1. **Research**
   1. Find paper(s) explaining
      1. Theory, and strengths and weaknesses of
         1. Logistic Regression
         2. Linear discriminant analysis
         3. Boosting? Maybe this link and/or any research? <https://towardsdatascience.com/what-is-boosting-in-machine-learning-2244aa196682/>
         4. Shap values? This makes nice visual that makes evaluation of variable importance easier
      2. Classification broadly
         1. Evaluation metrics… Accuracy, precision, recall, F1 score
2. **Data Exploration**
   1. High level summarizations.
      1. Frequencies
      2. Outliers
         1. Look at extremely high income… Do all of those people receive their loan? Is there a certain level of income where everyone gets approved? Or where loan pct income is below a certain level?
      3. Any anomalies we might need to control for (insanely high income)
   2. Data visualization
      1. Histogram
      2. Pairs plots
      3. Maybe overlay our outcome variable as a color? Maybe scatterplot of age/income with loan status as color?
3. **Data Cleaning**
   1. Based on the exploration step, should we:
      1. Do any kind of bucketing? For example, grouping age to <25, 25-35, etc.
      2. Control for outliers
      3. Create any new variables?
   2. Convert data types to types usable for modeling
      1. Adjusting to factors
      2. One-hot encoding (dummy variables) where necessary
   3. Standardize numeric variables (not strictly necessary for logistic regression, but will make our coefficients more interpretable)
4. **Training Models**
   1. **Lasso**
      1. Convert model
      2. Run lasso with Cross validation.
      3. Look at coefficients, what variables are being dropped
      4. Does this change based on any variables we are modifying from data cleaning step
      5. Make Lasso predictions
   2. **Logistic Regression**
      1. Split data into train/test data sets (80/20 split)
      2. Train model with all relevant variables
      3. Train model with only the variables selected by Lasso
         1. Some may need to be recoded with one-hot encoding rather than feeding through straight as a factor if we are only dropping one level
      4. Compare coefficients (or shap values) of different variables
      5. Make predictions with both types of models
   3. **Linear Discriminant Analysis**
      1. Same as above, potentially use lasso coefficients as well but not necessary
5. **Model Evaluation**
   1. Generate confusion matrix for each model
   2. Create table of accuracy measures, one row per model
   3. Generate shap values for each model, evaluate variable importance across models
6. **Paper**
   1. Set up and follow typical outline for research paper
   2. Select relevant visuals from code
   3. Probably should be a bit more introduction/research heavy than most papers, because the idea of our project is that we are introducing a new topic
7. **Presentation**
   1. Just pull relevant parts of paper to make a ppt for class on Monday