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 perform the following tasks:

- There are some quality issues with this data set. Your group is to:
  - Identify observations with clear issues and suggest ways to fix them.
  - Identify issues with one or more variables and suggest ways to fix them.
  - Decide on a way to fix each of these and actually do it.
- After fixing the data quality issues with the observations and variables with the data set, use linear regression to build a model that explains how prices of homes in King county Washington are based on the other variables.
- After fixing the data quality issues with the observations and variables with the data set, use logistic regression to build a model that helps predict if a home in King county Washington is of good quality. We will define a home to be of good quality if its condition value is greater than 3 and its grade value is greater than 7.

Your group will also produce data visualizations associated with both models.

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** (.html or .pdf file).
  - 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 involving the linear regression model and logistic regression model. You may also write about other interesting findings from your group as you were building these models. 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 data set.

3. A description of how your group found the observations that clearly are data entry errors, and explain what you will do with them. Also identify potential issues with using some of the variables in your linear regression model and logistic regression model as they are recorded and defined in the data set, and propose some solutions of these issues.

   Based on what your group found in section 3, implement your suggested solutions before proceeding to the remaining sections. Do not start working on the remaining sections until your group has finished with section 3.

4. Provide data visualizations that explore how prices of homes are related to the other variables.

5. A description of how you used linear regression to build a model that explains how prices of homes in King county Washington are based on the other variables.

6. Provide data visualizations that explore what characteristics are associated with good quality homes. We will define a home to be of good quality if its condition value is greater than 3 and its grade value is greater than 7.

7. A description of how you used logistic regression to build a model that helps predict if a home in King county Washington is of good quality. We will define a home to be of good quality if its condition value is greater than 3 and its grade value is greater than 7.

The audience for sections 2 to 7 is another classmate your client may hire to review your report.

Note: As you will be 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. Be sure to perform this split before working on sections 4 to 7. Section 3 should be done on the entire data set without this split. The code below provides an example of how your group may want to use to form 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

Sections 1 to 7 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 from linear regression. What are the key findings that the reader needs to take away?
- The practicality of the key findings from linear regression. Make it clear who would benefit from the knowledge gained from these key findings.
- Clearly describing the high-level results from logistic regression. What are the key findings that the reader needs to take away?
- The practicality of the key findings from logistic regression. Make it clear who would benefit from the knowledge gained from these key findings.
- Writing for the right audience.

Section 2

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

- Clearly describing the data and variables, as well as any variables that you created. Variables you created are clearly indicated as such.

Section 3

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

- Identifying the observations that clearly have a data entry error. Explain why these are clearly data entry errors. **Note:** Your group has to identify at least three types of data entry errors.
- Explaining what you will do with these observations, once you have identified them as having a data entry error. Be sure to implement these steps before proceeding with section 4 and beyond.
- Identifying variables in the data set that have issues with them as they are recorded and defined in the data set. In other words, some variables cannot be used in their current form. **Note:** Your group has to identify at least three variables (columns) or groups of variables that may be invalid predictors as they stand, or may present problems in multiple linear regression as discussed in class.
- Explaining what you would do with the variables, once you have identified them. Be sure to implement these steps before proceeding with section 4 and beyond.

Section 4

For section 4, you will be graded on these elements:

- Presenting appropriate univariate visualizations.
- Presenting appropriate bivariate visualizations.
- Presenting appropriate multivariate visualizations.
- Providing contextual commentary on the presented visualizations.

Section 5

For section 5, you will be graded on these elements:

- Giving clear reason(s) for the initial model(s) your group considered.
- Attempting 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) explains how prices of homes in King county Washington are based on the other variables.

Section 6

For section 6, you will be graded on these elements:

- Presenting appropriate univariate visualizations.
- Presenting appropriate bivariate visualizations.
- Presenting appropriate multivariate visualizations.
- Providing contextual commentary on the presented visualizations.

Section 7

For section 7, you will be graded on these elements:

- Giving clear reason(s) for the initial model(s) your group considered.
- Attempting 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) helps predict if a home in King county Washington is of good quality.

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 90 check marks from your elements, and coverted to the following points.

- 90 check marks: 100 points
- 88 to 89 check marks: 98 points
- 84 to 87 check marks: 95 points
- 81 to 83 check marks: 93 points
- 78 to 80 check marks: 90 points
- 75 to 77 check marks: 87 points
- 72 to 74 check marks: 83 points
- 69 to 71 check marks: 80 points
- every 3 less check marks will drop your score down a third of a letter grade.

Additional Grading Guidelines for Report

Your report should also 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.
- 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.