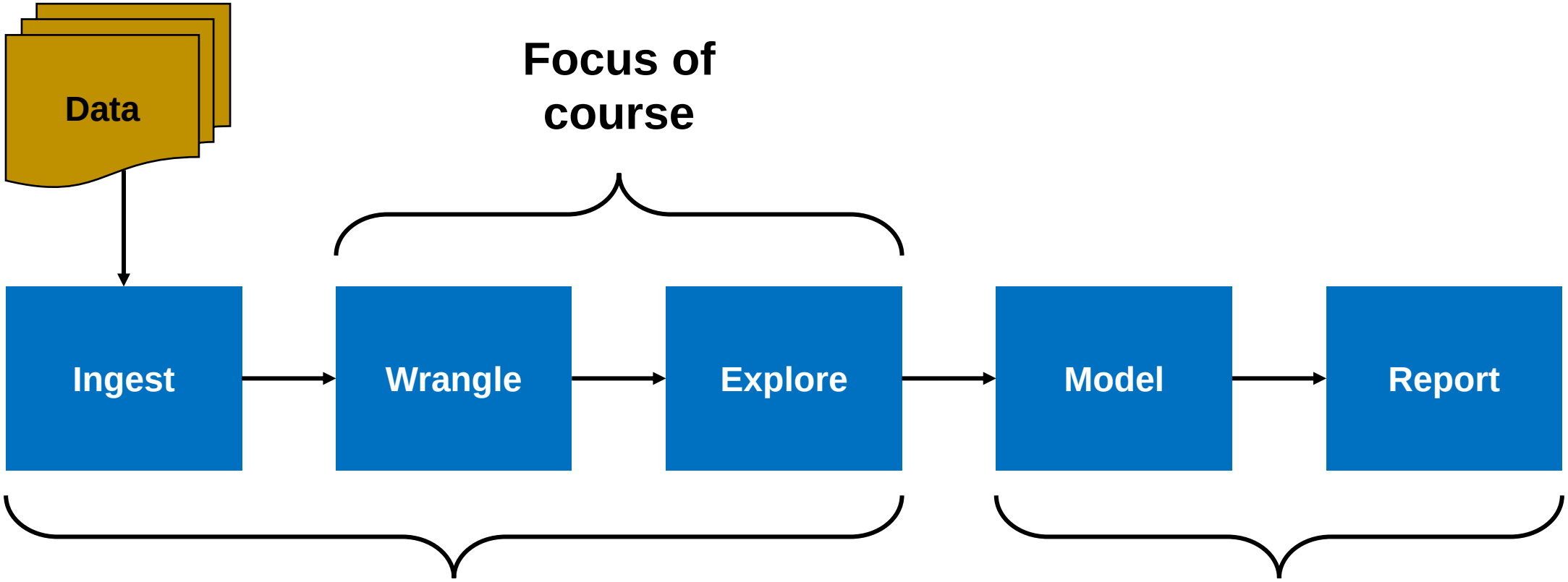
The Data Science Pipeline



`read_csv`

`dplyr`
`tidyr`
`stringr`
`tidytext`

`ggplot2`
`word.cloud`

Exploring the Titanic Dataset with R

- We will use the data wrangling and visualization skills to build an EDA on the titanic dataset

- The objective of the analysis is to understand the attributes that made it less/more likely to survive the disaster

Task # 1 (10 minutes)

- Download the titanic dataset from Blackboard (week 8)
- Load the dataset into R using `read_csv`
- Spend sufficient time inspecting the data (`str` or `glimpse` are handy)

Task # 1 Discussion

```
> glimpse(titanic_df)
Rows: 891
Columns: 12
$ PassengerId <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 2~
$ Survived    <dbl> 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 1, 0, 1, 0, 1, 0, 1, 1, 1, 0, 1, 0, 0, 1, 0~
$ Pclass      <dbl> 3, 1, 3, 1, 3, 3, 1, 3, 3, 2, 3, 1, 3, 3, 3, 2, 3, 2, 3, 3, 2, 2, 3, 1, 3, 3, 3, 1, 3, 3~
$ Name        <chr> "Braund, Mr. Owen Harris", "Cumings, Mrs. John Bradley (Florence Briggs Thayer)", "Heikk~
$ Gender      <chr> "male", "female", "female", "female", "male", "male", "male", "male", "female", "female"~
$ Age         <dbl> 22, 38, 26, 35, 35, NA, 54, 2, 27, 14, 4, 58, 20, 39, 14, 55, 2, NA, 31, NA, 35, 34, 15,~
$ SibSp       <dbl> 1, 1, 0, 1, 0, 0, 0, 3, 0, 1, 1, 0, 0, 1, 0, 0, 4, 0, 1, 0, 0, 0, 0, 0, 3, 1, 0, 3, 0, 0~
$ Parch       <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 2, 0, 1, 0, 0, 5, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 5, 0, 2, 0, 0~
$ Ticket      <chr> "A/5 21171", "PC 17599", "STON/O2. 3101282", "113803", "373450", "330877", "17463", "349~
$ Fare        <dbl> 7.2500, 71.2833, 7.9250, 53.1000, 8.0500, 8.4583, 51.8625, 21.0750, 11.1333, 30.0708, 16~
$ Cabin       <chr> NA, "C85", NA, "C123", NA, NA, "E46", NA, NA, NA, "G6", "C103", NA, NA, NA, NA, NA, NA, ~
$ Embarked    <chr> "S", "C", "S", "S", "S", "Q", "S", "S", "S", "C", "S", "S", "S", "S", "S", "S", "Q", "S"~
>
```

- Do you have good understanding of the variables based on the column names?
- Are the data types all appropriate? What would you change?
- Are the column names all appropriate? What would you change?
- Anything else that you notice?

Task # 2 (10 minutes)

Make the following changes to the dataset:

- Convert “Pclass” and “Survived” to a factor data type
- For “Embarked” map:
  - “S” to “Southampton”
  - “C” to “Cherbourg”
  - “Q” to “Queenstown”

Task # 2 Discussion

```
titanic_df <- read_csv("titanic.csv") %>%
                mutate(Survived = ifelse(Survived==1, "Yes", "No")) %>%
                mutate(Pclass = factor(Pclass)) %>%
                mutate(Embarked = case_when(
                  Embarked == "S" ~"Southampton",
                  Embarked == "Q" ~"Queenstown",
                  Embarked == "C" ~"Cherbourg"
                ))
```

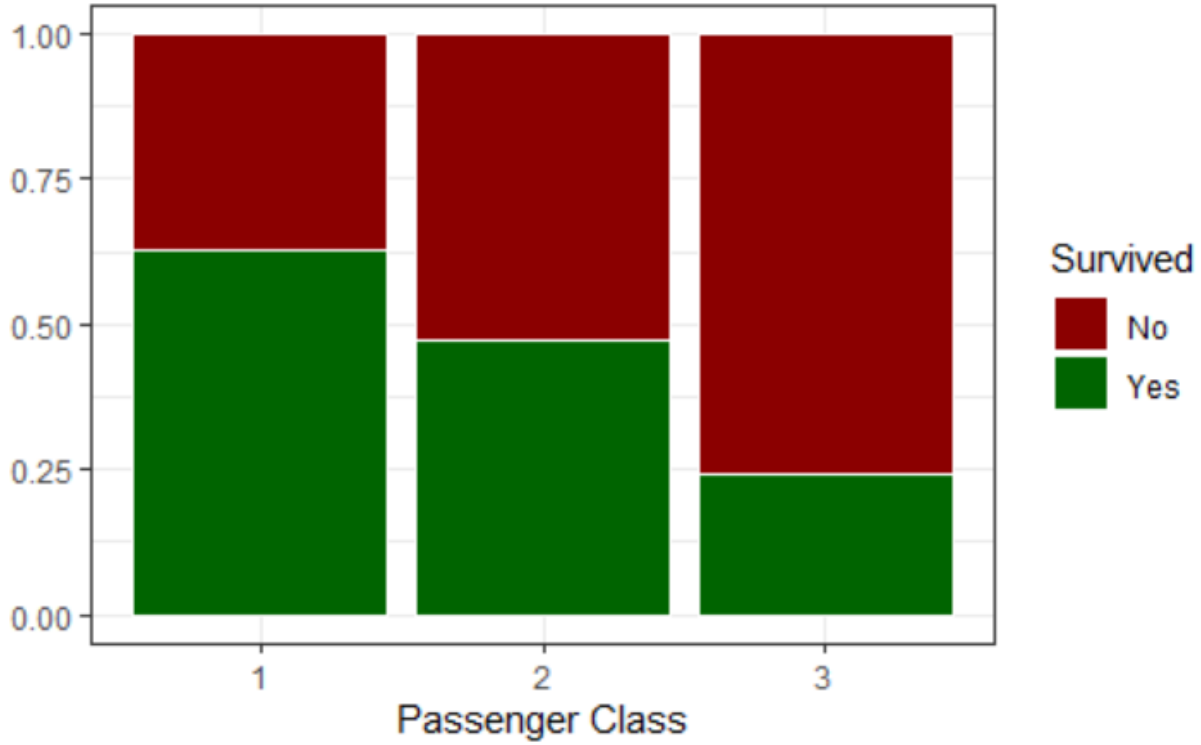
```
> glimpse(titanic_df)
Rows: 891
Columns: 12
$ PassengerId <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 2~
$ Survived    <chr> "No", "Yes", "Yes", "Yes", "No", "No", "No", "No", "Yes", "Yes", "Yes", "Yes", "No", "No~
$ Pclass      <fct> 3, 1, 3, 1, 3, 3, 1, 3, 3, 2, 3, 1, 3, 3, 3, 2, 3, 2, 3, 3, 2, 2, 3, 1, 3, 3, 3, 1, 3, 3~
$ Name        <chr> "Braund, Mr. Owen Harris", "Cumings, Mrs. John Bradley (Florence Briggs Thayer)", "Heikk~
$ Gender      <chr> "male", "female", "female", "female", "male", "male", "male", "male", "female", "female"~
$ Age         <dbl> 22, 38, 26, 35, 35, NA, 54, 2, 27, 14, 4, 58, 20, 39, 14, 55, 2, NA, 31, NA, 35, 34, 15,~
$ SibSp       <dbl> 1, 1, 0, 1, 0, 0, 0, 3, 0, 1, 1, 0, 0, 1, 0, 0, 4, 0, 1, 0, 0, 0, 0, 0, 3, 1, 0, 3, 0, 0~
$ Parch       <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 2, 0, 1, 0, 0, 5, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 5, 0, 2, 0, 0~
$ Ticket      <chr> "A/5 21171", "PC 17599", "STON/O2. 3101282", "113803", "373450", "330877", "17463", "349~
$ Fare        <dbl> 7.2500, 71.2833, 7.9250, 53.1000, 8.0500, 8.4583, 51.8625, 21.0750, 11.1333, 30.0708, 16~
$ Cabin       <chr> NA, "C85", NA, "C123", NA, NA, "E46", NA, NA, NA, "G6", "C103", NA, NA, NA, NA, NA, NA, ~
$ Embarked    <chr> "Southampton", "Cherbourg", "Southampton", "Southampton", "Southampton", "Queenstown", "~
>
```

Task # 3 (10 minutes)

Come up with an appropriate visualization to explore the relationship between Survival and Passenger Class

Task # 3 Discussion

```
titanic_df %>% ggplot(aes(x=Pclass, fill=Survived)) +
               geom_bar(position = "fill", col="white") +
               scale_fill_manual(values=c("dark red", "dark green")) +
               labs(x="Passenger Class", y="")+
               theme_bw()
```



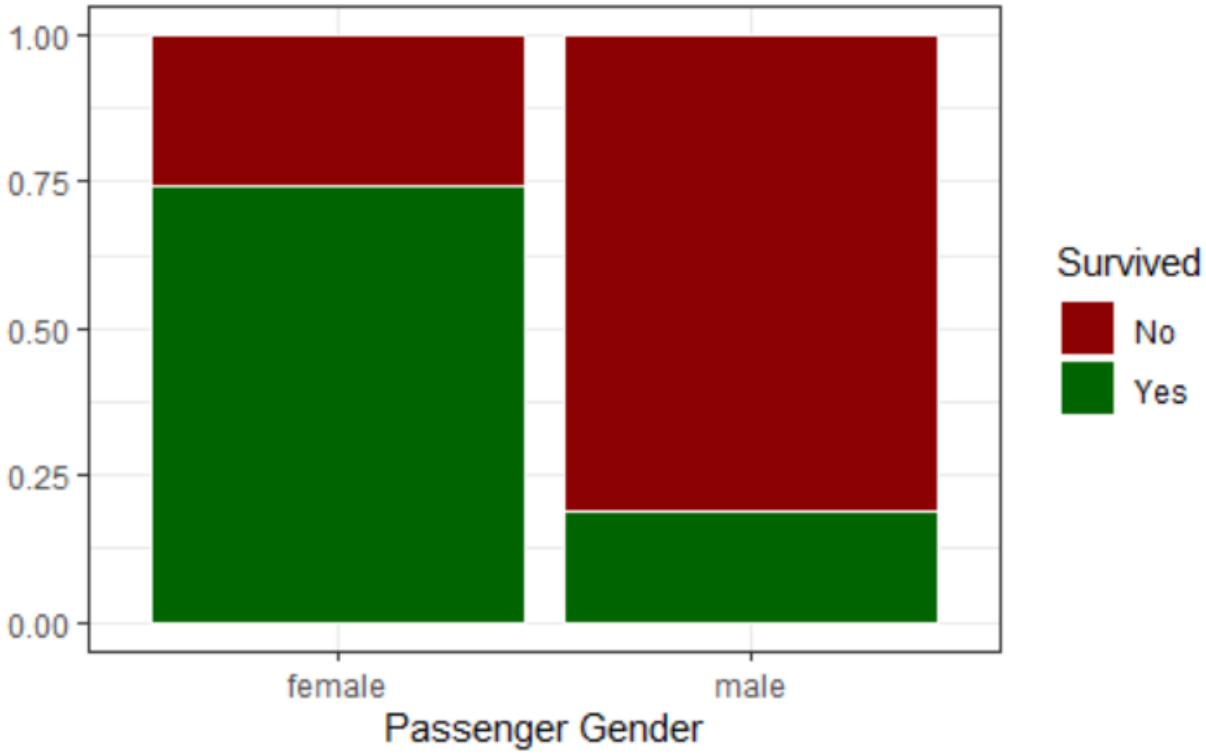
Stacked version

Task # 4 (5 minutes)

Come up with an appropriate visualization to explore the relationship between Survival and Gender

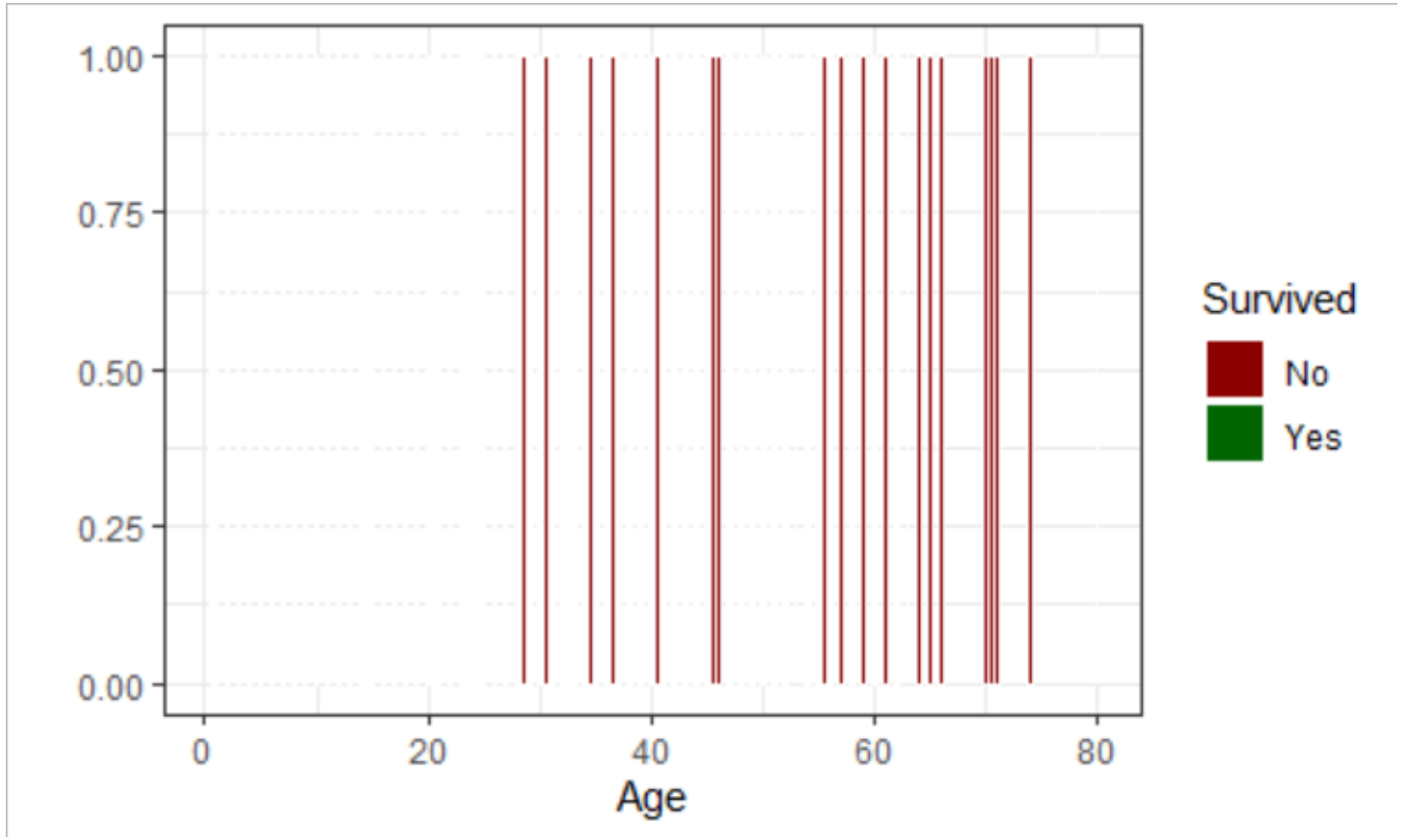
Task # 4 Discussion

```
titanic_df %>% ggplot(aes(x=Gender, fill=Survived)) +
          geom_bar(position = "fill", col="white") +
          scale_fill_manual(values=c("dark red", "dark green")) +
          labs(x="Passenger Gender", y="")+
          theme_bw()
```



Task # 4 Discussion

Relationship between Age and Survival



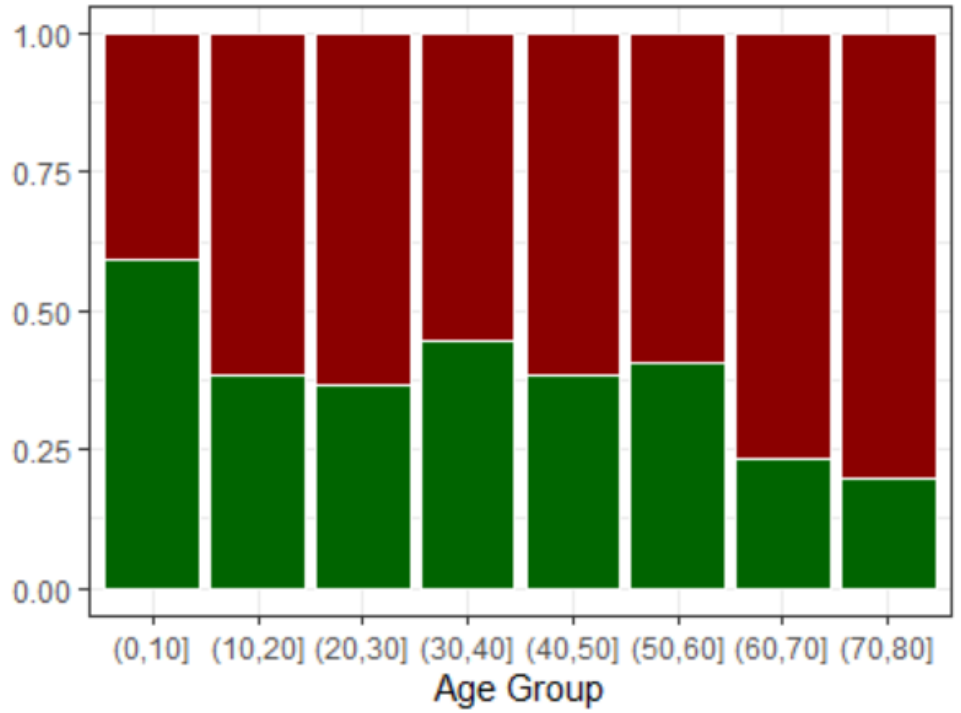
Task # 5 (15 minutes)

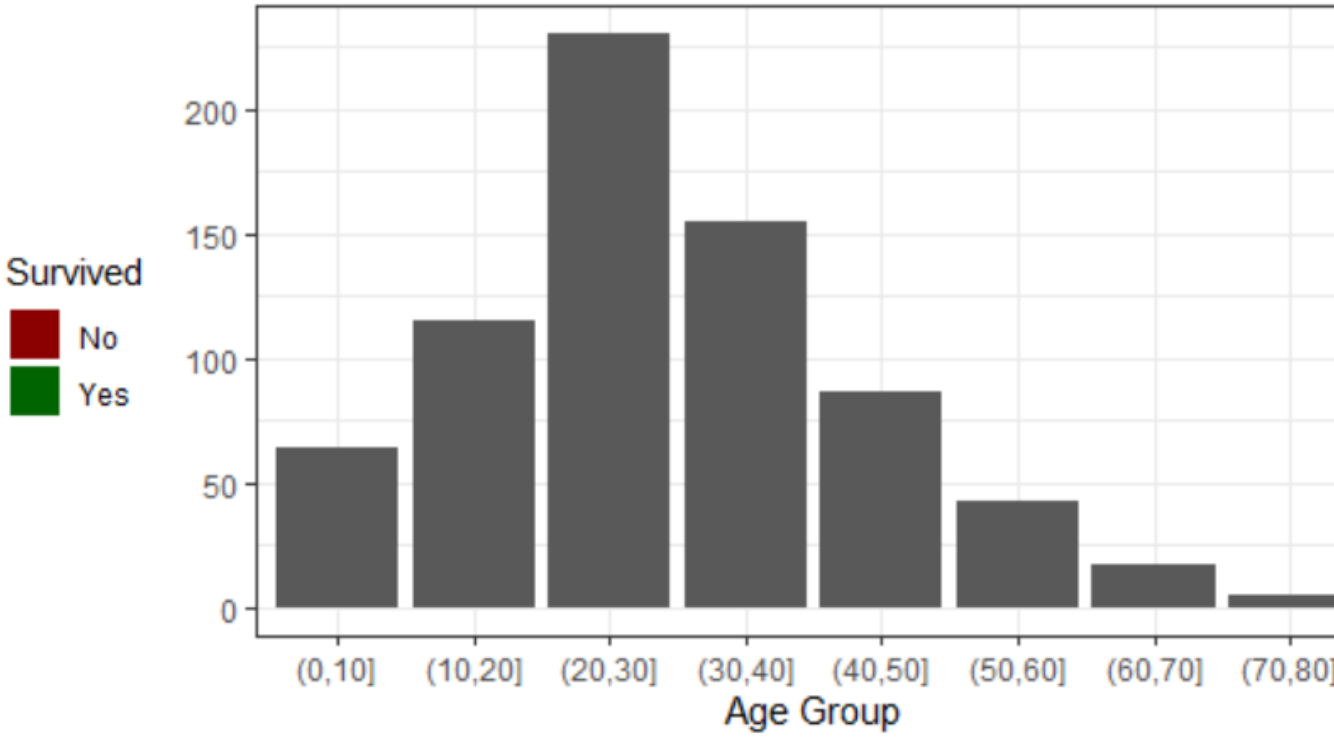
Come up with an appropriate visualization to explore the relationship between Survival and Age.

Hint: Create a new variable “age_group” that discretizes the age into 10-year buckets using the `cut` function

Task # 5 Discussion

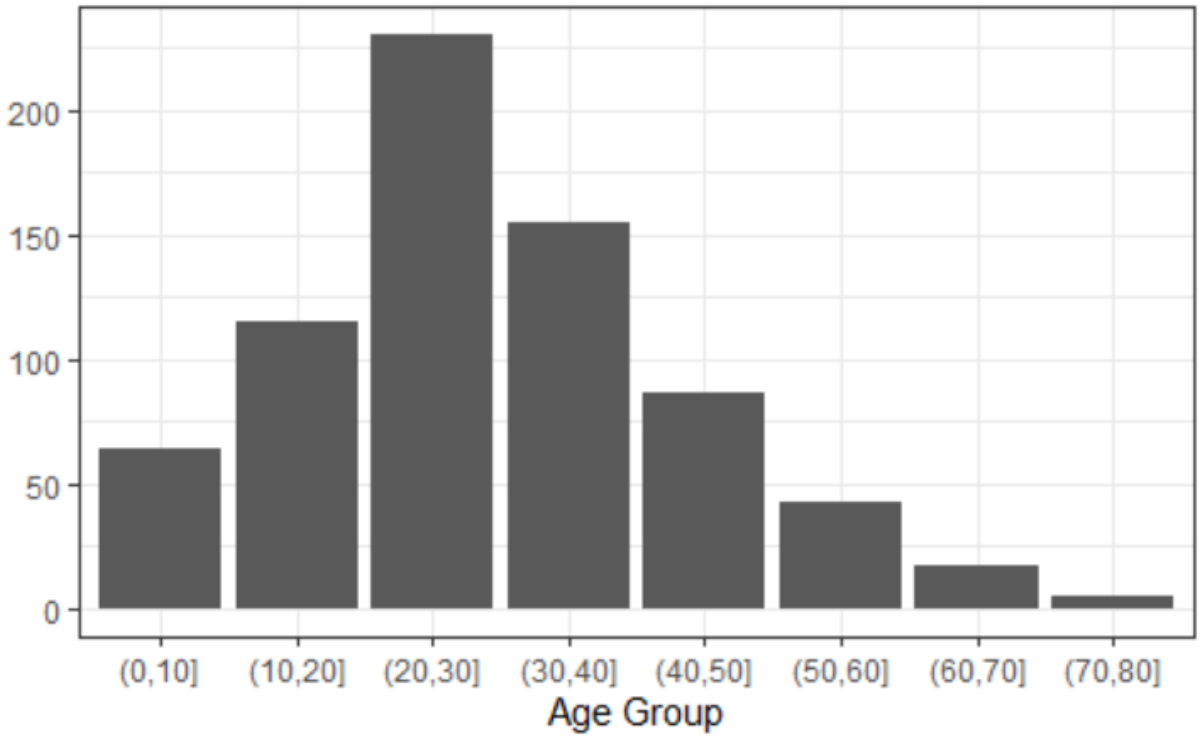
```
titanic_df %>%
        mutate(age_group = cut(Age, breaks=seq(0,100,10))) %>%
        select(Survived, age_group) %>%
        na.omit() %>%
        ggplot(aes(x=age_group, fill=Survived)) +
        geom_bar(position = "fill", col="white") +
        scale_fill_manual(values=c("dark red", "dark green")) +
        labs(x="Age Group", y="")+
        theme_bw()
```





Task # 5 Discussion

```
titanic_df %>%
        mutate(age_group = cut(Age, breaks=seq(0,100,10))) %>%
        select(Survived, age_group) %>%
        na.omit() %>%
        ggplot(aes(x=age_group)) +
        geom_bar() +
        labs(x="Age Group", y="")+
        theme_bw()
```

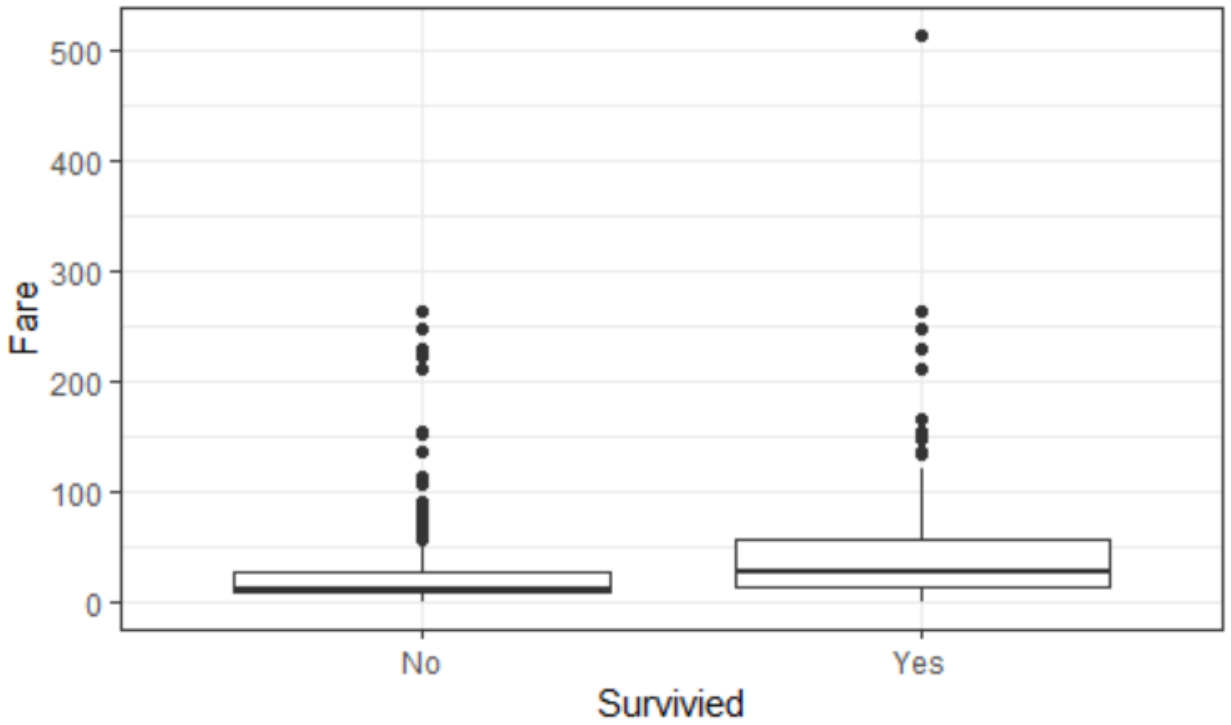


Task # 6 (10 minutes)

Come up with an appropriate visualization to explore the relationship between Survival and Fare

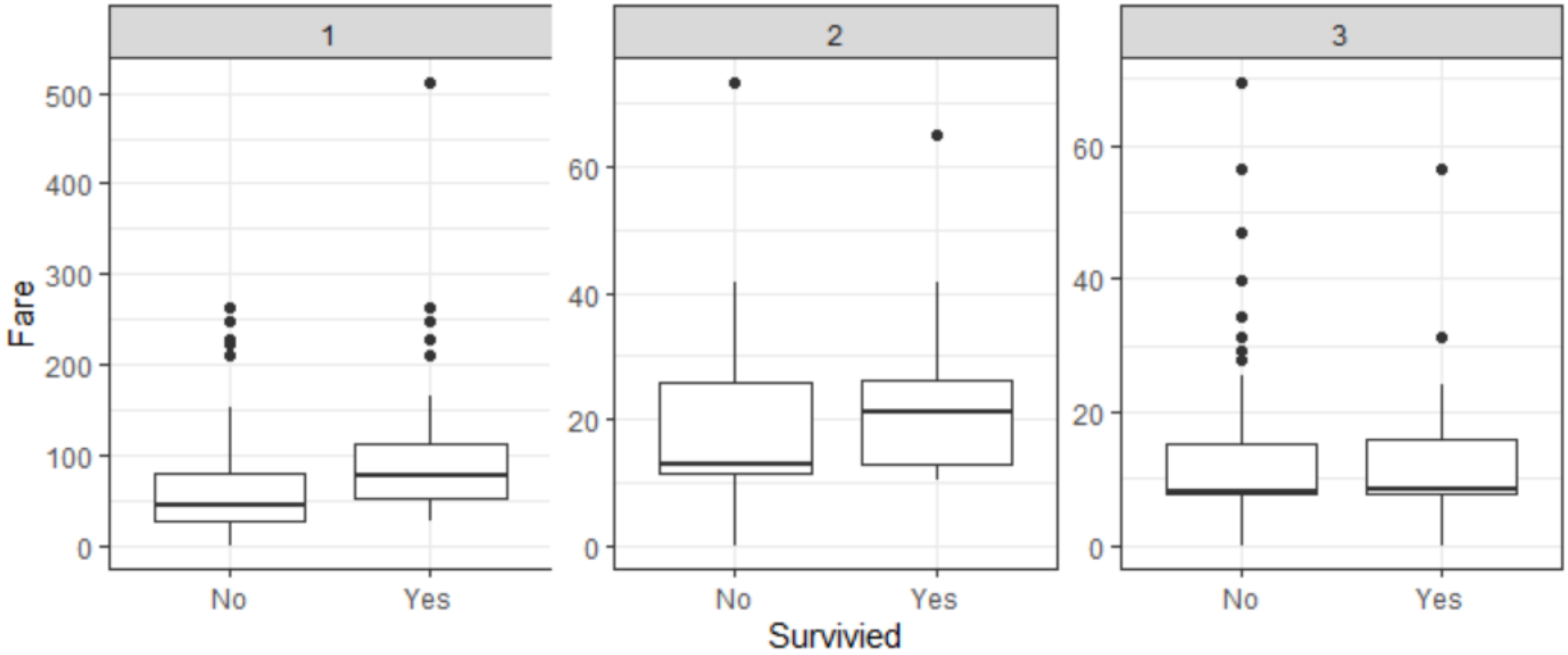
Task # 6 Discussion

```
titanic_df %>%
        select(Survived, Fare) %>%
        na.omit() %>%
        ggplot(aes(x=Survived, y=Fare)) +
        geom_boxplot()+
        labs(x="Survivied", y="Fare")+
        theme_bw()
```



Task # 6 Discussion

```
titanic_df %>%
  select(Survived, Fare, Pclass) %>%
  na.omit() %>%
  ggplot(aes(x=Survived, y=Fare)) +
  geom_boxplot()+
  facet_wrap(~Pclass, ncol=3, scale="free_y") +
  labs(x="Survivied", y="Fare")+
  theme_bw()
```

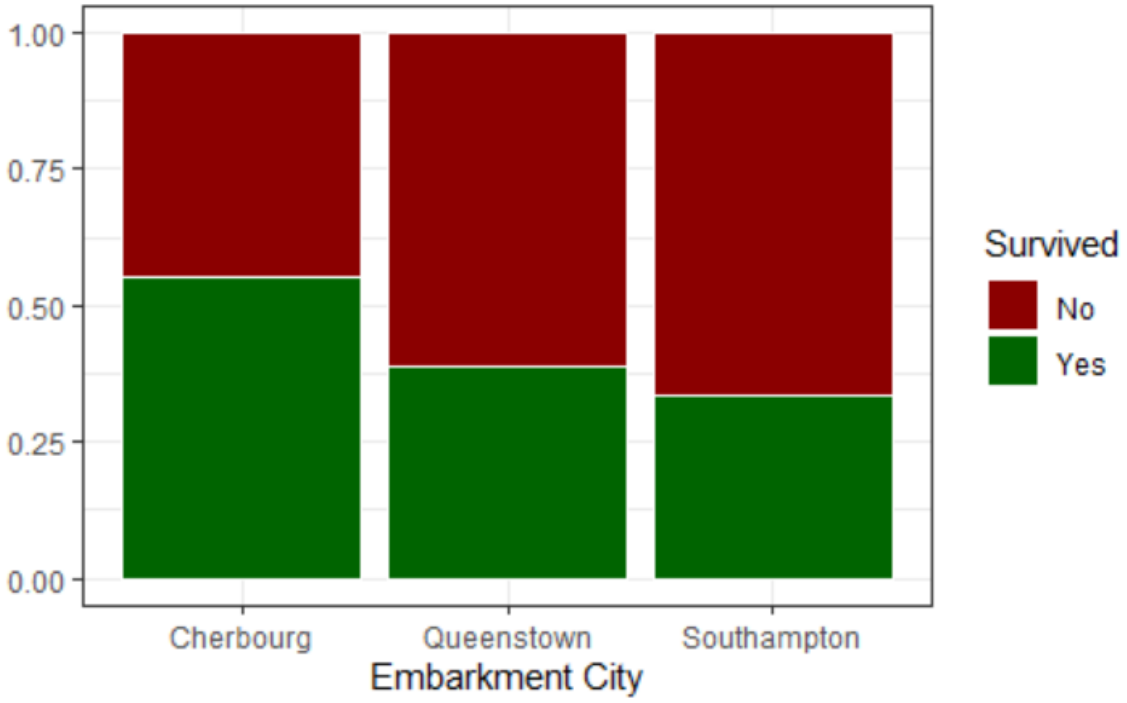


Task # 7 (10 minutes)

Come up with an appropriate visualization to explore the relationship between Survival and Embarkment city

Task # 7 Discussion

```
titanic_df %>%
  select(Survived, Embarked) %>%
  na.omit() %>%
  ggplot(aes(x=Embarked, fill=Survived)) +
  geom_bar(position = "fill", col="white") +
  scale_fill_manual(values=c("dark red", "dark green")) +
  labs(x="Embarkment City", y="")+
  theme_bw()
```



Task # 8 (10 minutes)

Use the tidytext package unnest_tokens to break up the passengers names column into separate words

Task # 8 Discussion

```
tidy_titanic_df <- titanic_df %>%
                       select(PassengerId, Survived, Name) %>%
                       unnest_tokens(word, Name)

glimpse(tidy_titanic_df)
```

```
> glimpse(tidy_titanic_df)
Rows: 3,638
Columns: 3
$ PassengerId <dbl> 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5~
$ Survived    <chr> "No", "No", "No", "No", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", ~
$ word        <chr> "braund", "mr", "owen", "harris", "cumings", "mrs", "john", "bradley", "f~
```

Task # 9 (10 minutes)

Find the top 10 names by survival status. Use your own stop words dataset if needed.

Task # 9 Discussion

```
stop_names <- tibble(word = c("mr", "miss", "mrs", "master"))

name_freq <- tidy_titanic_df %>% anti_join(stop_names) %>%
                     group_by(Survived, word) %>%
                     summarize(n = n()) %>%
                     top_n(20, n) %>%
                     arrange(-n) %>%
                     ungroup()
glimpse(name_freq)
```

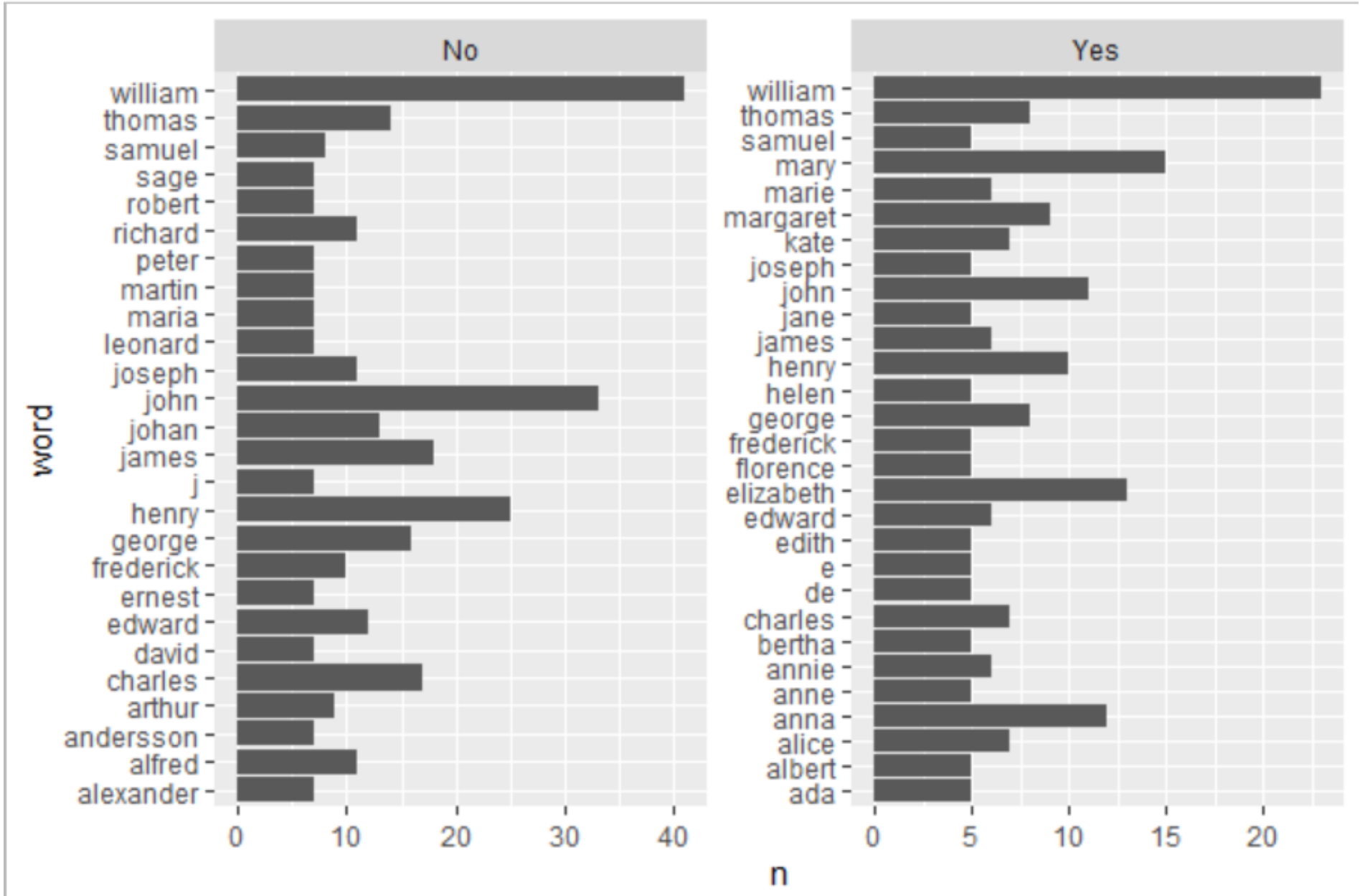
```
> glimpse(name_freq)
Rows: 55
Columns: 3
$ Survived <chr> "No", "No", "No", "Yes", "No", "No", "No", "Yes", "No", "No", "Yes", "No", "~
$ word     <chr> "william", "john", "henry", "william", "james", "charles", "george", "mary",~
$ n        <int> 41, 33, 25, 23, 18, 17, 16, 15, 14, 13, 13, 12, 12, 11, 11, 11, 11, 10, 10, ~
```

Task # 9 (10 minutes)

Visualize the top 10 names by survival status

Task # 10 Discussion

```
cale="free") +
```



Task # 10 Discussion

Last names only