Lab2

Bhavya Patel

2024-09-15

##Q1

```
a=range(movies$year)
paste("The oldest movie was in ",a[1],sep=" ")
```

```
## [1] "The oldest movie was in  1893"
```

```
paste("This latest movie was in",a[2],sep=" ")
```

```
## [1] "This latest movie was in 2005"
```

The Range of years of production of the movies is from 1893 to 2005 according to this dataset.

##Q2

```
bn=as.integer(table(is.na(movies$budget))[1])
bm=as.integer(table(is.na(movies$budget))[2])
p1=bn/(bn+bm)
p1=p1*100
p2=bm/(bn+bm)
p2=p2*100
paste("This is the proportion of movies that have their budget included in this
      database", round(p2,2),"%",sep=" ")
```

```
## [1] "This is the proportion of movies that have their budget included in this \n      database 91.13
```

```
paste("This is the proportion of movies that donot have their budget included in
      this database", round(p1,2),"%",sep=" ")
```

```
## [1] "This is the proportion of movies that donot have their budget included in \n      this database
```

```
print("Here's the lsit top 5 expensive movies")
```

```
## [1] "Here's the lsit top 5 expensive movies"
```

```
mb <- movies %>%
  select(title, budget) %>%
  arrange(desc(budget)) %>%
  head(5)

mb[,1]
```

```
## # A tibble: 5 x 1
##   title
##   <chr>
## 1 Spider-Man 2
## 2 Titanic
## 3 Troy
```

```
## 4 Terminator 3: Rise of the Machines
## 5 Waterworld
```

The proportion of the movies that included their budget is given above and the top 5 expensive movies list is also provided above.

##Q3

```
ml<- movies %>%
  select(title,length) %>%
  arrange(desc(length)) %>%
  head(5)
ml
```

```
## # A tibble: 5 x 2
##   title                                              length
##   <chr>                                               <int>
## 1 Cure for Insomnia, The                               5220
## 2 Longest Most Meaningless Movie in the World, The     2880
## 3 Four Stars                                           1100
## 4 Resan                                                 873
## 5 Out 1                                                 773
```

The list of top 5 longest movies is provided above.

##Q4

```
ss <- movies %>%
  filter(Short==1) %>%
  select(title,length) %>%
  arrange(length)  %>%
  head(1)
paste("The shorted Short film is",ss[1],sep=" ")
```

```
## [1] "The shorted Short film is 17 Seconds to Sophie"
```

```
ls <- movies %>%
  filter(Short==1) %>%
  select(title,length) %>%
  arrange(desc(length))  %>%
  head(1)
paste("The longest Short film is", ls[1], sep=" ")
```

```
## [1] "The longest Short film is 10 jaar leuven kort"
```

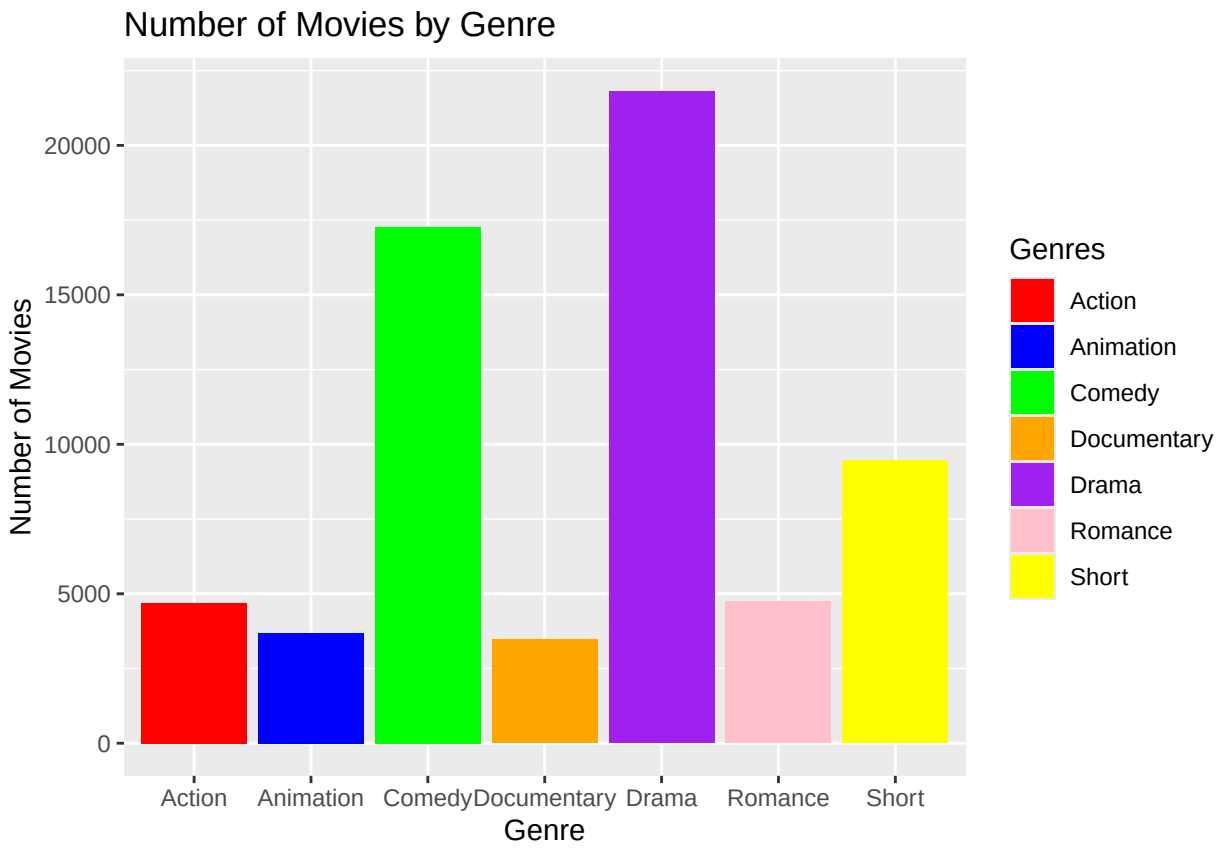
The shortest and longest Short film list is provided above.

##Q5

```
gc <- movies %>%
  pivot_longer(cols = Action:Short, names_to = "gc", values_to = "present") %>%
  filter(present == 1) %>%
  group_by(gc) %>%
  summarise(count = n())

ggplot(gc, aes(x = gc, y = count, fill = gc)) + geom_bar(stat = "identity") +
labs(title = "Number of Movies by Genre", x = "Genre", y = "Number of Movies") +
scale_fill_manual(values = c("Action" = "red", "Animation" = "blue",
    "Comedy" = "green", "Drama" = "purple", "Documentary" = "orange",
```

```
    "Romance" = "pink", "Short" = "yellow")) +
  guides(fill=guide_legend(title="Genres"))
```

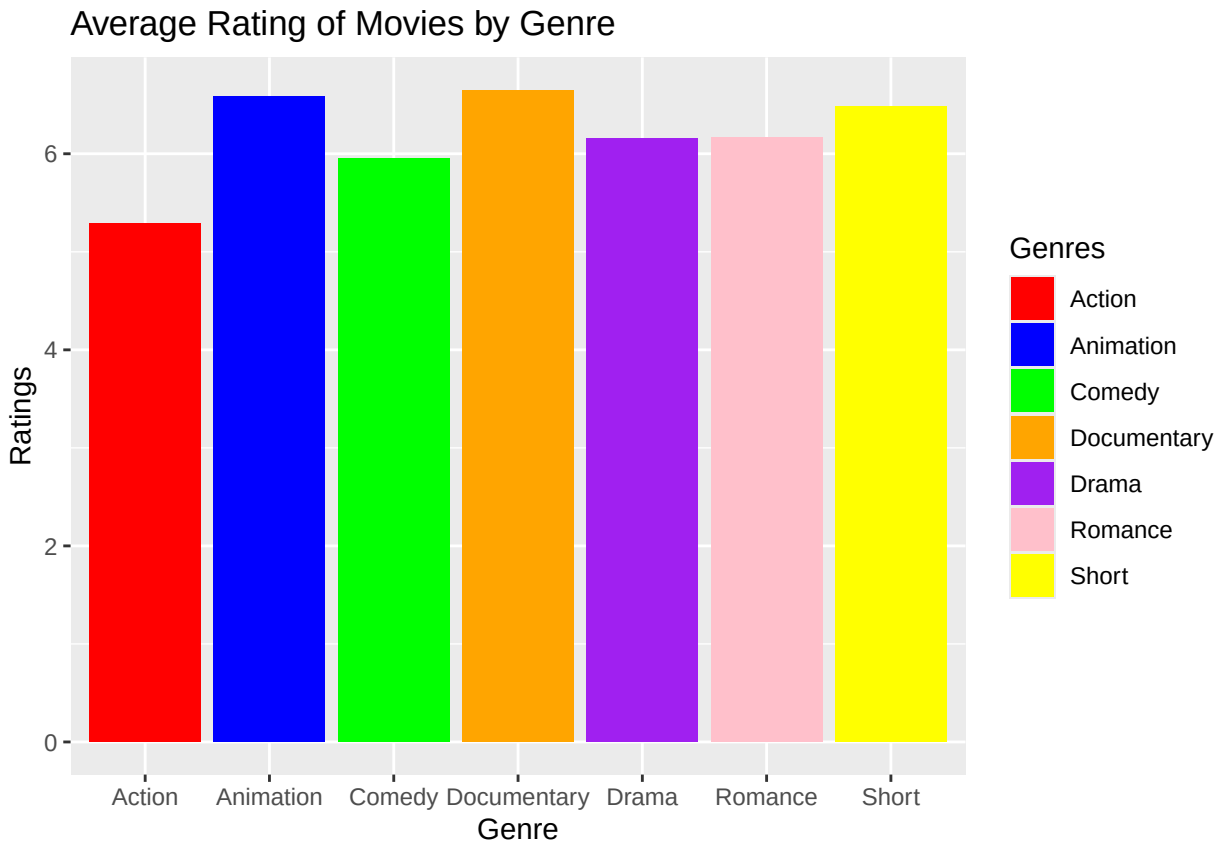


This bar chart represents number of movies in each Genre.

##Q6

```
gr <- movies %>%
  pivot_longer(cols = Action:Short, names_to = "gc", values_to = "present") %>%
  filter(present == 1) %>%
  group_by(gc) %>%
  summarise(avg = mean(rating, na.rm = TRUE))

ggplot(gr, aes(x = gc, y = avg, fill = gc)) + geom_bar(stat = "identity") +
  labs(title = "Average Rating of Movies by Genre", x = "Genre", y = "Ratings")+
  scale_fill_manual(values = c("Action" = "red", "Animation" = "blue",
  "Comedy" = "green", "Drama" = "purple", "Documentary" = "orange",
  "Romance" = "pink", "Short" = "yellow")) +
  guides(fill=guide_legend(title="Genres"))
```

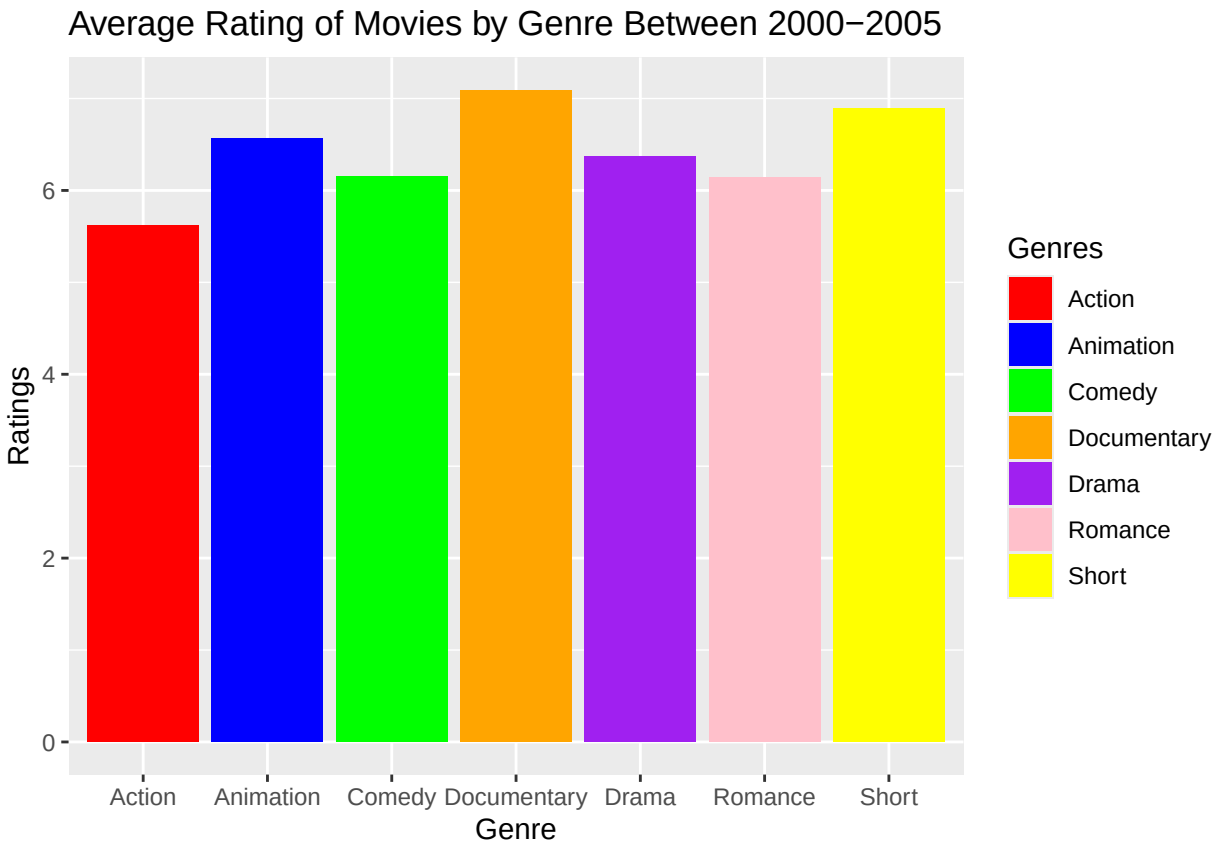


This bar chart represents average rating of movies in each Genre.

##Q7

```
gr2 <- movies %>%
  pivot_longer(cols = Action:Short, names_to = "gc", values_to = "present") %>%
  filter(present == 1, year >= 2000 & year <= 2005) %>%
  group_by(gc) %>%
  summarise(avg = mean(rating, na.rm = TRUE))

ggplot(gr2, aes(x = gc, y = avg, fill = gc)) + geom_bar(stat = "identity") +
  labs(title = "Average Rating of Movies by Genre Between 2000-2005", x = "Genre", y = "Ratings") +
  scale_fill_manual(values = c("Action" = "red", "Animation" = "blue",
  "Comedy" = "green", "Drama" = "purple", "Documentary" = "orange",
  "Romance" = "pink", "Short" = "yellow"))+
  guides(fill=guide_legend(title="Genres"))
```



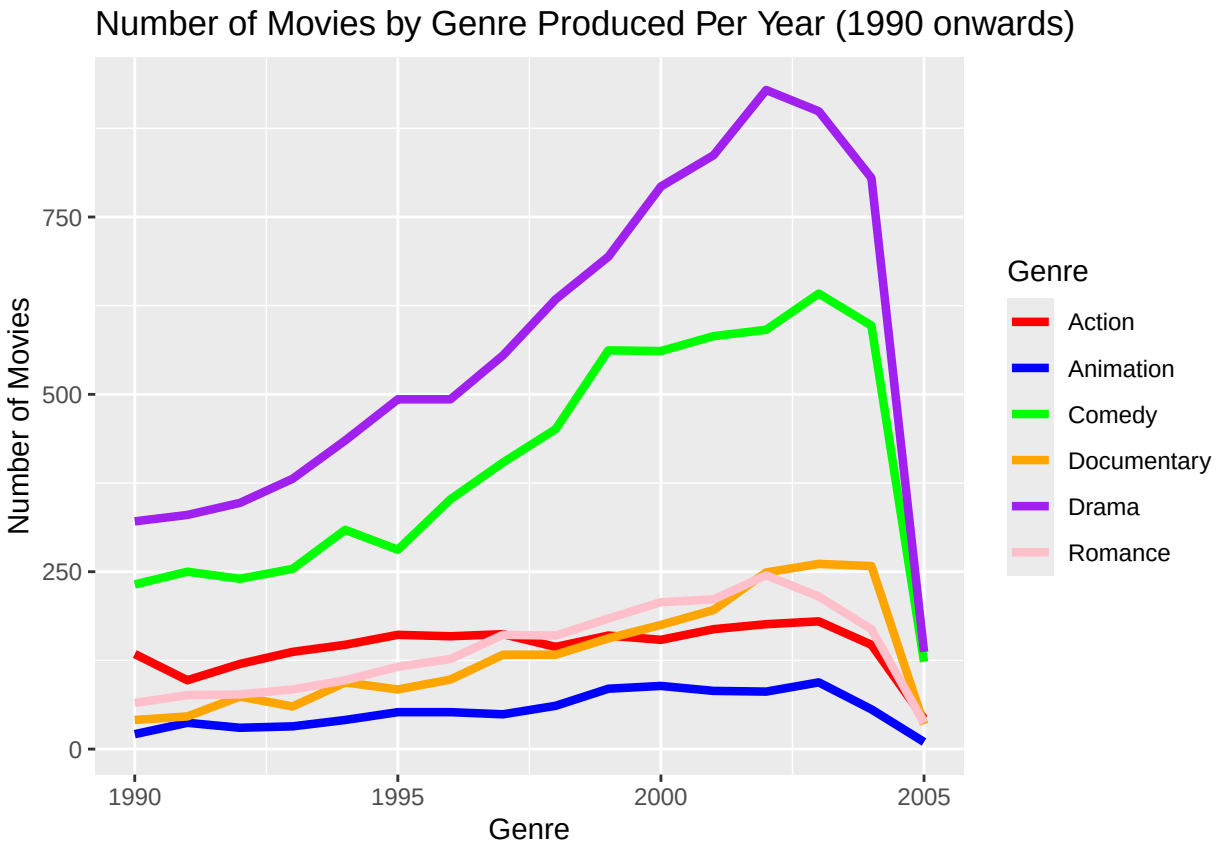
This bar chart represents average rating of movies in each Genre between 2000-2005.

##Q8

```
gm <- movies %>%
  pivot_longer(cols = Action:Romance, names_to = "gc", values_to = "present") %>%
  filter(present == 1 & year >= 1990) %>%
  group_by(year, gc) %>%
  summarise(count = n(), .groups = "drop")

ggplot(gm, aes(x = year, y = count, color = gc)) + geom_line(size = 1.5) +
  labs(title = "Number of Movies by Genre Produced Per Year (1990 onwards)",
       x = "Genre", y = "Number of Movies", color="Genre") +
  scale_color_manual(values = c("Action" = "red", "Animation" = "blue",
    "Comedy" = "green", "Drama" = "purple", "Documentary" = "orange",
    "Romance" = "pink"))
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



This line chart represents the number of movies produced yearly in each Genre except Short movies from 1990 onwards.

##Q9

```
paste("Question 1. What are the top 10 R-rated movies with best ratings?")
```

```
## [1] "Question 1. What are the top 10 R-rated movies with best ratings?"
```

```
tr <- movies %>%
  filter(mpaa=="R") %>%
  arrange(desc(rating)) %>%
  select(title, rating) %>%
  head(10)

#Answer 1:
paste("Top 10 best R-rated movies:")
```

```
## [1] "Top 10 best R-rated movies:"
```

```
tr[,1]
```

```
## # A tibble: 10 x 1
##    title
##    <chr>
##  1 Shawshank Redemption, The
##  2 Train Ride
##  3 Grey, The
##  4 Pulp Fiction
##  5 Schindler's List
```

```
##  6 Cidade de Deus
##  7 Memento
##  8 Usual Suspects, The
##  9 Cookers
## 10 Eternal Sunshine of the Spotless Mind
```

```
paste("Question 2. What movie has the best ratings in Action and Drama?")
```

```
## [1] "Question 2. What movie has the best ratings in Action and Drama?"
```

```
paste("How much was their budget, and what was the length of the movie?")
```

```
## [1] "How much was their budget, and what was the length of the movie?"
```

```
bad <- movies %>%
  filter(Action==1 & Drama==1) %>%
  arrange(desc(rating)) %>%
  select(title, rating, budget, length) %>%
  head(1)

#Answer2:
paste("The Best Rated movie in Action and Drama is", bad$title,
      "with ratings of", bad$rating, sep=" ")
```

```
## [1] "The Best Rated movie in Action and Drama is Morphin(e) with ratings of 9.7"
```

```
paste("Their budge was", bad$budget, "and the length of the movies was",
      bad$length, "minutes.", sep=" ")
```

```
## [1] "Their budge was 8000 and the length of the movies was 20 minutes."
```

```
paste("Question 3. What is the average budget for Action films over the past")
```

```
## [1] "Question 3. What is the average budget for Action films over the past"
```

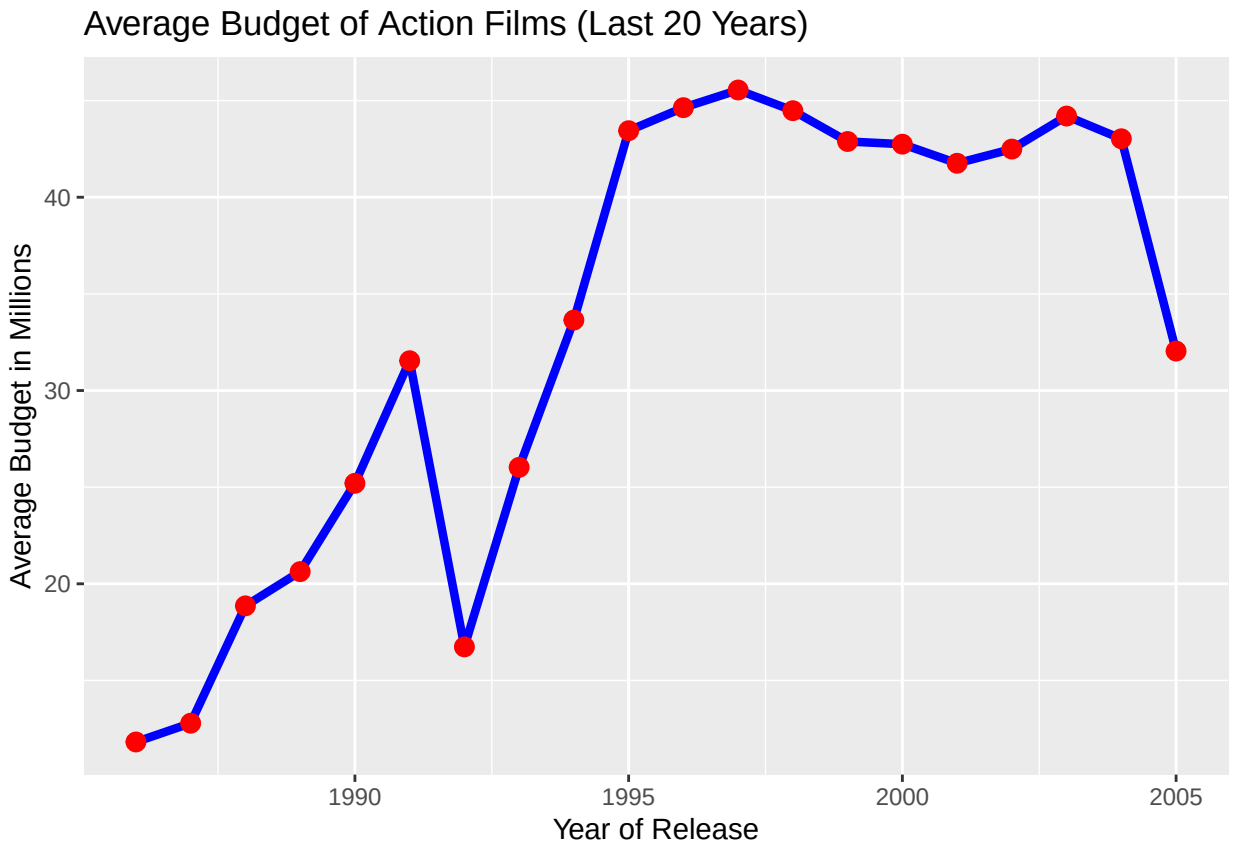
```
paste("20 years (use a line chart)?")
```

```
## [1] "20 years (use a line chart)?"
```

```
t1=max(movies$year)-20
aba <- movies %>%
  filter(year>t1, Action==1) %>%
  group_by(year) %>%
  summarise(avg = mean(budget, na.rm=T))

aba$avg <- aba$avg/1e6

ggplot(aba, aes(x = year, y = avg)) +
  geom_line(color = "blue", size = 1.5) +
  geom_point(color = "red", size = 3) +
  labs(title = "Average Budget of Action Films (Last 20 Years)",
       x = "Year of Release",
       y = "Average Budget in Millions")
```



Here I have added 3 of my own questions and answered them as well.