Stat 6021: Project 1

Background Information

You will be working in your assigned group of 3-4 students. Each group will work on the same data set. The data set that you will be working with describes more than 1,000 different diamonds that are for sale on `http://www.bluenile.com`. A .csv file of the data will be provided to you on Canvas. Please note that the .csv file contains a subset of the diamonds on Blue Nile as well as from other internet resources. The variables are:

- carat
- clarity
- color
- cut
- price

Detailed descriptions of these variables can be found on the diamond education page on Blue Nile.

Tasks

You have been approached by Blue Nile to perform the following tasks:

1. Use data visualizations to explore how price is related to the other variables (carat, clarity, color, cut), as well as how the other variables may relate to each other. Address the various claims on the diamond education page on Blue Nile.
2. Fit an appropriate simple linear regression for price against carat.

Deliverables

Your group will submit (one submission per group) on Canvas:

- A **report** (.html or .pdf file).
- An **R script** containing your code (.R or .Rmd file).

Each of you will also individually submit:

- A **Group Expectations Agreement**. Please see the Group Expectations Agreement document on Canvas for more information. Failure to upload this will result in a score of 0 for Project 1.
- A **Peer Evaluation**. Please see the Peer Evaluation document on Canvas for more information. The peer evaluation will be scored out of 20 points (in addition to the points for the project report).

Report Sections

The report should include the following sections:

1. A summary of findings that describes the high-level results of the analysis. This section should be written in a way that can be understood by a wide variety of readers, including readers with no background in statistics. A way to think about this is how newspaper articles report results from various studies, so avoid technical jargon. As an example, look at this article from the New York Times (paragraphs 10 and 11). If you are unable to access the article, a screenshot of the paragraphs is provided on Canvas. This section should be no more than 1 page.
2. A description of the data and the variables, as well as the data visualizations you created to address how price is related to the other variables as well as the claims made on the diamond education page. Be sure to provide contextual commentary on the visualizations.
3. A description of how you fitted the regression of price against carat, and the conclusions reached. If possible, be sure to provide some contextual commentary on the linear regression equation that you propose.

The audience for sections 2 and 3 is another classmate your client may hire to review your report.

Grading Guidelines

Your report will be graded A, B, C, D, or F and then converted to a 0-100 scale.

- A (90 to 100): the elements listed below are fully addressed and addressed well.
- B (80 to 89): a few elements listed below are missing or a few are not addressed well.
- C (70 to 79): some elements listed below are missing or some are not addressed well.
- D (60 to 69): a lot of elements listed below are missing or a lot are not addressed well.
- F (below 60): elements are generally missing or not addressed well.

Section 1

For section 1, you will be graded on:

- Clearly describing the high-level results of the analysis. What are the key findings that the reader needs to take away?
- Written for the right audience.

Section 2

For section 2, you will be graded on:

- Providing a description of the data and variables. This information can mostly be found in this document as well as the diamond education page on Blue Nile. If your group created any new variables based on existing variables, please include these variables in the description and clearly indicate that the variable was created by your group.
- Data visualizations provided to address how price is related to the other variables, including relevant comments.
- Appropriate univariate, bivariate, and multivariate visualizations are presented, including why these visualizations are being presented.
- Claims made on diamond education page on Blue Nile are addressed by your visualizations.

Section 3

For section 3, you will be graded on:

- A description of any transformation performed on the variables when fitting the SLR model, including reasons why these specific transformations were used.
- SLR assumptions checked.
- Contextual comments on how the SLR model inform us how price of diamonds are related to carat.

Additional Guidelines for Report

Your report should adhere to the following elements. Not following these will result in deduction of points (up to 5 points for each missing element).

- One member of the group will upload the report (.pdf or .html file) and the R script (.R or .Rmd file).
- Include the names of the group members and group number in the heading of your report.
- Have sections that are clearly labeled.
- Aim for no more than 20 pages. If you go over this limit a bit, that is fine.
- Do not use appendices as a way to work around the page limit. Anything that belongs in the main body of the report should be in the main body and not be tucked away in an appendix. I will not read anything in the appendix.
- The report should contain correct grammar, clear explanations, and professional presentation.
- The report should be cohesive.
- Your report should not include any R code. I should be able to repeat your analysis based on your description without looking at your R code.
- Relevant output from R (e.g. graphs, results from hypothesis tests, etc) should be included if the output is referenced to in the report.
- The text in your document should be readable after printing out on letter-sized paper.