**Midway Report**

Project in Data Science

Group: *G (Greedy Geckos)*

Aniela Marta Ciecierska

Francisco Gonçalves Medeiros Lemos Moreno

Jakub Piotr Gąsior

Jonas Drøivoldsmo Lesund

Michaela Macejovska

As a task for Projects in Data Science, the group analyzed the relevant pictures in order to take conclusions about possible occurrence of cancer. By researching several webpages (Seladi-Schulman, 2022) (Allan C. Halpern, 2021) (Osborn, s.d.) (Anadolu-Brasie R, 2008) (Michael J. Greco, 2023), the team gathered some information about different types of cancer and categorized the pictures into each disease. Since the images are being seen on a phone or laptop screen and the research is conducted by unexperienced students, some images were categorized into more than one group. The findings are as follows:

Actinic keratosis - a discolored, rough spot, up to 1 cm in diameter, often with hyperkeratotic

layers, sometimes taking the form of a cutaneous horn. 71

Basal cell carcinoma - most often a pale color nodule, well demarcated from the surrounding

skin, with translucent vessels. In more advanced forms with the presence of an ulcer with a ridged edge. 49

Melanoma - asymmetrical lesion, with irregular borders, non-uniform color (from light brown to

black) and diameter over 6 mm. 2

Nevus - symmetrical lesion, with even borders, uniform color (from light brown to black) and

diameter up to 6 mm. 13

Squamous cell carcinoma - from an erythematous, scaly papule to an ulcerated, disintegrating

Tumor. 22

Seborrheic keratosis - more or less raised lesions above the skin surface, sometimes

pedunculated, with a smooth, lumpy, or rough surface, pale or brown color. 25

**Is there some missing data?**

Each skin lesion is made up of a maximum of 26 features, where each line denotes a skin lesion, and each column denotes feature. It seems that there is a pattern of missing values. The same values are missing in every observation. There are some cells that contain "UNK" indicating unknown or missing data (“background\_father”, “background\_mother”, “grew”, “changed”, “itch”). Some rows have missing values, where the data is missing or was not recorded (“smoke”, “gender”, “cancer\_history”, “has\_piped\_water” and so on).

The reasons for missing data could range from data entry errors, non-response from the patients or medical professionals to systematic issues with how data is recorded or transferred.

Handling missing data depends on the context and the importance of the missing information. For the column, which seems to be a unique identifier for diagnosis, missing data might be more critical and could potentially indicate cases where the diagnosis was not recorded. Before handling the missing data, it's important to understand why it is missing. For instance, if the missing data is not random (i.e., there is some pattern to the missingness), simply deleting or imputing without investigating could introduce bias into any conclusions drawn from the data.