**Data cleaning:**

* Remove columns that have 100% missing values -> absolutely useless
  + Remove the top 5 columns that have an extremely high volume of missing values
  + Visualize the presence of missing values in all columns
* Visualize and observe the presence of missing values across all columns
  + Impute data for all columns that are necessary but have a high volume of missing values.
    - Segregate categorical variables and numerical variables by identifying columns with the attribute of ‘object’ and columns with the attribute of ‘int64’, ‘float64’.
    - Carry out imputations
  + Before carrying out imputation, replace all ‘not reported’ entries in the age/sex categories as NaNs
* Finding a way to visualize time, by default the change in recovery is shown by different ‘redcap\_event\_name’ events into different rows of observations with identical record ID for each patients
  + Instead of showing changes by time through record id, visualize it using ‘redcap\_event\_name’. Like visit\_1\_arm\_1, visit\_2\_arm\_2, visit\_2\_arm\_3. And see how the distribution looks like for the recovery rate for each visit.

**Inspiration along the way:**

* Our current plan is to create a dynamic dashboard, in which the user (physician, surgeons, or doctors) could select a group of data based on demographic variables, and then enter their current patient’s stats to compare their patient’s status with the average/median stats of all the patients in that particular demographic group.
  + This looks like if the doctor conducted an, either self-reported evaluation (KOOS reporting) or any other form of strength evaluation, the doctor could then enter that patient’s evaluation scores into the dashboard and compare the patient’s evaluation outcome to the general recovery outcome of all other patients based on the selected demographic

**Exploratory Data Analysis:**

* Replace NaN values in age\_groups\_dashboard, sex\_dashboard, visit\_sex to Not Reported
* Visualize graft type distribution by Age groups (as hue)
* Visualize graft type distribution by Sex
* Visualize distribution of age group by sex
* Utilize tss\_dashboard (time since surgery) as our column for looking at changes in evaluation
* Visualize distribution of pain score based on time since surgery
* Visualize, in small multiples of histograms, the distribution of different strength tests and limb symmetry indexes based on tss.
  + Make the legend false for all small multiples except for the last histogram