YIGA GILBERT

2024/HD05/21947U

2400721947

MSCS.

MCS 7103: Machine Learning

Exploratory Data Analysis Report

Key Analytical Questions for Dataset Evaluation

1. Data Quality and Completeness

- To what extent is the data complete? Are there any gaps in important columns such as lesion size or SCJ visibility, and how should missing data be addressed?

- Is the data consistent across different variables? Do all categorical columns maintain a uniform structure, or are there anomalies or irregularities?

- Does the dataset present a balanced distribution? For instance, do we observe an equal representation of positive and negative VIA assessments, varying image quality ratings, or SCJ visibility outcomes?

2. Image Quality

- What does the distribution of image quality ratings look like? Are the majority of images rated as "good," "poor," or "excellent," and is there any noticeable variation?

- How does the quality of the images affect their suitability for diagnostic purposes? Are images with lower quality less likely to be deemed diagnostically useful?

- Is there a connection between image quality and the accuracy of VIA assessments? Do better-quality images tend to lead to more accurate VIA outcomes?

3. VIA Assessment and Diagnostic Suitability

- What is the distribution of VIA assessment outcomes (positive vs. negative)? How frequently do we see positive assessments compared to negative ones?

- How do other variables, such as SCJ visibility, lesion size, or image quality, influence VIA assessments?

- What factors seem to contribute to positive VIA outcomes? Can any patterns be observed in the images or lesion sizes that correlate with positive assessments?

4. Lesion Size and Diagnosis

- How are lesion sizes distributed throughout the dataset? Are there common lesion sizes, and do the lesions tend to be larger or smaller in general?

- Does the size of the lesion impact VIA outcomes? For example, are larger lesions more likely to result in positive assessments, or is there a specific threshold that correlates with the likelihood of a positive diagnosis?

- How is lesion size related to image quality or SCJ visibility? Do larger lesions pose difficulties for capturing high-quality images or fully visible SCJ?

5. SCJ Visibility

- How frequently is the SCJ fully visible in the images, and does this visibility vary significantly?

- Does the visibility of the SCJ affect diagnostic suitability? Are images with fully visible SCJ more likely to be rated as suitable for diagnosis?

- What factors contribute to partial SCJ visibility? Is this issue related to poor image quality or larger lesion sizes?

6. General Trends and Patterns

- Are there any noticeable trends in the data over time, assuming there is a time-related component? For example, has the overall quality of the images or the accuracy of diagnoses changed over time?

- Are there any outliers or unusual patterns in the dataset? For example, are there specific cases where lesion sizes, images, or assessments deviate from the expected trends?

7. Recommendations for Improvement

- Based on the analysis, how can the overall quality of the images in the dataset be improved?

- What are the most critical factors for ensuring accurate VIA assessments? Which areas—such as image quality, lesion size, or SCJ visibility—should be prioritized to improve diagnostic accuracy?

- What strategies could be employed to address missing data, particularly for important fields such as lesion size?

8. Predictive Modeling Potential

- Can the data be leveraged to build a predictive model that assesses VIA outcomes based on variables such as image quality, lesion size, and SCJ visibility?

- What are the most important features that could help predict whether a VIA assessment will be positive or negative?

9. Comparative Analysis

- How does lesion size relate to VIA assessment results? Is there a consistent pattern where larger lesions correspond to positive VIA assessments, or does the proportion of the cervix area involved serve as a reliable predictor?

Introduction and Overview

The dataset consists of 371 rows and 10 columns, including critical categorical features like image quality, SCJ visibility, VIA assessments, and lesion size.
This report seeks to explore the relationships between these variables, identify key trends, and answer fundamental questions about the data completeness and diagnostic potential.

Overview of Categorical Features

	count	**unique**	**top**	**freq**	**relative freq**
image	371	335	040301_4461_20221102.jpg	2	0.54%
What is the quality of the picture?	371	2	good	265	71.43%
Is the quality of the picture good enough to make a diagnosis?	371	2	yes	335	90.30%
Is SCJ fully visible	337	3	fully visible	253	68.19%
What is the VIA assessment?	336	5	VIA negative	275	74.12%
What is the size of lesion (propotion of cervix area involved)?	48	4	1 quadrant (25%)	18	4.85%
Quality_Encoded	371	2	0	265	71.43%
Diagnosis_Encoded	371	2	1	335	90.30%
SCJ_Encoded	371	4	0	253	68.19%
VIA_Encoded	371	6	0	275	74.12%

1. Univariate Analysis

Univariate analysis focuses on the distribution of each variable in the dataset. By analyzing each variable individually, we gain insight into its characteristics.
For categorical variables, we look at how frequently each value occurs and assess data completeness.

image: This variable identifies the images in the dataset. Each image is used for diagnostic purposes. Most images are unique, though there are a few duplicates.

What is the quality of the picture?: Image quality is critical for making accurate diagnoses. A high percentage of the images (71.43%) are of 'good' quality, which is encouraging.

Is the quality of the picture good enough to make a diagnosis?: This field shows that 90.30% of the images are deemed suitable for diagnosis, suggesting that image quality generally meets diagnostic standards.

Is SCJ fully visible: SCJ visibility is essential for making reliable diagnoses. In this dataset, 68.19% of the images have fully visible SCJ, while 31.81% either do not or have missing data.

What is the VIA assessment?: VIA assessment indicates whether a visual inspection results in a positive or negative diagnosis. Most assessments (74.12%) are negative.

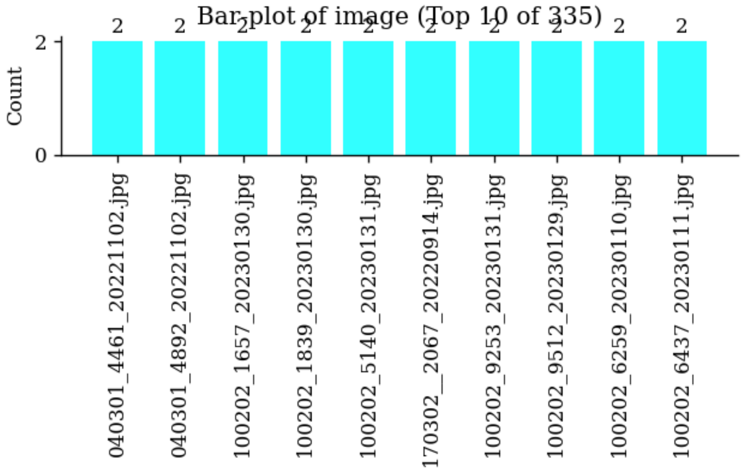
What is the size of lesion (proportion of cervix area involved)?: Lesion size is only reported in a small portion of cases (48 entries), with most indicating small areas of involvement.

1.1 Image

Image is a categorical variable with 335 unique values. None of its values are missing.

Summary Statistics

Mode (Most frequent)	060201_8214_20230220.jpg
Maximum frequency	2

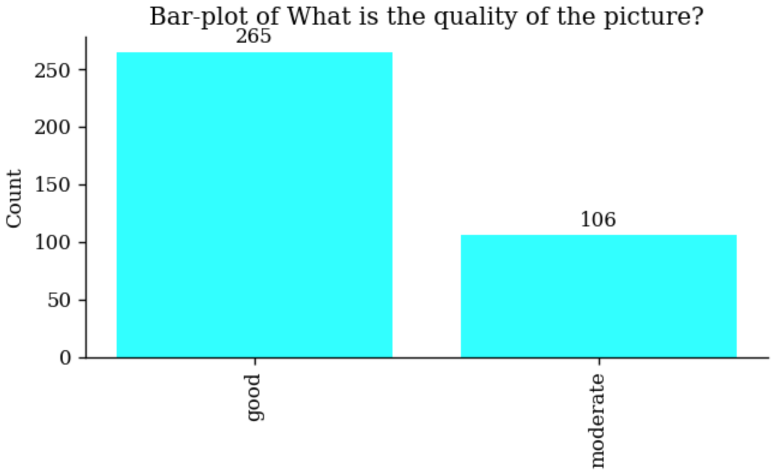


1.2 What Is The Quality Of The Picture?

What is the quality of the picture? is a categorical variable with 2 unique values. None of its values are missing.

Summary Statistics

Mode (Most frequent)	good
Maximum frequency	265

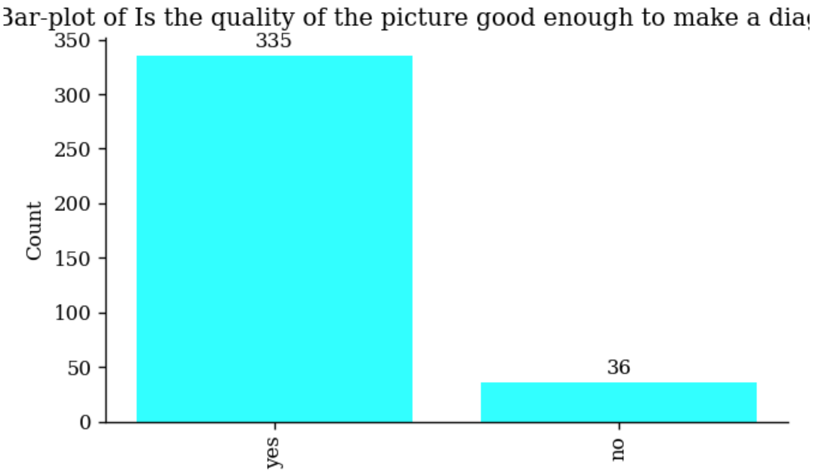


1.3 Is The Quality Of The Picture Good Enough To Make A Diagnosis?

Is the quality of the picture good enough to make a diagnosis? is a categorical variable with 2 unique values. None of its values are missing.

Summary Statistics

Mode (Most frequent)	yes
Maximum frequency	335

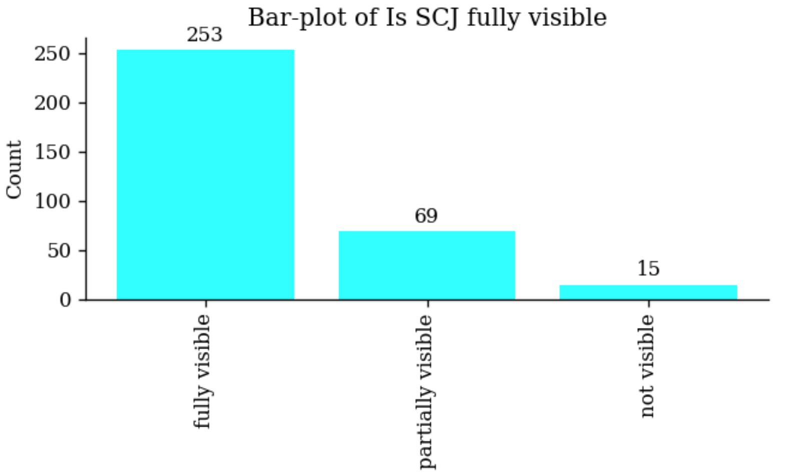


1.4 Is Scj Fully Visible

Is scj fully visible is a categorical variable with 3 unique values. 34 (9.16%) of its values are missing.

Summary Statistics

Mode (Most frequent)	fully visible
Maximum frequency	253

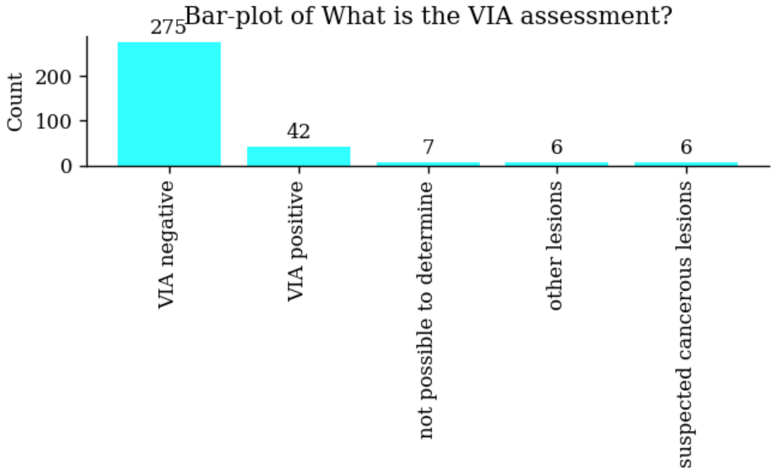


1.5 What Is The Via Assessment?

What is the via assessment? is a categorical variable with 5 unique values. 35 (9.43%) of its values are missing.

Summary Statistics

Mode (Most frequent)	VIA negative
Maximum frequency	275

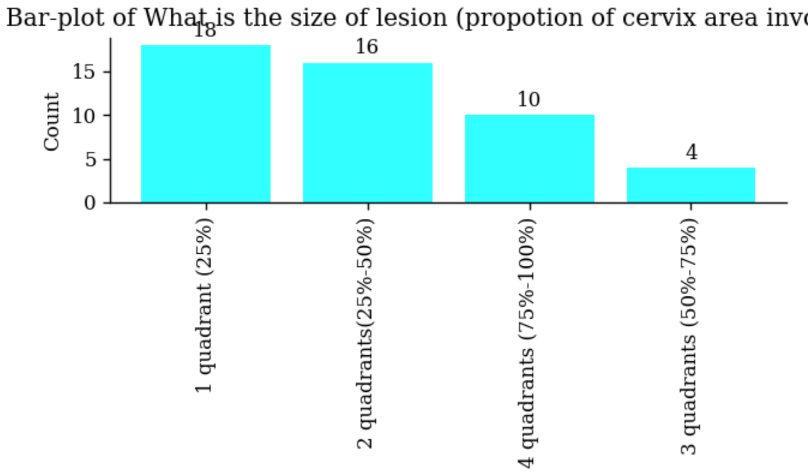


1.6 What Is The Size Of Lesion (Propotion Of Cervix Area Involved)?

What is the size of lesion (propotion of cervix area involved)? is a categorical variable with 4 unique values. 323 (87.06%) of its values are missing.

Summary Statistics

Mode (Most frequent)	1 quadrant (25%)
Maximum frequency	18

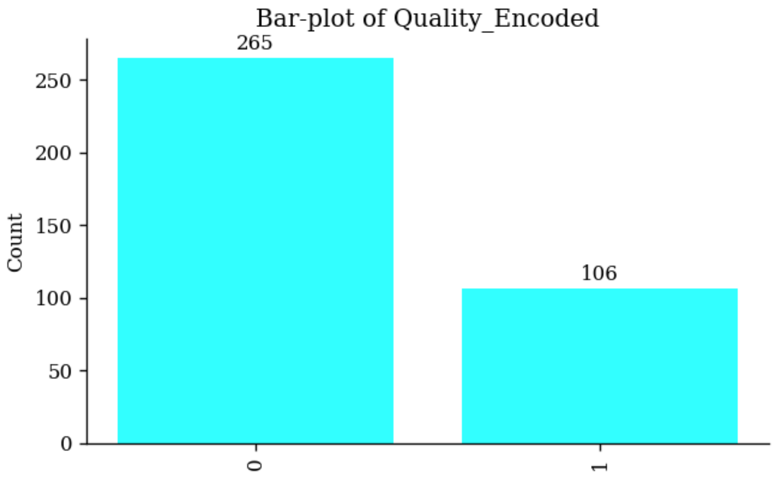


1.7 Quality_Encoded

Quality_encoded is a boolean variable with 2 unique values. None of its values are missing.

Summary Statistics

Mode (Most frequent)	0
Maximum frequency	265

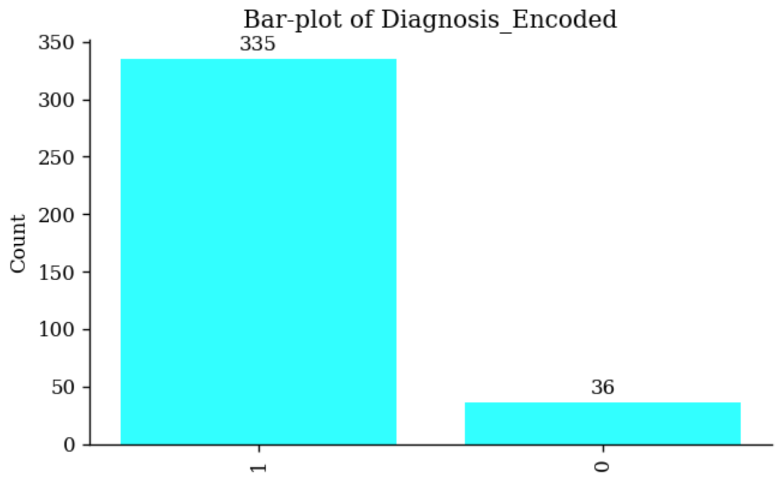


1.8 Diagnosis_Encoded

Diagnosis_encoded is a boolean variable with 2 unique values. None of its values are missing.

Summary Statistics

Mode (Most frequent)	1
Maximum frequency	335

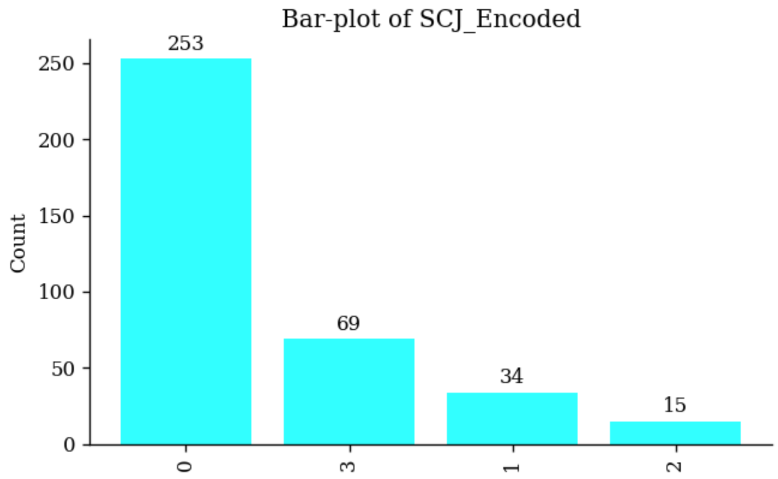


1.9 Scj_Encoded

Scj_encoded is a numeric (<=10 levels) variable with 4 unique values. None of its values are missing.

Summary Statistics

Mode (Most frequent)	0
Maximum frequency	253

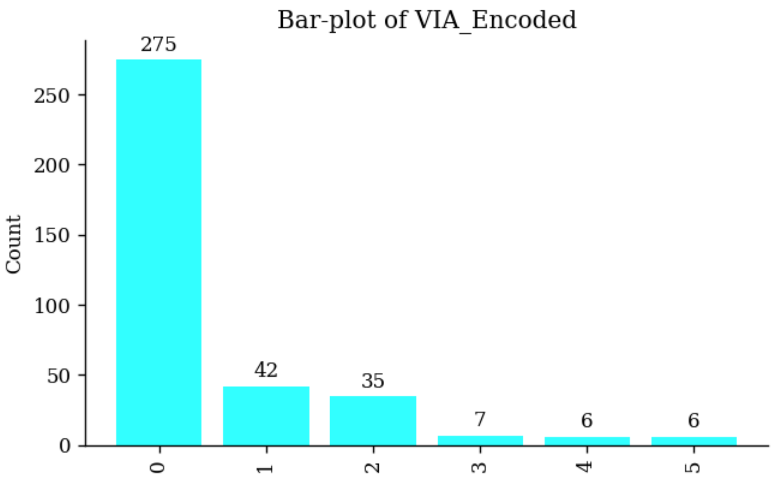


1.10 Via_Encoded

Via_encoded is a numeric (<=10 levels) variable with 6 unique values. None of its values are missing.

Summary Statistics

Mode (Most frequent)	0
Maximum frequency	275



2. Bivariate Analysis

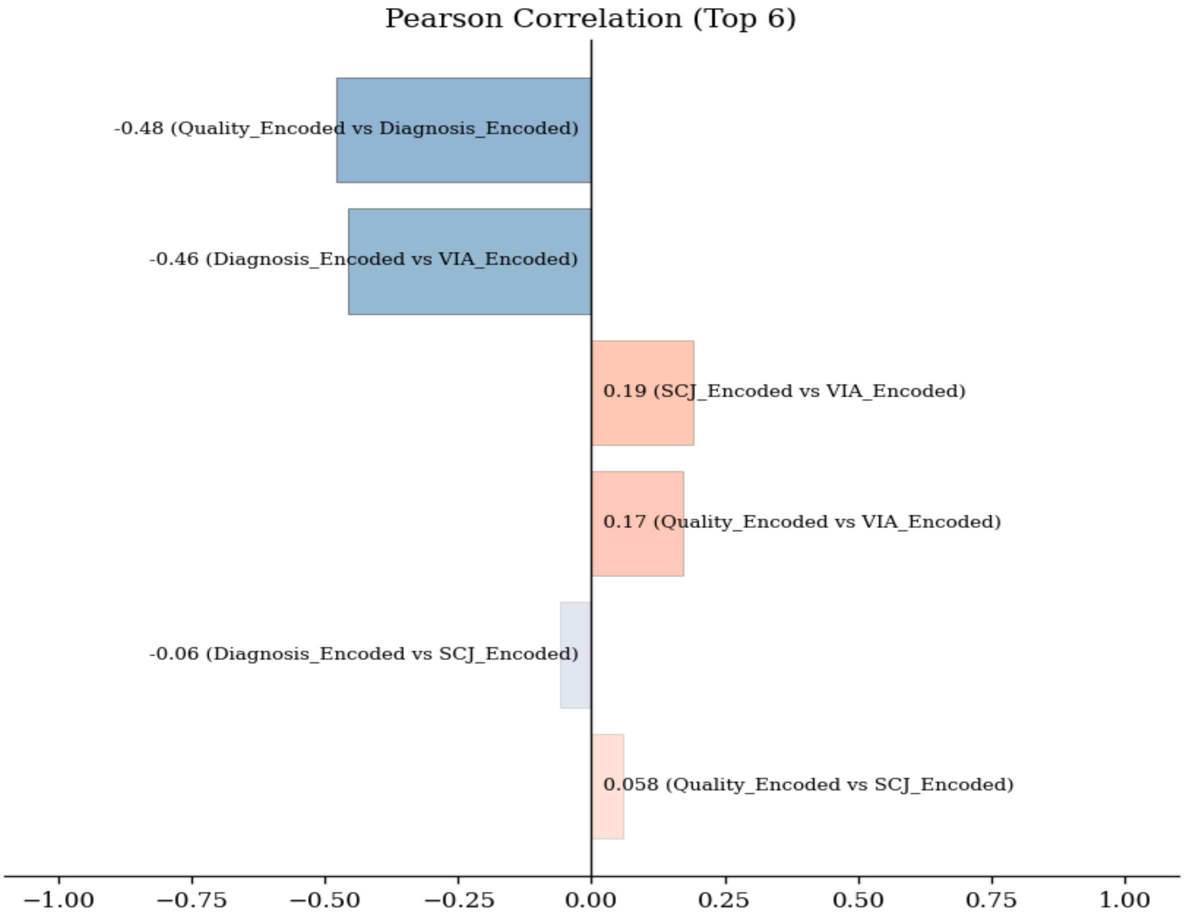
Bivariate analysis examines the relationships between two variables. This can help reveal correlations and dependencies, providing a deeper understanding of the interactions within the dataset.
We performed a correlation analysis between key variables and identified relationships that impact diagnosis outcomes.

Quality_Encoded vs. Diagnosis_Encoded: A moderate negative correlation (-0.48) was found, indicating that lower-quality images are less likely to be suitable for diagnosis.

Diagnosis_Encoded vs. VIA_Encoded: A moderate negative correlation (-0.46) suggests that as the image quality decreases, VIA-positive assessments are less frequent.

SCJ_Encoded vs. VIA_Encoded: A very weak positive correlation (0.19) was found, indicating a slight relationship between SCJ visibility and positive VIA outcomes.

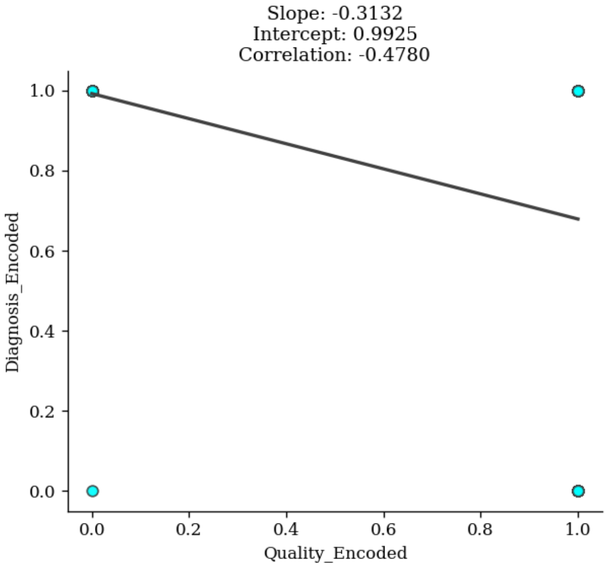
2.1 Overview



2.2 Regression Plots (Top 20)

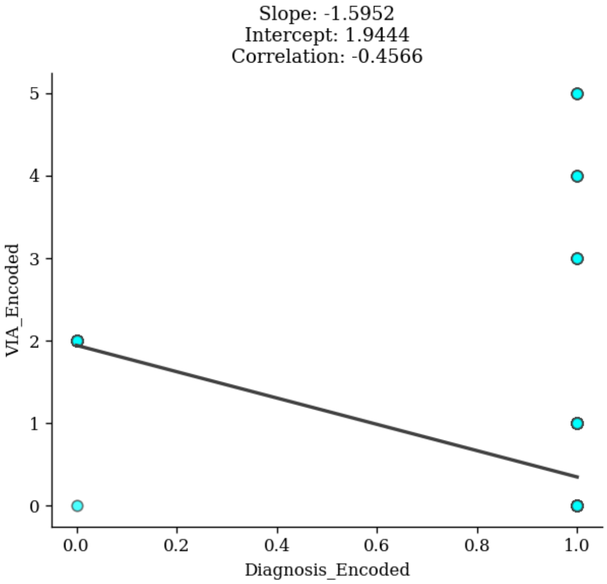
2.2.1 Quality_Encoded Vs Diagnosis_Encoded

Quality_Encoded and Diagnosis_Encoded have moderate negative correlation (-0.48).



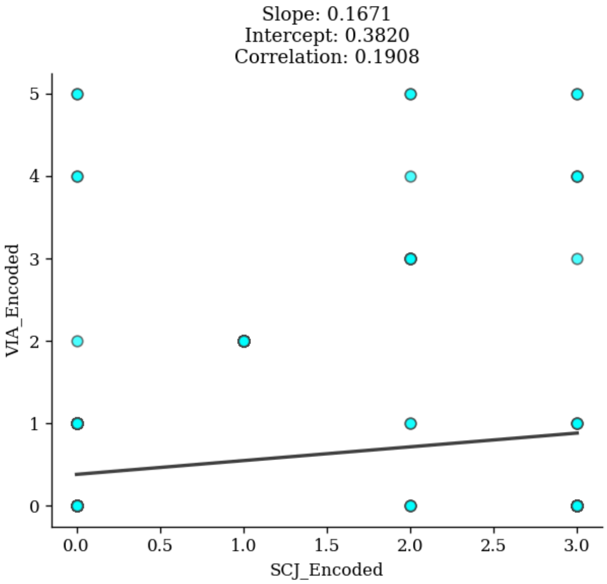
2.2.2 Diagnosis_Encoded Vs Via_Encoded

Diagnosis_Encoded and Via_Encoded have moderate negative correlation (-0.46).



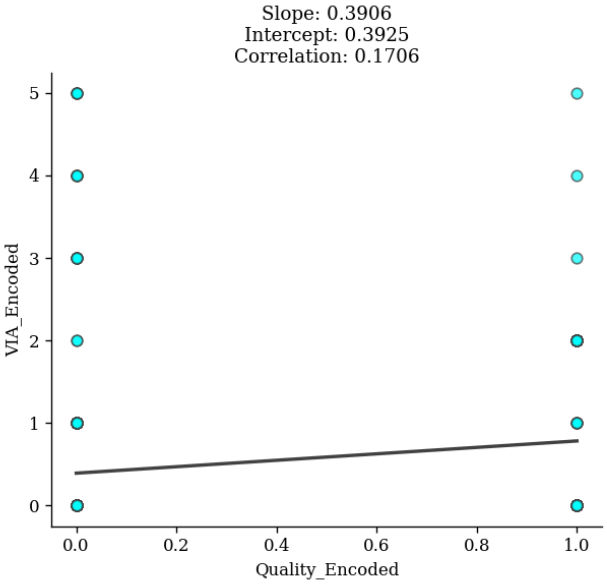
2.2.3 Scj_Encoded Vs Via_Encoded

Scj_Encoded and Via_Encoded have very weak positive correlation (0.19).



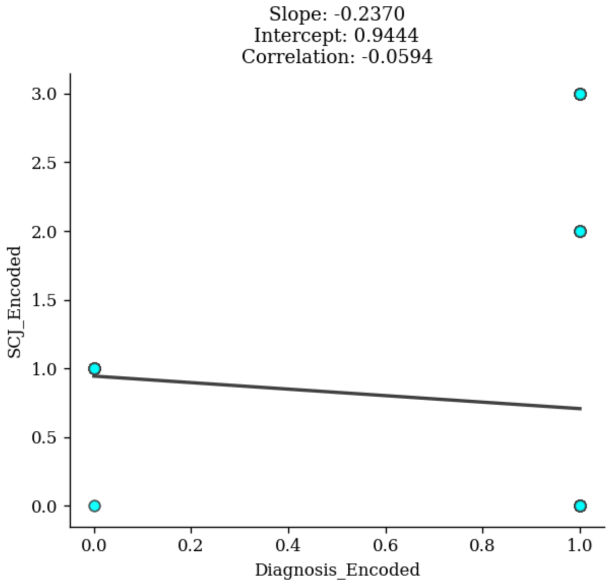
2.2.4 Quality_Encoded Vs Via_Encoded

Quality_Encoded and Via_Encoded have very weak positive correlation (0.17).



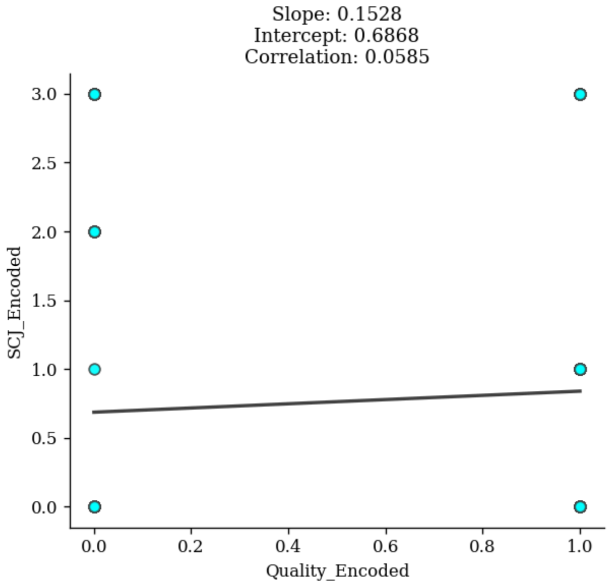
2.2.5 Diagnosis_Encoded Vs Scj_Encoded

Diagnosis_Encoded and Scj_Encoded have very weak negative correlation (-0.06).



2.2.6 Quality_Encoded Vs Scj_Encoded

Quality_Encoded and Scj_Encoded have very weak positive correlation (0.06).



Graph Explanations

Graphs are essential for visualizing data trends. In this section, we explain each graph included in the analysis and what insights they provide.

1. Distribution of Image Quality: This bar chart shows that most images are of good quality, which is critical for ensuring diagnostic accuracy. Poor-quality images make up a minority.

2. Distribution of VIA Assessments: The pie chart shows that most VIA assessments are negative. This could be due to the type of images being evaluated, or it may suggest that further follow-up is required for positive cases.

3. Quality vs. Diagnosis: This scatter plot reveals a clear negative relationship between image quality and diagnosis suitability, confirming the importance of high-quality images for accurate diagnosis.

4. SCJ Visibility vs. VIA Assessment: This plot shows that SCJ visibility slightly correlates with positive VIA assessments, suggesting that fully visible SCJ may aid in making accurate diagnoses.