**STAT 3220 – Summer 2025**

**Final Project (25% of grade)**

# Purpose

Often times in “the real world” when you are given a data set or are trying to answer a research question, you may not be equipped with the analysis skills to answer the question. You will have to research and learn a technique on your own. Additionally, you will need to communicate the understanding of the new technique and your analyses to any audience. Throughout the semester, you have had opportunities to develop the skills to be able to do just that. The purpose of the final project is to demonstrate your ability to learn a new analysis technique and present an analysis both verbally and in writing. Academic poster sessions are common-place events at most research fairs and conferences. So developing a poster, and presenting at these sessions are useful skills. To demonstrate your ability to communicate your findings orally and visually, you are to present your research and findings at our STAT 3220 presentation.

# Project Scope

You will work in a group of 3-4 students to research a question of interest to you and your group. You will be responsible for finding your own data set, establishing a research question, applying the techniques we’ve covered, while extending beyond the course, and presenting your findings. It is your responsibility to decide upon your group dynamic. You will perform a multiple linear regression analysis on data you compile. The following criteria are required for whichever option you choose.

You will include the following components in your poster presentation and analysis:

* **Introduction**: Brief summary of the research question you intend to answer. This section should include motivation behind your research question (why it is interesting and relevant) and what your question is. You will want to follow the structure of one general question and 2-3 specific research questions questions. You will include at least three relevant sources to support your claims and topic of interest.
  + Consider a topic which you find interesting and formulate 3 specific research questions about that topic. You will choose one response variable and select three variables of interest to formulate your specific hypothesized questions. At least one question must relate a quantitative variable to the response and at least one question must related a qualitative variable to the response.
  + **Some examples of specific vs non-specific questions:**
    - *Non-Specific A: Does the size of a home impact the selling price ?*
    - *Specific A: Do homes with a larger square footage in Charlottesville have a higher selling price on average?*
    - *Non-Specific B: Does political affiliation impact voter turnout?*
    - *Specific B: Do states with a democratic governor have a higher voter turnout than states with a republican governor?*
* **Data Summary:** This section will include the relevant information about your data. And the important exploratory data analysis. This may include aspects of part 1, if relevant.
  + **Data Sources**: Describe how the data were collected for the primary sources
    - You should describe the general metadata of each data set compiled.
      * Who, what, when, where, why was it collected?
    - Describe any manipulations you made to the data- variable transformations or alterations of categorical levels. Any specific subsetting used and why
    - Be sure to include the specifics about the observations/subjects in your sample
    - Potential issues you have found or are worried about in the data
    - Considerations of the reasonably trustworthiness of sources.
    - You should not use any “code language” to describe things like data scraping, merging, sorting, filtering data.
  + **Exploratory Data Analysis:** You will select graphical summaries that best show the key features of your data. Some specific guidelines:
    - You must include a **histogram of your response variable**
    - You must include **at least four graphical representations** and the report should address major features of the data.
      * You must include at least one summary that relate to each of your research questions in some way.
    - You have an option to include an additional two graphical summaries that you thought were interesting in your process.
    - In total, **you should not have more than 7 graphical summaries**. (Histogram + 6 others)
  + **EDA Summary**: First, in this section you will **interpret** the graphical summaries in writing. In your interpretations, you should include a numerical summary that corresponds with the graphical summary. Additionally, you should include context. This includes a description as to why your response variable is suitable for regression. If it is not suitable, you should discuss what you will need to do to make it suitable. This should include assessing multicollinearty and any variable screening, if you decide to include that.
* **Analysis- see next section for specifics:** You will perform a multiple linear regression for your data. You should summarize your model building and highlight important findings about your variables of interest. You should include relevant plots and limited output. Including the following components:
  + Include required analysis steps
  + Include the ”added technique” that you selected
  + Assessing the model.
  + Selecting a final ”best” model.
* **Conclusion:** Your conclusion should consist of 2 parts: 1) interpreting your results of the analyses in context of the problem and 2) commenting on areas of future improvements. Your interpretation would include the actual prediction equation, comments on its usefulness for prediction, and an example of actually using the model for prediction or estimation. The second part you would comment on a limitations of your data and how you would continue or improve on this research in the future.
* **Works Cited:** You should include any references you used in your project. This can be submitted directly on your poster, or an added pdf appendix submission to your presentation page.
* **Formatting:** Your poster should include summarized components from each of the following sections. Remember, there should not be paragraphs, instead using bulleted lists, tables, and relevant graphs to display and communicate your findings.

**Your written report will include the following:**

* The submitted report should be no more than 8-pages and may include additional appendices (see below). You should follow the project 2 Rmarkdown template. You should outline the following sections with headings.
  + **Introduction**: The research question(s) & Background/significance of the research.
  + **Data Summary**: This include the data source description, EDA plots, and written description of EDA
  + **Methods/Analysis**: Describe the analysis process step by step. This would include specific techniques and appropriate testing languages (appropriate order of analysis, assumptions, etc). You will not interpret your analysis here.
  + **Results**: The statistical interpretation of the final model. This should be in statistical terms and overall interpreting and assessing the statistical usefulness of the model. There should be no R output (that will go in the appendix). However, you will include your final model.
  + **Conclusions:** Your conclusion should consist of 2 parts: 1) interpreting your results of the analyses in context of the problem and 2) commenting on areas of future improvements. Your interpretation would include the actual prediction equation, comments on its usefulness for prediction, and an example of actually using the model for prediction or estimation. The second part you would comment on a limitations of this analysis and your data and how you would continue or improve on this research in the future.
* In addition to the 8-page paper, will have a **title page** which includes:
  + The title of the project and an image that best represents your project
* Appendices (do not count to page limit)
  + **Appendix A:** **Data Dictionary/ Code Book**: A description of each of the variables in your data (in table format). (does not count towards page limit)
    - Name of each variable including your response, and how you will refer to it in your report
    - Written description of each variable (imagine someone knows nothing of your data). Include the units of quantitative variables, levels of qualitative variables, if relevant
  + **Appendix B: Selection of Data rows** (does not count towards page limit)
    - You should use the **head()** function to output the first several rows of your data
  + **Appendix C: Final model output and plots.** This should only be a summary of your final model
  + **Appendix D: References**: APA guidelines (On its own page, does not count towards page limit)
    - Research background citations (min of 3)
    - Data source Citations (min of 2)
    - Additional Technique/Coding Help/Other (as many as needed)
* **Formatting** of the paper should include the following:
  + **Title page and appendices do not count towards page limit. All other sections do count.**
  + You will use the markdown template provided
  + Appropriate section headings (as above)
  + The introduction, data summary, and conclusions should be written in cohesive paragraphs with narration, not bulleted items.
  + The size of printed graphics may be modified if they remain readable.
  + The reference list and data appendix should begin on their own pages

# Additional Information on Data Compilation:

Find and compile data related to your topic that will reasonably answer your questions. Ultimately, you will have one final data table, with rows as your observations and columns as your variables. You are expected to minimally collect **8 explanatory variables** related to your topic of interest. Three of these variables are linked directly to your research questions and the remaining variables are other factors that you think will be important. You will have at least 3 qualitative and at least 3 quantitative predictors. You will have a minimum of 50 observations. Additionally, you will have a continuous response variable appropriate for MLR. Your data should be current in the last 5 years. You will have cross-sectional data, meaning you will NOT have time (year, moth, etc) as a factor. **You will compile your data from at least 1 primary source.** You may use Kaggle or another repository as a second source. However, you must have verification of the original data collection.

Some links to get started: <https://guides.lib.virginia.edu/stat3220>

* You should have approximately 50-100 observations (rows) in your data. If you have much more than 1000, you might consider taking a thoughtful sample
* You want to clearly define your population this will help guide your data collection, sampling, and/or inference
* You should have a clearly defined response variable. This should be continuous and quantitative. Your response should not be discrete with a small range of values (i.e. Likert scale, or mobile app prices)
* You should have a combination of quantitative and qualitative explanatory/predictor variables. Minimally 3 of each.
  + Your qualitative variables should not have more than 5 levels.
* Some data sets that are NOT suitable for this project:
  + World Happiness Data
  + Spotify Data
  + Airbnb New York
  + Data from packages in R
* You should use only publicly available data, or data which could be obtained in a straightforward manner. If you have questions about this, please contact me or Data Librarians, Jenn Huck, or Erich Purpur at [data@virginia.edu](mailto:data@virginia.edu)
* Finding Data + Citation Info
  + <https://guides.lib.virginia.edu/stat3220>
  + from Data to Viz - <https://www.data-to-viz.com/#sankey>
* [RMarkdown CheatSheet](https://rstudio.github.io/cheatsheets/html/rmarkdown.html?_gl=1*1g7ri4n*_ga*NjU4NDgwNTk1LjE2OTQwMzQ5OTc.*_ga_2C0WZ1JHG0*MTY5NDAzNDk5Ni4xLjAuMTY5NDAzNDk5Ni4wLjAuMA)

Analysis Choices + Requirements: You will select multiple linear regression or logistic regression as your analysis technique.

## Multiple Linear Regression Analysis Requirement

* Model building with significance testing (should be supported by EDA and/or variable screening)
* Identify and check for multicollinearity
* Residual analysis (assumptions + extreme observations) Make necessary adjustments as you see fit. If you are attempting to correct a violation, you only need to try up to three corrections, if the first doesn’t work. Include plots or output as needed.
* Final model selected should be assessed. It may not be great, but you will explain that in conclusion.
* Include at least ONE additional techniques from the list to add to your analysis: Weighted least squares (Ch 9.4), External Model validation (Ch 5.11), box -cox transformation (supplemental), or another technique or aspect of MLR you learned on your own.
  + **Weighted Least Squares (Ch 9.4).** In regular MLR each observation is equally weighted when building the model. Weighted least squares will “weight” some observations higher than others. This may be useful when you have data dealing with states or countries. (EX: is California “equal” to Rhode Island). This can be especially helpful when the observations are big outliers.
  + **External Model validation (Ch 5.11).** The most common way you will do this is through data-splitting. Data-splitting should be used when you have a larger data set (>100). Evaluating Jackknife MSE or R^2 or PRESS can be used for smaller data sets. Just evaluating parameter estimates and prediction intervals is not sufficient for this criteria.
  + **Box-Cox Transformation (supplemental).** This is used when you need to transform the response variable, but you are not certain what transformation to do. This procedure will give you the ideal power transformation, so that your data will meet the assumptions.
  + **Ridge or LASSO**. Is used when there are a lot of variables and severe multicollinearity.
  + **Poisson regression**. This is used when you have “count” data, particularly with a very unlikely response. Meaning you may have a lot of 0s.
* **Does our final model need to be “good”?** No. The purpose of this assignment is that you have completed a thorough analysis based on what we have covered in our course. So long as you have attempted though several iterations to improve your model, that is ok. You can explain in your conclusion why your model is not good.
* **What order should everything go in our analysis?** On the 7/3 project workday, we reviewed the MLR concept map. Refer to that as your guide for model building. Remember some steps happen simultaneously as others, and some steps need to be repeated.

# Logistics

## Due Saturday, July 12 at 10:30am (as a group, for all sections)

* **Copy of presentation:** Email Professor Varanyak a copy of your presentation- regardless of which time you are presenting. One per group copy all members on email

## Presenting January 12 during class time:

* The purpose of the presentation session is to present your results and findings of your research question, with less emphasis on “teaching” the analysis. You should be prepared to answer questions about the technique, but you should focus your presentation on answering your research question. Your audience is other students in this class. You should prepare a **10-12 minute presentation**.
* You can create whatever medium you’d like for a presentation- slides, poster, prezei, etc but it should be visually appealing and engaging.

## Due January 12 at 5:00pm

* **Executive Summary Paper:** Submit 8 page written report to Gradescope. One per group.

## Scoring: 350 Points

|  |  |  |  |
| --- | --- | --- | --- |
| **What?** | **How?** | **Who?** | **How much?** |
| **Presentation + Paper Includes Appropriate Summaries of Components** | Highlighted findings from Background, Data Summary, Analysis, and Conclusion all written coherently. The focus is on the research question, supported by the analysis. You do not need to include every single component of your analysis on the poster. | Scored Prof V  100 Poster  100 Paper | **200 points** |
| **Presentation:**  **Visual Aesthetic/ Formatting of Poster**  **Presentation** | There is not too much text or bulleted lists. Relevant graphs and related images are included. Overall, the poster is professional.  You demonstrate that you are knowledgeable about your topic and research question. Your presenting is clear and energetic | Average rating from your reviewers and me  \*If you are not present on your day to presentation, you will earn a 0 for this portion. | **15 points**  **35 points**  **See** [**Peer Review Rubric**](#_Peer_Review_Rubric:) |
| **Attending your presentation day** | You are able to answer questions by your reviewers | **You**  \*If you are not present on your day to presentation, you will earn a 0 for this portion. | **20 Points** |
| **Evaluation of other groups** | You will submit a formal evaluation of 2 other groups during the poster session | You submit thoughtful, appropriate feedback | **30 points** |

**50 Points:** Over the project work days, there will tasks to complete that will account for 50 points of your project Grade.

*Also be an end to term survey and group evaluation that may impact your individual score. Group concerns should be addressed ASAP.*

|  |  |  |  |
| --- | --- | --- | --- |
|  | Group 1 Eval | Group 2 Eval | Group 3 Eval |
| **Visual Aesthetic/ Formatting** |  |  |  |
| Appropriate amount of text and readable (Visual Aesthetic) |  |  |  |
| Graphics and Visuals **added** to the presentation and were appropriate (Visual Aesthetic) |  |  |  |
| The presentation is visually appealing (Visual Aesthetic) |  |  |  |
|  |  |  |  |
| **Presentation** |  |  |  |
| It was clear that the entire group understood the techniques used (Presentation) |  |  |  |
| It was clear that the group understood the content of the research topic (Presentation) |  |  |  |
| The group investigated an interesting research question(s) (Presentation) |  |  |  |
| The group used the presentation as an effective aid in presenting their research (Presentation) |  |  |  |
| The organization of the presentation was logical, and the material covered was complete (Presentation) |  |  |  |
| The group adequately answered your questions. (Presentation) |  |  |  |
| The group had effective presenting skills (energetic, spoke loudly and clearly) (Presentation) |  |  |  |
| Comments: |  |  |  |

## Peer Review Rubric

This will be submitted through a MS Form.

Rate each group on the following:

1 (Strongly Disagree) 2 (Disagree) 3 (Neutral) 4 (Agree) 5 (Strongly Agree)