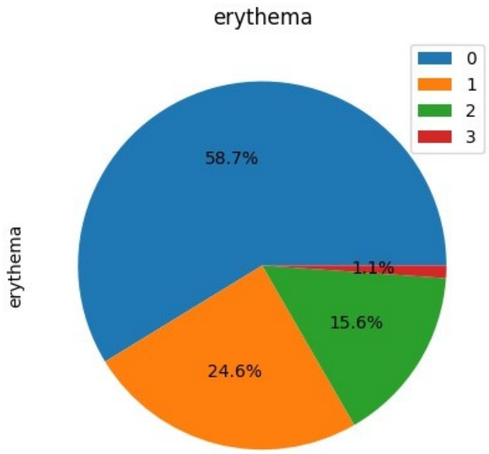
::Week 2 Insights::

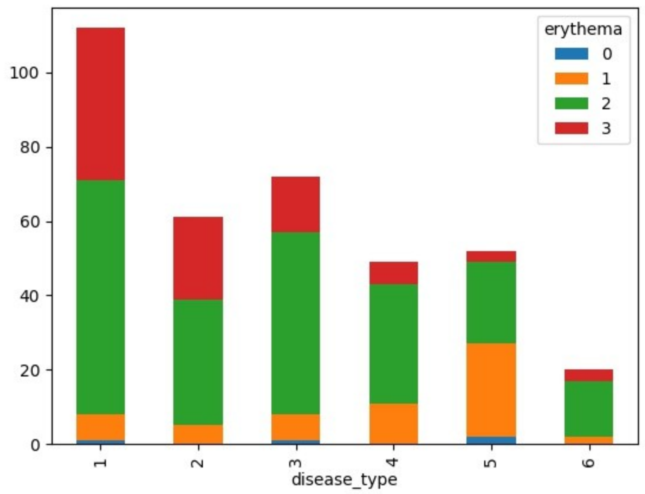
Week 1 recap:

- Understanding of domain
  - The dataset is a collections of symptoms which causes a particular type of skin disease
  - Moreover, the structure of cells at microscopic level is also given
- Univariate plots were made for each variable
- Correlation heatmap was made and studied
- The details were filled in the checklist file uploaded in git repo

This Week:

- Studying of univariate and bivariate models
  - Taking example of 1 column in dataset





Erythema is a type of symptom found more commonly in all 6 diseases given in the dataset.

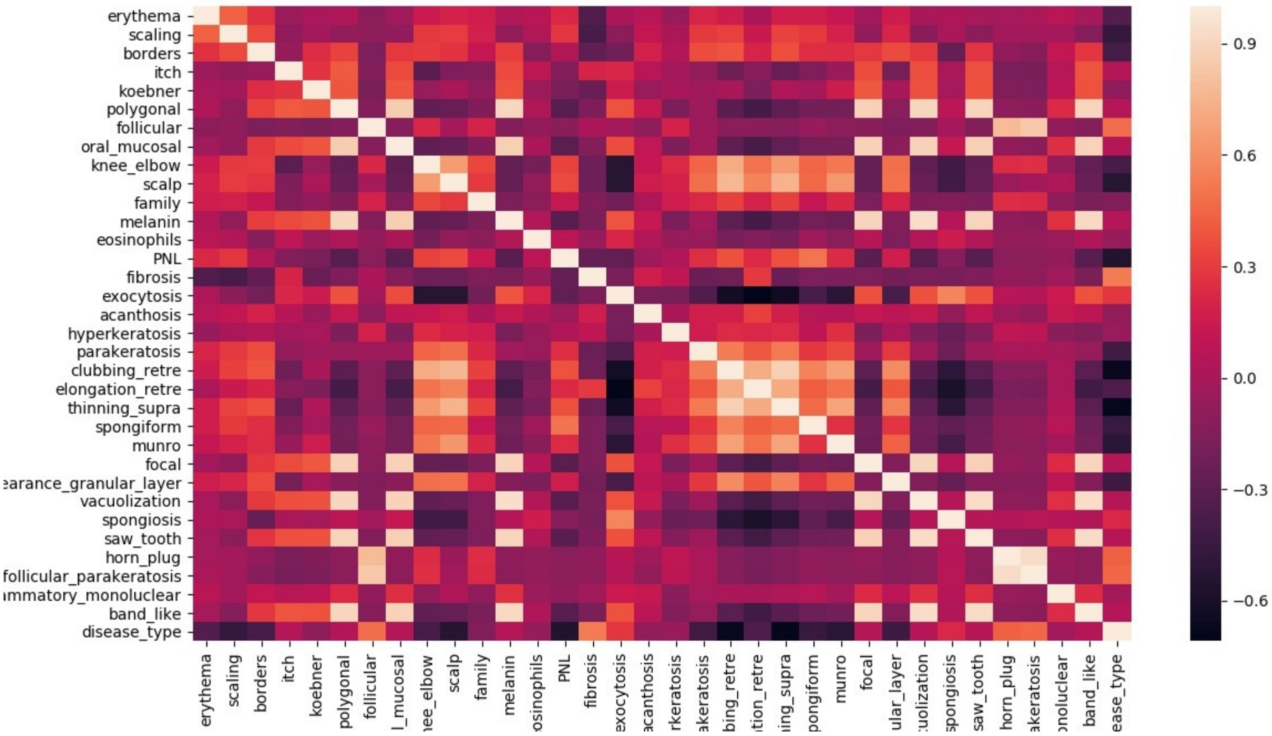
Checking the stacked plot of disease type vs level of erythema, we can fairly assume that erythema is one of the major symptom for having these diseases,considering that nearly all of the diseases have erythema as a major factor.

People not having erythema(58.7 %) have a much better chance of not contracting any type of disease.

It can also be assumed from the plots that people suffering from moderate (green-15.6 %) to high (red-1.1%) levels of erythema have a higher chance of contracting disease type 1 than other type of disease, particularly disease type 6.

Checked for every type of symptom and cell level with respect to disesase type.

- Heatmap Analysis



Looking at the heatmap, it can be fairly assumed that most of the symptoms and cellular level structures are generally correlated to each other.

- Feature Generation
  - Feature generation, however, is not possible as introducing newer symptoms and/or cell structures at microscopic level requires extreme scientific research.
  - Combining the results of two or more symptoms will also require a deeper understanding into the field of dermatology.
  - However, based on the insights, some symptoms can be dropped considering that there is least amount of variance of those symptoms across all types of diseases.

- Missing Value Treatment
    - There were only 8 missing values in the age (numeric) column
    - So, missing values were generated at specific points in the dataset
    - Taking example of erythema column in dataset- missing values were introduced at every 5th observation in this column
    - Imputed values by using knn imputation, central imputation and a package called miss forest
    - Miss forest and central imputation gave the best results for imputed values
    - Time taken by each method (in R):
        - time_knn is knn imputation from dmwr package
        - time_cen is central imputation from dmwr package
        - time _miss is missforest package imputation

time_knn	time_cen	time_miss
0.2197113 secs	0.002248287 secs	2.280271 secs

-
    - In Python (knn and simple fill used from fancy impute package)
        - time_mode is function self written to impute with mode

time_knn	time_mode	time_simple
0.054 secs	0.057 secs	0.008 secs

-
    -
        - Using fancyimpute package for imputation in python seems the best option
    - Central imputation seems to be the most optimal solution for this dataset

- Since all the features are categorical except for the age column, there is no need for standardizing the values