Genetic Correspondence and Comparative Spatial Analysis with Age and Gender Specification of ICMR Data on Cancer Incidence in India

Sarbojit Das (231080075)*, Swapnonil Mondal (231080098)*, Sameer Verma (220949)*, Sujash Krishna Basak (231080093)*, Sayanta Biswas (231080080)*

*Department of Mathematics & Statistics, Indian Institute of Technology Kanpur, India

April 3, 2024

Abstract

Cancer remains a significant global health challenge, characterized by immense suffering and often limited treatment options. We aim to analyze genetic data related to five common cancer types, each presenting unique challenges and implications for diagnosis and treatment. Our goal is to identify the probable genetic causes underlying these cancers, providing leads for early identification and reducing fatality rates. Focusing on India, we explore the spatial and genetic aspects of cancer, utilizing statistical concepts and analyzing data from specific years. We aim to shed light on cancer prevalence, distribution and genetic underpinnings within the Indian population. Additionally, we investigate gender-specific and age-specific cancer incidence to offer detailed insights into the cancer landscape in India.

Table of contents

1 Introduction

Cancer is a formidable adversary to human health. It persists as a significant global burden, causing immeasurable suffering and presenting substantial challenges to healthcare systems worldwide. Its insidious nature lies in the uncontrolled growth and dissemination of abnormal cells, disrupting the delicate balance of cellular regulation inherent in the human body's natural processes. Despite remarkable advancements in medical science, the complexities of cancer remain a formidable challenge, often accompanied by limited treatment options and devastating outcomes for affected individuals and their families.

This project endeavors to delve into the multifaceted realm of cancer. With a diverse population and a unique set of demographic, environmental and genetic factors, India provides a rich tapestry for exploring the spatial, genetic, gender-specific and age-specific dimensions of cancer prevalence and distribution. Furthermore, our study extends beyond genetic analyses to explore the spatial aspects of cancer incidence in India. Leveraging statistical concepts and data from specific years, we seek to illuminate patterns of cancer prevalence and distribution across different regions of the country. This spatial perspective is crucial for understanding the geographic disparities in cancer burden and for informing targeted interventions and resource allocation efforts. Such insights hold the potential to inform tailored prevention, screening and treatment strategies, thereby improving outcomes and mitigating the impact of cancer on individuals and society at large.

Through this comprehensive exploration of cancer incidence in India, our project aims to contribute to a deeper understanding of the disease landscape. By integrating these aspects, we endeavor to shed light on the complexities of cancer and to ultimately strive towards a future where the burden of cancer is significantly reduced.

2 Dataset

In this project, we deal with multiple datasets that are based on cancer patients and the nature of cancer in both male and female human bodies.

1. Cancer causing genes dataset: The input dataset contains 802 samples corresponding to 802 people who have been detected with different types of cancer. Each sample contains expression values of more than 20K genes. Samples are categorized into one of the following types of tumors: BRCA, KIRC, COAD, LUAD and PRAD.

Following is an extended description of the tumors associated with this dataset:

- BRCA (Breast Cancer): Breast cancer is one of the most common cancers among women. It originates in the breast tissue and can occur in both men and women.
- KIRC (Renal Cancer): Renal cell carcinoma, or kidney cancer, occurs in the lining of small tubes in the kidney.
- COAD (Colon Cancer): Colorectal cancer includes colon cancer (affecting the large intestine) and rectal cancer (affecting the rectum). Colon cancer is a major cause of cancer-related deaths.
- LUAD (Lung Cancer - *Adenocarcinoma*): Lung *adenocarcinoma* is a subtype of non-small cell lung cancer. It originates in the cells lining the airways and is one of the most common types of lung cancer.

- PRAD (Prostate Cancer): Prostate cancer occurs in the prostate, a small gland in men that produces seminal fluid. It is one of the most common cancers in men.

2. Estimated State-wise cancer incidences: The table provides estimated incidence of cancer cases in India by States and Union Territory-wise for all sites and for both sexes.

3. Gender-wise different sites of cancer: The table presents the estimated cancer incidence, number of cases, crude rate, and cumulative risk by sex and anatomical sites in India for the year 2022.

Following is an extended description of the terminologies associated with this dataset:

- Cum-risk: Cumulative risk of developing cancer in the age range of 0 to 74 years. It represents the likelihood or probability of an individual developing cancer within this specified age range.

$$\boxed{\text{Cumulative Risk} = \frac{\text{Number of Cancer Cases}}{\text{Total Population}} \times 100}$$

- Crude Rate (CR): Crude rate refers to the total number of cancer cases occurring in a population divided by the total population, expressed as a rate per a specific unit of time (usually per $100,000$ population). It provides a general overview of cancer incidence within a population, without considering factors such as age distribution.

$$\boxed{\text{Crude Rate} = \frac{\text{Number of Cancer Cases}}{\text{Total Population}} \times 100,000}$$

- Age-Adjusted Rate (AAR): Age adjusted rate is a standardized rate that takes into account the age distribution of a population. It allows for fair comparisons of cancer rates between different populations or over time by removing the influence of age as a confounding factor.

$$\boxed{\text{Age-Adjusted Rate } = \sum_{i=1}^{n} \left(\frac{\text{Age-specific rate}_i \times \text{Population}_i}{\text{Total Population}} \right) \times 100,000}$$

- Malig Imn.Prol D (Malignant Immunoproliferative Diseases): This term refers to a group of disorders characterized by the abnormal proliferation of immune cells, leading to the development of malignancies. These diseases involve the uncontrolled growth of cells of the immune system, such as lymphocytes or plasma cells, and can include conditions like lymphomas, leukemias and multiple myeloma.

4. Age-wise cancer incidences: The table provides gender-disaggregated, estimated top five leading sites of cancer (%) in India by age group ($0-14, 15-39, 40-64$ & 65^{+} age groups) for the year 2022.

3 Data Pre-processing and Cleaning

We conducted data pre-processing and cleaning tasks on two datasets: "`data.csv`" and "`labels.csv`". These datasets contain information related to cancer research, focusing specially on genetic data and associated labels.

1. Essential Packages: For this project, we utilized several essential packages in `R`, including `tibble`, `tidyverse` and `dplyr`.

2. Pre-processing: This step is very crucial pertaining to the complexity and volume of genetic data involved.

- Data Importing: Our code reads the CSV files "`data.csv`" and "`labels.csv`" into `R` as tibbles using the `read.csv()` function. These datasets contain genetic data and label information related to cancer research.
- Exploratory Data Analysis (EDA): We performed EDA to gain insights into the structure and content of the datasets. This involved examining the structure of both datasets using the `str()` function and displaying the first few entries of each dataset using the `head()` function.
- Categorical Variable Conversion: We converted some columns in the datasets from *character* type to *factor* type using the `as.factor()` function. This was necessary for categorical variables. We also converted the final dataframe into a more structured format, *tibble* to facilitate efficient analysis.
- Data Merging: We merged the two datasets into a single dataframe using the `cbind()` function, combining the label information with the genetic data.
- Data Quality Checks: We checked for null values (`is.null()`), NA values (`is.na()`), and duplicated entries (`duplicated()`) in the merged dataframe. The absence of such issues indicates clean data.
- Summary Statistics: We calculated summary statistics for the genetic data to understand its distribution and characteristics using the `summary()` function.

Table 1: Summary statistics for Genes 1 and 2

gene_1	gene_2
Min. :0.000	Min. :0.000
1st Qu.:2.299	1st Qu.:2.390
Median :3.144	Median :3.127
Mean :3.011	Mean :3.095
3rd Qu.:3.883	3rd Qu.:3.803
Max. :6.237	Max. :6.063

The processed dataframe contains both label information and genetic data. It is saved in the "`.Rdata`" format for further analysis and modelling.

4 Data Insights

The data we have, has 20, 534 columns with 800 rows, as mentioned in the description. And distribution of classes is given below:

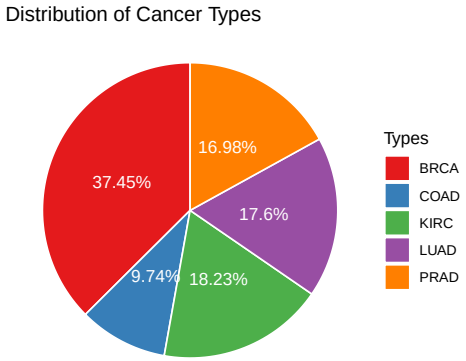


Figure 1: Proportion of Classes

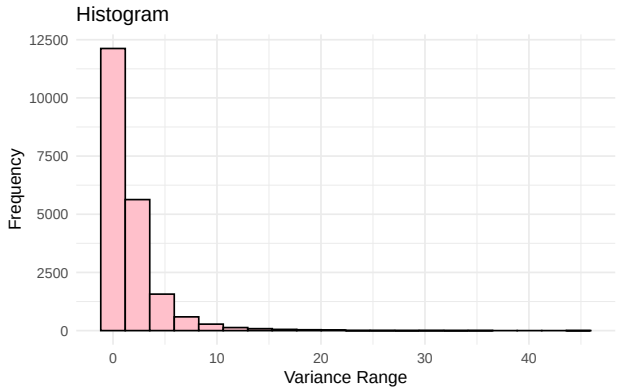


Figure 2: Histogram of Variance Range

From the pie chart, we can see that instances for class BRCA are more than that of others followed by KIRC, LUAD and PRAD and class COAD has the least number of instances.

- Variance Threshold: From the graph, it's easy to see that most of the columns have low variance. Features with high variance have data points that are spread out over a wide range. These features are considered significant because they contain valuable information that can help in discriminating between different classes or categories. Similarly, features with low variance have data points that have little variation. Such features may not carry much discriminative power and can be considered less significant.
- Distribution of average gene expression levels: We calculated the average expression values for each gene across all samples, sorting them in descending order based on these averages. The top 5 genes with the highest average expression values are displayed in a tabular format.

Table 2: Genes with highest mean expression value

Gene	Expression
gene_230	16.43044
gene_5380	16.38196
gene_232	15.96799
gene_18570	15.77775
gene_6857	15.71459

Following this, we generated a histogram and density plot to illustrate the distribution of average gene expression values.

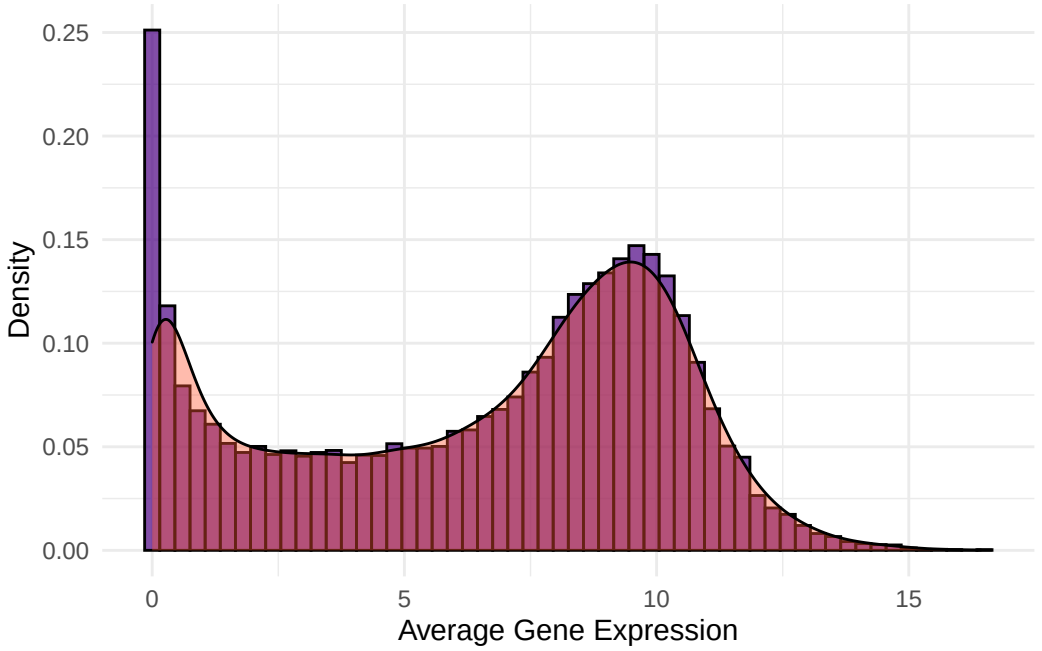


Figure 3: Distribution of mean gene expression levels

Boxplot of cases by gender:

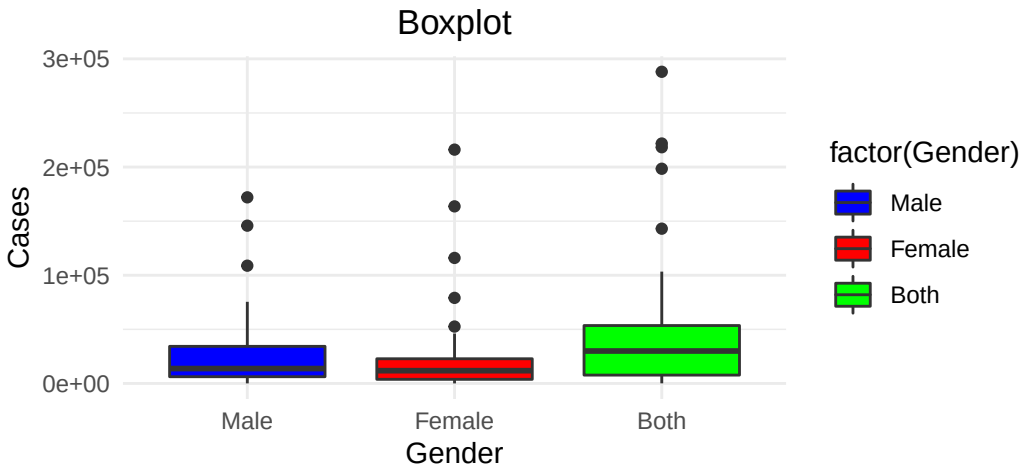
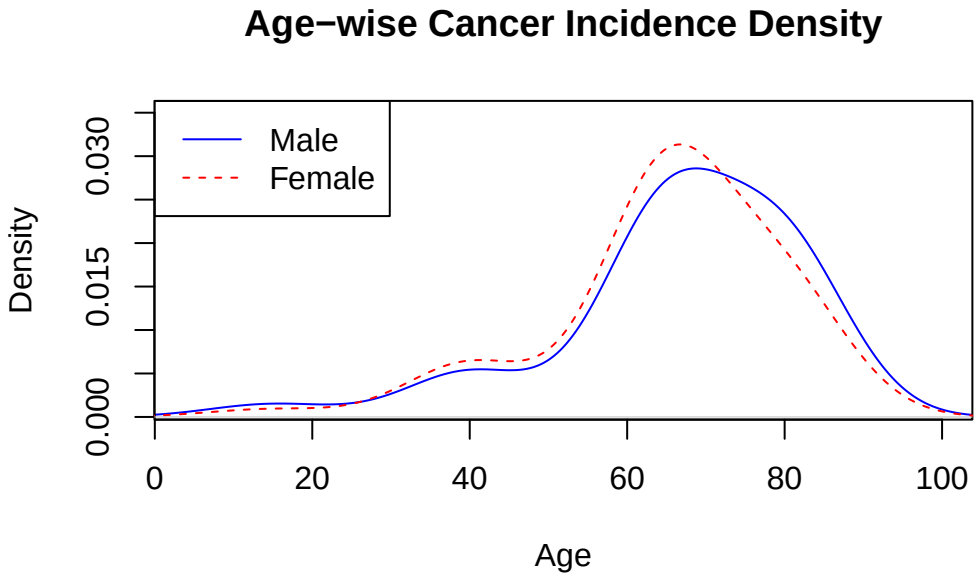


Figure 4: Cases by Gender

5 Visualizations

5.1 Age specific gender-wise cancer incidences

The shapes of the density curves represents the leptokurtic curve for females and mesokurtic curve for males.



5.2 Gender-wise top ten cancer sites

Leading Cancer Sites Among Females:

- Breast (17.89%): Major risk factors include female gender, age, family history, genetic mutations (e.g., BRCA1 and BRCA2), hormonal factors (e.g., early menarche, late menopause), reproductive history and lifestyle factors.
- Genital System (13.55%): This includes cancers of the cervix, uterus, ovaries, vagina and vulva. Risk factors vary by site but may include HPV infection, sexual activity, smoking, hormonal factors and genetic predisposition.
- Digestive System (9.60%): This includes dietary habits, alcohol consumption, tobacco use and chronic conditions such as obesity and *gastroesophageal reflux disease* (GERD).
- Uterine Cervix (6.55%): HPV infection is the primary risk factor, along with smoking, early sexual activity, multiple sexual partners and immuno-suppression.
- Oral Cavity and Pharynx (4.35%): Risk factors include including tobacco and alcohol use, HPV infection and poor oral hygiene.
- Ovary (3.82%): Risk factors include age, family history, nulliparity, infertility, hormonal factors and possibly endometriosis.

- Respiratory System (2.83%) - Lung and Bronchus (2.31%): Similar risk factors to those in males, primarily smoking and occupational exposures.
- Uterine Corpus (2.31%): Risk factors include hormonal factors (e.g., estrogen exposure), obesity, diabetes and certain genetic syndromes.
- Endocrine System (2.31%): This includes cancers of the thyroid, adrenal glands and other endocrine organs. Risk factors vary by site but may include radiation exposure, family history and certain genetic syndromes.

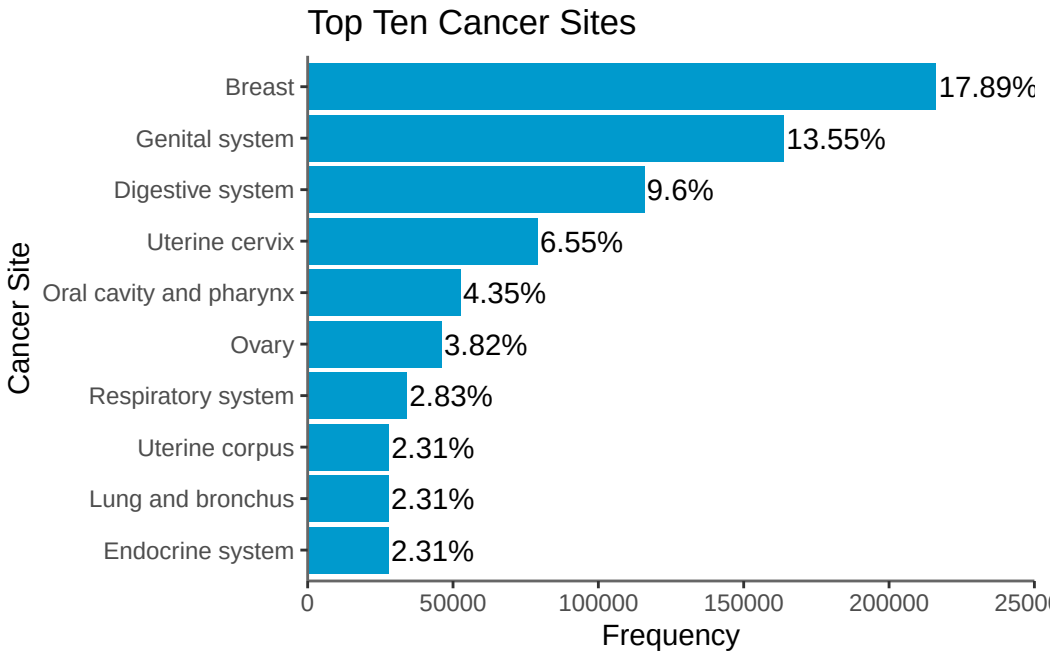


Figure 5: Top Ten Cancer Sites of Female

Leading Cancer Sites Among Males:

- Digestive System (13.06%): This includes cancers of the esophagus, stomach, liver, pancreas and colorectal region. Risk factors may include tobacco and alcohol use, dietary habits (e.g., high intake of processed meats) and chronic conditions such as *gastroesophageal reflux disease* (GERD) and *inflammatory bowel disease* (IBD).
- Oral Cavity and Pharynx (11.07%): Major risk factors include tobacco use (both smoking and smokeless tobacco), heavy alcohol consumption, *human papillomavirus* (HPV) infection and poor oral hygiene.
- Respiratory System (8.26%) - Lung and Bronchus (5.73%): Smoking, including exposure to secondhand smoke, is the primary risk factor for lung cancer. Occupational exposures to carcinogens such as asbestos, radon and diesel exhaust can also contribute.
- Mouth (4.57%): Similar risk factors to oral cavity and pharynx cancers, including tobacco and alcohol use, HPV infection and poor oral hygiene.

- Genital System (4.15%) - Prostate (3.32%): Prostate cancer is influenced by age, family history, and possibly dietary factors. Genetic predisposition and hormonal factors, particularly testosterone, play significant roles.
- Tongue (3.18%) - Other Oral Cavity (3.09%): Risk factors are similar to those for oral cavity and pharynx cancers, including tobacco and alcohol use, HPV infection and poor oral hygiene.
- Urinary System (2.65%): This includes cancers of the bladder, kidney and other urinary organs. Risk factors include smoking, occupational exposures, certain medications and genetic factors.

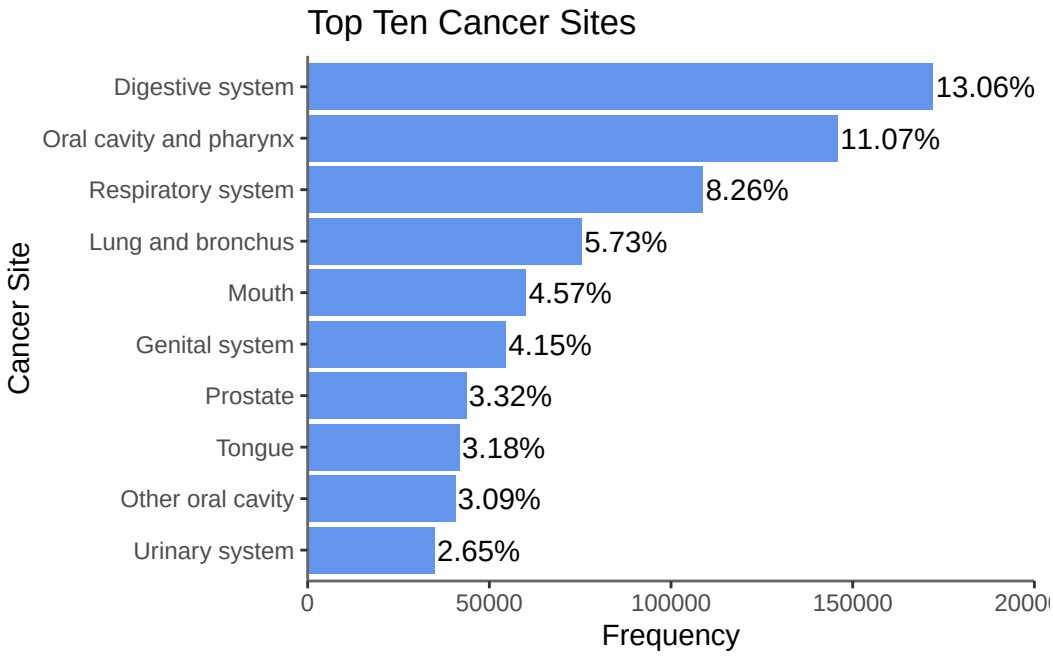


Figure 6: Top Ten Cancer Sites of Male

6 Spatial Analysis

- Northern and eastern parts of India show lower cancer incidence compared to regions like Uttar Pradesh, Maharashtra, Bihar, West Bengal, Hyderabad and Tamil Nadu. Hilly areas and Andaman Nicobar islands exhibit lower cancer rates compared to the plains. Western and southeastern parts of India demonstrate moderate cancer incidence.
- Uttar Pradesh stands out as having the highest cancer incidence among Indian states. Maharashtra, Bihar, West Bengal, Hyderabad and Tamil Nadu also exhibit high numbers of cancer cases.
- Factors such as pollution, industrial activities and exposure to carcinogens may contribute to higher cancer rates in certain regions.
- Clean and less polluted environments in hilly areas may contribute to lower cancer rates. Differences in lifestyle patterns, including dietary habits and physical activity levels, may also play a role in the observed regional disparities.

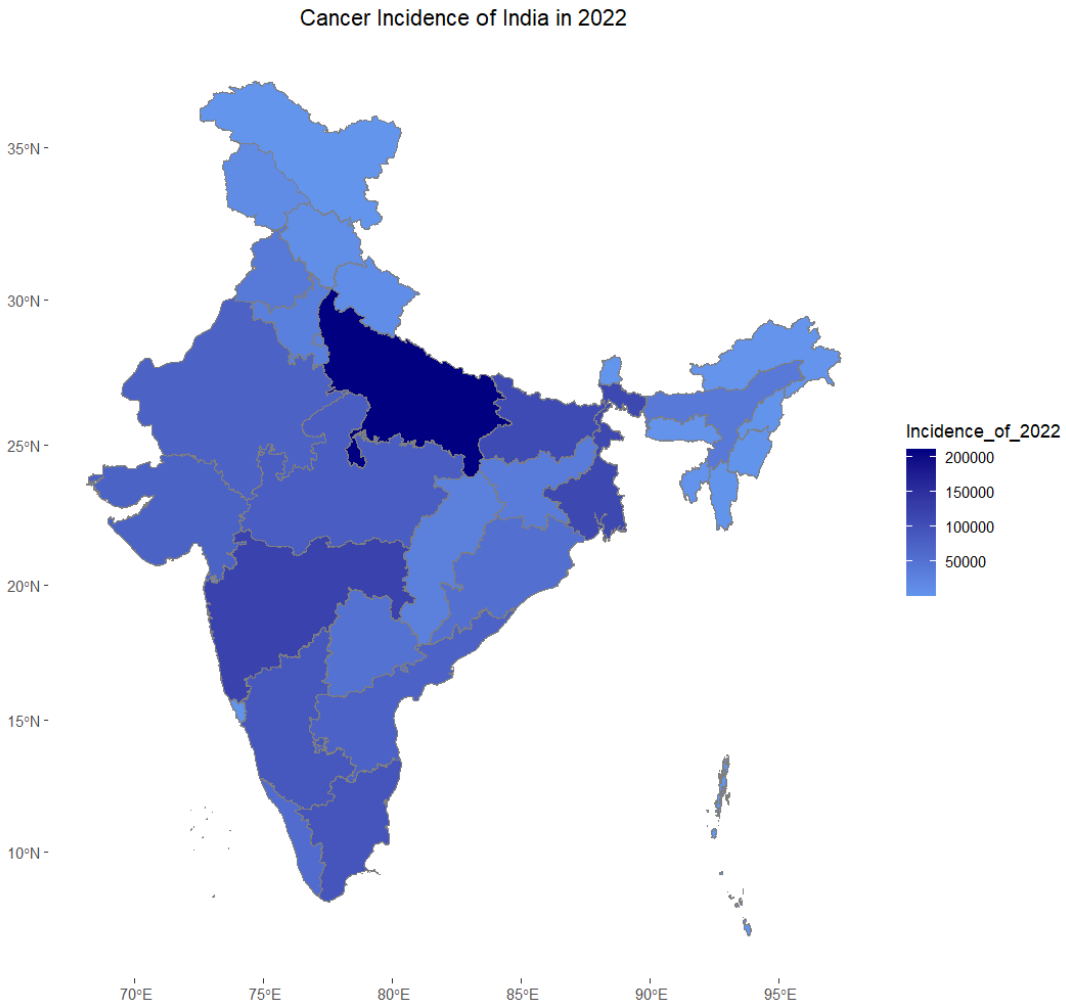
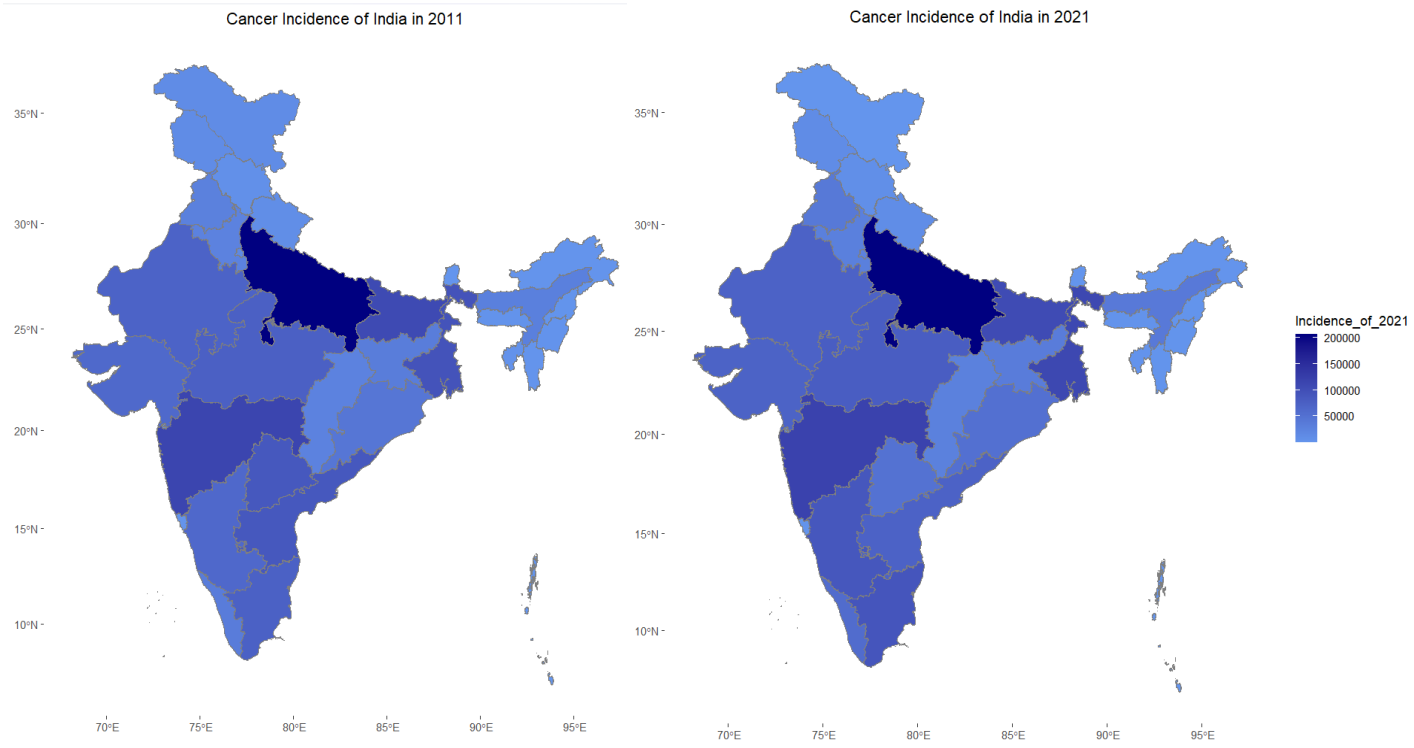


Figure 7: Cancer Incidence of India in the year 2022

We performed a comparative 10 year gap analysis with the cancer incidences in years, 2011 and 2021 respectively.

- We observed that Karnataka had improved over in cancer eradication whereas Hyderabad plunged to an increment in the number of cancer cases. The Telangana state, formed in 2 June, 2014, had less amount of cancer cases in 2021. Ladakh was formed as a Union Territory in 31 October, 2019. It showed few cancer patients in 2021, possibly due to its smaller population ratio.
- West Bengal displayed very slight changes in the cancer incidence ratio (i.e., $\frac{\text{Total no. of cancer patients in that state}}{\text{Total no. of cancer patients in India}}$) over the 10-year period.
- In a quick snapshot, there is a massive increment in the total number of cancer patients nationwide from approximately $160,000$ in 2011 to around $200,000$ in 2021. This substantial increase underscores the importance of heightened attention from both the public and government sectors towards cancer eradication efforts and mass awareness campaigns.



(a) Year 2011

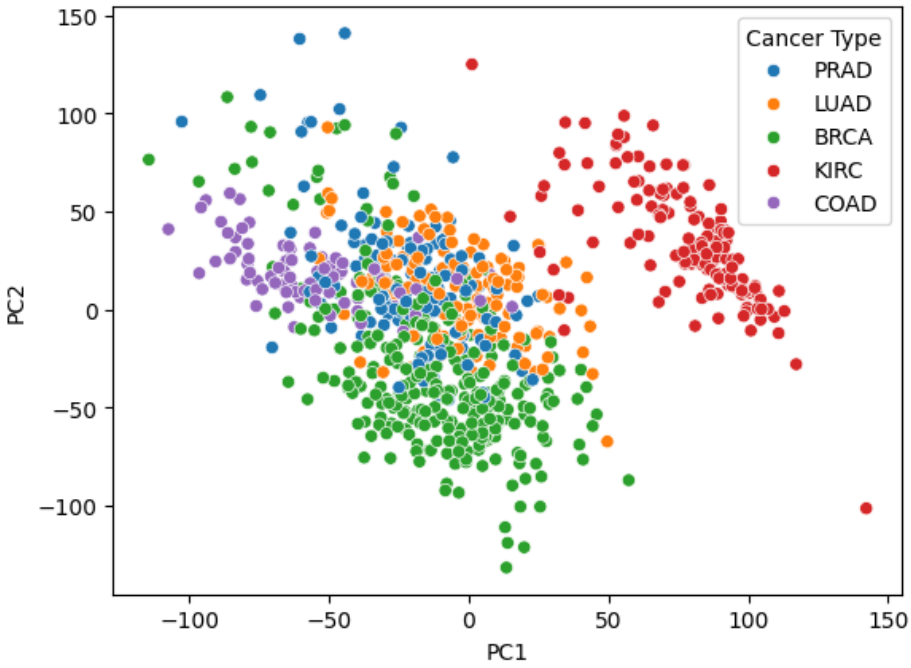
(b) Year 2021

Figure 8: A comparative 10 year gap cancer incidence in India

7 Principal Component Analysis

Having high dimensionality or feature space can lead model to perform poorly known as the "*Curse of Dimensionality*". To deal with this, we reduce the feature space by performing the dimension reduction technique, Principal Component Analysis (PCA) and we consider 95% variance explainability.

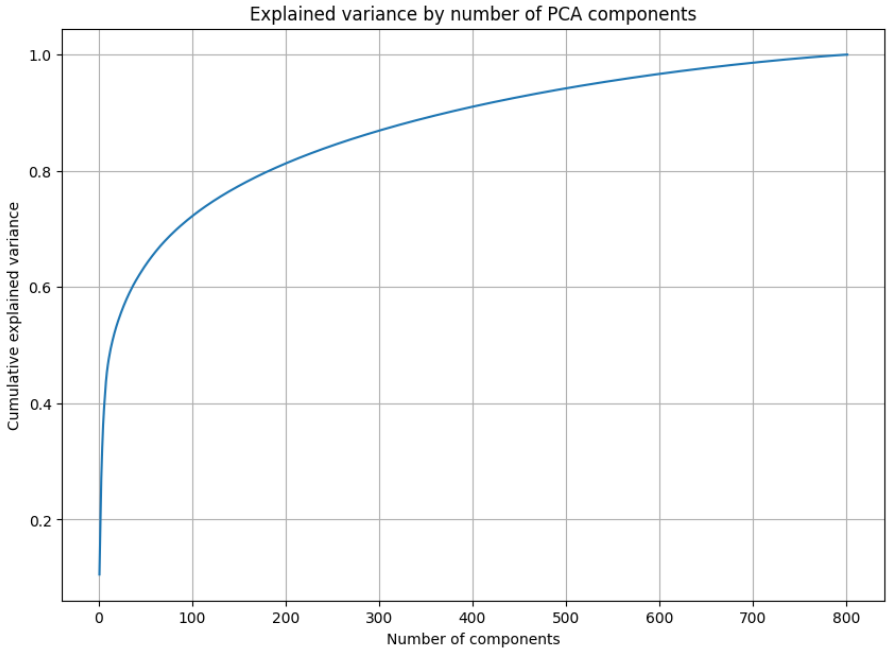
Each sample has expression values for around 20K genes. However, it may not be necessary to include all 20K gene expression values to analyze each cancer type. Therefore, we need to identify a smaller set of attributes which would then be used to fit multiclass classification models. So, the first task targets dimensionality reduction using PCA.



- PCA effectively reduces the high-dimensional data to 2-dimensions and provides insights into the distribution of different cancer types. The use of color to represent different cancer types allows for easy differentiation and understanding of the data.
- This plot shows that there are distinct clusters formed by different cancer types, indicating that the gene expression patterns vary across different cancer types.
- Although some genes overlap between the clusters, the plot does not show any clear separation between all cancer types.

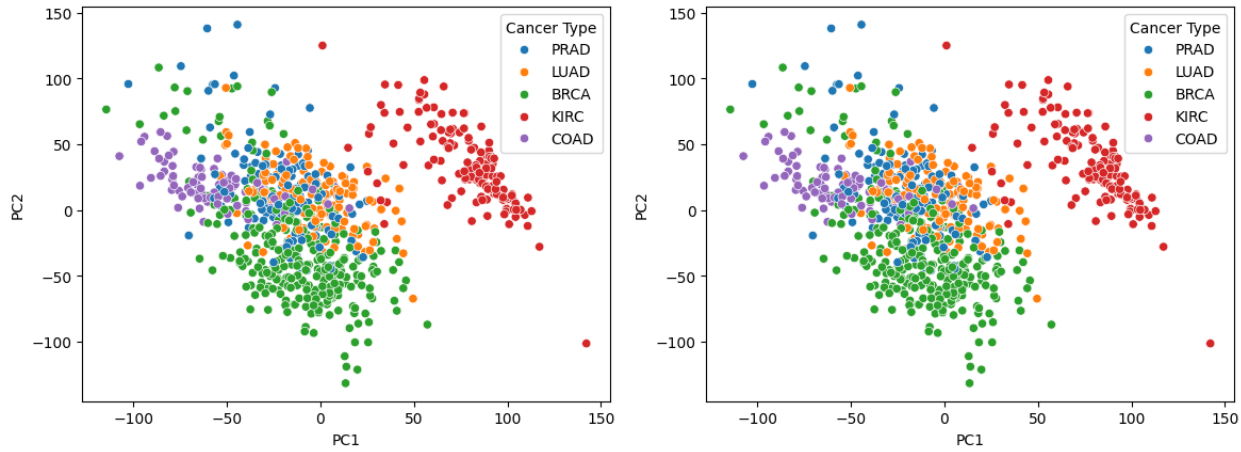
Cumulative Explained Variance Plot:

- This plot shows the cumulative sum of the explained variances for the components. The '*elbow*' or point where the curve starts to level off is often used to decide how many components one wants to keep.



- From this curve, we see that it starts to flatten around 150 - 200 PCs, suggesting a potential elbow point in this range.

Now, we try PCA with 600 and 400 components which gives 96.54% and 90.81% of variance explained respectively.



(a) 600 Components

(b) 400 Components

Figure 9: PCA with different components

8 Parametric Tests

8.1 Multiple F-tests

We want to determine whether means of gene information encoded by a particular gene (*response*) differ statistically significantly among the independent cancer groups (*covariates*). In this case, with reference to the `ICMR` dataset, the independent categorical variable is the '`Class`' column which represents the 5 different cancer types. The dependent variable is the gene expression levels of a specific gene. We analyze each gene individually in this setup, making it multiple F-tests.

We have,

- 5 groups (categories of cancer types)
- n_i observations in the i^{th} group (where $i = 1, 2, \ldots, 5$)
- N total observations ($N = n_1 + n_2 + \ldots + n_5$)
- X_{ij} is the observation in the i^{th} group and the j^{th} gene expression level ($j = 1, 2, \ldots, 20532$)
- $\overline{X_n} \overset{asym}{\sim} \mathcal{N}(\mu, \frac{\sigma^2}{n})$ i.e.,$\dfrac{\sqrt{n}(\overline{X_n} - \mu)}{\sigma} \xrightarrow{d} \mathcal{N}(0, 1)$ as $\boxed{n \to \infty}$ (by Central Limit Theorem) where $\overline{X_n}$ is a random variable with mean μ and variance $\frac{\sigma^2}{n}$ that denotes mean gene expression for a particular gene.

The null hypothesis (H_0) for each F-test is that there are no differences in mean gene expression levels among the different cancer types.

$$\text{H}_0 : \mu_1 = \mu_2 = \ldots = \mu_5$$

where: μ_i is the population mean of the i^{th} group.

The alternative hypothesis (H_1) is that at least one pair of mean gene expression levels is different.

$$\text{H}_1 : \text{At least one } \mu_i \text{ is different from the others.}$$

To test these hypotheses, we calculate the F-statistic: $\text{F} = \frac{\text{Between-group variance}}{\text{Within-group variance}}$

Rejection Criteria: If the p-value associated with the F-statistic is below a certain threshold (typically 0.05), we reject the null hypothesis.

Table 3: Top 3 high variance genes

Gene	F_statistic	p_value	variance
gene_9176	3463.550	0	44.76385
gene_9175	4194.489	0	36.36194
gene_15898	1905.191	0	34.50391

Table 4: Bottom 3 high variance genes

Gene	F_statistic	p_value	variance
gene_12668	3.163906	0.0137332	0.0014541
gene_12670	3.250183	0.0118692	0.0013394
gene_4834	2.446886	0.0450807	0.0005937

8.2 One vs. All t-test

In a one-vs-all t-test scenario, we compare the means of one cancer group (the "*one*" group) against the means of all other groups excluding that cancer group (the "*all*" group). It's commonly used in situations where one wants to compare a specific group against the rest of the data.

- Null Hypothesis (H_0): The null hypothesis assumes that there is no difference between the means of the "*one*" group and the means of the "*all*" group.

$$H_0 : \mu_{\text{one}} = \mu_{\text{all}}$$

- Alternative Hypothesis (H_1): The alternative hypothesis states that the means of the "*one*" group and the "*all*" group are different.

$$H_1 : \mu_{\text{one}} \neq \mu_{\text{all}}$$

- Test Statistic: $t = \dfrac{\bar{x_1} - \bar{x_2}}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$ where:
  - $\bar{x_1}$ and $\bar{x_2}$ are the means of the "*one*" group and the "*all*" group respectively.
  - s_1 and s_2 are the standard deviations of the "*one*" group and the "*all*" group respectively.
  - n_1 and n_2 are the sample sizes of the "*one*" group and the "*all*" group respectively.
- Rejection Criteria: If the p-value is smaller than a predetermined significance level (commonly 0.05), we reject the null hypothesis.
- Results: We perform this test for the top 3 high variance genes.

```
Performing 'one vs. all' t-tests for gene_9176 :
PRAD vs. All for gene_9176 : p-value = 0
LUAD vs. All for gene_9176 : p-value = 2.3192579e-50
BRCA vs. All for gene_9176 : p-value = 3.9070211e-28
KIRC vs. All for gene_9176 : p-value = 2.5060458e-42
COAD vs. All for gene_9176 : p-value = 0.00025240039
Performing 'one vs. all' t-tests for gene_9175 :
PRAD vs. All for gene_9175 : p-value = 2.0867204e-256
LUAD vs. All for gene_9175 : p-value = 4.2859671e-35
BRCA vs. All for gene_9175 : p-value = 8.2966265e-23
KIRC vs. All for gene_9175 : p-value = 4.8127605e-50
COAD vs. All for gene_9175 : p-value = 8.2634542e-27
Performing 'one vs. all' t-tests for gene_15898 :
PRAD vs. All for gene_15898 : p-value = 6.2853312e-37
LUAD vs. All for gene_15898 : p-value = 2.4837977e-99
BRCA vs. All for gene_15898 : p-value = 3.5424646e-33
KIRC vs. All for gene_15898 : p-value = 8.7197042e-28
COAD vs. All for gene_15898 : p-value = 7.6822615e-16
```

9 Non-Parametric Tests

9.1 Shapiro-Wilk Test

- We have performed this test to assess whether the datasets for male and female cancer patients come from a normal distribution. This is done using the `shapiro.test()` function in `R`.
- Null Hypothesis (H_0): The data are normally distributed.
- Alternative Hypothesis (H_1): The data are not normally distributed.
- Rejection Criteria: If the p-value is less than or equal to the chosen significance level (α), typically 0.05 then H_0 is rejected. This implies that there is sufficient evidence to conclude that the data do not come from a normally distributed population. Else, there is not enough evidence to conclude that the data do not come from a normally distributed population and further data analysis is required.
- This test is more powerful in detecting departures from normality over other normality tests, especially for small to moderate sample sizes.

Question: Do our samples come from a normal population?

Ans: Applying the Shapiro-Wilk's Test,

For **males**:

```
    Shapiro-Wilk normality test

data:  m
W = 0.95473, p-value = 1.992e-11
```

We reject H_0

For **females**:

```
    Shapiro-Wilk normality test

data:  f
W = 0.95473, p-value = 1.992e-11
```

We reject H_0

9.2 Kolmogorov-Smirnov Test

- We have performed the K-S test to compare to see if both the datasets come from the same distribution. This is done using the `ks.test()` function in "`R`".
- Null Hypothesis (H_0): The two datasets are drawn from the same continuous distribution.
- Alternative Hypothesis (H_1): The two datasets are not drawn from the same continuous distribution.
- Rejection Criteria: If the p-value is less than or equal to the chosen significance level (α), typically 0.05 then H_0 is rejected. This implies that there is sufficient evidence to conclude that the two datasets are not drawn from the same continuous distribution. Else, there is not enough evidence and further data analysis is required.

Questions:

1. Do both the samples come from the same distribution?

Ans: Applying K-S Test,

```
    Two-sample Kolmogorov-Smirnov test

data:  f and m
D = 0, p-value = 1
alternative hypothesis: two-sided
```

We fail to reject H_0

2. Does the ages of male and female cancer patients follow the Gamma distribution?

Ans: For **females**:

```
    One-sample Kolmogorov-Smirnov test

data:  f
D = 0.99482, p-value < 2.2e-16
alternative hypothesis: two-sided
```

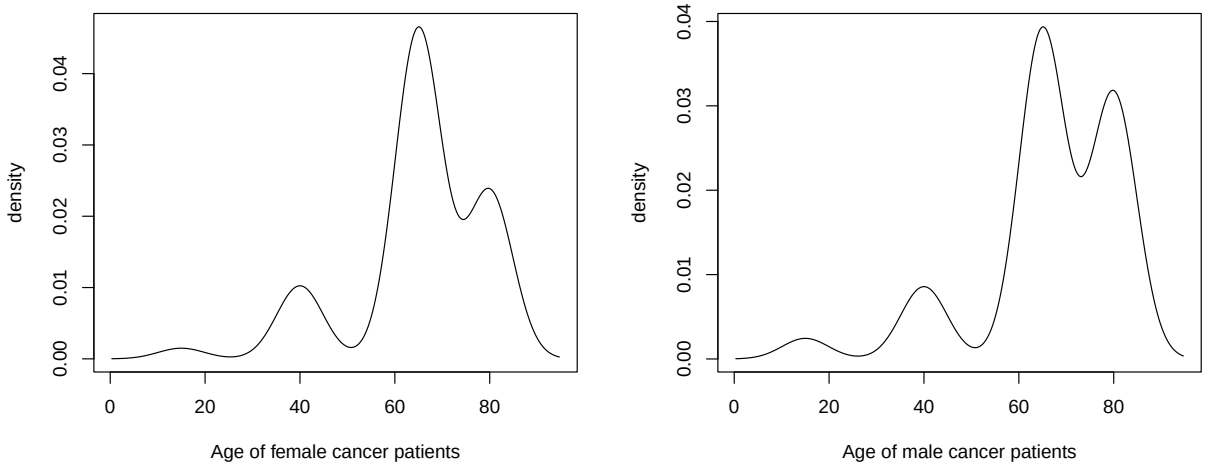
We reject H_0

For **males**:

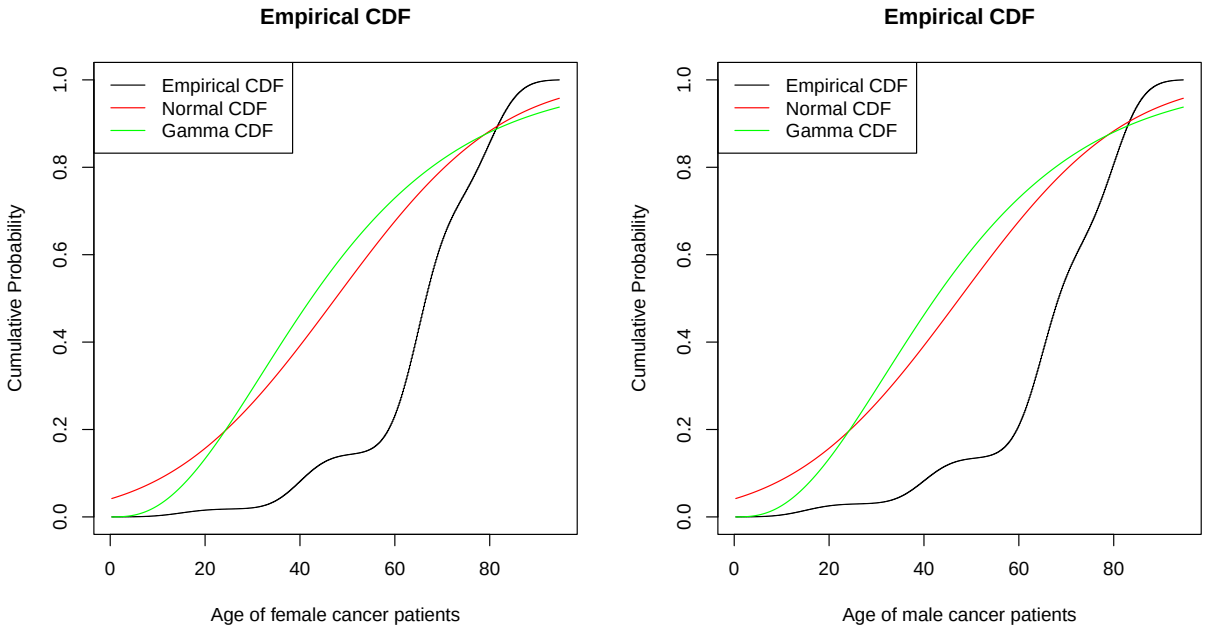
```
    One-sample Kolmogorov-Smirnov test

data:  m
D = 0.99482, p-value < 2.2e-16
alternative hypothesis: two-sided
```

We reject H_0

We have estimated the probability density function of both the datasets using kernel density estimation with a specified bandwidth 4.89 and plotted below.



We have calculated and plotted the empirical cumulative distribution function (ECDFs) for both the datasets. The theoretical cumulative distribution functions (CDFs) for normal and gamma distributions which are fitted to our datasets are also plotted alongside for visual comparision.



10 Conclusion

Acknowledgements: We express our sincere gratitude to Prof. Subhajit Dutta for his supervision and continuous support to complete this project.

11 References

- Fundamentals of Statistics (Volume One); A.M. Gun, M.K. Gupta, B. Dasgupta; The World Press Pvt. Ltd., 2019.
- The Cancer Breakthrough: A Nutritional Handbook for Doctors and Patients (Paperback) – 29 July, 2007