Stat 6021: Project 2

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 your group will be working on contains house sale prices for King County, Washington, which includes Seattle. It includes homes sold between May 2014 and May 2015.

You can download the data set on kaggle.com or on Canvas. More information on the variables in the data set can be found in this discussion thread on kaggle.

Note: This is a fairly popular data set that is used to practice building regression models. Please refrain from looking at ideas on the world wide web, prior to submitting your project.

Tasks

Your group is to come up with **two** questions of interest that you would answer using your data set.

- The first question should involve linear regression, so your response variable has to be quantitative.
- The second question should involve logistic regression, so your response variable has to be binary.

Your group will also produce data visualizations that address both questions of interest. There is some flexibility in terms of the questions your group can pursue. The more interesting the questions, the better.

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. Each student will individually submit.
- Part 2: Proposal. Please see the Proposal document on Canvas for more information. Failure to upload this will result in a score of 0 for the project. Each group will submit.
- Part 3: Report. Each group will submit:
  - A **report** (.html or .pdf file).
  - An **R script** containing your code (.R or .Rmd file).
- Part 4: 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. Be sure to address your two questions of interest. You may also write about other interesting findings from your group as you were investigating the two questions of interest. 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. This section should be no longer than one page, and definitely no longer than two pages.
2. A description of the data and the variables. Also, if you created any variables that your group used in your analysis, please include their descriptions as well and clearly describe how these were created. Make it clear that these variables were not part of the original dataset.
3. In this section, clearly state the two questions of interest your group is pursuing, as well as some motivation about why these questions are being pursued.
   - The first question should involve linear regression, so your response variable has to be quantitative.
   - The second question should involve logistic regression, so your response variable has to be binary.

   Clearly state the response variable for each question. This section should be no longer than one page.
4. Provide data visualizations that help answer your first question of interest.
5. A description of how you used linear regression to answer your first question of interest.
6. Provide data visualizations that help answer your second question of interest.
7. A description of how you used logistic regression to answer your second question of interest.

The audience for sections 4 to 7 is another classmate your client may hire to review your report.

Note: As you will assessing how your models perform on test data, you should randomly split your data in a training set and a test set. Data visualizations and model building should be done only on the training data. Use `set.seed(6021)` when splitting the data. The code below will perform this split.

```
Data<-read.csv("kc_house_data.csv", sep=",", header=TRUE)
set.seed(6021)
sample.data<-sample.int(nrow(Data), floor(.50*nrow(Data)), replace = F)
train<-Data[sample.data, ]
test<-Data[-sample.data, ]
```

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, as well as any variables that you created.
- Variables you created are clearly indicated as such.
- The descriptions should be in an easy-to-read format, for example: table or bullet points. Avoid putting descriptions of several variables into long paragraphs.

Section 3

For section 3, you will be graded on:

- Stating the two questions of interest clearly.
- Motivating the two questions of interest. Why are these questions worth investigating?
- Clearly stating the response variable for each question of interest.

Section 4

For section 4, you will be graded on:

- Presenting appropriate univariate, bivariate, and multivariate visualizations that help answer your first question of interest.
- Providing contextual commentary on the presented visualizations.

Section 5

For section 5, you will be graded on:

- Giving clear reason(s) for the initial model(s) your group considered.
- Attempts to improve the model (data transformations, adding terms, removing terms, etc), as well as reasons for decisions made on how to improve the model.
- Checking model diagnostics.
- Checking for influential observations, high leverages observations, and outliers, and how your group handled them.
- Assessing your model(s) in terms of predictive ability on test data.
- Providing any interesting or relevant interpretations of your model(s).
- Providing relevant conclusions on how your model(s) address your first question of interest.
- Relevant R output provided.

Section 6

For section 6, you will be graded on:

- Presenting appropriate univariate, bivariate, and multivariate visualizations that help answer your second question of interest.
- Providing contextual commentary on the presented visualizations.

Section 7

For section 7, you will be graded on:

- Giving clear reason(s) for the initial model(s) your group considered.
- Attempts to improve the model, as well as reasons for decisions made on how to improve the model.
- Assessing your model(s) in terms of predictive ability on test data.
- Providing any interesting or relevant interpretations of your model(s).
- Providing relevant conclusions on how your model(s) address your second question of interest.
- Relevant R output provided.

Additional Grading 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 30 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.