[CMP595] Final Project - Avengers, their Deaths and Subsequent Return to Life.

Introduction

This is a report on an analysis performed on a dataset that details the deaths (and subsequent return to life) of Marvel comic book superheroes that have joined the Avengers team, between the time they have joined and April of 2015. This dataset has been provided by an Internet website called FiveThirtyEight and is organized as a list of all members of the Avengers with some of their characteristics. For each Avenger there are binary variables that inform if the hero has died one, two, three, four or five times in the past. The same information is available if the hero has returned to life, from one to five times as well.

We analyze this dataset through three research questions:

- Is there some characteristic that makes an Avenger more likely to die?
- Is there some characteristic that makes an Avenger more likely to revive?
- Is the number of comic book appearances of a superhero correlated to the number of years since he joined the team?

Loading the CSV File and preprocessing the dataframe

Firstly, we download the CSV file from the FiveThirtyEight Github and save it as "avengers.csv"

```
file = "avengers.csv"
if(!file.exists(file)){
  download.file(
    "https://raw.githubusercontent.com/fivethirtyeight/data/master/avengers/avengers.csv",
    destfile=file)
}
```

Then, we load the file "avengers.csv" in the dataframe "avengers_df"

```
library(readr);
library(knitr);
library(tidyverse);
library(gridExtra);
avengers_df <- read.csv("avengers.csv", quote="\"");
kable(head(avengers_df %>% select(Name.Alias,Appearances,Death1,Return1)))
```

Name.Alias	Appearances	Death1	Return1
Henry Jonathan "Hank" Pym	1269	YES	NO
Janet van Dyne	1165	YES	YES
Anthony Edward "Tony" Stark	3068	YES	YES
Robert Bruce Banner	2089	YES	YES
Thor Odinson	2402	YES	YES
Richard Milhouse Jones	612	NO	

Changing the type of some columns

Based on an overall analysis of the data, we noticed that some columns have inadequate typing, specifically "URL", "Name.Alias" and "Notes" , which should be string. So, we convert them:

```
avengers_df$URL <-  as.character(avengers_df$URL)
avengers_df$Name.Alias <-  as.character(avengers_df$Name.Alias)
avengers_df$Notes <-  as.character(avengers_df$Notes)
```

Converting to Tidy data

As mentioned in the introduction, the dataset contains five columns that binarily register if the Avenger has died. Each column represents their first, second, third, fourth and fifth deaths. Similarly, there are also five columns that represent if the Avenger has returned to life. These 10 columns simply contain YES or NO values.

There is no information of when the Avenger has died. In fact, the dataset is essentially a list of all Avengers that have ever existed. Therefore, converting each death into different rows does not make sense.

Instead, to make the number of deaths and revivals simpler, we will convert the number of YES values for each variable into an integer.

```
avengers_tidydf <- avengers_df %>%
  mutate(Total.Deaths = avengers_df %>%
                        select(Death1,Death2,Death3,Death4,Death5) %>%
                        apply(1, function(x) length(x[x=="YES"]))
                        )   %>%
  mutate(Total.Return = avengers_df %>%
                        select(Return1,Return2,Return3,Return4,Return5) %>%
                        apply(1, function(x) length(x[x=="YES"]))
                        ) %>%
  subset(select=-c(Death1,Death2,Death3,Death4,Death5,
                   Return1,Return2,Return3,Return4,Return5))
kable(head(avengers_tidydf %>% select(Name.Alias,Appearances,Total.Deaths,Total.Return)))
```

Name.Alias	Appearances	Total.Deaths	Total.Return
Henry Jonathan "Hank" Pym	1269	1	0
Janet van Dyne	1165	1	1
Anthony Edward "Tony" Stark	3068	1	1
Robert Bruce Banner	2089	1	1
Thor Odinson	2402	2	1
Richard Milhouse Jones	612	0	0

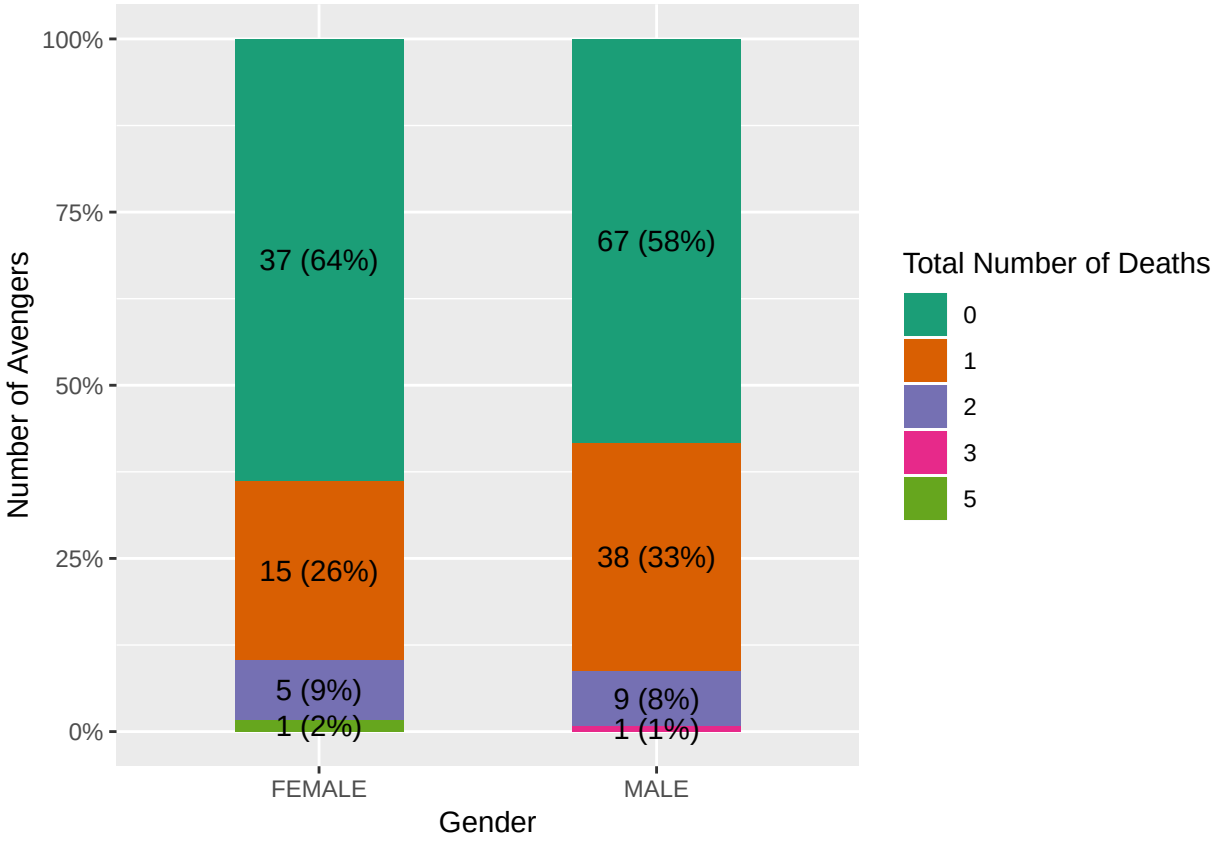
Question 1: Is there some characteristic that makes an Avenger more likely to have died?

We explored this question based on four characteristics available in the dataset: Gender, honorary status, number of years since joining the Avengers, and the number of comic book appearances of the hero. The plots in this section share a common presentaion method, which is the normalized stacked plot. This makes it easier to compare two groups of Avengers through diferences in the ratio of the total number of deaths in that group.

Gender

Let's start with gender. From the graph, we can see that women are more likely to not die than men (64% to 58%) compared to dying exactly one time (26% to 33%).

```
avengers_tidydf %>%
select(Gender,Total.Deaths) %>%
group_by(Gender,Total.Deaths) %>%
summarize(NumberOfAvengers = n()) %>%
mutate(percent= paste0(round(100*NumberOfAvengers/sum(NumberOfAvengers)), '%')) %>%
ggplot(aes(x= Gender, y = NumberOfAvengers, fill = as.factor(Total.Deaths))) +
geom_col(position = "fill", width = .5) +
scale_fill_brewer(name ="Total Number of Deaths", palette = "Dark2") +
geom_text(aes(label= paste0(NumberOfAvengers, ' (',percent,')')),
          position = position_fill(vjust=.5)) +
xlab("Gender") +
scale_y_continuous(name = "Number of Avengers", labels = scales::percent)#+
```

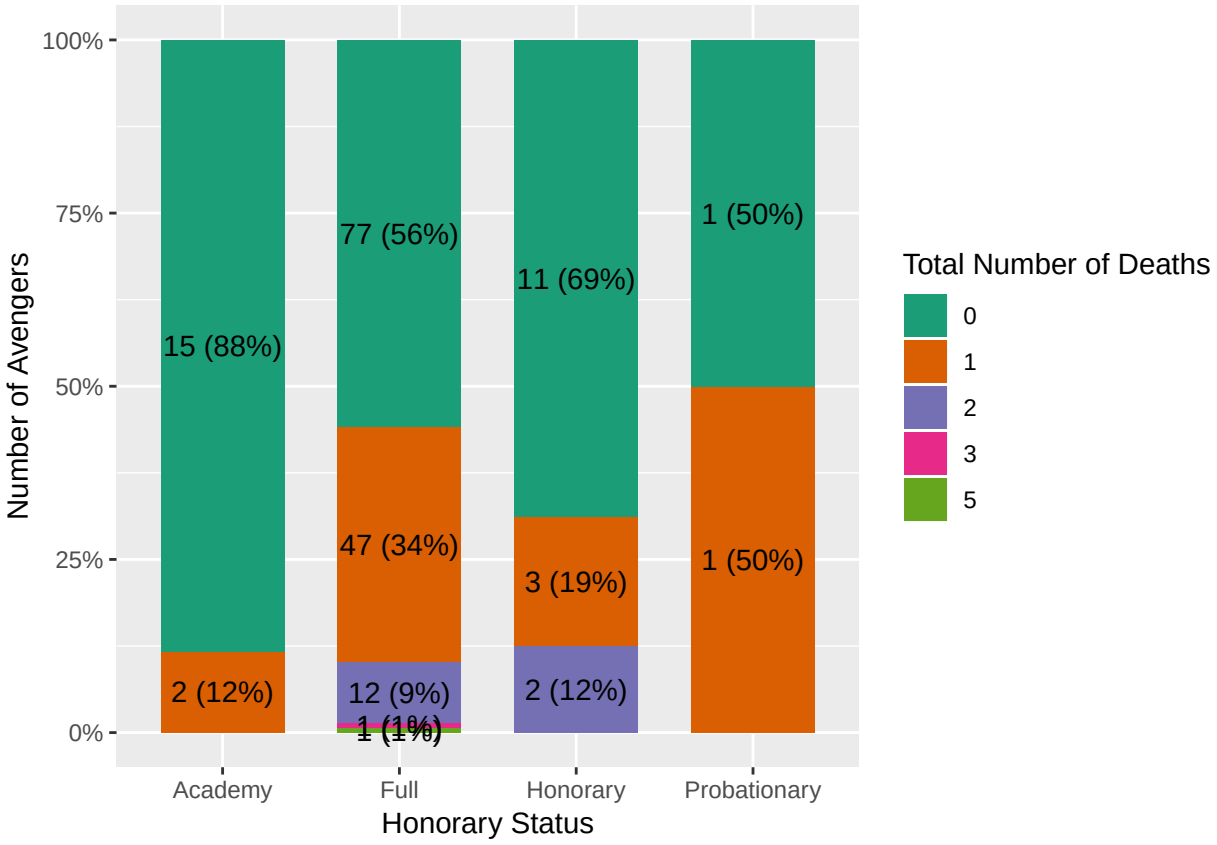


##Honorary Status

Let's try honorary status. As seen in the graph, Academic and Honorary Avengers are more likely to not have died compared to Full-time Avengers. Probationary has only two Avengers, so it is not possible to infer anything.

```
avengers_tidydf %>%
select(Honorary,Total.Deaths) %>%
group_by(Honorary,Total.Deaths) %>%
summarize(NumberOfAvengers = n()) %>%
```

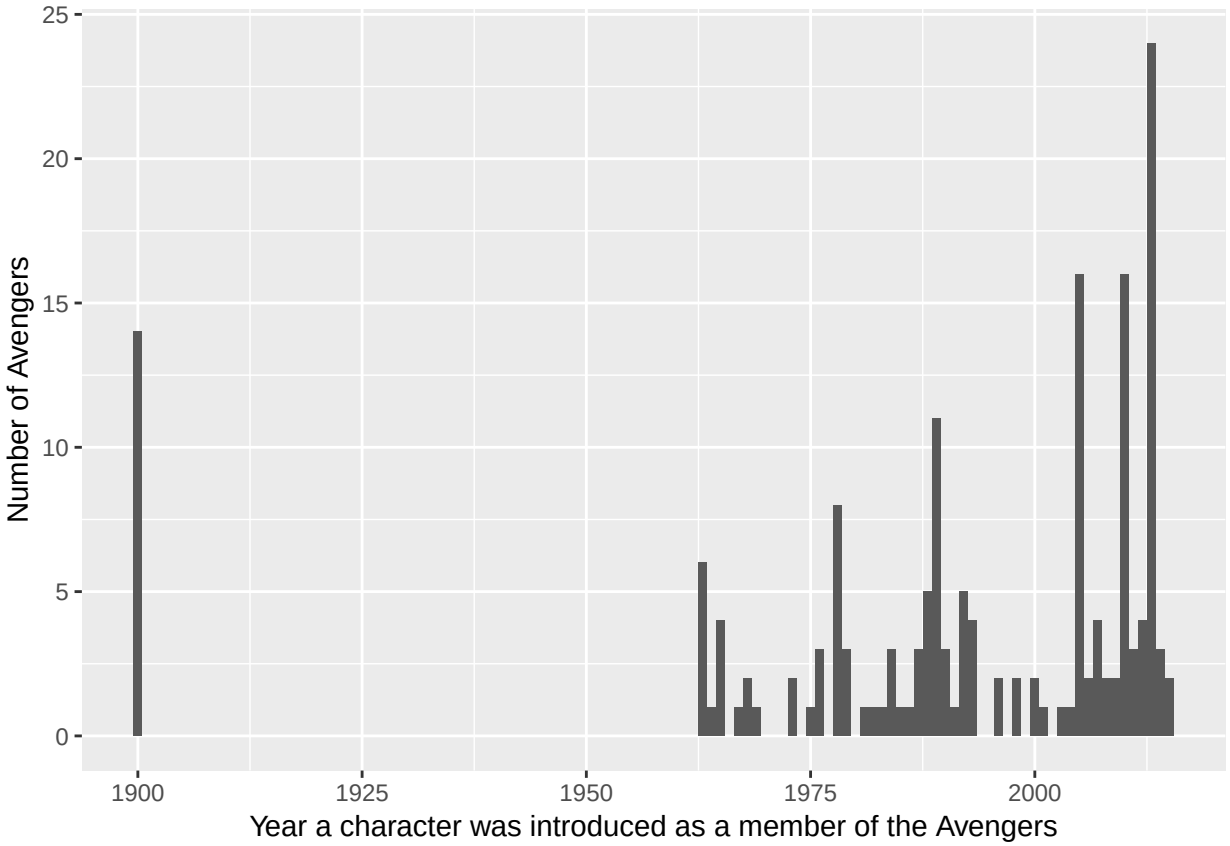
```
mutate(percent= paste0(round(100*NumberOfAvengers/sum(NumberOfAvengers)), '%')) %>%
ggplot(aes(x= Honorary, y = NumberOfAvengers, fill = as.factor(Total.Deaths))) +
geom_col(position = "fill", width = .7) +
scale_fill_brewer(name ="Total Number of Deaths", palette = "Dark2") +
geom_text(aes(label= paste0(NumberOfAvengers, ' (',percent,')')),
          position = position_fill(vjust=.5)) +
xlab("Honorary Status") +
scale_y_continuous(name = "Number of Avengers", labels = scales::percent)#+
```



Number of Years since joining the Avengers

Let's analyze the Years.since.joining variable. Years.since.joining is equal to, by the datasets definition, the year 2015 (when the dataset was created) minus the year the character was introduced as a member of the Avengers. Because of this, there is some problematic values in these columns, in that the year the character was introduced is 1900.

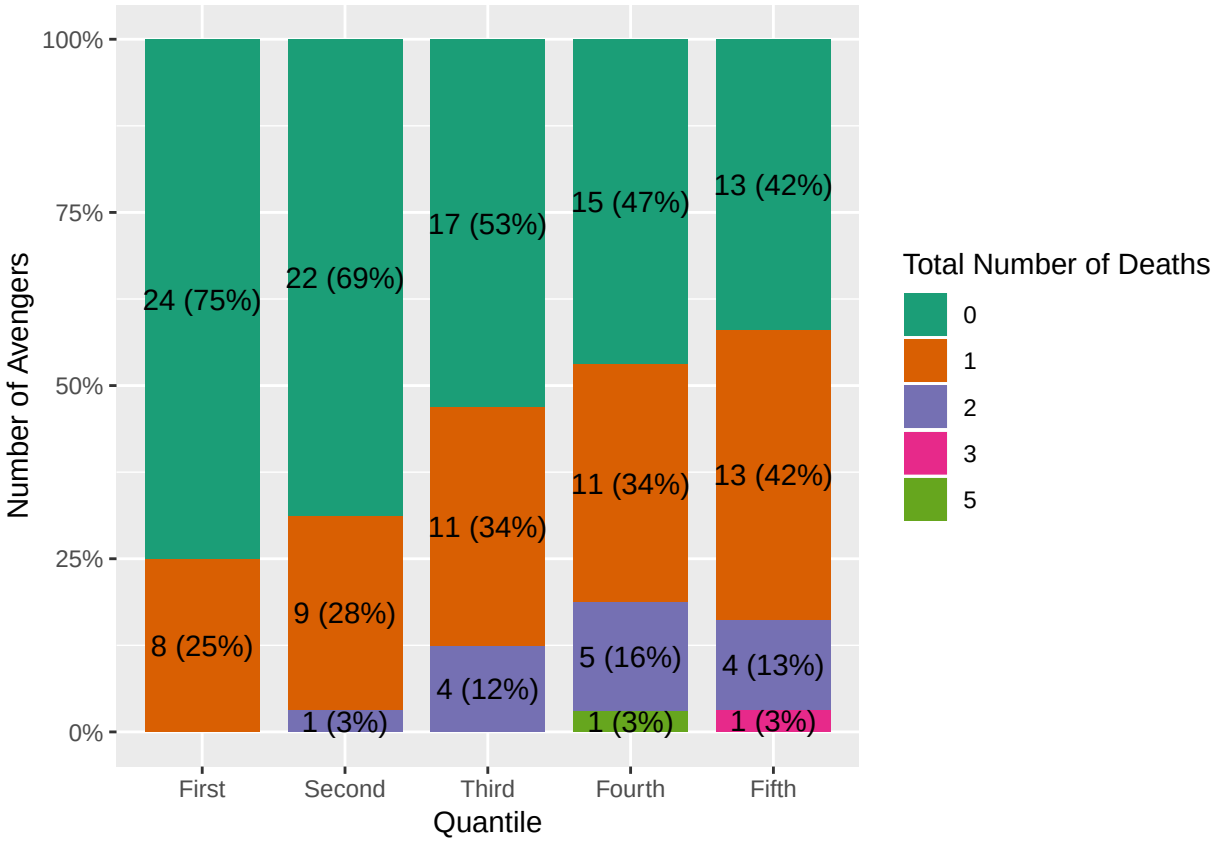
```
avengers_tidydf %>%
select(Year) %>%
group_by(Year) %>%
summarize(NumberOfAvengers = n()) %>%
ggplot(aes(x= Year, y = NumberOfAvengers, )) +
geom_col() +
scale_x_continuous(name= "Year a character was introduced as a member of the Avengers") +
scale_y_continuous(name = "Number of Avengers")#+
```



It is possible (and likely, looking at the Marvel wiki entries on these characters) that this value is simply the earliest possible year for the variable, which means that the year the character was introduced is unknown. Therefore, these plots should filter out these Avengers.

Since Years.since.joining is an integer variable, I grouped the rows by its 5 quantiles. Its subtle, but the longer an Avenger has been part of the team, the more likely it is for him to have died.

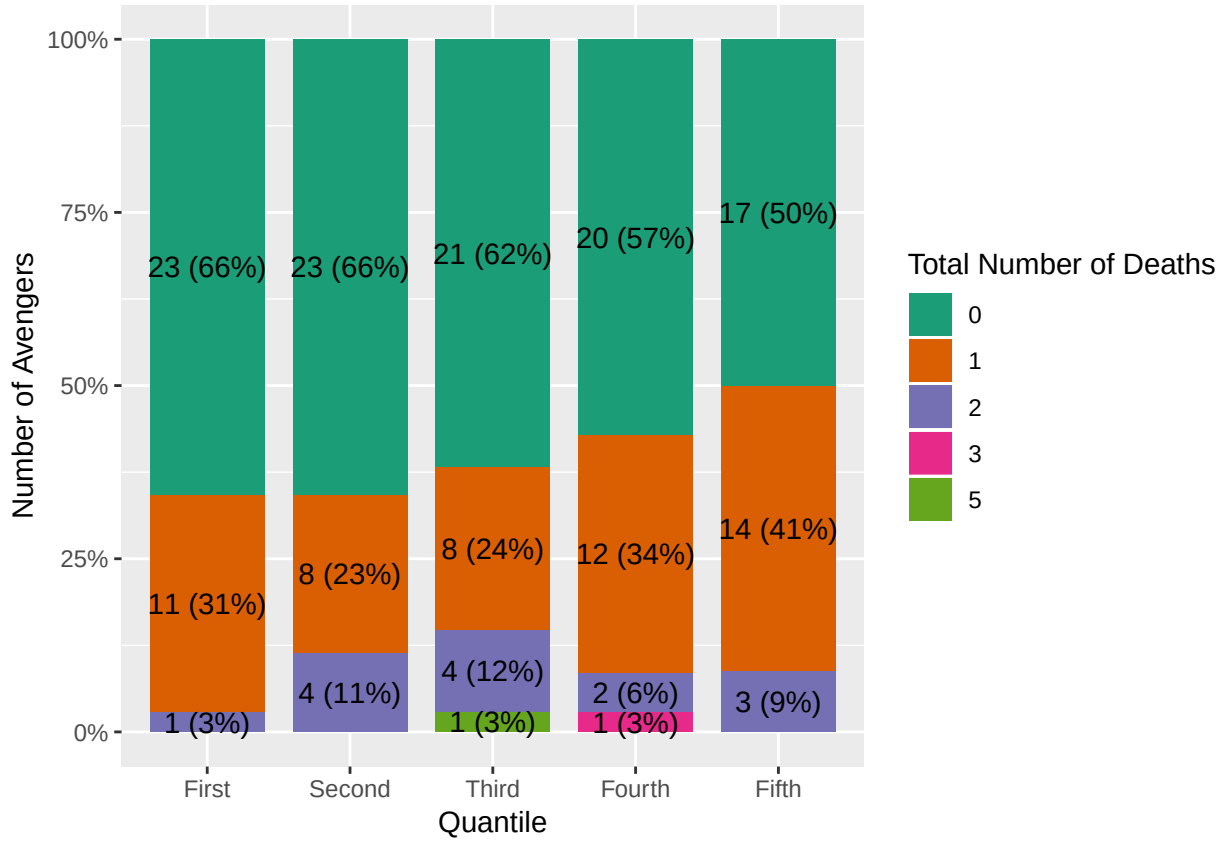
```
avengers_tidydf %>%
filter(Year != 1900) %>%
select(Years.since.joining,Total.Deaths) %>%
arrange(Years.since.joining) %>%
mutate(quartile = as.factor(ntile(Years.since.joining, 5))) %>%
group_by(quartile,Total.Deaths) %>%
summarize(NumberOfAvengers = n()) %>%
mutate(percent= paste0(round(100*NumberOfAvengers/sum(NumberOfAvengers)), '%')) %>%
ggplot(aes(x= quartile, y = NumberOfAvengers, fill = as.factor(Total.Deaths))) +
geom_col(position = "fill", width = .8) +
scale_fill_brewer(name ="Total Number of Deaths", palette = "Dark2") +
geom_text(aes(label= paste0(NumberOfAvengers, ' (',percent,')')),
          position = position_fill(vjust=.5)) +
scale_x_discrete(name= "Quantile",
                 labels = c("First", "Second", "Third", "Fourth", "Fifth")) +
scale_y_continuous(name = "Number of Avengers", labels = scales::percent)#+
```



Number of Appereances in comic books

I repeated the previous process for the number of appereances. The same thing can also be said. The more comic book appereances an Avenger has, the more likely it is for him to have died. This may imply that these variables have some correlation. We investigate this question in the last section of this report.

```
avengers_tidydf %>%
select(Appearances,Total.Deaths) %>%
arrange(Appearances) %>%
mutate(quartile = as.factor(ntile(Appearances, 5))) %>%
group_by(quartile,Total.Deaths) %>%
summarize(NumberOfAvengers = n()) %>%
mutate(percent= paste0(round(100*NumberOfAvengers/sum(NumberOfAvengers)), '%')) %>%
ggplot(aes(x= quartile, y = NumberOfAvengers, fill = as.factor(Total.Deaths))) +
geom_col(position = "fill", width = .8) +
scale_fill_brewer(name ="Total Number of Deaths", palette = "Dark2") +
geom_text(aes(label= paste0(NumberOfAvengers, ' (',percent,')')),
          position = position_fill(vjust=.5)) +
scale_x_discrete(name= "Quantile",
                 labels = c("First", "Second", "Third", "Fourth", "Fifth")) +
scale_y_continuous(name = "Number of Avengers", labels = scales::percent)#+
```



Question 2: Is there some characteristic that makes an Avenger more likely to have revived?

Filtering all Avengers that have not died

The method for investigating why an Avenger may return to life is similar to the method of the previous section. Thus, most plots from question 1 are repeated in this section, though altered to investigate "Total.Returns" instead of "Total.Deaths". There is, however, one preprocessing step, in which we filter all Avengers that have not died from the dataset. We do so:

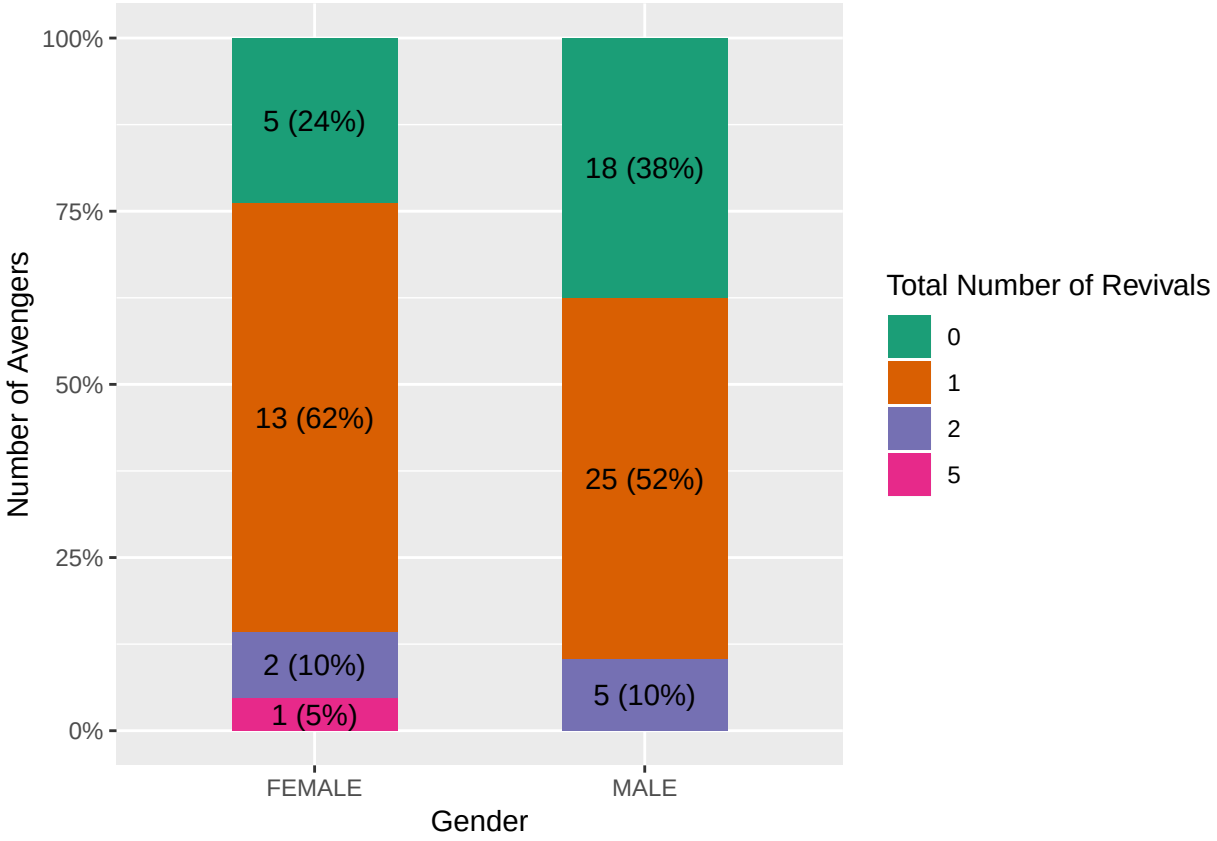
```
avengersThatDied_tidydf <- avengers_tidydf %>%
                           filter(Total.Deaths!=0)
kable(head(avengersThatDied_tidydf %>%
           select(Name.Alias,Appearances,Total.Deaths,Total.Return)))
```

Name.Alias	Appearances	Total.Deaths	Total.Return
Henry Jonathan "Hank" Pym	1269	1	0
Janet van Dyne	1165	1	1
Anthony Edward "Tony" Stark	3068	1	1
Robert Bruce Banner	2089	1	1
Thor Odinson	2402	2	1
Steven Rogers	3458	1	1

Gender

Let's start with Gender. Similarly with deaths, women are more likely to have returned to life than men.

```
avengersThatDied_tidydf %>%
select(Gender,Total.Return) %>%
group_by(Gender,Total.Return) %>%
summarize(NumberOfAvengers = n()) %>%
mutate(percent= paste0(round(100*NumberOfAvengers/sum(NumberOfAvengers)), '%')) %>%
ggplot(aes(x= Gender, y = NumberOfAvengers, fill = as.factor(Total.Return))) +
geom_col(position = "fill", width = .5) +
scale_fill_brewer(name ="Total Number of Revivals", palette = "Dark2") +
geom_text(aes(label= paste0(NumberOfAvengers, ' (',percent,')')),
          position = position_fill(vjust=.5)) +
xlab("Gender") +
scale_y_continuous(name = "Number of Avengers", labels = scales::percent)#+
```

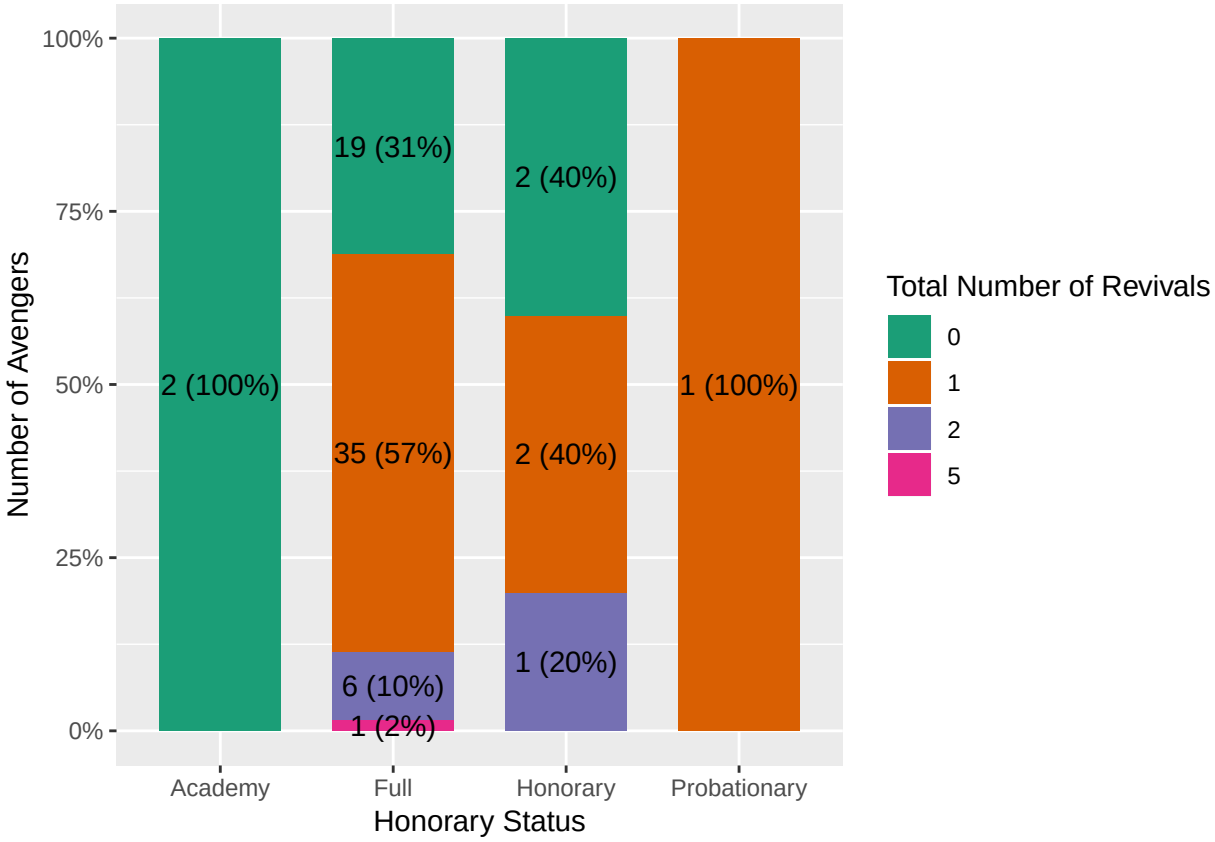


##Honorary Status

In the case of honorary status, proportionaly FUll-Time Avengers are more likely to return to life, but other statuses have so few members its not really safe to conclude anything.

```
avengersThatDied_tidydf %>%
select(Honorary,Total.Return) %>%
group_by(Honorary,Total.Return) %>%
summarize(NumberOfAvengers = n()) %>%
mutate(percent= paste0(round(100*NumberOfAvengers/sum(NumberOfAvengers)), '%')) %>%
ggplot(aes(x= Honorary, y = NumberOfAvengers, fill = as.factor(Total.Return))) +
```

```
geom_col(position = "fill", width = .7) +
scale_fill_brewer(name ="Total Number of Revivals", palette = "Dark2") +
geom_text(aes(label= paste0(NumberOfAvengers, ' (',percent,')')),
          position = position_fill(vjust=.5)) +
xlab("Honorary Status") +
scale_y_continuous(name = "Number of Avengers", labels = scales::percent)#+
```

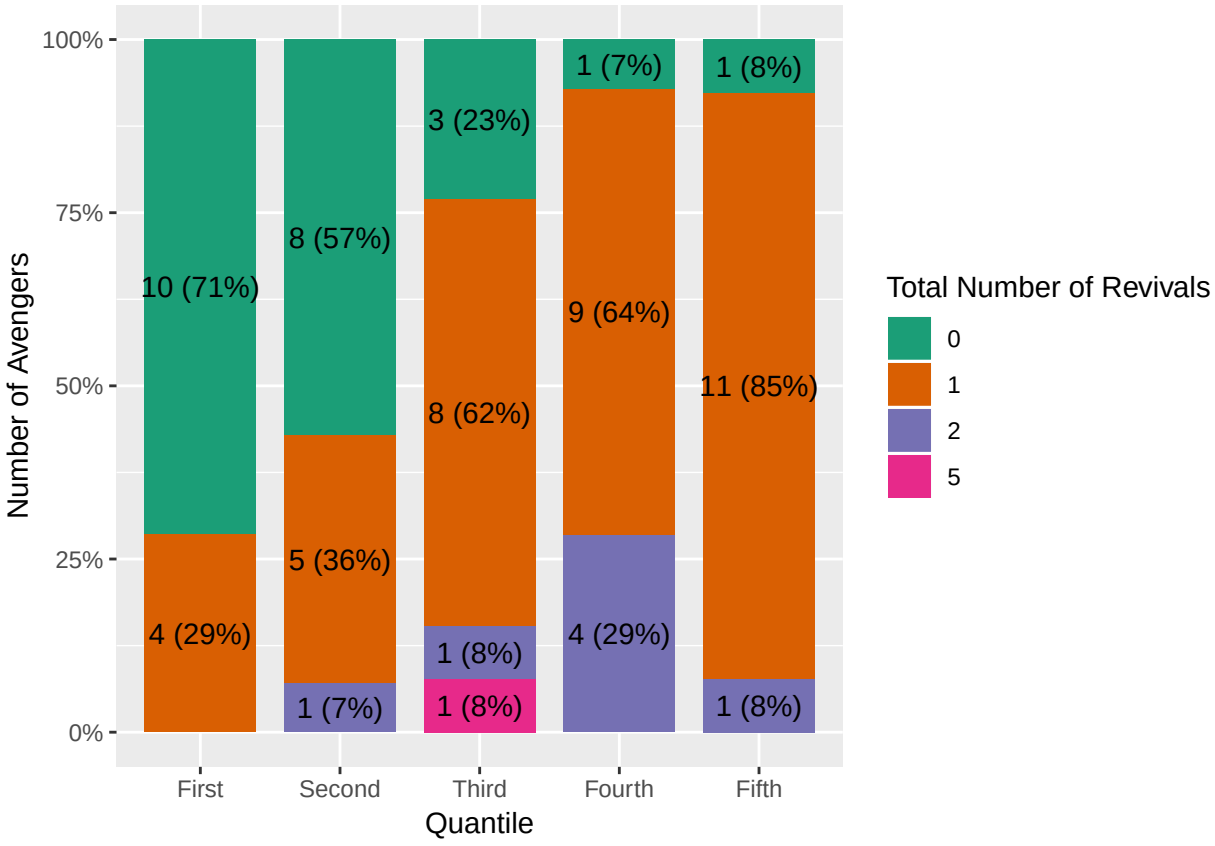


Number of Years since joining the Avengers

Just as an Avenger is more likely to have died the longer he has been one, so too is he more likely to have returned to life.

```
avengersThatDied_tidydf %>%
filter(Year != 1900) %>%
select(Years.since.joining,Total.Return) %>%
arrange(Years.since.joining) %>%
mutate(quartile = as.factor(ntile(Years.since.joining, 5))) %>%
group_by(quartile,Total.Return) %>%
summarize(NumberOfAvengers = n()) %>%
mutate(percent= paste0(round(100*NumberOfAvengers/sum(NumberOfAvengers)), '%')) %>%
ggplot(aes(x= quartile, y = NumberOfAvengers, fill = as.factor(Total.Return))) +
geom_col(position = "fill", width = .8) +
scale_fill_brewer(name ="Total Number of Revivals", palette = "Dark2") +
geom_text(aes(label= paste0(NumberOfAvengers, ' (',percent,')')),
          position = position_fill(vjust=.5)) +
```

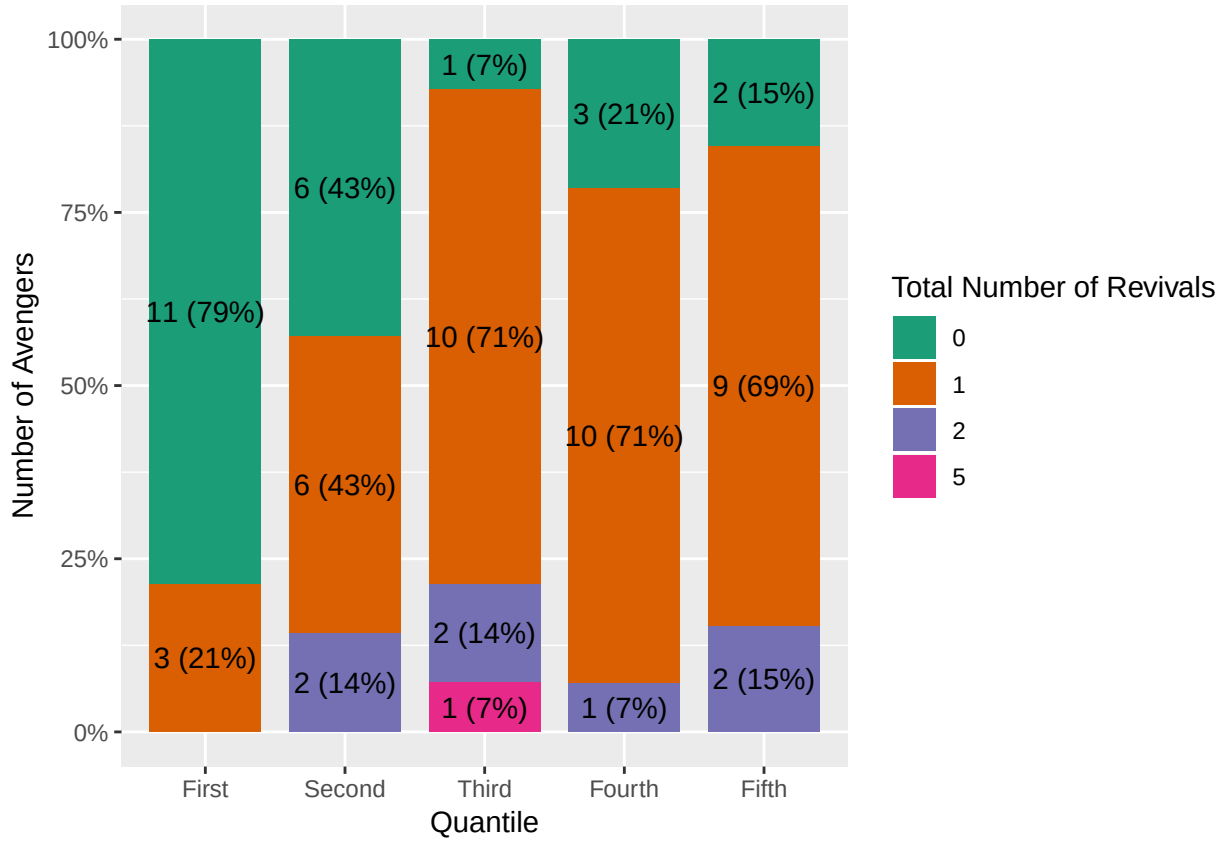
```
scale_x_discrete(name= "Quantile",
                 labels = c("First", "Second", "Third", "Fourth", "Fifth")) +
scale_y_continuous(name = "Number of Avengers", labels = scales::percent)#+
```



Number of Appereances in comic books

In case of comic book appearances, there is also a positive correlation between this variable the how many times an Avenger has returned to life. It seems obvious, as dead heroes cannot appear in any more comics.

```
avengersThatDied_tidydf %>%
select(Appearances,Total.Return) %>%
arrange(Appearances) %>%
mutate(quartile = as.factor(ntile(Appearances, 5))) %>%
group_by(quartile,Total.Return) %>%
summarize(NumberOfAvengers = n()) %>%
mutate(percent= paste0(round(100*NumberOfAvengers/sum(NumberOfAvengers)), '%')) %>%
ggplot(aes(x= quartile, y = NumberOfAvengers, fill = as.factor(Total.Return))) +
geom_col(position = "fill", width = .8) +
scale_fill_brewer(name ="Total Number of Revivals", palette = "Dark2") +
geom_text(aes(label= paste0(NumberOfAvengers, ' (',percent,')')),
          position = position_fill(vjust=.5)) +
scale_x_discrete(name= "Quantile",
                 labels = c("First", "Second", "Third", "Fourth", "Fifth")) +
scale_y_continuous(name = "Number of Avengers", labels = scales::percent)#+
```

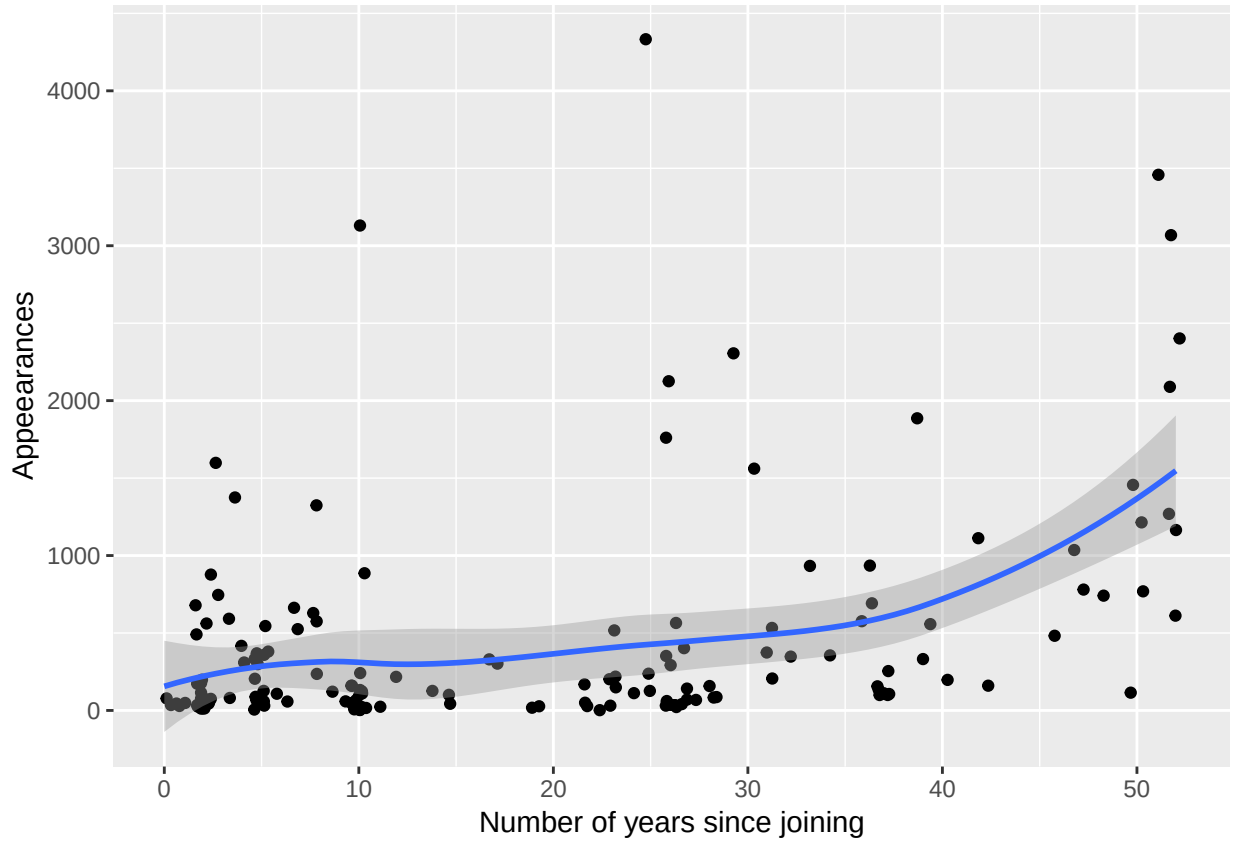


Question 3: Is the number of comic book appearances of a superhero correlated to the number of years since he joined the team?

Again, the Avengers with invalid Year values were filtered out. Then, a point was plotted for each Avenger in relation to the two variables. Based on the points, it is possible to observe a hard limit of 52 on the number of "years.since.joining", which is natural considering the first team of Avengers was formed in 1963.

Also plotted is a "locally estimated scatterplot smoothing" with geom_smooth(). This shows a trend between the two variables

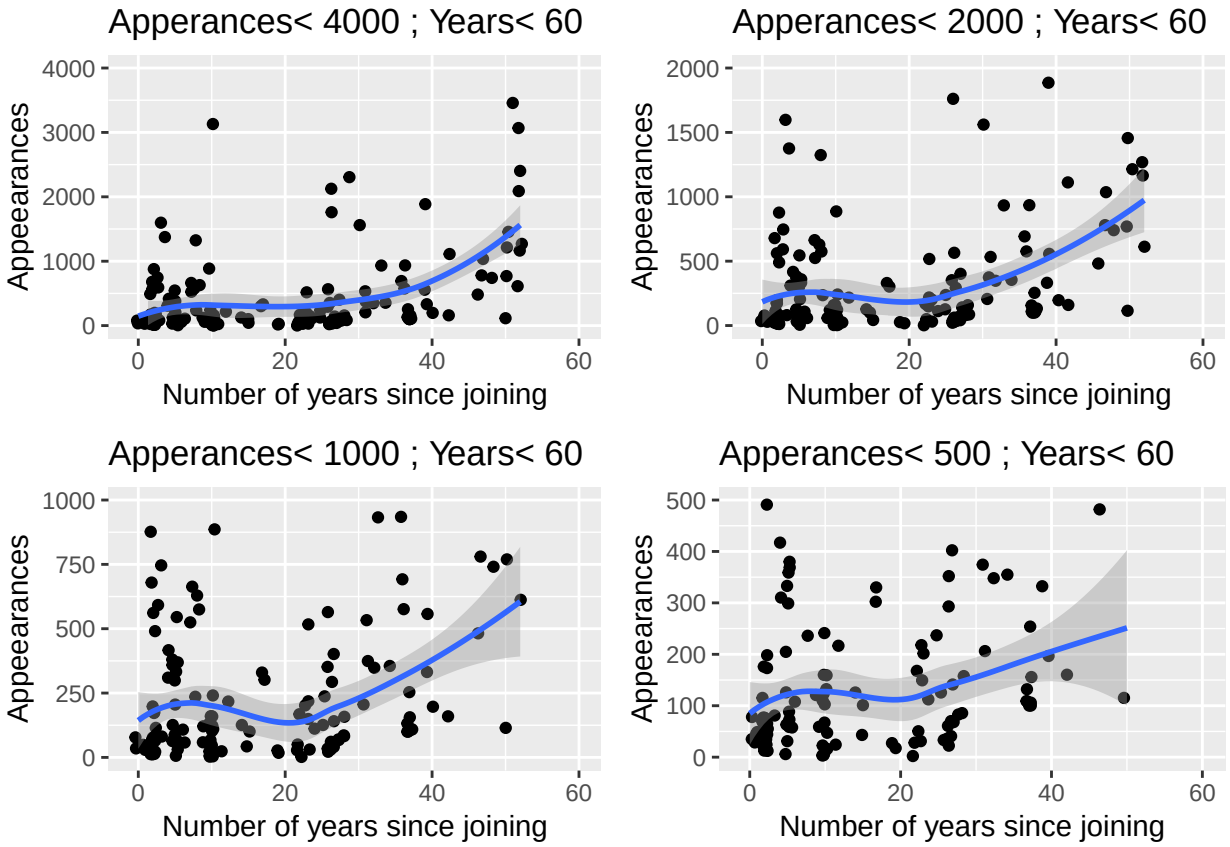
```
avengers_tidydf %>%
filter(Year != 1900) %>%
ggplot(aes(x= Years.since.joining, y = Appearances)) +
geom_jitter() +
geom_smooth(method = loess) +
scale_y_continuous(name = "Appeearances") +
scale_x_continuous(name = "Number of years since joining")
```



This plot shows a positive correlation between these two variables, but there might be some abnormality because of more popular superheros, such as Spider-Man. By filtering out the Avengers above 4000, 2000, 1000, and 500 appearances, we can see that Avengers between 10-20 years experience some loss of popularity, while newer and older Avengers see a positive correlation between these two variables.

```
scatterplotAppereancesYearsSinceJoining<- function (n=5000, m = 60){
  avengers_tidydf %>%
  filter(Year != 1900, Appearances<n ,Years.since.joining < m) %>%
  ggplot(aes(x= Years.since.joining, y = Appearances)) +
  geom_jitter() +
  geom_smooth(method = loess) +
  ggtitle(paste("Apperances<", n, "; Years<", m)) +
  scale_y_continuous(limits = c(-1,n), name = "Appeearances") +
  scale_x_continuous(limits = c(-1,m), name = "Number of years since joining")
}

lapply(list(4000,2000,1000,500),
       function(n){scatterplotAppereancesYearsSinceJoining(n)}) %>%
  grid.arrange(ncol = 2,grobs=.);
```


Apperances< 4000 ; Years< 60
Appeearances
4000
3000
2000
1000
0
0
20
40
60
Number of years since joining
Apperances< 2000 ; Years< 60
Appeearances
2000
1500
1000
500
0
0
20
40
60
Number of years since joining
Apperances< 1000 ; Years< 60
Appeearances
1000
750
500
250
0
0
20
40
60
Number of years since joining
Apperances< 500 ; Years< 60
Appeearances
500
400
300
200
100
0
0
20
40
60
Number of years since joining