


Exploratory Data Analysis

Open RStudio and create a new project under your Module 3 folder and call it **Mod3Assignment2**. For this assignment, you will be creating an R Markdown document that will include the creation of plots and basics around exploratory data analysis (EDA). Once completed, all you need to do is submit the word document that is created.

Create the R Markdown Document

- 1.) In RStudio, select *File -> New File -> Text File*. This will create a blank text file in the same area that scripts were created in previous assignments (upper left panel). Save this file to your project as **Mod3Assign2Answer.rmd** (it is important to save with the .rmd extension as this saves the text file as an R Markdown file).
- 2.) Create a Header 1 with the title: **Module 3 - Assignment 2**
- 3.) Create a Header 2 with the title: **Last Name, First Name** (replace with your name)
- 4.) Create a Header 3 with the title: **Exploratory Data Analysis**
- 5.) Click on the dropdown arrow next to the Knit icon  at the top of the R Markdown Pane in RStudio and select Knit to Word.
- 6.) Notice that you now have a document in your files for the project named AssignmentAnswer.docx. This is the file you will be uploading later to Canvas.
- 7.) Add a chunk of code that will load the tidyverse package.

Does diamond color impact price?

- 1.) For this assignment, we will be using the diamonds data frame that comes with the tidyverse package
- 2.) Create a Header 4 with the title: **Diamond Color and Price**
- 3.) In the console, type `?diamonds` to see the help document for diamonds. Notice that this dataset includes prices, carat, cut, color, clarity, etc.
- 4.) As practice, in your R Markdown document, add a new chunk of code and create a scatterplot using this dataset with carat for the x-axis and price for the y-axis.
- 5.) After this chunk of code, add the following questions and provide the answers:
 - 1.) What do you notice from the scatterplot as the carat size increases?
 - 2.) From the scatterplot, what carats are most represented within the diamonds dataset?

Module 3: Assignment #2

- 6.) If you have ever researched diamonds, these results should not be surprising. However, we might not know if color is also tied to price.
- 7.) Create a new chunk of data and add the code to create a scatterplot that displays color on the x-axis and price on the y-axis. Right below this chunk of data provide a short explanation of what you see in your plot and if it is useful. Do you see any difference based solely on color?
- 8.) Next, create a new chunk of code and copy the code for the plot you did in instruction 4 (carat and price) but add an additional entry to display the diamond color on the plot.
- 9.) Based on these results, add a paragraph after your chunk of code and answer the following:
 - 1.) Does color impact the price?
 - 2.) Are certain colors associated with carat size? Provide an example.
- 10.) If you have noticed, we are dealing with a lot of data within this dataset. To make it easier to analyze, you may want to choose a random sample of data from the dataset. To do this, create a new chunk of data and execute the following code to create a random sample of 100:

```
dsample <- diamonds[sample(nrow(diamonds), 100), ]
```

- 11.) This creates a smaller set of data (dsample) that will help in exploring the data further. Add the previous code you created in instruction 8 to the code chunk that you just created the sample set with. This should be much easier to read. Also, add an additional line of code to create a fitted line in your plot [hint: `geom_smooth()`]
- 12.) This will create a fitted line for each of the color diamonds so it may look different from other plots we have done thus far. However, it might provide insight into the trends for certain colors as far as carat by price is concerned. **NOTE: Because this will create a random sample of data each time this is run, the output when you knit the file may be different from when you run the code in RStudio.**
- 13.) Using this same dataset and the three variables we have been working with, create an additional plot (e.g., boxplot, linechart) in a new chunk of data and explain what you think is going on with the data.