Homework 6

Grace Okamoto

10/5/2021

Problem 1

In your own words, describe what is measured by each of the following and give an example of a dataset that it could reasonably be applied to analyze:

(A) Pearson Correlation

The Pearson Correlation is the measurement of linear correlation between two sets of data. One of the examples I saw was the relationship between income in one particular year vs income in the other year.

(B) Spearman's Rank Correlation

Spearman's Rank Correlation is another measurement of correlation, except that it measures the correlation between the ranks of variables in the data rather than the values.

(C) Contingency Table

A contingency table is a table that displays the association between two variables.

Problem 2

Pick and read one of the articles presented on: https://fivethirtyeight.com/tag/hollywood-taxonomy/ (each article clusters the movies starring a specific person). Provide brief responses to the following:

I chose the Four Types of Idris Elba Movies, because we share the same birthday and also because he was one of the only actors who I had seen in more than one movie. Forgive me for being uncultured, I'm working on it.

(A) Do these seem like reasonable choices to you?

Well I can't intuitively group the movies myself as I've only seen a few, but from the graph it seems like a reasoanble distinction to make. It is important to take into consideration the genre of movie as well, as there are some highly grossing movies on the chart that would appear to be outliers or ungroupable otherwise.

(B) What other metrics or dimensions might you use to compare movies starring the same individual?

I know a popular metric I see when I'm looking at the highest grossing movies of all time is the gross adjusted for inflation. For instance, Avatar is the highest grossing movie of all time, but Gone With the Wind is the highest grossing movie adjusted for inflation, as it was released 70 years earlier. It would be interesting to compare gross adjusted for inflation with the year the movies were released.

(C) Is it clear from the text how the author selected the number of clusters? If so, do you agree with their choice? If not, does their choice seem supported by the scatterplot? Justify your responses.

The graph itself is quite clear to understand, and the reasoning behind the clustering also seems apparent. I'm not a movie expert at all so I don't think it would be wise to disagree with the author. However, one interesting thing to note is that The Dark Tower, which hadn't yet come out at the time the article was written, would be a considerable outlier for the "Ensemble Fantasy and Sci-Fi" cluster, since it had a considerably dismal Rotten Tomatoes score, and an equally dismal box office gross. Perhaps it would be better clustered with "Crap" but it interesting to see how some movies are considered to be in one cluster even though they may fit in more than one cluster.

(D) What positive aspects do the visualizations have? Is there anything that could be improved about the visualizations?

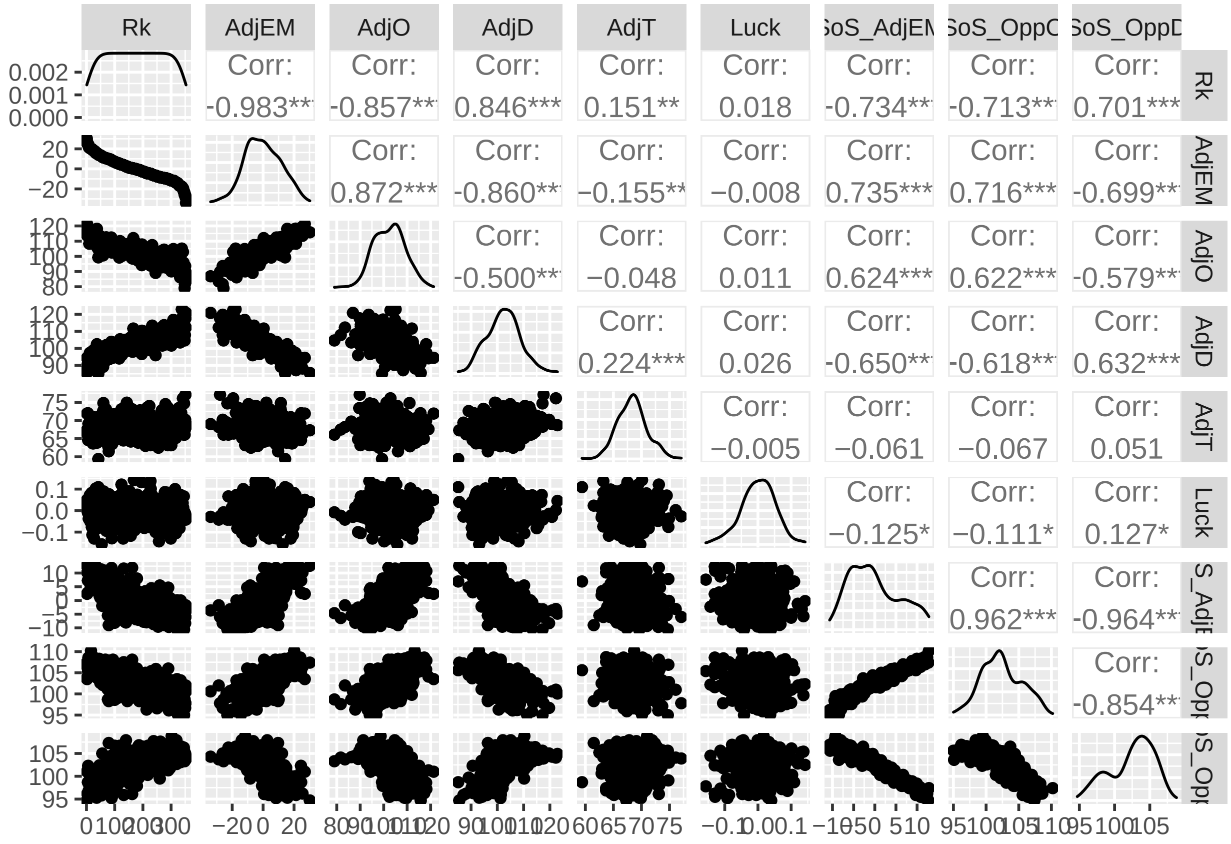
I appreciate that the visualization is relatively clear and easy to read, in addition to being visually appealing. One aspect I do wish was included was more movie names, as a simple dot on the graph doesn't seem to quite give the details I would like.

Problem 3

The basketball data we looked at in class has been loaded in as a dataframe called BB2020 in the first cell.

(A) Construct a correlation table for columns c(1,5,seq(6,18,2)) of the full dataset.

```
ggpairs(BB2020, columns = c(1,5,seq(6,18,2)))
```



(B) Which of these columns are most strongly correlated?

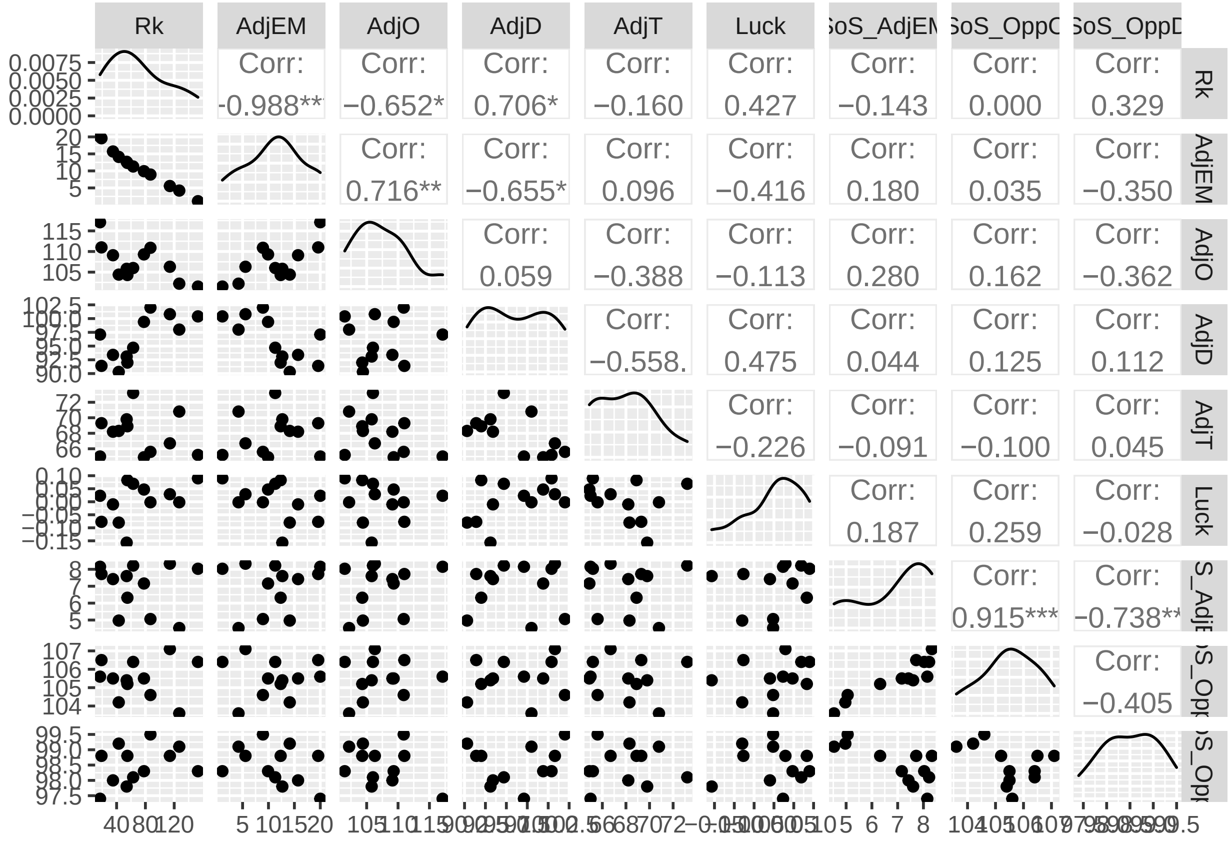
The columns Rank and AdjEM are very strongly correlated. Other columns with strong correlation values are SoS_OppO and S_AdjEM and SoS_OppD and S_AdjE.

(C) Which of these columns are least strongly correlated?

Some of the columns with extremely weak correlation are AdjT and AdjO, any of the luck columns, SoS_AdjEM and AdjT, SoS_OppO and AdjT, and SoS_OppD and AdjT.

(D) Construct a new correlation table for the same columns but just the rows corresponding to teams in the PAC12.

```
BB2020<- BB2020[BB2020$Conf=="P12",]
ggpairs(BB2020, columns = c(1,5,seq(6,18,2)))
```



(E) Which of these columns are most strongly correlated in the new table?

I have no idea if I did this right but it looks as if the Rank and AdjEM columns again have the strongest correlation, along with SoS_OppO and S_AdjEM.

(F) Which of these columns are least strongly correlated in the new table?

AdjD and AdjO, SoS_AdjEM and AdjD, SoS_AdjEM and AdjT, and SoS_OppD and Luck all displayed weak correlation, with SoS_OppO and Rank having the clearest lack of correlation.

(G) What differences do you notice between the two correlation matrices?

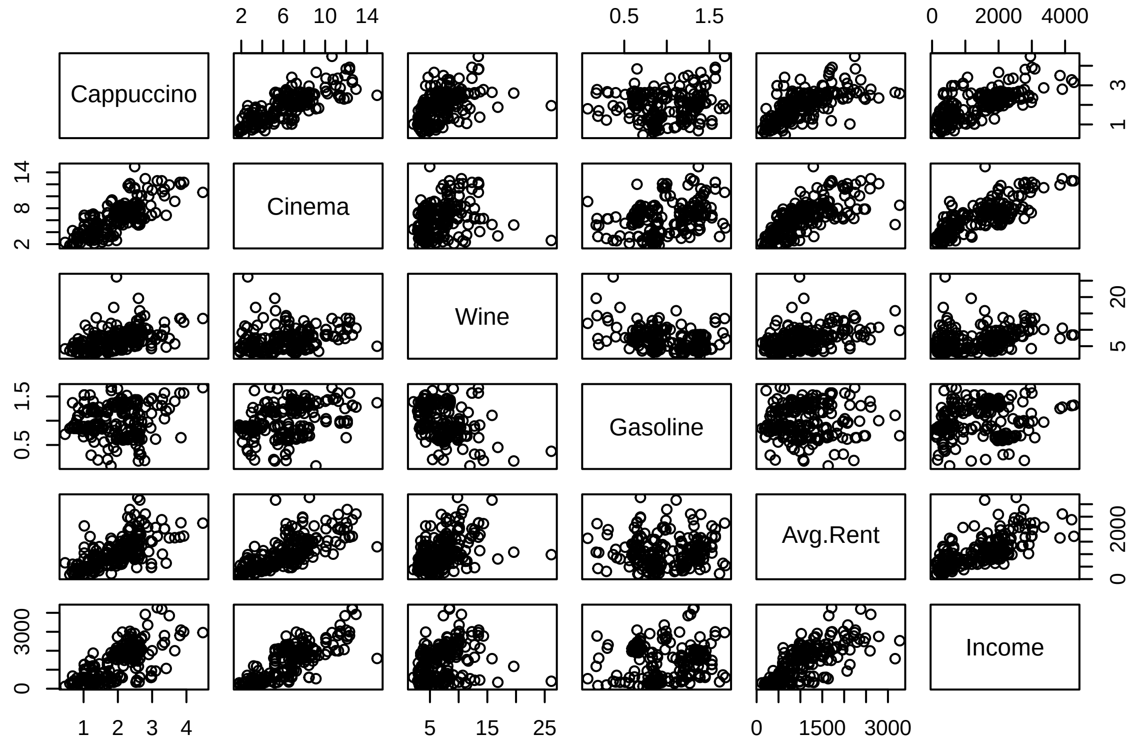
Well for one, the second one looks messed up. That's probably a me issue though. There's a much greater correlation between luck and all the other variables, which might say less of the abilities of those in the PAC-12 conference. To me, the scatterplots are much more difficult to decipher, as there doesn't seem to be enough information anymore to determine any patterns.

Problem 4

The cost of living data we looked at in class has been loaded in as a dataframe called COL in the first cell.

(A) Make a scatter plot matrix of the numeric columns

```
pairs(COL [,c(2:7)])
```



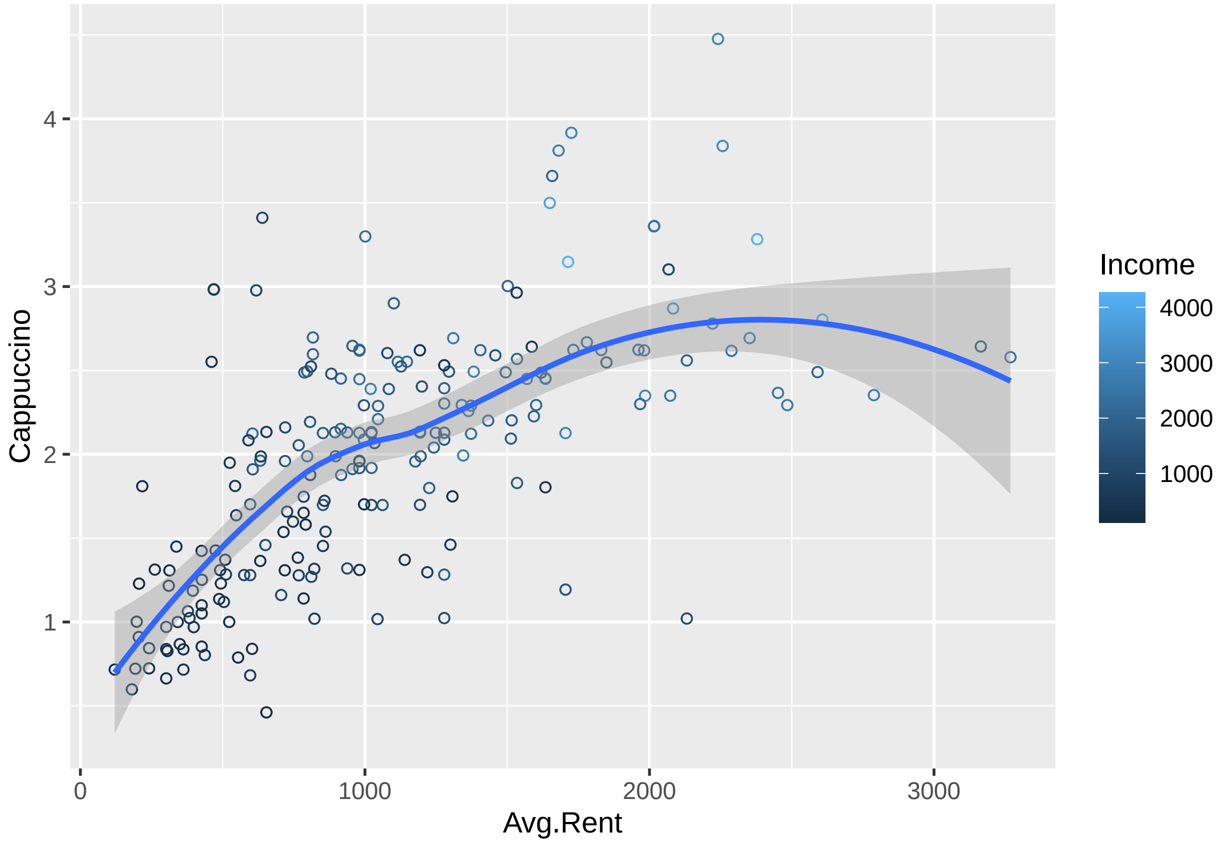
(B) Write a brief (no more than three sentences) summary of what you observe in the plot in (A)

Some of the variables have much stronger correlations to each other than other variables. The transparency of the circles is both disturbing and confusing as it makes the scatterplots rather difficult to read.

(C) Choose a single subplot that seems most interesting to you and make a separate scatterplot of just those two columns with the points colored by the income value.

```
ggplot(data = COL, mapping = aes(x=Avg.Rent, y=Cappuccino,color=Income)) + geom_jitter(shape=1) +
  geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



(D) Write a brief (no more than two sentences) summary of what you observed in the plot in (C)

Cappuccinos and rent are quite loosely correlated, although the correlation tapers off as rent grows higher. In addition, the lower the income, the lower the average rent and cappuccino prices, although there are some notable outliers to this graph.