ADVANCED EDA
FOR GENOMIC DATA ANALYSIS

PROJECT REPORT

Abstract

Exploratory Data Analysis (EDA) plays a crucial role in understanding genomic datasets, identifying patterns, and uncovering insights before applying predictive modeling. This project focuses on Advanced EDA techniques applied to genomic data, ensuring a comprehensive structure, quality assessment, and feature exploration. The analysis is implemented as an interactive web application using **Streamlit**, allowing users to dynamically explore and visualize the data.

Motivation

Genomic data analysis is a vital component of bioinformatics, helping researchers discover genetic markers, mutations, and correlations in disease prediction. However, handling large-scale genomic datasets requires **robust preprocessing, visualization, and interactive analysis tools**. This project aims to bridge the gap between raw genomic datasets and meaningful insights through an **Advanced EDA framework**.

Problem Statement

The key challenges in genomic data analysis include:

- Handling missing, duplicate, or erroneous data entries.
- Understanding the relationships between various genomic features.
- Identifying skewed distributions and outliers in genomic features.
- Providing an **interactive** and **user-friendly** interface for dynamic exploration of genomic datasets.

To address these challenges, this project integrates **automated EDA, interactive visualizations, and statistical summaries** within a Streamlit-based web application.

Literature Survey

Several studies highlight the importance of **data preprocessing and visualization in genomics**:

- **Tukey (1977)** introduced the concept of EDA as a fundamental step before modeling.
- **Biecek & Kosinski (2017)** discussed the application of EDA in large-scale biological datasets.
- **PCA & t-SNE (Van der Maaten, 2008)** have been widely used for dimensionality reduction in genomic data.
- **Python Libraries (Pandas, Matplotlib, Seaborn, Plotly)** offer powerful visualization tools to explore complex datasets.

This project builds upon these methodologies, incorporating **interactive analysis** to provide deeper insights into genomic data.

Dataset Exploration

The dataset used in this project is **Breast Cancer Genomic Data**, containing multiple features related to tumor characteristics and patient demographics. Key aspects analyzed include:

- **Structure Investigation:**
    - Shape of dataset
    - Data types of each feature
    - Summary statistics
- **Quality Investigation:**
    - Handling duplicate values
    - Identifying missing values
    - Detecting inconsistent entries
- **Feature Analysis:**
    - Distribution of numerical and categorical features
    - Feature relationships (correlation matrix, heatmaps, pair plots)
    - Outlier detection and handling

Project Pipeline

The project follows a structured **EDA pipeline**:

1. **Data Ingestion** – Loading and previewing the dataset.
2. **Data Cleaning** – Handling missing values, duplicates, and data inconsistencies.
3. **Feature Exploration** – Statistical summaries, distributions, and relationships.
4. **Visualization** – Interactive charts, histograms, boxplots, heatmaps, and PCA analysis.
5. **Interactivity** – Streamlit-based web app for dynamic user exploration.
6. **Report Generation** – Automated reporting of insights and findings.

Framework Module

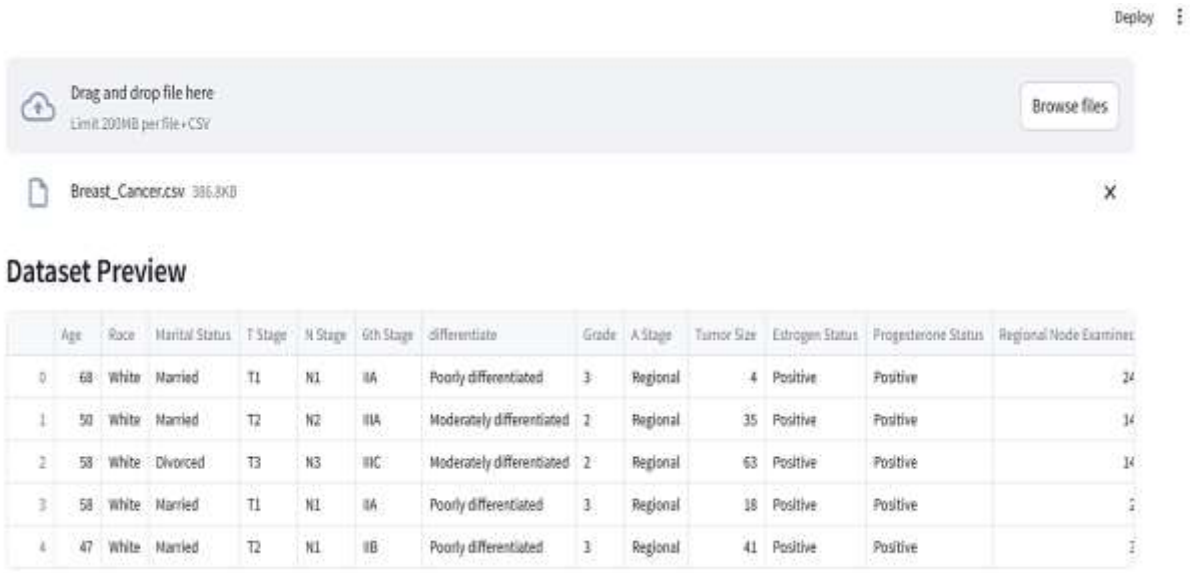
The framework consists of the following **core modules**:

- **Data Loader**: Reads the dataset into a Pandas DataFrame.
- **Preprocessing Module**: Handles missing values, duplicates, and unwanted entries.
- **Visualization Module**: Generates dynamic plots using Matplotlib, Seaborn, and Plotly.
- **Feature Engineering Module**: Computes correlations, outliers, and feature patterns.
- **Streamlit Interface**: Provides a user-friendly interactive experience for analysis.

Snapshots

1. Web Application Interface.

2. Dataset Overview



	Age	Race	Marital Status	T Stage	N Stage	6th Stage	differentiate	Grade	A Stage	Tumor Size	Estrogen Status	Progesterone Status	Regional Node Examined
0	68	White	Married	T1	N1	IIA	Poorly differentiated	3	Regional	4	Positive	Positive	24
1	50	White	Married	T2	N2	IIIA	Moderately differentiated	2	Regional	35	Positive	Positive	14
2	58	White	Divorced	T3	N3	IIIC	Moderately differentiated	2	Regional	63	Positive	Positive	14
3	58	White	Married	T1	N1	IIA	Poorly differentiated	3	Regional	18	Positive	Positive	2
4	47	White	Married	T2	N1	IIB	Poorly differentiated	3	Regional	41	Positive	Positive	3

Dataset Structure

Shape of Dataset: `(4024, 16)`

Data Types:

	0
Age	int64
Race	object
Marital Status	object
T Stage	object
N Stage	object
6th Stage	object
differentiate	object
Grade	object
A Stage	object
Tumor Size	int64

Data Quality Check

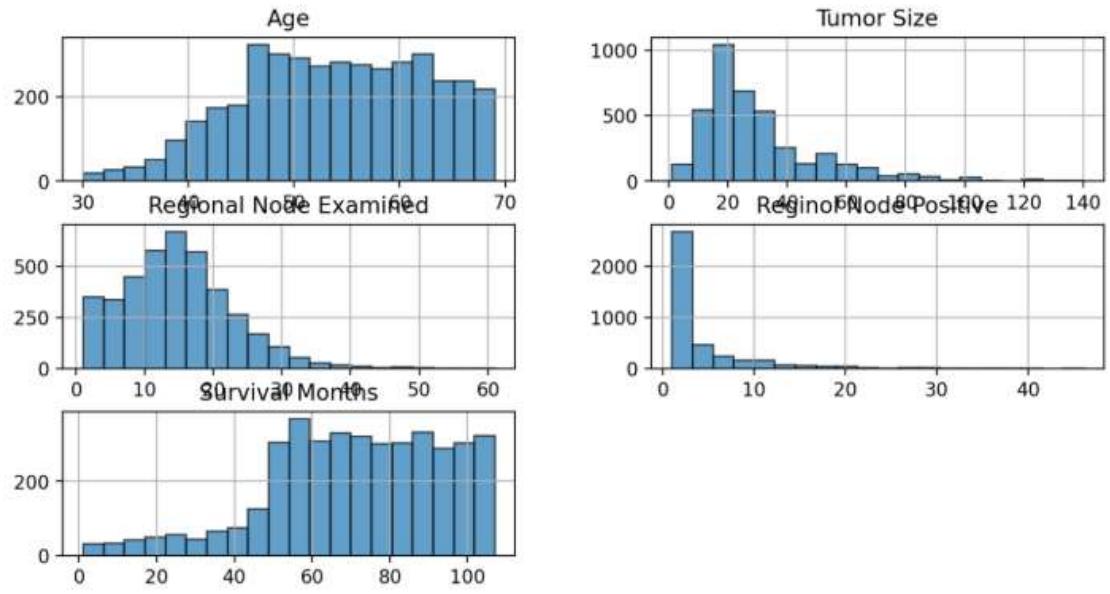
Missing Values:

	0
Age	0
Race	0
Marital Status	0
T Stage	0
N Stage	0
6th Stage	0
differentiate	0
Grade	0
A Stage	0
Tumor Size	0

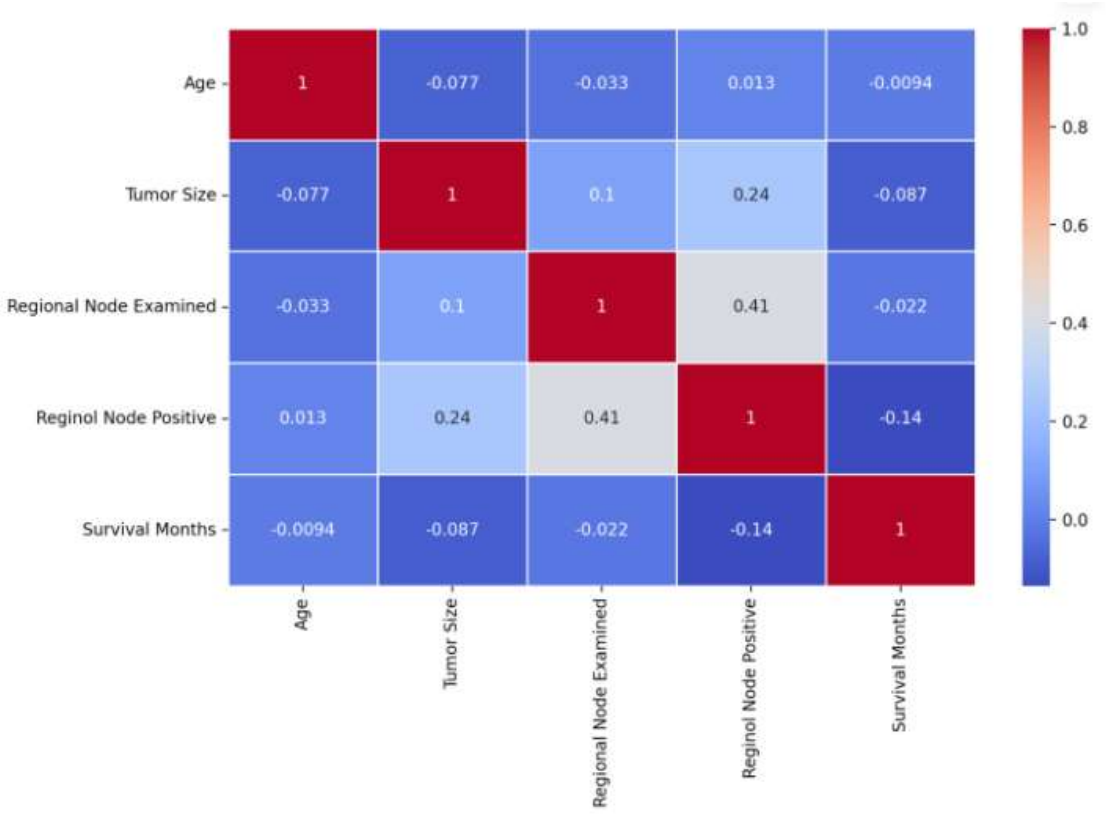
Duplicate Rows: `1`

3. Feature Distributions

Feature Distribution



4. Correlation Heatmap



5. Data Summary Statistics

Data Summary Statistics

	Age	Tumor Size	Regional Node Examined	Reginol Node Positive	Survival Months
count	4,024	4,024	4,024	4,024	4,024
mean	53.9722	30.4737	14.3571	4.1581	71.298
std	8.9631	21.1197	8.0997	5.1093	22.9214
min	30	1	1	1	1
25%	47	16	9	1	56
50%	54	25	14	2	73
75%	61	38	19	5	90
max	69	140	61	46	107

EDA Completed Successfully!

Conclusion

This project successfully demonstrates **Advanced EDA techniques** for genomic data using an interactive Streamlit web application. Key takeaways include:

- **Comprehensive Data Exploration:** Statistical summaries, distributions, and feature relationships.
- **Interactive & User-Friendly:** Streamlit-based UI for seamless data analysis.
- **Scalability:** The framework can be extended to analyze other genomic datasets.

Future work includes integrating **machine learning models** for predictive analysis and expanding the application to support **larger genomic datasets**.