Laptop Data Analysis

2024-04-22

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.4.4 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(openintro)

## Loading required package: airports  
## Loading required package: cherryblossom  
## Loading required package: usdata

library(ppcor)

## Loading required package: MASS  
##   
## Attaching package: 'MASS'  
##   
## The following objects are masked from 'package:openintro':  
##   
## housing, mammals  
##   
## The following object is masked from 'package:dplyr':  
##   
## select

library(lm.beta)  
library(DataExplorer)  
library(corrgram)

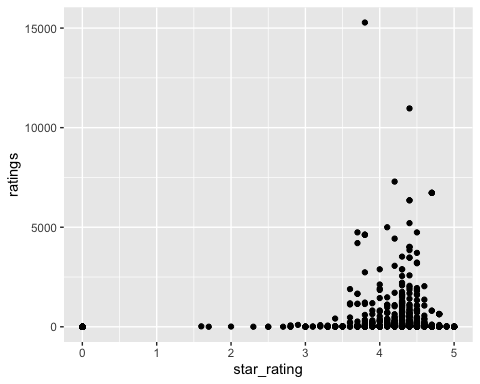
library(dplyr)

laptop <-read.csv("Cleaned\_Laptop\_data.csv")

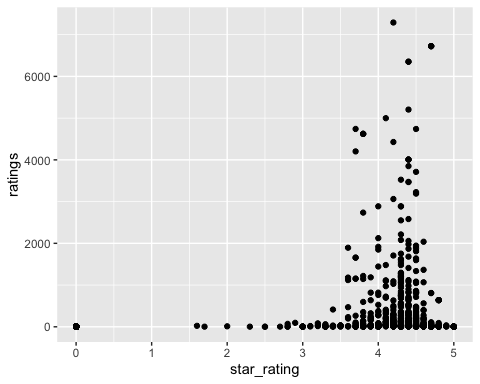
##Introduction My project is on Laptop data that was taken in March of 2022. It contains a ton of variables whether it be the Brand of laptop, specific parts in the laptop including: processor, ram, sdd, and display size. The processor is what runs a laptops operating system and applications. The ram is a form of electronic computer memory. The ssd is a solid-state drive providing persistent data storage. The display size is the size of the laptop screen in inches. A couple other variables I make use of are star\_rating, the number of stars people rated the laptop 0-5, ratings, the number of ratings the laptop got, and old\_price and latest\_price, which are the prices of the laptop when first on the market to now(which is march 2022), and the units for the prices are twp decimal places to the left so 16990 is actually $169.90. In this project I am mainly trying to find relationships between the computer parts and the star ratings that those laptops received.

#Diving into the data The first variables that I noticed and found to be the most interesting were the star\_rating and ratings and I wanted to see what variables affect these and to see if they are related in anyway.

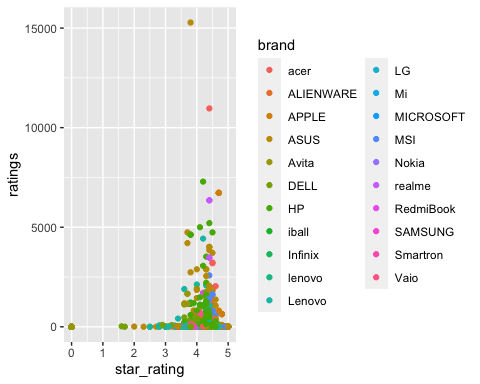
ggplot(data = laptop, mapping= aes(star\_rating, ratings))+  
 geom\_point()

 Seeing how there are two outliers I wanted to create a plot that was easier to see because the points seem to be bunched up there at the bottom. But just by looking at this it seems that the more reviews a laptop gets the better the star rating is on average. I also may want to look at those two individual outliers later to see what they are and why they got so many reviews.

laptop %>%   
 filter(ratings < 7500) %>%   
 ggplot(mapping= aes(star\_rating, ratings))+  
 geom\_point()

 It looks like this plot just further supports my observations in the first it just looks a little bit better without those outliers. So now I want to look at if the brand has anything to do with the star\_ratings how they’re affected by it.

ggplot(data = laptop, mapping= aes(star\_rating, ratings))+  
 geom\_point(aes(color = brand))

 Although we can gather a ton of information from this plot, it seems like there are so many brands and I want to clean this up for only the biggest competitors, so lets see which brands have the most laptops in this dataset and work from there.

laptop %>%   
 group\_by(brand) %>%   
 summarise(n= n()) %>%   
 arrange(desc(n))

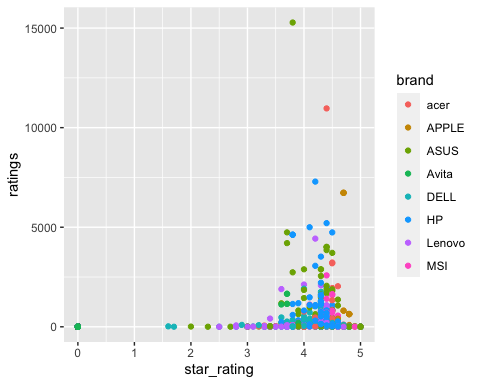
## # A tibble: 21 × 2  
## brand n  
## <chr> <int>  
## 1 ASUS 254  
## 2 DELL 154  
## 3 Lenovo 148  
## 4 HP 142  
## 5 acer 58  
## 6 MSI 52  
## 7 APPLE 28  
## 8 Avita 18  
## 9 LG 5  
## 10 Vaio 5  
## # ℹ 11 more rows

As we can see from this there are 13 Laptop brands that only have 5 entries or less meaning if I take the top 8 Brands my analysis would be more significant, so I’m going to create a new dataframe with just these top 8 brands and call it heavy hitters.

heavy\_hitters <- laptop %>%   
 filter(brand %in% c("ASUS","DELL","Lenovo","HP","acer","MSI","APPLE","Avita"))

So this new data frame only takes out 42 observations but it will help clean up further analysis with those outliers gone from the equation. So now we can redo our last plot to see if its easier to look at.

ggplot(data = heavy\_hitters, mapping= aes(star\_rating, ratings))+  
 geom\_point(aes(color = brand))

 As we can see this plot is way less clustered and we can that some brands more consistently be higher ratings than others. Every HP, acer, APPLE, and MSI laptops don’t have any star-ratings under 3 whereas Lenovo, Avita, ASUS, and DELL have laptops that trickle under 3. By looking at where the 0 star rating is on the plot we can see that no laptop has a rating of 0 because they have no reviews. We can confirm this using an arrange function and filtering by ratings above 0.

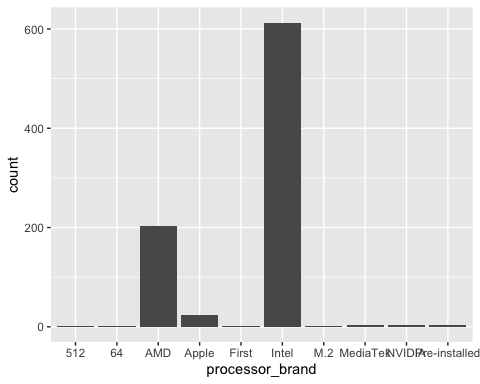
heavy\_hitters %>%   
 dplyr::select(brand, star\_rating, ratings) %>%   
 filter(ratings > 0) %>%   
 group\_by(star\_rating) %>%   
 arrange(star\_rating)

## # A tibble: 595 × 3  
## # Groups: star\_rating [29]  
## brand star\_rating ratings  
## <chr> <dbl> <int>  
## 1 DELL 1.6 23  
## 2 DELL 1.7 3  
## 3 ASUS 2 12  
## 4 ASUS 2.3 3  
## 5 DELL 2.5 2  
## 6 Lenovo 2.5 2  
## 7 ASUS 2.7 3  
## 8 HP 2.8 4  
## 9 Lenovo 2.8 4  
## 10 Lenovo 2.8 70  
## # ℹ 585 more rows

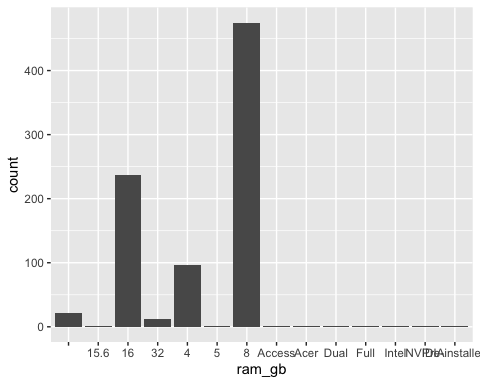
By looking at this we can see that DELL has the two lowest rated laptops in our heavy hitters set but Lenovo has the lowest rated Laptop with 70 reviews meaning it had to be pretty bad to get 70 reviews.

I want to now look at what these heavy hitters have in common in regards to their specifications to see if they’re similar or all over the place. For this I’m going to do a series of bar charts look at the counts of each computer part I want to look at.

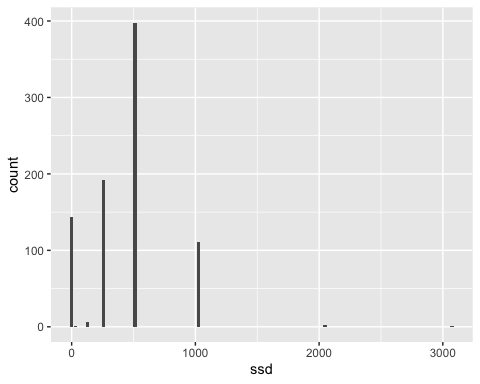
ggplot( data = heavy\_hitters) +  
 geom\_bar( aes(processor\_brand))



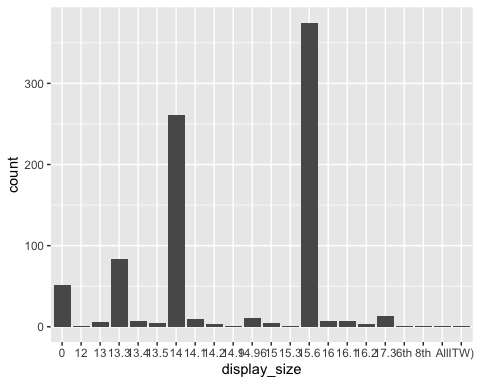
ggplot( data = heavy\_hitters) +  
 geom\_bar( aes(ram\_gb))



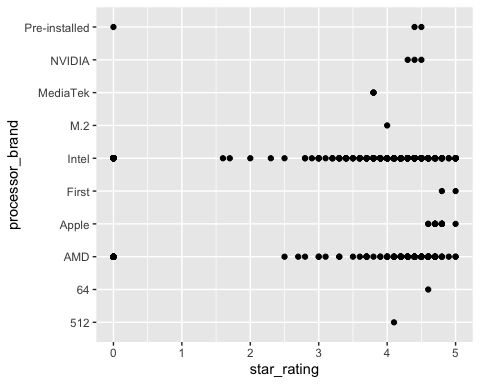
ggplot( data = heavy\_hitters) +  
 geom\_bar( aes(ssd))



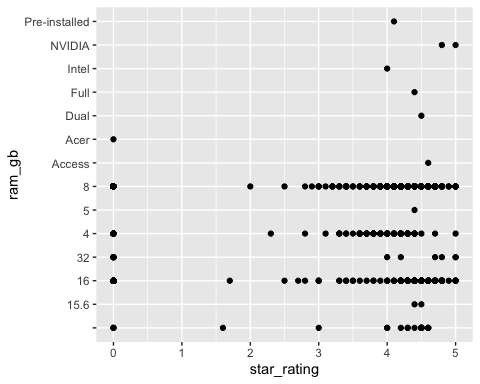
ggplot( data = heavy\_hitters) +  
 geom\_bar( aes(display\_size))

 As we can see from the following bar charts there are a few parts that have a significant lead compared to the rest. It seems that Intel processors, 8 Gigabytes of Ram, 512 ssd, and 15.6 inch screen displays are the most popular among the heavy hitters. I wonder if these variables have any correlation with the star\_rating? Lets find out.

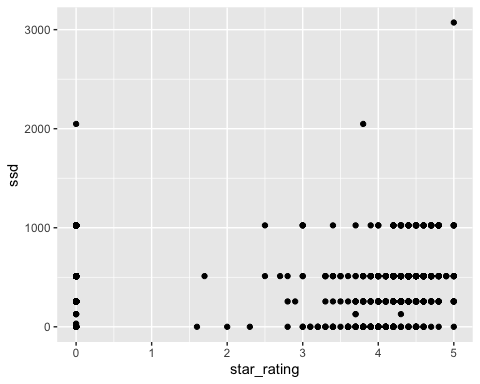
ggplot(heavy\_hitters) +  
 geom\_point(aes(x = star\_rating, y = processor\_brand))



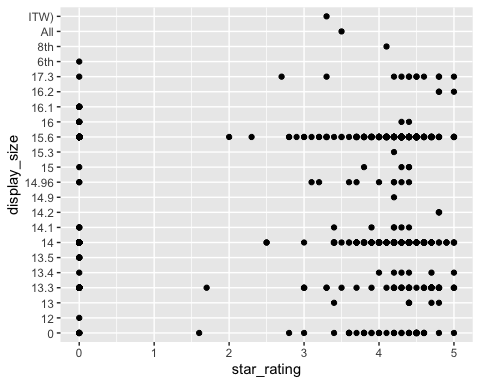
ggplot(heavy\_hitters) +  
 geom\_point(aes(x = star\_rating, y = ram\_gb))



ggplot(heavy\_hitters) +  
 geom\_point(aes(x = star\_rating, y = ssd))



ggplot(heavy\_hitters) +  
 geom\_point(aes(x = star\_rating, y = display\_size))

 After trying to find if these variables were correlated with star\_rating I found that this sort of test wouldn’t work because they are both categorical variables. So now I will compute a chisquare test to see if they are significant for star ratings above 3.5 and the top 2 or 1 when it comes to count in each variable (ex: processor brand = intel and AMD)

(chisq\_data\_proBrand <-heavy\_hitters %>%   
 filter(processor\_brand %in% c("Intel","AMD")) %>%   
 filter(star\_rating >= 3.5) %>%   
 dplyr::select(processor\_brand, star\_rating) %>%   
 table())

## star\_rating  
## processor\_brand 3.5 3.6 3.7 3.8 3.9 4 4.1 4.2 4.3 4.4 4.5 4.6 4.7 4.8 4.9 5  
## AMD 1 1 6 1 2 6 6 15 14 23 22 15 9 3 1 2  
## Intel 4 8 10 21 21 27 25 45 63 63 46 18 15 9 2 13

(chisq\_data\_ram <-heavy\_hitters %>%   
 filter(ram\_gb %in% c(8,16)) %>%   
 filter(star\_rating >= 3.5) %>%   
 dplyr::select(ram\_gb, star\_rating) %>%   
 table())

## star\_rating  
## ram\_gb 3.5 3.6 3.7 3.8 3.9 4 4.1 4.2 4.3 4.4 4.5 4.6 4.7 4.8 4.9 5  
## 16 2 1 1 2 1 3 3 9 11 20 18 13 14 15 1 5  
## 8 2 3 8 14 16 21 23 40 58 57 41 19 18 7 2 8

(chisq\_data\_ssd <-heavy\_hitters %>%   
 filter(ssd %in% c(512)) %>%   
 filter(star\_rating >= 3.5) %>%   
 dplyr::select(ssd, star\_rating) %>%   
 table())

## star\_rating  
## ssd 3.5 3.6 3.7 3.8 3.9 4 4.1 4.2 4.3 4.4 4.5 4.6 4.7 4.8 4.9 5  
## 512 3 1 2 5 8 8 13 22 35 61 41 21 21 8 3 6

(chisq\_data\_disp <-heavy\_hitters %>%   
 filter(display\_size %in% c(15.6,14)) %>%   
 filter(star\_rating >= 3.5) %>%   
 dplyr::select(display\_size, star\_rating) %>%   
 table())

## star\_rating  
## display\_size 3.5 3.6 3.7 3.8 3.9 4 4.1 4.2 4.3 4.4 4.5 4.6 4.7 4.8 4.9 5  
## 14 1 2 6 7 10 12 10 16 17 33 17 13 8 2 3 3  
## 15.6 2 1 7 13 9 17 19 29 50 35 40 15 12 7 0 7

chisq.test(chisq\_data\_proBrand)

## Warning in chisq.test(chisq\_data\_proBrand): Chi-squared approximation may be  
## incorrect

##   
## Pearson's Chi-squared test  
##   
## data: chisq\_data\_proBrand  
## X-squared = 26.674, df = 15, p-value = 0.03151

This shows that processor brand is significant with a p-value of 0.03151 so we reject the null hypothesis, telling us there is a relationship between high star rating and processor brand.

chisq.test(chisq\_data\_ram)

## Warning in chisq.test(chisq\_data\_ram): Chi-squared approximation may be  
## incorrect

##   
## Pearson's Chi-squared test  
##   
## data: chisq\_data\_ram  
## X-squared = 48.321, df = 15, p-value = 2.257e-05

This shows that ram\_gb is very significant with a p-value of 2.257e-05 so we reject the null hypothesis, telling us there is a relationship between high star rating and ram\_GB size.

chisq.test(chisq\_data\_ssd)

##   
## Chi-squared test for given probabilities  
##   
## data: chisq\_data\_ssd  
## X-squared = 265.29, df = 15, p-value < 2.2e-16

This shows that ssd is very significant with a p-value of < 2.2e-16 so we reject the null hypothesis, telling us there is a relationship between high star rating and ssd.

chisq.test(chisq\_data\_disp)

## Warning in chisq.test(chisq\_data\_disp): Chi-squared approximation may be  
## incorrect

##   
## Pearson's Chi-squared test  
##   
## data: chisq\_data\_disp  
## X-squared = 20.028, df = 15, p-value = 0.1709

This shows that display size is not significant with a p-value of 0.1709 so we fail to reject the null hypothesis, telling us there is a no relationship between high star rating and display size.

Although 3 out of the 4 of these are significant the X-squared values aren’t that high with the highest being 265.29 in the ssd test. From here I wanted to try and checkout the t-test of these same groups, with the star\_rating versus the top two processor brands, ram\_gb, ssd’s, and display sizes.

tBrandData <- heavy\_hitters %>%   
 filter(processor\_brand %in% c("Intel","AMD"))  
   
t.test(star\_rating ~ processor\_brand, data = tBrandData)

##   
## Welch Two Sample t-test  
##   
## data: star\_rating by processor\_brand  
## t = -0.2429, df = 333.98, p-value = 0.8082  
## alternative hypothesis: true difference in means between group AMD and group Intel is not equal to 0  
## 95 percent confidence interval:  
## -0.3640291 0.2840075  
## sample estimates:  
## mean in group AMD mean in group Intel   
## 2.843350 2.883361

Based off of this t-test we find that the p-value is not less than 0.05 so we fail to reject the null hypothesis stating that there is not a significant difference between star rating of the processor brands AMD and Intel.

tRamData <- heavy\_hitters %>%   
 filter(ram\_gb %in% c(8,16))  
   
t.test(star\_rating ~ ram\_gb, data = tRamData)

##   
## Welch Two Sample t-test  
##   
## data: star\_rating by ram\_gb  
## t = -4.8252, df = 409.36, p-value = 1.977e-06  
## alternative hypothesis: true difference in means between group 16 and group 8 is not equal to 0  
## 95 percent confidence interval:  
## -1.1307220 -0.4761037  
## sample estimates:  
## mean in group 16 mean in group 8   
## 2.378903 3.182316

Based off of this t-test we find that the p-value is way less than 0.05 meaning we reject the null hypothesis indicating that there is a significant difference between star ratings of ram size of 8GB and 16 GB.

tSsdData <- heavy\_hitters %>%   
 filter(ssd %in% c(256,512))  
   
t.test(star\_rating ~ ssd, data = tSsdData)

##   
## Welch Two Sample t-test  
##   
## data: star\_rating by ssd  
## t = 1.4384, df = 407.59, p-value = 0.1511  
## alternative hypothesis: true difference in means between group 256 and group 512 is not equal to 0  
## 95 percent confidence interval:  
## -0.08936682 0.57679354  
## sample estimates:  
## mean in group 256 mean in group 512   
## 3.192708 2.948995

Based off of this t-test we find that the p-value is not less than 0.05 so we fail to reject the null hypothesis stating that there is not a significant difference between star rating of the SSD size of 512GB and 256GB.

tDispData <- heavy\_hitters %>%   
 filter(display\_size %in% c(15.6,14))  
   
t.test(star\_rating ~ display\_size, data = tDispData)

##   
## Welch Two Sample t-test  
##   
## data: star\_rating by display\_size  
## t = -2.4965, df = 529.85, p-value = 0.01285  
## alternative hypothesis: true difference in means between group 14 and group 15.6 is not equal to 0  
## 95 percent confidence interval:  
## -0.71416501 -0.08517292  
## sample estimates:  
## mean in group 14 mean in group 15.6   
## 2.737931 3.137600

Based off of this t-test we find that the p-value is less than 0.05 meaning we reject the null hypothesis indicating that there is a significant difference between star ratings of display size of 15.6in and 14in.

After finding some significance with the t-test I wanted to move on and look to see if there would be some other variables that might be correlated to star\_rating so I went to figure that out with the variables old\_price, latest\_price, and ratings.

cor(heavy\_hitters$latest\_price, heavy\_hitters$star\_rating)

## [1] -0.1736646

cor(heavy\_hitters$old\_price, heavy\_hitters$star\_rating)

## [1] -0.07907447

cor(heavy\_hitters$ratings, heavy\_hitters$star\_rating)

## [1] 0.213709

And surprisingly after looking at the correlation between all of these variables there isn’t any significant correlation between star rating and latest\_price, old\_price, and ratings with the highest correlation being 0.21 between star\_rating and ratings which I thought would be the highest but it is still a pretty weak correlation. While we’re on the topic of pricing I’m curious to see the average latest price and old price of laptops with a star rating 4 and above.

heavy\_hitters %>%   
 filter(star\_rating >= 4) %>%   
 dplyr::select(brand, star\_rating, old\_price) %>%   
 group\_by(brand, star\_rating) %>%   
 summarise(oldPmean = mean(old\_price)) %>%   
 arrange(desc(star\_rating))

## `summarise()` has grouped output by 'brand'. You can override using the  
## `.groups` argument.

## # A tibble: 64 × 3  
## # Groups: brand [7]  
## brand star\_rating oldPmean  
## <chr> <dbl> <dbl>  
## 1 APPLE 5 329900   
## 2 ASUS 5 106993.  
## 3 DELL 5 95168.  
## 4 HP 5 128522.  
## 5 Lenovo 5 54450.  
## 6 ASUS 4.9 122990   
## 7 MSI 4.9 93489   
## 8 acer 4.9 89999   
## 9 APPLE 4.8 189733.  
## 10 ASUS 4.8 126658.  
## # ℹ 54 more rows

heavy\_hitters %>%   
 filter(star\_rating >= 4) %>%   
 dplyr::select(brand, star\_rating, latest\_price) %>%   
 group\_by(brand, star\_rating) %>%   
 summarise(latPmean = mean(latest\_price)) %>%   
 arrange(desc(star\_rating))

## `summarise()` has grouped output by 'brand'. You can override using the  
## `.groups` argument.

## # A tibble: 64 × 3  
## # Groups: brand [7]  
## brand star\_rating latPmean  
## <chr> <dbl> <dbl>  
## 1 APPLE 5 309990   
## 2 ASUS 5 147392.  
## 3 DELL 5 88395   
## 4 HP 5 105727.  
## 5 Lenovo 5 40494.  
## 6 ASUS 4.9 99990   
## 7 MSI 4.9 79990   
## 8 acer 4.9 53990   
## 9 APPLE 4.8 176323.  
## 10 ASUS 4.8 126357.  
## # ℹ 54 more rows

So after looking at both of these it might be weird to look at but these pricing values are pretty deceiving. The laptops are not actually hundreds of thousands of dollars the decimal place is just supposed to be two places to the left. After analyzing the data it seems that typically the higher star rating the laptop has the more expensive it is. Of course this isn’t always the case but it would make sense, the more you pay equals the better quality laptop. Now that got me thinking because I ran the previous correlations on the whole heavy hitters dataframe maybe there wasn’t much correlation but here there seems to be. Lets test it out.

newlatestprice\_stars <- heavy\_hitters %>%   
 filter(star\_rating >= 4) %>%   
 dplyr::select(star\_rating, latest\_price)

cor(newlatestprice\_stars)

## star\_rating latest\_price  
## star\_rating 1.0000000 0.4383429  
## latest\_price 0.4383429 1.0000000

newoldprice\_stars <- heavy\_hitters %>%   
 filter(star\_rating >= 4) %>%   
 dplyr::select(star\_rating, old\_price)

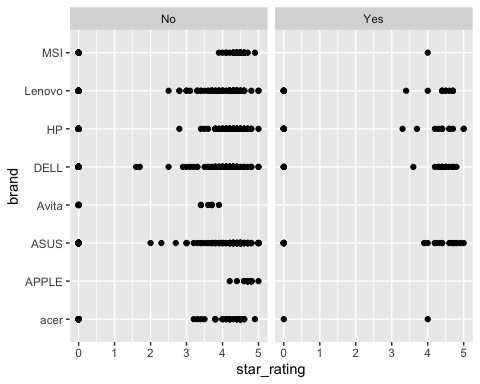
cor(newoldprice\_stars)

## star\_rating old\_price  
## star\_rating 1.0000000 0.3744273  
## old\_price 0.3744273 1.0000000

So after running the correlation on these new dataframes we see the correlation went up a good amount for both latest\_price and old\_price when being compared to star\_ratings that are above or equal to 4. Although the correlation amount isnt anything crazy it is still a good amount to see a difference.

After seeing little correlation between these variables I wanted to look at something else in the dataframe. Theres a boolean variable of touchscreen that I’d like to check out because it seems so random but could be cool. I’m thinking that brands of laptops with a touch screen could have lower ratings overall because they could be gimmicky but lets check it out.

heavy\_hitters %>%  
 ggplot()+  
 geom\_point(mapping = aes(star\_rating,brand)) +  
 facet\_wrap(~Touchscreen)

 So from our plots it looks like laptops with touchscreens have over all good ratings (again the ratings with 0 have no reviews so they are default to 0). Another interesting thing here is that Apple and Avita have no laptops with touchscreens, opposed to every other brand which has at least tried once.

##Conclusion Through this analysis I found that it was pretty difficult to find relationships between any of the variables in this data set and the star\_rating the laptops received. I really wanted to see if there was a specific reason why these laptops were getting higher star ratings than others but with the lack of lower star ratings it made it a bit more difficult than I had originally thought. A lot of my predictions were false when I looked at the analysis but that is all apart of analyzing data and its more fun when you get proved wrong anyway. If I were to do more analysis on this data I would go see if the prices and the different specifications of the laptops have any correlation finding out which parts are the most expensive compared to others.

##Websites that helped <https://www.quora.com/What-are-the-hardware-components-of-a-laptop>