Group Members: Quentin Shin (ttvy6kv), Taylor Decina (mvu2ab), Keeley Messick (vmv6mp)

DS 3001

2/27/2025

01: Project Wrangling/EDA

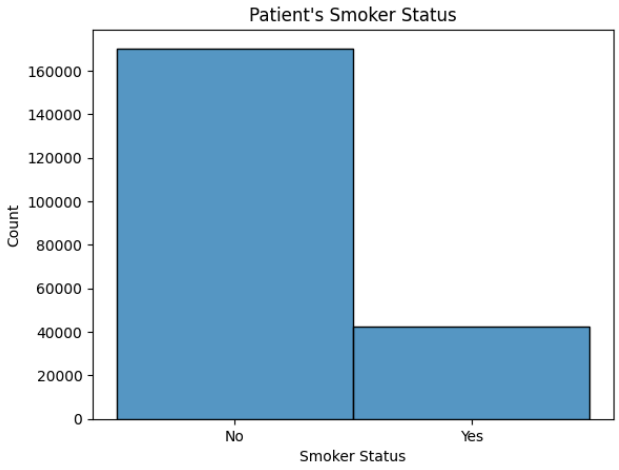
1. What is in your data?

- Our data is on demographic and clinicopathologic variables. These include Patient Id, Age, Gender, Country, Ethnicity, Family History, Radiation Exposure, Iodine Deficiency, Smoking, Obesity, Diabetes, TSH Level, T3 Level, T4 Level, Nodule Size, Thyroid Cancer Risk, and Diagnosis. The risk of thyroid cancer serves as our dependent variable. This risk is measured in terms of low, medium, and high. The diagnosis variable contributes by describing the risk's severity. The others act as independent variables which aim to predict thyroid cancer. The data comes from a Kaggle dataset from 2025. It was collected over the span of 15 years. Each patient was followed for at least 10 years.

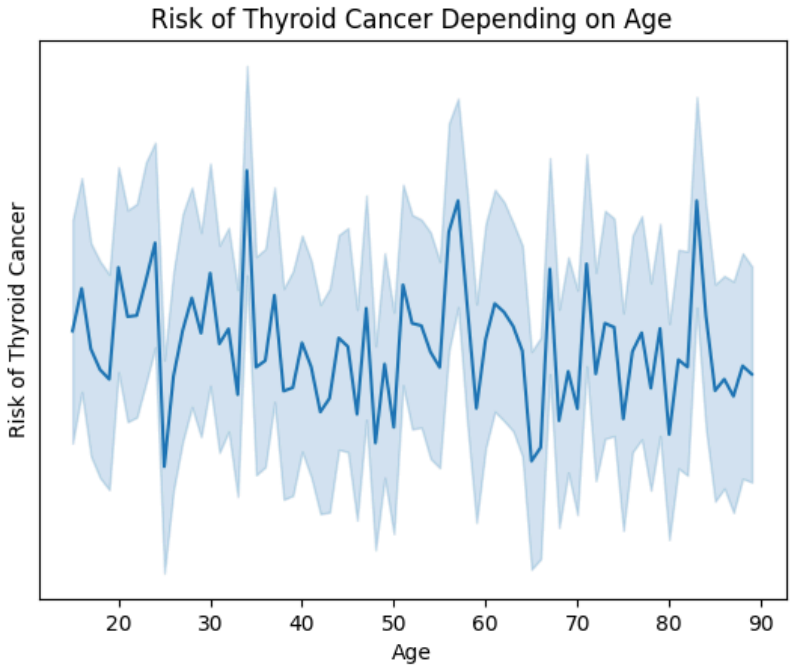
2. How will these data be useful for studying the phenomenon you're interested in?

- Thyroid cancer is extremely treatable when caught early. It is important to identify risk factors that contribute to its cause to better inform people about potential exposures. We are curious about which variables play a significant role in the risk of thyroid cancer. Knowing this information would greatly benefit the public when making lifestyle choices. Knowing whether you are predisposed to be at risk or that your actions are contributors can be crucial to preventing thyroid cancer or at the very least, keeping its risk low.
- We have done some minor exploratory data analysis on a variety of variables to see any patterns of interest in this topic.

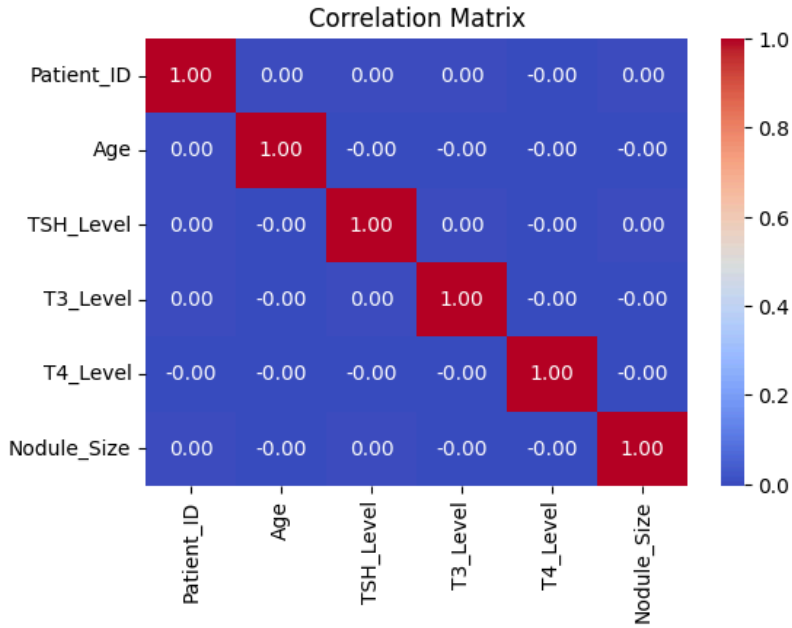
Histogram of Smoker Status



Line Plot of Risk of Thyroid Cancer Depending on Age



Correlation Matrix



How Many of Each Type Grouped by Country

Country	Patient_ID	Age	Gender	Ethnicity	Family_History	Radiation_Exposure	Iodine_Deficiency	Smoking	Obesity	Diabetes	TSH_Level	T3_Level	T4_Level	Nodule_Size	Thyroid_Cancer_Risk	Diagnosis
Brazil	21413	21413	21413	21413	21413	21413	21413	21413	21413	21413	21413	21413	21413	21413	21413	21413
China	31978	31978	31978	31978	31978	31978	31978	31978	31978	31978	31978	31978	31978	31978	31978	31978
Germany	10557	10557	10557	10557	10557	10557	10557	10557	10557	10557	10557	10557	10557	10557	10557	10557
India	42496	42496	42496	42496	42496	42496	42496	42496	42496	42496	42496	42496	42496	42496	42496	42496
Japan	16867	16867	16867	16867	16867	16867	16867	16867	16867	16867	16867	16867	16867	16867	16867	16867
Nigeria	31918	31918	31918	31918	31918	31918	31918	31918	31918	31918	31918	31918	31918	31918	31918	31918
Russia	21297	21297	21297	21297	21297	21297	21297	21297	21297	21297	21297	21297	21297	21297	21297	21297
South Korea	14965	14965	14965	14965	14965	14965	14965	14965	14965	14965	14965	14965	14965	14965	14965	14965
UK	10642	10642	10642	10642	10642	10642	10642	10642	10642	10642	10642	10642	10642	10642	10642	10642
USA	10558	10558	10558	10558	10558	10558	10558	10558	10558	10558	10558	10558	10558	10558	10558	10558

How Many of Each Type Grouped by Gender

Gender	Patient_ID	Age	Country	Ethnicity	Family_History	Radiation_Exposure	Iodine_Deficiency	Smoking	Obesity	Diabetes	TSH_Level	T3_Level	T4_Level	Nodule_Size	Thyroid_Cancer_Risk	Diagnosis
Female	127527	127527	127527	127527	127527	127527	127527	127527	127527	127527	127527	127527	127527	127527	127527	127527
Male	85164	85164	85164	85164	85164	85164	85164	85164	85164	85164	85164	85164	85164	85164	85164	85164

Proportion Grouped by Gender

Gender	Patient_ID	Age	Country	Ethnicity	Family_History	Radiation_Exposure	Iodine_Deficiency	Smoking
Female	59.958813	59.958813	59.958813	59.958813	59.958813	59.958813	59.958813	59.958813
Male	40.041187	40.041187	40.041187	40.041187	40.041187	40.041187	40.041187	40.041187

Obesity	Diabetes	TSH_Level	T3_Level	T4_Level	Nodule_Size	Thyroid_Cancer_Risk	Diagnosis
59.958813	59.958813	59.958813	59.958813	59.958813	59.958813	59.958813	59.958813
40.041187	40.041187	40.041187	40.041187	40.041187	40.041187	40.041187	40.041187

3. What are the challenges you've resolved or expect to face in using them?

- There are still some potential issues with this data. One being that our dependent variable is in terms of levels (low, medium, high) so we will have to visually accommodate for that in our charts. Additionally, some of the data is not common knowledge, like T4_Level, so that will require explanations to the common user to allow for better understanding. A final challenge might be needing to change the categorical variables to binary. There are six variables that are “Yes” and “No” answers and it would be easier to display these numerically in our graphs.