Analyzing global cancer statistics through visualization

1. Introduction

1.1 Context and Objectives

Cancer has been one of the most serious diseases throughout history. According to the World Health Organization, cancer is the second leading cause of death worldwide. Millions of people die from it yearly and has caused a major burden to the society and economy, affecting individuals and families physically, psychologically and financially. In this project we are going to create a dashboard that visualizes cancer mortality trends and distributions according to different demographic and geographic factors. For the analysis of data, various supervised models will be applied to extract information. We will be focusing on Power BI and Python to conduct our task in the project.

1.2 Significance

Cancer comes in all types and forms, they are abnormal cells that can start and grow in almost any place of the body. When they start to invade important organs, complications may arise and eventually result in death. So it is important to be able to diagnose in the early stage and provide treatment to minimize the damage as much as possible. In this project we aim to have a deeper understanding of the underlying factors correlated with cancer death.

To reach our goal, the dataset will be processed through visualization and supervised learning. In specific, we will use decision tree models and linear regression models to analyze the dataset to gain more insight. Different parameters will be utilized to produce results such as future trends and probability prediction. With the results in hand, it is hoped that this project can equip the readers with more knowledge regarding cancers, and that the findings can help regulators to better allocate social resources.

2. Data Overview

2.1 Data Description

In this project, we used 2 datasets to analyze the global cancer statistic.

The first dataset used in this project is the Cancer and Deaths Dataset : 1990~2019 Globally published in Kaggle, which is a free online platform that provides access to different datasets and machine learning resources. It is a comprehensive dataset containing information on cancer incidence and mortality rates across the world from 1990 to 2019. [3] The dataset contains 9 csv files which includes information on cancer types, age, death rate, and region, etc. This dataset is mainly for Data Visualization and Insights and regression.

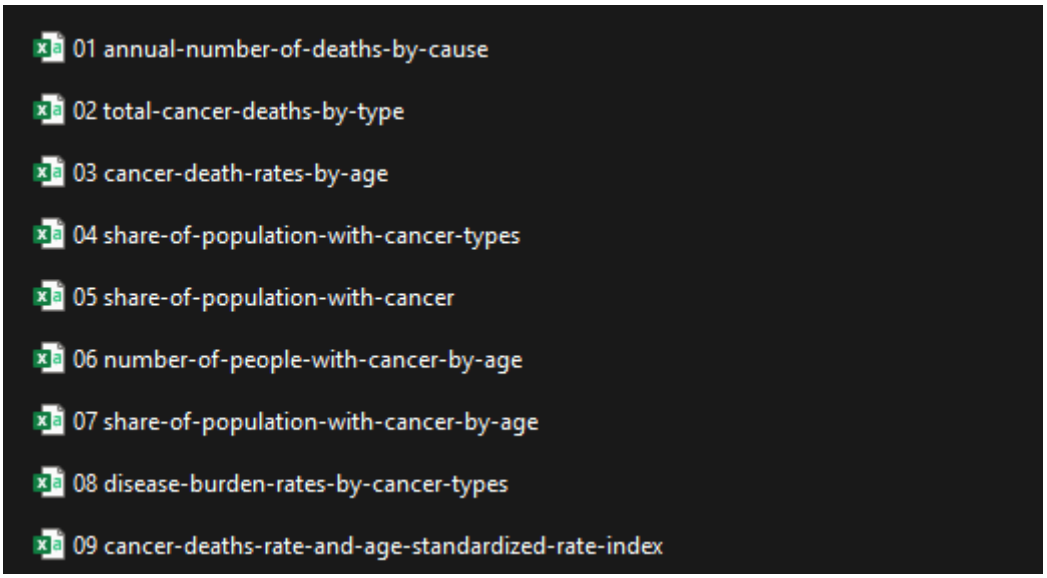


Figure 1: Dataset in CSV format

Another dataset is the Cancer data about the Benign and malignant cancer data, which is also published in Kaggle. This dataset contains information about 570 cancer cells and 30 features to determine whether the cancer cells in our data are benign or malignant. The cancer data contains 2 types of cancers: 1. benign cancer (B) and 2. malignant cancer (M). [6] This dataset is mainly for Decision Tree.

2.2 Data Preprocessing

Data preprocessing is an important step in data science. refers to the cleaning, transforming, and integrating of data in order to make it ready for analysis. [5] Since the csv files may contain null value, we need to do data cleaning to ensure the data quality. Besides, transforming and integrating the dataset is required to make the analysis process more feasible. Below are the steps of data preprocessing applied in this project:

- Data cleaning:

If the dataset contains some null value, when we are trying to do calculation on it, the code will encounter errors. Therefore, in this project, we treat all the null values in the csv files as zero, and convert the cleaned dataframe to new csv.

Source code:

```
import pandas as pd
import numpy as np
```
✓ 0.0s

```
df1 = pd.DataFrame(pd.read_csv('01 annual-number-of-deaths-by-cause.csv'))
df2 = pd.DataFrame(pd.read_csv('02 total-cancer-deaths-by-type.csv'))
df3 = pd.DataFrame(pd.read_csv('03 cancer-death-rates-by-age.csv'))
df4 = pd.DataFrame(pd.read_csv('04 share-of-population-with-cancer-types .csv'))
df5 = pd.DataFrame(pd.read_csv('05 share-of-population-with-cancer.csv'))
df6 = pd.DataFrame(pd.read_csv('06 number-of-people-with-cancer-by-age.csv'))
df7 = pd.DataFrame(pd.read_csv('07 share-of-population-with-cancer-by-age.csv'))
df8 = pd.DataFrame(pd.read_csv('08 disease-burden-rates-by-cancer-types.csv'))
df9 = pd.DataFrame(pd.read_csv('09 cancer-deaths-rate-and-age-standardized-rate-index.csv'))
df = [df1, df2, df3, df4, df5, df6, df7, df8, df9]
```
✓ 0.1s

```
for i in range(len(df)):
    if df[i].isnull().values.any():
        df[i].fillna(0, inplace=True)
```
✓ 0.0s

```
df1.to_csv('res1.csv')
df2.to_csv('res2.csv')
df3.to_csv('res3.csv')
df4.to_csv('res4.csv')
df5.to_csv('res5.csv')
df6.to_csv('res6.csv')
df7.to_csv('res7.csv')
df8.to_csv('res8.csv')
df9.to_csv('res9.csv')
```

Figure 2: Codes from data cleaning python notebook

- Transforming data:

In the 2nd csv file, which is the total cancer deaths by type, has the deaths of each cancer type initially. However, if we want to analyze the total death by cancer, we need to sum up all the values of death in the same row. In excel, we can simply just add a new column, total, where its value is the summation of all the values in the same row:

V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG
Deaths - Ce	Deaths - Br	Deaths - No	Deaths - Pa	Deaths - Es	Deaths - Te	Deaths - Na	Deaths - Ot	Deaths - Co	Deaths - Ne	Deaths - Me	Total
332	422	996	138	529	3	66	37	539	25	6	10386
340	438	1010	137	535	4	67	37	542	26	7	10558
353	472	1040	137	546	4	69	38	550	26	7	10894
367	505	1062	139	560	5	72	39	555	27	7	11241
380	526	1069	140	575	5	73	39	554	28	8	11484
389	541	1074	139	584	6	74	39	552	28	8	11649
399	559	1081	139	595	6	75	39	550	28	8	11843
410	580	1090	139	606	6	75	39	550	29	8	12045
419	588	1094	140	615	6	76	39	549	29	9	12165
428	591	1098	141	626	6	76	39	548	29	9	12275
438	600	1100	142	633	6	76	39	549	30	9	12429
453	623	1113	145	642	7	76	39	557	31	10	12736
456	636	1126	148	639	7	75	39	567	32	10	12846
464	703	1165	152	641	8	75	39	585	33	10	13262
475	740	1202	157	646	9	76	39	603	34	11	13604
480	754	1228	161	646	10	76	39	618	35	11	13774
485	761	1257	167	649	10	76	39	635	36	12	13925
489	768	1290	174	652	10	76	40	655	37	12	14100
496	767	1330	181	659	11	77	41	678	38	12	14308
499	770	1376	190	664	11	77	41	703	39	13	14516
506	789	1425	201	672	12	78	42	729	40	13	14809

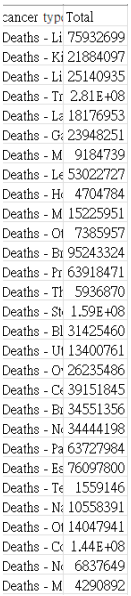
Figure 3: summation of deaths value to obtain total deaths of all types of cancer

Besides doing summation of all the values in the same row, if we want to analyze the total death of each cancer type and find out the most common and dangerous cancer type, we need to transpose this excel. After transposing the excel and obtaining an excel with cancer type as row and country as column, we can sum up the value in the same row and find the total deaths caused by cancer type from 1990 - 2019.

Entity		Afghanista	Afghanista	Afghanista	Afghanista	Afghanista	Afghanista	Afghanista	Total
Code	AFG	AFG	AFG	AFG	AFG	AFG	AFG	AFG	
Year	1990	1991	1992	1993	1994	1995	1996	1997	
Deaths - Li	851	866	890	914	933	946	962	976	75932699
Deaths - Ki	66	66	68	70	71	71	72	73	21884097
Deaths - Li	89	89	91	93	94	94	95	95	25140935
Deaths - Tr	983	982	989	995	996	995	994	993	2.81E+08
Deaths - La	260	263	268	275	282	286	291	295	18176953
Deaths - Ga	180	182	185	189	193	195	198	201	23948251
Deaths - M	47	48	51	53	54	56	57	58	9184739
Deaths - Le	1055	1089	1171	1252	1296	1323	1358	1398	53022727
Deaths - He	102	108	118	126	130	135	139	143	4704784
Deaths - M	108	109	110	113	115	116	118	119	15225951
Deaths - Ot	16	17	17	17	17	17	17	17	7385957
Deaths - Br	526	538	560	582	599	610	622	635	95243324
Deaths - Pr	351	356	362	368	374	377	379	382	63918471
Deaths - Th	56	58	60	62	63	65	66	68	5936870
Deaths - St	2200	2238	2300	2375	2447	2502	2564	2624	1.59E+08
Deaths - Bl	263	264	267	271	273	275	277	278	31425460
Deaths - Ut	61	62	63	64	65	66	67	68	13400761
Deaths - Ov	79	80	82	84	85	86	88	90	26235486
Deaths - Ce	332	340	353	367	380	389	399	410	39151845
Deaths - Br	422	438	472	505	526	541	559	580	34551356
Deaths - No	996	1010	1040	1062	1069	1074	1081	1090	34444198
Deaths - Pa	138	137	137	139	140	139	139	139	63727984
Deaths - Es	529	535	546	560	575	584	595	606	76097800
Deaths - Te	3	4	4	5	5	6	6	6	1559146
Deaths - Na	66	67	69	72	73	74	75	75	10558391
Deaths - Ot	37	37	38	39	39	39	39	39	14047941
Deaths - Co	539	542	550	555	554	552	550	550	1.44E+08
Deaths - Ne	25	26	26	27	28	28	28	29	6837649
Deaths - Me	6	7	7	7	8	8	8	8	4290892

Figure 4: Transposed table and summation result

Since we only want to analyze the deaths caused by each cancer type, we can simply delete all the columns in the middle and form a final excel:



cancer typ	Total
Deaths - Li	75932699
Deaths - Ki	21884097
Deaths - Li	25140935
Deaths - Tr	2.81E+08
Deaths - La	18176953
Deaths - G	23948251
Deaths - M	9184739
Deaths - Le	53022727
Deaths - H	4704784
Deaths - M	15225951
Deaths - Ot	7385957
Deaths - Bi	95243324
Deaths - Pr	63918471
Deaths - Th	5936870
Deaths - St	1.59E+08
Deaths - Bl	31425460
Deaths - Ut	13400761
Deaths - O	26235486
Deaths - Ce	39151845
Deaths - Bi	34551356
Deaths - N	34444198
Deaths - Pa	63727984
Deaths - Es	76097800
Deaths - Te	1559146
Deaths - Na	10558391
Deaths - Ot	14047941
Deaths - C	1.44E+08
Deaths - N	6837649
Deaths - M	4290892

Figure 5: final excel

The above steps can be completed with excel easily.

3. Analysis and Results

3.1 Analysis

3.1.1 Data Visualization and Insights

Data visualization is a method to represent data and information using different kinds of visuals. The primary goal is to present complicated data in a way such that it is easier to understand and interpret. Viewers can easily understand the message such as trend, distribution, patterns and insights from the visuals.

In this part, Power BI, which is a microsoft interactive data visualization platform, is used for building a dashboard to visualize the data. In order to visualize the data and analyze them, we first import all the cleaned and transformed dataset to Power BI. Then, we can create relationships between the csv files in order to display the reports accurately and interactively. By adding multiple useful visuals to the page, a dashboard will finally be built. Below are the overview of the dashboard and details of the steps:

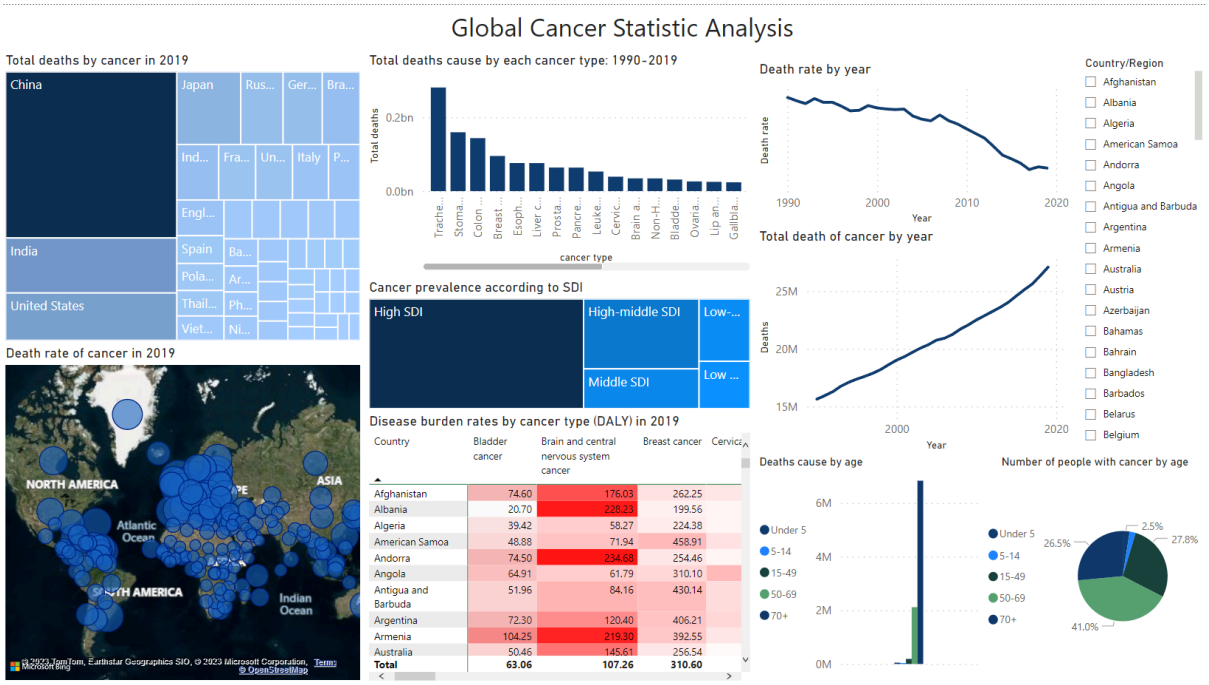


Figure 6: overview of the dashboard

The dashboard contains 9 different visuals, including heatmap, map, table, pie chart, etc. In addition, there is also a filter for selecting countries to analyze. The principle of designing data visualization is to keep everything simple. Therefore, in this dashboard, the color chosen is very simple, not more than 5 types of color so that the dashboard can be more clearer and consistent. Making the dashboard interactive is also important since we can have more important information on the dataset. For example:

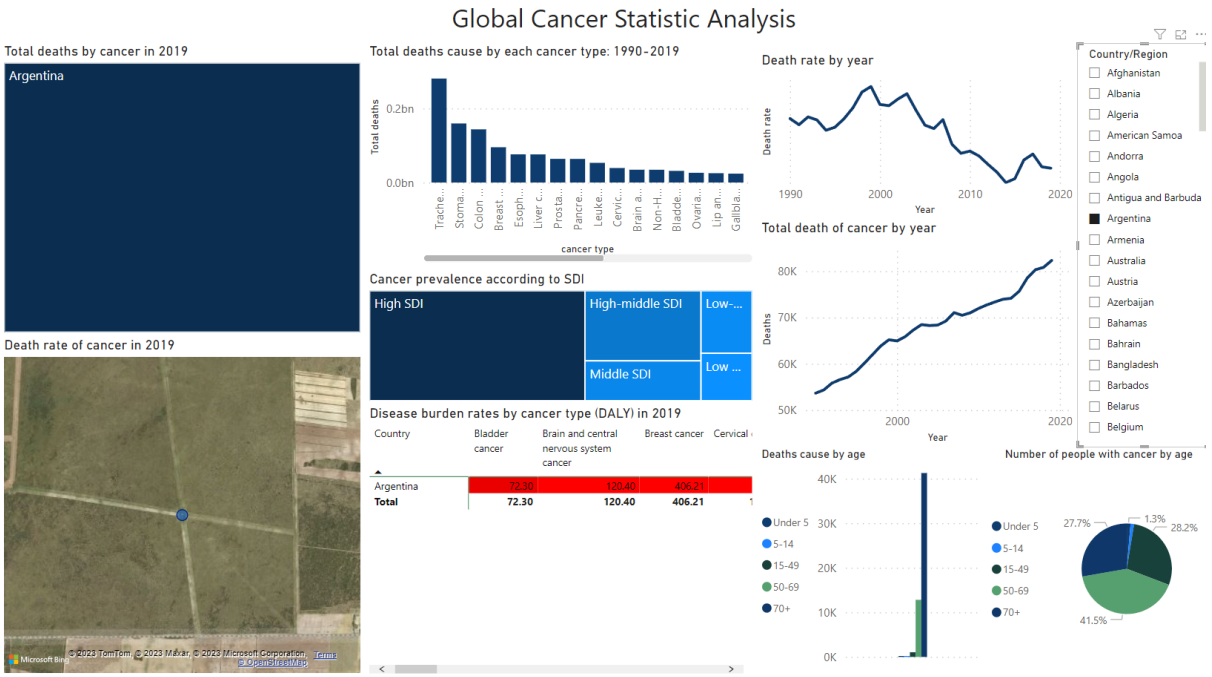


Figure 7: dashboard after randomly select a country

After selecting a country randomly on the filter, the visuals on the dashboard will interact with the filter and provide cancer information about that country more precisely. In order to do so, we need to manage the relationships between different tables, since there are multiple tables in the dataset. Power BI provides a function to manage the relationship. We can relate the entity, which is the country/region in these tables, between the tables. The type of relationship type is many-to-many.

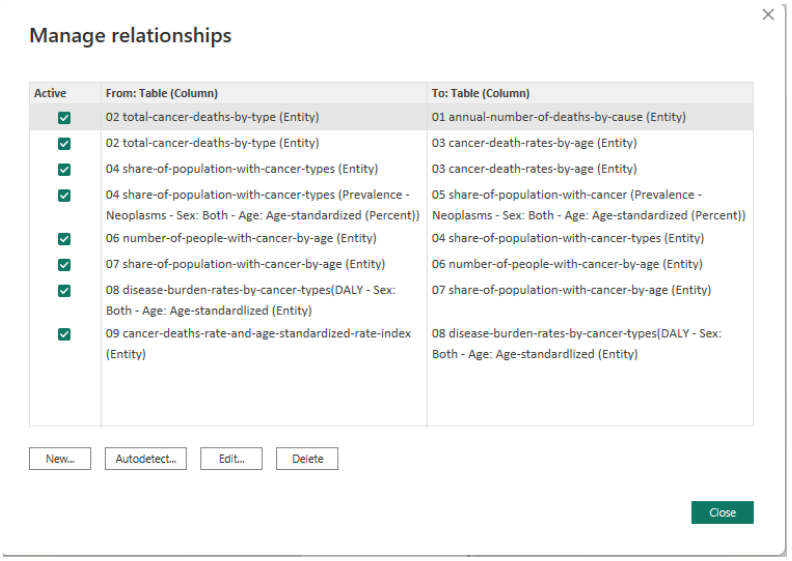


Figure 8: relationship between tables

Here are the details of each visuals:

1. Heatmap of total deaths by cancer in 2019:

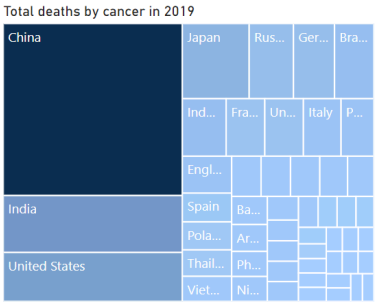


Figure 9: Heatmap

This visual shows the total deaths by cancer in 2019 of the countries, and the countries are sorted descending, where countries with most deaths have a larger area in the heatmap. The color is using the gradient format where darker means higher in value and lighter means lower in value.

2. Map of death rate of cancer in 2019:

Figure 10: Map

This visual is a map showing the death rate by a bubble of the countries/regions in the world. Larger bubble size means that the countries have a higher death rate of cancer. Users can use mouse scroller to zoom in or out in this visual.

3. Bar chart of Total deaths caused by each cancer type: 1990-2019:

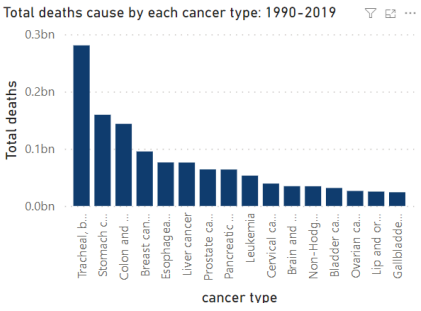


Figure 11: Bar chart

A bar chart showing the total deaths caused by each cancer type. The cancer type is sorted descending by the total deaths. To view the complete cancer type, users can place their mouse pointer on the bar.

4. Tables of disease burden rates by cancer type in 2019:

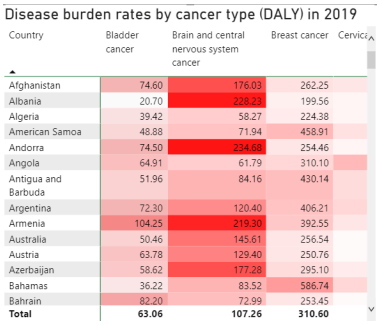


Figure 12: Table

This visual shows the burden rate of each cancer type of a country/region. Cells with redder background means that the country/region is having a high burden rate of that cancer type.

5. Heatmap of cancer prevalence according to DSI:

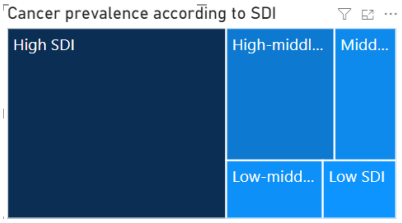


Figure 13: Heatmap

SDI refers to the Social Development Index. The heatmap shows the cancer prevalence according to SDI by the size of the rectangular box in heatmap and the color.

6. Line chart of death rate by year:

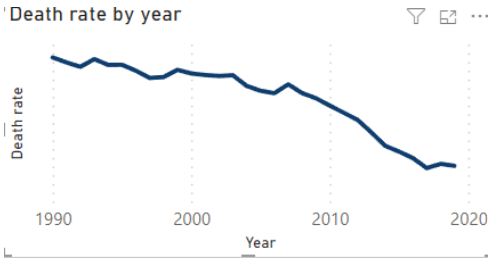


Figure 14: Line chart

This line chart shows the death rate of cancer with respect to year.

7. Line chart of total cancer deaths by year:

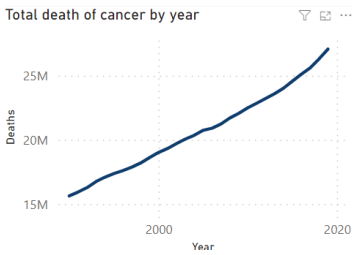


Figure 15: Line chart

This line chart shows the trend of death of cancer with respect to year.

8. Bar chart

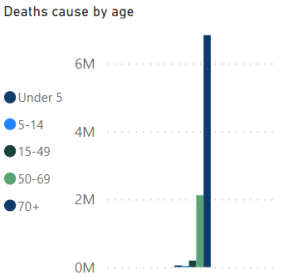


Figure 16: Bar chart

This bar chart shows the deaths caused by cancer of 5 age groups.

9. Pie chart

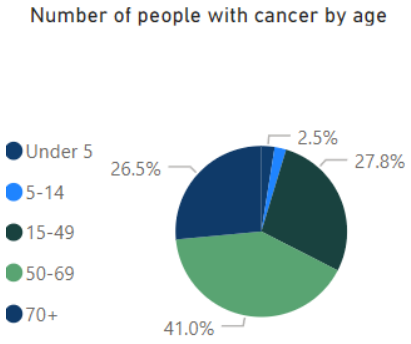


FIgure 17: Pie chart

This pie chart shows the distribution of people with cancer of 5 age groups.

3.1.2 Decision Tree

The Decision Tree is a type of supervised classification technique. It can be used to predict whether the cancer cells are benign or malignant, according to the following visual characteristics of the cancer: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry and fractal dimension. Each characteristic also has three statistical variants which are the mean, the standard error, and the worst. So we will build three decision trees according to the three variants with each having 10 attributes. Weka is used to build and visualize the decision tree.

In this dataset, 357 entries are diagnosed with Benign cancer while only 212 entries are diagnosed with malignant cancer, which makes the number of entries between the two classes highly imbalanced. A training set consisting of asymmetric numbers of data from the

two classes may result in a biased decision tree classifier towards the majority class, which in this case is the Benign cancer class. The classifier may assign more test cases to the majority class and lead to an optimistic accuracy estimation. Therefore, we need to restore balance on the training set by undersampling the Benign cancer class, this can prevent the bias from occuring in the first place.

Undersampling is the technique of removing entries of the larger class so that it is the same size as the smaller class. In this case, we used the pandas package in Python, which is a package that provides functions to work with data structures of relational data. We read in the .csv file, splitted it into two separate dataframes, randomly sampled 212 entries from the Benign cancer dataframe, then merged the two dataframes back together, and finally exported it back to a .csv file.

3.1.3 Regression

Regression techniques are crucial for the prediction and forecasting for the future trend of cancer. This technique aims to find the best fitting line which indicates the relationship between different variables based on the input data and we can use the predicted line to predict the future trend.

In order to avoid overfitting, which refers to the phenomenon which the prediction model is too fit to the given data but have a poor ability on predicting future observations effectively [7], as the amount of provided data points is limited, we just applied the polynomial regression in degree 1 (linear regression) and 2 (quadratic regression) to our data to enhance the performance and accuracy of our prediction. For linear regression, we measured the Pearson correlation coefficient as an evaluation metric [7]. The following is its mathematical formula:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where,

r = Pearson Correlation Coefficient

x_i = x variable samples

y_i = y variable sample

$\bar{x}$ = mean of values in x variable

$\bar{y}$ =mean of values in y variable

Figure 18: Mathematical formula of Pearson correlation coefficient (r) [8]

The closer it gets to 1/-1, the higher the linear correlation of the two variables will be. We only looked into the linear regression if $abs(r)$ > 0.8 to ensure the data that we investigated has a high correlation. As for quadratic regression, we applied the coefficient of determination, R^2 as another evaluation metric [9]. It ranges from 0 to 1. The more it closes to 1, the higher the quadratic correlation of the two variables will be. Below is the formula of it:

$$R^2 = 1 - \frac{RSS}{TSS}$$

R^2 = coefficient of determination
RSS = sum of squares of residuals
TSS = total sum of squares

Figure 19: Mathematical formula of coefficient of determination (R^2) [9]

We only looked into the quadratic regression if R^2 > 0.9 to ensure the data that we further investigated has a high regression correlation. In this project, we applied regression techniques on the number of deaths, death rate (Death rate per 100,000 people) , prevalence (percentage of the population affected by cancer) and the DALY (Disability-Adjusted Life Years) rate of cancer in the world versus time (year) in order to predict the future trend on cancer in the world. To investigate further, apart from only using the overall data of the world, we also used overall data obtained from countries having different Social Development Index (SDI), including high, Middle and Low SDI, and different incomes, including High, Upper Middle, Lower Middle and Low income to find out if there are any correlation between different SDIs/ incomes and cancer. We used numpy.polyfit() to facilitate the process of polynomial regression.

3.2 Results

3.2.1 Result on data visualization

After building the dashboard on Power BI, we can find out the pattern and trend in cancer statistics easily.

Firstly, we can observe that countries with higher populations have more total cancer deaths in general. In the heatmap of total deaths by cancer in 2019, China, India, US have the highest total deaths in cancer, and China has the highest since it has the largest population.

Besides, according to the map, the bubble size in Europe is relatively larger, therefore, we can conclude that the death rate in Europe is higher than other areas. Additionally, higher SDI areas have higher cancer prevalence than that of the lower SDI areas. Therefore, we can observe the locational distribution of cancer statistics by the above 3 visuals.

For age group, from the dashboard, we can see that for both death cases and people with cancer, people with higher age have much more percentage in both cases, while teenagers and children only have near zero single digit percentage in both visuals. Therefore, we can observe that higher age groups, especially elderly, have a higher chance of getting cancer and cancer is more deadly to them. This is due to their weaker immune system.

For the cancer trend, it is obvious that the total death of cancer is increasing linearly by years due to the increasing population. Although the total death of cancer is increasing, the death rate of cancer is decreasing year by year. This is probably due to the development and improvement of modern medical technology.

Last but not least, we can observe that the most common type of cancer is tracheal, bronchus, and lung cancer.

3.2.2 Result on decision tree

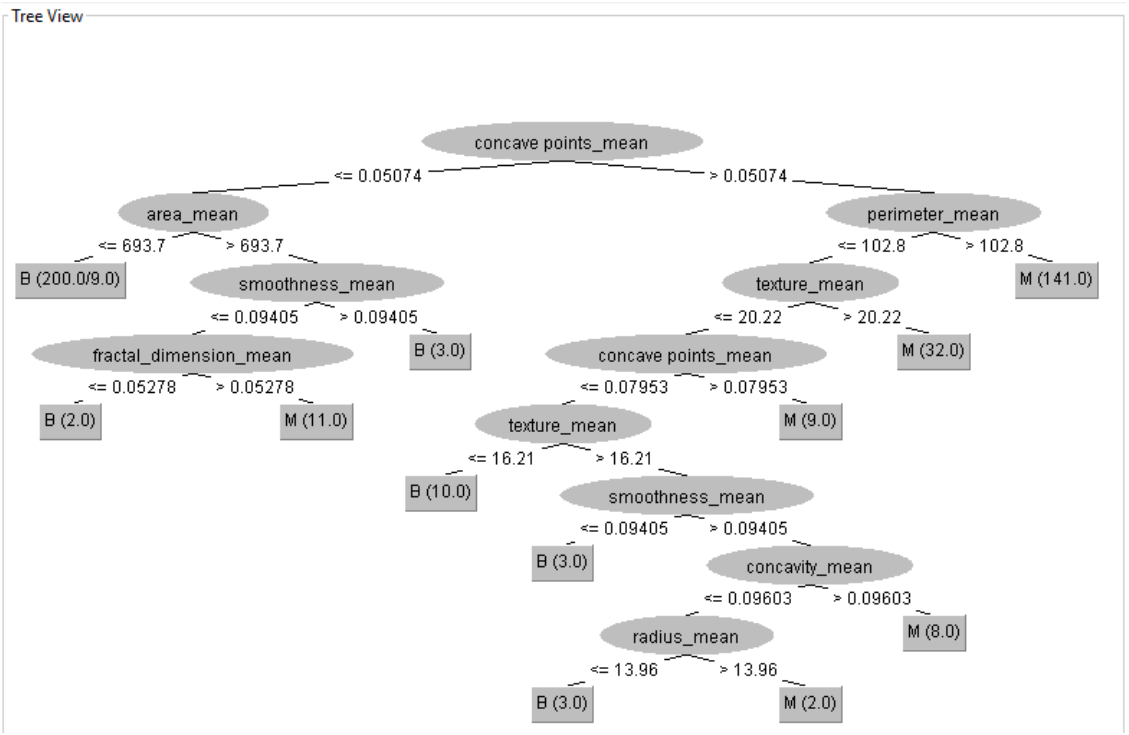


Figure 20: Decision tree using the mean metric

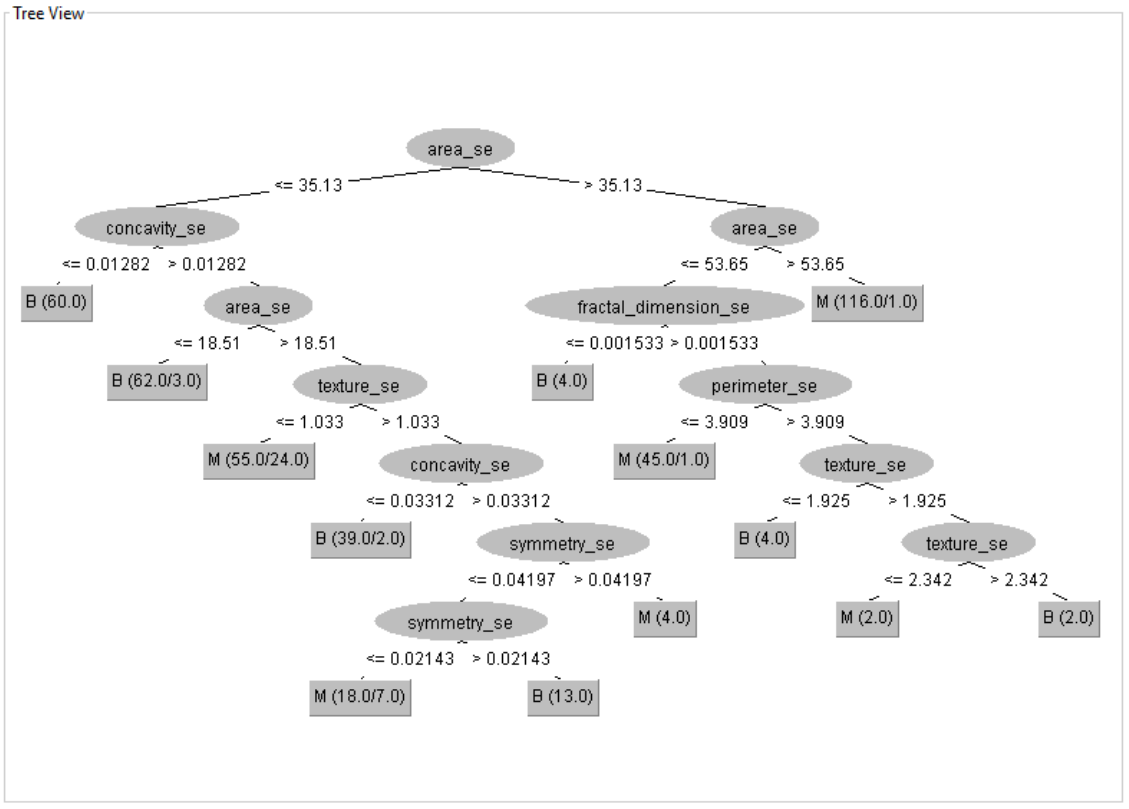


Figure 21: Decision tree using the SE metric

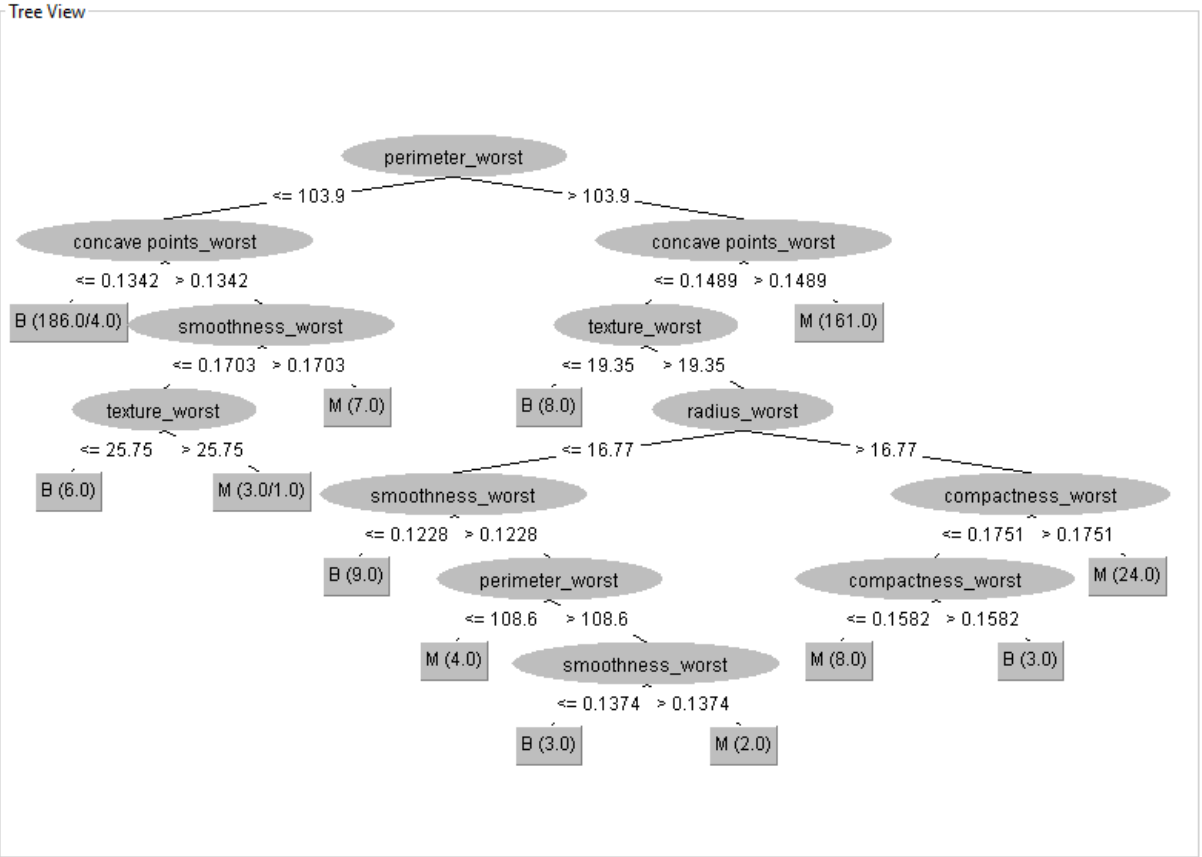


Figure 22: Decision tree using the worst metric

For evaluation, instead of only using the typical accuracy, we also evaluate the trees with the balanced accuracy. Which represents the result better as its use case is to deal with imbalanced data. The balanced accuracy reduces to the vanilla accuracy if the decision tree performs equally well on both classes. If the decision tree takes advantage of the imbalanced data set, the balanced accuracy will drop compared to the vanilla accuracy to reflect it.

$$\text{balanced accuracy} = \frac{sensitivity + specificity}{2}$$

where

$$sensitivity = \text{true positive rate} = \frac{TP}{TP + FN}$$

$$specifity = \text{true negative rate} = \frac{TN}{FP + TN}$$

Figure 23: balanced accuracy equations

We will also further calculate the precision and recall of each decision tree as additional metrics. Recall is an important metric in our case as the cost of false negatives, wrongly classifying a malignant cancer as benign, is high. The patient may miss the best time for treatment.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Figure 24: precision and recall equations

Decision Tree	Precision	Recall	Balanced Accuracy	Accuracy
Mean	0.890	0.956	0.926	0.924
SE	0.726	0.791	0.766	0.764
Worst	0.822	0.968	0.905	0.896

First of all, from the above table we can observe that the difference between the balanced accuracy and the vanilla accuracy is very small for all three decision trees, <0.01, which implies that the undersample technique is able to prevent the classifiers from building bias towards the larger class.

Secondly, the SE tree underperforms the other two trees in all three metrics. Which suggests that the SE metric may not be a faithful way to represent visual characteristics of a cancer cell when it comes to malignant cancer diagnosis. Both the mean metric and the worst metric performs well with >0.8 in all three metrics. Although the worst metric has a lower precision than the mean metric, it does not matter much in this case as the cost of false positive is low, the patient will only need to do a follow-up checking to give a better idea of their health condition. On the other hand, the recall is a much more important metric as the cost of missing an underlying malignant cancer is high, it may lead to delayed treatment, in this case the worst metric has a slightly better performance than the mean metric. But overall, the mean metric seems to be the best way to represent visual characteristics of a cancer cell as it performs similarly well in recall and accuracy when compared to the worst matric, but

has a higher precision which can help minimize unnecessary medical checkup costs as it produces less false positives.

Overall, the decision tree classifier does a good job on predicting whether a cancer cell is malignant, with an accuracy of over 90% especially when using visual characteristics of a cancer cell represented in the mean metrics.

3.2.3 Result on regression

Results and analysis for regression

After applying linear and quadratic regression, computing the sum of squares error for each degree, plotting both regression lines on the same graph which also contain original data and extending the plotted lines to the year of 2030 for future prediction, we find out in general, cancers have an increasing trend on number of deaths and prevalence, as well as a decreasing trend on death rate and DALY Rate in the future.

General Observations and assumptions:

In most of the cases, the regression results are having pearson correlation coefficient r > 0.8 and coefficient of determination R^2 > 0.9, showing that there is a strong correlation between number of deaths/death rate/prevalence/DALY rate of cancer and time. Also, we can see that in most of the cases, both linear and quadratic regression agree on the same trend. Since we are just interested in the future trend of cancer (either increasing or decreasing after 2020), we just focus on the results of quadratic regression, without focusing much on the linear regression result. For the case that both linear and quadratic regression had a high correlation, but linear and quadratic regression disagree on the future trends, e.g. quadratic regression suggests an increasing trend and linear regression suggests a decreasing trend in the future, we assume that no conclusion can be drawn but we will still try to interpret the results. It is also observed that the sum of squares error can be quite high sometimes, especially when the data in the y axis is large (like data of the number of deaths). It becomes reasonable when we apply regression on a rate instead of a large number over time.

Number of deaths (Data: World, different incomes)

By observing the results produced using the world data, It is discovered that all types of cancers are causing an increasing number of deaths over time, with the only exception of esophageal cancer. Also, this general trend applies to people of different incomes, with

several exceptions. For instance, there is a decreasing trend in the number of deaths caused by Hodgkin lymphoma in the high-income class suggested by linear regression.

```
Pearson Correlation Coefficient:  -0.8418354006130906

Correlation coefficient r^2:  [0.98811778]
```

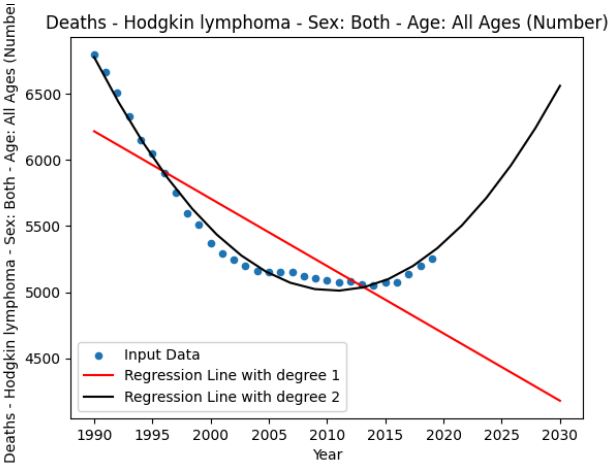


Figure 25: High income countries, Hodgkin lymphoma

One possible interpretation is there is a positive correlation between socioeconomic status and the survival rate regarding this specific disease [10]. However, since the prediction of quadratic regression suggests there will be an increasing trend, we are not confident to draw any conclusions here. There are also some similar cases in which the predictions from linear and quadratic regression disagree with each other, including stomach cancer, cervical cancer etc.

Death rate per 100,000 people (Data: World, different incomes)

It is discovered that the death rate per 100,000 people is generally decreasing over time in most groups of different ages (age <5, age 5-14, age 15-49, age 50-69, age 70+). Showing that although the number of deaths was increasing, the death rate was decreasing. One possible interpretation for this phenomenon is that the world's population is increasing, so the number of deaths caused by cancer also increases as more people are likely to get cancer and die of cancer. However, as there is improvement in medical technology, the fraction of people killed by cancer is reduced.

However, when we further investigate the chart of low income countries with age >70, the result is as follows:

```
Pearson Correlation Coefficient:  0.9780051905755514

Correlation coefficient r^2:  [0.97572859]
```

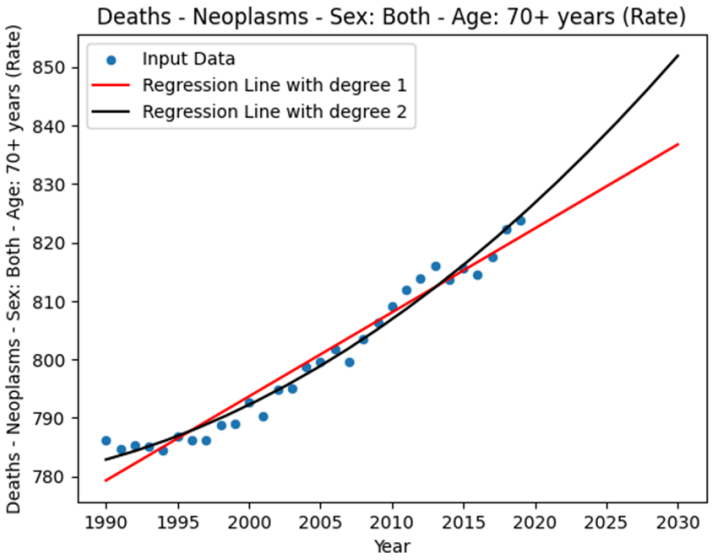


Figure 26: Low income countries, age >70

We noticed a steadily increasing trend in both linear and quadratic regression.

Another outcome that is also worth investigating is the age cluster 15-49 in the same income class. According to the quadratic regression result, the trend will be increasing after 2020. However, since the Pearson Correlation Coefficient in this case was high, and the linear regression suggests a decreasing trend, we cannot make concrete conclusions for this insightful observation.

```
Pearson Correlation Coefficient:  -0.8906612315030958

Correlation coefficient r^2:  [0.98486638]
```

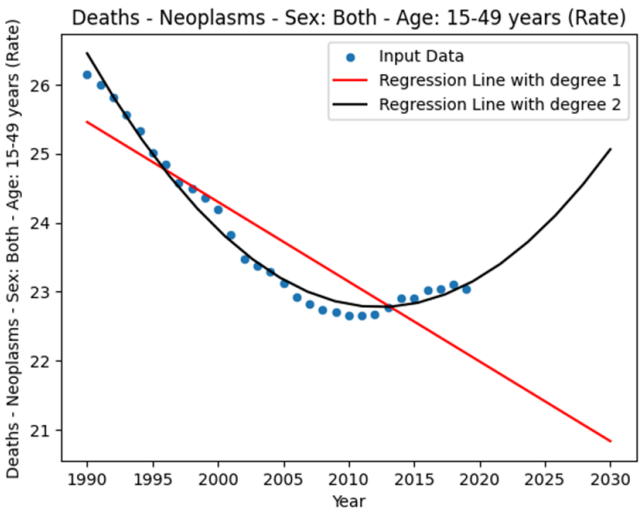


Figure 27: Low income countries, age 15-49

Prevalence (percentage of the population affected by cancer) (Data: World, different SDIs).

In general, the prevalence of cancer is increasing globally. Similar observations could be seen in countries with different SDIs, except high SDI countries with ages 15-49. The graph of it is shown below:

```
Pearson Correlation Coefficient:  0.9572300493240524

Correlation coefficient r^2:  [0.9662507]
```

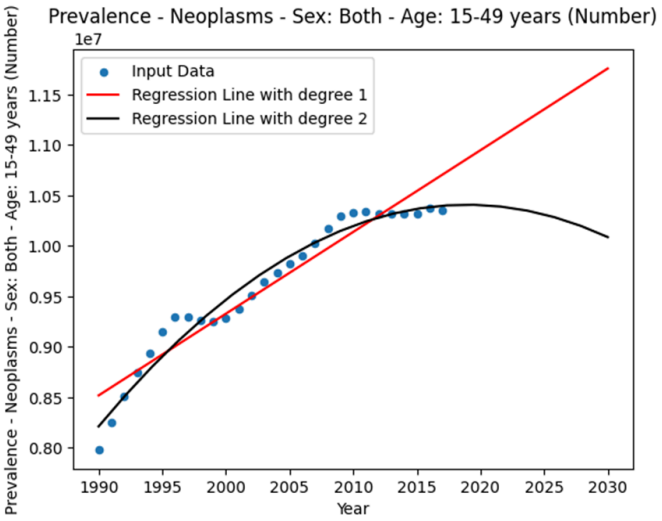


Figure 28: High SDI countries, age 15-49

From the above graph, we can conclude that the prevalence of cancer in this cluster has a decreasing/non-increasing trend. However, this conclusion is uncertain as both linear and quadratic regression suggest different future trends.

DALY Rate (Disability-Adjusted Life Years) (Data: World, different incomes)

It is discovered that the DALY rate of most types of cancers has a decreasing trend globally except Hodgkin lymphoma, it has an increasing trend over time.

Below is a summary for some key observations regarding DALY Rate:

Income of a country	Name of the disease which have increasing trend of DALY
High Income	Colun and rectum cancer, Hodgkin lymphoma
Upper Middle Income	Hodgkin lymphoma
Lower Middle Income	Breast, gallbladder, kidney cancer
Low Income	Breast cancer, nasopharynx, non-melanoma skin cancer, other malignant neoplasms, Hodgkin lymphoma, non Hodgkin lymphoma

Based on the observations, one preliminary conclusion that we can draw is that low income countries are more likely to have an increasing trend on DALY rate on different types of cancer compared with higher income countries.

4. Conclusion

In conclusion, we have built a dashboard which can visualize the global cancer data effectively. At the same time, we have also implemented regression techniques to verify the correlations between death rate, DALY rate, prevalence and time (year). Furthermore, as an extension of our work, we developed a decision tree which can predict whether the cancer cells are benign or malignant based on an extra dataset in order to gain more discerning insight related to cancer.

According to the visualization results in the dashboard and regression analysis, tracheal, bronchus, and lung cancer are the most prevalent types of cancer currently and it is believed that it is due to cigarette smoking and secondhand smoke exposure [11]. Besides, we can conclude that the population is directly proportional to the number of deaths caused by cancer, and the death rate of cancer in Europe is relatively higher. Although the number of deaths caused by cancer is increasing, the death rate of cancer is decreasing, mainly due to the improvement of modern medical technology. However, the overall prevalence of cancer has an increasing trend. At the same time, higher SDI areas tend to have higher cancer prevalence, which disagrees with our initial expectations that low SDI areas will have a higher cancer prevalence. We are also able to conclude that cancer has a more significant impact on the elderly in both prevalence and death rates, mainly owing to their weaker immune system.

Apart from that, based on the results derived from regression analysis, we are able to conclude that in general, cancers have an increasing trend of prevalence, as well as a decreasing trend in DALY Rate in the future. However, results obtained from regression are unsatisfactory as the number of data points is not sufficient, it is difficult to have an accurate prediction on future trends and only few conclusions could be drawn.

Moreover, the decision tree can predict whether a cancer cell is malignant with over 90% accuracy, which can act as a reliable preliminary diagnosis tool for potential cancer patients, especially potential cancer patients in areas where the mortality rate and prevalence are

high. Once their tumors are classified as malignant, they could seek medical assistance, such as cancer screening and diagnosis, as soon as possible. As a consequence, the adverse impact of cancer worldwide can be minimized ideally.

However, there is still room for improvement in our project. To get a more favorable insight related to cancer, one possible refinement is we can include other related datasets such as datasets containing cancer patient information. With the combination of such datasets and our chosen dataset, we can build a prediction model of cancer prognosis to try to estimate the odds of a person getting cancer. However, due to the limitation of our dataset, and because of our visualization-oriented focus, we regret that we are not able to include it in our work.

All in all, most of the derived results meet our preliminary expectation in our project proposal phase. We hope that the insight in our work can contribute to the reduction of adverse effects caused by cancer, one of the world's largest health problems.

5. Reference and Appendix

5.1 Reference

[1] Xiaomei M., Herbert Y. Global Burden of Cancer. Yale J Biol Med. From: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1994799/

[2] Freddie B., Mathieu L., Et al. The ever-increasing importance of cancer as a leading cause of premature death worldwide. CancerScope. From: https://acsjournals.onlinelibrary.wiley.com/doi/full/10.1002/cncr.33587

[3] Belatey H., Cancer and Deaths Dataset : 1990~2019 Globally. Kaggle. From: https://www.kaggle.com/datasets/belayethossainds/cancer-and-deaths-dataset-19902019-globally?resource=download&select=06+number-of-people-with-cancer-by-age.csv

[4] World Health Organization. Cancer - Screening and early detection. From: https://www.who.int/europe/news-room/fact-sheets/item/cancer-screening-and-early-detection-of-cancer

[5] Data Preprocessing in Data Mining. deepak_jain. Geeksforgeeks. From: https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/

[6] Erdem Taha., Cancer Data. Kaggle. From: https://www.kaggle.com/datasets/erdemtaha/cancer-data

[7] E. Alpaydin, Introduction to machine learning. The MIT Press, 2020

[8] A. Patil. Beginner's Guide to Pearson Correlation Coefficient. Analytics Vidhya. From: https://www.analyticsvidhya.com/blog/2021/01/beginners-guide-to-pearsons-correlation-coefficient/

[9] A. Kumar. R-squared, R2 in Linear Regression: Concepts, Examples. Data Analytics. From: https://vitalflux.com/r-squared-explained-machine-learning/

[10] C. Hung, C. Ou, H. Lai, Y. Chen, C. Lee, S. Li, Y. Su. High combined individual and neighborhood socioeconomic status correlated with better survival of patients with lymphoma in post-rituximab era despite universal health coverage. Journal of Cancer Research and Practice 3 (2016). pp.118-123

[11] Cancer. National Center for Chronic Disease Prevention and Health Promotion (NCCDPHP). From: https://www.cdc.gov/chronicdisease/resources/publications/factsheets/cancer.htm

5.2 Appendix

Google Colab notebook for dataset undersampling: https://colab.research.google.com/drive/149LOkqtBB4zmEzE3H7spSk1inu4CAzGM?usp=sharing

Google Colab notebook for regression analysis: https://colab.research.google.com/drive/1u4hh1EAWaGxUmhrOKEPR75_A6NWXbHLz?usp=sharing