

TEMPLATE for STAT-4000 DS Project: Standard Academic Structure Focus: Methodology, rigorous analysis, results interpretation. Good for projects emphasizing model building, comparison, or hypothesis testing.

This template is generally good but for reference only, you don't have to follow it strictly, we encourage you to modify the template, or even adopt a different structure based on your need.

[Your Project Title Here]

[Student Name 1, Student Name 2, ...]

[Course Name, e.g., STAT 4000]

Department of Mathematics and Statistics

Auburn University

[date]

1 Introduction

1.1 Background and Motivation

Briefly introduce the background or context of the dataset(s) used. e.g., What problem are you addressing? Why is this area interesting or important?

1.2 Problem Statement and Research Questions

Clearly state the specific question(s) you aim to answer/explore using the dataset(s). Be precise and clear. (e.g., “Can we predict wine quality based on physicochemical properties?”, “Are there distinct customer segments based on credit card usage?”)

1.3 Dataset Description

Describe the dataset(s) used. Source, number of observations, number of features. Mention key variables relevant to your research questions.

1.4 Report Outline

Briefly state what each subsequent section covers.

2 Methodology

2.1 Data Preprocessing

Describe all steps taken to clean and prepare the data. (e.g., Handling missing values, feature scaling/normalization, encoding categorical variables, feature engineering). Justify why these steps were necessary/appropriate for your data and chosen methods.

2.2 Exploratory Data Analysis (EDA)

Describe and summarize key insights from EDA (e.g., histogram, Density Plot, Box Plot, etc., detailed plots can go in Appendix or Results). Mention distributions, correlations, or initial patterns discovered relevant to your project goal.

2.3 Modeling Approaches / Algorithms Used

Describe the specific algorithms or statistical methods applied (e.g., Linear Regression, SVM, K-Means, PCA, Decision Trees, Neural Networks, etc.). Provide a brief theoretical explanation of how each chosen method works. Justify why each method is appropriate for your research question(s) and data type, and how you tune the parameter of the models.

2.4 Evaluation Metrics

State the metrics used to evaluate model performance or analysis quality. (e.g., Accuracy, Precision, Recall, F_1 -score, R-squared, Adjusted R-squared, MSE, Silhouette measure, etc.). Explain why these metrics are suitable for your problem.

2.5 Tools and Libraries

Mention the primary software (e.g., Python) and libraries (e.g., Pandas, Scikit-learn, Matplotlib, Seaborn, TensorFlow/PyTorch) used.

3 Results

Present the main findings of your analysis, and how your analysis addresses your research questions.

4 Discussion & Conclusion

Interpret the results presented in the previous section. What do they mean in the context of your problem statement? Discuss the significance and implications of your findings. Were there any surprising or unexpected results? Acknowledge the limitations of your study (e.g., data limitations, assumptions made, methods' weaknesses). Suggest potential areas for future work or improvements.

Summarize the key findings and reiterate the answers to your research questions. Provide a concise takeaway message from your project.

References

List any external resources, papers, or significant online sources consulted (provide links/references as required by project guidelines).

A Detailed Exploratory Data Analysis (if any)

Include additional plots and summaries from your EDA that support your preprocessing or analysis but weren't central to the main results section.

B Additional Results (e.g., Model Diagnostics)

Include supplementary results, diagnostic plots (e.g., residual plots), or details not essential for the main part.

C Code Snippets (Optional)

Include key code snippets if necessary. Full code should be submitted separately.