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

- Part 1: 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 the project.
- Part 2: Report. Each group will submit:
  - A **report** (.pdf file preferred, .html file ok).
  - An **R script** containing your code (.R or .Rmd file).
- Part 3: 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). Each student will individually submit.

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

Sections 1 to 3 will be graded based on the following elements:

Section 1

For section 1, you will be graded on these elements:

- Clearly describing the high-level results of the analysis. What are the key findings that the reader needs to take away?
- Addressing your key findings as they relate to various claims made by Blue Nile.
- Writing for the right audience.

Section 2

For section 2, you will be graded on these elements:

- 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.
- Providing data visualizations to address how price is related to the other variables, including relevant comments.
- Addressing claims made on diamond education page on Blue Nile with your visualizations.
- Providing univariate, bivariate, and multivariate visualizations, as appropriate.

Section 3

For section 3, you will be graded on these elements:

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

Each of the elements listed above will be graded on a four-tiered scale and awarded 0 to 3 check marks:

- Good (3 check marks): the element is fully addressed and addressed well.
- Satisfactory (2 check marks): the element is addressed but can be improved upon.
- Unsatisfactory (1 check mark): the element is poorly addressed and needs to be reworked.
- Major issues (0 check marks): there are major issues with the element, or the element is not addressed at all.

Your report will be scored out of 30 check marks from your elements, and coverted to the following points.

- 30 check marks: 100 points
- 29 check marks: 98 points
- 28 check marks: 96 points
- 26 to 27 check marks: 93 points
- 24 to 25 check marks: 90 points
- 22 to 23 check marks: 87 points
- 20 to 21 check marks: 83 points
- 18 to 19 check marks: 80 points
- every 2 less check marks will drop your score down a third of a letter grade.

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 file preferred, .html file ok) 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.
- Be careful with using extremely long paragraphs. Using tables and / or bullet points can be useful in summarizing key pieces of information.
- 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.