STAT207 – Data Science Exploration - Project – 75 Points

Due: Tuesday, December 6 11:59pm CST on 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 should work in groups of 3 people.**

- **In your group of 3, you must do at least 25% 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, *without looking at the code!*
        - 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.
- **Professionalism**
    - Your report and findings should be well-explained and written in **paragraphs** and **complete sentences** and in the **markdown cells (not in code blocks or in comments).**
    - Do not just spit out code and expect your reader to automatically know:
        - Why you chose to use this code, what its purpose is, what you're doing in the code block, and what you want them to notice in the result.
        - Why the output of your code is important.
        - How your code answers any relevant questions.
    - Any paragraphs, sentences, and explanations that you write should be considered satisfactory to, say, your high school writing teacher.

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 [50 pt]

Deadline: Tuesday, December 6 by 1:30pm CST on Canvas.

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 [17 pt]

Presentation Date:

- During your final lab section time **Wednesday December 7 in-person.**
- If you foresee an issue in presenting on 12/7 let me know asap!

Format:

- **Your presentation should be no more than 10 minutes.**
- You must present some part of the presentation 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.

Peer Evaluation [8 pts]

- **Deadline**: **Friday, December 9 11:59pm CST** on **Canvas**.
- **Purpose**:
  - For presenters/report writers:
    - The purpose of this final part of the project **for the presenters** and **report writers** is to give constructive feedback on:
      - how **clearly** you were able to communicate and answer your research questions with your analyses
      - how well you were able to **motivate** your research to a peer, and
      - how **reproducible** your analysis was.
  - For listeners/readers:

        - The purpose of this final part of the project **for the presentation listeners** and **report readers** is to:
            - **gain practice** being able to extract the most important parts of an oral research presentation or report and
            - **get ideas** as to how to make your own presentation/report delivery better.
- **<u>Steps</u>**:
    - On the day of your presentation, you (as an individual) will be randomly assigned to:
        - Watch another group's presentation
        - Read another group's report.
    - After watching the group's presentation and reading their report you will fill out a survey form on **Canvas**, which will ask you the following questions (see last page of this document).
    - 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.
- **<u>Graded</u>**:
    - For completeness

Dataset Options

You can choose your own dataset or you can choose from one of the four supplied datasets discussed in the next page. 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!

Less-Adventurous Options

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

More-Adventurous Option

4. **Download your Own Data from the Spotify API**
   a. The attached Downloading Your Own Spotify Datasets.ipynb file guides you through how to download your own data from the **Spotify API**. Specifically, it shows you how to do the following.
      i. Download information about an **album** listed on Spotify.
      ii. Download information about an **artist** listed on Spotify.
      iii. Download audio information about a **song** listed on Spotify.
      iv. Download a **"pseudo-random" sample of songs** from a given **genre** and **release time period** listed on Spotify.
   b. You can download and analyze any (or multiple) datasets of songs that you would like in this analysis, BUT **any part of your analysis that involves *inference* should be conducted on a (pseudo-)random sample of a larger population.**

Project Report Specifications

Your report should include the analyses, code, and explanations detailed in each of the following sections. ***(There's a rubric below this part with more detailed point breakdown).***

1. Introduction [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. *(At least a paragraph here).*

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. Explain how the dataset was collected and briefly discuss what the variables are/represent.

2. Descriptive Analytics Research Question Set [8 points]

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

You should ask at least one question of each the following two types.

- *"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. In the beginning of this section, clearly state your research question you will answer with your analysis in this section. Remember, descriptive analytics only involves describing relationships in the dataset that you have, so your research question should be *just* about the dataset.
2. Explain why/how you chose these 3+ variables that you are exploring here.
3. Use all appropriate visualizations needed to **thoroughly** answer this question. **You should have at least one visualization that incorporates 3 or more variables in the same plot.**
4. Use all summary statistics needed to **thoroughly** answer this question as well.
5. Finally, in the conclusion part of this section clearly describe what you see in your visualizations and summary statistics, what they tell you, and how they help answer your research question. When describing associations, you should be **thorough.**

i. When discussing a relationship between two numerical variables, you should discuss:
    1. Strength (correlation)
    2. Shape
    3. Direction
    4. Outliers
    5. Slope and intercept of best fit line.

ii. If you are comparing the relationship between two numerical variables (for various different sets of data), then you should be prepared to *compare* the strength (correlation), shape, direction, outliers, and slope/intercept of the best fit line.

iii. When describing a numerical variable distribution, you should discuss:
    1. Shape
    2. Appropriate measure of center
    3. Appropriate measure of spread
    4. Outliers

iv. When comparing numerical variable distributions (for various different sets of data), then you should be prepared to *compare* the shapes, appropriate measures of centers, appropriate measures of spread, and outliers for the different distributions.

3. Inference Research Question Set [7 points]

For your second research question set, you should pick two variables (one of them categorical with two levels) and explore the relationship between these variables *in a population* (that your dataset is a random sample of).

You should ask at least one question of each the following type.

- *"Is there sufficient evidence to suggest that there is an association between x and y in INSERT LARGER POPULATION HERE?"*

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

Hypothesis	***Can you help you answer the question***
$H_A: \mu_1 - \mu_2 \neq 0$ *μ_1 =average response variable value for level 1 of categorical variable* *μ_2 =average response variable value for level 2 of categorical variable*	*Is there sufficient evidence to suggest that there is an association between the numerical variable and categorical variable 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)*	*Is there sufficient evidence to suggest that there is an association between the two categorical variables in the population?*

1. In the beginning of this section, clearly state your research question(s) that you will answer with your analysis. Remember, inferential statistics involves answering research questions about populations given a random sample from that population. So your research question should be about a larger population your dataset was randomly sampled from.
2. Explain why/how you chose your response variable(s) and your explanatory variable(s) that you are exploring in this question.
3. 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(s).
    c. Calculate a p-value (or confidence interval) for this hypothesis test and use it to state your conclusion.
4. Finally, in the conclusion part of this section clearly discuss how your hypothesis test conclusion answers your research question.

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)
```

4. Linear Regression Research Question Set [10 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 in the dataset as well as the population.

You should ask at least one question of each the following two types.

- *"Is there a linear relationship between y and x1,x2,x3, and x4 in the sample?"*
- *"Is there sufficient evidence to suggest that there is a linear relationship between y and x1,x2,x3, and x4 in INSERT POPULATION HERE?"*

1. In the beginning of this section, clearly state your research questions you will answer with your analysis. Explain why/how you chose these 4+ explanatory variables to use as well as your numerical response variable.
2. Use at least one linear regression to answer your research questions. 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. Use two methods that we discussed in this class to assess the fit of the model.
    e. Which slopes in your model do we have sufficient evidence to suggest are non-zero in the population model? Explain your answer.

3. Finally, in the conclusion part of this section use your linear regression analysis to formally answer your research questions.
    a. *"Is there a linear relationship between y and x1,x2,x3, and x4 in the sample?"*
        i. *Consult (2.d) results*
    b. *"Is there sufficient evidence to suggest that there is a linear relationship between y and x1,x2,x3, and x4 in INSERT POPULATION HERE?"*
        i. *Consult (2.c and 2.e) results*

5. Logistic Regression Research Question Set [12 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.

You should ask at least one question of each the following two types.

- *"What explanatory variables should we include in the model that predicts this response variable to build a parsimonious model?""*
- ***[For a given logistic regression model that you have selected]:*** *"Is there a linear relationship between the log-odds of the success level of y and x1,x2,x3, and x4 in the INSERT POPULATION HERE?*

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. In the beginning of this section, clearly state your research question you will answer with your analysis and why you chose to explore these 4+ explanatory variables and the response variable that you selected.
    a. *If you chose to create a categorical response variable from a numerical variable, state why you chose to use the threshold that you did.*
2. Use at least one logistic regression model to answer: *"What explanatory variables should we include in the model that predicts this response variable to build a parsimonious model?""* When answering this question, 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 (or forward selection 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. Use your logistic regression model to calculate the ROC and AUC of your **test dataset.** What does this ROC and AUC tell you about your model's ability to make predictions on new data?
g. Use your ROC to pick a good predictive probability threshold. Explain why this is a good predictive probability threshold, *given your research goals.*
h. 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. Use at least one logistic regression model to answer: *"Is there a linear relationship between the log-odds of the success level of y and x1,x2,x3, and x4 in the INSERT POPULATION HERE?* (It doesn't necessarily have to be the final logistic regression model that you chose in (2) above. If it's different, explain how/why you chose these different explanatory variables to use). When answering this question, you should do the following.
    a. Explain which slopes in your **final** model you have sufficient evidence to suggest are non-zero in the population model. Explain your answer.
4. Finally, in the conclusion part of this section, clearly state the answer to your 2+ research questions based on your analyses.

6. Conclusion [4 points]

1. **Summarization:** You should summarize the findings of each of your individual research questions here in your conclusion and discuss how your findings relate back to your overarching motivation. (At least a paragraph here).
2. **Future Work:**
    a. 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?
    b. Also, discuss any shortcomings that you might be aware of when it comes to your analysis. *(Almost analyses have some inevitable shortcoming*). Is there any part of your analysis/data/decisions that may impact the confidence that you have in the answers to your research questions? *(No analyses are 100% perfect!)*

7. Overall Report Professionalism [4 points]

- Report is organized
- Report has a title and appropriate section headings
- Report is written in complete sentences
- Explanations are in markdown boxes (not code/comment).
- Findings are well-explained.
- Report is not just code spit out on a page, that makes the reader do all the work in terms of figuring out what you're trying to say/do in the analysis.

STAT207 Project Report Rubric

Section	Total Section Points	Evaluation Criteria	Subpoints
1. Introduction	5	Thorough and Clear Motivation	3
		State the research questions and how they plan to answer them	1
		Display dataset and discuss how it was collected and what the variable names mean	1
2. Descriptive Analytics	8	* Clearly states the research questions that they intend to answer in the beginning of the section. * Questions are well-formulated and something that they can answer.	1
		* Gives all visualization(s) that will thoroughly answer the question. * Thoroughly explains what is shown in the visualizations(s). * Interpretations are correct and the visualizations that they use are sensible given the question/data.	3
		* Gives all summary statistics that will thoroughly answer the question. * Thoroughly explains what the summary statistics imply. * Interpretations are correct and the summary statistics that they use are sensible given the question/data.	2.5
		* Summarizes how their visualizations and summary statistics answer the research question. * Explicitly answers their research question stated at the beginning of the section.	1.5
3. Inference	7	* Clearly states the research questions that they intend to answer in the beginning of the section. * Questions are well-formulated and something that they can answer. * Explains why they chose the explanatory and response variables that they did.	1
		* Clearly states hypothesis tests and they are correct. * Hypothesis test is suitable to use to answer the question.	1
		* Correctly checks and interprets conditions	1.5
		* Correctly calculates p-value or confidence interval	1.5
		* Makes a conclusion based on the p-value/confidence interval, and it is correct, thorough, uses correct language.	1

		* Uses their conclusion to clearly answer the question stated at the beginning of the section.	1
4. Linear Regression	10	* Clearly states the research questions that they intend to answer in the beginning of the section. * Questions are well-formulated and something that they can answer. * Explains why they chose the explanatory variables and response variable that they did.	1
		* Summary output table shown	0.5
		* Linear regression equation given * equation in the right format	1.5
		* Correctly checks and interprets conditions	2
		* Correctly assesses the fit of the linear regression model using two methods	2
		* Correctly conducts hypothesis tests on ALL population slopes in the model. Correctly explains the conclusions of each of these hypothesis tests.	1.5
		* Uses their analyses in this section to clearly answer the questions stated at the beginning of the section.	1.5
5. Logistic Regression	12	* Clearly states the research questions that they intend to answer in the beginning of the section. * Questions are well-formulated and something that they can answer. * Explains why they chose the explanatory variables and response variable that they did.	1
		* Correct train test split.	1
		* Correct use of backwards elimination.	2
		* Shows the final model, fitted with just the training data.	0.5
		* Correctly displays the logistic regression model equation for the final model.	1
		* Correctly calculates ROC and AUC with test data. * Correctly interprets the ROC and AUC with respect to model performance.	1.5
		* Picks a threshold * Explains why this is a good threshold based on research goals.	1.5
		* Classifies test data with model/threshold. * Calculates fpr and tpr of test data.	1

		* Correctly conducts hypothesis tests on ALL population slopes in the model. Correctly explains the conclusions of each of these hypothesis tests.	1.5
		* Uses their analyses in this section to clearly answer the questions stated at the beginning of the section.	1
6. Conclusion	4	* Clear summarization of findings and answers the research questions. * Discusses how their findings relate back to their overall motivation discussed in their introduction.	2
		* Discussion of future work * Discussion of shortcomings	2
Professionalism of Report	4	* Report is organized * Report has a title and appropriate section headings * Report is written in complete sentences * Explanations are in markdown boxes (not code/comment) * Findings are well-explained. * Report is not just code spit out on on a page, that makes the reader do all the work in terms of figuring out what you're trying to say/do in the analysis.	4
Total	**50**	I usually like to give lots of additional comments (not graded) that can help hone your future research report writing skills.	50

STAT207 Project Presentation Rubric

Grade: / 17

Team Members: _____

SLIDES / 11

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 (3)

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

Layout (5)

- ☐ Content is well organized, fonts are easy to read.
- ☐ Slides are engaging and not too wordy.
- ☐ No code shown on the slides
- ☐ No irrelevant code output is shown on the slides.

PRESENTATION / 6

Narrative / Motivation (4)

- ☐ 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 (2)

- ☐ 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.

STAT430 Peer Evaluation Questions [8 points]

Deadline: **Friday, December 9 11:59pm CST** on **Canvas**.

On the day of the project presentations, you will be randomly assigned to watch another groups presentation, and you will be randomly assigned to read another groups report. These two groups will be able to see your feedback. There will be a form on **Canvas** for you to give your answers to the questions below.

Presentation Peer Evaluation Questions [4 points]:

1. What is the **motivation** for the analyses in this presentation? Or 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 was 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 an interesting question for future work. *For instance:*
    - *"How did you make <this decision here> in your analysis?"*
    - *"I noticed <this particular thing> in one of your visualizations, it'd be interesting to try out <this particular analysis> in future work, because <reasoning/curiosity here>."*

Presentation Peer Evaluation Questions [4 points]:

1. What is the **motivation** for the analyses in this presentation? Or in other words, why should you (or someone else) care about the analysis that you just read/listened to?

2. How easy would it be to **reproduce** this group's entire analysis on your own in Python *without looking at their code*? If it was not extremely easy (or straightforward), what was not straightforward about it?

3. Were there any analysis decisions made in this report, in which you were unsure of **why they chose to do this** (ie. they didn't explain)? What were they?

4. Did the analyses and conclusions **answer the research questions** that was stated at the beginning of the report/presentation? If so, how? What were the answers to these research questions?

5. How would the results/answers to these research questions be **useful** to someone?

6. Are there any **shortcomings** that you can think of (the report should have mentioned some) in which their analyses may not have provided *perfect* answers to their research questions.