**Capstone Project 2 Proposal**

**Title:** Clothing Reviews

**Problem:**

The current reality of the apparel industry is that less and less consumers are frequenting brick and mortar establishments. Customers are finding themselves drawn more towards online shopping and e-commerce for its ease and convenience. A unique problem that is arising from e-commerce, particularly for the apparel industry, is how customers are deciding on specific products without physically seeing, touching, or trying on articles of clothing. As a result, reviews are the go to resource when making decisions on whether to buy something or not. Reviews are not only important for consumers, but also for any online retail. They rely on reviews to generate interest and profit on a particular item.

Reviews can be extensive and oftentimes contain irrelevant information. Can we compress the data to improve the product? Can we take the reviews and predict other consumer feedback? From reviews, online retailers can improve the quality of their product to satisfy consumers and increase profit and credibility.

**Who might care?**

* Online Retailers (Amazon, Target, Walmart, Ebay, etc.)
  + Use this report to improve products and increase profit and credibility.
* Online customers
  + Consider this report as a reference for quality reviews and key words for choosing a product or online retailer.

**Questions to analyze?**

What clothing item gets reviewed most?

What item gets recommended the most?

Are consumers in a certain age range more likely to leave reviews? Either positive or negative?

Are there any patterns in the consumer review narratives?

Can we predict recommendation from reviews?

How to improve the likelihood of an item being recommended?

**Data:**

Data is obtained from a Kaggle competition:

<https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews/version/1>

**Examples of features:** Clothing ID, Age, Review Text, Recommended, Positive Feedback Count, etc.

**Approach to solving this problem:**

Python numpy and pandas packages will be used for data wrangling. Dataset will be checked for any inconsistencies or missing/duplicate data.

Python's Scipy and statsmodel modules will be used to find statistic values and significance for this data set. Matplotlib and seaborn packages will be used for exploratory data analysis (EDA). Comparisons and correlations between variables will be analyzed and visualized. At this point, preliminary questions should be answered.

It would be interesting to see if there is a pattern in review text and positive feedback count. To do this, the application of statistical modeling and machine-learning will be tested with Python's scikit-learn module. There will be Natural Language Processing (NLP) using the review text and positive feedback features of the data set and other classification methods on other non-text features.

**Deliverables:**

1. Python code on GitHub:
   1. Data wrangling
   2. Data exploration analysis
   3. machine learning model development
2. Report on capstone project