# Progress Update 2 Part I

Data collection, cleaning and preparation is major part of data science domain. There isn’t much available pre-hand for skin care product besides list of products from website and websites such as Kaggle. Data preparation for this project include scrapping websites for data of item/products they are selling, their reviews, ratings, images, description, and user related data for that specific product. This requires using Scrapy, a Python framework that collects data from website based on the entered CSS and or Xpath. Plus, engineering and cleaning the data obtained from Kaggle to match the ideal format for the model to be used

## Data Preparation

Scrapy has been used to crawl and collect data from [lookfantastic](https://www.lookfantastic.com/). The features kept in mind for model processing are Product name, brand, rating, description, image, and ingredients. This will in turn help for content-based classification for recommendation. Attached [Skincare Products](#products_list) sheet contains the sample of 2926 product from a list of multinational cosmetics and skincare brands. These have been extracted from the web as well collected from Kaggle and the required features extracted. Although this is still in the works as you can see the review tabs are empty which will be filled with user semantic analysis from the reviews of these product which are collected in a separate sheet with user data, to be used for collaborative feedback. The second sheet, [User Reviews](#user_list) contains 8000+ user data of reviews and there features such as skin, hair and eye tone and their desired effect. More of this data is to be scraped from the lookfantastic website with the features to be kept in mind are user rating, product id, review, and skin type. This sheet is mainly still in works due to the sheer number of reviews to be extracted and passed in to the semantic analyzer that will classify the overall product rating as good or bad which will be used in the products sheet as a normalized feature to help identify the right product for the user.

SQL server was used to perform the required analysis for extracting the right requirements from the bulk raw data from the internet. Any null values will have to be taken care after the data has been prepared for consumption.

 