Stat 6021: Project 2

Background Information

The dataset 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 should download the dataset on kaggle.com. More information on the variables in the dataset can be found in this discussion thread on kaggle.

Note: This is a fairly popular dataset 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, 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. You are advised to run your ideas with me or the GTA before proceeding. The more interesting the questions, the better.

Deliverables

This project will take place over the last few days of the semester with various components due at different dates. The list of deliverables is shown below:

- Part 1: Group Expectations Agreement (.pdf file), due Tuesday August 9, by end of class time. Every member submits via Assignments.
- Part 2: Typed Report (.pdf or .html file), including R script (.R or .Rmd file), due Thursday August 11. One submission per group via Assignments.

- Part 3: Video Presentation (provided by a link to the zoom recording or a link to any free video hosting site), due Thursday August 11. One submission per group via Assignments.
- Part 4: Feedback on Classmates' Presentation, due Friday August 12. Everyone types in their feedback to a textbox on Assignments.
- Part 5: Self- and Peer-Evaluation on Project (due Friday August 12). Everyone submits via Test & Quizzes.

More information on how to complete each part and how each part will be evaluated is provided in the sections below.

Part 1: Group Expectations Agreement

Please see the effective group work document for more information, on Collab.

Submission

Everyone should submit a copy of their group's Group Expectations Agreement via Assignment. Failure to do so will result in getting a 0 for Project 2. Documents within a group should be identical. Submission indicates agreement with the document.

Part 2: Typed Report (100 points)

Your group is to type up a report for this project. One member of your group is to upload the report and the R script via Assignments on Collab. The report should include the following sections:

1. An executive summary that describes the high-level results of your group's analysis. This executive summary 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 more than 2 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. In this section, also clearly state the two questions of interest your group is pursuing, as well as some motivation about why these questions are being pursued. Also provide relevant data visualizations and how these help your group gain insight into your questions of interests.
3. A detailed description of how you used linear regression to answer your first question of interest.

4. A detailed description of how you used logistic regression to answer your second question of interest.

The audience for sections 2, 3, and 4 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.

Grading Guidelines

Each section, 1, 2, 3 and 4, will be graded A, B, C, D, or F and then converted to a 0-100 scale.

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

Your score for the report will be the average score from these four sections.

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.
- Clearly stating the two questions of interest your group is pursuing.
- Providing motivation as to why your group is pursuing these two questions of interest.
- Data visualizations provided to provide insight into your questions of interest, including relevant and contextual comments.
- Appropriate univariate, bivariate, and multivariate visualizations are presented, including why these visualizations are being presented.

Section 3

For section 3, 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 4

For section 4, 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).

- Include the names of the group members and group number in the heading of your report.
- Have sections that are clearly labeled. You may create subsections within the four sections listed above if needed.

- 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.
- Uploading the files in the right file format: .pdf or .html for the report, and .R or .Rmd for the script. No other formats and no zipped files.

Submission

Please submit your group's report (.pdf or .html file) and R script (.R or .Rmd file) via Assignments. One upload per group.

Part 3: Video Presentation (100 points)

Each group will provide a video recording of a presentation that is **no more than 15 minutes long**. This presentation should be designed to be understandable by anyone familiar with the topics covered in the course, but who has not worked on this dataset and has not read your typed report. Each group is free to organize who talks about which topics during the presentation. Not everyone needs to talk, but all group members should be active contributors to the presentation materials.

Presentation Guidelines

- The presentation should use PowerPoint or something similar as a visual.
- Presentation should be viewed as a summary of the report. You will not be able to present everything that is in the report.
- General rule of thumb: 1 to 2 minutes per slide.
- Each slide should be clear and easy to read.
- The presentation should be clear, with good pace and logical flow.

- R output should be clearly labeled.
- The recording should be hosted on zoom (preferred. Directions provided on Collab). Other popular video hosting sites such as youtube and vimeo are acceptable (as long as the viewer doesn't have to register or pay to view).

Grading Guidelines

Your group's presentation will be graded on three criteria: big picture, data visualizations, and models. Each criteria will be graded A, B, C, D, or F and then converted to a 0-100 scale.

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

Your score for the presentation will be the average score from these three criteria.

Big Picture Criteria

You will be graded on:

- The questions of interest are clearly stated.
- High-level results of the analysis are presented and connected to the questions of interest.

Data Visualizations Criteria

You will be graded on:

- Description of data and relevant variables provided. If a variable is not presented, you don't have to describe it.
- Relevant visualizations and / or numerical summaries are presented.
- Visualizations are clear and viewer knows what to look at.

Models Criteria

You will be graded on:

- Relevant models are presented.
- Some information about the model building process provided.
- How your linear regression model(s) answer the first question of interest.
- How your logistic regression model(s) answer the second question of interest.

Additional Grading Guidelines for Presentation

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

- Include the names of the group members and group number on your first slide.
- Slides should be easy to read. Consider using bullet points instead of paragraphs.
- Slides should not be overcrowded. The viewer should know what to look at in the slides and not have to guess or squint.
- The presentation should be delivered at a comfortable pace. With a short presentation, presenters have a tendency to rush.
- Adhere to the time limit: **15 minutes**.
- Providing the needed information to access the link for your group's recording. If password protected, please provide the password.
- If hosting the recording outside of zoom, be sure that the viewer doesn't have to pay or register in order to view the recording.

Submission

Please submit your group's presentation (link to where the recording is hosted. If your link is password protected, please provide the password as well) via Assignments. One upload per group.

Part 4: Feedback on Classmates' Presentation (10 points)

More information provided on the Project 2 page on Collab.

Submission

Each student will submit the feedback via Assignments.

Part 5: Self- and Peer-Evaluation of Group Participation in Project 2 (10 points)

- You will anonymously evaluate each group member's contributions to the project.
- Complete this by giving an honest of your own performance and that of your group members in project 2.

Submission

Each student will submit the group participation evaluation via Test & Quizzes.