Hussain Sarfraz

Dr. Kates

DATA-101-610-2021C

09/21/2021

Module 3: Exploratory Data Analysis Homework

This set of homework problems will let you test the knowledge you gained by working through Module 3 on Exploratory Data Analysis, which provides additional practice working with ggplot and dplyr.

1. Above, we analyzed the effect of carat, cut, and color on a diamond’s price. But what about the 4th c, clarity?

**Part A**: How does clarity matter to a diamond’s price?

**Part B**: How is clarity related to the other Cs?

**Part C**: What would you conclude based on your analysis?

If you need more information about clarity, you can consult <https://www.americangemsociety.org/page/clarityscale>.

**NOTE**: 0 is the highest diamond grade while 10 is the lowest diamond grade. (this information was taken from AGS Diamond Grading System site <https://www.americangemsociety.org/buying-diamonds-with-confidence/ags-diamond-grading-system/>)

Brief Explanation

Clarity is related to a diamonds color and cut. If a diamond has a high scale of any of these variables, then a diamonds price would increase.

A diamonds weight (carat) is not related to clarity, color, and cut. This means that if a diamond has a low weight that does not mean the price of the diamond would be low. The scale of the diamond’s clarity, color, and cut would increase the price even with low weight. On the other hand, even if a diamond has a low grade of clarity, color, and cut that does not guarantee a low price for the diamond since the diamonds weight could determine a high price.

I reached this conclusion based off my analysis of the graphs I made which displayed the diamond data. I have included details in the detailed explanation of this section.

**Part A:** How does clarity matter to a diamond’s price?

Detailed Explanation

To answer this question, I followed these steps:

* 1. Create a bar chart using the clarity variable to see which clarity diamond types does our dataset hold the most.
  2. Create a boxplot that compares the clarity and price variables to see which clarity type has a higher price than the other diamonds. Having a general observation about this will help me with my analysis on step 3.
  3. Use the carat and price scatter plot and add a color aesthetic which represents the various clarity types in our dataset.
  4. Compared all the carat and price scatter plots, each having a different aesthetic of color, cut, and clarity
  5. Created a geom\_count graph that compares clarity with color and cut

STEP 1

I created a bar chart that would count the number of diamond clarity types that our dataset holds. I did this to get a better understanding of the dataset and create observations that could potentially help with my analysis.

This is what I typed in R (Input): This was my output:

Text

Description automatically generated Chart, bar chart, histogram

Description automatically generated

Based off my output I can see that the diamonds dataset has a lot of diamonds that have a clarity ranking of 5 (slightly included/SI1). The second highest number of diamonds are those with a clarity ranking of 4 (very slightly included/VS2). There are a low number of diamonds with a clarity ranking of 8 (included/I1) and 0 (flawless/IF).

It would be interesting to see the price of diamonds with a clarity ranking of 5 and 4 (the diamonds that appears the most) , and 8 and 0 (the diamonds that appeared the least).

Do the diamonds that appeared the most have a higher price than the diamonds that appeared the least? Or is it the other way around? I will find my answer to this observation in step 2.

STEP 2

Following my observations from step 1, I have created a boxplot that compares diamond clarity and price. I did this to investigate further and see which diamond clarity types are sold at a higher price. Knowing this would also help me answer my questions made in step 1.

This is what I typed in R (Input): This was my output:

Text

Description automatically generated Chart

Description automatically generated

It seems as if my hypothesis from step 1 was not correct because after looking at the medians (the bolded lines) in the boxplot I notice that Diamonds with a clarity type of 6 (Silightly Included/SI2) seems to be the most expensive. Diamonds with a clarity type of 8 (included/I1) and 5 (slightly included/SI1) are priced a bit lower than the clarity type 6 diamond price.

The lowest selling diamonds are the ones with a clarity type of 0 (flawless/IF) and 1 (very very slightly included/VVS1).

I would like to point out that all of the boxplots have many outliers, meaning that although the medians of some of the diamonds seem to be low (such as clarity types 0 and 1), the outliers cause the price for these diamonds to be high . The high number of outliers in all the boxplots suggest that there might be another factor that is causing the the outliers/high price to form. This means that there is probably a high chance that clarity does not have a significant influence in determining the price of a diamond.

Also, since 0 is the highest possible grade it would make sense that the price for diamonds with a clarity type 0 should have a high price. However, this was not the case in the boxplot. This means that I need to see the relationship that clarity has with other elements (such as carat, cut, and color). I am going to start off my examining the relationship between carat,price, clarity.

**Part B:** How is clarity related to the other Cs?

STEP 3

I will now be investigating the relationship between carat, price, and clarity. I am going to be doing this through using a carat and price scatter plot and adding a color aesthetic that represents the various clarity types

This is what I typed in R (Input): This was my output:

Text

Description automatically generated Chart, scatter chart

Description automatically generated

The scatterplot shows that diamonds with a clarity type 0 and 1 have a high price but there are a few points that represent these clarity types (I am saying this because I see many more blue, green, and yellow points that make up the graph). This is probably why clarity types 0 and 1 had a low median in the boxplot in step 2. Also, my observations from step 1 were true because the barchart showed that diamonds with clarity types 5 and 4 have the highest count. This is shown in the scatter plot since there are a lot of dark and light green points.

I also noticed that diamonds with a highest possible clarity grade (zero) weigh less than diamonds with a low clarity grade (5-8). The diamonds with a clarity grade of 5,6, and 8 have similar prices as diamonds with a clarity grade of zero. As mentioned before, I expected the diamonds with a low clarity grade to have a low price. I was surprised when I saw these results in the scatterplot. I believe that there are other factors (cut and color) that cause the price for these diamonds to be high despite having a low clarity grade.

Also, I do belive that because the weight of the diamonds (with a low clarity grade) were higher, the price increased. I am saying this because I have noticed that as the weight of the diamond decreases, the price of low clarity grade diamonds decreases. Diamonds that are between 1-2 carats verify this observation because the pink and blue points are higher than the other colors.

STEP 4

I then compared each graph of carat and price with a different color aesthetic of clarity, cut, and color. I did this so I can observe some general patterns between the different components of the diamond grading system.

Carat-Price plot w/ Clarity Types: Carat-Price plot w/ Diamond Color: Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

Carat-Price plot w/ Diamond Cut:

Chart, scatter chart

Description automatically generated

**Part C:** What would you conclude based on your analysis?

What I noticed by comparing all the scatterplots is that diamonds that weigh lass have a high price because they have a high grade of cut, clarity, and color which makes the diamonds price high despite the low weight (carat). On the other hand, diamonds that have a high weight have low grades of cut, clarity, and color but the high weight brings the diamonds price up to the level of the low weight diamonds with the high grades.

I came to this conclusion by looking at the scale in all three scatter plots. For example, in the Carat-Price scatterplot with Clarity types as an aesthetic the highest grade of clarity (zero) is represented by pink. Pink is shown at the beginning of the graph where the diamond weight is small.

In the Carat-Price scatterplot with color as an aesthetic, D is the highest grade of color and it is at the beginning of the graph where the diamond weight is approximately 1 carat. As the carat weight increases the lowest grade of color, J is shown more frequently.

The same goes with In the Carat-Price scatterplot with cut as an aesthetic. Although, in this graph the pattern is not too strong as it is in the previous two graphs. This is because there are only a few fair diamonds that weigh a lot. Majority of the fair diamonds weigh less which is why majority of the fair diamonds have a low price. But many ideal and premium diamonds are listed with a high price even with a weight of 1 carat (approximately).

The code I used to add the Diamond Cut and color aesthetic in the Carat-Price scatterplots above:

This is what I typed in R (Input): This was my output:

Text

Description automatically generatedChart, scatter chart

Description automatically generated

This is what I typed in R (Input): This was my output:

Text

Description automatically generatedChart, scatter chart

Description automatically generated

STEP 5

I wanted to explore the data further by using the geom\_count graph and seeing what type of diamonds are the most in the dataset. I started off my creating a geom\_count graph for color and clarity.

This is what I typed in R (Input): This was my output:

Text

Description automatically generated A picture containing chart

Description automatically generated

This graph shows that the dataset contains many diamonds that have a low grade clarity and high-low color grade. This explains why the scatterplots in step 4 had more colors that other points.

This is what I typed in R (Input): This was my output:

Text

Description automatically generated with medium confidence A picture containing chart

Description automatically generated

This graph shows that the dataset contains many diamonds that have a low grade clarity and high-low cut grade. This explains why the scatterplots in step 4 had more colors that other points.

Both graphs do not tell us much or add on to my analysis but they do give me more insights about the data I am dealing with. What I have learned is that the there are more low clarity grade (6,5,4) diamonds in the dataset. On the other hand, there are a mix of diamonds with various cut and clarity grades. There is no specific diamond cut and clarity grade that that is prominent in the dataset because the big points are equally distributed vertically.

1. Calculate the number of flights for each destination in the NYC flights data. Plot that against the average arrival delay for each destination. What, if anything, would you conclude? Is this surprising? Why or why not?

*Feedback:  Great plot here. You correctly interpreted the relationship--one thing to remember is that as you had more flights per carrier, we saw the average delay level out between 0-10 because as our data increases, variability in our arrival delay decreases. Individual cases of larger delays will have less effect on the overall average when you have lots and lots of data points. 1.5/2*

To solve this problem, I followed these steps:

* 1. Create an object that only includes values of non-cancelled flights
  2. Calculate the number of flights in each destination and average arrival delays and display this through a scatterplot

STEP 1

I created an object called ‘not\_cancelled’ which only contains flights that do not have a blank/NA value in the departure and arrival delay column. I am choosing delays because if the rows delay values are not filled then that means the flight was cancelled.



STEP 2

I then counted the amount of flights in each destination using the count and group\_by functions. I also calculated the average arrival delays through using the mean function. I then made a scatterplot which showed the number of flights in each destination (x-axis) and average arrival delay (y-axis). Additionally, I added a goem\_smooth function to the graph to get a better visualization between the relationship between the variables. The code I used to graph and my output is displayed below:

This is what I typed in R (Input): This was my output:

Text, letter

Description automatically generated Chart, scatter chart

Description automatically generated

Based on my output, I came to the conclusion that as the number of flights increase, for each destination, the average arrival delay decreases and slowley moves towards zero. I reached this conslusion by looking at how the geom\_smooth line gradually goes down. I also saw that there were many high average arival delays when the number of flights was close to zero. As the number of flights reached 2500 and onwards, the high average of arrival delays went down from 30 to around 10.

1. This will ask you to solve some problems with the nycflights dataset.
   1. Plot average speed against distance. What do you conclude?

**NOTE**: I have used the not\_cancelled object since it only contains flights that are not cancelled. I talk about the object not\_cancelled in question 2.

This is what I typed in R (Input): This was my output:

Text

Description automatically generated with low confidence Chart, scatter chart

Description automatically generated

My output shows that as a airplanes average speed increases the airplane gets to travel a greater distance. There are more outliers such as the airplanes which have a speed of 10-12. These planes do not travel a big distance. In fact, the planes travel a short distance.

A possible explaination for this is that airplanes with a high average speed probably did not travel a great distance because maybe the planes engine was too large or small? There could be a possibility that there were some airplanes that had to re-route or take a long way to their destination (this could potentially ecplain why they are some outliers in the upper part of the graph. Weather could also be another factor that prevented planes, with high average speeds, to not go far.

* 1. Create a new variable that records whether or not a flight arrived on time. *Hint: define a flight to be “on time” if it gets to the destination on time or earlier than expected, regardless of any departure delays.* Then, determine the on-time arrival percentage based on whether the flight departed on time/early or not (i.e., whether there was a departure delay). What percent of flights that were “delayed” departing arrive “on time”?

27.7 percent of flights that were delayed for departure arrived on time. I got this answer from the code below:

This is what I typed in R (Input): This was my output:

Text

Description automatically generated Text

Description automatically generated with medium confidence

I categorized the arrival flights that were on time from a 1 to 0 scale (this was done so the proportion of early/on-time arrival flights can be calculated. I also did the same for flights that had a departure delay.

Since the question was asking for flights that were delayed and still arrived on time I had to made sure I was getting the proportion of arrived flights that were also delayed. This is why I added the filter funtion (to only have the data of delayed flights).

After that, I used the summarize function to calculate the proportion/percantage of flights that had a depature delay and still arrived early/on-time.

* 1. Building on the question above, determine (roughly) what the cutoff point is for departure delays where you can still expect to get to your destination on time.

The rough cutoff point for departure delays is 7.47. This is a rough estimate that if you departure delay is 7.47 minutes long then there is still a change that your flight would arrive on time/early. I reached this conclusion from the code I wrote below.

This is what I typed in R (Input): This was my output:

Text

Description automatically generated Text

Description automatically generated with medium confidence

I kept the code similar to the one in part B. The detail that I did change was that I filtered delayed flights and flights that arrived early/on-time. I did this because the question was asking for a rough estimate of a cutoff point for departure delays that could gauruntee a early/on-time arrival. After I did this, I got the mean/average of the departure delays that arrived early/on-time. Calculating this would give me a good estimate of the average departure delay cutoff that would still highly gauruntee a on-time/early arrival.