Homework 4

Grace Okamoto

22/9/2021

Problem 1

Write a brief description in your own words of each of the types of chart or graph below:

(a) Bar Chart

A bar chart is a chart that groups data based off category and visualizes it as horizontal or vertical bars that are proportionate to the values of the data.

(b) Scatter Plot

A scatter plot is a plot that uses two-dimensional graphing to display individual values for data.

(c) Line Plot

A line plot is similar to a scatter plot, except that it has a line connecting the majority of the data points together, generally eliminating outliers and giving the data a cleaner and more readable feel.

(d) Box Plot

A box plot, or a box-and-whiskers plot, is a plot useful for visualizing groups of data and their average values. It is useful for studying outliers and maximum and minimum values, the first and third quartile, and the mean.

(e) Histogram

A histogram is similar to a bar chart, but instead of the values being separated categorically, the data is grouped by intervals to give the chart a smoother and more coherent appearance.

(f) Pie Chart

A pie chart is a chart that is circular and partitioned to demonstrate proportions, generally in percentages.

Problem 2

The following problems ask you to analyze some ‘bad’ visualizations using the frameworks from class.

(a)

Pick one of the ‘bad’ visualizations on http://www.perceptualedge.com/examples.php (click on the images to see the descriptions) and summarize in your own words the changes that were made to update the plot.

I chose the causes of untimely death chart, not because it was the first option (well, not just because it was the first option, that is), but also because I despise charts that have nonlinear elements to them. This chart causes me additional anger because of the unlabled sections left in the design. Why are there blank sections? What are they for? Do they have a meaning or are they just in there for aesthetic purposes? Instead, the author of the article decides to improve this map by visualizing the data as three bar graphs, with the y variable being the various causes of death, and the three columns being the years of life lost in the particular year of 2010, the percent change in years of life lost between the range of 2005-2010, and the deaths in 2010.

(b)

Select one of the bad visualizations on https://www.businessinsider.com/the-27-worst-charts-of-all-time-2013-6 and provide brief responses to the following questions:

- (i) What is the visualization trying to communicate?

The visualization is trying to communicate the percent of LGBT-identifying individuals in each state.

- (ii) Why does it fail to be effective?

For some reason the whole map is green, meaning that the map is rendered as useful only for those who are unaware of what the United States looks like.

- (iii) How would you modify the plot to improve it according to the principles we discussed in class?

I believe this chart would be made more effective by using a gradient coloring scheme to denote higher vs. lower percentages of LGBT-identifiying individuals. Althought that was most likely what the chart was originally attempting to accomplish, it seems as if their percentage range for the coloring system was faulty. If they were to reduce the range to only that of the percentages encompassed in the data, then the spectrum of color would be utilized more successfully.

Problem 3

Load the data in GME_Stock.csv into R as a data frame or tibble. For parts (a), (b), and (c) below, create the plots using this dataset, making sure to label the axes and follow the other principles of good figure design:

```
library("tidyverse")
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

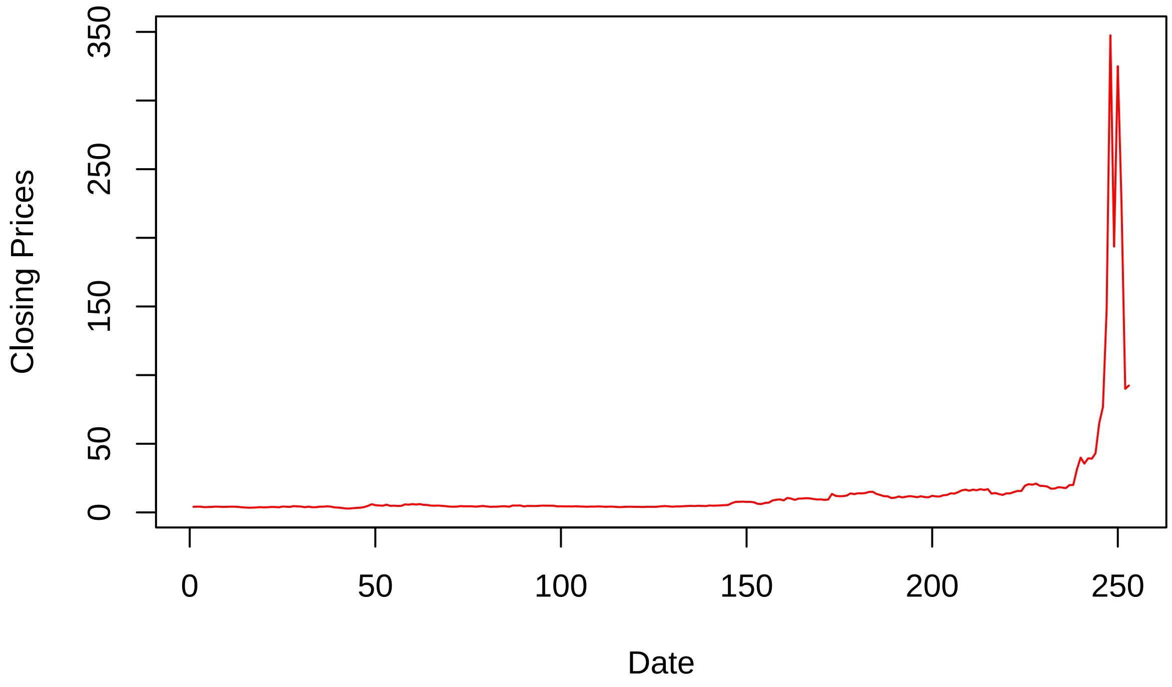
```
gme <- read.csv("GME_Stock.csv")
head(gme)
```

```
##        Date Open High  Low Close Adj.Close  Volume
## 1  2/4/2020 4.03 4.25 3.97  4.07      4.07 3563100
## 2  2/5/2020 4.15 4.41 4.14  4.18      4.18 2641700
## 3  2/6/2020 4.20 4.30 4.14  4.14      4.14 1510300
## 4  2/7/2020 4.11 4.13 3.77  3.81      3.81 2742300
## 5 2/10/2020 3.85 4.10 3.74  3.94      3.94 2777000
## 6 2/11/2020 3.98 4.24 3.95  4.02      4.02 3415000
```

(a)

Plot the values in the closing prices column as a line plot

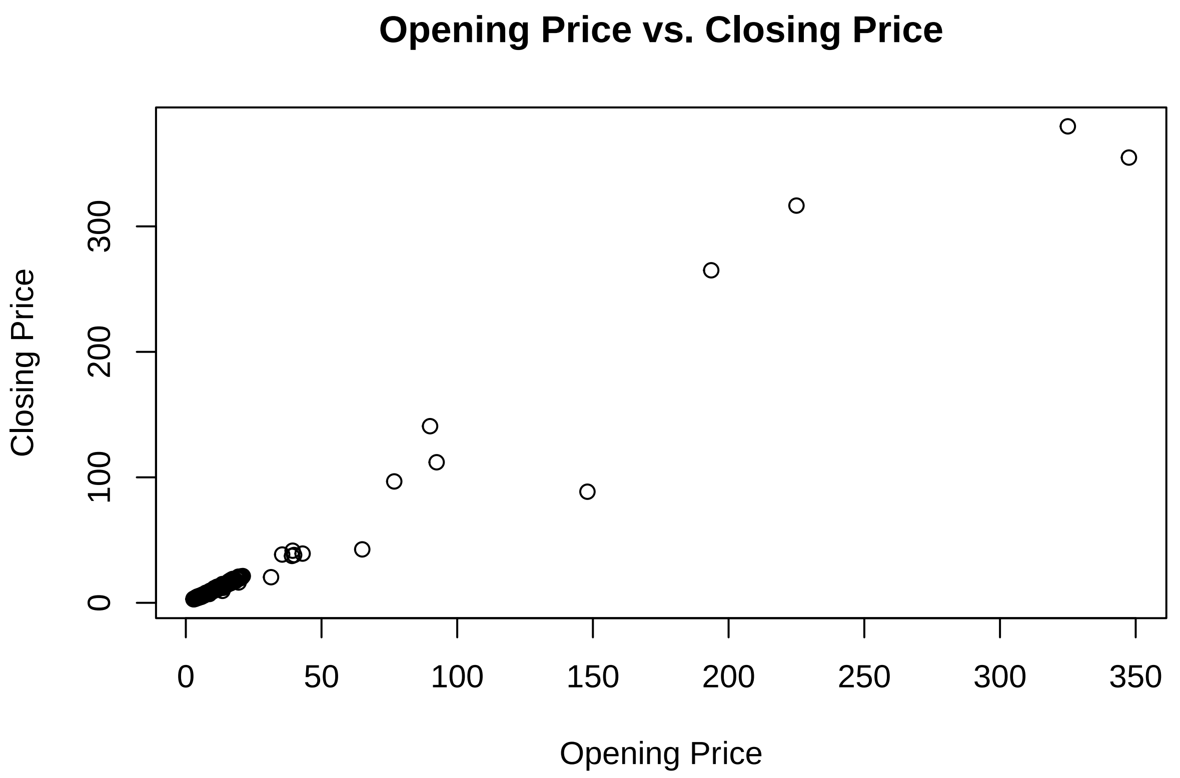
```
plot(gme$Close,type="l",ylab="Closing Prices",xlab="Date",col="red")
```



(b)

Make a scatterplot of opening prices vs. closing prices

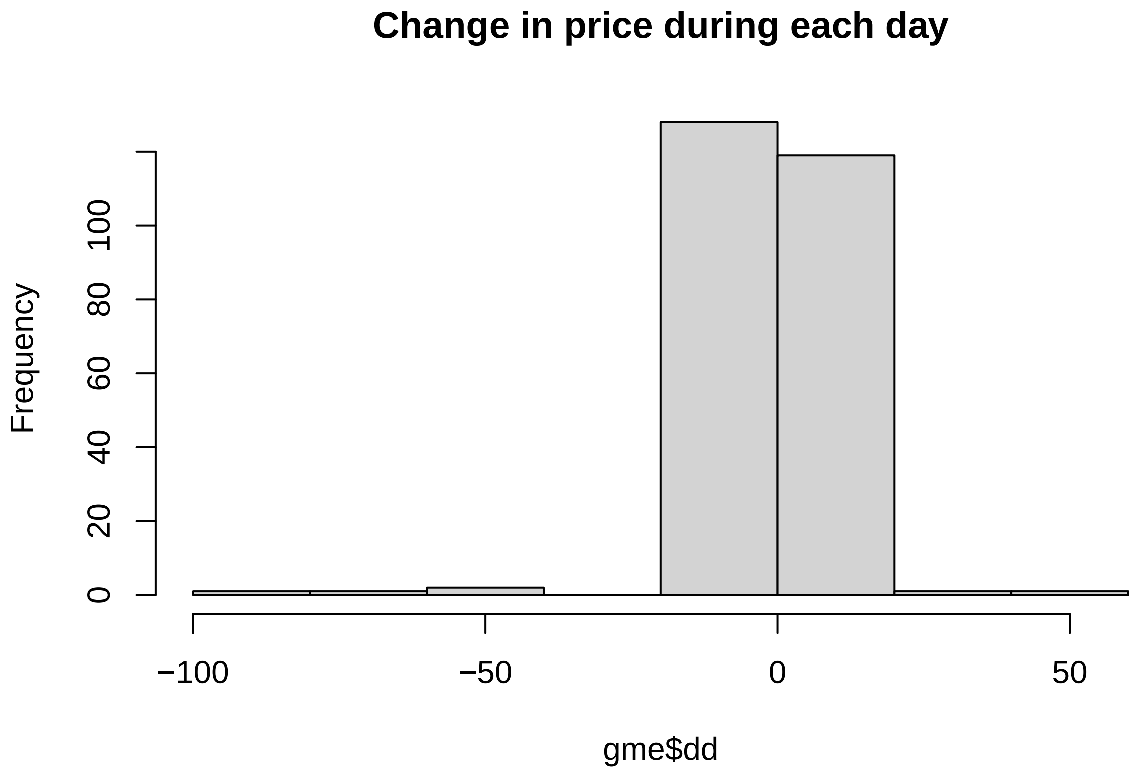
```
plot(gme$Open ~ gme$Close, xlab="Opening Price", ylab="Closing Price",main= "Opening Price vs. Closing
```



(c)

Add a new column to your dataframe whose rows represent the change in price during each day and make a histogram of these values.

```r
gme$dd <- gme$Close-gme$Open
hist(gme$dd, main = "Change in price during each day")
```



(d)

Make a bar chart presenting the in-state tuition data presented in this table:

$

School	WSU	EWU	UW	UI	CWU
Tuition	11841	7526	11465	8304	8273

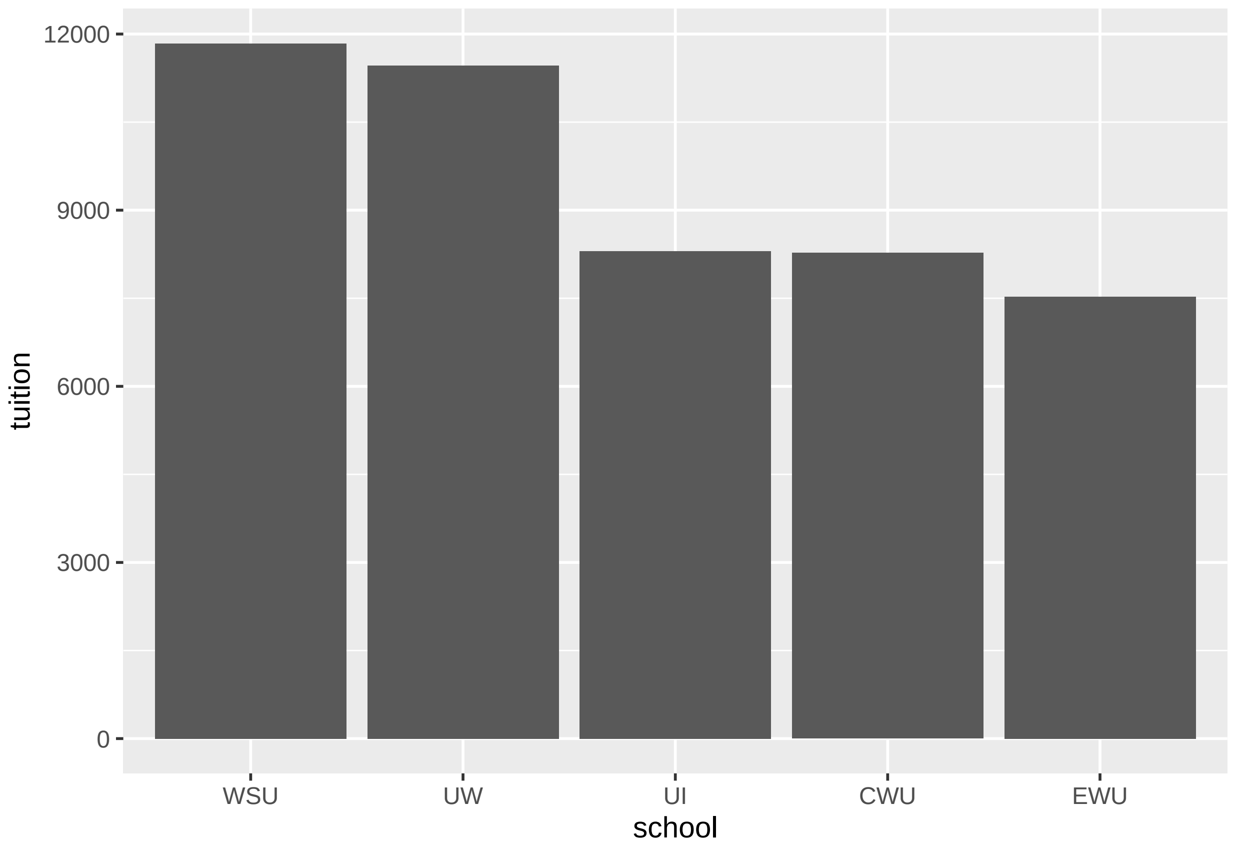
\ $

```
install.packages("ggplot2")
```

```
## Warning: package 'ggplot2' is in use and will not be installed
```

```
library("ggplot2")
bcd = data.frame(
  school = c("WSU","EWU","UW","UI","CWU"),
  tuition = c(11841,7526,11465,8304,8273)
)

bcd1 <- bcd
bcd1$school <- factor(bcd1$school,levels = bcd1$school[order(bcd1$tuition, decreasing = TRUE)])
ggplot(bcd1, aes(school, tuition)) + geom_bar(stat = "identity")
```



(e)

Provide a brief justification for the order you placed the bars from left to right in (d).

Honestly I have no idea what's going on anymore but I wanted to order it from increasing to decreasing tuition, since there didn't seem to be a logical way to order it by school. Therefore I did that, somehow, by visiting 20 different stackoverflow pages and copying their code until something looked okay. It turns out WSU is the most expensive, which is great for my wallet I suppose.

Problem 4

Read the Readme for the MovieRatings Dataset and pick one of the sub tables (such as byyearbyage or disneymovies) that interest you the most to do some exploratory analysis.

(a)

Load your chosen table as a dataframe or tibble in R.

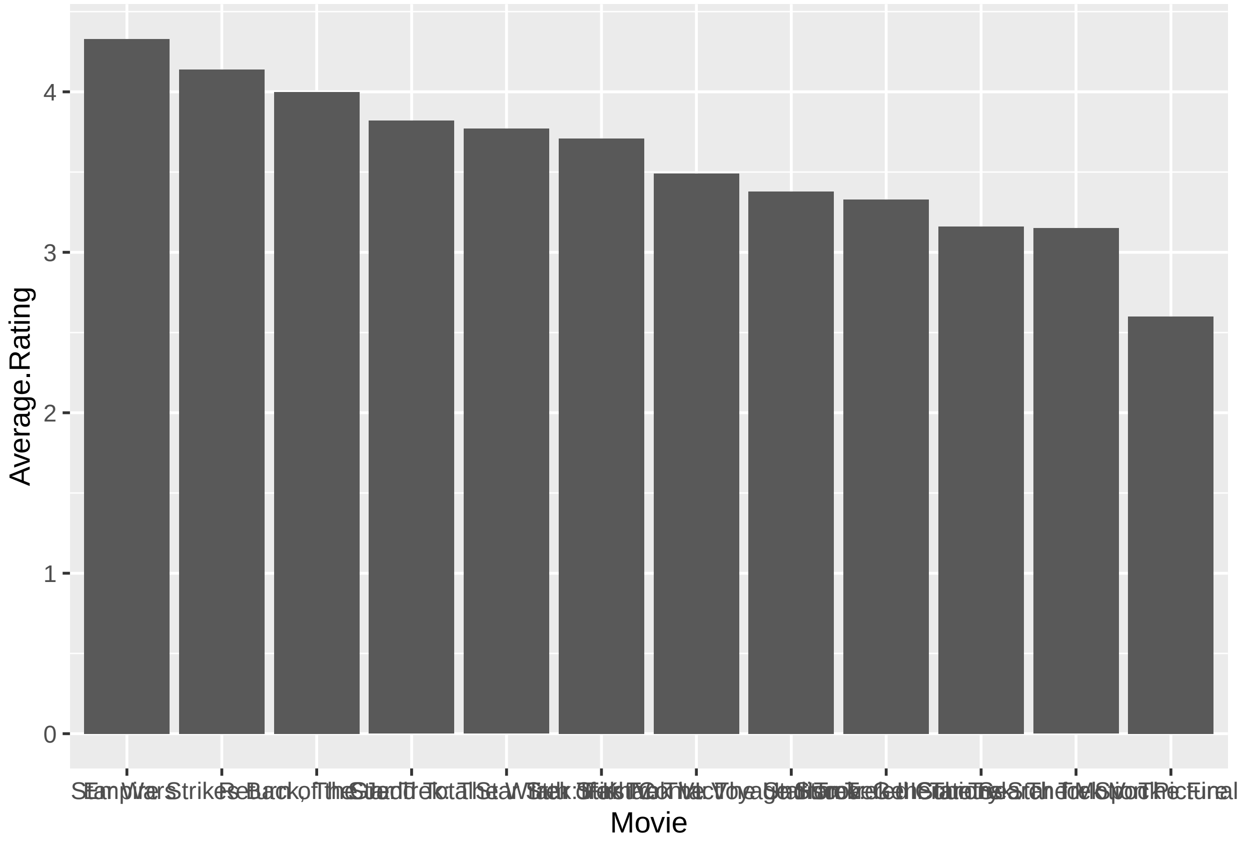
```
mr <- read.csv("Movie_Ratings/MovieRating_starTrekVsStarWars.csv")
head(mr)
```

```
##                                     Movie Number.of.Ratings Number.of.Women
## 1          Star Trek: The Motion Picture                73               8
## 2            Star Trek: The Wrath of Khan               152              25
## 3     Star Trek III: The Search for Spock               118              15
## 4          Star Trek IV: The Voyage Home               134              21
## 5         Star Trek V: The Final Frontier                43               6
## 6 Star Trek VI: The Undiscovered Country                  96              12
##   Number.of.Men Average.Rating Avg..Rating.Women Avg..Rating.Men
## 1            65           3.15              2.38            3.25
## 2           127           3.77              3.08            3.91
## 3           103           3.16              2.80            3.21
## 4           113           3.49              2.95            3.58
## 5            37           2.60              2.33            2.65
## 6            84           3.38              2.83            3.45
```

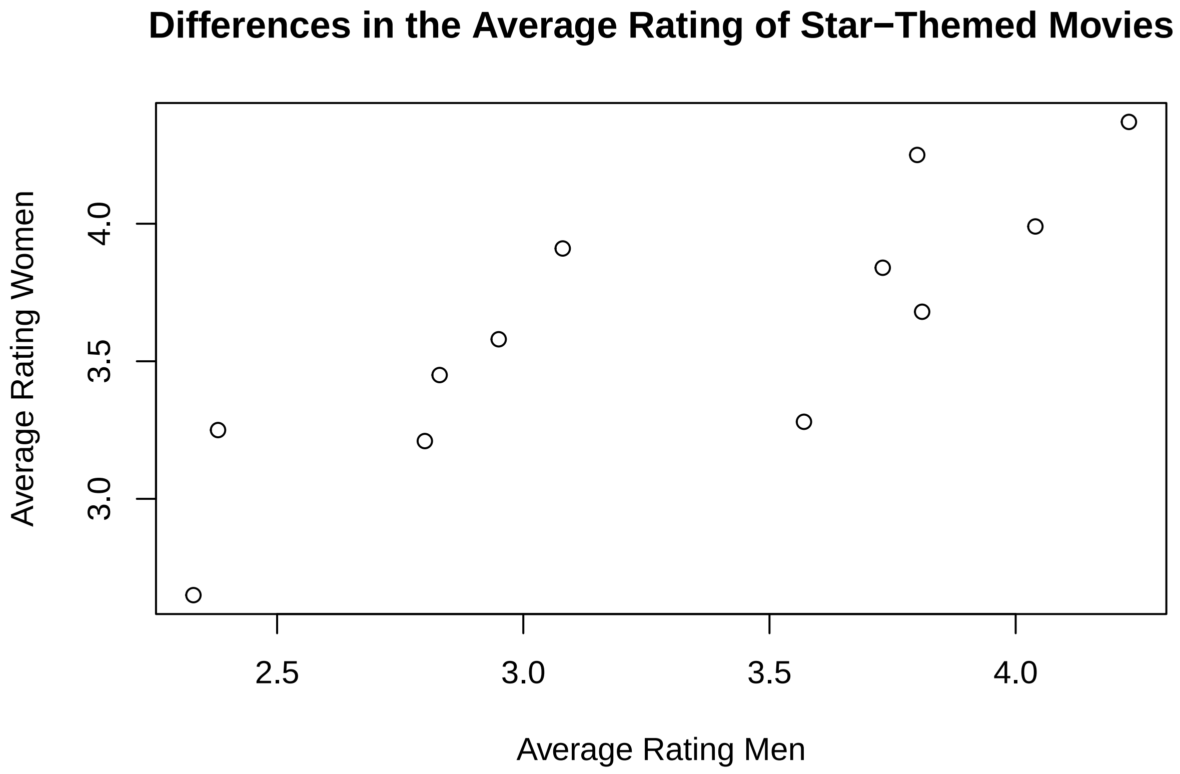
(b)

Explore the data by making at least two of the charts or plots we discussed this week.

```
mr$Movie <- factor(mr$Movie,levels = mr$Movie[order(mr$Average.Rating, decreasing = TRUE)])
ggplot(mr, aes(Movie, Average.Rating)) + geom_bar(stat = "identity")
```



```
plot(mr$Avg..Rating.Men ~ mr$Avg..Rating.Women, xlab="Average Rating Men", ylab="Average Rating Women",
```



(c)

Select one pattern that is revealed by your visualizations.

I suppose it could be deduced that for movies with Star in the title, men are slightly more likely to rate the movie higher. This is probably due to the fact that men like stars more.

(d)

Write a paragraph describing the data, the pattern that you observed, and how your plot displays that pattern.

In the first graph, the bar chart, the Star Wars and Star Trek movies are sorted by rating, where it can be seen (if you ignore the fact that the labels are going through some though circumstances) that Star Wars movies, on average, rate higher than Star Trek movies. In the scatter plot of the second graph, we can see that men are slightly more likely to rate these movies higher in most cases. Therefore, it seems reasonable to conclude that on average, Star Wars movies rate higher than Star Trek movies, and that men rate this genre of movie higher as well.

(e)

Come up with a potential explanation for the trend that you described in (d) and write a couple of sentences describing the what data you could use to test you explanation further.

Well, I suppose the variables I chose to study have brought up the difficult subject of gender differences, so I'll try to broach this as peacefully as possible. Regardless of the causes behind this gender desparity, whether it is gender differences in a person's innate nature, sociatal upbringing, or the demographics which the movies were marketed to, it would be helpful to explore more star-related series. For instance, BattleSTAR Galactica would be an interesting variable to add. In addition, it would also be interesting to see the other abominations that Star Wars and Star Trek may come up with in the future, and analyze how these fit in with the current trends.