Title: Cancer Incidence in the East of England

Subtitle: A Data-Driven Insight into Regional Cancer Diagnoses

Introduction

The statistic that one in two people will develop cancer in their lifetime underscores the growing importance of understanding cancer patterns, both globally and regionally. This figure reflects advancements in life expectancy, with age being the most significant risk factor for cancer. It also points to the increasing need for early detection, effective prevention, and well-resourced healthcare systems to manage this rising burden.

Research by Cancer Research UK highlights that over 60% of cancer cases occur in individuals aged 65 and older. Factors like smoking cessation, regular physical activity, moderate alcohol consumption, and maintaining a healthy weight could prevent over 40% of cancer cases each year in the UK. These findings emphasise the dual role of medical advancements and public health initiatives in addressing the cancer epidemic.

Understanding regional cancer trends, such as those in the East of England, becomes critical for planning targeted healthcare responses. Geographic-specific data can help allocate resources, identify high-incidence cancer types, and tailor prevention campaigns to regional demographics.

(Ahmad et al., British Journal of Cancer, 2015), which analysed how lifetime cancer risks have shifted over decades.

Objective

Analyse cancer diagnoses data for the East of England to uncover trends and actionable insights.

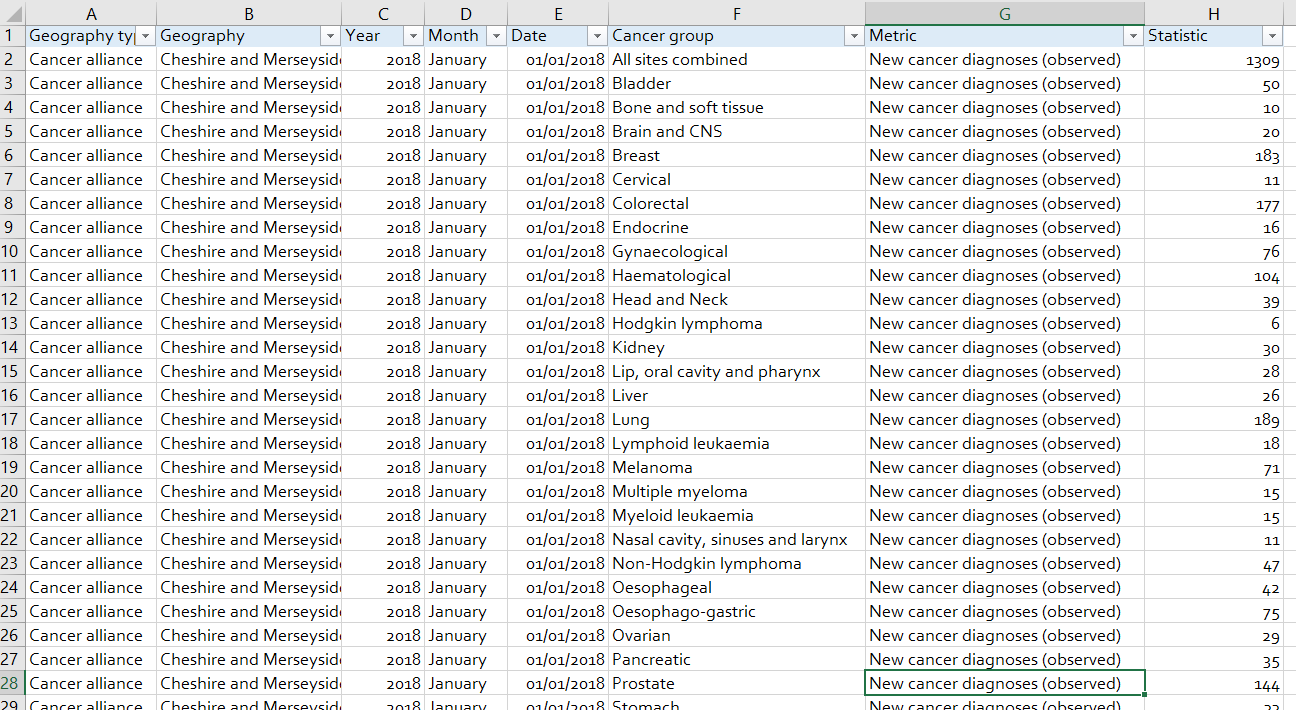
Aim

To investigate regional patterns of cancer incidence in the East of England, identify trends in diagnoses by cancer type, and provide actionable insights for healthcare planning and targeted interventions.

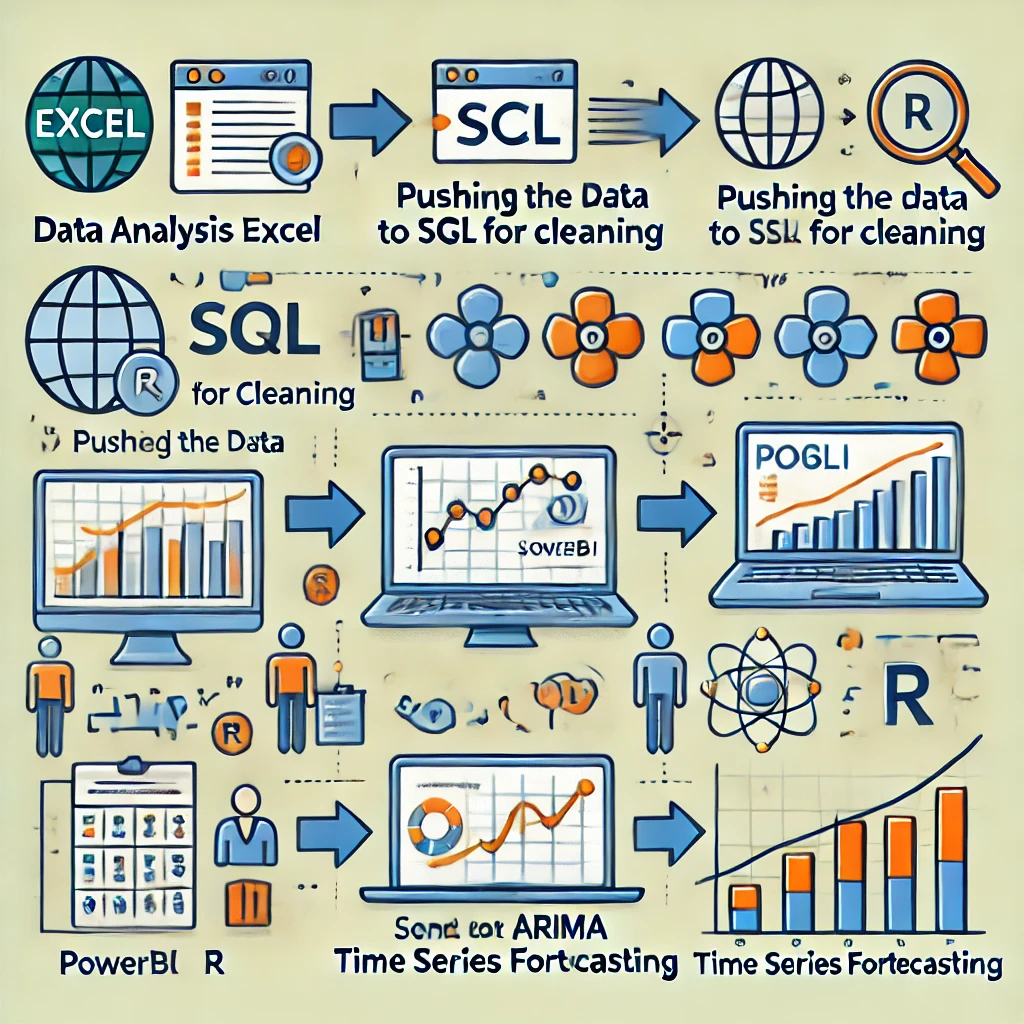
This aim focuses not only on presenting data but also on interpreting it for meaningful action. It considers public health implications, emphasizing regional and national comparisons, and highlights the importance of targeted responses.

Data Analysis Approach for Cancer Incidence in the East of England

Data contain 8 columns (Geography type, Geography, Year, Month, Date, Cancer group, Metric, Statistic) and 134380 rows.



This analysis utilizes a multi-tool approach, incorporating Excel, SQL, Power BI, and R to address the data quality, visualization, and forecasting requirements as outlined in the job description. The goal is to analyze cancer diagnosis data and provide insightful trends and forecasts.



Excel: Exploratory Data Analysis (EDA)

In this phase, Excel is used for Exploratory Data Analysis (EDA) to assess and clean the dataset before further processing. Key tasks include:

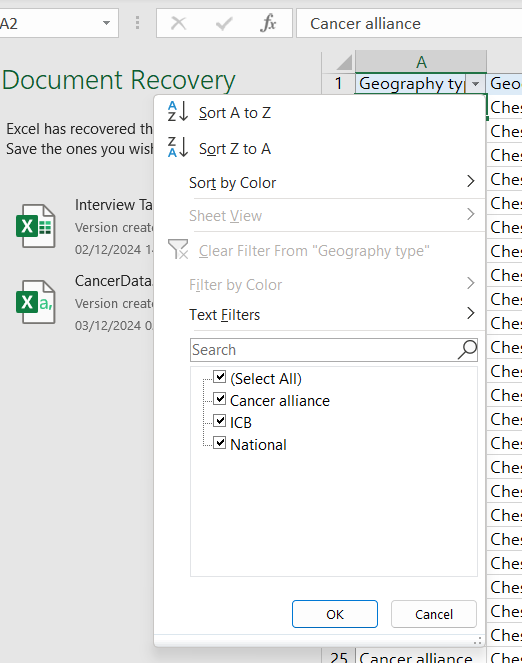
Data Overview and Inspection: Row and Column Count: To ensure completeness, check the total number of records and variables.

Filtering and Consistency Check: Filter the data to identify relevant columns for analysis, removing errors and inconsistencies.

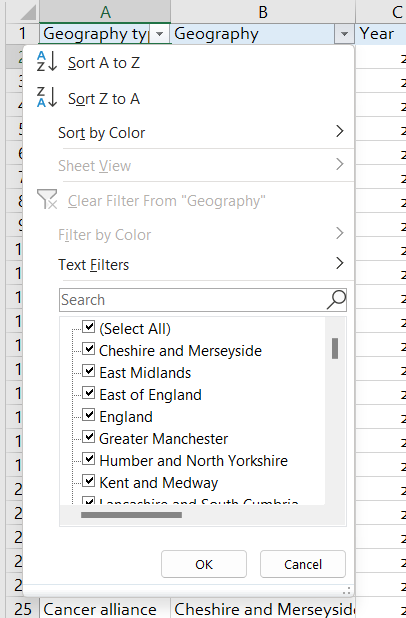
Distinct Value Check: Identify distinct values in each column to spot discrepancies and remove irrelevant data, such as erroneous or duplicate entries.

Key Columns Overview:

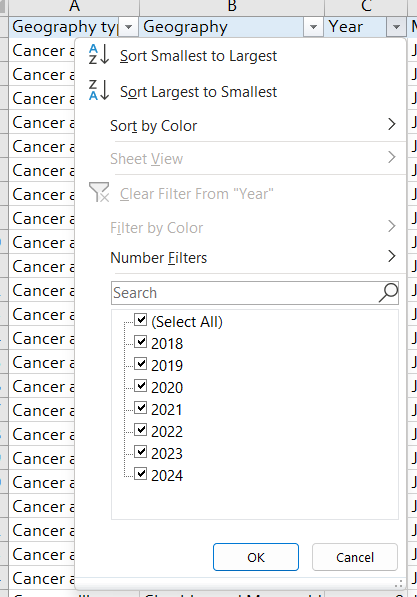
Geography Type: Categorized into three types: Cancer Alliance, ICB (Integrated Care Board), and National.



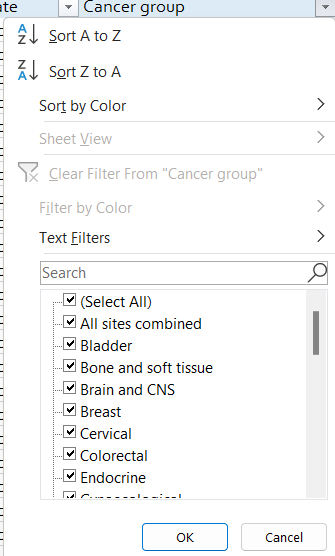
Geography: Contains 63 distinct regions across the East of England.



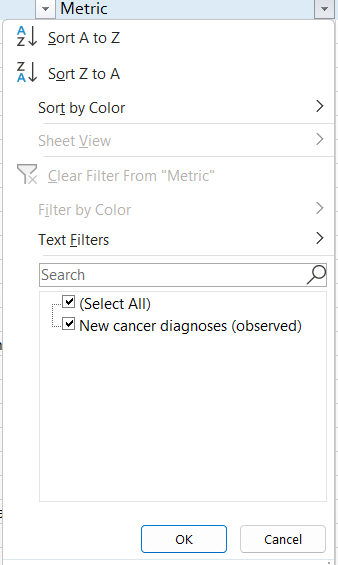
Date Information: Time-based information covering the years 2018 to 2024, with data recorded on the first day of each month.



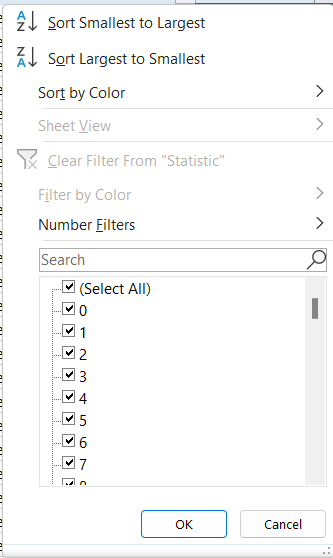
Cancer Group: Contains 33 cancer types. The 'ALL SITES COMBINED' entry, an outlier with high values, will be excluded for more accurate insights.



Metric: Describes the statistical measurement of cancer diagnoses (observed) with just one records for each value.



Statistic: Numeric values representing the actual cancer diagnoses from 0 to 30216.

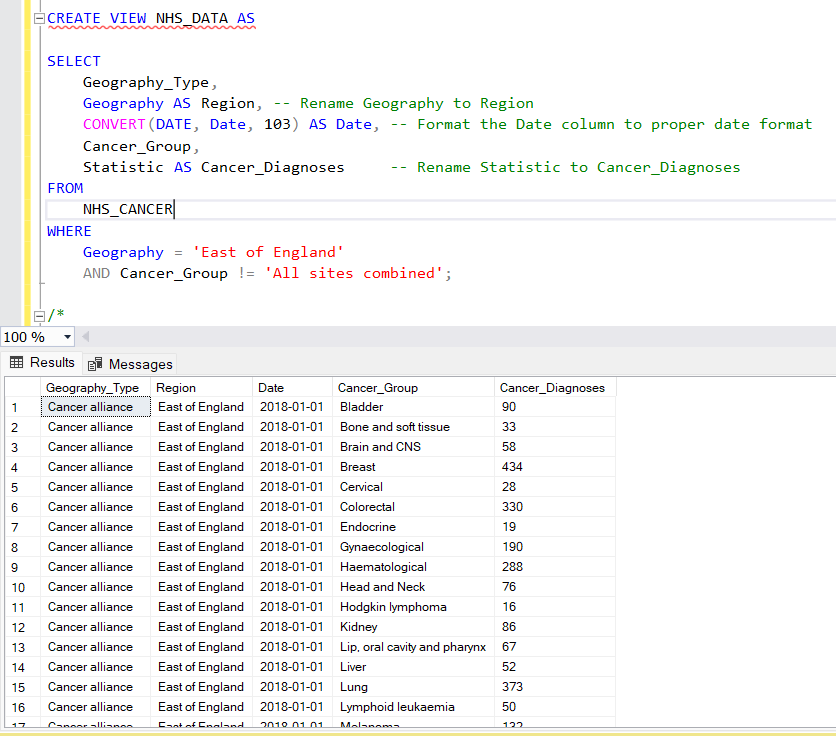


**SQL: Data Cleaning and Quality Assurance**

After performing the initial data check in Excel, SQL will be used for further data cleaning and ensuring quality before exporting the cleaned dataset to Power BI.

Data Cleaning Steps:

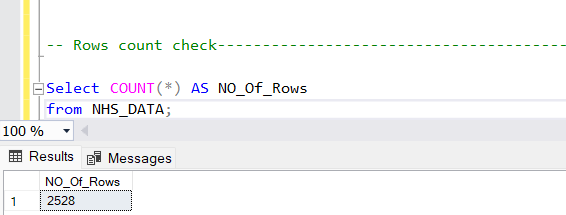
* Remove Unnecessary Columns: Remove irrelevant columns such as Year, Month, and Metric to focus on critical data.
* Filter Data for East of England: Extract data specific to the East of England region from the geography column (Geography\_type, Geography, Date, Cancer\_group and Statistic).
* Format Date Column: Ensure the Date column is in the appropriate date format for consistent analysis (year-month-day=yyy/mm/dd).
* Remove Outliers: Exclude the outlier 'ALL SITES COMBINED' from the Cancer Group column to prevent skewed results.
* Rename Columns: Rename Geography to Region and Statistic to Cancer\_Diagnoses for clarity.
* View: Create a view so the that can be pushed to power bi.



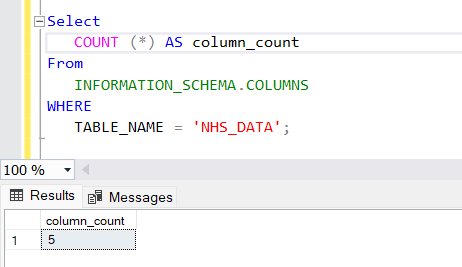
SQL CODE FOR DATA CLEANING

Data Quality Check Steps

Row Count: The dataset should not exceed 134,380 rows.



Column Count: Verify that the dataset has exactly 5 columns: Geography Type, Region, Date, Cancer Group, and Cancer Diagnoses.



Data Type Checks: Ensure proper data types:

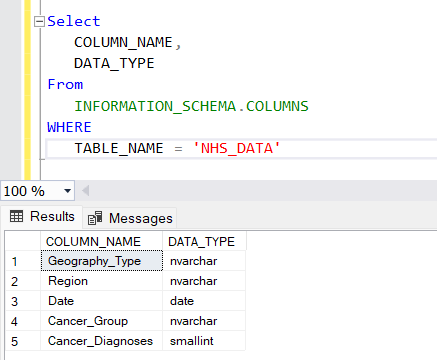
Geography\_Type: varchar

Region: varchar

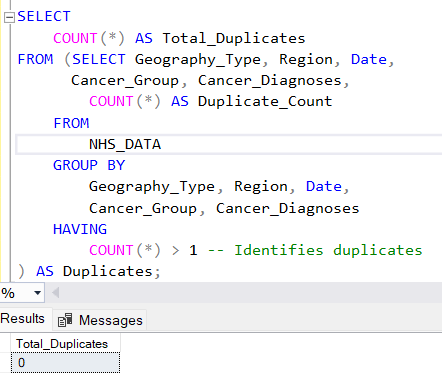
Date: date format

Cancer\_Group: varchar

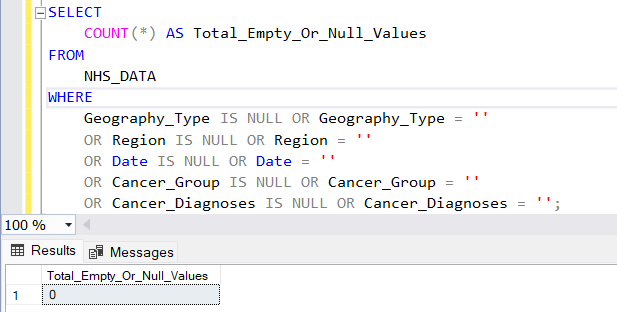
Cancer\_Diagnoses: numerical/integer



Duplicate Check: Ensure there are no duplicate records.



Null or Empty Value Check: Ensure there are no missing or empty values in any critical column.



Outlier Check: There should be no more than 10 outliers per Cancer Group.

---- Outliers check--------------------------------

-- Step 1: Calculate Q1, Q3, and IQR for each Cancer\_Group

WITH Quartiles AS (

SELECT

Cancer\_Group,

PERCENTILE\_CONT(0.25) WITHIN GROUP (ORDER BY Cancer\_Diagnoses)

OVER (PARTITION BY Cancer\_Group) AS Q1,

PERCENTILE\_CONT(0.75) WITHIN GROUP (ORDER BY Cancer\_Diagnoses)

OVER (PARTITION BY Cancer\_Group) AS Q3

FROM NHS\_DATA

),

IQR\_Calculation AS (

SELECT DISTINCT

Cancer\_Group,

Q1,

Q3,

(Q3 - Q1) AS IQR,

(Q1 - 1.5 \* (Q3 - Q1)) AS Lower\_Bound,

(Q3 + 1.5 \* (Q3 - Q1)) AS Upper\_Bound

FROM Quartiles

),

Outlier\_Flagged AS (

SELECT

d.Cancer\_Group,

d.Cancer\_Diagnoses,

CASE

WHEN d.Cancer\_Diagnoses < i.Lower\_Bound OR d.Cancer\_Diagnoses > i.Upper\_Bound THEN 1

ELSE 0

END AS Is\_Outlier

FROM NHS\_DATA d

JOIN IQR\_Calculation i

ON d.Cancer\_Group = i.Cancer\_Group

)

-- Step 2: Count the number of outliers for each Cancer\_Group

SELECT

Cancer\_Group,

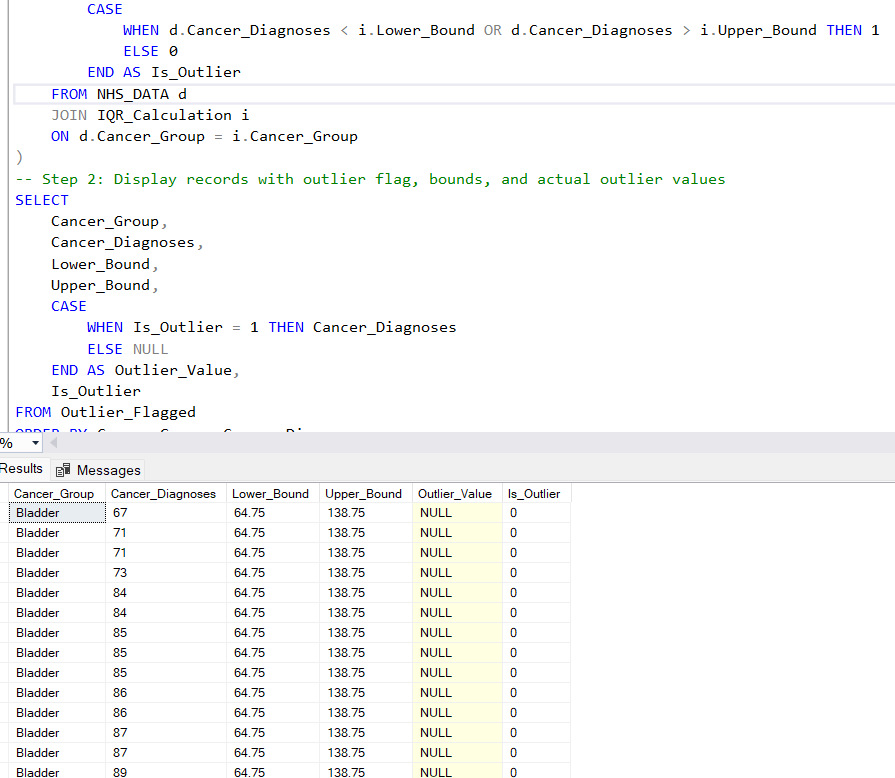
SUM(Is\_Outlier) AS Number\_Of\_Outliers

FROM Outlier\_Flagged

GROUP BY Cancer\_Group;

**Code for outlier**

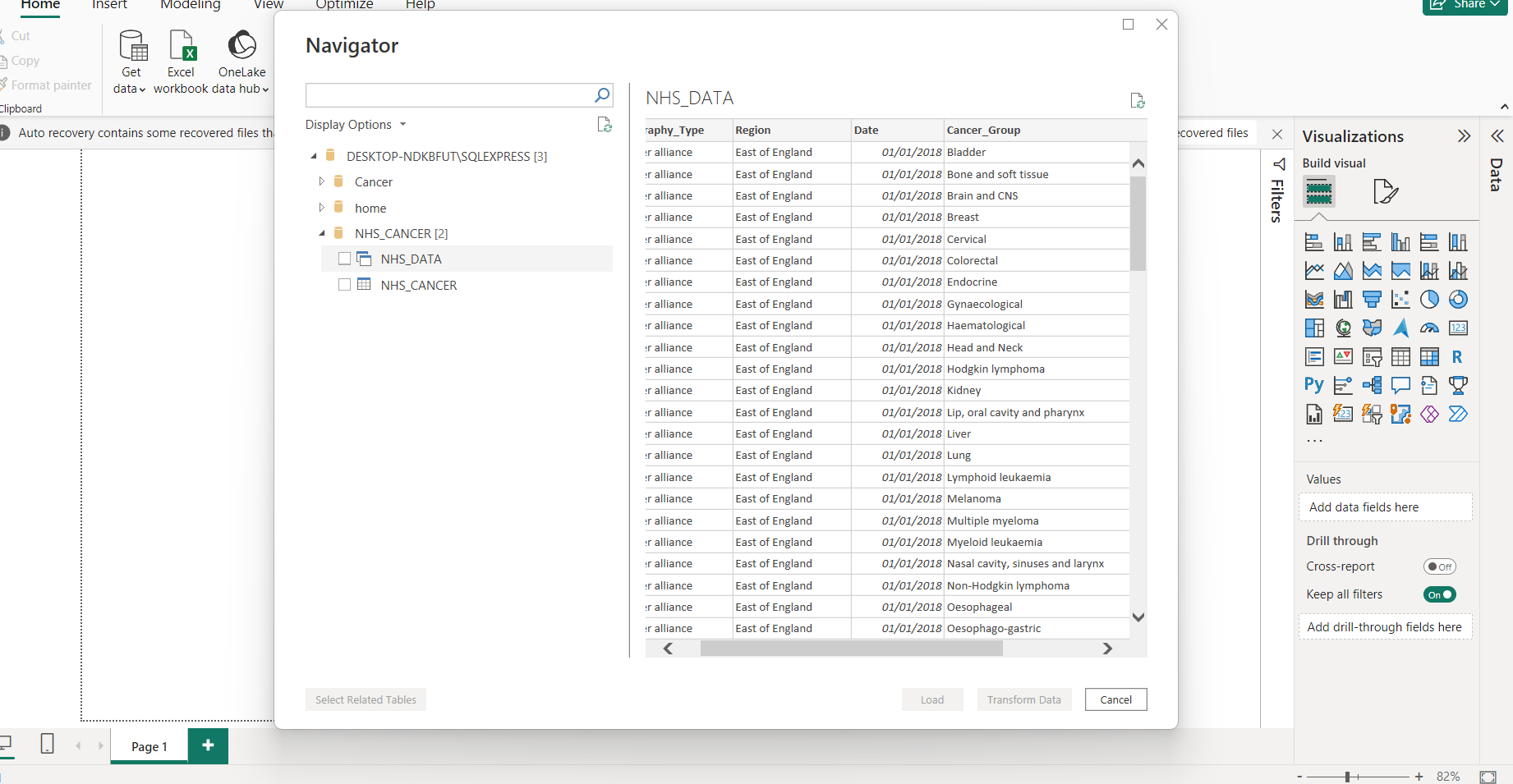
Bladder 1, Bone and soft tissue 0, Brain and CNS 1, Breast 3, Cervical 2, Colorectal 2, Endocrine 0, Gynaecological 4, Haematological 1, Head and Neck 0, Hodgkin lymphoma 1, Kidney 1, Lip, oral cavity and pharynx 0, Liver 1, Lung 0, Lymphoid leukaemia 1, Melanoma 1,Multiple myeloma 0, Myeloid leukaemia 0, Nasal cavity, sinuses and larynx 1, Non-Hodgkin lymphoma 2, Oesophageal 4, Oesophago-gastric 3, Ovarian 0, Pancreatic 0, Prostate 4, Stomach 4, Testicular 0, Unknown 1, Upper GI excl OG 0, Urological excl prostate 2, Uterine 1



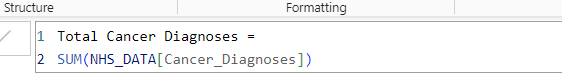
After meeting these quality expectations, the cleaned data will be pushed to Power BI for advanced analysis.

**Power BI: Data Visualization and Analysis**

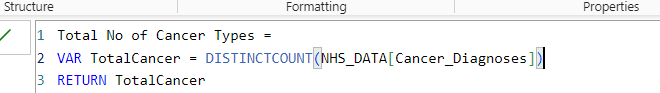
Power BI will be used to create interactive dashboards that provide a deeper understanding of cancer incidence trends. The following metrics will be calculated using DAX (Data Analysis Expressions):



* Sum of Diagnoses: Total diagnoses for each Cancer Group.



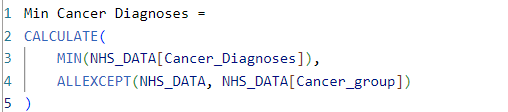
Total number of cancer type



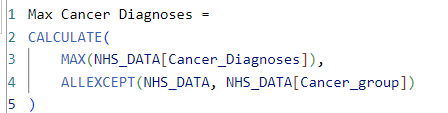
Total number of days



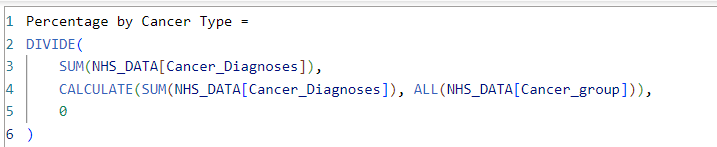
* Maximum and Minimum Diagnoses: Identify the highest and lowest diagnoses recorded.

Minimum  


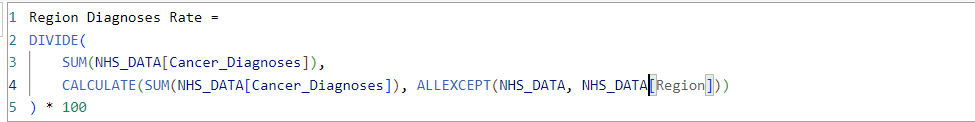
Maxi,um



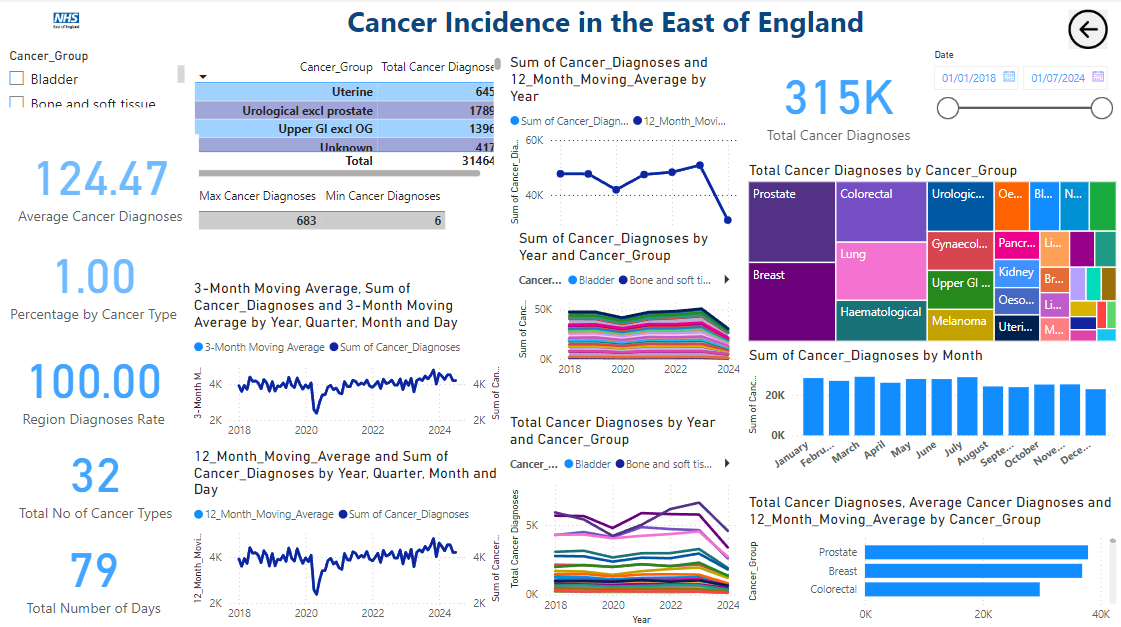
* Percentage by cancer type: refers to the proportion of total cancer diagnoses that belong to a specific type of cancer. In simple terms, it tells you how much of the total cancer diagnoses in a dataset are associated with each cancer type.



* Region diagnoses rate: Show the rate of diagnoses of each cancer type. refers to the frequency or rate at which cancer cases are diagnosed in a specific geographic area, such as a region or locality. This rate helps to understand how common cancer diagnoses are within a particular region, providing insights into the health burden in that area.



* Trends Over Time: Visualize changes in cancer diagnoses over the time series (2018-2024).
* Moving Average: Compute the suitable moving average for diagnoses over time to smooth out fluctuations.
* Comparison of Cancer Groups: Compare the number of diagnoses across different cancer types and regions over time.
* An interactive dashboard will be created to allow stakeholders to explore these trends visually and make informed decisions.



Observations from the Graphs

There is a decline in diagnoses in 2020 visible in both moving averages. This could be linked to the global impact of the COVID-19 pandemic, which disrupted healthcare services, causing fewer screenings and diagnoses. After 2020, the charts suggest a recovery trend, with diagnosis rates increasing in both short and long terms, indicating a resumption of healthcare activities. The 3-month moving average shows seasonal variability, with some months experiencing higher rates of diagnoses than others. The 12-month moving average points to an overall stable or slightly increasing trend post-2020, suggesting improvements in cancer diagnosis rates over time.

Actionable Recommendations for Stakeholders

Resource Allocation:

Use the 3-month average to identify periods of high demand for diagnostic services and allocate resources accordingly (e.g., hiring more staff during peak times or increasing diagnostic capacity).

Policy Adjustments:

Analyze the drop in 2020 to ensure contingency plans for similar disruptions in the future. Monitoring and Evaluation: Use the 12-month average to evaluate the long-term effectiveness of public health campaigns or diagnostic initiatives.

Public Awareness Campaigns: Focus campaigns during months of low diagnoses (as seen in the 3-month average) to address potential gaps in awareness or service access.

1. Sum of Cancer Diagnoses and 12-Month Moving Average by Year:

Trend Analysis: The graph indicates a fluctuating pattern in the sum of cancer diagnoses, with a noticeable dip in recent years.

12-Month Moving Average: The moving average shows a smoother trend over the years, which could help identify broader patterns without being influenced by short-term fluctuations.

Insight: There appears to be a significant decline in diagnoses after 2021, which could be worth investigating further, possibly to identify changes in healthcare access or diagnostic practices. A drop in diagnoses could also be due to factors such as changes in reporting, delays caused by external events like the COVID-19 pandemic, or other public health impacts.

2. Sum of Cancer Diagnoses by Year and Cancer Group:

Data Breakdown by Cancer Group: This graph breaks down the cancer diagnoses by specific types (e.g., Bladder, Bone, and soft tissues). Insight: The graph shows varying trends across different cancer types. For example, Bladder cancer diagnoses seem to have the highest sum of cases in most years, while Bone and soft tissue cancers might show a smaller proportion but still demonstrate a steady increase.

Trend Variations: The behavior of each cancer type may suggest that diagnostic patterns or healthcare responses differ across cancer groups. It would be useful to explore why certain cancers are increasing more rapidly than others or why some are declining.

3. Total Cancer Diagnoses by Year and Cancer Group:

Overall Trend: This graph tracks the total cancer diagnoses over the years, which includes all cancer groups combined. The sum fluctuates but seems to exhibit a slight increase up to 2022 and then potentially stabilizing or slightly decreasing.

Insight: The graph suggests that, despite fluctuations in the total number of diagnoses, some specific cancer groups are seeing increased numbers, while others are either stable or decreasing. This could indicate a shift in cancer types being diagnosed more frequently.

Summary:

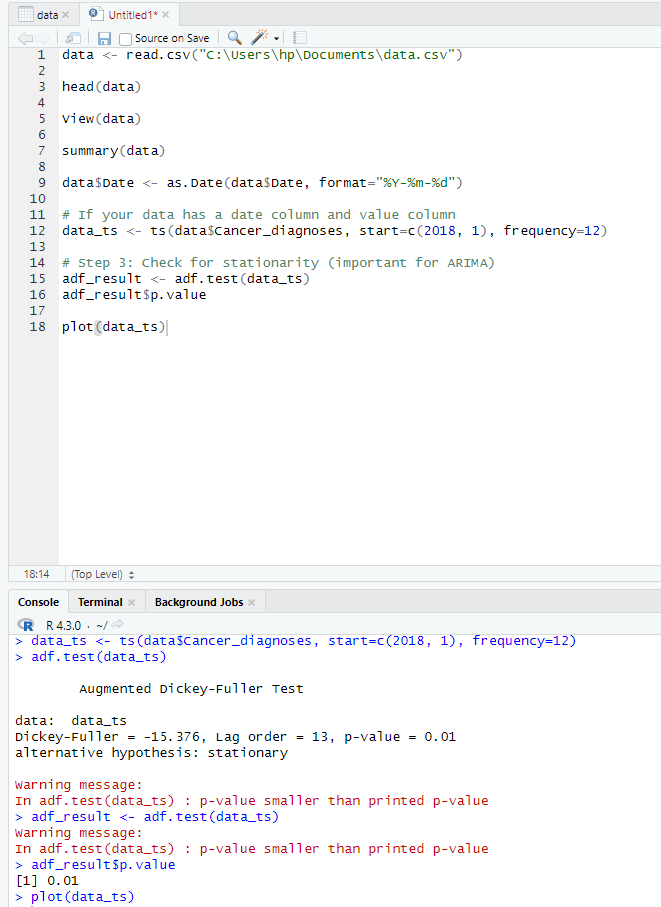
Yearly Fluctuations: Overall, there’s a visible fluctuation in the data, which could be impacted by multiple factors like changes in diagnostic practices, public health policies, or even improvements in early detection techniques.

Cancer Type-Specific Analysis: Different cancer types have different diagnostic patterns. Identifying which cancer types are more prevalent could help guide public health initiatives or resource allocation.

Potential Decline: The decline in diagnoses seen in the first graph should be further explored. It might indicate a broader trend related to the healthcare system or environmental factors.

R: Time Series Forecasting

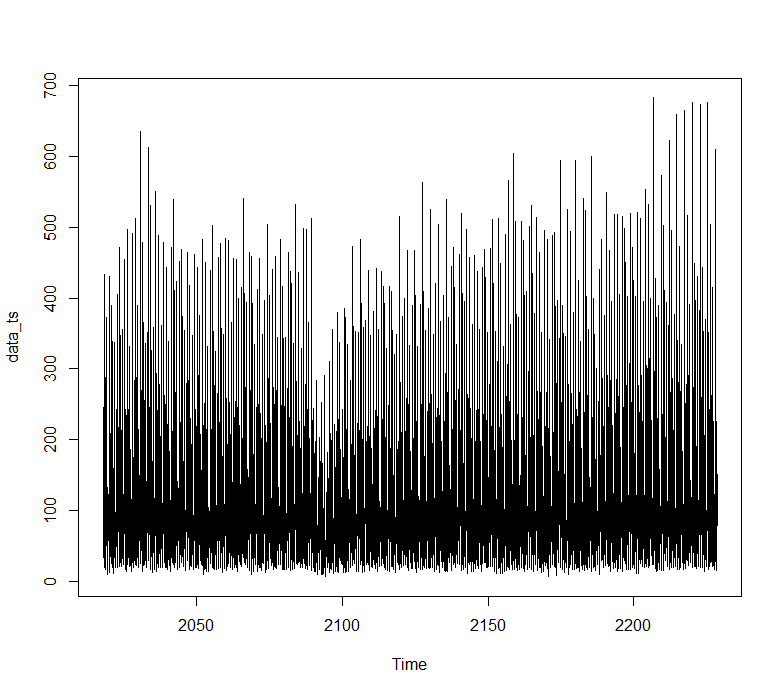
Once the data has been processed and visualized in Power BI, R will be used to conduct time series analysis. The goal is to forecast future cancer diagnoses based on historical trends. Key steps include:



p-value from the Augmented Dickey-Fuller (ADF) test is 0.01. Since your p-value (0.01) is less than 0.05 (the common significance level), reject the null hypothesis. This means that there is strong evidence to suggest that the time series is stationary. A p-value of 0.01 indicates that the time series is likely stationary and does not have a unit root. This is a good result, as stationarity is typically a necessary assumption for models like ARIMA.

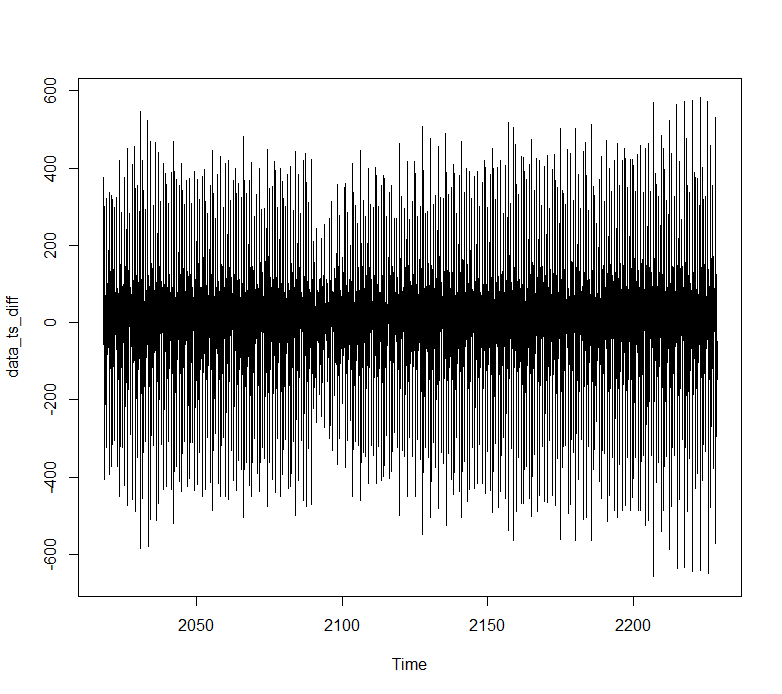
Plot the time series to visualize trends and seasonality

plot(data\_ts)



Differencing since necessary to make the series stationary

data\_ts\_diff <- diff(data\_ts)



Time Series Decomposition: Decompose the time series data to analyze trends, seasonality, and irregular components.

Modeling: Build forecasting models, such as ARIMA (AutoRegressive Integrated Moving Average), to predict future cancer diagnoses in the East of England.

> summary(model)

Series: data\_ts

ARIMA(4,1,1)(2,0,2)[12]

Coefficients:

ar1 ar2 ar3 ar4 ma1 sar1 sar2 sma1 sma2

-0.2956 -0.0505 -0.1187 -0.1404 -0.9886 0.0876 0.0586 -0.2168 -0.2360

s.e. 0.0201 0.0211 0.0247 0.0515 0.0032 0.3632 0.1530 0.3626 0.2001

sigma^2 = 14611: log likelihood = -15700.96

AIC=31421.91 AICc=31422 BIC=31480.26

Training set error measures:

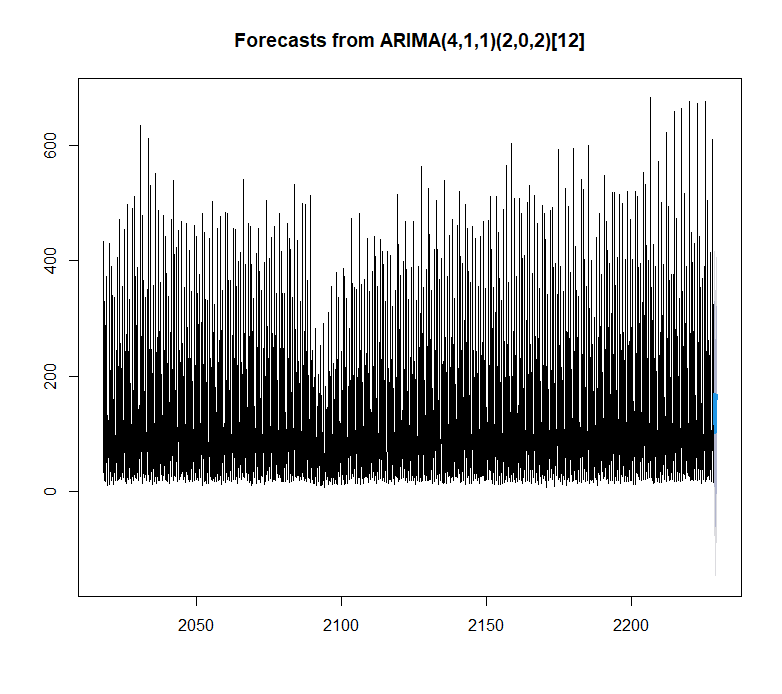
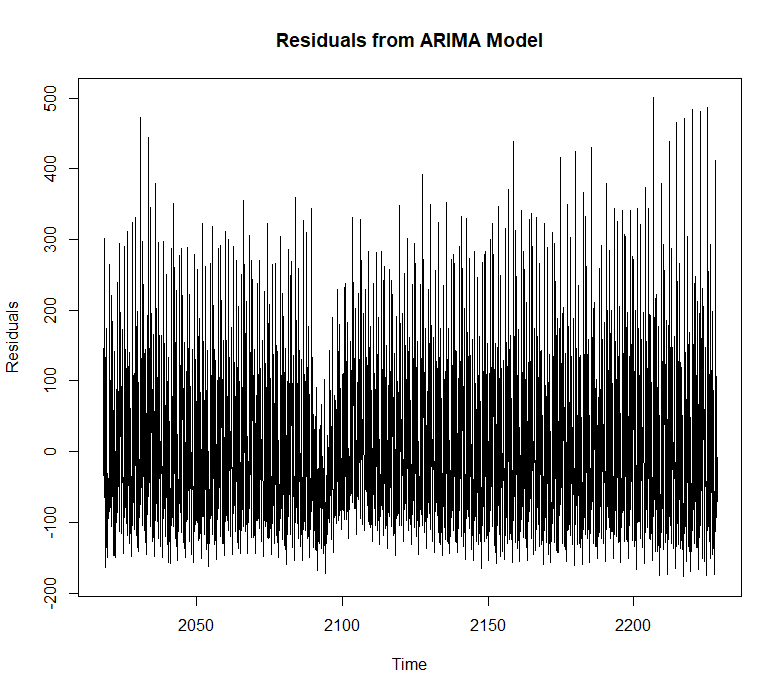
ME RMSE MAE MPE MAPE MASE ACF1

Training set 1.120409 120.6369 94.03525 -120.1777 162.522 0.6851107 0.008158782

This ARIMA model fits the data with some areas of improvement, as indicated by higher error metrics like **MAPE**.

Observations:

The forecast suggests that the values in the future will likely follow a similar pattern to the historical data, but the variability in the forecast increases with time, especially in the distant future (shown by the wider confidence intervals). The model has captured the seasonality well, as reflected in the oscillations of the forecast aligned with the seasonal patterns in the data. The high volatility (spikes) in the forecast toward the end of the series may indicate unusual events or outliers that the model has trouble predicting precisely.

Insight Generation: Provide actionable insights and trends, allowing decision-makers to understand how cancer diagnoses might evolve in the future.

Structure and Behavior of Residuals:

The residuals appear to fluctuate around a mean of approximately zero, which is desirable in a well-fitted ARIMA model. This suggests that the model is unbiased on average.

However, there seems to be considerable variability in the residuals, with some periods having higher spikes than others.

Variance Consistency:

The residuals show relatively consistent variance throughout the time series, which indicates that the assumption of homoscedasticity (constant variance) may hold. Any noticeable increase or clustering in variance over time would suggest issues in the model.

Extreme Values:

There are some noticeable extreme spikes, both positive and negative, which could represent outliers in the data or periods where the model underperformed.

Shapiro-Wilk normality test

data: residuals

W = 0.88591, p-value < 2.2e-16