Homework 6

Rhea Toves

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 measure of actual linear dependence between variables. This could be applied to a dataset of ages.

(B) Spearman’s Rank Correlation:

- The measure of linear dependence between rank order between variables. This could be applied to datasets that can rank variables.

(C) Contingency Table:

- Explores the relationship between variables. This could be applied to any dataset with two or more variables.

_

Problem 2 - Pick and read one of the articles presented on: https://fivethirtyeight.com/tag/hollywood-taxonomy/ (https://fivethirtyeight.com/tag/hollywood-taxonomy/) (each article clusters the movies starring a specific person). Provide brief responses to the following:

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

- I chose to read the article, “The Three Types of Dwayne ‘The Rock’ Johnson Movies” and the three types the author chose to break down were (1) Dwayne Johnson, (2) Dwayne ‘The Rock’ Johnson, and (3) The Rock. I would say these are reasonable types to describe his movies. The author breaks down each point; the category “The Rock” are movies that have “both review scores and box-office returns below or equal to The Rock’s median”. The second category, “Dwayne ‘The Rock’ Johnson” consists of “films with either below-median reviews or below median box office” and the category, “Dwayne Johnson” are “movies that had both above-median reviews and above-median box office”.

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

- The author focused on the box office gross and rotten tomatoes score for this analysis. The author could have studied the the ratings for each individual movie, any other well known actors starring in the movies with Dwayne, or could have collected opinions/ratings from the general public.

(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.

- It is very clear how the author selected the number of clusters and I agree with their choice. I say this because the author goes into depth about their reasoning for each of the three categories. The movies are displayed on the scatterplot graph based on their box office gross and rotten tomatoes score; the author chose to cluster the movies based on its distance from the median reviews.

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

- The initial scatterplot at the beginning of the article is the base for everything the author discusses in their analysis; therefore this plot is very helpful and clear for the audience to follow along. Towards the end of the article, the author includes a bar graph that displays the different roles Dwayne Johnson has played in his movies. I wish the author did more with this information, such as, analyze it, connect the data to the overall theme of the article, etc.

_

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.

```
##                Rk        AdjEM        AdjO        AdjD         AdjT
## Rk     1.00000000 -0.983360883 -0.85703180  0.84625347  0.151094672
## AdjEM -0.98336088  1.000000000  0.87208622 -0.85989328 -0.154511640
## AdjO  -0.85703180  0.872086222  1.00000000 -0.50011911 -0.048045515
## AdjD   0.84625347 -0.859893283 -0.50011911  1.00000000  0.223876595
## AdjT   0.15109467 -0.154511640 -0.04804552  0.22387659  1.000000000
## Luck   0.01820671 -0.007768704  0.01082173  0.02551344 -0.004538131
##               Luck   SoS_AdjEM    SoS_OppO    SoS_OppD
## Rk     0.018206709 -0.73443481 -0.71304665  0.70105832
## AdjEM -0.007768704  0.73465402  0.71622544 -0.69876628
## AdjO   0.010821732  0.62364623  0.62230500 -0.57930509
## AdjD   0.025513435 -0.64953934 -0.61828092  0.63232545
## AdjT  -0.004538131 -0.06147465 -0.06685047  0.05138593
## Luck   1.000000000 -0.12504380 -0.11148159  0.12747400
```

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

- The columns, “Rk”, “AdjEM”, “AdjO”, “AdjD”, “AdjT”, “SoS_AdjEM”, “SoS_OppO”, and “SoS_OppD” are strongly correlated.

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

- The columns, “AdjT” and “Luck” are least strongly correlated.

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

```
##                Rk       AdjEM        AdjO        AdjD         AdjT
## Rk     1.00000000 -0.983360883 -0.85703180  0.84625347  0.151094672
## AdjEM -0.98336088  1.000000000  0.87208622 -0.85989328 -0.154511640
## AdjO  -0.85703180  0.872086222  1.00000000 -0.50011911 -0.048045515
## AdjD   0.84625347 -0.859893283 -0.50011911  1.00000000  0.223876595
## AdjT   0.15109467 -0.154511640 -0.04804552  0.22387659  1.000000000
## Luck   0.01820671 -0.007768704  0.01082173  0.02551344 -0.004538131
##               Luck   SoS_AdjEM    SoS_OppO    SoS_OppD
## Rk     0.018206709 -0.73443481 -0.71304665  0.70105832
## AdjEM -0.007768704  0.73465402  0.71622544 -0.69876628
## AdjO   0.010821732  0.62364623  0.62230500 -0.57930509
## AdjD   0.025513435 -0.64953934 -0.61828092  0.63232545
## AdjT  -0.004538131 -0.06147465 -0.06685047  0.05138593
## Luck   1.000000000 -0.12504380 -0.11148159  0.12747400
```

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

- The columns, “Rk”, “AdjEM”, “SoS_AdjEM”, “SoS_OppO”, and “SoS_OppD” are strongly correlated.

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

- The columns, “AdjO”, “AdjD”, “AdjT”, and “Luck” are least strongly correlated.

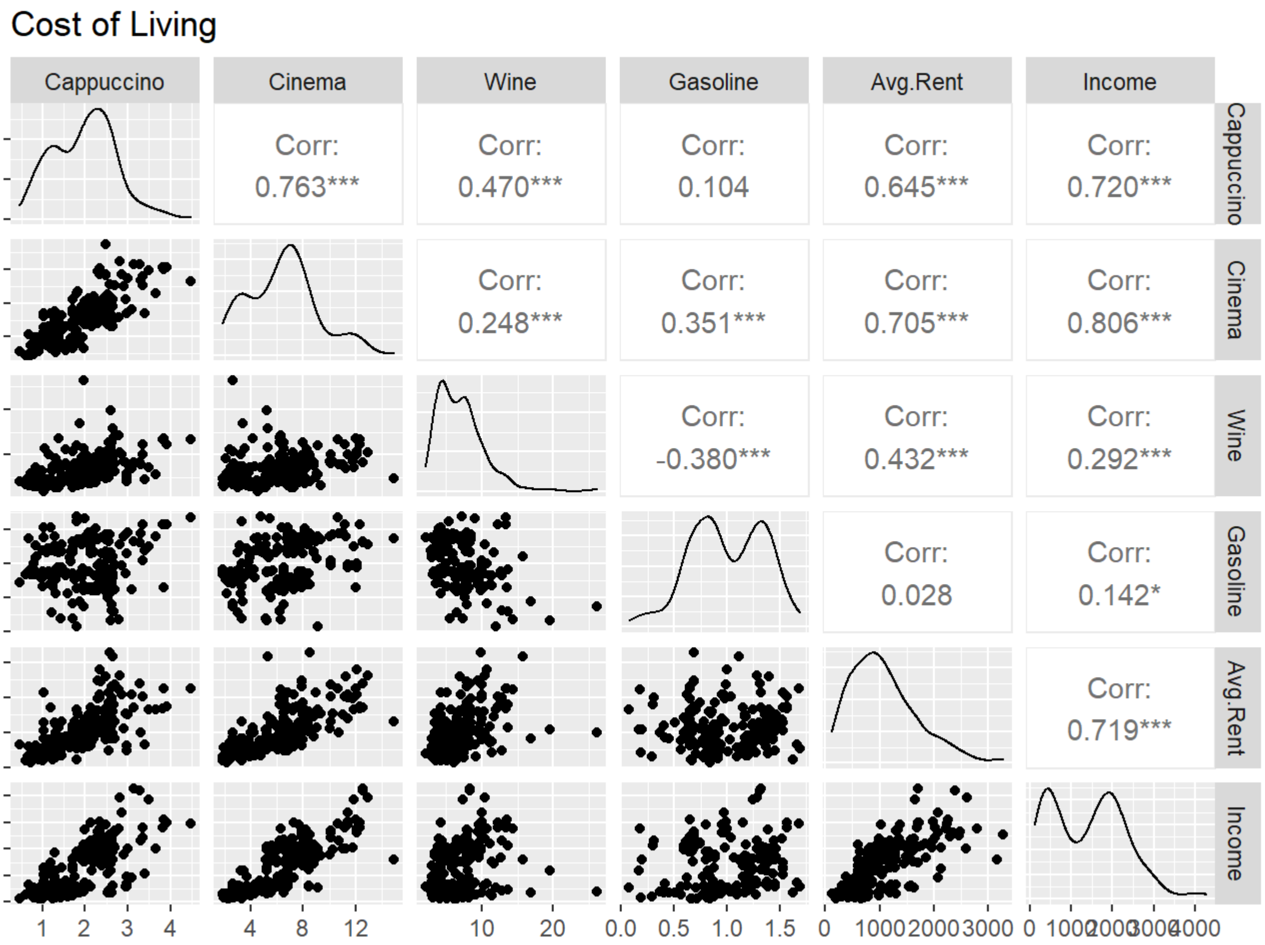
(G) What differences do you notice between the two correlation matrices?

- It is interesting that the strongly correlated columns changed between the two correlation matrices. The common least strongly correlated columns between the two were “AdjT” and “Luck”.

_

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.

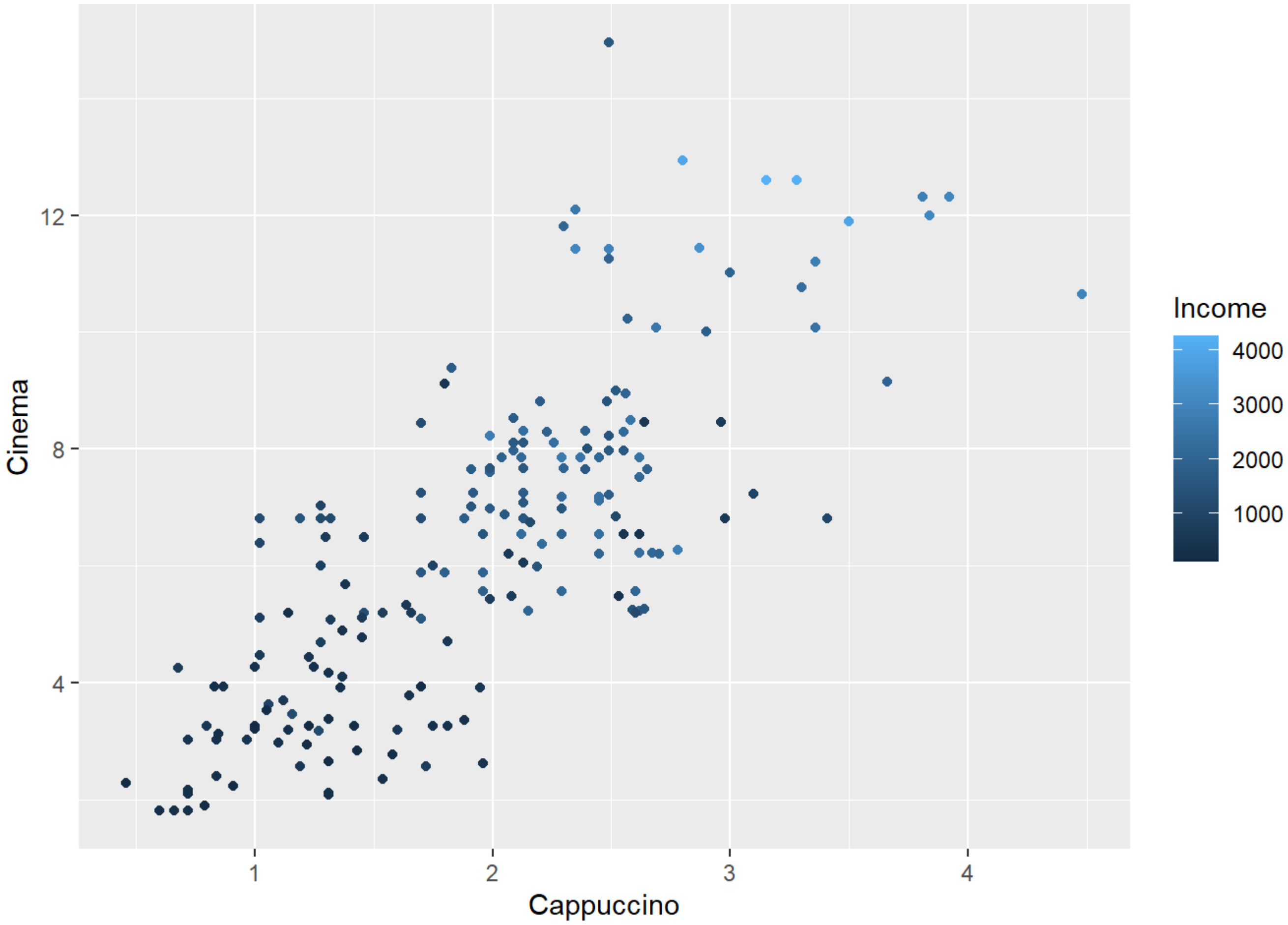
(A) Make a scatter plot matrix of the numeric columns



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

- Plot (A) focuses on the Cappuccino column. The points of the plot create a couple of bell-shaped curves, this displays the movement of each data point.

(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.



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

- I focused on the two columns, “Cinema” and “Cappuccino”. There are darker-colored points towards the bottom of the plot and the points become lighter towards the top of the graph.