Summary and Recommendations

As part of my learning in data analytics and healthcare informatics, I completed an **exploratory data analysis (EDA)** project based on a comprehensive cancer patient dataset from **Denmark**. This dataset consisted of **3,000 patient records**, each containing detailed information about cancer types, stages, treatment plans, lifestyle factors, genetic conditions, and survival durations.

🎯 Project Objective

The primary goal of this project was to:

- Understand how **demographic** and **clinical factors** affect cancer survival.
- Explore **regional trends** in cancer types and outcomes across Denmark.
- Analyze how variables like **BMI**, **smoking status**, **cancer stage**, and **treatment types** correlate with patient survival and relapse.
- Practice data wrangling and visualization techniques to tell a compelling story with healthcare data.

🛠️ Tools & Technologies

To carry out the analysis, I used the following tools:

- **Python (Jupyter Notebook)**: For all data preprocessing, exploration, and visualization.
- **Pandas**: To clean, filter, and manipulate data efficiently.
- **Matplotlib & Seaborn**: For creating clear, interpretable visualizations including histograms, box plots, heatmaps, and bar charts.

📊 Key Explorations & Insights

🔹 **Cancer Type Distribution**:
Certain cancers like breast, lung, and prostate were more common in specific Danish regions, suggesting geographical influences or demographic clustering.

🔹 **Survival by Stage & Treatment**:
Patients diagnosed at early stages (Stage I & II) had significantly longer survival durations than those at later stages. Treatment types like immunotherapy and chemotherapy showed different survival patterns depending on the cancer type and patient profile.

🔹 **Lifestyle Impact**:

- **Smoking Status** had a noticeable effect on survival, with current smokers showing reduced survival months.
- **BMI** extremes (underweight or obese) also impacted patient outcomes negatively.

🔹 **Correlation Matrix**:
A heatmap of numeric features revealed strong correlations between tumor size, number of positive lymph nodes, and survival duration.

🔹 **Region-Wise Survival**:
Average survival months varied by region, possibly reflecting differences in healthcare access, early detection programs, or environmental factors.
