STAT 311 – Final Project

The final project for this course involves performing a data analysis project focused on a regression problem using a real dataset.

Project Objectives

The objective of this project is to provide you and your partner(s) with an opportunity to explore the challenges of data analysis, including missing data, measurement accuracy, reporting, and the depth of analysis. You will apply the techniques covered in STAT 311 to analyze relationships among variables using statistical procedures you have learned. Selecting the appropriate procedure is part of the learning process. This project will strengthen your skills in building statistical models and effectively presenting your findings both in writing and through a formal presentation. Clearly communicating statistical interpretations is a crucial aspect of this course.

Project Requirements

1. Choose Your Data

You must find a dataset for the project, either from a publicly available online data repository or another suitable source.

- The dataset should have a relatively large number of observations and at least six predictors (excluding polynomial and interaction terms).
- If you are unsure of a dataset's suitability, consult with me.
- Avoid datasets with time or spatial coordinates as predictors, as they typically do not satisfy the model assumptions discussed in class.

Suggested Data Sources:

- StatCrunch Datasets
- Awesome Public Datasets (GitHub)
- Kaggle datasets
- Google dataset search.
- World Bank data
- Reddit
- https://github.com/caesar0301/awesome-public-datasets
- UCI Machine Learning Repository
- https://guides.emich.edu/data/free-data

2. Define Your Research Questions

Based on your chosen dataset, identify the specific questions your analysis will address.

3. Conduct Data Analysis

- Select the variables to include in your model based on your research questions.
- Explore distributions and summarize variables.
- Identify outliers or influential observations.
- Model Building
- Consider higher-order models.
- Model Selection
- Assessing statistical adequacy
- Considering higher order terms.
- Model comparison
- Residual Analysis
- Model Validation

4. Written Report (Due: Monday, December 1st, Midnight)

The report should be 10-15 pages (including figures and tables) and prepared in **LaTeX or Word**.

Report Structure:

- **Front Cover:** Title, name(s)/group number, date, reference to "STAT 311 Regression Analysis, Fall 2025."
- **Introduction:**
  - Define the research question(s).
  - Explain the significance of the question(s).
  - Describe your approach to answering them.
- **Data Description:**
  - Describe the dataset, including its source, generation, and collection process.
  - Provide a narrative on the data-generating process.
- **Exploratory Data Analysis:**
  - Provide structured examples of the data.
  - Utilize graphs, tables, and summary statistics.
- **Methods:**
  - Describe the study design.
  - Detail the statistical methods used, including model selection and validation.
- **Results:**
  - Report findings, including graphical and tabular representations.
  - Discuss relationships among variables.
  - Justify why your final model is appropriate.
- **Conclusions and Discussion:**
  - Summarize findings in a non-technical manner.
  - Limit this section to **one page** with no figures.
- **Appendix:**
  - Include supplementary JMP output.
- **References:**
  - Cite all sources appropriately.

5. Presentation (Due: Tuesday, December 2nd , 5:30 PM)

- The presentation should be **10 minutes**, followed by **3-5 minutes of questions**.
- Assume the audience consists of analysts with limited prior knowledge of your topic.
- Each presenter must:
  - Keep **the camera on and show your face** throughout.
  - Speak for a part of each section: introduction, methods, results (JMP output), and conclusion.
  - **No prerecorded videos**.
- Slides must cover:
  - Research questions
  - Chosen variables
  - Methods used
  - Summary of results (tables and graphs)
  - Conclusions in context
- The presentation has to be conducted during the allocated time and are not allow to bring a recorded presentation to play during that time.
- All presenter are rquired to devide the slides so ech one of your explain a part of introduction, a part of methods, a part of summary of results using JMP output and a part of coclusion.

Project Submission

The final project submission must include **four files**, uploaded separately to **D2L (not in a zip folder):**

1. **Final Report:** PDF format, named **"GroupNumber_Report"**.
2. **JMP Analysis:** Saved data and outputs.
3. **Dataset:** If downloadable, include the file; otherwise, provide instructions for obtaining the original dataset.
4. **Final Presentation Slides:** Named **"GroupNumber_Presentation"**.

Additionally, upload the report and presentation to **GitHub** and share the repository link either in your final presentation slide or via the project dropbox in D2L. Instructions for creating a GitHub repository are available in the D2L project folder.

References

Ensure all sources are properly cited throughout your work.

Checkpoint 1 – Dataset & Research Question (due Nov 10)

1. What dataset have you chosen, and where did you find it?
2. How many variables and observations does it contain?
3. What is your main research question (dependent variable + predictors)?
4. Have you verified that your dataset meets regression assumptions (no time/spatial coordinates)?
5. How will you handle missing data or data cleaning?

Checkpoint 2 – Project Proposal (due Nov 15th before midnight)

1. What is your **project title**?
2. What is your **topic area** (e.g., health, economics, environment, marketing)?
3. What is your **dataset source** (e.g., Kaggle, StatCrunch, UCI, GitHub, World Bank)?
4. Provide a short **abstract (3–5 sentences)** summarizing your research idea.
5. State your **dependent variable** and at least **six predictors** you plan to explore.
6. What potential challenges do you foresee (e.g., missing data, measurement issues)?
7. Have you included a properly formatted **reference to "STAT 311 Regression Analysis, Fall 2025"**?
8. Did you attach or link the dataset file (if downloadable)?

Checkpoint 3 – Model Building & EDA (due November 27th before midnight)

1. What exploratory plots and summary statistics have you generated?
2. Have you identified any outliers or influential points?
3. What model(s) have you tried so far, and what selection criteria are you using (AIC, BIC, R^2, adj R^2)?
4. How are you validating the model (e.g., cross-validation, train/test split)?
5. Have you checked model assumptions (residual normality, multicollinearity, etc.)?

Checkpoint 4 – Report Draft & Presentation Planning (due November 29th before midnight)

1. Have you completed all major analysis sections (EDA, Model Building, Results)?
2. Have you written the draft of your Introduction and Methods sections?
3. Are your graphs and tables formatted clearly for the report and presentation?
4. Have you distributed presentation responsibilities among group members?
5. Have you tested your presentation to stay within the 10-minute limit?

Checkpoint 5 – Final Submission Review (before Dec 1–2)

1. Are all four required files named correctly and uploaded separately to D2L?
2. Is the GitHub repository link included in your slides or dropbox submission?
3. Have you verified your JMP output and dataset are readable and clean?
4. Does each group member know their presentation part and technical setup?
5. Have you rehearsed your presentation with camera on and screen sharing enabled?