Team 2: Adam, Babatola, Iqra and Thinley

Thyroid Cancer Risk Analysis

Providing an in-depth analysis of dataset and uncovering key insights into which demographics are at high risk of having thyroid cancer.

code institute

Overview

Providing an in-depth analysis of dataset and uncovering key insights into which demographics are at risk of having thyroid cancer:

- **Hypothesis 1**: Age and gender influence thyroid cancer risk, with older individuals and females having a higher probability
  - **Validation**: Use box plots and regression analysis to explore how cancer risk varies across different age groups and gender distributions

- **Hypothesis 2**: Certain countries and ethnicities have a higher prevalence of thyroid cancer due to genetic and environmental factors
  - **Validation**: Conduct geospatial analysis and visualize the distribution of thyroid cancer cases across different regions.

Planning & Design

Ideation

Project Goal: Build interactive dashboards for data analysis

Medical Use Case: Improve decision-making through data insights of thyroid cancer risk based on demographics

Target Audience: Medical professionals, analysts, WHO and decision-makers

Design

User Stories: "As a Medical practitioner, we want to explore demographics of thyroid cancer risk dynamically."

Intuitive UI: Clean layouts, easy navigation

Accessibility: Readable colours, labeling

Interactivity: Clickable filters, zoomable charts and maps

Hypothesis: Which demographic has high risk of thyroid cancer

Technologies

Tools: Visual Studio Code, Jupyter Notebook, Power BI, PowerPoint

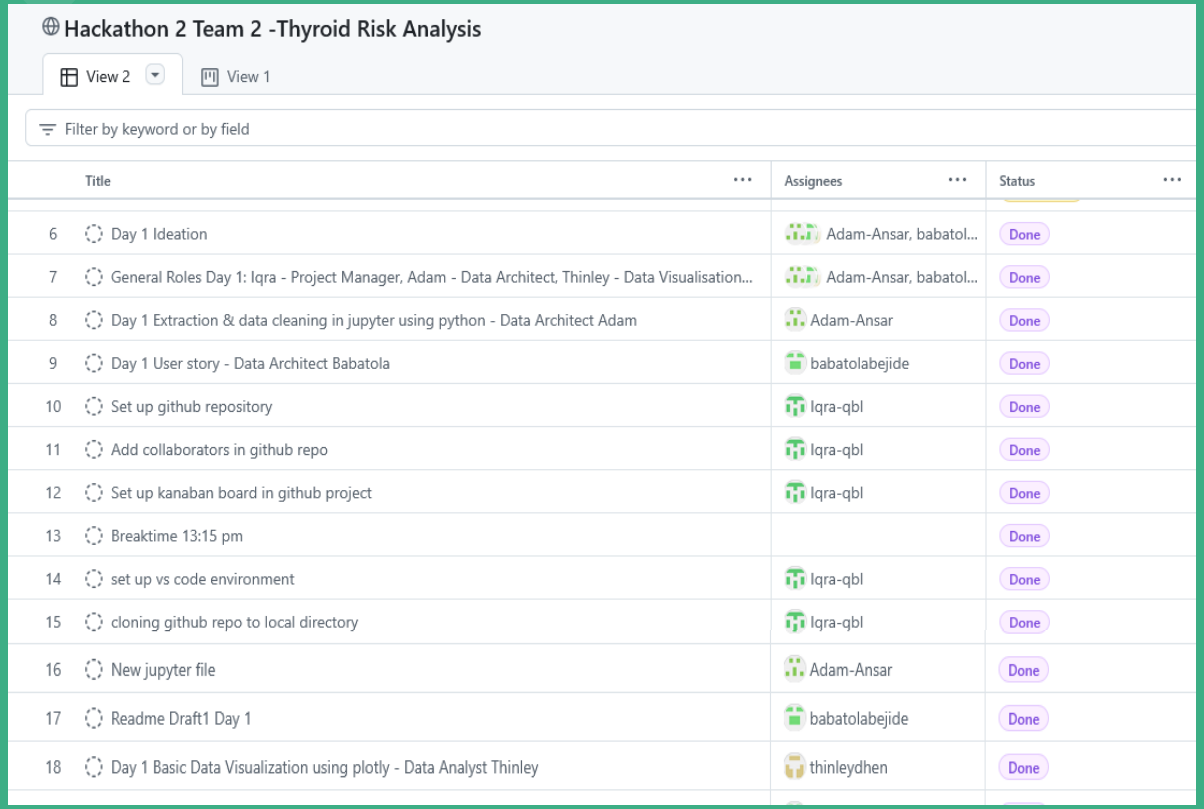
Wireframing: Balsamic Wireframes

Project Management: GitHub Projects, Google Meets

Version Control: GitHub for collaboration

Libraries & Frameworks: Python (Pandas, NumPy, Plotly, Seaborn, Ipywidgets, StatsModels), Power BI

Project Board



Hackathon 2 Team 2 -Thyroid Risk Analysis

View 2 | View 1

Filter by keyword or by field

	Title	Assignees	Status
6	Day 1 Ideation	Adam-Ansar, babatol...	Done
7	General Roles Day 1: Iqra - Project Manager, Adam - Data Architect, Thinley - Data Visualisation...	Adam-Ansar, babatol...	Done
8	Day 1 Extraction & data cleaning in jupyter using python - Data Architect Adam	Adam-Ansar	Done
9	Day 1 User story - Data Architect Babatola	babatolabejide	Done
10	Set up github repository	Iqra-qbl	Done
11	Add collaborators in github repo	Iqra-qbl	Done
12	Set up kanaban board in github project	Iqra-qbl	Done
13	Breaktime 13:15 pm		Done
14	set up vs code environment	Iqra-qbl	Done
15	cloning github repo to local directory	Iqra-qbl	Done
16	New jupyter file	Adam-Ansar	Done
17	Readme Draft1 Day 1	babatolabejide	Done
18	Day 1 Basic Data Visualization using plotly - Data Analyst Thinley	thinleydhen	Done

Assigned Tasks: 30

MoSCoW Prioritisation:

- **Must have:** an interactive dashboard and proper documentation
- **Should have:** simple and followable code
- **Could have:** nicely stylized code and dashboard
- **Would have:** utilize the dataset to full potential

Project Backlog: Setting VS Code and Github collaboration

Team 2 - Thyroid Risk Analysis

XX % High Risk

XX % Low Risk

Country A

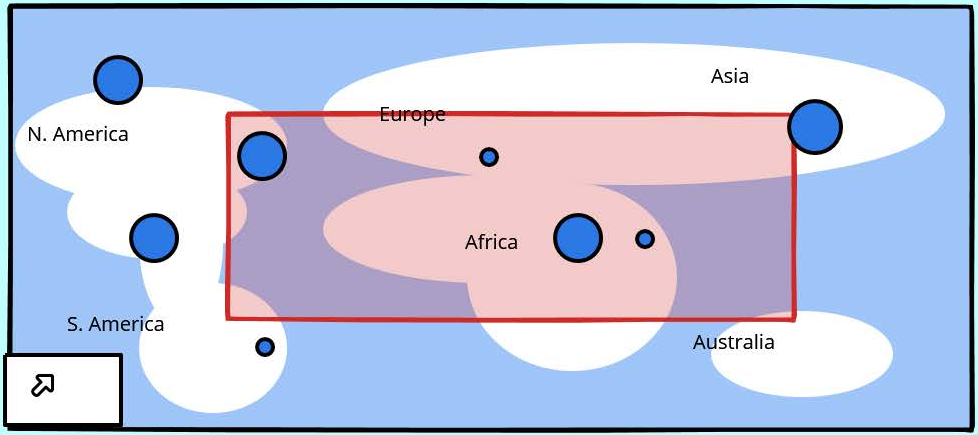
Country B

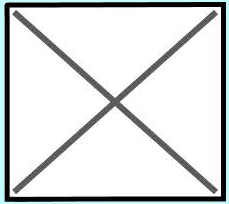
Country C

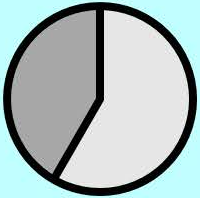
Country D

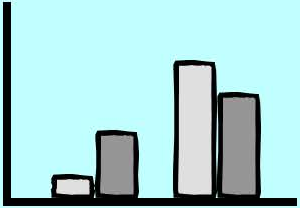
Country E

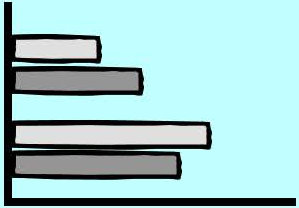
Country F

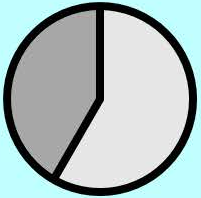










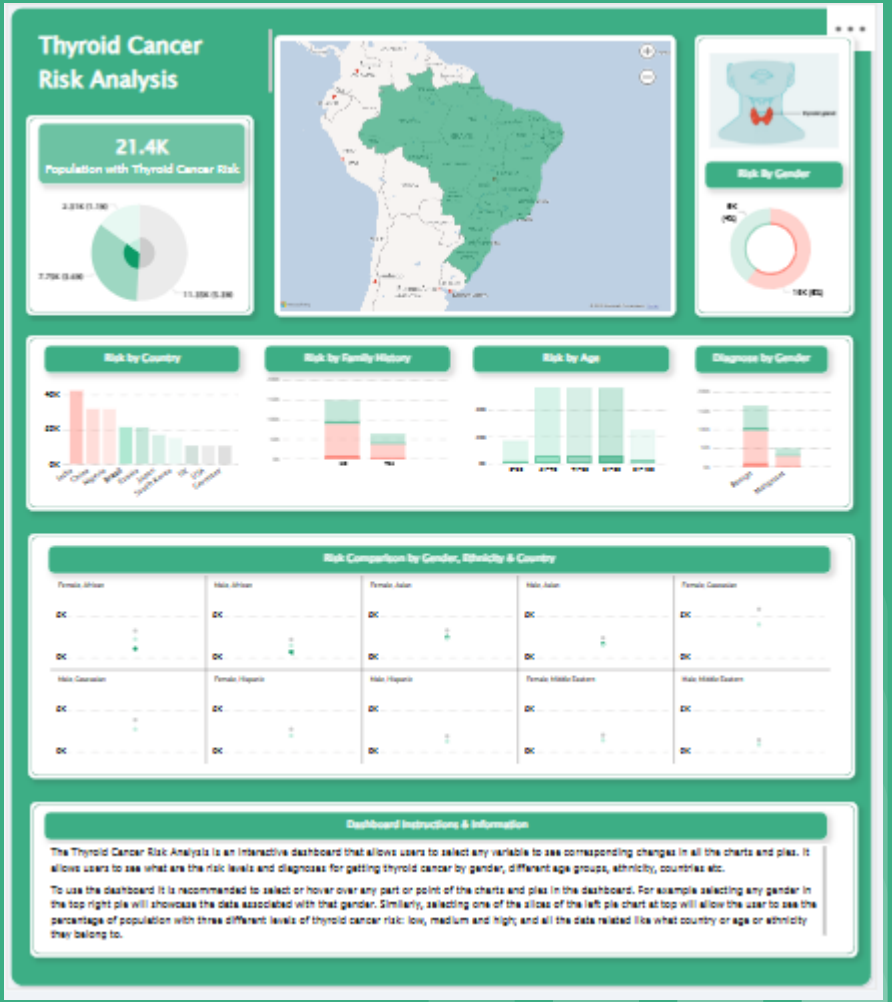


Features

1. **Interactive:** all variables i.e. age, gender, country, ethnicity, diagnosis are interlinked so selecting one will show all the relevant data across different charts and map

2. At the bottom there are instructions on using the different features of the dashboard

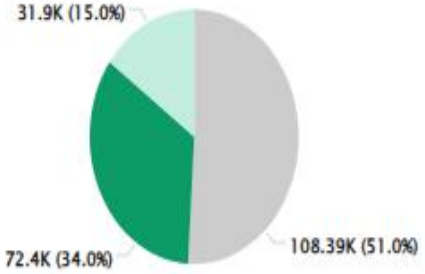
3. Hovering over any chart will further explain the data specific e.g. regarding gender, amount of population at risk, country etc

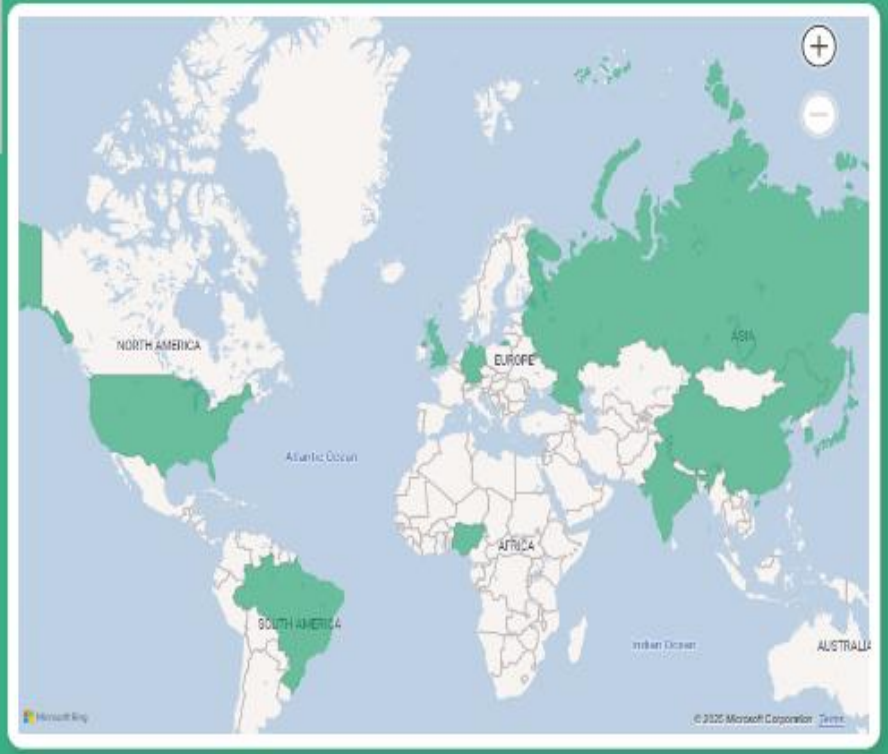


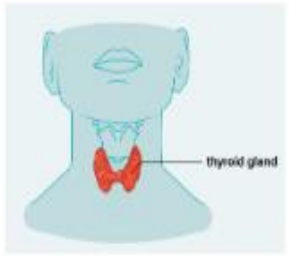
Thyroid Cancer Risk Analysis

212.7K

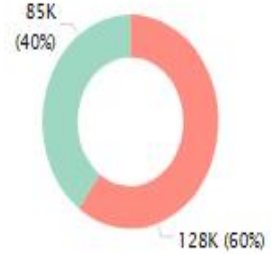
Population with Thyroid Cancer Risk







Risk By Gender

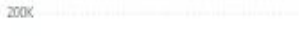


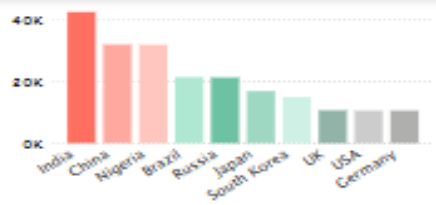
Risk by Country

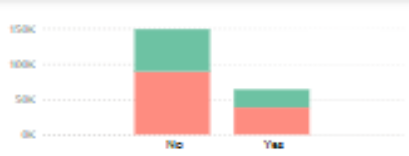
Risk by Family History

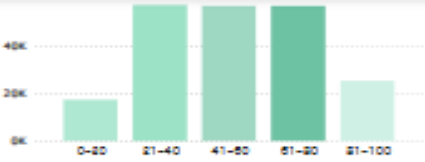
Risk by Age

Diagnose by Gender

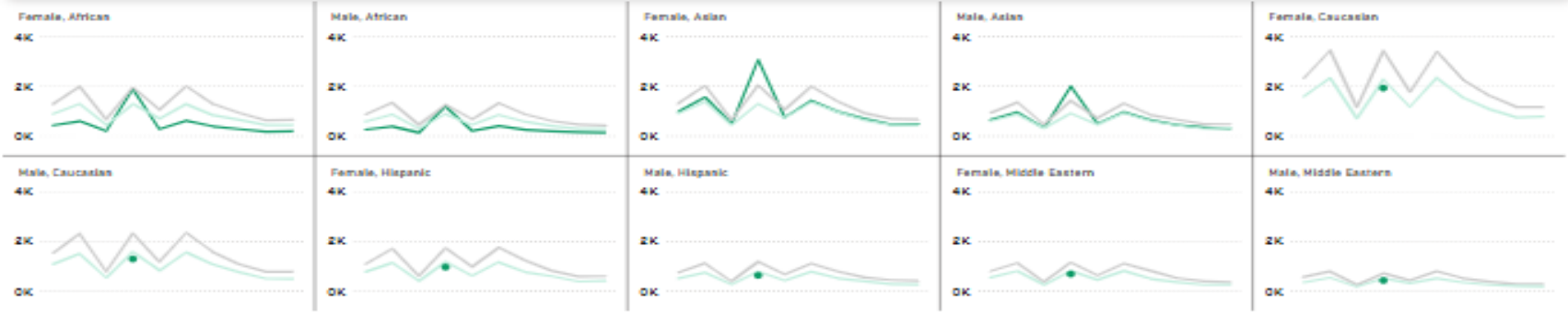








Risk Comparison by Gender, Ethnicity & Country

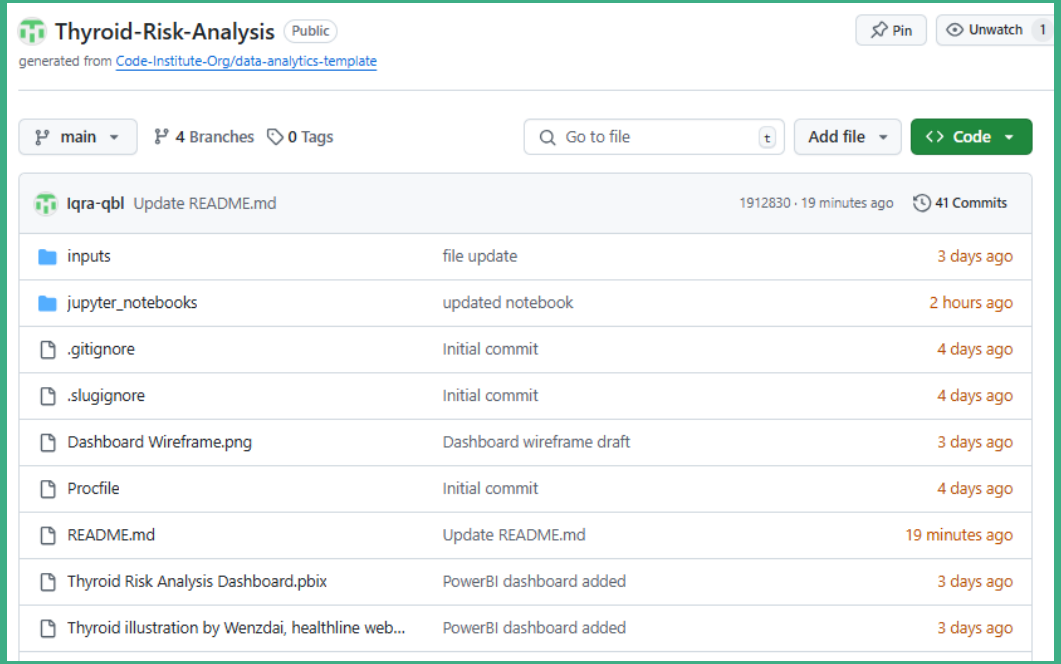


Dashboard Instructions & Information

The Thyroid Cancer Risk Analysis is an interactive dashboard that allows users to select any variable to see corresponding changes in all the charts and pies. It allows users to see what are the risk levels and diagnoses for getting thyroid cancer by gender, different age groups, ethnicity, countries etc.

To use the dashboard it is recommended to select or hover over any part or point of the charts and pies in the dashboard. For example selecting any gender in the top right pie will showcase the data associated with that gender. Similarly, selecting one of the slices of the left pie chart at top will allow the user to see the percentage of population with three different levels of thyroid cancer risk: low, medium and high; and all the data related like what country or age or ethnicity they belong to.

Documentation, Testing & Version Control



Thyroid-Risk-Analysis Public

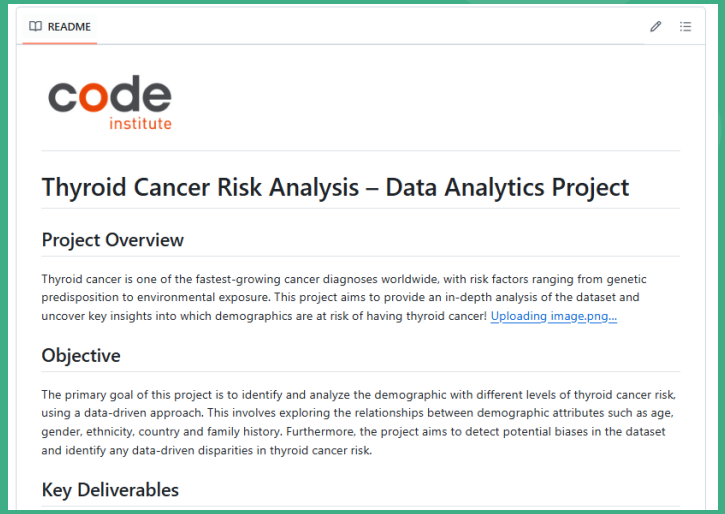
generated from Code-Institute-Org/data-analytics-template

Pin | Unwatch 1

main | 4 Branches | 0 Tags | Go to file t | Add file | Code

Iqra-qbl Update README.md — 1912830 · 19 minutes ago — 41 Commits

inputs	file update	3 days ago
jupyter_notebooks	updated notebook	2 hours ago
.gitignore	Initial commit	4 days ago
.slugignore	Initial commit	4 days ago
Dashboard Wireframe.png	Dashboard wireframe draft	3 days ago
Procfile	Initial commit	4 days ago
README.md	Update README.md	19 minutes ago
Thyroid Risk Analysis Dashboard.pbix	PowerBI dashboard added	3 days ago
Thyroid illustration by Wenzdai, healthline web...	PowerBI dashboard added	3 days ago



README

code institute

Thyroid Cancer Risk Analysis – Data Analytics Project

Project Overview

Thyroid cancer is one of the fastest-growing cancer diagnoses worldwide, with risk factors ranging from genetic predisposition to environmental exposure. This project aims to provide an in-depth analysis of the dataset and uncover key insights into which demographics are at risk of having thyroid cancer! Uploading image.png...

Objective

The primary goal of this project is to identify and analyze the demographic with different levels of thyroid cancer risk, using a data-driven approach. This involves exploring the relationships between demographic attributes such as age, gender, ethnicity, country and family history. Furthermore, the project aims to detect potential biases in the dataset and identify any data-driven disparities in thyroid cancer risk.

Key Deliverables

GitHub Repository: https://github.com/Iqra-qbl/Thyroid-Risk-Analysis

Section 1 : Data Extraction, Transformation, and Loading (ETL)

Setting up & Importing Python packages that we will be using in this project to carry out the analysis. For example Numpy to compute numerical operations and handle arrays, Pandas for data manipulation and analysis, Matplotlib, Seaborn and Plotly to create different data visualisations.

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
sns.set_style('whitegrid')
import plotly.express as px
```

✓ 3.8s Python

Data Extraction

Loading the CSV dataset containing the data collected previously and extracting it into dataframe using pd.read_csv() function.

```python
df = pd.read_csv("thyroid_cancer_risk_data.csv")
```

Section 2: Bias Detection

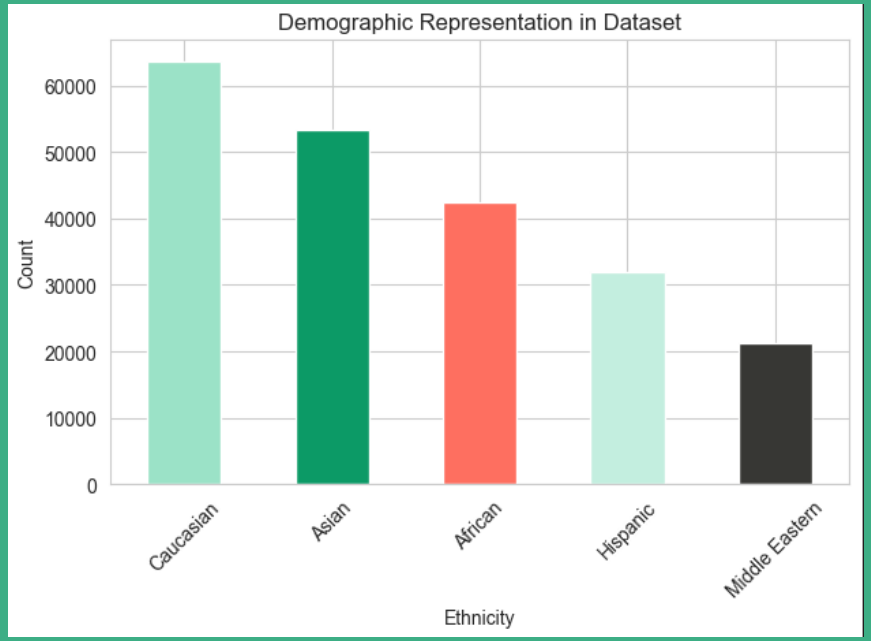
Checking the data for any bias in any of the classes

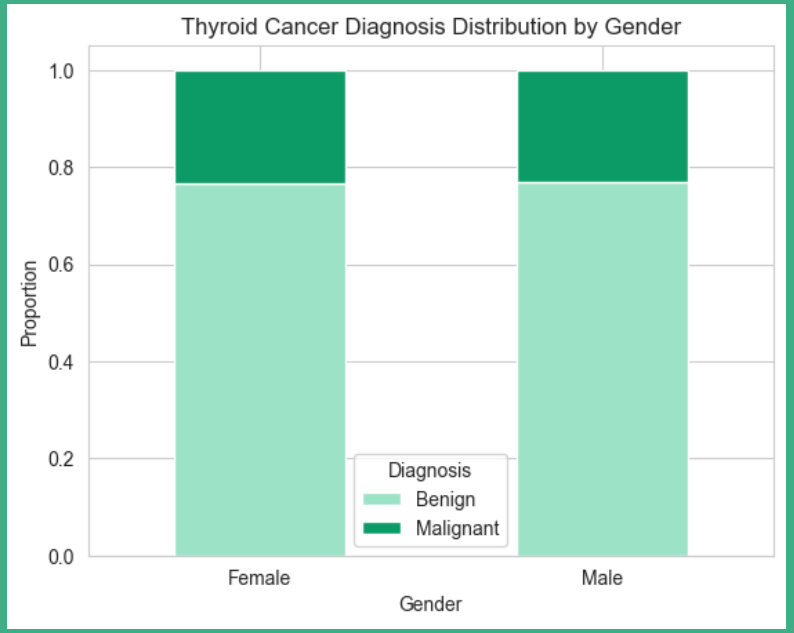
Bias in data can significantly affect the outcomes and interpretations of our analysis. It is crucial to identify and address any biases to ensure the validity and fairness of our results. In this section, we will check for potential biases in the dataset across different classes such as gender, age, ethnicity, and family history.

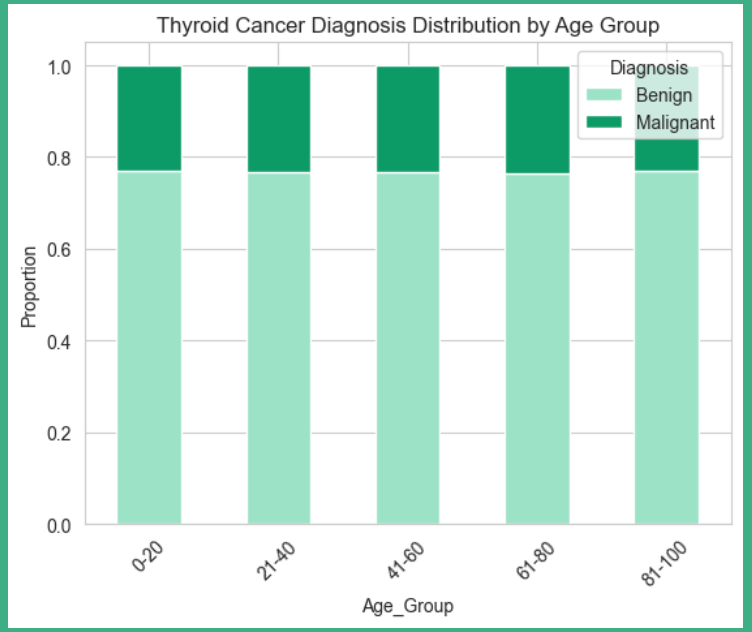
Load the cleaned dataset for Bias Analysis

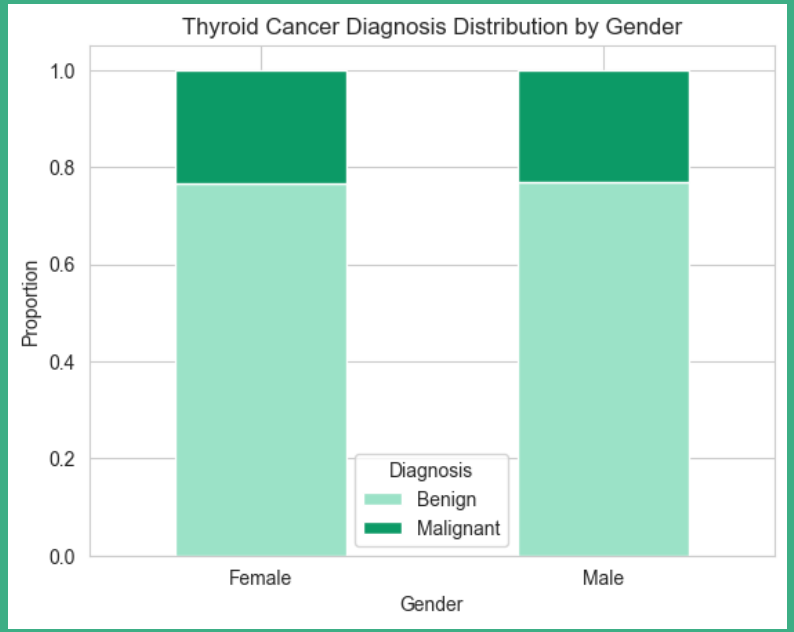
```
# Analyze demographic representation
data = pd.read_csv('thyroid_cancer_risk_data_cleaned.csv')

demographic_counts = data['Ethnicity'].value_counts()
print(demographic_counts)
```

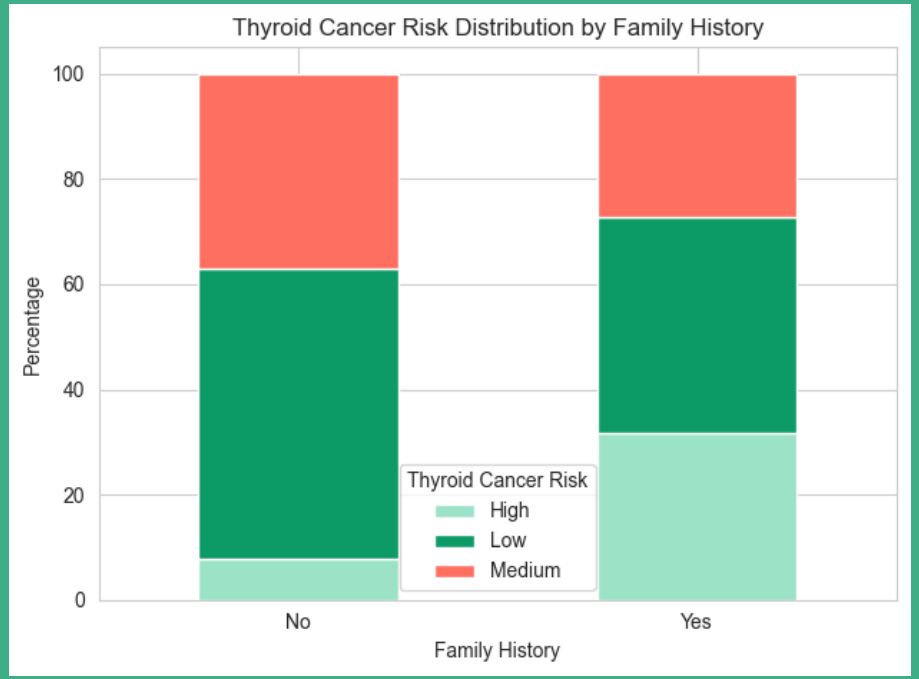
[15] ✓ 0.6s Python


Demographic Representation in Dataset
Count
60000
50000
40000
30000
20000
10000
0
Caucasian
Asian
African
Hispanic
Middle Eastern
Ethnicity


Thyroid Cancer Diagnosis Distribution by Gender
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
Diagnosis
Benign
Malignant
Female
Male
Gender


Thyroid Cancer Diagnosis Distribution by Age Group
Diagnosis
Benign
Malignant
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
0-20
21-40
41-60
61-80
81-100
Age_Group


Thyroid Cancer Diagnosis Distribution by Gender
Diagnosis
Benign
Malignant
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
Female
Male
Gender


Thyroid Cancer Risk Distribution by Family History
100
80
60
40
20
0
Percentage
Thyroid Cancer Risk
High
Low
Medium
No
Yes
Family History

Section 3: Analysis

Hypothesis 1: Age and gender influence thyroid cancer risk, with older individuals and females having a higher probability.

Validation: Use box plots and regression analysis to explore how cancer risk varies across different age groups and gender distributions.
Logistic Regression Modeling: Fitting a logistic regression model to quantify the effect of age and gender on thyroid cancer probability.

Empty markdown cell, double-click or press enter to edit.

[20]
```
from statsmodels.miscmodels.ordinal_model import OrderedModel

# Convert categorical variables to numerical values
df['Thyroid_Cancer_Risk_Numeric'] = df['Thyroid_Cancer_Risk'].map({'Low': 1, 'Medium': 2, 'High': 3})
df['Gender_Binary'] = df['Gender'].map({'Female': 1, 'Male': 0})

# Define independent variables (without intercept)
X_ordinal = df[['Age', 'Gender_Binary']]
```

```
Optimization terminated successfully.
         Current function value: 0.994911
         Iterations: 13
         Function evaluations: 16
         Gradient evaluations: 16
                              OrderedModel Results
==============================================================================
Dep. Variable:     Thyroid_Cancer_Risk_Numeric   Log-Likelihood:          -2.1161e+05
Model:                            OrderedModel   AIC:                       4.232e+05
Method:                  Maximum Likelihood      BIC:                       4.233e+05
Date:                     Mon, 10 Feb 2025
Time:                             20:18:29
No. Observations:                   212691
Df Residuals:                       212687
Df Model:                                2
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Age              0.0003      0.000      1.504      0.133   -8.72e-05       0.001
Gender_Binary    0.0178      0.008      2.111      0.035       0.001       0.034
1/2              0.0640      0.012      5.354      0.000       0.041       0.087
2/3              0.5284      0.003    155.718      0.000       0.522       0.535
==============================================================================
```

[21] ✓ Python

Age: 36

Gender: Female

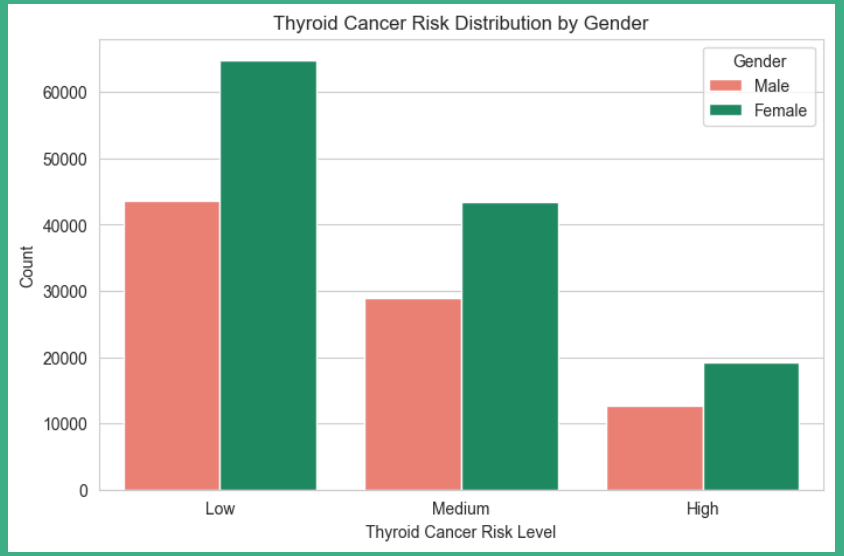
```
'For Age: 36, Gender: Female'

'Probability of Low Risk: 0.5230'

'Probability of Medium Risk: 0.1126'

'Probability of High Risk: 0.3643'
```

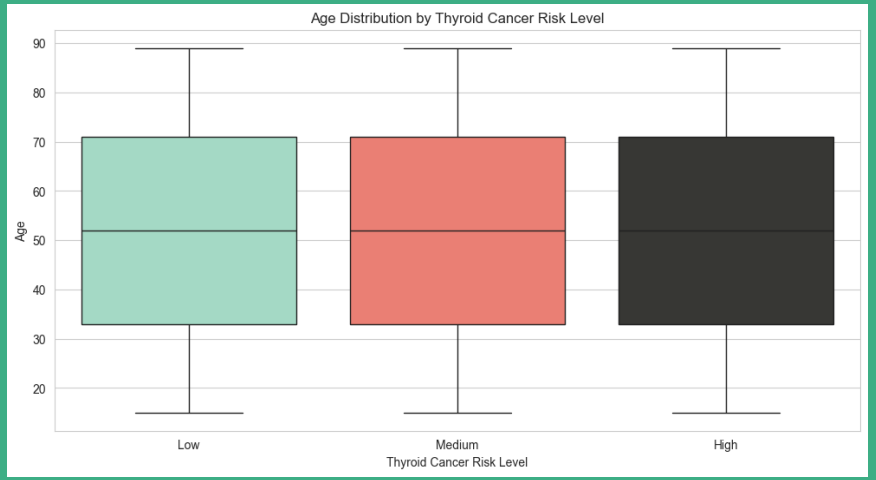
[21] ✓ Python

Age: 50

Gender: Male

```
'For Age: 50, Gender: Male'

'Probability of Low Risk: 0.5196'

'Probability of Medium Risk: 0.1129'

'Probability of High Risk: 0.3675'
```


Thyroid Cancer Risk Distribution by Gender
Gender
Male
Female
Count
60000
50000
40000
30000
20000
10000
0
Low
Medium
High
Thyroid Cancer Risk Level


Age Distribution by Thyroid Cancer Risk Level
Age
90
80
70
60
50
40
30
20
Low
Medium
High
Thyroid Cancer Risk Level

Insights & Finding

Key Data Insights

- Caucasians are the most represented group, with 63,669 entries, while the Middle Eastern individuals are the least represented, with only 21,335 entries, Asians with 53,261 entries, Africans with 42,414 entries, and Hispanics with 32,012 entries, showing an uneven distribution
- Females are more likely to have thyroid cancer especially Asian females
- Chi-square test p-value = 0.5102 ($p > 0.05$), which means no statistically significant difference in the diagnosis proportions across the genders
- Chi-squared test p-value for age groups: 0.5586, analysis does not show any bias in thyroid cancer diagnosis based on age groups within this dataset
- Most countries show a similar risk distribution: Low risk (~53-54%) with most common category, Medium risk (~35-36%) follows, High risk (~10-11%) is the least frequent in most countries. This suggests that thyroid cancer risk is generally low in most global regions
- India Shows a Unique Risk Pattern High-risk individuals make up 32.86% of the population (significantly higher than other countries)
- Japan (10.06%) and South Korea (10.50%) have the lowest high-risk proportions

Collaboration & Outcomes

Outcomes

Are you happy with the final product?
Yes

What do you hope to achieve in the next development cycle?

Fully utilize the dataset to its full potential

What would you do differently if you could start again?

We identified the bias but didn't take any action on it, we left the dataset as it was because of time constraints and no experience with handling this situation.

Development Problems

Problems that arose during development?:

Git collaboration

In group conflicts and resolutions?
No

Did you find any of the behaviour related content useful? Teamwork, problem solving etc?

Yes, mindful collaboration and problem- solving, work division

Interactivity: Overall good

Summary

Overall group dynamic:
Good, friendly, professional

Overall satisfaction: 9/10 (-1 for git)

What we learned: github collaboration

Our experiences: We had many issues with git commands but we had good troubleshooting and good mindset for the project

Q&A