SKIN CANCER ANALYSIS

ISHA SINGH
ISINGH4@DONS.USFCA.EDU

HTTPS://ISINGH4.GITHUB.IO/PROJECT.HTML
ALPHA RELEASE

TABLE OF CONTENTS

BACKGROUND AND MOTIVATION

AN EVERYDAY ROUTINE SHOULD CONSIST OF YOURSELF TAKING GOOD CARE OF YOUR SKIN. WHEN YOUR SKIN IS NOT GIVEN PROPER CARE, IT COULD RESULT IN SKIN CANCER. OF COURSE, THAT IS NOT ALWAYS THE CASE BUT THERE IS ALWAYS THE POSSIBILITY.

MY DAY REVOLVES AROUND SKINCARE AND DATA SCIENCE. SINCE ALWAYS I HAVE FOUND INTEREST IN SKINCARE AND KNOWING MORE ABOUT WHAT IMPORTANT FEATURES CAN HELP REDUCE SKIN CANCER. READING MORE ABOUT THE PREVENTIONS GAVE ME A CLEAR UNDERSTANDING OF WHY IT'S NECESSARY TO ALWAYS CARE ABOUT THE SKIN. DURING MY FREE TIME, I ENJOY EXPERIMENTING AND CREATING HOMEMADE MASKS, USING NATURAL INGREDIENTS. FURTHERMORE, I WANTED TO FIGURE OUT WHAT ARE OTHER ASSOCIATIONS THAT CAN INCREASE THE CHANCES OF HAVING SKIN CANCER.

I HAVE DECIDED TO DO THIS RESEARCH PROJECT AS I WOULD LIKE TO COMBINE MY PASSION, SKINCARE, WITH DATA, TO ANALYZE AND DISCUSS THE THE CORRELATION ON WHY AND HOW SKIN CANCER OCCURS.

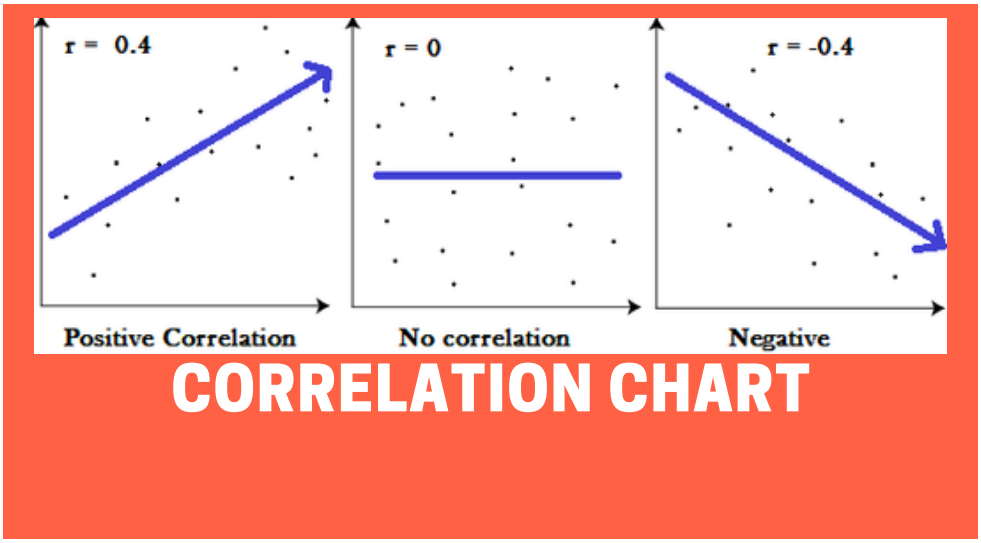
.

PROJECT OBJECTIVES

The objective of this project is to understand what other features have a relationship with increasing the chances of getting skin cancer, or in other words, if they have some sort of association.

In specific, these are some of the objectives that will be focused on in the research itself:

- Is there a correlation between the location with the highest UV index and where there are a lot of skin cancer patients?
- At what age are most people getting skin cancer.
- Which gender has a higher chance of being diagnosed with cancer and at what age?



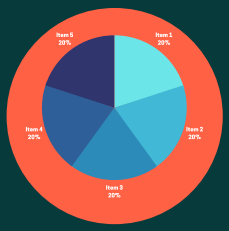
CORRELATION CHART

DATA

Here are some relevant data that I will be using during the following research project.

01 — SKIN CANCER MNIST HAM10000

This dataset consists of large collection of multi-source dermatoscopic images of pigmented lesions.
This data can be useful as it shows factors such as sex, age, and localization. I can use these features to understand my hypothesis on which sex has a more significantly higher rate for getting skin cancer. I can also understand which Sex and what age has more risk in getting skin cancer.
Link: https://www.kaggle.com/kmader/skin-cancer-mnist-ham10000



Skin Lesion Analysis Towards Melanoma Detection

The following dataset focuses mainly on understanding and focusing on Melanoma. Main features include sex, age, and location. https://challenge2020.isic-archive.com/

Dataset #3 [In process]

Looking for a data set, which will assist to determine where are the underlying factors of skin cancer by the most important way to find that out would be by using UV radiation. The UV radiation would help signify if there is a correlation on if there is a high number of diagnosis of skin cancer, is there a high number of UV index's

note: there maybe dataset#4 required as per this project if location seems to be an important factor.

DATA PROCESSING

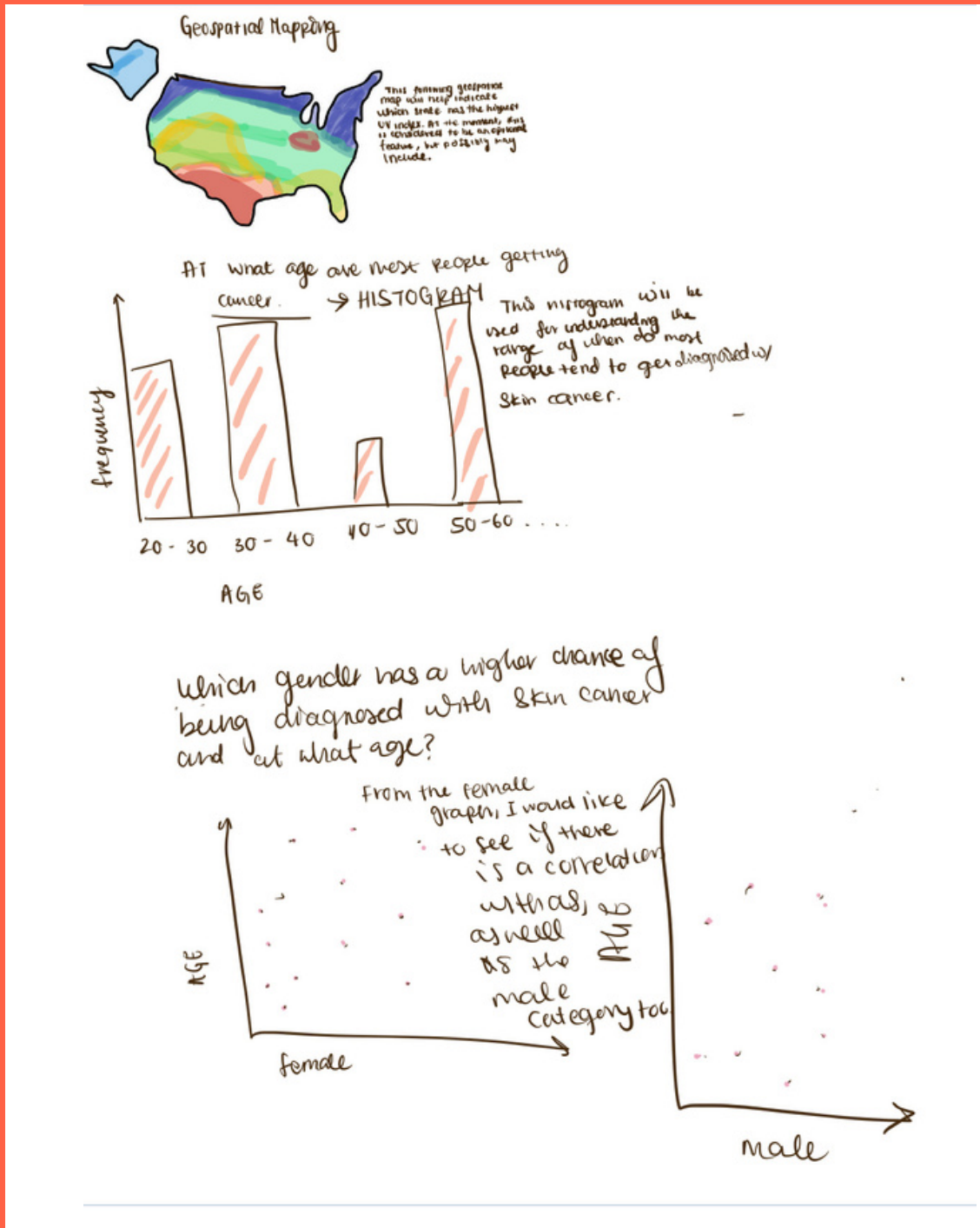
The way NAs can be removed is by the action of pre-processing. This is the most important aspect when taking a look at the data itself and experimenting with the data. It is very unclean data if one does not tend to remove the data itself. In order to do that, one important way is to replace the NA values with the mean (or yet the average) of the column itself. That procedure can be done in Python and more specifically its library, Pandas.

```
import numpy as np
import pandas as pd
import sklearn as sk
metadata=pd.read_csv("metadata.csv")
```

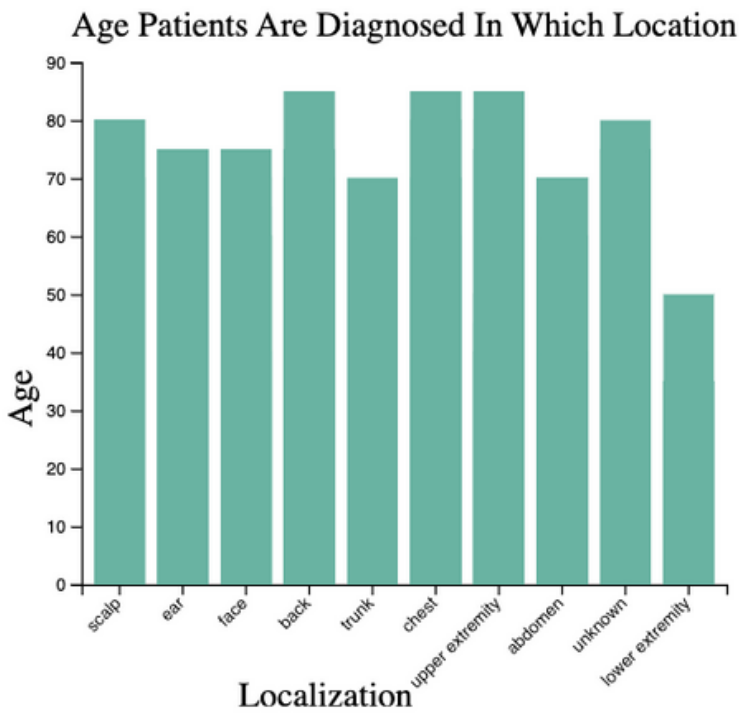
```
metadata=metadata.dropna()
```

```
metadata.to_csv('new_meta_data.csv', index = False)
```

VISUALIZATION DESIGNS PART I (DRAFT)



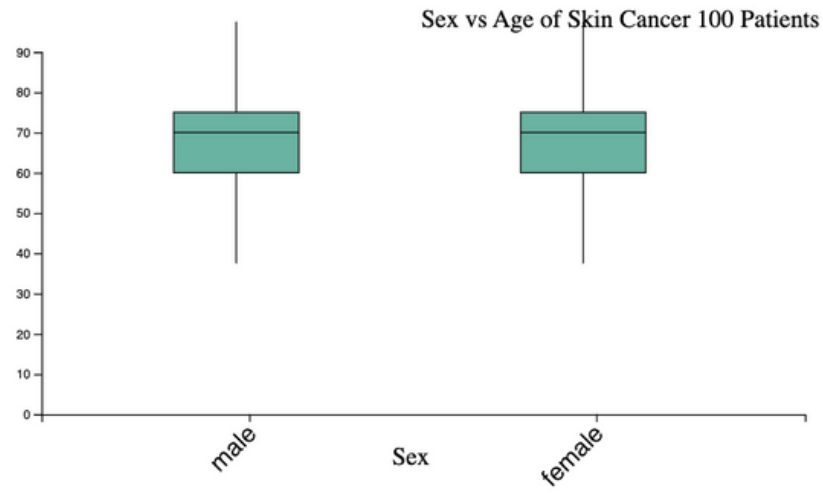
VISUALIZATION DESIGNS CONTINUED PART II



Hypothesis I: One factor that I wanted to understand is at what age do most people tend to get diagnosed with skin cancer. From this bar chart, which was done with the assistance of D3, we see that most people get diagnosed with skin cancer on the back around the age of around 85. The lowest age of getting diagnosed is age 50 for lower extremities. Overall, this bar chart shows at what age are most patients getting what type of skin cancer.

Source: https://vizhub.com/isingh4/a50d32c4ffe144ac9887b171d31eb8ad?edit=files&file=index.html

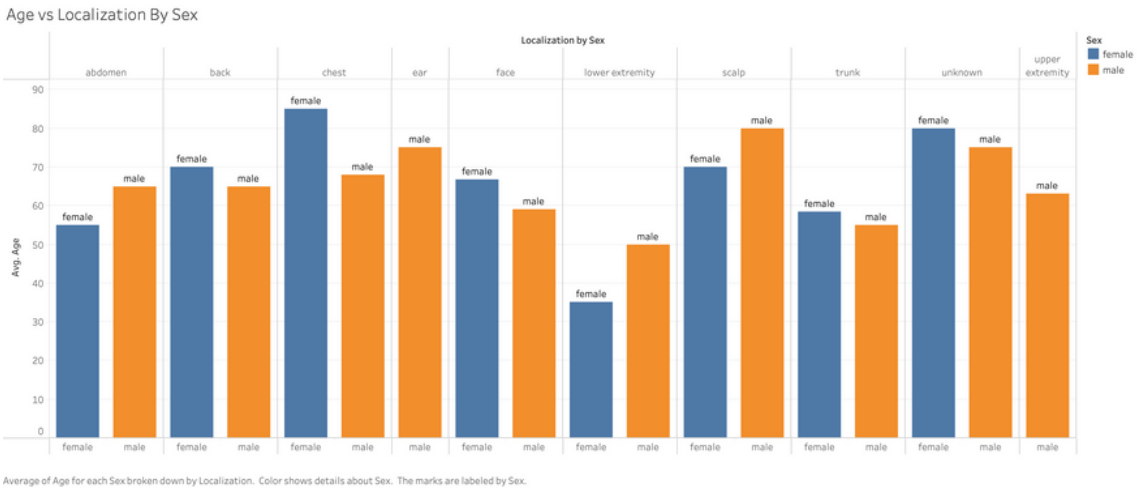
VISUALIZATION DESIGNS CONTINUED PART III



Hypothesis II: Another piece of information that I wanted to understand is there at an earlier age males and/or females are being diagnosed with skin cancer. My hypothesis was that possibly males have a higher chance of being detected with skin cancer earlier as females tend to care for their skin more than males do. Unfortunately, I was wrong about this. The age of most patients was fairly the same for both male and female.

https://vizhub.com/isingh4/dcc0376d2bf149ff914185d399dfb0e5?edit=files&file=index.html

VISUALIZATION DESIGNS CONTINUED PART IV



Hypothesis III: Another main important factor is having a comparison of when do most females and males get detected at what age with skin cancer. This is an important representation it shows all important factors.

MUST HAVE FEATURES

BOX PLOT

The most necessary plot as of right now I believe is a box plot to understand if there are any

Frequency Bar Chart

The second most necessary plot will of age and its frequency to determine what age are most people being diagnosed with skin cancer and which specific location do most skin cancers are diagnosed.

OPTIONAL FEATURES

GEOSPATIAL MAPPING

One optional feature that possibly may make the research more interesting is to have two sets of maps describing data One would be understanding which state has the highest UV index and which state has the highest number of skin cancer cases.

RELATED WORK

Bhalla, Sherry, et al. “Prediction and Analysis of Skin Cancer Progression Using Genomics Profiles of Patients.” Nature News, Nature Publishing Group, 31 Oct. 2019, https://www.nature.com/articles/s41598-019-52134-4.

Holtz, Yan. “The D3 Graph Gallery – Simple Charts Made in d3.Js.” The D3 Graph Gallery – Simple Charts Made with d3.Js, https://d3-graph-gallery.com/.

McDowell, Sandy. “Researchers Identify States Where Improved Sun Protection Could Prevent the Most Melanomas.” American Cancer Society, American Cancer Society, 17 Feb. 2020, https://www.cancer.org/latest-news/researchers-identify-states-where-improved-sun-protection-could-prevent-the-most-melanomas.html#:~:text=The%20southern%20states%2C%20California%2C%20and,a%20relatively%20high%20UV%20Index.

“Melanoma of the Skin Statistics.” Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 8 June 2021, https://www.cdc.gov/cancer/skin/statistics/index.htm.

Park, Young Ji, et al. “A Retrospective Study of Changes in Skin Cancer Characteristics over 11 Years.” Archives of Craniofacial Surgery, Korean Cleft Palate-Craniofacial Association, Apr. 2020, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7206466/.

UPCOMING IMMEDIATE MILESTONES

- Add more key and legends to the graph
- If time, work on geospatial.

PROJECT SCHEDULE

Requirement	Details of Requirement	Deadline
Project Proposal	Basic Information, background and motivation, project objectives, data, data processing, visualization design, must-have features, optional features and project schedule.	11 March 2022
Revised Proposal, related work, and website	Update the project from the advised suggestions that the project has given.	23 March 2022
Task 1	Clean the data using pandas	1 April 2022
Alpha Release		8 April 2022
Task 2	Start Working on Visualizations	8 April 2022
Task 3	Complete Visualizations	15 April 2022
Beta Release		20 April 2022
Final Project Presentation	Present the project to the class	16 May 2022