Final Course Project
STAT 207 - Data Science Exploration

Due: Wednesday, December 6 by 11:59 pm on GitHub & Canvas

Main Goal of Analysis

The main goal of this project is to tell a compelling story based on the data science analyses you will perform on a dataset.

You are required to perform three main analytical tasks:

1. Linear Regression Research Question ~ pick a numerical response variable and at least four explanatory variables that you suspect might affect your response variable. Explore whether there is a linear relationship between these explanatory variables and the response variable.
2. Logistic Regression and Classification Research Question ~ pick (or make) a categorical response variable with two levels and at least four explanatory variables that you suspect might affect your response variable. Explore whether there is a linear relationship between these explanatory variables and the log-odds of the success level of the response variable.
3. Inference Research Question ~ pick two variables (at least one of them categorical) and explore the relationship between these variables *in the dataset* followed by *in a population*.

Additional descriptions for these tasks can be found later in this document. If you think that there are additional questions or analyses that would add additional insights to your overall research goal, you're more than welcome to pursue these in addition to what is required in this document.

Project Format

The project will have three components. Additional details can be found below.

1. Project Report [80 points]
2. Project Presentation [15 points]
3. Student Summarization for Another Group Presentation [5 points]

Group Structure

You can work in groups of up to 3 people, or you can work individually.

- If you work with a group of 3, you must do at least 25% of the work in order to get full credit.
- If you work with a group of 2, you must do at least 33% of the work in order to get full credit.

Dataset Options

You can choose your own dataset, or you can choose from one of the three supplied datasets below. There are several places you can go to find interesting datasets, but here are some places to start:

- https://www.kaggle.com/datasets
- https://archive.ics.uci.edu/ml/datasets.php

Choosing your own data:

If you choose your own data, it must meet the following specifications.

1. It must be a sample (ideally random, or reasonable to assume random) from a larger population. You are allowed to take a random sample of a population dataset and use that as your sample.
2. It must have at least five variables total
    a. Variables that have uninformative information don't count and won't be useful. Examples of uninformative variables include those that records a row name, row id, or is a linear combination of the other variables in the datasets. If you aren't sure, come ask!
3. It must have at least one categorical variable
4. It must have at least one quantitative variable

Provided datasets:

Some of these datasets might have missing values.

1. Spotify (spotify_top_200_sample.csv)
    a. This is a random sample of song attributes for songs that have been on the Top 200 Weekly Global charts of Spotify in the years 2020 and 2021.
    b. Read more about this data here: https://www.kaggle.com/datasets/sashankpillai/spotify-top-200-charts-20202021
2. Body Dimensions (bdims.csv)
    a. This dataset is comprised of various body dimensions of a random sample of physically active adults
    b. While we have used this dataset briefly in a previous lecture, there are many other research questions that you can explore with this dataset. **You should not choose to perform an analysis that we have already done.**
    c. Read more about this data here: https://www.openintro.org/book/statdata/?data=bdims
3. Ames, Iowa Housing (ames.csv)
    a. This is a (assume random) sample of residential home sales in Ames, Iowa between 2006 and 2010 with characteristics about the homes and the sale.
    b. Read more about this data here: https://www.openintro.org/book/statdata/?data=ames

Project Report Specifications

Deadline: Wednesday, December 6 by 11:59 pm on GitHub
Format:

- Jupyter notebook
- This should be a clean data analysis report that you could submit to an employer or client (not a homework assignment). At the very least, your report should have a title, headings for each section, and be written in paragraphs and with complete sentences.
- You can use and modify the attached project_template.ipynb file as a template for this report if you'd like. You can add and delete as many cells as you'd like in this file.

1. **Introduction [10 points]**
    a. Title: Give your research report a title
    b. Motivation: Describe the motivation for why you or someone else would want to explore your dataset (or a dataset of this type). You can give background research (with citations) if this would help back up your motivation.
    c. Research Questions: You will answer at least three **sets** of research questions. In your introduction, you should briefly describe why you (or someone else) would be interested in answering these research questions. How could the answers to these research questions be used?
    d. Dataset: Display at least 6 observations from your dataframe in this section and show how many rows and columns it has.
    e. Think Critically: What limitations exist within your data? What information is included in your data? What information (observations or variables) is not included in your data? What might you want to add? Are there any practical or ethical considerations for this data? You don't need to answer all of these questions, but you should answer some.

2. **Linear Regression Research Question Set [20 points]**
    a. Goal: You should pick a numeric response variable and at least four explanatory variables that you suspect might affect your response variable. Explore whether there is a linear relationship between the explanatory variables and the response variable, in the dataset and in the population.
        i. For instance, you could ask two questions like: "Is there a linear relationship between y and x1, x2, x3, and x4 in the sample? What explanatory variables should we include in the model to build a parsimonious model?"
    b. State your research question(s) you'd like to answer with your analysis.
    c. Use at least one linear regression model to answer this question. Include the following:
        i. Show the summary output for your linear regression.
        ii. Write out the linear regression equation for your model. Use appropriate notation.
        iii. Check the linear regression conditions of your model.

iv. Use at least one technique to perform feature selection for your model to build a parsimonious model. Motivate why you selected your feature selection technique. Share your final set of predictors for your model.

v. Interpret one of the slope coefficients (not the intercept) for your linear model.

vi. Discuss what percent of variability in your response variable is explained by this model in the dataset. Is this high? Is this low?

vii. Make at least one prediction with your model.

d. Finally, discuss how your linear regression analysis helps answer your research question.

3. **Logistic Regression and Classification Research Question Set [20 points]**

a. Goal: You should pick (or make) a categorical response variable with two levels and at least four explanatory variables that you suspect might affect your response variable. Explore whether there is a linear relationship between the explanatory variables and the log-odds of the success level of the response variable, in the dataset and in the population.

i. For instance, you could ask questions like: “Is there a linear relationship between the log-odds of the success level of y and x1, x2, x3, and x4 in the sample? How does a classifier built on this model perform on new data?”

b. State your research question(s) you’d like to answer with your analysis.

c. Use at least one logistic regression model to answer this question. Include the following:

i. Split your dataset into a training dataset and testing dataset.

ii. Fit your logistic regression model to the training data.

iii. Show the summary output for your logistic regression model.

iv. Write out the logistic regression equation for your final model.

v. Use an ROC curve on the training data to pick a good predictive probability threshold. Explain why this is a good predictive probability threshold, given your research goals.

vi. Use your logistic regression model to calculate the ROC curve and AUC on your test dataset. How does this compare to the corresponding values for the training data?

vii. Use the predictive probability threshold to classify your test data. What is the accuracy rate, sensitivity, and specificity of your classification of the test data?

d. Finally, discuss how your logistic regression analysis helps answer your research question.

4. **Inference Research Question Set [20 points]**

a. Goal: You should pick two variables (at least one categorical) and explore the relationship between these variables *in the dataset* (descriptive analytics) and *in a population* (inference).

i. For instance, you could ask “What is the relationship between x and y in this dataset? Is there an association between x and y in my population?”

ii. Choose a parameter that incorporates a difference, like a difference in two population proportions, difference of two population means, or population mean of paired data (differences).

b. State your research question that you will answer with your analysis. Remember, descriptive analytics only involves describing relationships in the dataset that you have, so your first research question should be *just* about the data. Then, inferential statistics involves answering research questions about populations given a random sample from that population. Your second research question should reference the population from which your data were collected.

c. Generate at least one visualization to observe the behavior of your variable(s) of interest in the data. Describe what you see in your visualization, what it tells you, and how it helps answer your research question.

d. Complete at least one simulation-based hypothesis test to answer this research question, including

i. Stating your hypotheses,

ii. Checking the conditions for this test,

iii. Simulating a sampling distribution for the statistic of interest

iv. Calculating a p-value (or confidence interval) for this test by hand, and

v. Using your results to state a conclusion.

e. Finally, discuss how your conclusion answers your research question.

5. **Conclusion [10 points]**

a. Summarization: Summarize the findings of your individual research questions in the conclusion. Provide at least a paragraph. (This will likely be a restatement of what you have already included in your report).

b. Limitations: What limitations did you face in your analysis, results, or interpretations? What challenges did you face in your data analysis? What contextual information is important before you make strong claims from these results?

c. Future work: If you (or someone else) were to conduct future work based on these analyses, what kind of research questions or analyses might that entail?

Project Presentation

In lab on Wednesday, December 6

Format:

- Your presentation should be between 4-6 minutes long. If your lab section has fewer groups, you may be able to extend your presentation. If your lab section has a large number of groups, this guideline may be strictly enforced.
- You will **not** be able to present every component of your project. Instead, you should select one of the three main research questions to present.
- You must present some part of the presentation (if you’re in a group) in order to get full presentation credit.

- Presentation should be presented in **slides** (not the Jupyter notebook). I suggest preparing 2-5 slides.
- You should submit a rough draft of your slides, including at least your cover slide to the Canvas assignment for the Project Presentation. This will not be graded but is done to enable the Project Summarization assignments through the Peer Review tool.

Grading Breakdown:

1. Slides [10 points]
    a. Content [8 points] ~ Your team should present *some* content for the following topics. You will be graded on correctness, including analyses being appropriate for the data and results being interpreted correctly.
        i. Introduction (3 points)
        ii. Select one of the three main research questions from the project report (3 points)
            1. Linear Regression Research Question
            2. Logistic Regression and Classification Research Question
            3. Inference Research Question
        iii. Conclusion (2 points)
    b. Layout [2 points]
        i. Content is well organized, fonts are easy to read
        ii. Slides are engaging and not too wordy
2. Presentation [5 points]
    a. Narrative/Motivation [3 points]
        i. Clearly explain motivation for the analysis
        ii. Clearly state research question and how it relates to the motivation
        iii. Clearly summarize answer to research question that were discovered from the analyses
    b. Presentation [2 point]
        i. All team members speak and present some portion of the material
        ii. Team members speak loud enough for everyone to hear
        iii. Team members understand the material; they are not reading directly from a notecard or script (you may have some reference materials, including the slides with you).

Student Summarization for Another Group Presentation

Deadline: Wednesday, December 6 by 11:59 pm on Canvas

Purpose:

- For presenters: to give the presenting teams constructive feedback on how clearly they were able to communicate and answer their research question with their analyses and how well they were able to motivate their research to a peer
- For listeners: to gain practice being able to extract the most important parts of an oral research presentation

Steps:

- On the day of the presentations, you (as an individual) will be randomly assigned to another group presentation through a Peer Review on Canvas

- After watching this group’s presentation, you should respond to the five questions in the “Student Summarization of Presentation” document and submit it through the Peer Review on Canvas
- The group that you summarized in this report will be able to see the constructive feedback and your summarization
- If you are unclear about how to answer the questions for this document, you are encouraged to reach out to the group that you were assigned to for clarification

Questions to Answer:

1. What is the motivation for the analysis in this presentation? In other words, why should you (or someone else) care about the analysis that you just read/listened to?
2. Did the analyses and conclusions answer the research questions that were stated at the beginning of the report/presentation? If so, how? What were the answers to these research questions?
3. How would the results/answers to these research questions be useful to someone?
4. After watching this presentation, what is one follow up question that you would have for this group? This could be a question about the work that they already did or a suggestion of an interesting question for future work.
5. Any other supportive or constructive feedback that you’d like to give this group on their work? (Not required)

Graded: [5 points] for completion

Intended Audience/Reader of your Project

The intended audience of your report/presentations should be someone who has the same level statistical/python knowledge as you and your STAT 207 classmates. **Theoretically, you should be able to send/present your report to one of your classmates (who is not on your team), and they should be able to understand everything that you did and the claims that you are making.**

Grading

In addition to being graded for correctness and completion (as noted), this project will be graded on a qualitative basis. Qualitatively, we will be looking for the following things:

- **Clarity about Analyses, Algorithms, and Data Choices**
    - Someone who has taken a STAT207-level class should be able to read through your report and/or watch your presentation and easily be able to do the following.
        - Replicate what you did in your analyses.
        - Know why you made the choices that you did in your analyses.
- **Clarity about Motivation (i.e. the “so what?”) of your analyses**
    - Beginning of the Report & Presentation
        - Someone who is **about to** read your report and watch your presentation should be able to clearly answer the questions.
            - Why should I (or someone else) care about the report that I am about to read/listen to?
            - What research questions do they intend to answer?
            - How do these research questions relate to their motivation?
        - Therefore, in the introduction of your report and presentation you should make this clear.
    - Middle of the Report & Presentation:
        - While **in the middle of** your report and presentation, your audience should be able to clearly answer the question:
            - How do each of these analyses/algorithms/data choices that they’re making/using tie back into the overarching motivation of this whole analysis?
        - Therefore, each new analysis/model/algorithm/data choice that you make, you should explain this and make it clear to your audience.
    - End of the Report & Presentation:
        - Someone who has **just finished** reading your report and watching your presentation should be able to clearly answer the questions:
            - Why should I (or someone else) care about the analysis that I just read/listened to?
            - Did their analyses and conclusions answer the research questions that they stated at the beginning of the report/presentation? If so, how? What were the answers to these research questions?
            - How would the results/answers to these research questions be useful to someone?
        - Therefore, in the conclusion of your report and presentation you should make this clear.