Title

Section I: Dataset Description & Reading

- Dataset's info (about hospitalization etc)
- Dataset's shape
- Link to description of each variable + possible values
- Reading the SAS file and showing the head

Section II: Selected Subset & Target of Dataset and Justification

- Reasons we chose the specific subset we ended up with
- What was our target label and why?
- Describe any limitations we faced (e.g. departments)

Section III: Touch on Dataset Quality in General

- Comment on missing values
- Comment on columns that expressed really similar things
- What features should have been included (are missing in general)?
- Any additional comments on what is wrong with this dataset. [Possible feature engineering needed?]

Section IV: Dataset Exploration for Selected Subset

- The groupings of the dataframes and justification - general. Add the label to each dataframe. Next, for each dataframe...
- Before each dataframe, add the variable's description and possible values.
- Don't forget to fix the data types if needed.
- If there is time, try to clean the data a bit at least, and state what's missing to be done. Or at least list HOW the data needs to be cleaned (outliers, consistency issues, missing values that are complex to handle and how, etc).
- Plot the correlation heatmap for each of the variables against the label.
- Check if any variables are too dependent on each other. Remove those.
- Now we have our "filtered" data. Visualize if needed. [heatmap, regression lines, scatter plot, etc]
- Show Statistics .describe

Section V: Touch on Data Quality of the Selected Subset

- Comment on feature availability. What issues did you find with the data? [POSSIBLE FEATURE ENGINEERING NEEDED]? List any ideas that you believe would improve the dataset quality and issues. Maybe how to remedy them too. Or what extra steps would you do for a better analysis? Some ideas :)

Section VI: Potentially Important Features & Justification

- A final dataset with the filtered features from the earlier steps.
- Some visuals to reveal our insights on which data mattered the most. [heatmap, regression lines, scatter plot, etc]
- How we could have done our exploration better? What steps could we have followed or resources we could have reached for? Any way to do any of our sub-steps better?