SOURCING OPEN DATA

Data Immersion - Exercise 6.1

Contents

1. Sourcing Data

Dataset: "Lung Cancer Mortality Datasets v2"

1.1. Data Sourcing

This dataset was sourced from Kaggle (this link directs to the data card)

This is an artificially generated dataset that is publicly available and was posted by a Kaggle user named MasterDataSan. The dataset bio describes the data being made "as close a representation of reality as possible".

1.2. Data Collection

This dataset is an example of artificially generated data designed to closely represent reality. The basis for this dataset is unknown, although I have inquired about this to the data owner.

My current assumption is that the dataset is modelled off clinical records from various European countries. Given the data grain is on a patient level, the original data may be non-shareable due to privacy laws.

1.3. Data Trustworthiness

At the moment, without a point of reference as to what the data is based on, and who MasterDataSan is, the data should be considered untrustworthy but has been considered acceptable for use in this project by my Mentor.

1.4. Data Contents

This dataset is a comprehensive collection of data relating to individuals diagnosed with lung cancer. Each row focuses on a single patient, providing demographic details as well as health-related variables. The geographical distribution is limited to European countries. The diagnosis date for all patients ranges from June of 2014 to June of 2024. (See 2.2.2. column descriptions for further details)

1.5. Reason for Selection

Having a health science background, I believe I am equipped to interpret this data accurately and draw meaningful insights. Also, having checked the data thoroughly, I know it meets all project requirements making it viable for the advanced analytical techniques to be employed. I also believe this dataset may present interesting challenges, such as integrating data from sources like World Bank, to broaden the analysis.

2. Data Profile

2.1. Data Cleaning Process

2.1.1. Initial data exploration:

- Checked data structure and general characteristics
- Made overview of numerical variables using df.describe() and visualised their distributions using histograms and pie charts for binary variables.
- Checked unique values and distributions of categorical variables

2.1.2. Missing values:

- Checked all columns for missing values (no missing values)

2.1.3. Duplicates:

- Checked for duplicate rows (no duplicates)

2.1.4. Mixed data and transformations:

- Checked for mixed datatypes in all columns (no mixed data types)
- Changed column name “beginning_of_treatment_date” to “start_treatment_date” for simplicity.
- Datatype conversions made:

Column Name	**From**	**To**	**Reason**
“diagnosis_date”	object	datetime64[ns]	Columns all contain dates.
“start_treatment_date”			
“end_treatment_date”			
“hypertension”	int64	bool	Boolean values applicable and saves memory.
“asthma”			
“cirrhosis”			
“other_cancer”			
“survived”			
“family_history”	object	bool	
“age”	float64	int8	Save memory
“bmi”	float64	float16	
“cholesterol_level””	Int64	int16	

2.1.5. Filtering and Dropping Data

- Dropped "id" column since it won't be useful for future analysis
- Date ranges checked. Found some dates went past current date (likely due to the data being artificially made). I decided to consider this data as erroneous for the sake of data manipulation experience.
    - Cutoff date set at June 1st, 2024, and all rows where "end_treatment_date" is after the cutoff were dropped.
- Checked to ensure that the diagnosis date, start treatment date, and end treatment date were chronologically correct in each row – created "chronologically_sound" flag.
    - All rows = True

2.1.6. Outliers

- Checked numerical columns, "age", "bmi", and "cholesterol_level", for outliers.
    - Outliers were present within the age column but were not dropped since the distribution of patient ages is reflective of a population (normally distributed).
    - No outliers in bmi or cholesterol_level

2.1.7. Final Checks and Export

- Filtered data checked for shape, statistics on numeric columns, and value counts on categorical variables.
- Dataset exported as pickle file since it will be worked on only in Python.

2.2. Shape and General Info

- **Rows:** 2,842,404
- **Columns:** 18
- **Numerical Variables:** 3
- **Categorical Variables:** 12 (including Booleans)
- **Date Variables:** 3
- **Memory Usage:** 292.2+ MB

2.3. Column Descriptions

2.3.1. Numerical columns

Name	Type	Value Range	Description
age	int8	4 – 104	Hypothetical patient’s age at time of diagnosis.
bmi	float16	16.0 – 45.0	Hypothetical patient’s Body Mass Index at the time of diagnosis.
cholesterol_level	int16	150 – 300	Hypothetical patient’s cholesterol level measured in total milligrams of cholesterol per decilitre of blood (mg/dL).

2.3.2. Categorical columns

Name	Type	Unique Values	Description
gender	object	2	Hypothetical patient’s biological sex.
country	object	27	Hypothetical patient’s country of residence (within Europe).
cancer_stage	object	4	The stage of lung cancer at the time of diagnosis (I, II, III, IV).
family_history	Bool	2	Indicates whether there is a family history of cancer.
smoking_status	object	4	The smoking status of the patient.
hypertension	Bool	2	Indicates whether the patient has high blood pressure.
asthma	Bool	2	Indicates whether the patient has asthma.
cirrhosis	Bool	2	Indicates whether the patient has cirrhosis of the liver.
other_cancer	Bool	2	Indicates whether the patient has had any other type of cancer in addition to the primary diagnosis.
treatment_type	Object	4	The type of treatment the patient received.
survived	Bool	2	Indicates whether the patient survived.
chronologically_sound	Bool	1	Flag for determining whether the date variables make sense chronologically.

2.3.3. Date columns

Name	Date Range	Description
diagnosis_date	4/6/2014 to 1/12/2023	The date on which the patient was diagnosed with lung cancer.
start_treatment_date	5/6/2014 to 15/12/2023	The date on which the patient started their treatment.
end_treatment_date	5/12/2014 to 1/6/2024	The date on which the patient finished their treatment or died.

2.4. Summary Statistics

	Min	Q1	Q2	Q3	Max	Mean	Std Dev
Age	4	48	55	62	104	55	10.0
BMI	16.0	23.2	30.5	37.7	45.0	30.5	8.37
Cholesterol_level	150	196	242	271	300	233.7	43.4
Hypertension	0	1	1	1	1	0.75	0.43
Asthma	0	0	0	1	1	0.47	0.50
Cirrhosis	0	0	0	0	1	0.23	0.42
Other_cancer	0	0	0	0	1	0.09	0.28
Survived	0	0	0	0	1	0.22	0.41

2.5. Limitations and Ethical Considerations

2.5.1. Limitations

Being artificial in nature, the data is subject to many possible limitations. Primarily, since the method of generation is unknown, it is possible some components of this data will not reflect reality. Hence, the insights drawn from this analysis, if they were to be used outside of this project, would merely direct attention to further analysis on real data. Also, being based on real data, this dataset carries an indeterminate number of limitations, possibly including:

- **Inconsistent reporting standards:** Different European countries may have varying reporting standards which makes comparisons across these countries less reliable.
- **Inconsistent medical practices:** Quality of care, expertise, and treatment practices may vary by country, and even by hospital, making analysis more difficult.
- **Unclear treatment labels:** The “combined” treatment category is inherently vague and could mean any number of things, such as switched treatments or underwent multiple treatments.
- **Limited treatment information:** As there is no information regarding variables within treatments (except for start and end dates), limited insights can be drawn about effective practices.

Furthermore, the data is stated to be a close representation of reality. It is unclear if this refers to the data already mirroring some well understood relationships between variables or providing proportionally realistic counts and distributions of all variables (although this is doubtful given the close to equal distribution of patients in each European country despite different populations).

2.5.2. Ethics

- **Privacy:** Again, being artificial in nature, the data can’t be used to identify real people thus ensuring privacy to be a non-issue. Additionally, fake names and pseudonyms have not been used either to ensure mistaken connections to real people can’t be made.
- **Sensitivity:** Whilst sensitivity also feels like a non-issue in an artificial dataset, information regarding family history, smoking status, treatment type, and survival would necessitate careful handling in real-life as to avoid misinterpretation and stigmatization of patients.

2.5.3. Bias

- **Sample Bias:** Most categorical variables like country and cancer stage are too equally distributed which indicates that some demographic groups and countries are certainly underrepresented and overrepresented, impacting analysis accuracy.
- **Collection Bias:** In reference to the data this dataset is based off, possible inconsistencies and errors made when recording/measuring. Some patients might lie for various reasons.

3. Defining Questions to Explore

The following is a list of questions this dataset may explore. Some questions may only be possible with additional data:

3.1. Demographic

- Which health markers are most strongly associated with lung cancer survival?
- Is there an association between treatment received and lung cancer survival?
- Which age groups are most affected by lung cancer?

3.2. Geographic

- How distributed is the prevalence of lung cancer across Europe?
- Which countries have higher rates of lung cancer survival? Why might this be?

3.3. Temporal

- What is the average treatment duration for individuals with lung cancer
  - ➢ How does this vary based on whether the patient survived?
- Are there any notable trends or seasonal patterns in the diagnosis and treatment dates?

3.4. Predictive Modelling

- Can a predictive model be developed to determine an individual’s likelihood of surviving lung cancer based on health markers, treatment, and/or demographical variables?
  - ➢ Which variables contribute most to lung cancer survival?
- Can a predictive model be developed to determine an individual’s chance of survival based on the duration of their treatment?