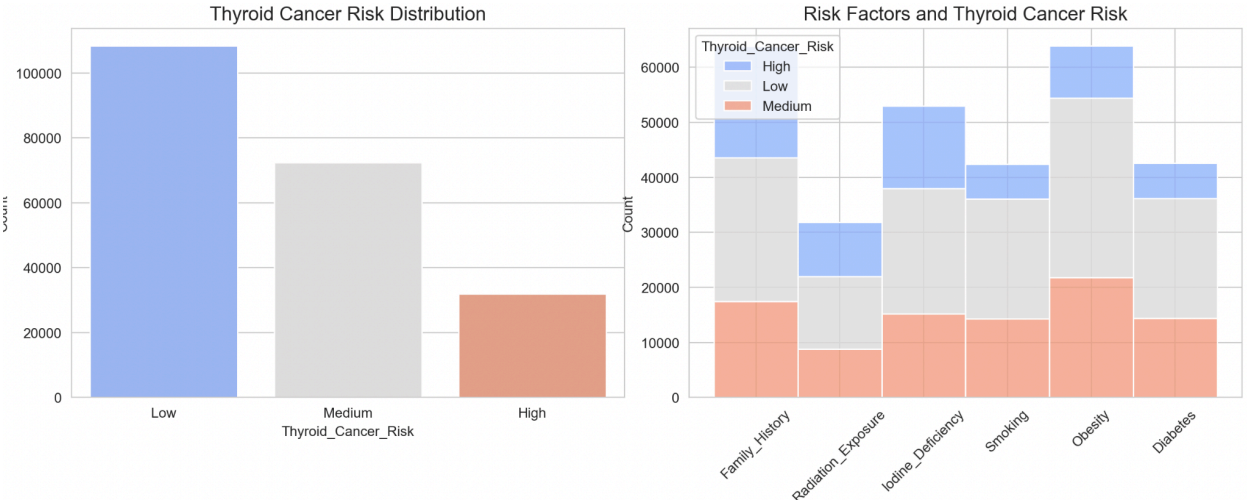
Jerek Stegman
Visualization Final
Data Visualization
March 12, 2025

I. Thyroid Cancer Risk Analysis

II.

Figures 1 and 2.



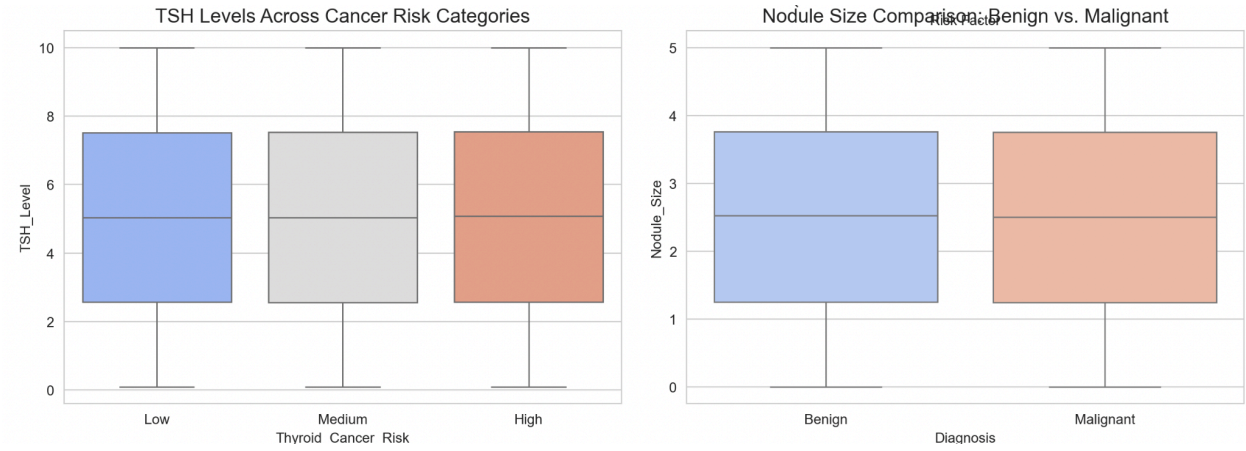
(Figure 1)

- **Bars** represent the count of individuals who are categorized into three risk groups: low risk, medium risk, and high risk of thyroid cancer.
- **Colors** represent risk level - blue = low, grey = medium, orange = high
- **X-Axis** - Thyroid Cancer Risk Category
- **Y-Axis** - Count of Individuals

(Figure 2)

- **Stacked Bars** represent each risk factor, which is further broken down into cancer risk levels.
- **Colors** represent risk level - blue = low, grey = medium, orange = high
- **X-Axis** - Risk Factors
- **Y-Axis** - Count of individuals associated with a risk factor and risk level

Figures 3 and 4.



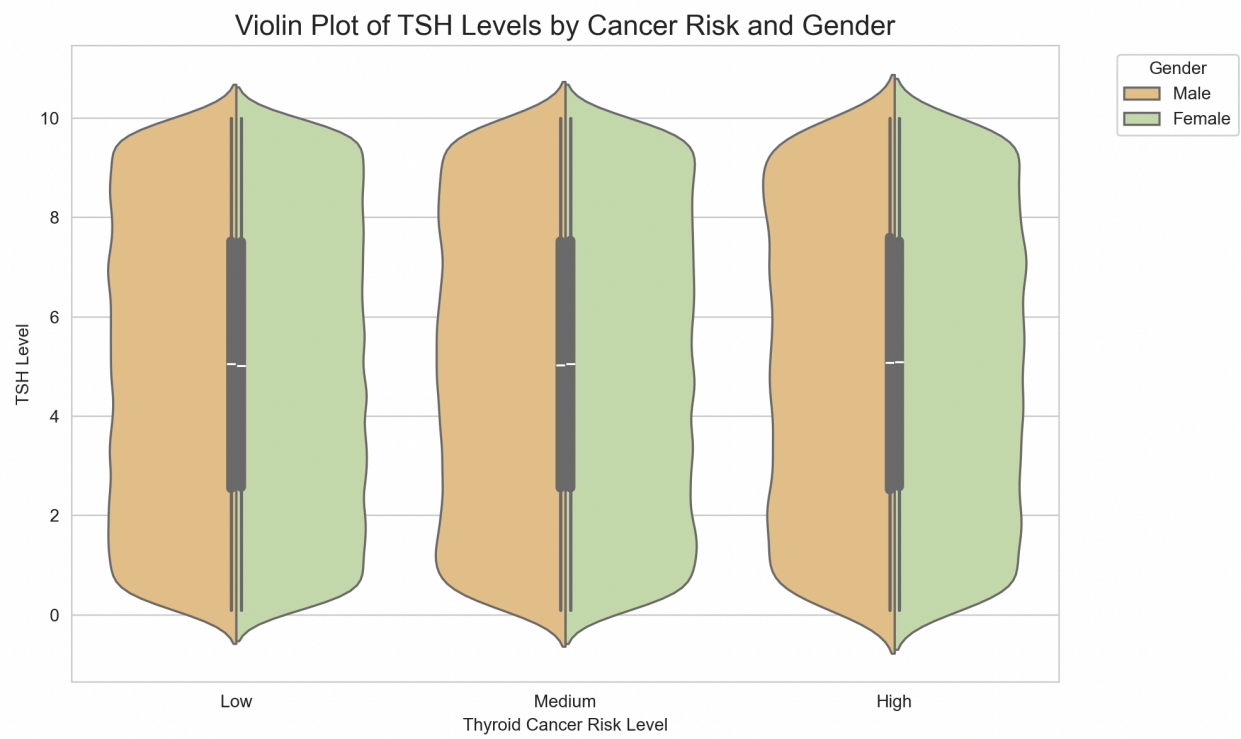
(Figure 3)

- **Boxplots** represent the distribution of Thyroid-Stimulating Hormone (TSH) levels associated with low, medium, and high risk groups.
- **Colors** represent risk level - blue = low, grey = medium, orange = high
- **X-Axis** - Thyroid Cancer Risk Categories
- **Y-Axis** - TSH Levels
- **Other Details** - Boxplots are used in this visualization to show median, quartiles, and outliers of TSH levels across different risk groups.

(Figure 4)

- **Boxplots** represent a comparison of nodule sizes between benign and malignant diagnoses.
- **Colors** - blue = Benign, orange = malignant
- **X-Axis** - Diagnosis
- **Y-Axis** - Nodule Size - assumed unit is centimeters
- **Other Details** - the medians and spread help indicate larger nodules in malignant cases than in benign cases.

Figure 5.



(Figure 5)

- **Violin Plots** represent the distribution of TSH levels within each cancer risk category and are split according to gender.
- **Colors** - orange = Male, green = Female
- **X-Axis** - Thyroid Cancer Risk Level
- **Y-Axis** - TSH Levels
- **Other Details** - The inner boxplots show the medians and interquartile ranges, and the violin shows the distribution of density

III. Findings

The visualizations created in this project provide key insights into many different areas, including:

- Risk Distribution
    - The majority of individuals fall into the low-risk category, and a small number fall into the high-risk
- Risk Factor Association
    - Family history, obesity, and iodine deficiency show a higher correlation to high cancer risk levels than other lifestyle factors.
- TSH Level Patterns
    - The levels of TSH actually appear consistent between risk groups.
- Nodule Size Difference
    - Malignant nodules tend to appear larger than benign nodules, which ould serve as a diagnostic indicator
- Gender Differences
    - TSH trends remain consistent between gender groups.

IV. Data and Method

Data

The data set used for this project was pulled from Kaggle and consisted of multiple variables in thyroid cancer diagnoses. The dataset was created in a way to replicate real-world cancer research.

Methods

- Preparation
    - Cleaning and organizing the data using Python Libraries
    - Standardized variable names
- Visualization
    - Use of common and aesthetically pleasing visualizations found within several Python libraries
- Tools Used
    - Python
    - Libraries: Pandas, Matplotlib, Seaborn, Numpy

V. Importance of Data and Analysis

This project and the analysis it presents through data visualization can help aid in a visual examination of medical risks. This project can help identify groups of people at elevated risk of Thyroid cancer, understand TSH levels and nodule size, and how they relate to cancer diagnoses, and recognize gender based differences in cancer diagnoses.