STAT207 – Data Science Exploration - Project – 50 Points

Due: Friday, May 13 1:30pm CST on Github.

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 can work in groups of up to 3 people. Or you can work by yourself.**

- **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.**

To receive full credit, you should follow the steps and answer the questions given in this document for your project. However, 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 stipulated in this document.

In addition to being graded for correctness and completion, 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 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 (ie. the "so what?") of your Analyses**
    - *Beginning of the Report and 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 and 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, for 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 and 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.

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 STAT207 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.**

Project Format

This project will have three components.

Project Report [35 pt]

Deadline: Friday, May 13 by 1:30pm CST on Github.

Should contain: Everything stipulated in the **Project Report Specifications** discussed below.

Format:

- Jupyter notebook.
- This should look like a clean data analysis report that you would theoretically submit to an employer (not a homework assignment). Thus, at the very least, your report should have:
  - a title
  - headings for each of your sections
  - You should write paragraphs and in complete sentences.
- You can use and modify the attached project **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.

Graded:

- See "Project Report Specifications" section below for point breakdown.

Project Presentation [12 pt]

Presentation Dates:

- Option 1: During your final lab section time **Thursday April 28 in-person**
- Option 2: During our designated "final exam period" **Friday May 13 1:30-4:30pm CST ONLINE**
- If for some reason neither of these options are viable, please let Dr. Ellison know ASAP!

Format:

- Ideally, keep your presentation within 7-10 minutes long.
- 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).

Graded:

- See attached **presentation rubric** for what you should present and how you will be graded.

Student Summarization for Another Group Presentation [3 pts]

- **Deadline**: **Friday, May 14 2:30pm CST** on **Canvas**.
- **Purpose**:
  - For presenters:
    - The purpose of this final part of the project **for the presenters** is to give the presenting teams constructive feedback on how clearly they were able to communicate and answer their research questions with their analyses and how well they were able to motivate their research to a peer.
  - For listeners:
    - The purpose of this final part of the project **for the listeners** is to gain practice being able to extract the most important parts of an oral research presentation.
- **Steps**:
  - On the day of ***your*** presentation (April 28 or May 13), you (as an individual) will be randomly assigned to another group presentation.
  - After watching this group's presentation, you should fill out the **"Student Summarization of Presentation"** document and submit it individually 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 in this document, you are encouraged to reach out to the group that you were assigned to for clarification.
- **Graded**:
  - For completeness

Dataset Options

You can choose your own dataset or you can choose from one of the three supplied datasets below. The csvs for each of these datasets are located in the same folder that this document is in. There is more information about each of these datasets below.

There are several places you can go to to find interesting datasets, but here are some places you can start.

https://www.kaggle.com/datasets

https://archive.ics.uci.edu/ml/datasets.php

https://corgis-edu.github.io/corgis/csv/

https://data.world/datasets/regression

https://github.com/fivethirtyeight/data

For students interested in sports data:

- NFL: https://www.nflfastr.com/
- MLB and other baseball: https://billpetti.github.io/baseballr/
- CFB: https://saiemgilani.github.io/cfbfastR/index.html
- More sports stuff: https://sportsdataverse.org/

Choosing your Own Dataset

If you decide to choose your own dataset, it must meet the following specifications.

1. It must be a random sample from a larger population (ideally with a sample size that is less than 10% of this population size). You are allowed to take a random sample of a non-random dataset and use that as your sample.
2. It must have at least one categorical variable.
3. It must have at least one numerical variable.
4. It must have at least five variables total.
    a. *Variables that have a different value for every row don't count and won't be useful.*

Pre-Selected Dataset Options

Some of these datasets might have missing values! You should always check!

1. **Quality of Life and Cost of Living City Data** The observations in the movehub_data.csv file contain various cost of living and quality of life metrics for a **random sample of 100 global cities**. This data was extracted from movehub.com.
   The full dataset and more information about the full dataset can be found here: https://www.kaggle.com/datasets/blitzr/movehub-city-rankings?select=movehubqualityoflife.csv
   a. You can assume that the cost of living prices given in the dataset are listed in US dollars.
   b. You can assume that the quality of life metrics are supplied and calculated by movehub.com.

2. **Body Dimensions Dataset** (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 in previous lectures and assignments, 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 dataset here: https://www.openintro.org/book/statdata/?data=bdims

3. **Ames, Iowa Housing Dataset** (ames.csv)
   a. This is a (assume random) sample of residential home sales in Ames, Iowa between 2006 to 2010 and properties about the homes and the sale.
   b. Read more about this dataset here: https://www.openintro.org/book/statdata/?data=ames

Project Report Specifications

Your report should include the analyses, code, and explanations detailed in each of the following sections.

1. Introduction [3.5 points]

Title: Give your research report a title.

Motivation: After picking a dataset describe the motivation for why you or someone else would want to explore this dataset or a dataset of this type. You can give background research (with citations) if this would help back up your motivation.

Research Questions: In this report you will answer at least four **sets** of research questions. In your introduction you should briefly discuss each of your research questions that you plan to answer using this dataset and how you plan to answer that research question (ie. what analysis will you use). Finally, you should briefly describe why you (or someone else) would be interested in answering this research question. How could the answers to these research questions be used?

Dataset: Display your dataframe in this section and show how many rows and columns it has.

2. Descriptive Analytics Research Question Set [6.5 points]

For your first set of research question(s), you should pick three or more variables and explore the relationship between these variables *in the dataset.*

For instance, you should ask a research question set like: "What is the relationship between x and y in this dataset? And furthermore, how does this relationship between x and y change for different values of z?"

1. State your research question you will answer with your analysis. Remember, descriptive analytics only involves describing relationships in the dataset that you have, so your research question should be *just* about the dataset.
2. Use at least one visualization to answer this question. **You should have at least one visualization that incorporates 3 or more variables in the same plot.**
3. Use at least one set of summary statistics to help answer this question as well.
4. Describe what you see in your visualizations and summary statistics, what they tell you, and how they help answer your research question.

3. Inference Research Question Set [6.5 points]

For your second research question, you should pick two variables (one of them categorical) and explore the relationship between these variables *in a population.*

For instance, you should ask a research question like: "Is there an association between x and y in INSERT POPULATION HERE?"

Choose one of the hypotheses below to help you answer this research question.

Hypothesis	***Equivalent to saying…***
H_A: $\mu_1 - \mu_2 \neq 0$ *μ_1 =average response variable value for level 1 of categorical variable* *μ_2 =average response variable value for level 2 of categorical variable*	*There is an association between the <u>numerical variable</u> and <u>categorical variable</u> in the population.*
<u>ANOVA</u> *H_A: at least one pair of population means are different from each other* *μ_i =average response variable value for level i of categorical variable (for i=1,2,…,p)*	*There is an association between the <u>numerical variable</u> and <u>categorical variable</u> in the population.*
H_A: $p_1 - p_2 \neq 0$ *p_1 =proportion of categorical variable level 1 values that are the success level (of the other categorical variable)* *p_2 = proportion of categorical variable level 1 values that are the success level (of the other categorical variable)*	*There is an association between the <u>two categorical variables</u> in the population.*

1. State your research question that you will answer with your analysis. Remember, inferential statistics involves answering research questions <u>about populations</u> given a random sample from that population. So your research question should be about the population your dataset was randomly sampled from.
2. Use at least one hypothesis test to answer this research question.
    a. Make sure you state your hypotheses.
    b. Make sure you check your conditions for this hypothesis test.
    c. Calculate a p-value (or confidence interval) for this hypothesis test and use it to state your conclusion.
3. Finally, discuss how your conclusion answers your research question.

<u>Hint</u>: You can create a 0/1 categorical variable from a numerical variable in a given dataframe df by using/modifying the code below.

```
df['new_cat_var'] = 1*(df['num_var']>=some_number)
```

4. Linear Regression Research Question Set [7 points]

For your third research question set, you should pick a numerical response variable and at least 4 explanatory variables that you suspect might affect your response variable and then explore whether there is a linear relationship between these explanatory variables and the response variable <u>in the dataset</u> as well as <u>the population</u>.

For instance, you should ask two research question like: "Is there a linear relationship between y and x1,x2,x3, and x4 in the sample? What about in the <u>INSERT POPULATION HERE</u>?"

1. State your research question you will answer with your analysis.

2. Use at least one linear regression to answer this research question. Make sure you do the following as well.
    a. Show the summary output for your linear regression.
    b. Write out the linear regression equation for your model. Use appropriate notation.
    c. Check the linear regression conditions. If they are not met, try transforming one of the variables (maybe with a ln()) and see if that helps meet the conditions. If you have multicollinear explanatory variables, try dropping one.
    d. Discuss what percent of variability of your response variable is explained by the model. Is this high? Is this low?
    e. Make at least one prediction with your model.
    f. Which slopes in your model do we have sufficient evidence to suggest are non-zero in the population model? Explain your answer.
3. Finally, discuss how what you did your linear regression analysis here helps answer your research question.

5. Logistic Regression Research Question Set [8 points]

For your fourth research question set, you should pick (or make) a categorical response variable with two levels and at least 4 explanatory variables that you suspect might affect your response variable and then explore whether there is a linear relationship between these explanatory variables and the log-odds of the success level of the response variable in the dataset as well as the population.

For instance, in this section you should ask research question set like the following. "Is there a linear relationship between the log-odds of the success level of y and x1,x2,x3, and x4 in the sample? What about the INSERT POPULATION HERE? What explanatory variables should we include in the model to build a parsimonious model?""

Hint: You can create a 0/1 categorical variable from a numerical variable in a given dataframe df by using/modifying the code below.

```
df['new_cat_var'] = 1*(df['num_var']>=some_number)
```

1. State your research question you will answer with your analysis.
2. Use at least one logistic regression to answer this research question. When fitting your logistic regression model, you should do the following.
    a. Split your dataset into a training dataset and test dataset.
    b. Starting with these 4+ explanatory variables and using your **training dataset**, perform a backwards elimination algorithm (using AIC or BIC) to help you find a parsimonious logistic regression model. (We will discuss this next week).
    c. Then fit your **final** parsimonious logistic regression model with just your **training dataset.**
    d. Show the summary output for your **final** logistic regression.
    e. Write out the logistic regression equation for your **final** model.
    f. Which slopes in your **final** model do we have sufficient evidence to suggest are non-zero in the population model? Explain your answer.

g. Use your logistic regression model to calculate the ROC and AUC of your **test dataset.**
h. Use your ROC to pick a good predictive probability threshold. Explain why this is a good predictive probability threshold, *given your research goals.*
i. Then use this predictive probability threshold to classify your **test dataset**. What is the false positive rate and the true positive rate of your classification of the test dataset?

3. Finally, discuss how what you did your logistic regression analysis here helps answer your research question.

6. Conclusion [3.5 points]

1. **Summarization:** You should summarize the findings of each of your individual research questions here in your conclusion. (At least a paragraph here).
2. **Future Work:** Finally, if you (or someone else) were to conduct future work based on these analyses, what kind of research questions or analyses might that entail?

STAT430 Project Presentation Rubric

Grade: / 12

Team Members: _____

SLIDES / 8

Content (3) – You should present *some* content on each of these topics

- (0.5) Introduction
- (0.5) Research Question 1 and Analysis that Answers it
- (0.5) Research Question 2 and Analysis that Answers it
- (0.5) Research Question 3 and Analysis that Answers it
- (0.5) Research Question 4 and Analysis that Answers it
- (0.5) Conclusion

Correctness (2.5)

- Analyses are appropriate for the data, results are interpreted correctly.

Layout (2.5)

- Content is well organized, fonts are easy to read.
- Slides are engaging and not too wordy.

PRESENTATION / 4

Narrative / Motivation (3)

- Clearly explain motivation for the analysis.
- Clearly stated research questions in the beginning and how these questions relate back to the motivation.
- Clearly summarize answers to research questions that were discovered from the analyses.

Presentation (1)

- All team members speak and present some portion of the material.
- Team members speak loud enough for everyone to hear
- Team members understand the material, they are not reading directly from a notecard or script.