CMDA-3654

Homework 5

Eduardo Salvador

Due as a .pdf upload

Instructions:

Delete the Instructions section from your write-up!!

I have given you this assignment as an .Rmd (R Markdown) file.

- Change the name of the file to: `Lastname_Firstname_CMDA_3654_HW5.Rmd` , and your output should therefore match but with a `.pdf` extension.
- You need to edit the R Markdown file by filling in the chunks appropriately with your code. Output will be generated automatically when you compile the document.
- You also need to add your own text before and after the chunks to explain what you are doing or to interpret the output.
- Feel free to add additional chunks if needed. I **will not** be providing assignments to you like this for the entire semester, just long enough for you to learn how to do it for yourself.

Required: The final product that you turn in must be a .pdf file.

- You MUST Knit this document directly to a PDF conversion from another format is not permitted.

Problem 1: [50 pts] Tidying Data with `dplyr`

Unicorns on unicycles

This dataset is thought to have been recorded by an amateur scientist, a natural philosopher by the name of Rudolphus in the 17th century in The Netherlands. This scientist recorded the annual population of unicorns in western Europe over a century and also recorded the sales of unicycles in that same time period. Although not much of the accompanying text remains of the original documents, what we can read is the tables and the idea that Rudolphus thought there was some sort of relationship between uncorns and unicycle sales.

The documents were recently unearthed from a hidden chest in Delft and seem to be written by Rudolphus Hogervorstus, my great great great uncle, in 1681. These documents show that he was studying the then roaming herds of unicorns in the area around Delft. Unfortunately these animals are extinct now.

His work contains multiple tables, carefully written down, documenting the population of unicorns over time in multiple places and related to that the sales and numbers of unicycles in those countries.

There are also some text describing the hypotheses of Rudolphus, but most of the texts were eaten by moths and have decayed over time. Some of the remaining text suggest that, according to Rudolphus, the unicorn populations and unicycles were related: “The presence of the cone on the unicorn hints at a very defined sense of equilibrium, …[some missing text]….. it is therefore only natural to assume unicorns ride unicycles”.

Your task

We want to discover the relationship between unicorns and unicycles. As part of the archival process these tables were copied in excel.

Try to read the data in, if you are using R, try the package `readxl` . It also helps if you look at the data before reading it in.

Load the two files, try to join them together.

Is there a relation between unicorns and unicycles?

About the data

Consider the `sales.xlsx` and `observations.xlsx` data sets.

- `observations.xlsx` contains information about the population of unicorns for specific years different countries.
- `sales.xlsx` contains two tables one of them consists of the total number of unicycles sold. The other is the revenue from the unicycle sales.

Problem Instructions

a. Join the two files together to create a single data frame, called `unicorns.cycling` . (Be sure to take precautions concerning country names being slightly different in the two files). Create a “tidy” data frame with appropriately named variables. (i.e. country, year, variable1, variable2, etc). Display the first 10 rows of the final data frame.

```
library(tidyverse)
library(readxl)
library(dplyr)
#Bring up Excel files using read_xlsx
obs <- read_xlsx("/Users/eduardosalvador/Desktop/FINAL\ Spring\ Semester\ 2021/CMDA\ /As
signments/HW5/observations.xlsx")
sales <- read_xlsx("/Users/eduardosalvador/Desktop/FINAL\ Spring\ Semester\ 2021/CMDA\ /
Assignments/HW5/sales.xlsx")
#Selected colums 6 to 8 for turnover and 1 to 3 for sales
total_turnover <- select(sales,6:8)
unicycle <- select(sales,1:3)
#Set up to be able to merge
names(unicycle)[1] <- "countryname"
names(unicycle)[2] <- "year"
names(total_turnover)[1] <- "countryname"
names(total_turnover)[2] <- "year"
#Merging dataframe by countryname and year in sales
sales <- merge(total_turnover,unicycle,by = c("countryname","year"),all.x=T)
#Getting population observations
obs$countryname <- toupper(obs$countryname)
#Displaying first 10 rows of dataframe by left join sales and observations
unicorns.cycling<-left_join(sales,obs)
unicorns.cycling[1:10,]
```

	countryname <chr>	year <dbl>	total_turnover <dbl>	unicycles <dbl>	pop <dbl>
1	AUSTRIA	1670	5274.0	60	85
2	AUSTRIA	1671	5186.1	59	83
3	AUSTRIA	1674	4658.7	53	75
4	AUSTRIA	1675	5098.2	58	82
5	AUSTRIA	1676	4922.4	56	79
6	AUSTRIA	1677	4395.0	50	70
7	AUSTRIA	1678	5010.3	57	81
8	AUSTRIA	1680	5010.3	57	80
9	FRANCE	1673	4321.8	49	70
10	FRANCE	1674	4939.2	56	79

1-10 of 10 rows

b. Report a summary table of this new “tidy” data frame.

```
#Used summary function
summary(unicorns.cycling)
```

```
 countryname             year      total_turnover   unicycles
 Length:42          Min.   :1670   Min.   :3758   Min.   :44.00
 Class :character   1st Qu.:1673   1st Qu.:4663   1st Qu.:53.00
 Mode  :character   Median :1676   Median :4938   Median :56.00
                    Mean   :1676   Mean   :4989   Mean   :57.29
                    3rd Qu.:1678   3rd Qu.:5288   3rd Qu.:60.00
                    Max.   :1680   Max.   :6380   Max.   :73.00
      pop
 Min.   : 63.00
 1st Qu.: 75.25
 Median : 80.00
 Mean   : 81.71
 3rd Qu.: 86.50
 Max.   :104.00
```

c. Use `dplyr`, create new variables that calculate the mean, min, and max of both `unicorns` and `unicycles` by year.

```
library(dplyr)
#Groupping together unicorn.cycling with the year as a variable
together<-group_by(unicorns.cycling,year)
#Using summerise function to get mean,min and max
summarise(together,"Mean population of unicorns"=mean(pop))
```

year <dbl>	Mean population of unicorns <dbl>
1670	85.00000
1671	77.33333
1672	85.66667
1673	81.25000
1674	82.50000
1675	80.80000
1676	81.00000
1677	78.60000
1678	84.00000
1679	86.33333

1-10 of 11 rows Previous **1** 2 Next

```
summarise(together,"Min population of unicorns"=min(pop))
```

year <dbl>	Min population of unicorns <dbl>

year <dbl>	Min population of unicorns <dbl>
1670	85
1671	70
1672	79
1673	69
1674	75
1675	64
1676	68
1677	63
1678	75
1679	79

1-10 of 11 rows Previous **1** 2 Next

```
summarise(together,"Max population of unicorns"=max(pop))
```

year <dbl>	Max population of unicorns <dbl>
1670	85
1671	83
1672	98
1673	96
1674	94
1675	99
1676	104
1677	99
1678	101
1679	93

1-10 of 11 rows Previous **1** 2 Next

```
summarise(together,"Mean unicycles"=mean(unicycles))
```

year <dbl>	Mean unicycles <dbl>

year <dbl>	Mean unicycles <dbl>
1670	60.00000
1671	54.00000
1672	59.66667
1673	56.75000
1674	58.00000
1675	56.60000
1676	56.80000
1677	55.20000
1678	59.00000
1679	60.33333

1-10 of 11 rows Previous **1** 2 Next

```
summarise(together,"Min unicycles"=min(unicycles))
```

year <dbl>	Min unicycles <dbl>
1670	60
1671	49
1672	54
1673	48
1674	53
1675	45
1676	48
1677	44
1678	53
1679	55

1-10 of 11 rows Previous **1** 2 Next

```
summarise(together,"Max unicycles"=max(unicycles))
```

year <dbl>	Max unicycles <dbl>

year <dbl>	Max unicycles <dbl>
1670	60
1671	59
1672	69
1673	68
1674	66
1675	70
1676	73
1677	70
1678	71
1679	66

1-10 of 11 rows Previous **1** 2 Next

d. Write your own R function to calculate the max number of unicorns and unicylces for each year. Return a list/data frame with the year and values.

```
#Calling function and creating if and else statement to get max num of unicorns and unic
ycles
grande<- function(unicorns.cycling){
  for(i in 1:nrow(unicorns.cycling)){
    if(unicorns.cycling$unicycles<unicorns.cycling$pop){
      get_max<-unicorns.cycling$pop
    }
    else{
      get_max<-unicorns.cycling$unicycles
    }
    unicorns.cycling["get_max"]<-get_max
  }
  unicorns.cycling %>%
    select(year,get_max)
}
#Using function on unicorns.cycling
grande(unicorns.cycling)
```

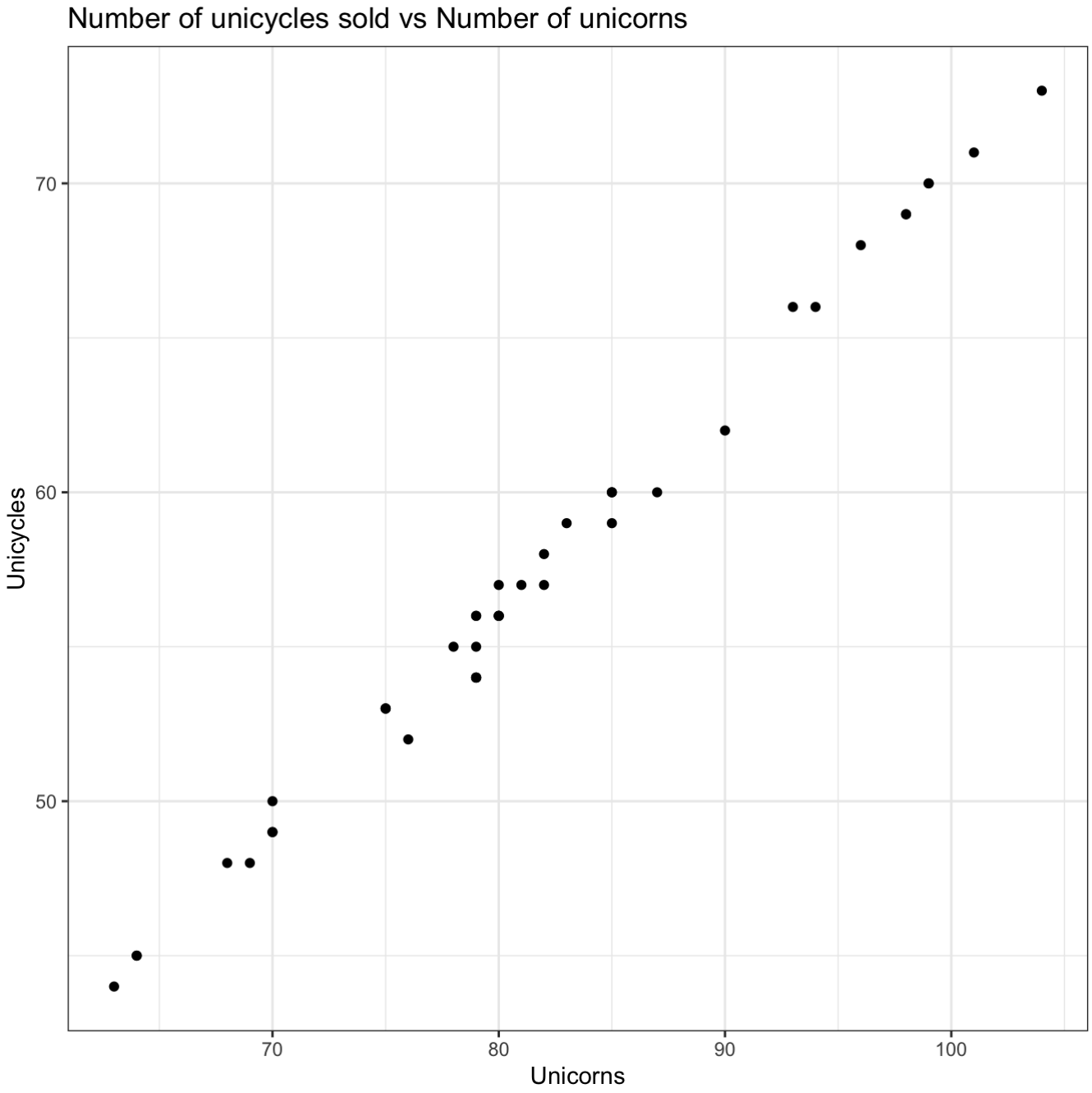
year <dbl>	get_max <dbl>
1670	85
1671	83
1674	75
1675	82

year <dbl>	get_max <dbl>
1676	79
1677	70
1678	81
1680	80
1673	70
1674	79

1-10 of 42 rows Previous **1** 2 3 4 5 Next

e. Use `ggplot()` to plot a scatterplot for the number of unicycles sold versus the number of unicorns.

```
#Dsiplaying ggplot with layers
 ggplot(data = unicorns.cycling,aes(x = pop, y = unicycles))+
  ggtitle("Number of unicycles sold vs Number of unicorns")+
  scale_x_continuous("Unicorns") +
   scale_y_continuous("Unicycles")+geom_point() + theme_bw()
```



f. Comment on how you feel using 1 for this type of problem differed from using functions in `Base R`. Do you think this process was easier or harder? There's no wrong answer here, every R programmer has their own personal preference, just give your own personal opinion on the `tidyverse`.

```
#I don't like it, I prefer using base R but, will try to get used to it.
```

Problem 2: [50 pts] More Practice with EDA and Data Wrangling.

The data comes from Marvel Wikia and DC Wikia.

The data is split into two files, for DC and Marvel, respectively: `dc-wikia-data.csv` and `marvel-wikia-data.csv`.

Each file has the following variables:

Variable}	Definition
page_id	The unique identifier for that characters page within the wikia
name	The name of the character
	urlslug
ID	The identity status of the character (Secret Identity, Public identity, [on marvel only: No Dual Identity])
ALIGN	If the character is Good, Bad or Neutral
EYE	Eye color of the character
HAIR	Hair color of the character
SEX	Sex of the character (e.g. Male, Female, etc.)
GSM	If the character is a gender or sexual minority (e.g. Homosexual characters, bisexual characters)
ALIVE	If the character is alive or deceased
APPEARANCES	The number of appareances of the character in comic books (as of Sep. 2, 2014. Number will become increasingly out of date as time goes on.)
FIRST APPEARANCE	The month and year of the character's first appearance in a comic book, if available
YEAR	The year of the character's first appearance in a comic book, if available

Note: These data sets have plenty of missing values so you will need to handle them appropriately depending on the situation.

a. Let's do some very basic data cleaning. `FIRST APPEARANCE` is formatted differently in the two data sets. They both contain the month and year for first appearance. Additionally the variable `YEAR` contains the Year of first appearance (so there is a redundancy). Rename the `FIRST APPEARANCE` variable to `MONTH` and fix the observations so that they will be the Month given by the full word, i.e. March instead of Mar.

```
#Additional tidyverse package
library(lubridate)
#Loading excel files
DC_Wikia <- read.csv("/Users/eduardosalvador/Desktop/FINAL\ Spring\ Semester\ 2021/CMDA\
/Assignments/HW5/dc-wikia-data.csv", header = T)
Marvel_Wikia <- read.csv("/Users/eduardosalvador/Desktop/FINAL\ Spring\ Semester\ 2021/C
MDA\ /Assignments/HW5/marvel-wikia-data.csv", header = T)
#Renaming column 12 to MONTH
names(Marvel_Wikia)[12] <- "MONTH"
Marvel_Wikia$MONTH <- format(my(Marvel_Wikia$MONTH),"%B")
names(DC_Wikia)[12] <- "MONTH"
DC_Wikia$MONTH <- format(my(DC_Wikia$MONTH),"%B")
#Displaying headers
head(DC_Wikia)
```

	page_id	name	urlslug	ID
	<int>	<chr>	<chr>	<chr>
1	1422	Batman (Bruce Wayne)	\\/wiki\\/Batman_(Bruce_Wayne)	Secret I
2	23387	Superman (Clark Kent)	\\/wiki\\/Superman_(Clark_Kent)	Secret I
3	1458	Green Lantern (Hal Jordan)	\\/wiki\\/Green_Lantern_(Hal_Jordan)	Secret I
4	1659	James Gordon (New Earth)	\\/wiki\\/James_Gordon_(New_Earth)	Public I
5	1576	Richard Grayson (New Earth)	\\/wiki\\/Richard_Grayson_(New_Earth)	Secret I
6	1448	Wonder Woman (Diana Prince)	\\/wiki\\/Wonder_Woman_(Diana_Prince)	Public I

6 rows | 1-5 of 14 columns

```
head(Marvel_Wikia)
```

	page_id	name	urlslug
	<int>	<chr>	<chr>
1	1678	Spider-Man (Peter Parker)	\\/Spider-Man_(Peter_Parker)
2	7139	Captain America (Steven Rogers)	\\/Captain_America_(Steven_Rogers)
3	64786	Wolverine (James \\"Logan\\" Howlett)	\\/Wolverine_(James_%22Logan%22_Howlett)
4	1868	Iron Man (Anthony \\"Tony\\" Stark)	\\/Iron_Man_(Anthony_%22Tony%22_Stark)
5	2460	Thor (Thor Odinson)	\\/Thor_(Thor_Odinson)
6	2458	Benjamin Grimm (Earth-616)	\\/Benjamin_Grimm_(Earth-616)

6 rows | 1-4 of 14 columns

b. Determine how many missing observations we have for each of the variables in the two data frames and display this in a table.

```
#Showing hos many missing obsevations there are for DC and Marvel (is.na)
table(is.na(Marvel_Wikia))
```

```
 FALSE   TRUE
210162   2726
```

```
table(is.na(DC_Wikia))
```

```
FALSE  TRUE
82328  7320
```

c. Has the proportion of Female Characters improved over time (in years)? Investigate this question using tables and plots for both DC and Marvel comics. Is DC or Marvel doing better with regards to Female Characters overall?

```
#Marvel female character grouping by year
FMarvel <- Marvel_Wikia %>%
  group_by(YEAR) %>%
  summarize(proportion = sum(SEX == "Female Characters"))
  FMarvel
```

YEAR <int>	proportion <int>
1939	10
1940	33
1941	15
1942	14
1943	13
1944	12
1945	12
1946	12
1947	13
1948	14

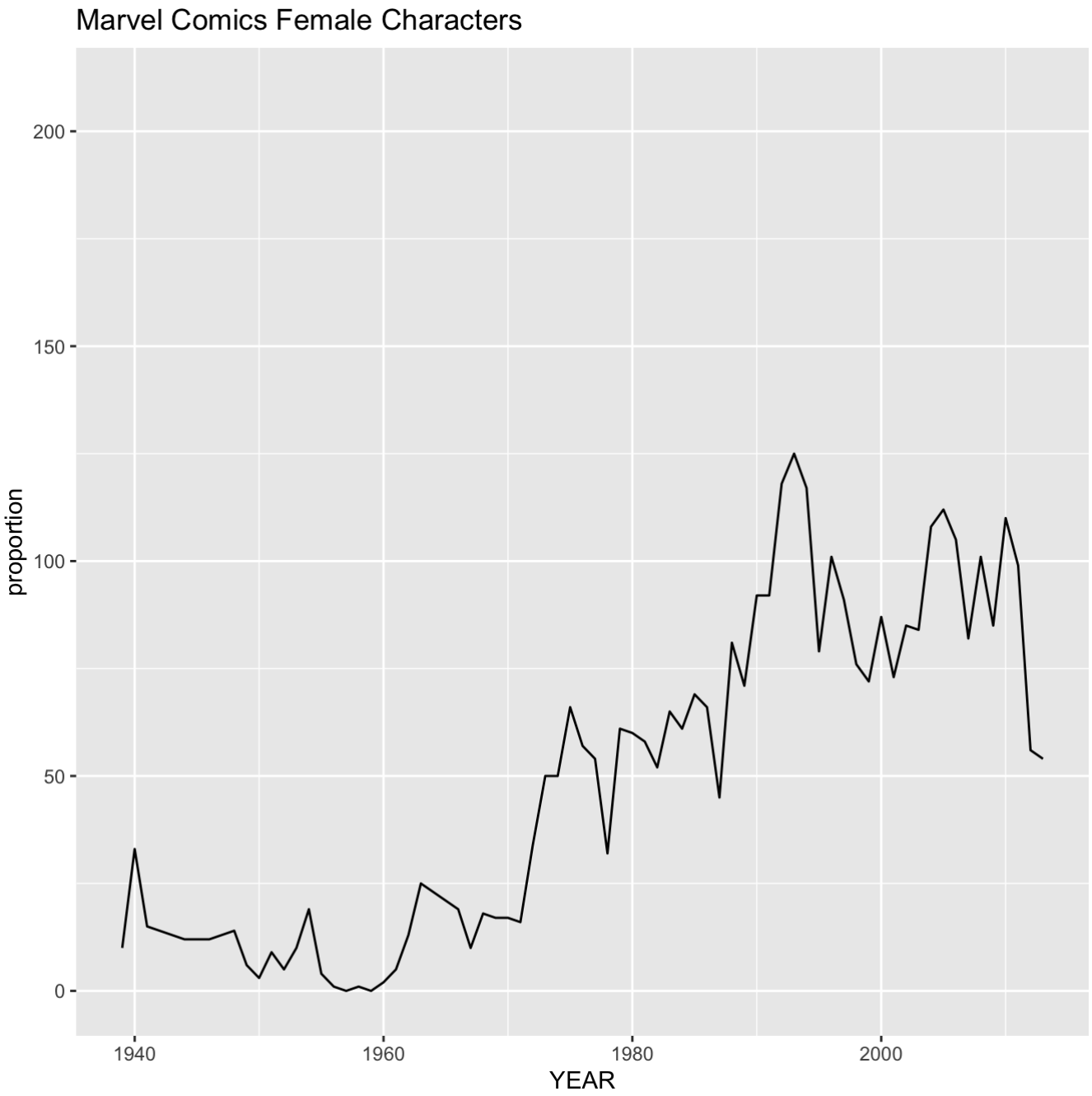
1-10 of 76 rows Previous **1** 2 3 4 5 6 … 8 Next

```
#DC females grouped by year
FDC <- DC_Wikia %>%
  group_by(YEAR) %>%
  summarize(proportion = sum(SEX == "Female Characters"))
  FDC
```

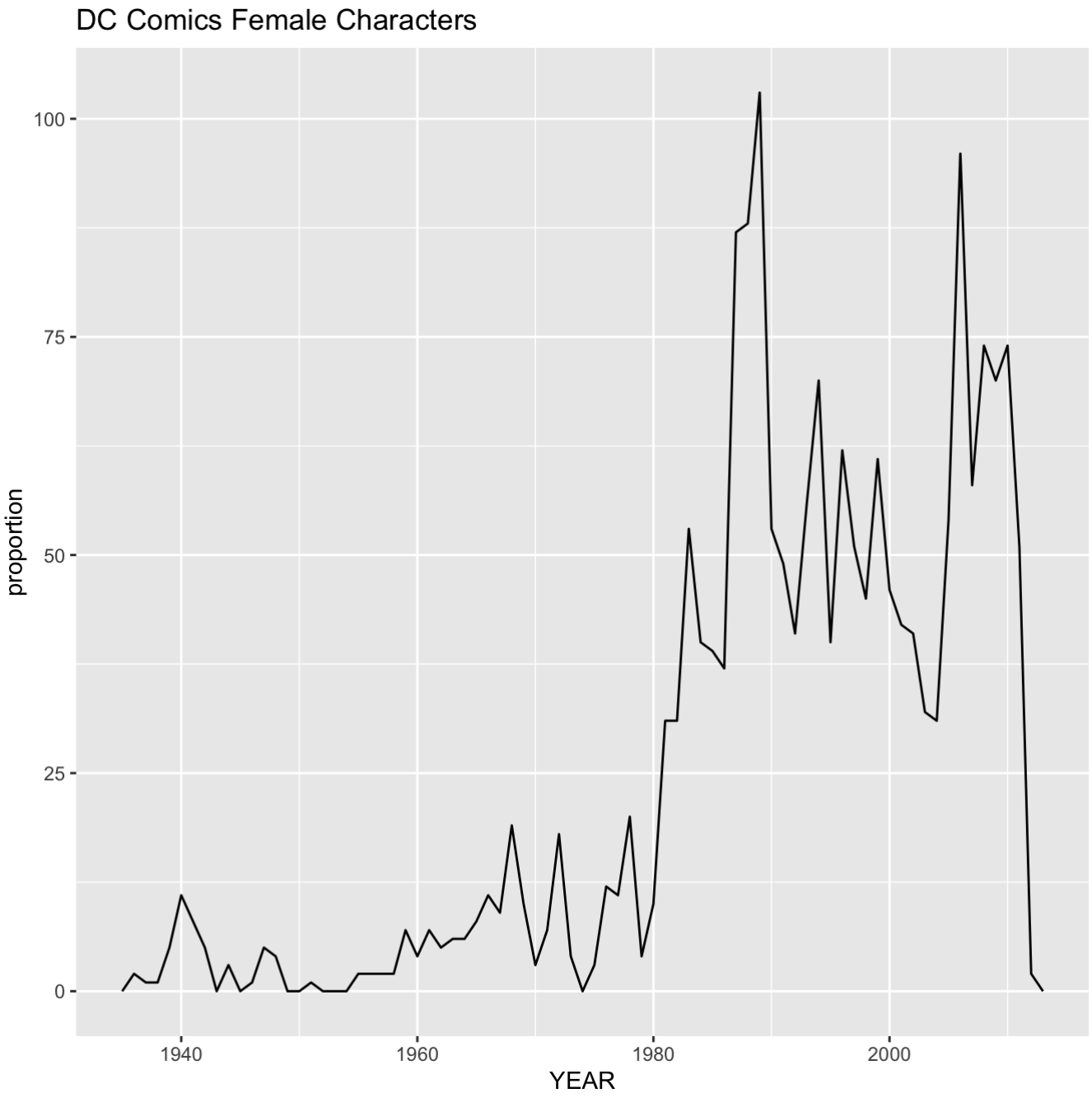
YEAR <int>	proportion <int>
1935	0
1936	2
1937	1
1938	1
1939	5
1940	11
1941	8
1942	5
1943	0
1944	3

1-10 of 80 rows Previous **1** 2 3 4 5 6 … 8 Next

```
#Displaying Marvel females character
ggplot(FMarvel) + geom_line(aes(x=YEAR,y=proportion)) + ggtitle("Marvel Comics Female Ch
aracters")
```




```
#Displaying DC female characters
ggplot(FDC) + geom_line(aes(x=YEAR,y=proportion)) + ggtitle("DC Comics Female Character
s")
```



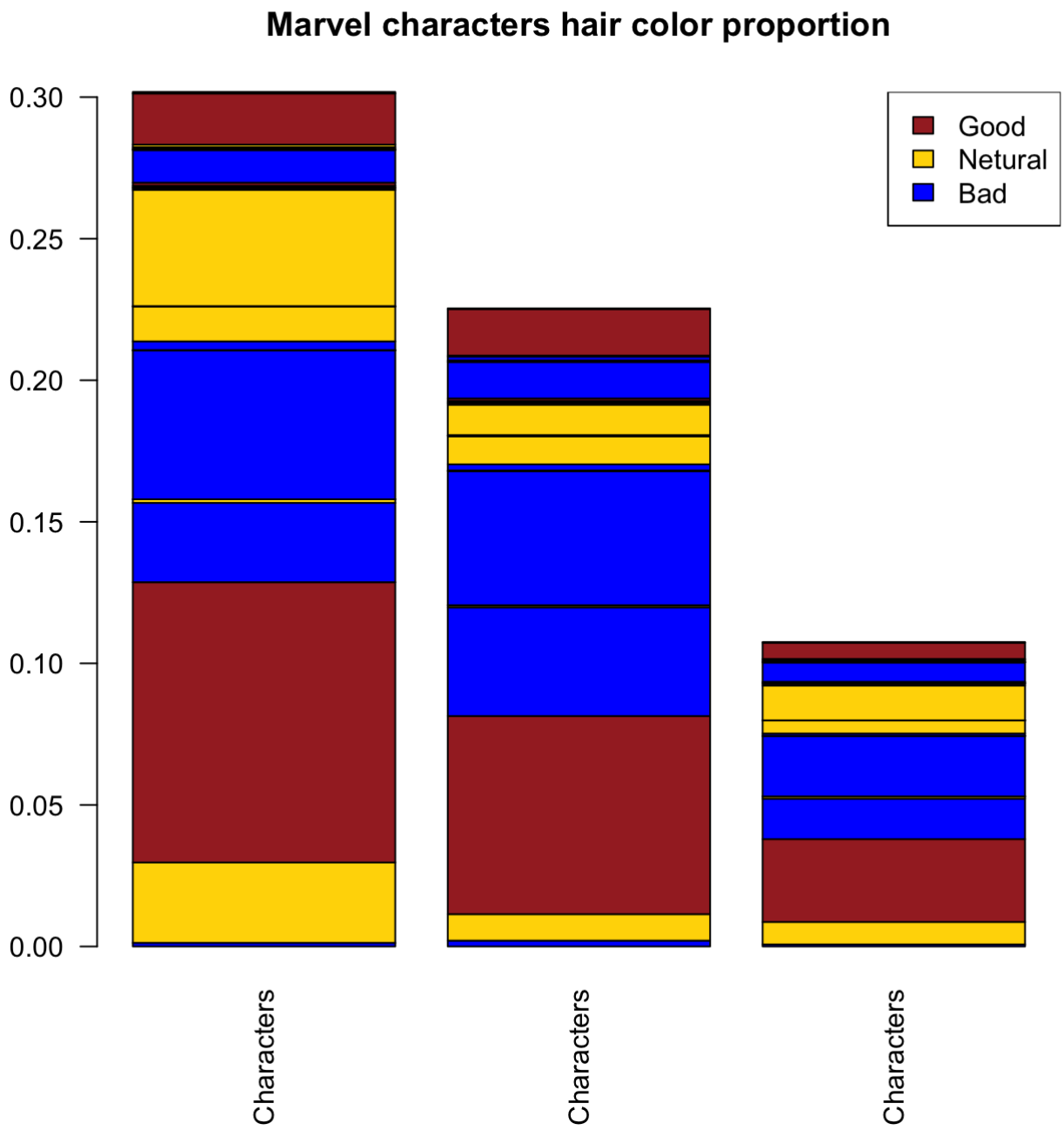

```
#DC is doing better with regards to female characters overall
```

d. Do artists prefer to associate different hair color with Good/Neutral/Bad characters? Determine the proportion of characters in the DC, Marvel, and Combined Datasets that have the different hair color (or bald) for the different alignments.Display this in a table and a stacked relative frequency barplot with Alignment on the x-axis.

```
#Creating table for Marvel
MTable<-table(Marvel_Wikia$HAIR,Marvel_Wikia$ALIGN)
MTable<-prop.table(MTable)

#Nonclassified characters excluded
MTable<-MTable[2:nrow(MTable),2:ncol(MTable)]

#Creating plot for Marvel
barplot(MTable,
        main="Marvel characters hair color proportion",
        las=2,
        col=c("blue","gold","brown"))
legend("topright", legend =c("Good","Netural","Bad"),fill=c("brown","gold","blue"))
```

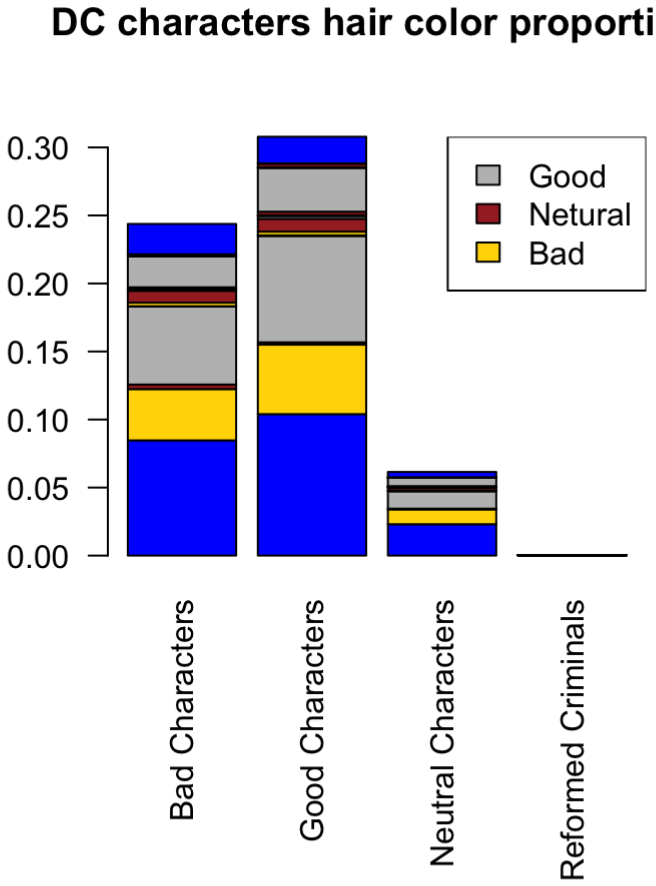


```
#Setting parameters
par(mar=c(18,15,6,6))

#Creating table for DC
DCTable<-table(DC_Wikia$HAIR,DC_Wikia$ALIGN)
DCTable<-prop.table(DCTable)

#Nonclassified characters excluded
DCTable<-DCTable[2:nrow(DCTable),2:ncol(DCTable)]

#Creating plot for Marvel
barplot(DCTable,
        main="DC characters hair color proportion",
        las=2,
        col=c("blue","gold","brown","grey"))
legend("topright", legend =c("Good","Netural","Bad"),fill=c("grey","brown","gold","blue"
))
```


DC characters hair color proportion
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Good
Netural
Bad
Bad Characters
Good Characters
Neutral Characters
Reformed Criminals

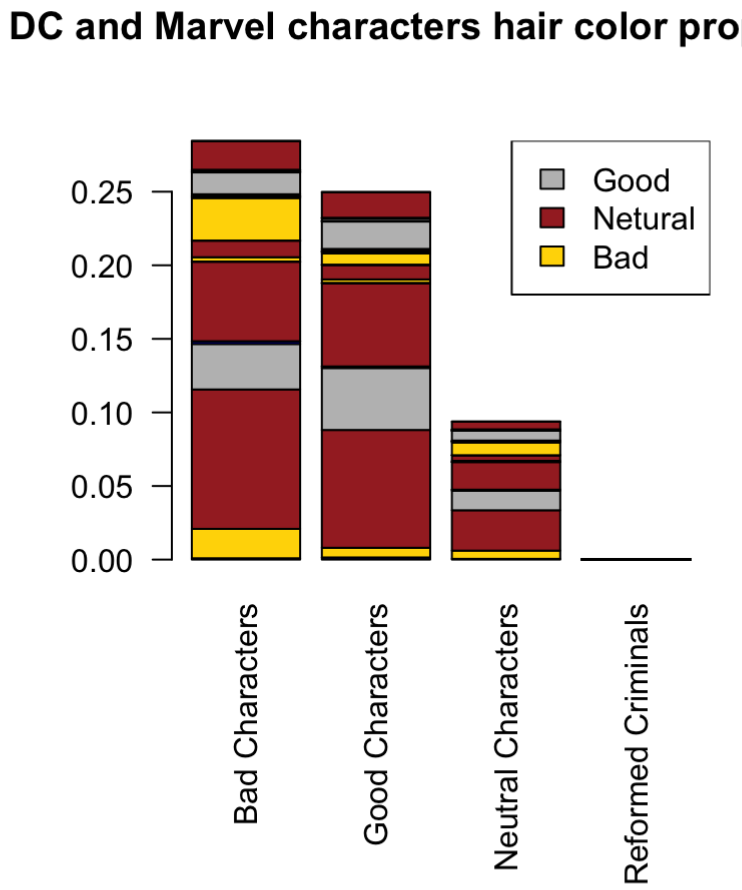
```
#Setting parameters
par(mar=c(18,15,6,6))

#Binding dataframes
MDC<-rbind(Marvel_Wikia,DC_Wikia)

#tables for Marvel and DC
MergeTab<-table(MDC$HAIR,MDC$ALIGN)
MergeTab<-prop.table(MergeTab)

#Nonclassified characters excluded
MergeTab<-MergeTab[2:nrow(MergeTab),2:ncol(MergeTab)]

#Creating plot for Marvel
barplot(MergeTab,
        main="DC and Marvel characters hair color proportion",
        las=2,
        col=c("blue","gold","brown","grey"))
legend("topright", legend =c("Good","Netural","Bad"),fill=c("grey","brown","gold","blue"
))
```



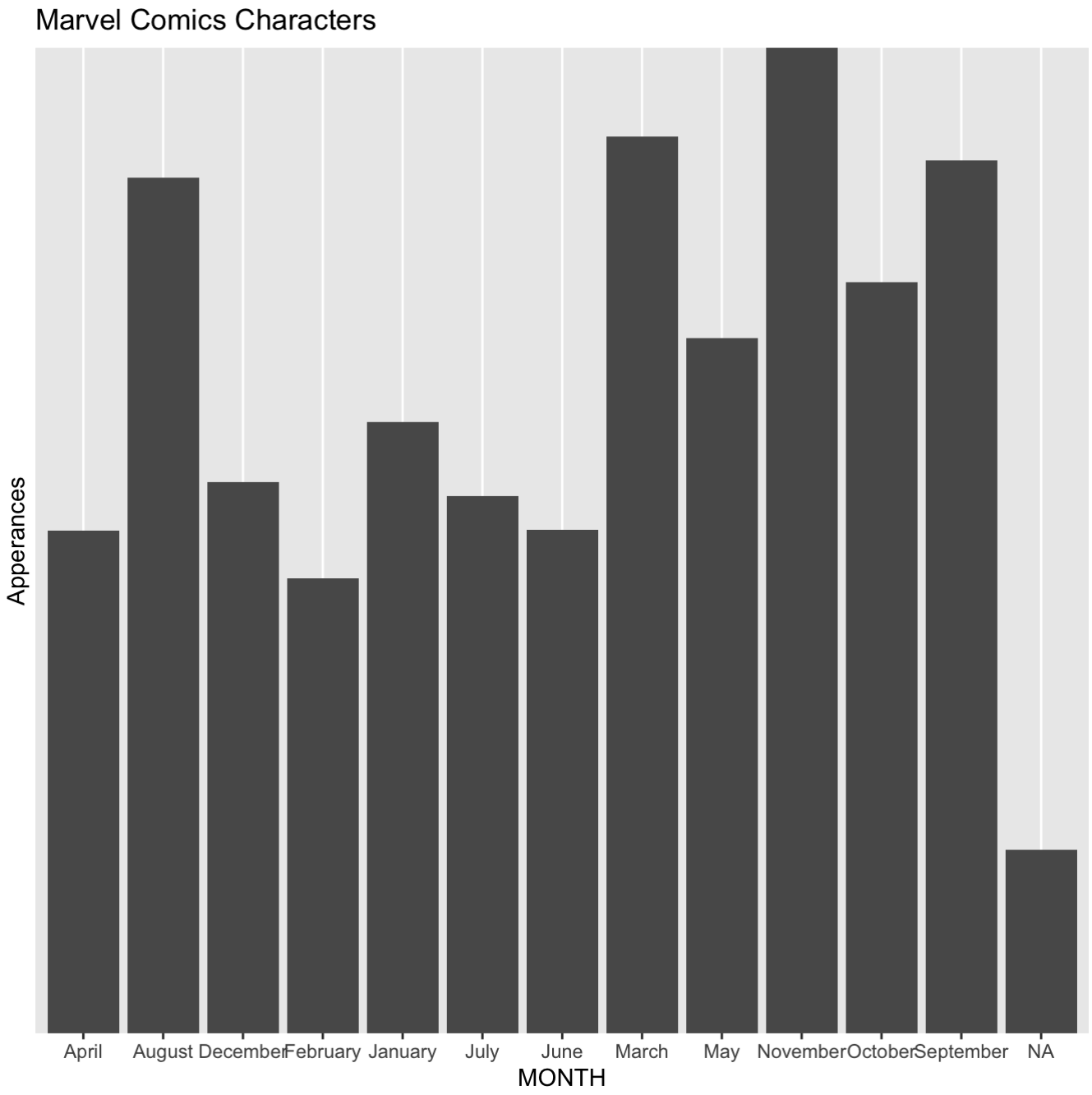

```
#resetting parameters
par(mar=c(18,15,6,6))
```

e. In what month do the most popular or most successful characters get introduced? To investigate this, determine the total number of appearances for all characters who were first introduced in January, February, and so forth. Do this for both DC and Marvel and present the results in a table and barplot. Which months tend to produce the “most popular” characters?

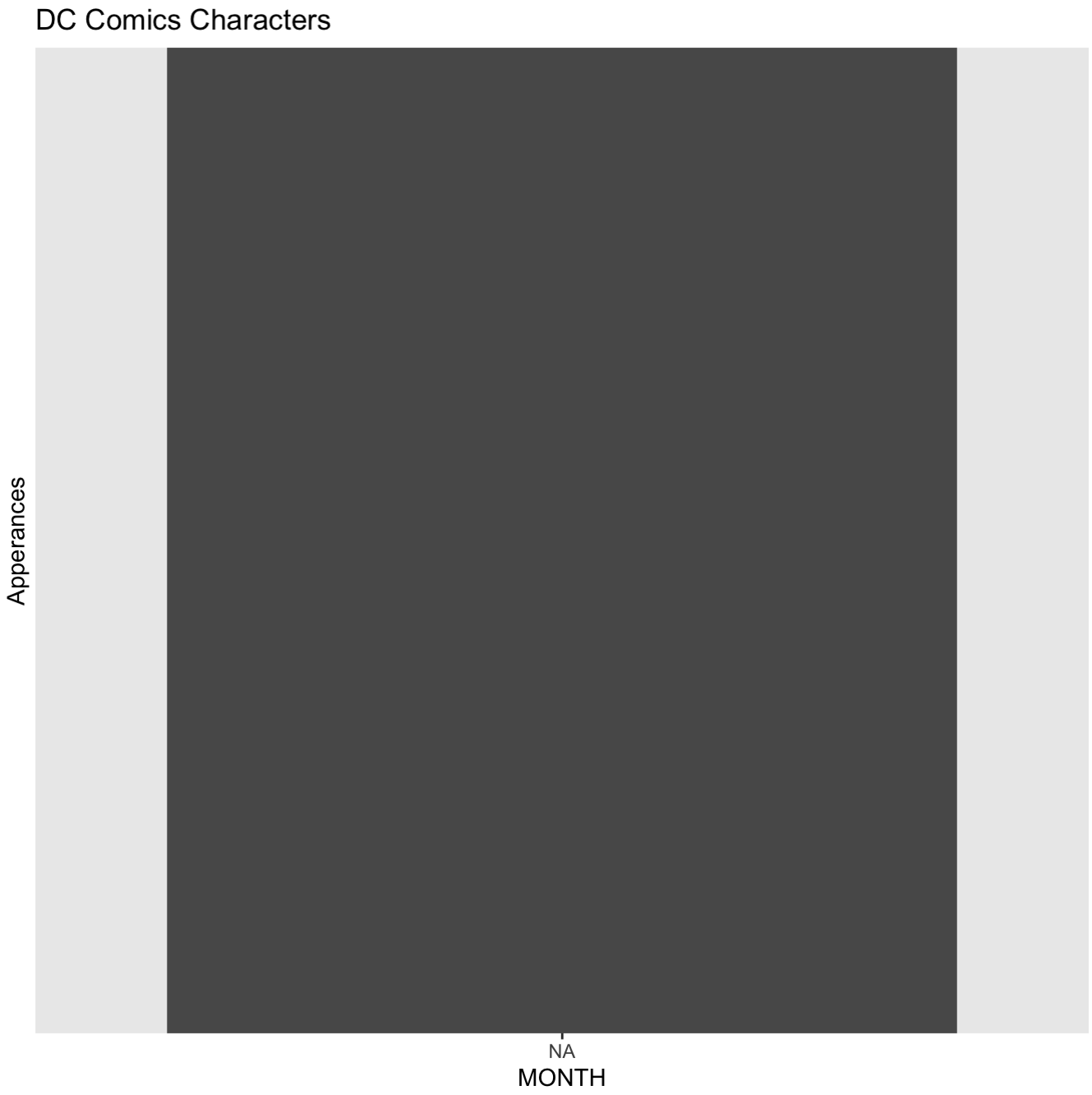
```
#Marvel population grouped by Month
Marvel_pop <- Marvel_Wikia %>%
  group_by(MONTH) %>%
  summarize(Everyap = sum(APPEARANCES,na.rm = T))

#DC population grouped by Month
DC_pop <- DC_Wikia %>%
  group_by(MONTH) %>%
  summarize(Everyap = sum(APPEARANCES,na.rm = T))


#Displaying Marvel population character
ggplot(Marvel_pop,aes(x=MONTH,y=Everyap))+geom_bar(stat="identity") +
  ggtitle("Marvel Comics Characters")+
  scale_x_discrete("MONTH") + scale_y_discrete("Apperances")
```




```
#Displaying Marvel population character
ggplot(data = DC_pop,aes(x=MONTH,y=Everyap))+geom_bar(stat="identity") +
  ggtitle("DC Comics Characters")+
   scale_x_discrete("MONTH") + scale_y_discrete("Apperances")
```




```
#Marvel produces more characters
```

Problem 3: [15 pts **Extra Credit**] Recoding Factors

Turn your attention to `adult.csv`. Notice that we have a slew of different categories in the marital status variable.

a. Turn your attention to the Marital variable. What are the unique factor levels? Do you feel as if we could combine some of these levels into more broad levels? Why or why not? If so, what do you suggest?

b. Using `Base R` functions only, convert this column into three categories: `Married`, `FormerlyMarried` and `NeverMarried`. Report a table of these new factor levels reporting number of observations at each level.

c. Repeat the same process as above, but using functions in the `dplyr` package.
