Achievement 6

Exercise 6.1

Poojaben Thummar

DATA SOURCING

This dataset contains data about lung cancer Mortality. This database is comprehensive collection of patient information, specifically focused on individuals diagnosed with cancer. It is designed to facilitate the analysis of various factors that may influence cancer prognosis and treatment outcomes. The database includes a range of demographic, medical, and treatment-related variables, capturing essential details about each patient's condition and history.

DATA COLLECTION

This dataset is an example of artificially generated data designed to closely represent reality. 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.

kaggle

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.

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.

REASON FOR SELECTION

I have been living in Europe for the past two years so I am curios to digging in this data, I believe I am equipped to interpret this data accurate 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.

DATA PROFILE

Data Cleaning Process

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

Missing values:

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

Duplicates:

- Checked for duplicate rows (no duplicates)

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:

Columns name	It was	Changed to
diagnosis_date	object	datetime64[ns]
start_treatment_date	object	datetime64[ns]
end_treatment_date	object	datetime64[ns]
hypertension	int64	bool
asthma	int64	bool
cirrhosis	int64	bool
other_cancer	int64	bool
survived	int64	bool
“family_history”	object	bool
“age	float64	int8
bmi	float64	float16
“cholesterol_level””	Int64	int16

Filtering and Dropping Data

- Dropped “id” column since it will be not 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.
- Checked to ensure that the diagnosis date, start treatment date, and end treatment date
were chronologically correct in each row – created “chronologically_sound” flag.

Outliers:

- Checked numerical columns, “age”, “bmi”, and “cholesterol_level”, for outliers.
distribution of patient ages is reflective of a population (normally distributed).

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.

Shape and General Info

- Rows: 2,842,404
- Columns: 1

Numerical columns

age : Hypothetical patient's age at time of diagnosis.
bmi : Hypothetical patient's Body Mass Index at the time of diagnosis.
cholesterol_level :Hypothetical patient's cholesterol level measured in total milligrams of cholesterol per decilitre of blood (mg/dL).

Categorical columns

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

Summary Statistics

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

Limitations and Ethical Considerations

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.

DEFINING QUESTIONS TO EXPLORE

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

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?

Geographic

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

Temporal

- What is the average treatment duration for individuals with lung cancer
- Are there any notable trends or seasonal patterns in the diagnosis and treatment dates?