Design Challenge 2: Product Review Exploration

Yingzhou "Joe" Liu, yliu945

I. Descriptions of My Visualization

My visualization is essentially a R Shiny product that represents an interactive web app involving tons of interactions between viewers and designers. To see and manipulate my visualization, click on the web app link https://yingzhouliu.shinyapps.io/CS765DesignChallege2/, or open it in a browser.

The primary goal of this app is to show viewers how general attributes of a product like price, average rating, quantity sold and sales interact between each other and to investigate the underlying correlations between them. The secondary goal of this app is to let viewers decide what they want to see based on their preferences for some specific attributes of a product like high sales, or high average rating or both, and to allow views to subset the initial data based on their willingness, to obtain the detailed information and to study the relationships among those attributes with a particular subset of data. From the perspective of either primary goal or second goal of this product, this app emphasizes on each unique product and its characteristics, rather than each reviewer.

This app was built on two main data sets including *CDs and Vinyl 5.csv* and *Music Instruments.csv*. By preprocessing the initial data sets and their corresponding *mega.csv* data sets, attributes like overall rate of a product by each reviewer, product ID, review ID, category, description, and review time were achieved. Since the goals of this app are more relevant to each product, quantity sold was calculated for each product by counting how many reviews each product had. Moreover, average rating was calculated through the sum of all overall ratings of a product divided by quantity sold or the number of reviews of that product. Furthermore, sales was calculated for each product by the price of that product times the quantity sold.

From data processing, price, average rating, quantity sold and sales can range from 0 to 599, from 2.7 to 5, from 0 to 163, and from 0 to 3,725, respectively. Since they are all ratio or quantitative variables, a slider widget was constructed for each variable to select a range of numeric values from the range of that variable. The interval between each value on the slider that can be selected for price, average rating, quantity sold and sales was set as 0.1, 0.1, 1 and 20, respectively. It is of great convenience to have both upper and lower bounds for each variable, since in that way can viewers not only specify the upper and lower bounds, but also fix the range and move the slider of a specific range and change the upper and lower bounds by the same number. This subsetting strategy has been widely applied when people deal with scalability.

From the web app, viewers can also notice three available boxes for text input for category, description and product ID, respectively. This app provides a couple of examples that you can try to specify a category, description or a product ID. However, views do not have to type in exact name of any category, description or product ID. Instead, as long as viewers type in a substring or text of the initial name of any of them, the app will filter out those data that do not contain that substring in that variable and present the rest. This kind of entry of unstructured text values can be related to the searching strategy and subsetting strategy when people think about scalability.

The core of this app is the interactive plot on the right hand side, which converts the filtered data set by previous variables into a visualisation, setting up default mappings between variables in the data set and visual properties. Regardless of interactive property, this visualization is essentially a scatterplot of two variables X and Y which can be selected from the bottom selection boxes. For each one of those two variables, views have four choices including price, average rating, quantity sold and sales which are essentially the basic attributes viewers are interested in when they focus on each product and its properties. Each dot in the scatterplot encodes a product, and both x and y positions encode two of four attributes of a product.

With the respective of interaction in the plot, hovering strategy was applied in this app to show all details of products which are encoded using dots in the plot. Besides, correlation value of selected variables, the number of products and the number of categories shown in the plot are given as values in the right bottom panel.

Moreover, as viewers filter the initial data set using those available variables, the number of dots may decrease significantly. When the number of dots are small, it would be easy to use color encoding to visualize different categories. Therefore, color encoding for category was added when the number of categories is larger than zero but no more than ten. Color encoding is much easier for people to get the rough number of a specific category under some constraints of other variables. Here, color brewer is used with enough contrasts between different colors, since different categories are visualized here and the colors used to encode them should differ significantly between each other.

Last but not least, this app allows views to choose different data sets to realize the visualization.

II. Task and Case Evaluation

A variety of tasks can be solved and realized using this app. The corresponding rationale for why this visualization can address the tasks.

- Viewer could use my visualization to identify which products have few quantity sold, but are highly rated.

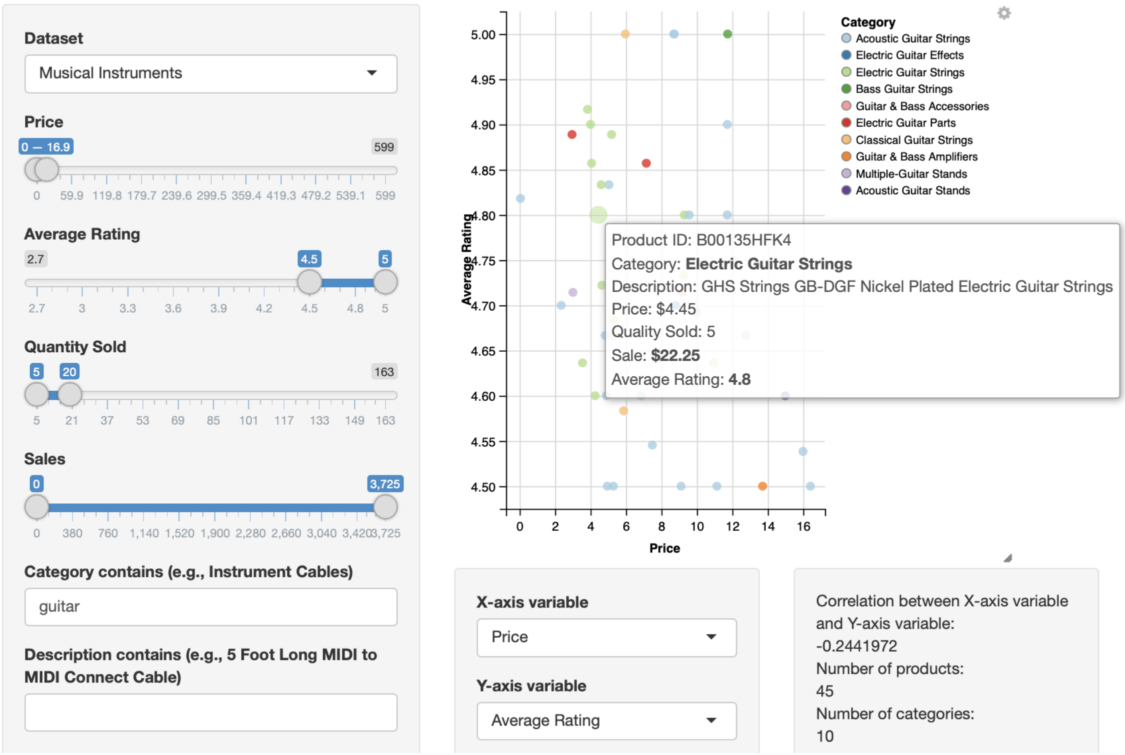
For this task, *Music Instruments* data set has been chosen. Viewers can actually define a range of low values of quantity sold and a range of high values of average ratings by themselves. Besides, viewers can search for any type of category they are interested in.

For instance, as shown in Figure 1, a viewer restricts the values of quantity sold to a range between 5 and 20 and the values of average rating to a range between 4.5 and 5. Besides, the viewer limits categories to those only related to guitar and prices between 0 to 16.9, which are low prices.

From the plot, as shown in Figure 1, the viewer can easily see which guitar-related products are highly rated with few quantity sold and low prices. Moreover, the viewer can use the function of hovering to check all details of those products including specific category, description, product ID, average rating, sales, price and quantity sold.

Furthermore, the view can easily know how many products are guitar-related under those restrictions and how many categories fill in this subset of all products. Correlation is negative and small here,which means a product with a lower price might have a slight higher average rating, which makes sense.

Since the number of total categories in the plot is ten, the view gets a chance to obtain a rough awareness of which categories has a large number of products among all categories in the plot. It is obvious that acoustic guitar strings and electric guitar strings are the most common guitar-related low-price products with few quantity sold but high rating.



Dataset

Musical Instruments

Price

0 — 16.9 599

Average Rating

4.5 5

Quantity Sold

5 20

Sales

0 3,725

Category contains (e.g., Instrument Cables)

guitar

Description contains (e.g., 5 Foot Long MIDI to MIDI Connect Cable)

Product ID: B00135HFK4
Category: **Electric Guitar Strings**
Description: GHS Strings GB-DGF Nickel Plated Electric Guitar Strings
Price: $4.45
Quality Sold: 5
Sale: **$22.25**
Average Rating: **4.8**

X-axis variable

Price

Y-axis variable

Average Rating

Correlation between X-axis variable and Y-axis variable:
-0.2441972
Number of products:
45
Number of categories:
10

Figure 1: Products with few quantity sold but high rating

- Given a category and price band, is there a relationship between number of reviews and average rating?

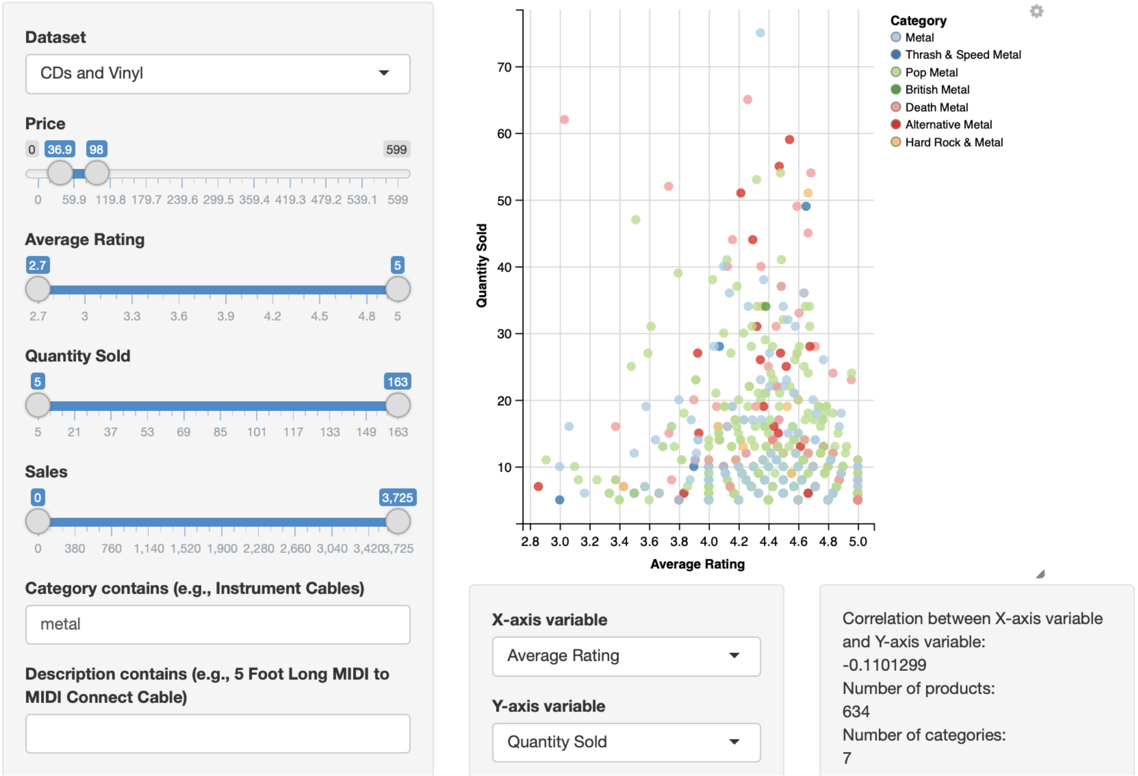
For this task, *CDs and Vinyl 5* data set has been chosen. Viewers can search for any type of category they are interested in, and select a price band as well.

For instance, as shown in Figure 2, a viewer restricts the values of price to a range between 36.9 and 98. Besides, the viewer limits categories to those only related to medal.

Average rating and quantity sold have been selected as X and Y variable respectively, since they are both relevant variables to the task. From the plot, as shown in Figure 2, the viewer can easily see that most dots representing products concentrate when average rating is high while quantity sold, here equal to the number of reviews, is low. Moreover, several products with low rating and low number of reviews exist, but the total number of those products is much lower than the number of products with high rating but low number of reviews. Similarly, several products with high rating and high number of reviews exist, but the total number of those products is much lower than the number of products with high rating but low number of reviews. Besides, the number of products with high number of reviews but low rating is the least in all scenarios, which makes sense.

Furthermore, the view can easily know how many products are metal-related under those restrictions and how many categories fill in this subset of all products. Correlation is negative and small here,which means a product with a lower quantity sold or number of reviews might have a slight higher average rating, which enhances the conclusion that given a category and price band, products with higher rating generally tend to have fewer reviews.

Since the number of total categories in the plot is ten, the view gets a chance to obtain a rough awareness of which categories has a large number of products among all categories in the plot. However, different than previous example, color encoding for category is not necessary here, since the focus of our question is unrelated to the variety of categories.



Dataset

CDs and Vinyl

Price

0 36.9 98 599

Average Rating

2.7 5

Quantity Sold

5 163

Sales

0 3,725

Category contains (e.g., Instrument Cables)

metal

Description contains (e.g., 5 Foot Long MIDI to MIDI Connect Cable)

X-axis variable

Average Rating

Y-axis variable

Quantity Sold

Correlation between X-axis variable and Y-axis variable:
-0.1101299
Number of products:
634
Number of categories:
7

Figure 2: Given a category and price band, products with higher rating tend to have fewer reviews

- Do cheaper products have lower ratings?

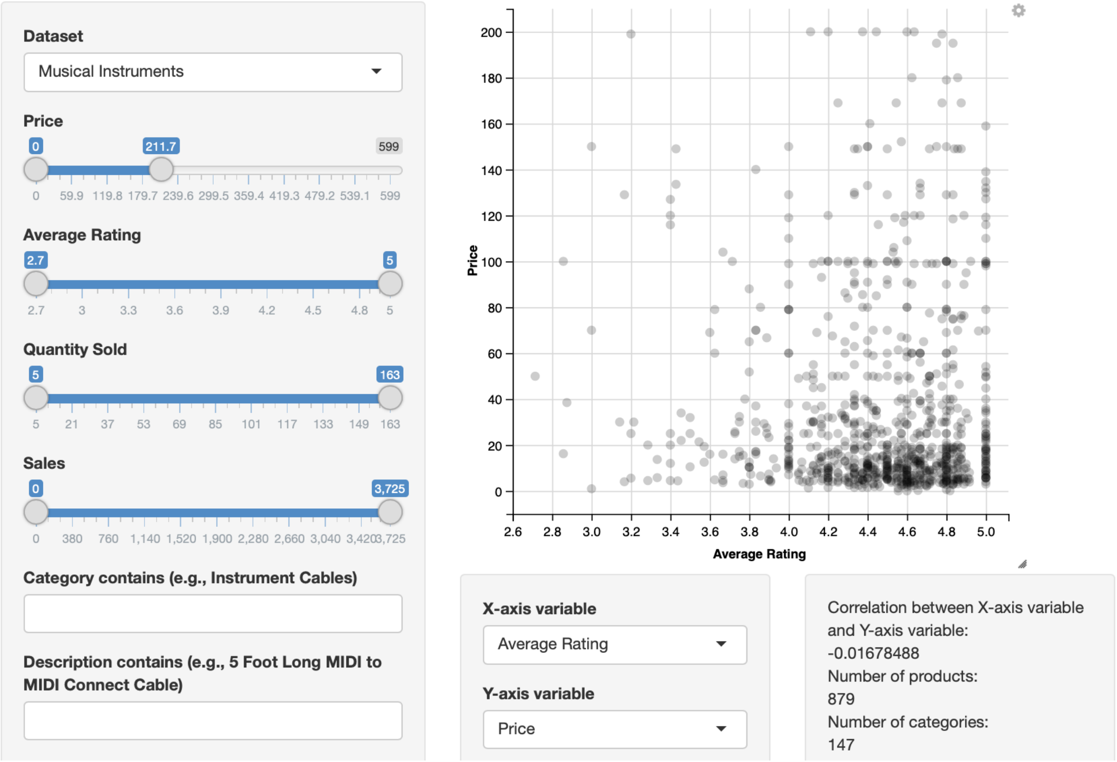
For this task, *Music Instruments* data set has been chosen, since it has fewer observations.

Since this task emphasizes on correlation between rating and price, we do not have to apply many filters to see the correlation.

For instance, as shown in Figure 3, a viewer only restricts the values of price to a general range between 0 and 200.

Average rating and price have been selected as X and Y variable respectively, since they are both relevant variables to the task. From the plot, as shown in Figure 3, the viewer can easily see that most dots representing products concentrate when average rating is high, regardless of whether the price is high or low. Although when price gets higher, the number of products gets lower as the number of dots decrease, but it is still high than the number of dots when average rating is low. Therefore, either cheaper or more expensive products tend to higher ratings.

Furthermore, the correlation of average rating and price is negative and extremely small here, which means a product with a lower price might have a slight higher average rating but it is extremely trivial, which enhances the conclusion that no significant correlation between price and average rating.



Dataset

Musical Instruments

Price

0 – 211.7

Average Rating

2.7 – 5

Quantity Sold

5 – 163

Sales

0 – 3,725

Category contains (e.g., Instrument Cables)

Description contains (e.g., 5 Foot Long MIDI to MIDI Connect Cable)

X-axis variable

Average Rating

Y-axis variable

Price

Correlation between X-axis variable and Y-axis variable:
-0.01678488
Number of products:
879
Number of categories:
147

Figure 3: Either cheaper or more expensive products tend to higher ratings

- Are cheaper products more popular?

For this task, *Music Instruments* data set has been chosen, since it has fewer observations.

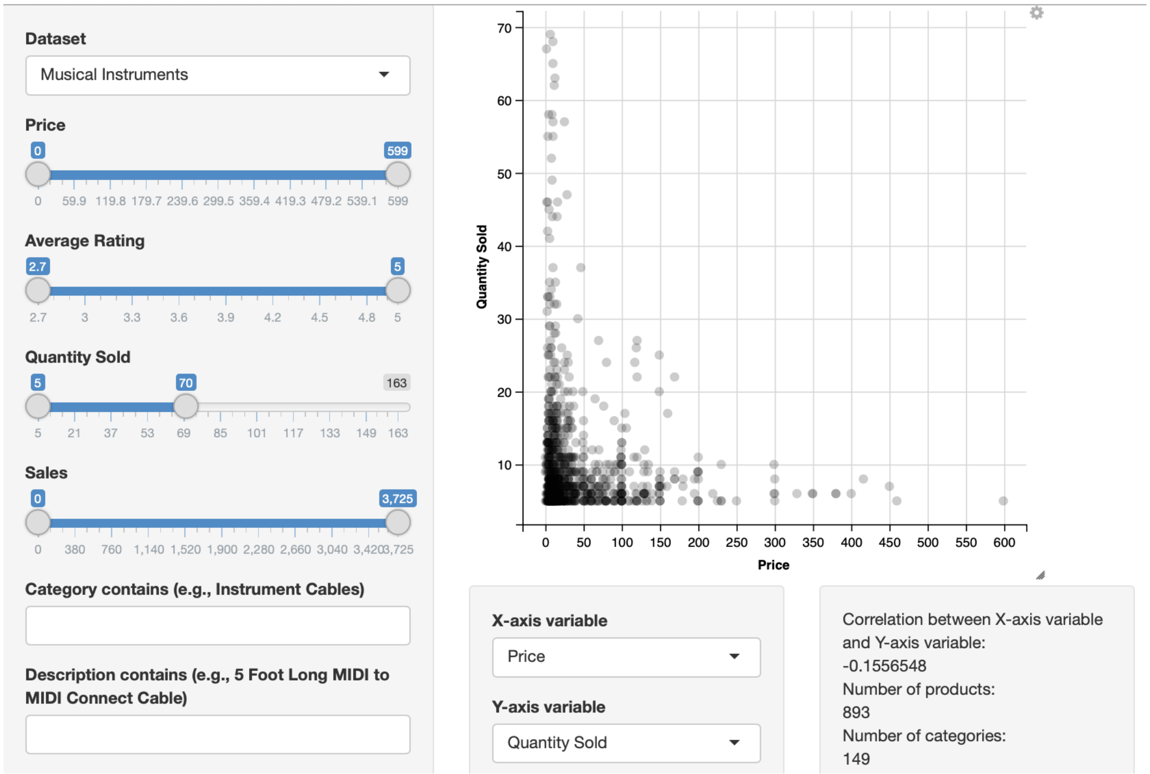
Since this task emphasizes on correlation between popularity and price, we do not have to apply many filters to see the correlation.

Popularity can be represented by the number of reviews which is shown as quantity sold in this app.

For instance, as shown in Figure 4, a viewer only restricts the values of quantity sold to a general range between 0 and 70.

Quantity sold and price have been selected as Y and X variable respectively, since they are both relevant variables to the task. From the plot, as shown in Figure 4, the viewer can easily see that more dots representing products concentrate when quantity is high, when the price is low. Although when price gets higher, the number of products gets lower as the number of dots decrease, but it is obvious that the quantity sold of a product with higher price is lower than a product with lower price. Therefore, cheaper products tend to higher quantity sold.

Furthermore, the correlation of average rating and price is negative and small here, which means a product with a lower price might have a slight higher average rating, which enhances the conclusion that cheaper products slightly tend to higher quantity sold.



Dataset
Musical Instruments

Price
0 – 599

Average Rating
2.7 – 5

Quantity Sold
5 – 70

Sales
0 – 3,725

Category contains (e.g., Instrument Cables)

Description contains (e.g., 5 Foot Long MIDI to MIDI Connect Cable)

X-axis variable
Price

Y-axis variable
Quantity Sold

Correlation between X-axis variable and Y-axis variable:
-0.1556548
Number of products:
893
Number of categories:
149

Figure 4: Cheaper products slightly tend to higher quantity sold

- Given a category and price band, is there a relationship between sale and average rating?

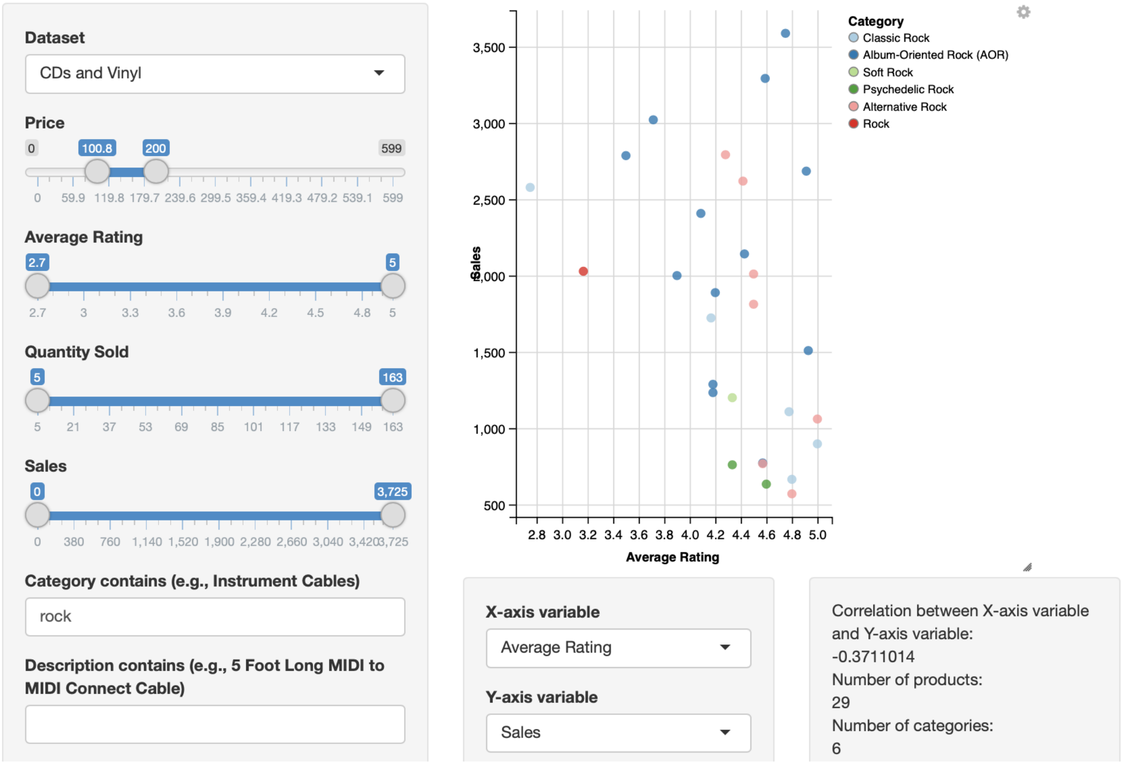
For this task, *CDs and Vinyl 5* data set has been chosen. Viewers can search for any type of category they are interested in, and select a price band as well.

For instance, as shown in Figure 5, a viewer restricts the values of price to a range between 100.8 and 200. Besides, the viewer limits categories to those only related to rock.

Average rating and sale have been selected as X and Y variable respectively, since they are both relevant variables to the task. From the plot, as shown in Figure 5, the viewer can see that when the sale is high, it is hard to tell a product related to rock tend to have high rating or low rating. However, when the sale is low, it is obvious that a product tend to have high rating.

Furthermore, the view can easily know how many products are rock-related under those restrictions and how many categories fill in this subset of all products. Correlation is negative and small here, since when the sale is low, it is obvious that a product tend to have high rating, but it is not when the sale is high.

Since the number of total categories in the plot is ten, the view gets a chance to obtain a rough awareness of which categories has a large number of products among all categories in the plot.



Dataset

CDs and Vinyl

Price

0 100.8 200 599

Average Rating

2.7 5

Quantity Sold

5 163

Sales

0 3,725

Category contains (e.g., Instrument Cables)

rock

Description contains (e.g., 5 Foot Long MIDI to MIDI Connect Cable)

X-axis variable

Average Rating

Y-axis variable

Sales

Correlation between X-axis variable and Y-axis variable:
-0.3711014
Number of products:
29
Number of categories:
6

Figure 5: Given a category and price band, it is obvious that a product tend to have high rating, but it is not when the sale is high.

- Viewer could use my visualization to identify which products make the most money but currently have low quantity sold.

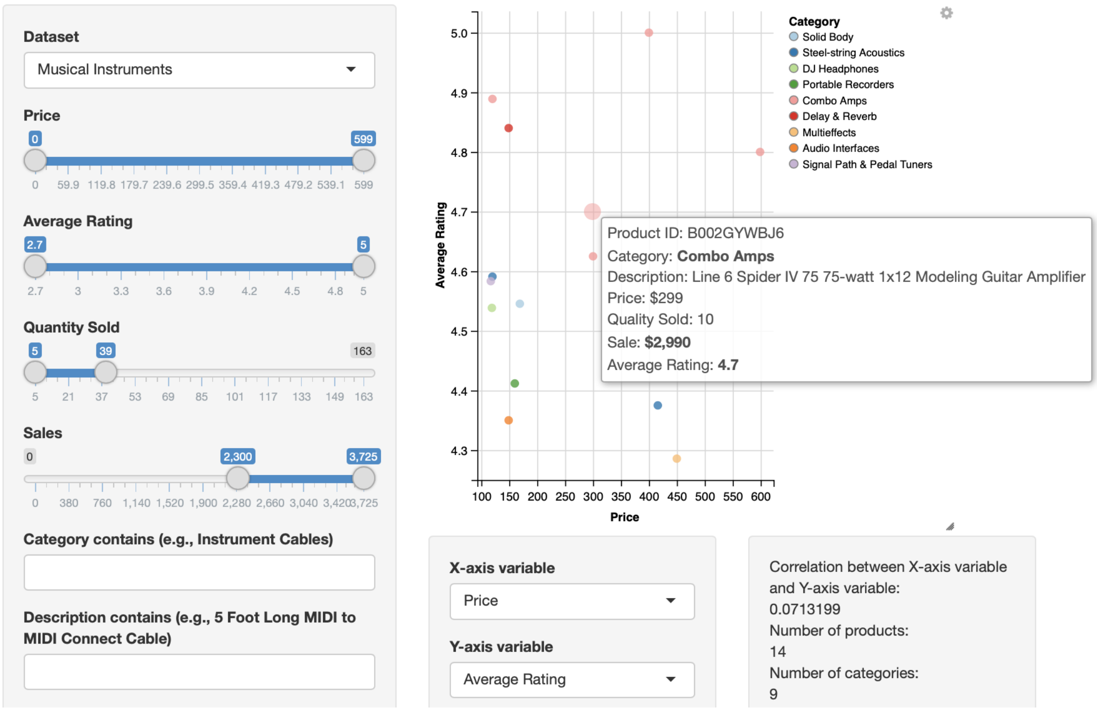
For this task, *Music Instruments* data set has been chosen. Viewers can actually define a range of low values of quantity sold and a range of high values of sale by themselves.

For instance, as shown in Figure 6, a viewer restricts the values of quantity sold to a range between 5 and 39 and the values of sale to a range between 2,300 and 3,725.

From the plot, as shown in Figure 6, the viewer can easily see which products have high sale with few quantity sold. Moreover, the viewer can use the function of hovering to check all details of those products including specific category, description, product ID, average rating, sales, price and quantity sold.

This function will be very helpful for manufacturers since they can focus on those products that are not sold a lot but can make a large amount of money overall.

The ability to hover have a vital importance in this visualization, since viewers need to know the details of those products.



Dataset: Musical Instruments

Price: 0 – 599

Average Rating: 2.7 – 5

Quantity Sold: 5 – 39

Sales: 2,300 – 3,725

Category contains (e.g., Instrument Cables)

Description contains (e.g., 5 Foot Long MIDI to MIDI Connect Cable)

Product ID: B002GYWBJ6
Category: **Combo Amps**
Description: Line 6 Spider IV 75 75-watt 1x12 Modeling Guitar Amplifier
Price: $299
Quality Sold: 10
Sale: **$2,990**
Average Rating: **4.7**

X-axis variable: Price

Y-axis variable: Average Rating

Correlation between X-axis variable and Y-axis variable:
0.0713199
Number of products:
14
Number of categories:
9

Figure 6: Identify which of their products make the most money

III. Scalability

Scalability is the capacity to change in size or scale to manage complexity of visualizations. There are several basic strategies we are capable of using to fix the issue in regard of more data points than pixels in visualization, including scanning serially, selecting subsets and summarizing.

This app basically uses scanning serially, selecting subsets and summarizing, thus it excels at dealing with scalability.

With respect to scanning serially, it uses searching, since three available boxes are designed for text input for category,description and product ID, respectively.

With respect to selecting subsets, it uses filtering based on price, average rating, quantity sold and sales by changing each slider widget. Besides, searching boxes for text input for category,description and product ID will eventually filter the data set as well.

With respect to summarizing, it uses correlation value of selected variables, the number of products and the number of categories shown in the plot are given as values to summarize a large number of points and their underlying relationship.

Overall, this app realizes the function of dealing with a huge data and by interacting with viewers, it can scale the initial data well with good alignment with the decision of a viewer.

IV. Programs

R Shiny[1] is an R package that makes it easy to build interactive web apps straight from R.

[1] https://shiny.rstudio.com

If a viewer wants to use this product, he does not have to run the source code. Instead, click on the web app link https://yingzhouliu.shinyapps.io/CS765DesignChallege2/, or open it in a browser.

If someone wants to modify or use the source code and run the initial program, please first download the folder named R Shiny Product inside the folder named Joe Liu DC2 that is shared to Aditya Barve.

This app is based on R, thus please make sure R is installed. Check how to install R Studio here (https://rstudio.com/products/rstudio/download/).

The following packages need to be installed by running the following in R:

```
install.packages(c('shiny','ggvis','dplyr','RSQLite','RColorBrewer'))
```

After all these packages are installed, run this app by entering the directory of the folder named R Shiny Product i, and then running the following in R:

```
shiny::runApp()
```

V. Interaction

This app is all about interaction with viewers.

For those continuous variables, interaction is applied since we have those sliders to select a specific range of values in that attribute to filter the entire data, which means we have the data that actually fills in that range of values of that attribute. For those categorical variable, interaction is utilized as well since we have those search boxes to search certain ID, category or description.

Besides, this app shows a scatterplot of two variable selected for two of a list including price, average rating, quantity sold and sales, aiming to present relationship or correlation between those two variables. Also, hover has been also applied to show the detailed values of a specific point.

Interaction involved here solves the challenge of filtering data based on multiple constraints of multiple variables and presenting pairwise relationship between multiple variables. If only static visualization is allowed, designers may need tons of them to show all the relationships on all those possible filtered data based on possible constraints of available variables. By incorporating interaction, designers never have to make decisions but users have to. Users can choose what they want to see from a powerful and comprehensive visualization. Interaction helps dealing with the scalability as well.

VI. Data sets

The data sets chosen to use have been shown in each image of each example. This app allows users to switch between two different data sets but it still needs improvement to read in a new data set.

VII. Self-Assessment

I have some experiences about designing a simple R Shiny, but not a complicated one. Therefore, I spent some time on going through tutorials, and a lot of time on debugging. Besides, I also spend a large amount of time figuring out how to design effectively and efficiently.