Capstone Project 1 Milestone Report

Problem Statement:

Every day more than 9,500 people are diagnosed with skin cancer, making it the most common cancer in America. As of 2011, the annual cost of treating skin cancer is $8.1 billion, and the Center for Disease Control (CDC) called it “a growing public health issue”.

Skin cancer has a high survival rate when caught early on; however, if left untreated, invasive cancerous cells can migrate to other parts of the body via the blood or lymph supply and become lethal. Thus, early detection and treatment is crucial for optimal recovery. The problem is, many benign conditions of the skin appear similar to skin cancer, and for an untrained individual it may be difficult to differentiate between a such harmless condition and something more malign.

In this capstone project, I would like to create an application of sorts for a non-health professional individual who suspects that he/she may have skin cancer. The individual can upload a picture of their skin condition, and the system will return the most likely diagnosis among those included in the dataset. Using this information, the individual can make a more informed decision on whether or not to pursue further treatment.

Dataset Description and Wrangling:

My dataset is the HAM10000 dataset from the Medical University of Vienna (available for download at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DBW86T).

The dataset contains 10015 photos of cancer-like skin conditions and a .csv metadata with the lesion id, the image id, the diagnoses, the age of the patient, the sex of the patient, and the location of the condition. Refer to Table 1 for a summary of the skin conditions contained in the dataset, Figure 1 for an example image to be analyzed, and Figure 2 for the first ten rows in the metadata.

Table 1: A description of the diagnoses included in the HAM1000 dataset .

Abbreviation	Diagnosis	Non/Cancerous	Description	Counts

akiec	Actinic keratoses and intraepithelial carcinoma	Noncancerous	A pre-cancerous area of dry skin caused by UV damage.	327
bcc	Basal cell carcinoma	Cancerous	Manifests as a bump and is usually caused by prolonged exposure to UV rays	514
bkl	Benign keratosis-like lesions	Noncancerous	A benign skin condition on the upper extremity that can be patchy or dry in appearance.	1099
df	Dermatofibroma	Noncancerous	Small modules that form in the upper extremities.	115
mel	Melanoma	Cancerous	Cancerous proliferation of melanocytes. The most common type of skin cancer and manifests as a darkened patch.	1113
nv	Melanocytic nevi	Noncancerous	Noncancerous proliferation of melanocytes. Commonly known as a “mole”.	6705
vasc	Vascular lesions	Noncancerous	Occurs after injury to the skin. Commonly known as a “bruise”.	142

Figure 1. Example of an image to be used for the model.

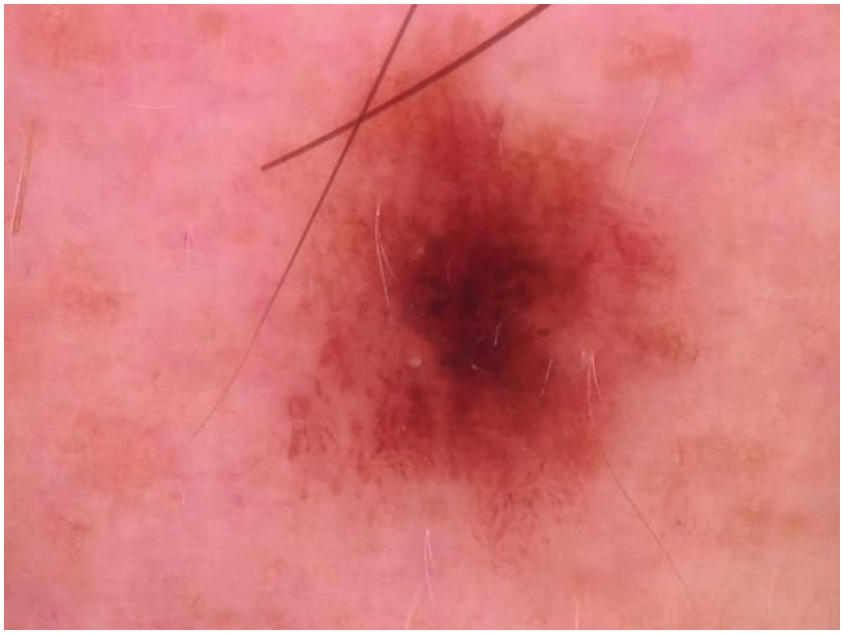


Figure 2. The first ten rows of the tabular metadata.

	lesion_id	image_id	dx	dx_type	age	sex	localization
0	HAM_0000118	ISIC_0027419	bkl	histo	80.0	male	scalp
1	HAM_0000118	ISIC_0025030	bkl	histo	80.0	male	scalp
2	HAM_0002730	ISIC_0026769	bkl	histo	80.0	male	scalp
3	HAM_0002730	ISIC_0025661	bkl	histo	80.0	male	scalp
4	HAM_0001466	ISIC_0031633	bkl	histo	75.0	male	ear
5	HAM_0001466	ISIC_0027850	bkl	histo	75.0	male	ear
6	HAM_0002761	ISIC_0029176	bkl	histo	60.0	male	face
7	HAM_0002761	ISIC_0029068	bkl	histo	60.0	male	face
8	HAM_0005132	ISIC_0025837	bkl	histo	70.0	female	back
9	HAM_0005132	ISIC_0025209	bkl	histo	70.0	female	back

The metadata contains 10015 rows, corresponding with the 10015 images in the dataset. The column names are as follows:

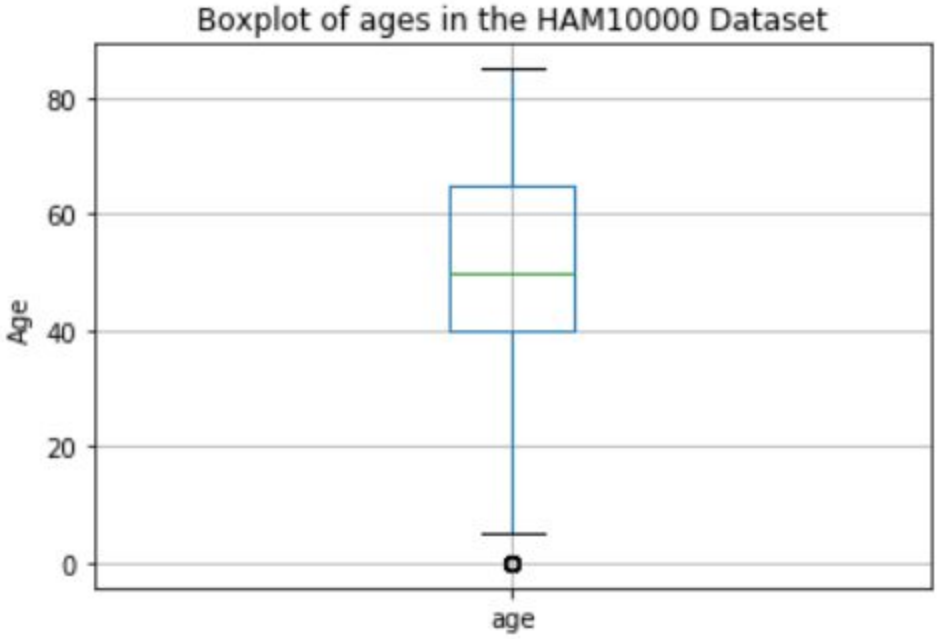
- **lesion_id** - Lesion ID (identification code for the skin condition, may or may not be unique for every picture as multiple pictures may be taken for the same condition
- **image_id** - Image ID (identification code for the image itself, is unique for every picture

- **dx** - Diagnosis
- **dx_type** - How the Diagnosis was made (histology, confocal viewing, at a follow-up, or by consensus)
- **age** - Age of the patient
- **sex** - Sex of the patient
- **localization** - Where the skin condition was found

All of the columns were categorical aside from age, which is numerical. At initial look, it appears that the only column which contains any null values is the “age” column with 57 nulls. However, it is entirely possible that missing values in other columns were not encoded as “NaN”, and were instead put in as something else.

First, we take a look at the numerical data. In figure 3, we have a box-and-whisker plot that displays both the spread and the existence of any outliers in patient age.

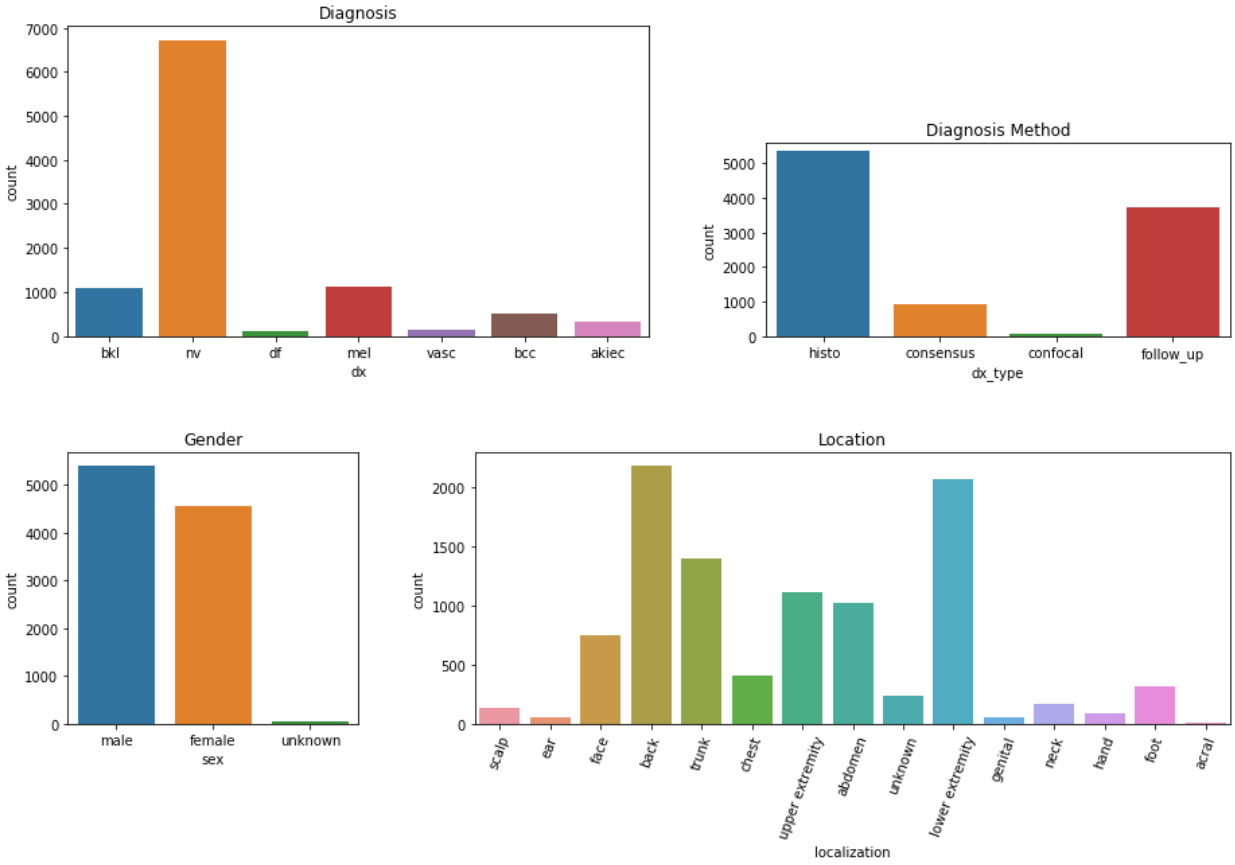
Figure 3. Boxplot of the age column of the HAM10000 metadata.



Patients’ ages range from 0 to 85, with a mean of 52 and a standard deviation of 17 years. There is only one outlier at 0, referring to newborn patients, that seems suspicious and warrants further examination. Upon examining the 39 entries where the patient age is 0, the diagnoses are either "nv" (nevi, or a mole/birthmark), "vasc" (vascular lesion, or a bruise), or "bkl" (benign keratosis-like lesions), none of which are cancerous. This information makes sense, could be useful, and doesn't need to be dropped.

Then we can take a look at the categorical variables. Figure 4 is a distribution of unique categories for the categorical variables (Diagnosis, Diagnosis Method, Gender, Location) in the HAM10000 metadata. I excluded the image id and lesion id columns because these columns have many values, with most of them unique for every entry, making it meaningless to plot.

Figure 4. Distribution of data in categorical variables.



One notable observation to be made about the diagnosis column is that the data is very imbalanced, with one category ("nv") occupying more than 6000 of the 10015 entries. This could have effects on our model testing later on.

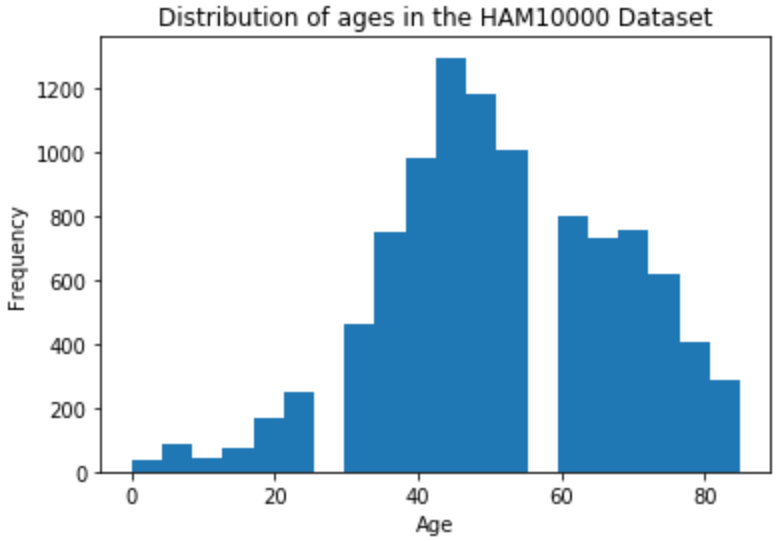
There is one category in the gender and location columns, "unknown", that refers to values not known. I inputted values for these columns, as well as the null values in the age column, using Random Forest.

Initial findings:

Figure 5 is a histogram of the age column. It appears that patient age does not follow a normal distribution, and the data skew to the left with most of the patients being older than 40 years of

age. This makes sense, since the risk for all types of cancer, including skin cancer, increases with age and is much higher for people older than 40 years than for those younger.

Figure 5. Distribution of patient ages in the HAM10000 dataset

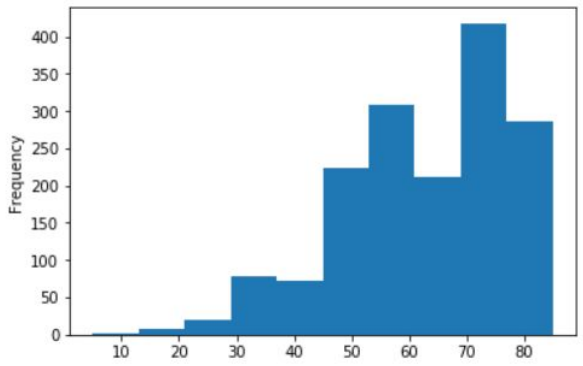


In particular, I was curious if patients with cancerous diagnoses (Melanoma or Basal cell carcinoma) are older than patients with noncancerous diagnoses (Actinic keratoses, Benign keratosis-like lesions, Dermatofibroma, moles, or bruises). To test this hypothesis, I split the data into cancerous and non-cancerous and compared the distribution of patient age for the two groups.

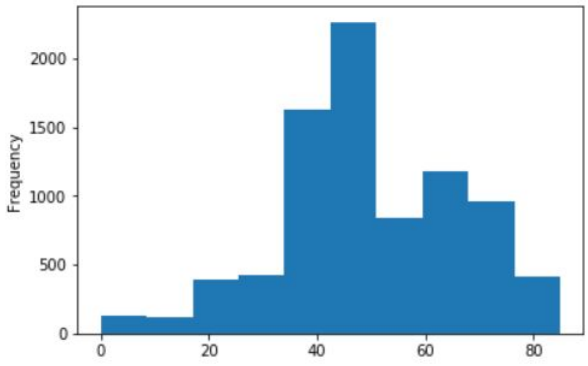
There are 1637 cancerous cases and 8388 non-cancerous cases, with the mean age for patients with cancerous conditions being 62.65 years with a standard deviation of 15 years and that of patients with non-cancerous conditions 49.77 years with a standard deviation of 16.53 years. The distributions can be seen below in Figure 6: as shown, the cancerous data is heavily skewed to the left with the majority of the data concentrated in the higher ages, whereas the non-cancerous data has a more normal distribution and most of the data is centered and symmetrical around the mean.

Figure 6. Distribution of ages for cancerous data (left) and noncancerous data (right)

```
cancer['age'].plot(kind='hist')
plt.show()
```



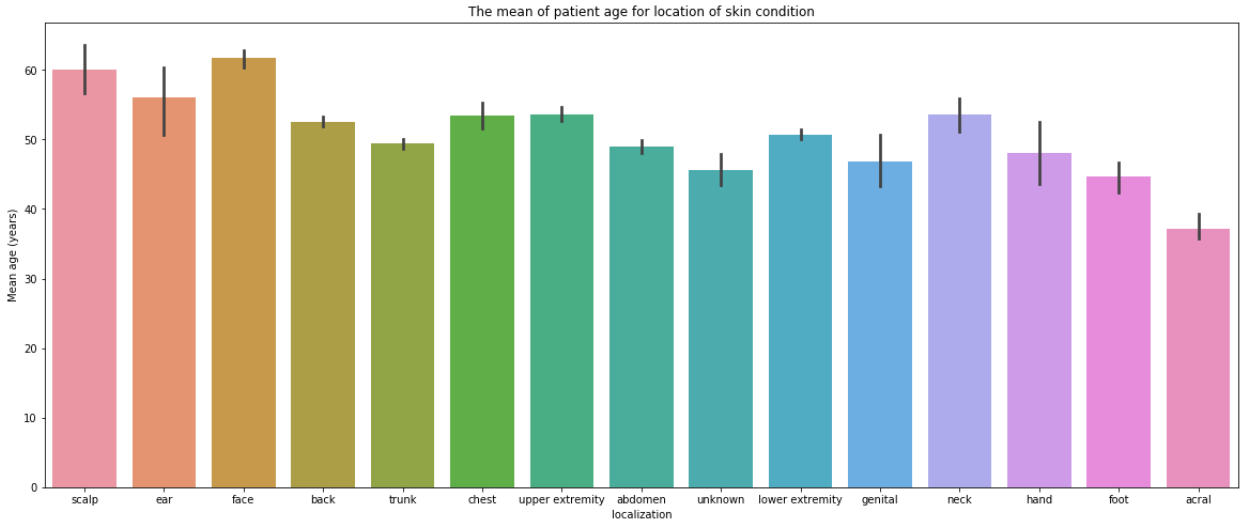

```
noncancer['age'].plot(kind='hist')
plt.show()
```



To see if the difference is statistically significant, I used a frequentist approach and a student t-test with an alpha set to 0.05. My null hypothesis was that there was no difference between patient ages between the two groups, and my alternate hypothesis was that non-cancerous patients were younger than cancerous patients. Running a one-way hypothesis test, I got a t-value of 31.11 and a p-value less than 0.05, meaning that indeed, cancerous patients are significantly older than non-cancerous patients.

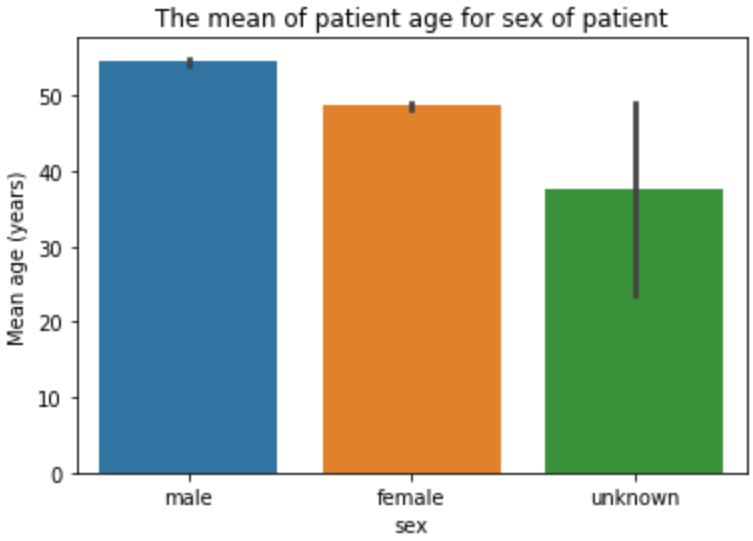
I also took a look at the mean ages for different locations of condition and genders in figures 7 and 8, respectively.

Figure 7. Mean patient ages for different locations of skin condition.



Generally, most ages were around 50 years, but patients whose condition was on the scalp and face tended to be older (around 60 years old) while those with conditions in the foot, unknown, or acral regions tended to be younger (around 45 years old).

Figure 8. Mean patient ages by gender



It appears that male patients are generally older than female ones, and that those with their sex unknown has the greatest variation. This makes sense, since those with unknown sex can either be male or female.