CANCER PREVALENCE IN THE US

1. BUSINESS UNDERSTANDING

1.1 BUSINESS OVERVIEW

Cancer arises from the transformation of normal cells into tumor cells in a multi-stage process that generally progresses from a precancerous lesion to a malignant tumor. These changes are the result of the interaction between a person's genetic factors and three categories of external agents.
These are:

1. Physical carcinogens, such as ultraviolet and ionizing radiation.
2. Chemical carcinogens, such as asbestos, components of tobacco smoke, alcohol, aflatoxin (a food contaminant), and arsenic (a drinking water contaminant)
3. Biological carcinogens, such as infections from certain viruses, bacteria, or parasites.

Cancer is a leading cause of death worldwide, accounting for nearly 10 million deaths in 2020, or nearly one in six deaths.
In the United States in 2018, 1,708,921 new cancer cases were reported and 599,265 people died of cancer. For every 100,000 people, 436 new cancer cases were reported and 149 people died of cancer (CDC).

1.2 PROBLEM STATEMENT

In the US, the different types of cancer affect various groups differently. It has been observed that cancer incidences and deaths vary depending on race, age, gender and even State.
In 2018, 1,708,921 new cancer cases were reported and 599,265 people died of cancer in the US.
It is estimated that for every 100,000 people, 436 new cancer cases were reported and 149 people died of cancer in 2018.
By 2019, there were 599,601 cancer deaths; 283,725 were among females and 315,876 among males (CDC - Cancer).
The main aim of this study is to find out the most prevalent type of cancer across all States in the USA.

1.3 BUSINESS OBJECTIVES

1. To determine the most prevalent type of cancer across all States (incidences).
2. To determine which Gender is most diagnosed with cancer (incidences).
3. To determine which State has the highest number of cancer incidences.
4. To determine which cancer type is prevalent in this State.
5. To determine which State has the highest number of cancer deaths
6. To determine which cancer type is prevalent in this State.
7. To estimate the average number of deaths and cancer incidences for the year 2022.

1.4 PROJECT PLAN

This project is going to follow the six phases of the Cross Industry Standard Process For Data Mining (CRISP - DM).
The JIRA Kanban board will be used to assign, track and manage the different tasks involved.
Link

1.5 PROJECT JUSTIFICATION

The effects of cancer are deep as it does not only affect patients but also family and friends who are the main caregivers to cancer patients.
Cancer has adverse Social as well as Financial impacts on patients, family and friends.
The emotional effects of cancer often lead the patients into distress, and sometimes eventually, death. This leaves most caregivers seeking emotional support for not only the patients but also themselves.
Cancer treatment and management, i.e Chemotherapy, has left many patients financially burdened as they have to constantly pay for treatment as well as transport costs to and from the healthcare facilities.
According to a report released by the National Cancer Institute in 2019, NIH - Report, it is seen that the national patient economic burden associated with cancer care was $21.09 billion, made up of patient out-of-pocket costs of $16.22 billion and patient time costs of $4.87 billion.
In the US, there are 71 NCI-Designated Cancer Centers, located in 36 states and the District of Columbia NCI-Designated Cancer Centers. These are seen to sufficiently serve patients across the States.
In a study done by The American Society of Clinical Oncology (ASCO) in 2020, it was found that there are 12,940 oncologists practicing in the United States. The article indicated that the workforce is concentrated in nonrural areas, with only 11.6% of oncologists practicing in rural areas ASCO Report. This indicates that cancer patients in rural areas have less access to oncologists.
In 2020, an estimated 1,806,590 new cases of cancer were diagnosed in the United States and 606,520 people died from the disease.

By finding out the percentage rise of cancer cases among children, we will help the US government investigate the probable causes for the same and come up with preventive measures that could reduce the infection rate for future generations.

2. DATA UNDERSTANDING.

In this study we used four datasets.
These are sample datasets for the period 2014 - 2019.
The datasets used in this study are from the American Cancer Society
(Cancer Stats Center).

1. US Cancer Death Rates Link

This dataset contains information on deaths that occurred as a result of cancer infections within the period 2015 - 2019.

2. US Cancer Incidences Rates Link

This dataset contains information on cancer diagnoses that occured in the USA in the period 2014 - 2018. We used sheet one and two of this dataset.

3. US Estimated New Cases Link

This dataset gives an estimation of different types of cancer incidences in the USA for the year 2022.

4. US Estimated Cancer Deaths Link

This dataset gives an estimation of deaths caused by different types of cancer in the USA for the year 2022.

2.1 DESCRIBING AND EXPLORING DATA

Below is the description for the above datasets with the fields explained:
The first data shows information about; number of cases, total population of those affected, rate of infections of different age groups, gender and race. It also shows different types of cancers and how the infections are manifested by age, gender and race.
The second dataset shows estimated deaths by type of cancer and gender.

2.11. Fields of the DeathRate dataset.

This dataset illustrates **death rates** of cancer patients per 100, 000 people in the US.
The column names show deaths caused by the different types of cancer in either / both genders.

1. State
2. All cancer types combined / Both sexes combined
3. All cancer types combined / Female
4. All cancer types combined / Male
5. Brain and other nervous system / Both sexes combined
6. Brain and other nervous system / Female
7. Brain and other nervous system / Male
8. Breast / Both sexes combined
9. Breast / Female
10. Breast / Male
11. Cervix / Both sexes combined
12. Cervix / Female
13. Cervix / Male
14. Colorectum / Both sexes combined
15. Colorectum / Female
16. Colorectum / Male
17. Esophagus / Both sexes combined
18. Esophagus / Female
19. Esophagus / Male
20. Hodgkin lymphoma / Both sexes combined
21. Hodgkin lymphoma / Female
22. Hodgkin lymphoma / Male
23. Kidney and renal pelvis / Both sexes combined
24. Kidney and renal pelvis / Female
25. Kidney and renal pelvis / Male
26. Larynx / Both sexes combined
27. Larynx / Female
28. Larynx / Male
29. Leukemia / Both sexes combined
30. Leukemia / Female
31. Leukemia / Male
32. Liver and intrahepatic bile duct / Both sexes combined
33. Liver and intrahepatic bile duct / Female
34. Liver and intrahepatic bile duct / Male
35. Lung and bronchus / Both sexes combined
36. Lung and bronchus / Female
37. Lung and bronchus / Male
38. Melanoma of the skin / Both sexes combined
39. Melanoma of the skin / Female

40. Melanoma of the skin / Male
41. Myeloma / Both sexes combined
42. Myeloma / Female
43. Myeloma / Male
44. Non-Hodgkin lymphoma / Both sexes combined
45. Non-Hodgkin lymphoma / Female
46. Non-Hodgkin lymphoma / Male
47. Oral cavity and pharynx / Both sexes combined
48. Oral cavity and pharynx / Female
49. Oral cavity and pharynx / Male
50. Ovary / Both sexes combined
51. Ovary / Female
52. Ovary / Male
53. Pancreas / Both sexes combined
54. Pancreas / Female
55. Pancreas / Male
56. Prostate / Both sexes combined
57. Prostate / Female
58. Prostate / Male
59. Stomach / Both sexes combined
60. Stomach / Female
61. Stomach / Male
62. Testis / Both sexes combined
63. Testis / Female
64. Testis / Male
65. Thyroid / Both sexes combined
66. Thyroid / Female
67. Thyroid / Male
68. Urinary bladder / Both sexes combined
69. Urinary bladder / Female
70. Urinary bladder / Male
71. Uterine corpus / Both sexes combined
72. Uterine corpus / Female
73. Uterine corpus / Male

2.12. Fields of the IncRate dataset - Sheet one (All US).

These columns show a summary of the incidences of the different types of cancer in either or both genders.

1. Cancer Type

2. Both sexes combined
3. Female
4. Male

2.13. Fields of the IncRate dataset - Sheet two (State).

This dataset gives the records for **incidences / diagnoses rates** of cancer patients per 100, 000 people in the US in the period 2014 - 2018.
The column names show diagnosis caused by the different types of cancer in either / both genders.

1. State
2. All cancer types combined / Both sexes combined
3. All cancer types combined / Female
4. All cancer types combined / Male
5. Brain and other nervous system / Both sexes combined
6. Brain and other nervous system / Female
7. Brain and other nervous system / Male
8. Breast / Both sexes combined
9. Breast / Female
10. Breast / Male
11. Cervix / Both sexes combined
12. Cervix / Female
13. Cervix / Male
14. Colon (excluding rectum) / Both sexes combined
15. Colon (excluding rectum) / Female
16. Colon (excluding rectum) / Male
17. Colorectum / Both sexes combined
18. Colorectum / Female
19. Colorectum / Male
20. Esophagus / Both sexes combined
21. Esophagus / Female
22. Esophagus / Male
23. Hodgkin lymphoma / Both sexes combined
24. Hodgkin lymphoma / Female
25. Hodgkin lymphoma / Male
26. Kidney and renal pelvis / Both sexes combined
27. Kidney and renal pelvis / Female
28. Kidney and renal pelvis / Male
29. Larynx / Both sexes combined
30. Larynx / Female

31. Larynx / Male
32. Leukemia / Both sexes combined
33. Leukemia / Female
34. Leukemia / Male
35. Liver and intrahepatic bile duct / Both sexes combined
36. Liver and intrahepatic bile duct / Female
37. Liver and intrahepatic bile duct / Male
38. Lung and bronchus / Both sexes combined
39. Lung and bronchus / Female
40. Lung and bronchus / Male
41. Melanoma of the skin / Both sexes combined
42. Melanoma of the skin / Female
43. Melanoma of the skin / Male
44. Myeloma / Both sexes combined
45. Myeloma / Female
46. Myeloma / Male
47. Non-Hodgkin lymphoma / Both sexes combined
48. Non-Hodgkin lymphoma / Female
49. Non-Hodgkin lymphoma / Male
50. Oral cavity and pharynx / Both sexes combined
51. Oral cavity and pharynx / Female
52. Oral cavity and pharynx / Male
53. Ovary / Both sexes combined
54. Ovary / Female
55. Ovary / Male
56. Pancreas / Both sexes combined
57. Pancreas / Female
58. Pancreas / Male
59. Prostate / Both sexes combined
60. Prostate / Female
61. Prostate / Male
62. Rectum / Both sexes combined
63. Rectum / Female
64. Rectum / Male
65. Stomach / Both sexes combined
66. Stomach / Female
67. Stomach / Male
68. Testis / Both sexes combined
69. Testis / Female
70. Testis / Male

71. Thyroid / Both sexes combined
72. Thyroid / Female
73. Thyroid / Male
74. Urinary bladder / Both sexes combined
75. Urinary bladder / Female
76. Urinary bladder / Male
77. Uterine corpus / Both sexes combined
78. Uterine corpus / Female
79. Uterine corpus / Male

2.14. Fields of the New Case Estimates dataset.

This dataset gives the records for **new cases estimates** of cancer patients per 100, 000 people in the US in 2022.
The column names show estimated incidence rates of different types of cancer in both genders.

1. State
2. All cancer types combined
3. Acute lymphocytic leukemia
4. Acute myeloid leukemia
5. Anus, anal canal and anorectum
6. Bones and joints
7. Brain and other nervous system
8. Breast
9. Cervix
10. Chronic lymphocytic leukemia
11. Chronic myeloid leukemia
12. Colon (excluding rectum)
13. Colorectum
14. Digestive system
15. Endocrine system
16. Esophagus
17. Eye and orbit
18. Gallbladder and other biliary
19. Genital system
20. Hodgkin lymphoma
21. Kidney and renal pelvis
22. Larynx
23. Leukemia
24. Liver and intrahepatic bile duct
25. Lung and bronchus

26. Lymphoma
27. Melanoma of the skin
28. Mouth
29. Myeloma
30. Non-Hodgkin lymphoma
31. Oral cavity and pharynx
32. Other and unspecified primary sites
33. Other digestive organs
34. Other endocrine
35. Other leukemia
36. Other nonepithelial skin
37. Other oral cavity
38. Other respiratory organs
39. Ovary
40. Pancreas
41. Penis and other male genital
42. Pharynx
43. Prostate
44. Rectum
45. Respiratory system
46. Skin (excluding basal and squamous)
47. Small intestine
48. Soft tissue (including heart)
49. Stomach
50. Testis
51. Thyroid
52. Tongue
53. Ureter and other urinary organs
54. Urinary bladder
55. Urinary system
56. Uterine corpus
57. Vagina and other female genital
58. Vulva

2.15. Fields of the Death Estimates dataset

This dataset gives the records for **estimated deaths** due to cancer per 100, 000 people in the US in 2022.

The column names show diagnosis of different types of cancer in both genders.

1. State
2. All cancer types combined

3. Acute lymphocytic leukemia
4. Acute myeloid leukemia
5. Anus, anal canal and anorectum
6. Bones and joints
7. Brain and other nervous system
8. Breast
9. Cervix
10. Chronic lymphocytic leukemia
11. Chronic myeloid leukemia
12. Colorectum
13. Digestive system
14. Endocrine system
15. Esophagus
16. Eye and orbit
17. Gallbladder and other biliary
18. Genital system
19. Hodgkin lymphoma
20. Kidney and renal pelvis
21. Larynx
22. Leukemia
23. Liver and intrahepatic bile duct
24. Lung and bronchus
25. Lymphoma
26. Melanoma of the skin
27. Mouth
28. Myeloma
29. Non-Hodgkin lymphoma
30. Oral cavity and pharynx
31. Other and unspecified primary sites
32. Other digestive organs
33. Other endocrine
34. Other leukemia
35. Other non-epithelial skin
36. Other oral cavity
37. Other respiratory organs
38. Ovary
39. Pancreas
40. Penis and other male genital
41. Pharynx
42. Prostate

43. Respiratory system
44. Skin (excluding basal and squamous)
45. Small intestine
46. Soft tissue (including heart)
47. Stomach
48. Testis
49. Thyroid
50. Tongue
51. Ureter and other urinary organs
52. Urinary bladder
53. Urinary system
54. Uterine corpus
55. Vagina and other female genital
56. Vulva

2.2 VERIFYING DATA QUALITY

Missing data: Puerto Rico has no values recorded. Some other states i.e Alaska, Delaware, DC, Maine, New Hampshire, North and South Dakota, Rhode Island, Vermont and Wyoming also have missing data for some types of cancer. This makes every column have missing data.

For the third data set, columns with Cervical, Uterine, Testicular and Prostate cancers are null for the opposite gender.

Data errors: There are no typographical errors in all datasets.
Bad Metadata: There is no bad metadata.

3. DATA PREPARATION.

3.1. DATA SELECTION AND CLEANING.

In all datasets, there were Null values in every column, removing these columns would lead to a huge loss in valuable data.
Therefore, the following null column values were replaced with zero then dropped:

1. Acute lymphocytic leukemia
2. Acute myeloid leukemia

3. Anus, anal canal and anorectum
4. Bones and joints
5. Chronic lymphocytic leukemia
6. Chronic myeloid leukemia
7. Colon (excluding rectum)
8. Digestive system
9. Endocrine system
10. Eye and orbit
11. Gallbladder and biliary
12. Genital system
13. Lymphoma
14. Mouth
15. Other and unspecified primary sites
16. Other digestive organs
17. Other endocrine
18. Other leukemia
19. Other non epithelial skin
20. Other oral cavity
21. Other respiratory organs
22. Rectum
23. Respiratory system
24. Skin (excluding basal and squamous)
25. Small intestine
26. Soft tissue (including heart)
27. Tongue
28. Ureter and other urinary organs
29. Urinary system
30. Vulva

Columns with string values mixed with numeric values have been converted to numeric figures
Rows with no values have been filled with zeros.

4. ANALYSIS.

This project was analyzed using a python notebook using pandas and numpy libraries.
The graphics were created using matplotlib and

Below is attached a copy of the python notebooks that was used for the data preparation procedures and analysis:

https://github.com/nderitu-ndegwa/DSFT14_Grp2_Wk5

Also attached are the links to streamlit and jira kanban boards respectively.

https://share.streamlit.io/nderitu-ndegwa/dsft14_grp2_wk5/main/data_preview.py
https://nderituvincent.atlassian.net/jira/software/projects/DGW/boards/3

5. CONCLUSION

The most prevalent type of cancer is Breast cancer.
The gender that registered the highest number of cancer incidences was the male gender.
The state that had the highest number of cancer deaths was Carlifonia with 60970 deaths.
The most prevalent type of cancer in California is Lung and Bronchus with 9660 deaths.
Average number of estimated total deaths for 2022 is 609360.
The State that is estimated to have the highest number of incidence rates in 2022 is Kentucky

6. RECOMMENDATION

We recommend the US government to add more cancer equipment and gear up cancer awareness campaigns to mostly Kentucky which is projected to register the highest cases.
The government should also move quickly to invest more in cancer research with special interest on breast cancer, as the cancer incidence and deaths rates estimates are on an alarming upward trend as observed from the data.

Problem Definition
Objectives and goals
Project Plan
Data Sourcing
Data Preparation and Quality
Data Cleaning
Analysis
Conclusion, Recommendation, Next steps