Midway Check in (Team) Event - Evaluation Criteria

**A brief discussion of the current project status:**

Our team has been actively preparing the data from the MrOS dataset. We have successfully moved from a divide and conquer strategy of each team member analyzing various forms (such as bone mineral density, lifestyle, and clinical notes) to this week combining those cleaned forms with our target variables from dataset titled FAFEB23. The new merged dataset located "/dsa/groups/casestudy2023su/team03/model\_data/mros\_merged.csv" has also been split into dependent and independent variable and then test and train for modelling to come.

**A brief description of planned future work including evolving data story:**

This current week will be focused on using what we learned in the EDA process to clean the datasets, merge them, and create a final data set for use in predictive modeling. We would also like to implement outlier detection and feature reduction, with the goal to begin model fitting in earnest by the end of the week or beginning of next week. Feature importance from tree-based models will be used to assess the importance of different features in predicting our target variable. Based on discussions with our mentor, we are working to actively reduce dimensionality in an appropriate way to maintain the ability to tell a strong descriptive data story.

**Critical assessment of team workflow and process:**

The team is working efficiently, with tasks clearly divided among team members and regular communication maintained via Slack. The team also makes sure to review and document each step of their work in Jupyter notebooks, ensuring transparency and reproducibility. All data files are available to each team member and the cleaned, merged dataset has been made available as a CSV for all teammates to perform feature reduction and modeling going forward.

**Summary of team member contributions and performance:**

Karen-Created correlation plots for MU, TU and NF forms in the V1 dataset. Concatenated dataset subset from MU, TU, and NF forms with FAFEB23 data subset in preparation for target analysis.

David- Updated cleaning being done on Medication Use (MU) and Nottingham Power Rig (NP) forms. Built on Tyler’s merging of the datasets to begin feature reduction analysis (highly correlated feature pairs and recursive feature elimination).

Josh- Updated EDA for bone mineral densities and GI and FF forms. Collaborated in team meetings. Worked on midway check-in document.

Tyler- annotated form review, export of clean datasets, merging of datasets, minor multivariate correlation EDA on target variables

**Discussion of any roadblocks or domain questions/problems or technical issues that need to be addressed:**

The team has done well addressing most of the technical issues we have been able to fill most of the NA's and the data set. We are now working to start modeling the data to predict our outcome variables. However, the outcome variables are counts of fractures, so to use a classification problem we would have to convert them to binary. We feel comfortable with the domain questions, but there are a lot of features in our dataset so far and we are in discussions on reducing dimensionality while retaining the ability to tell a strong descriptive data story.

**Link to Mentor recorded mentor meeting and key meeting takeaways provided:**

Link to recording: [6-30-23\_Mentor\_Meeting\_Team03.mp4](https://mailmissouri-my.sharepoint.com/:v:/r/personal/jwj8c8_umsystem_edu1/Documents/SU23_DSA8080%20Casestudy/Zoom%20Meetings/6-30-23_Mentor_Meeting_Team03.mp4?csf=1&web=1&e=fax21f)

Link to Meeting Notes: [6/30/2023 - Dr. Green](onenote:https://mailmissouri-my.sharepoint.com/personal/jwj8c8_umsystem_edu1/Documents/SU23_DSA8080%20Casestudy/SU23_DSA8080%20Casestudy/Mentor%20Meetings.one#6/30/2023%20-%20Dr.%20Green&section-id={3D12AE69-253A-46DF-A3B5-D58BA7327F52}&page-id={6700CC64-2323-419F-92A0-209DD6B4669E}&end) ([Web view](https://mailmissouri-my.sharepoint.com/personal/jwj8c8_umsystem_edu1/_layouts/OneNote.aspx?id=%2Fpersonal%2Fjwj8c8_umsystem_edu1%2FDocuments%2FSU23_DSA8080%20Casestudy%2FSU23_DSA8080%20Casestudy&wd=target%28Mentor%20Meetings.one%7C3D12AE69-253A-46DF-A3B5-D58BA7327F52%2F6%5C%2F30%5C%2F2023%20-%20Dr.%20Green%7C6700CC64-2323-419F-92A0-209DD6B4669E%2F%29))

* Dataset and Features:
  + The dataset contains over 160 features from survey data, with approximately 20 related to Bone Mineral Density (BMD) and 500 rows.
    - Feature reduction should be performed on the entire dataset, and then a comparison can be made with sparsely populated features.
    - Some features may be important for specific subsets of the population, suggesting the possibility of running analyses on subsets based on these features.
  + There are two variables in the outcome: traumatic fractures and non-traumatic fractures.
  + Null values in features can still be useful if a small percentage is strongly correlated.
  + Bone Mineral Density (BMD) is considered a crucial feature, as it is supported by literature and correlates with fracture risk.
* Audience Considerations:
  + The intended audience includes clinicians, patients, and caretakers.
  + Using a black-box model may present challenges in delivering the results to the audience.
  + The number of features affects the ability to make meaningful statements, considering the context of all features.
* Modeling Approach:
  + Full regression may not be feasible due to limited data.
  + The focus is on predicting one or more fractures.
  + Tree-based models like Random Forest, LightGBM, or XGBoost are suggested for feature importance analysis.
  + Logistic Regression is recommended to assess the change in probability associated with each feature.
* Fracture Distribution:
  + The count data for fractures is predominantly skewed, with 70% having no fractures, 20% having one fracture, and the rest having up to six fractures.
    - Since dichotomous independent variables, converting them into binary instead of ordinal may be appropriate.