Getting Started: Diamonds

Erin Dowd 2020-07-11

Purpose: Throughout this course, you'll complete a large number of *exercises* and *challenges*. Exercises are meant to introduce content with easy-to-solve problems, while challenges are meant to make you think more deeply about and apply the content. The challenges will start out highly-scaffolded, and become progressively open-ended.

In this challenge, you will go through the process of exploring, documenting, and sharing an analysis of a dataset. We will use these skills again and again in each challenge.

Grading Rubric

Unlike exercises, **challenges will be graded**. The following rubrics define how you will be graded, both on an individual and team basis.

Individual

Category	Unsatisfactory	Satisfactory
Effort	Some task **q**'s left unattempted	All task **q**'s attempted
Observed	Did not document observations	Documented observations based on analysis
Supported	Some observations not supported by analysis	All observations supported by analysis (table, graph, etc.)

Category	Unsatisfactory	Satisfactory
Code Styled	Violations of the style guide hinder readability	Code sufficiently close to the style guide

Team

Category	Unsatisfactory	Satisfactory
Documented	No team contributions to Wiki	Team contributed to Wiki
Referenced	No team references in Wiki	At least one reference in Wiki to member report(s)
Relevant	References unrelated to assertion, or difficult to find related analysis based on reference text	Reference text clearly points to relevant analysis

Due Date

All the deliverables stated in the rubrics above are due on the day of the class discussion of that exercise. See the Syllabus for more information.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   1.0.0
## v tidyr   1.1.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts ------------------------------------------ tidyverse_conflicts()
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

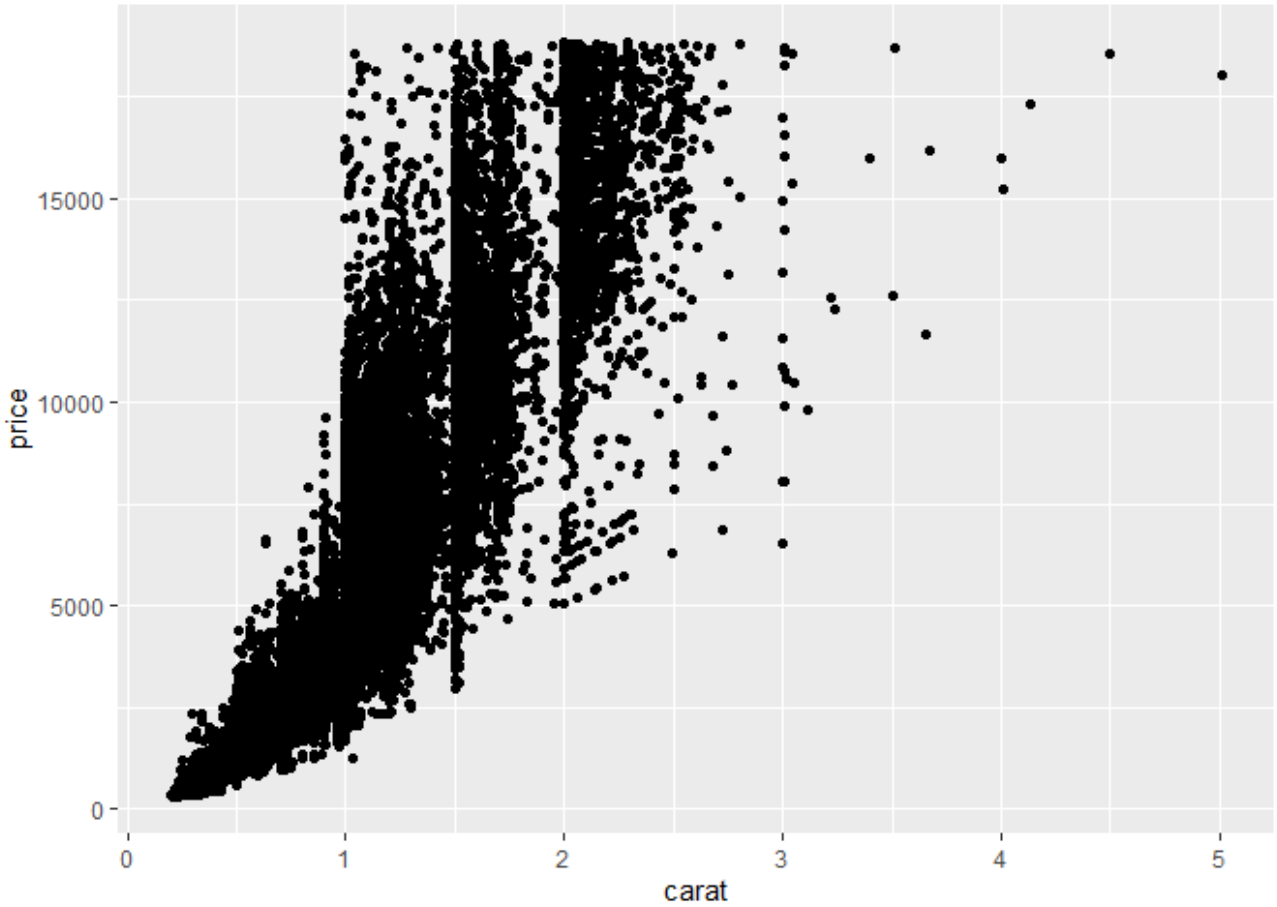
Data Exploration

In this first stage, you will explore the `diamonds` dataset and document your observations.

q1 Create a plot of `price` vs `carat` of the `diamonds` dataset below. Document your observations from the visual.

Hint: We learned how to do this in `e-vis00-basics` !

```
ggplot(data = diamonds)+
      geom_point(mapping = aes(x = carat, y = price))
```



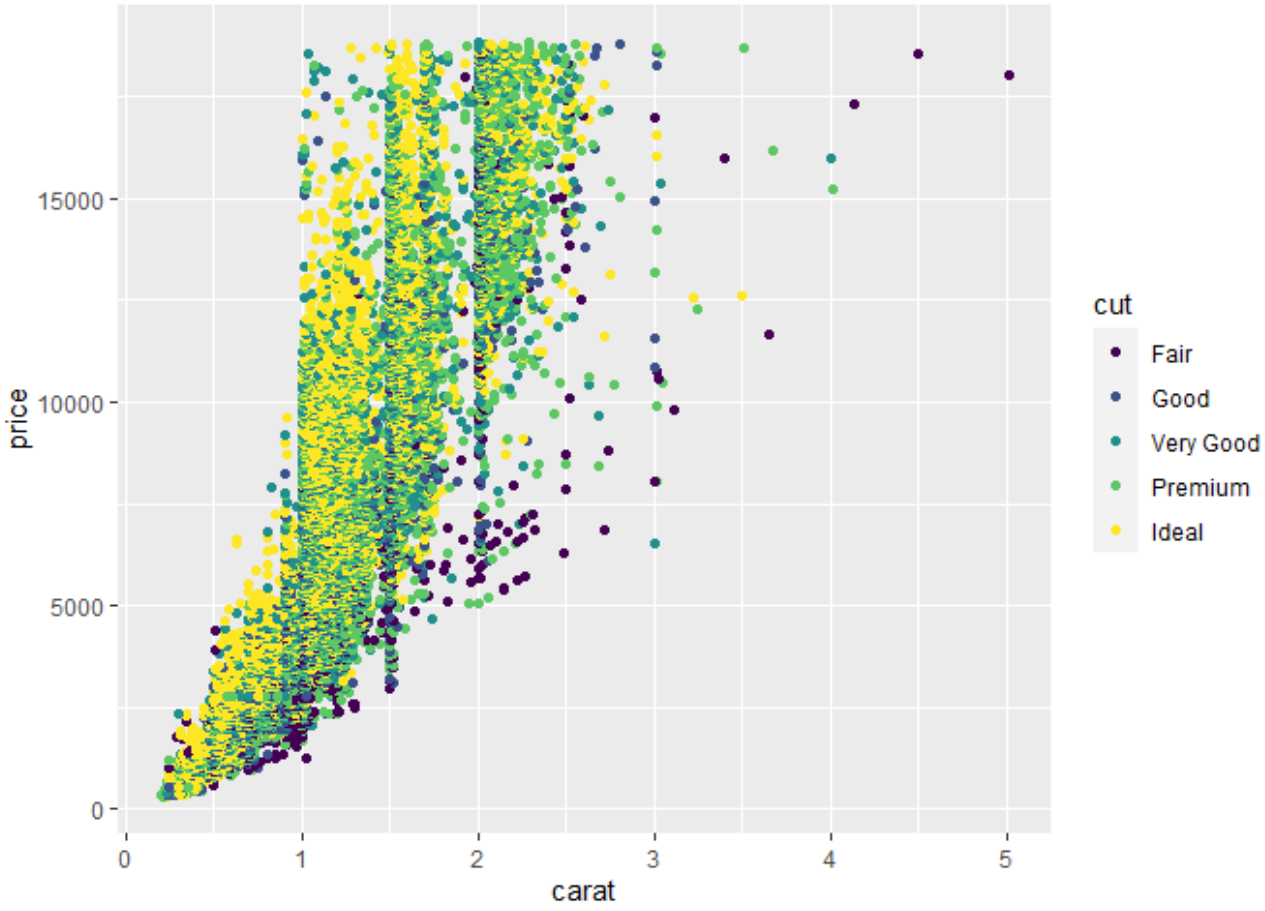
(q1-task-1.png) **Observations**:

- As carat increases, so does price; however, as many diamonds in a given carat - band have a large range in prices, there are likely additional factors that
- impact diamond price. Additionally, while there appears to be an even -distribution of diamonds at all fractions of 1 carat below 1 carat, more -diamonds in the dataset appear to be clustered at whole-number carat values -than at carat values between whole numbers over 1 carat.There is a relatively -small range of prices for diamonds up

to 1 carat, between 1 dollar and 5000 -dollars, a range of approximately 1000 to 5000 dollars for diamonds between 1 -and 1.5 carats, while diamonds between approximately 1.8 carats to 3 carats -have a range in price of approximately 5000 to approximately 19000 dollars. -Above 2.5 carats, the range in price continues to decrease, and all of teh -diamonds over 4 carats have prices of over 15000 dollars. The maximum price of -a diamond in this dataset is 18,823 dollars.

q2 Create a visualization showing variables `carat`, `price`, and `cut` simultaneously. Experiment with which variable you assign to which aesthetic (`x`, `y`, etc.) to find an effective visual.

```
ggplot(data = diamonds)+
      geom_point(mapping = aes(x = carat, y = price, color=cut))
```



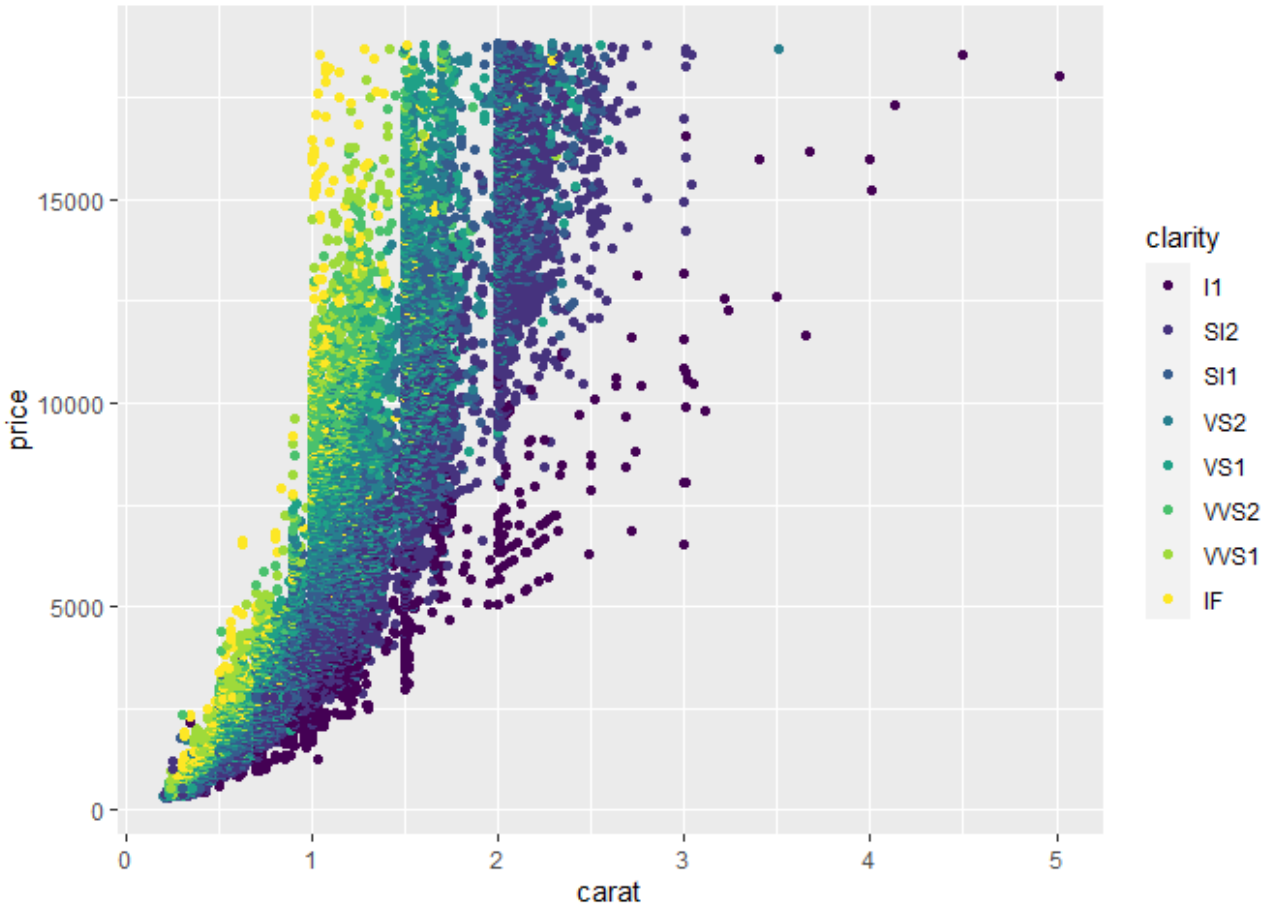
(q2-task-1.png) **Observations**:

From looking at this graph, higher price correlates strongly with higher carat, with some apparent cluster effects at whole numbers.“Ideal” cuts, represented in yellow, appear to be more frequent in smaller-carat diamonds, i.e. carats up -to 2.5, “good” are present throughout carat categories, and “fair” cuts make up a larger and larger proportion of the

diamonds in the sample as carat increases. For larger carat diamonds, the cut does not appear to be the most important factor for price for diamonds over 3 carats, diamonds with the highest prices were more likely to be “fair” than to be “ideal” cuts

Additional Analysis: Clarity

```
ggplot(data = diamonds) +
   geom_point(mapping = aes(x = carat, y = price, color=clarity))
```



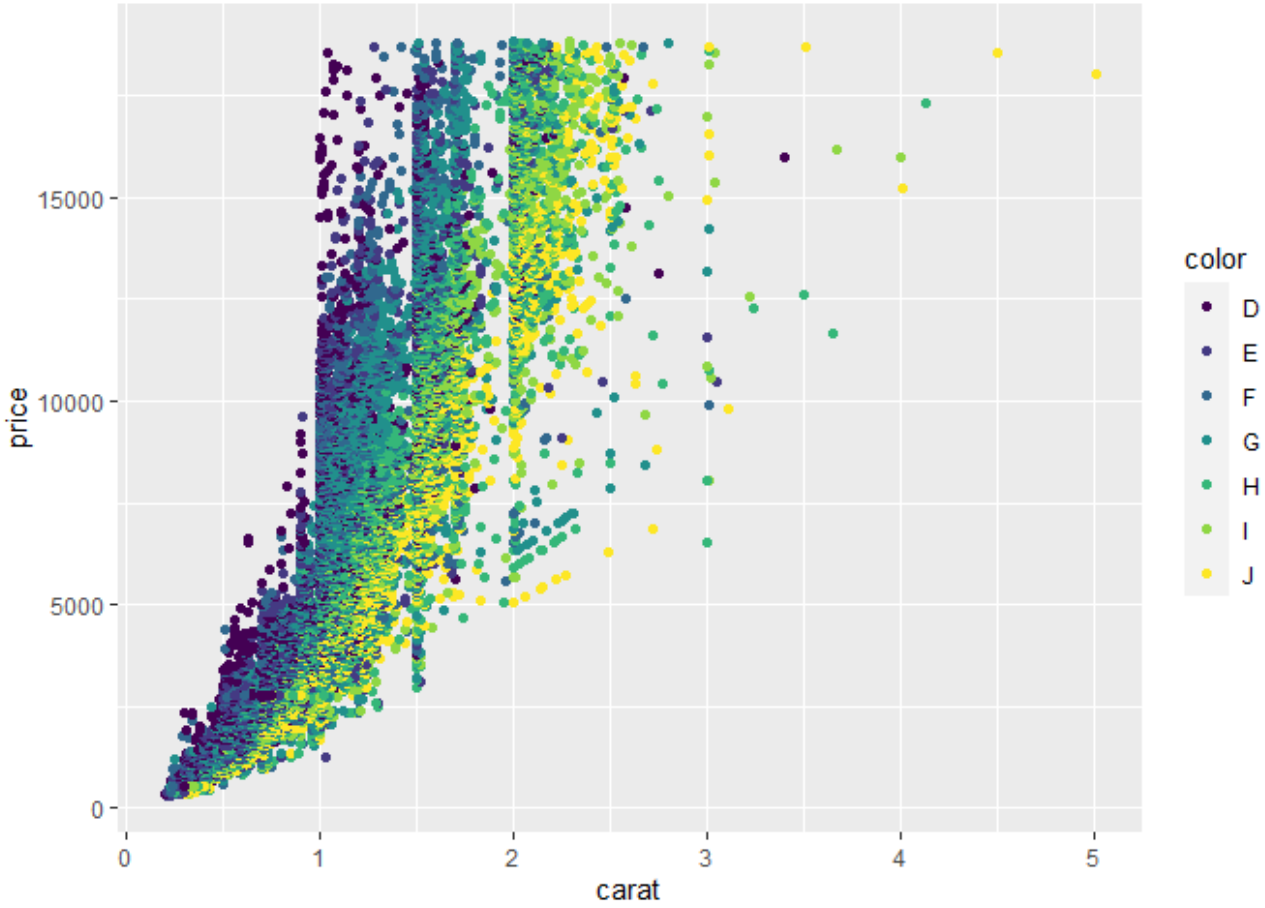
(Clarity-1.png)

Clarity Observations Continuing to examine additional factors that may predict what creates this range in price at similar carat values, when clarity is mapped by color, indicated by the ratings, from best to worst, of IF (flawless), VVS1, VVS2 (very very slight inclusion, grades 1 and 2), VS1, VS2, (very slight inclusion, grades 1 and 2), SI1, SI2 (slight inclusion, grades 1 and 2), I1 (included). Most diamonds with a flawless, IF, rating are clustered in the group of diamonds of 1.5 carats or smaller. As carat increases, lower grades of clarity make up more and more of the diamonds at each band, which can be observed in a clear color gradient. In diamonds of 3 carats or higher in this dataset, diamonds of 3 carats or higher are mostly of the I1, SI1,SI2 grades, but still command high prices. Thus, while clarity is

clearly associated with higher prices, especially in lower carat diamonds, large diamonds appear to claim high prices even with lower clarity ratings.

Additional Analysis: Color

```
ggplot(data = diamonds) +
   geom_point(mapping = aes(x = carat, y = price, color=color))
```

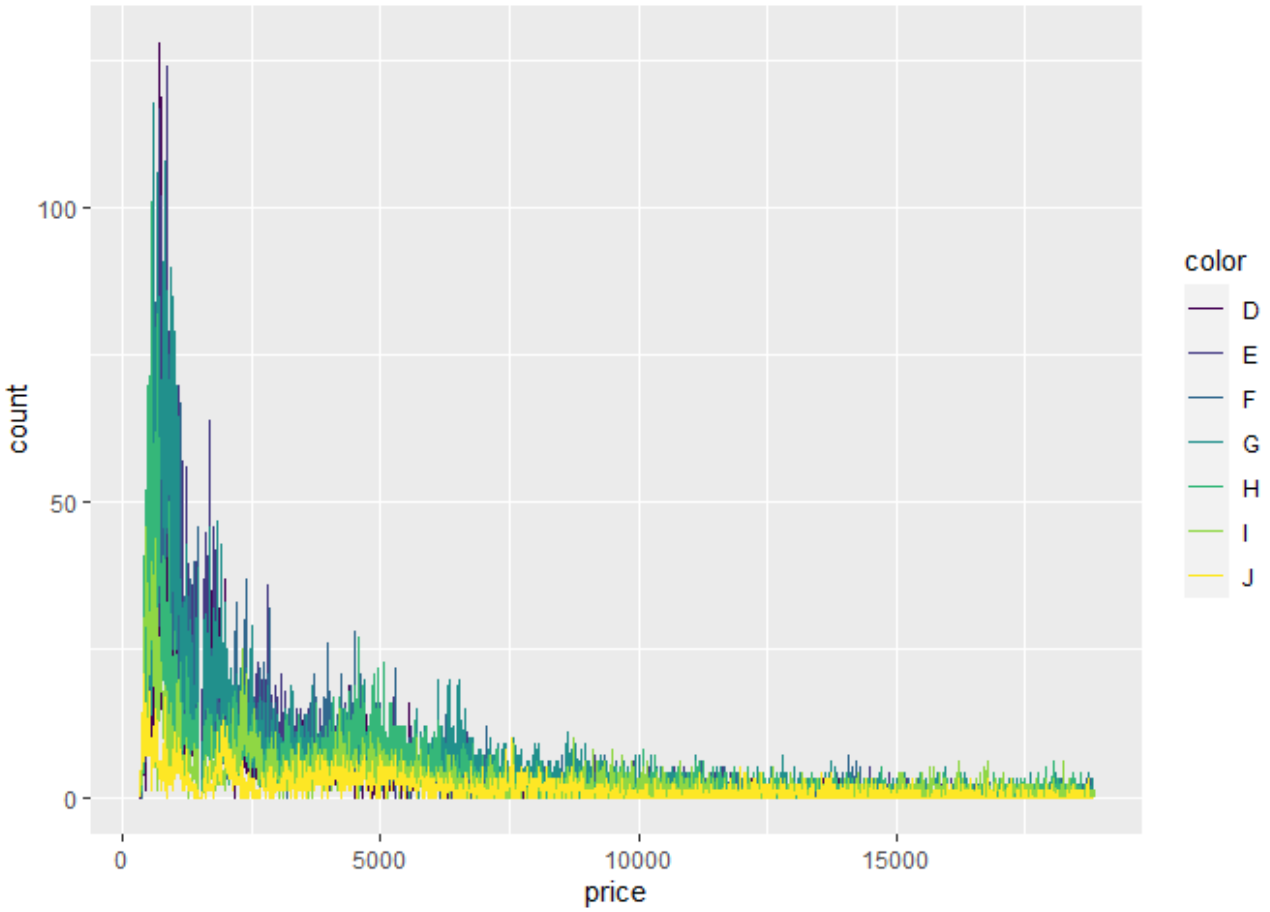


(Color-1.png)

Color Observations When the same data is color-coded for color, instead of clarity, using color ratings of D through J, where D is considered the best and J the worst, A general cluster of diamonds of D and E rankings is observed in the higher price bands of diamonds of below 2 carats, but over 2 carats even diamonds with a J rating for color have high priced, and some higher-carat diamonds with D color rating have relatively low prices, suggesting that color is not a strong predictor of the price of a large carat diamond.

Frequency of diamonds of different color and clarity, at different price points
Frequency ploygon of the number of diamonds at a given price, where their color is indicated by color

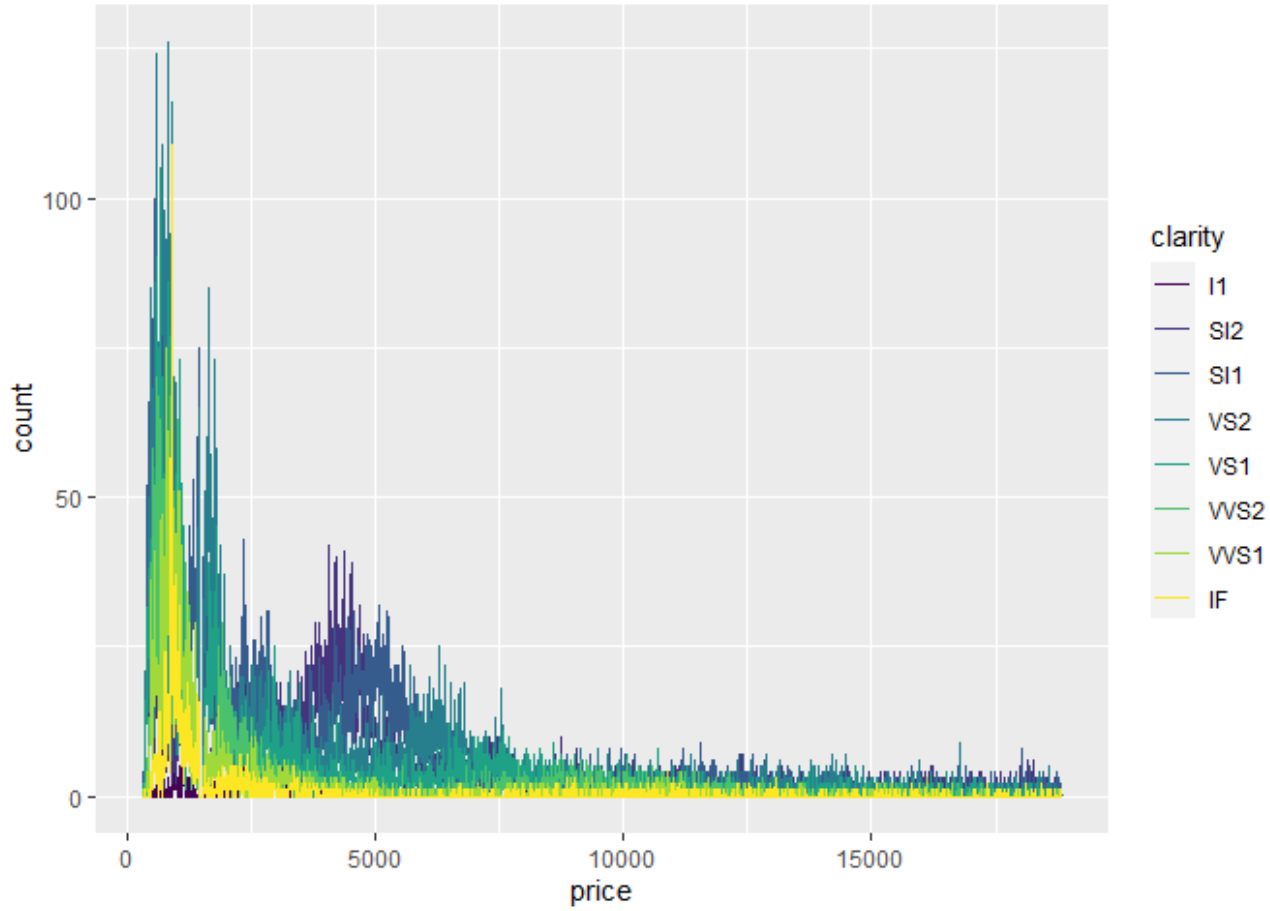
```
ggplot(diamonds, aes(price, color = color)) +
      geom_freqpoly(binwidth = 10)
```



(Color Frequency Distribution-1.png)

Discussion: Color Frequency At Different Price Points When examining a frequency polygon of the distribution of different color ratings, diamonds of a higher grade of color are much more common at the lower price points, while lower-grade color ratings form a larger and larger portion of available diamonds at higher price points. ***Frequency polygon of the number of diamonds at a given price, where the color indicates clarity***

```
ggplot(diamonds, aes(price, color = clarity)) +
      geom_freqpoly(binwidth = 10)
```



(Clarity Distribution-1.png) **Discussion: Clarity Frequency At Different Price Points** In contrast to color, the distribution of diamonds of different clarity grades have several different peaks and clusters. Low clarity graded diamonds have the first peak in quantity per price point, followed by a peak in high-grade clarity diamonds, followed by another peak in quantity of mid-quality clarity diamonds. There appears to be a small peak in low to mid quality clarity -diamonds around 2500 dollars, and a large number of low to mid quality clarity diamonds in the price range of 3000 to 5000 dollars. After 7500 dollars, the proportion of different clarity diamonds appears to even out as the net number of diamonds of all clarity ratings decreasing at the higher price points. # Communication

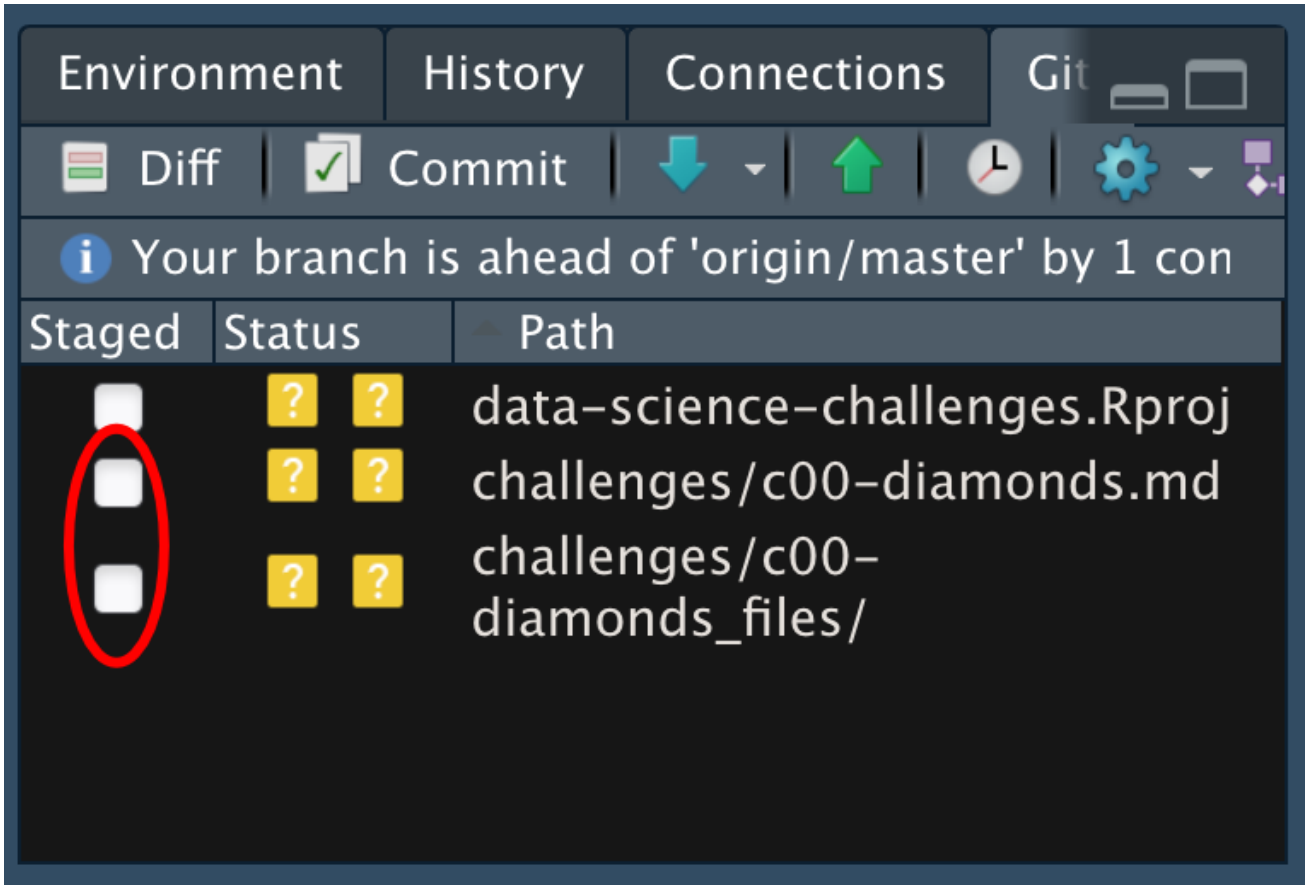
In this next stage, you will render your data exploration, push it to GitHub to share with others, and link your observations within our Data Science Wiki.

q3 *Knit* your document in order to create a report.

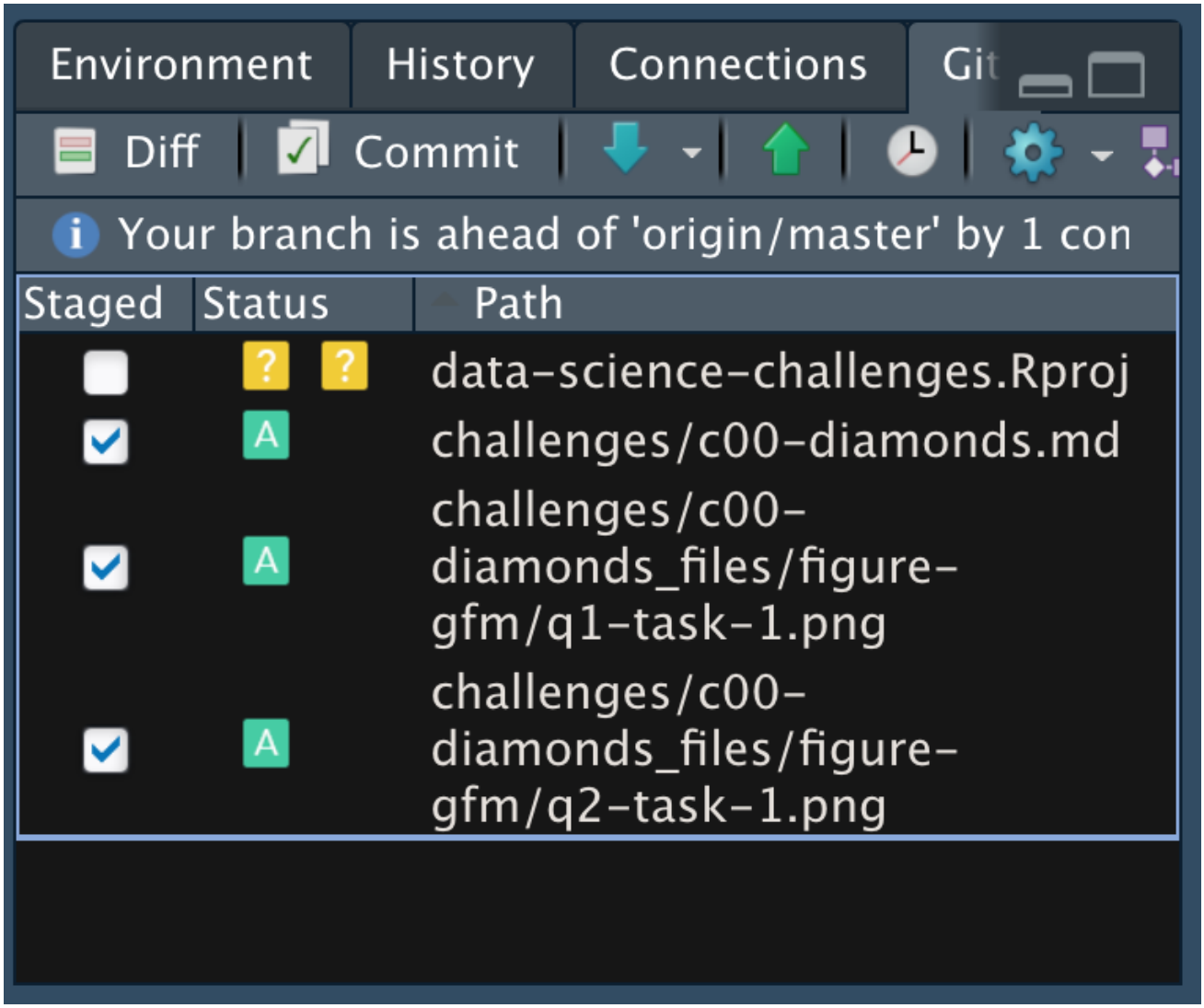
You can do this by clicking the “Knit” button at the top of your document in RStudio.

This will create a local `.md` file, and RStudio will automatically open a preview window so you can view your knitted document.

q4 *Push* your knitted document to GitHub.



You will need to stage both the `.md` file, as well as the `_files` folder. Note that the `_files` folder, when staged, will expand to include all the files under that directory.



q5 *Document* your findings in our Wiki. Work with your learning team to come to consensus on your findings.

The Datasets page contains lists all the datasets we've analyzed together.

q6 *Prepare* to present your team's findings!

q7 Add a link to your personal data-science repository on the Repositories page. Make sure to file it under your team name!