https://www.kaggle.com/datasets/doshiharmish/311servicerequestboston2021/data - non-emergency request line in Boston
https://www.kaggle.com/datasets/vinodkumarcvk/healthcareticketingsystem/data - healthcare tickets ranging from basic medical diagnoses to emergency calls

Notes about EDA:

- EDA stands for Exploratory Data Analysis and consists of several parts
    - Data visualization
        - We have to pick a way to represent our data in a way that makes sense
        - Would this be a scatter plot to observe trends in data? Distribution graphs or histograms to see grouping of data? Box plots to measure variation in data?
        - To pick a method for visualizing data, we would have to decide on what kind of statistic and trends we would be tracking
    - Data cleaning
        - Missing values
            - In real-world data, there are often missing values
            - These can be interpolated using several methods
                - Mean estimation - using the average of existing values
                - Fixed value - setting a fixed value (such as 0) for missing values
                - Forward/backward fill - replace with the previous/next value
                - Regression interpolation - linear, quadratic, etc.
            - This will help our data and predictions stay smooth and will prevent the need to delete large amounts of data
        - Numerical and Categorical
            - Numerical values often need to be normalized using their mean and standard deviation
            - Categorical data can be handled using one-hot encoding, which is a method for converting categories into usable vectors
            - Quick explanation: instead of categories numbered like [1, 2, 3], one-hot encoding will create three separate vectors: [1,0,0], [0,1,0], and [0,0,1]
            - One-hot encoding also helps eliminate any ordering or precedence given to certain categories
    - Splitting into training, validation, and testing sets
        - The training set is what the model will use to learn the best parameters
        - Valid set is mostly used to track the real loss/accuracy and implement things like early stopping
        - Test set is to evaluate the model's accuracy after training is complete
        - There are several ways to split train and test sets:
            - Percentage of data

            - K-fold cross validation
            - Others?
        - Choosing specific methods for data cleaning and processing will become more apparent as we solidify our datasets and model choice

- Model selection
    - Recurrent neural network - simpler, won't work as well for long sequences
    - LSTM (long short-term memory) - more complicated, better for analyzing longer sequences
    - The choice will depend on what length of text we will be targeting

Finalized dataset: ***customer service tickets [kaggle link]***