Project Proposal Template

Question/need:

- The purpose of this project is to build a model using Natural Language Processing techniques such as Topic Modeling and sentiment analysis in order to understand common myths about the Covid vaccine and how sentiments around them have changed over time. Next, with myths that are still popular, we can recommend them as a theme for NYSDOH'S next vaccine myth debunking ad. The data used will be comments from reddits r/VaccineMyths subreddit.
- The findings of this model will benefit NYSDOH's aim to educate New Yorkers on the effectiveness/ reliability of the vaccine by creating a commercial that debunks popular myths about it.

Data Description:

- I will be downloading a publicly available vaccine myth dataset off of kaggle.

  Link: https://www.kaggle.com/gpreda/reddit-vaccine-myths

- The dataset includes 1531 rows of reddit comments

Tools:

- In order to complete this project, I plan on using the following tools: NLTK, spaCy, gensim, pandas, numpy, seaborn, matplotlib,

Methodology:

- Data Cleaning/ exploratory data anaysis
- Pre-processing comments (Tokenization, Lemmetization, stop words, stemming, named entity recognition, etc)
- Topic Modeling to find most popular myths
- Sentiment analysis on the most popular myths from the previous step- do redditors agree/ disagree (feel positively/negatively about) the

popular myths, how has the sentiment changed (gotten more/less positive) over time.

MVP Goal:

- An MVP for this project would be a jupyter notebook with preprocessing steps, EDA and sentiment analysis done