Team Approach

- Screen share approach for AWS/modeling.
- Divide and conquer the PowerPoint (Use Canva; export to PDF) and Documentation (We will tell Denis what is appropriate, per Denis on 4/7).

Day 1 - Friday

- Select a Kaggle dataset (At least 10 columns)
    - ~~Pokemon Image Classification (each class, 100 images)~~
    - ~~USDA Food (53, 8790)~~
    - Chat GPT Sentiment Analysis
        - ChatGPT sentiment analysis | Kaggle
        - `import torch`
`from **transformers** import AutoTokenizer, AutoModel`
        - **Bert** picks up the context – try w and w/o StopWords
        - Find a simple bag of words encoder
            - Look at **Word2Vec**
    - ~~Airbnb Regression (26 cols)~~
    - ~~Life Happiness Prediction (12 cols)~~
    - ~~Predict Promotion based on Factors (12 cols)~~
    - ~~Life Expectancy (22 cols)~~
    - ~~Nutrition Physical Acitvity (33 cols)~~
- Define the business problem to be solved.
    - **Understanding what sentiment, if any, surrounding Chatgpt tweets so we can anticipate potential pushback from clients during implementation phase.**
        - Building a model off Chatgpt and as new statements come in for Chatgpt, we can predict sentiment.
        - Booz Allen implications of determining sentiment before something goes out.
    - This is not a sentiment that we can use for Twitter, etc. because it's trained on Chatgpt.
    - Defense perspective – Embracing technology has been slow due to security/commercialization.
- Define the project goals and scope.
    - Hope to create a model that has over 80% accuracy of predicting sentiment of Chatgpt tweets.
    - ONLY specifically Chatgpt tweets. However, in the future we could tailor this specifically to Booz Allen needs.
    - 1 month of tweets from November 2022.
- Perform an EDA and create plots depicting important aspects of your dataset (Individual presentation)
    - Finished
- AutoML

Day 2 - Monday

- Create a project plan that includes the phases from exploratory data analysis to deploying the model.
    - Complete
- Create a presentation outline.
    - Complete
- Create an AWS S3 bucket and upload the raw dataset into it.
    - Note this should be an *organization* S3 bucket so we can all access.
    - Complete
- Create an AWS SageMaker notebook for EDA.
    - Complete
- Perform EDA and create plots depicting important aspects of your dataset.
    - Complete
- Additional Things to Add to Model:
    - WordCloud: Complete
    - Emojis: Complete
    - Capitalization – lowercase: Complete

Day 3 - Tuesday

- Select (or create) a baseline model.
    - **Model Iterations: (Baseline and Champion)**
        - 2,000 entries -> 61% accuracy, no cleaning [BASELINE]
        - 10,000 entries -> 81% accuracy, no cleaning
        - 20,000 entries -> 83% accuracy, no cleaning
        - 20,000 entries -> 84% accuracy, cleaning
        - 20,000 entries -> 84% accuracy, cleaning AND stopwords
        - 50,000 entries -> 88% accuracy, cleaning
        - 60,000 entries -> 88% accuracy, cleaning AND stopwords
        - Full Dataset -> 92% accuracy, cleaning AND NO stopwords
        - Full Dataset -> 91% accuracy, cleaning AND stopwords
        - AutoML H20 -> 44% accuracy
        - AutoPilot Sagemaker -> 87% accuracy
- Define model architectures to try or consider.
- Run first experiment and create a first model.
- **Internal Goals**
    - Put artifacts in S3 bucket and make new inferences off endpoint.
    - Start an AutoML model with Sagemaker Studio.
    - Begin Documentation

Day 4 - Wednesday

- Run the rest of the experiments.
- Version, score and evaluate the models created.

- Select the final model for production environment.
- Deploy the final model to production.
  - Manually change a couple of the numbers/made up data point to show new inferences. Show that the model can make predictions; pass in new data and retrieve a prediction. If we choose to do live we can have Sagemaker up, but we need to have a backup demo.

Day 5 - Thursday

- Wrap up any remaining tasks.
- Finalize the presentation.
  - We need a good reasoning for choosing the notebook instance type!
- Dry run presentation with the instructor
  - SHOOT FOR 15 MINUTES. It is very bad to go past 25 minutes.
- Pre-record any live demos for backup.
  - Use Open Broadcast Software (https://obsproject.com/) or QuickTime

Rashod: README & Clean up
Ashleigh/Alyssa: Post-Mortum
Jules: PowerPoint