Technical Report: GLOBAL CANCER PATIENT REPORT (Microsoft Excel Analysis Project)

Outline

This report presents a detailed analysis of the Global Cancer Patient dataset using Microsoft Excel. The goal is to uncover meaningful insights about cancer prevalence, demographics, survival rates, and patterns across countries and genders.

It follows a comprehensive structure:

1. Introduction
2. Story of Data
3. Data Splitting and Preprocessing
4. Pre-Analysis
5. In-Analysis
6. Post-Analysis and Insights
7. Data Visualizations & Charts
8. Recommendations and Observations
9. Conclusion

Introduction

The purpose of this analytical project is to understand global cancer distribution and patient characteristics, exploring patterns related to age, gender, and geographic factors. The project also aims to identify countries and demographics with higher cancer burdens to inform healthcare resource allocation and awareness initiatives.

Problem Being Addressed

Cancer is one of the world's most significant public health challenges. However, raw datasets often obscure the key stories behind patient outcomes and disease concentration. This analysis uses Excel to:

- Uncover demographic and geographic cancer patterns.
- Highlight gender disparities in cancer cases.
- Offer data-backed insights for global health strategies.

Key Datasets and Methodologies

The dataset includes:

- Patient ID
- Age
- Gender
- Country

- Cancer Type
- Survival Status
- Diagnosis Date

Excel methodologies used:

- **PivotTables** to summarize cancer type by gender and country.
- **Statistical formulas** (AVERAGE, COUNTIF, PERCENTILE) to compute survival averages and rates.
- **Charts** (Bar, Pie, Histogram) to visualize demographic patterns.
- **Conditional Formatting** to highlight critical age ranges or high-mortality regions.

Story of Data

This dataset captures a decade-long view (2015–2024) of global cancer patients. It includes patient demographics like age group and gender, medical aspects like cancer type, stage, and genetic risk, as well as cost-related elements such as treatment expenses and survival years. The data provides a comprehensive understanding of cancer prevalence, costs, and patient outcomes across different regions

Data Source

This data was downloaded from Kaggle.com

Data Structure

Each row represents a unique patient record, and each column corresponds to a specific attribute like age, gender, or cancer type.

| Patient ID | Age | Gender | Country | Cancer Type | Diagnosis Date | Survival Status |

Important Features and Their Significance

- **Age**: Indicates age-specific vulnerability to certain cancers.
- **Gender**: Helps analyze male vs. female incidence rates.
- **Country**: Reveals regional health disparities and cancer prevalence.
- **Cancer Type**: Determines common and rare cancer categories.
- **Survival Status**: Measures mortality and recovery trends.

Data Splitting and Preprocessing

Data Cleaning

Key steps:

- Removed duplicated patient entries using Data → Remove Duplicates.

- Used `=IF(ISBLANK(Age),AVERAGE(AgeRange),Age)` to fill missing ages.
- Standardized country names using `PROPER()` function.
- Corrected inconsistent gender entries (e.g., "M" vs. "Male").

Handling Missing Values

- Missing **survival statuses** replaced with "Unknown."
- Missing **ages** imputed with median values using `=MEDIAN(AgeRange)`.

Data Transformations

- Created a new field **"Age Group"** using nested `IF` formulas:
  `=IFS(B2:B50001<=30,"Young",B2:B50001<=55,"Mid Age",B2:B50001>=56,"Old")`
- Derived **"Survival Rate (%)"** using:
  `=COUNTIF(SurvivalStatusRange,"Alive")/COUNTA(SurvivalStatusRange)*100`

Data Splitting

- **Dependent Variable:** Survival Status
- **Independent Variables:** Age, Gender, Country, and Cancer Type

Industry Context

This study falls under global healthcare analytics, a sector where Excel-based descriptive analysis is crucial for reporting and policy formulation.

Stakeholders

- **World Health Organizations and Ministries of Health**
- **Hospitals and Oncology Departments**
- **Medical Researchers and Policy Makers**

Value to the Industry

This project demonstrates how structured Excel analysis can transform complex health data into clear, evidence-based insights for cancer prevention, funding prioritization, and global awareness campaigns.

Pre-Analysis Insights

- Mid-age and elderly groups might carry the highest cancer burden, leading to higher treatment costs and lower survival years compared to younger populations.
- Countries with advanced healthcare systems may show higher treatment costs but also longer survival years due to better interventions.
- High genetic risk scores for cancers like colon and prostate could explain higher treatment costs and poorer survival rates without early interventions.

- Gender differences in treatment costs or survival rates may be minimal, but specific cancer types might show gender-linked prevalence patterns.
- Cancers detected at earlier stages are likely to incur lower treatment costs and yield better survival outcomes than late-stage cancers.
- Yearly analysis may reveal cost spikes linked to the introduction of new technologies or therapies that initially raise costs but improve survival years over time.
- Certain countries might show low treatment costs but shorter survival years, indicating gaps in healthcare quality or access.
- High genetic risk in some cancers might not always lead to poor survival outcomes if early screening programs are in place.
- Rising treatment costs over time may reflect increasing healthcare demands, aging populations, or adoption of precision medicine approaches.
- Cancer types like lung cancer may show high prevalence but lower genetic risk, indicating environmental or lifestyle factors drive their occurrence.
- Advanced stages of cancer are likely to require more aggressive and expensive treatments with reduced survival years compared to early-stage diagnoses.

In-Analysis Observations

- Cancer prevalence shows no major gender differences, with males, females, and others contributing nearly equal proportions, indicating cancer risk factors cut across all gender categories.
- The mid-age group bears the highest cancer burden, suggesting lifestyle exposures, occupational hazards, and aging cells likely contribute heavily during this stage of life.
- The elderly group also shows a high prevalence, emphasizing the role of aging and declining immunity in cancer development.
- Young individuals record the lowest cases, possibly due to shorter exposure durations and fewer age-linked biological changes.
- Treatment costs have shown fluctuations across the 10-year period, peaking in 2016 and dropping in 2022, pointing to varying healthcare demands or changes in treatment approaches.
- The rebound in treatment costs by 2024 suggests rising patient numbers or adoption of costlier modern therapies.
- Colon and prostate cancers account for the highest treatment costs, highlighting their complexity or tendency to be diagnosed at later stages when care becomes expensive.
- Skin cancer records the lowest treatment costs, implying its treatments may be simpler, shorter, or diagnosed earlier.
- Genetic risk is highest for colon and prostate cancers, suggesting strong hereditary links, making family history an important risk assessment factor.
- Lung cancer shows low genetic risk, meaning environmental and lifestyle causes like smoking and pollution likely drive its high prevalence.
- Russia, USA, China, Brazil, and Australia emerge as cancer hotspots, possibly due to industrialization, pollution, dietary habits, or better detection systems.
- Mid-age adults dominate across all cancer types, with lung cancer leading, reinforcing the mid-age group as the most vulnerable population segment.

- Countries with low recorded cancer prevalence may face underreporting challenges, lack of screening, or limited healthcare infrastructure to detect cases early.

Analysis Techniques Used in Excel

- **PivotTables:** Cancer Type vs. Gender and Country.
- **COUNTIFS**:
  `=COUNTIFS(GenderRange,"Male",SurvivalStatusRange,"Dead")`
  to compute male mortality count.
- **AVERAGEIFS**:
  `=AVERAGEIFS(AgeRange,CancerTypeRange,"Lung",SurvivalStatusRange,"Dead")`
  to estimate average age of death for lung cancer patients.
- **Charts:**
  - Bar charts for gender distribution
  - Pie charts for cancer type share
  - Line graphs for age vs. survival rates

Post-Analysis Recommendations

- Expand early detection programs focusing on mid-age and elderly populations to reduce late-stage diagnoses and lower treatment costs.
- Implement genetic counseling and screening for families with a history of colon and prostate cancers to identify at-risk individuals early.
- Invest in public health awareness campaigns emphasizing lifestyle modifications such as quitting smoking, healthy eating, and regular exercise to prevent cancers linked to environmental factors.
- Develop workplace wellness programs offering free cancer screening for employees in high-risk age groups to encourage early diagnosis.
- Strengthen cancer data registries to improve accuracy, track trends, and guide evidence-based healthcare policies.
- Introduce nationwide cancer prevention education in schools and workplaces to build awareness from a young age.
- Expand cancer care infrastructure, especially in high-prevalence countries, to handle growing patient demands effectively.
- Negotiate affordable drug prices and promote local production of chemotherapy drugs to make treatments cost-effective.
- Establish cancer-specific health insurance plans to protect families from catastrophic treatment expenses.
- Support research on cost-efficient treatments, early diagnostic technologies, and cancer vaccines to reduce long-term costs and mortality.
- Deploy telemedicine and mobile screening units in rural or underserved regions for better healthcare reach.
- Encourage cross-country collaboration for sharing best practices in cancer prevention, screening, and treatment innovations.
- Introduce subsidies or tax breaks for hospitals investing in early detection technologies to promote proactive healthcare systems.

- Launch anti-pollution policies and urban health initiatives in industrial areas to reduce environment-linked cancer risks.
- Organize annual cancer awareness weeks offering free check-ups, especially targeting mid-age adults and high-risk groups.

Comparison with Initial Findings

Initial beliefs that cancer affected both genders equally were disproved, the male-to-female ratio was nearly 4:1.
The link between age and mortality was confirmed with strong correlation (visually observed via scatter plot).

Data Visualizations & Charts

Visuals produced using Excel:

- **Bar Chart:** Gender Distribution of Cancer Patients
- **Pie Chart:** Cancer Type Proportion
- **Line Graph:** Age vs. Survival Rate
- **Column Chart:** Country-wise Patient Count

Dashboards

An Excel dashboard was created using **Slicers** to filter by:

- Gender
- Country
- Cancer Type
  This allowed stakeholders to explore survival rates interactively.

Explanation of Visualizations

"The bar chart highlights that male patient dominate cancer statistics globally, especially in the 60+ age group.
The pie chart reveals that breast, lung, and colon cancers contribute over 60% of total cases."

General Recommendations and Observations

Actionable Insights

- **Male-focused interventions:** Launch awareness campaigns targeting older men.
- **Resource allocation:** Prioritize screening programs in countries with high male incidence.
- **Female-specific care:** Expand breast cancer screening in developing countries.

Optimizations or Business Decisions

- Support public health dashboards for real-time cancer surveillance.
- Collaborate internationally for standardized cancer data reporting.

Unexpected Outcomes

Despite global awareness, the female cancer population remains significantly underrepresented, possibly due to data collection gaps or limited access to healthcare in certain regions.

Conclusion

Key Learnings

This Excel-based project underscores the power of spreadsheet analytics in transforming medical data into public health insights. The findings highlight:

- A strong age-gender disparity in global cancer cases.
- The importance of regional health data monitoring.
- The potential for Excel dashboards to simplify medical reporting.

Limitations

- Dataset may not cover all global regions equally.
- Lack of variables such as lifestyle, treatment, or income levels limits interpretation.

Future Research

Future projects could:

- Integrate machine learning forecasts (using Python or Power BI alongside Excel).
- Add treatment response data to link cancer type with survival outcomes.