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

A Comprehensive Analysis on Cancer Causing and Incidence in India

Group – The Outlaws

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

PURPOSE

We want to accomplish the following goals in this project:

1. To determine the cancer causing genes in the human body.

2. To analyse if there is any dependency of those genes in affecting different types of cancer.

3. To compare the estimated cancer incidence in India according to the states on some coherent years and to infer region-wise dependencies diagrammatically.

4. To predict the possibility of the occurrence of 5 different types of cancer for an individual.

5. To find components that capture maximum variability through the Principal Component Analysis.

6. To find how the different sites of cancer are distributed among the different sexes.

DATASET

Cancer Gene Dataset

This dataset contains 802 samples for the corresponding 802 people who have been detected with different types of cancer. Each sample contains expression values of more than 20K genes. Samples have one of the types of tumours : BRCA, KIRC, COAD, LUAD and PRAD.

Estimated state wise Cancer Incidence

The dataset provides estimated incidence of cancer cases in India by States and Union Territory wise for all sites and for both sexes.

Gender wise different sites of Cancer

The dataset presents the estimated cancer incidence, number of cases, crude rate and cumulative risk by sex and anatomical sites in India for the year 2022.

Age wise Cancer Incidence

The dataset provides gender disaggregated, estimated top 5 leading sites of cancer (%) in India by age group (0-14, 15-39, 40-64 & 65+ age groups) for the year 2022.

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:

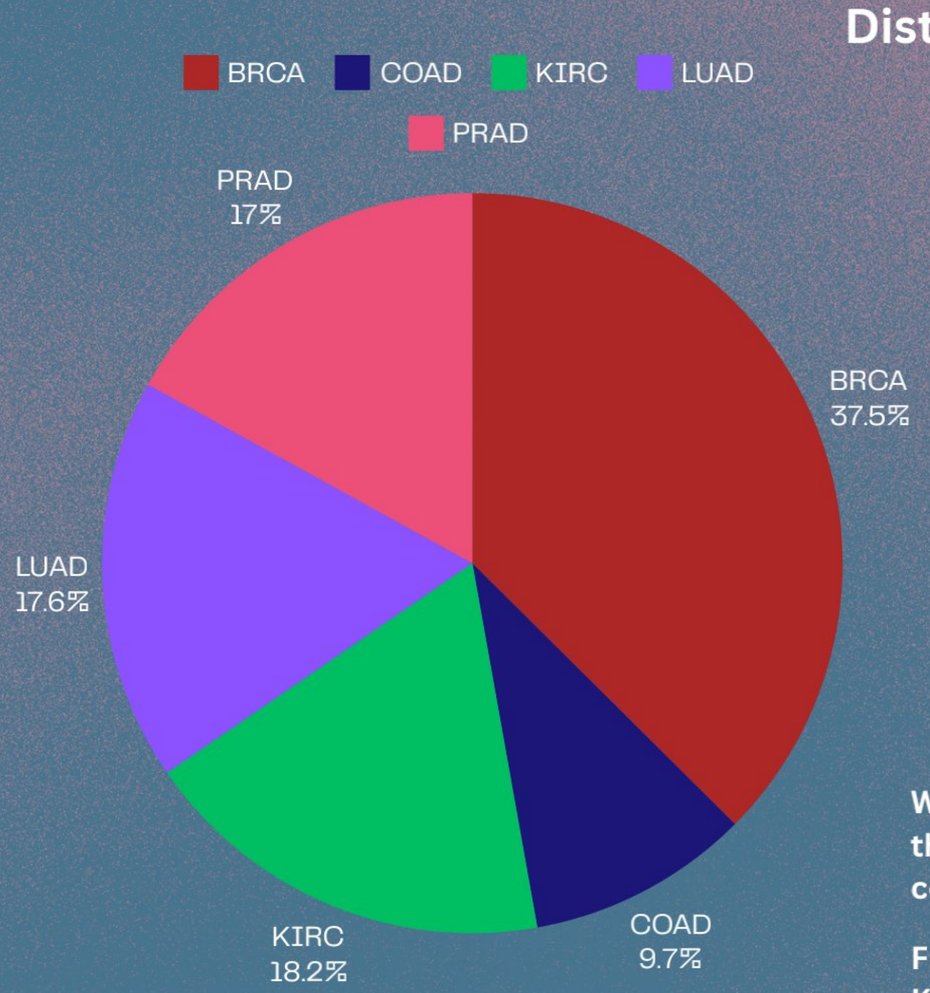


Figure 1: Proportion of Classes

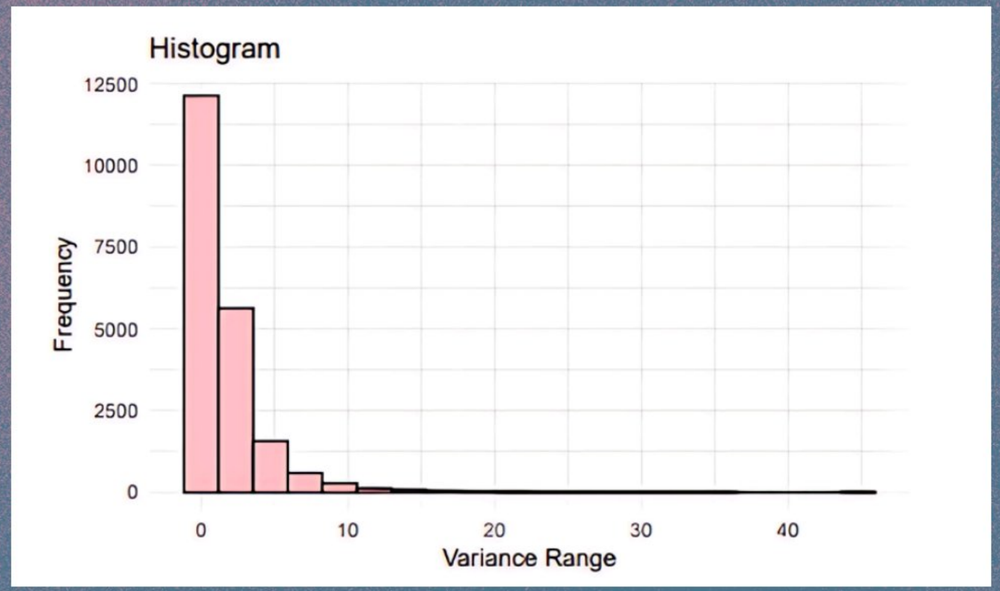


Figure 2: Histogram of Variance Range

We see that most gene columns have low variance range. Since these features capture less variability, they can be considered to be less significant. Features that capture maximum variability are less in count and cannot be disregarded as they have maximum power in discriminating between classes.

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. our exploration of different cancer types demonstrated non-uniform distribution

Data Insights

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

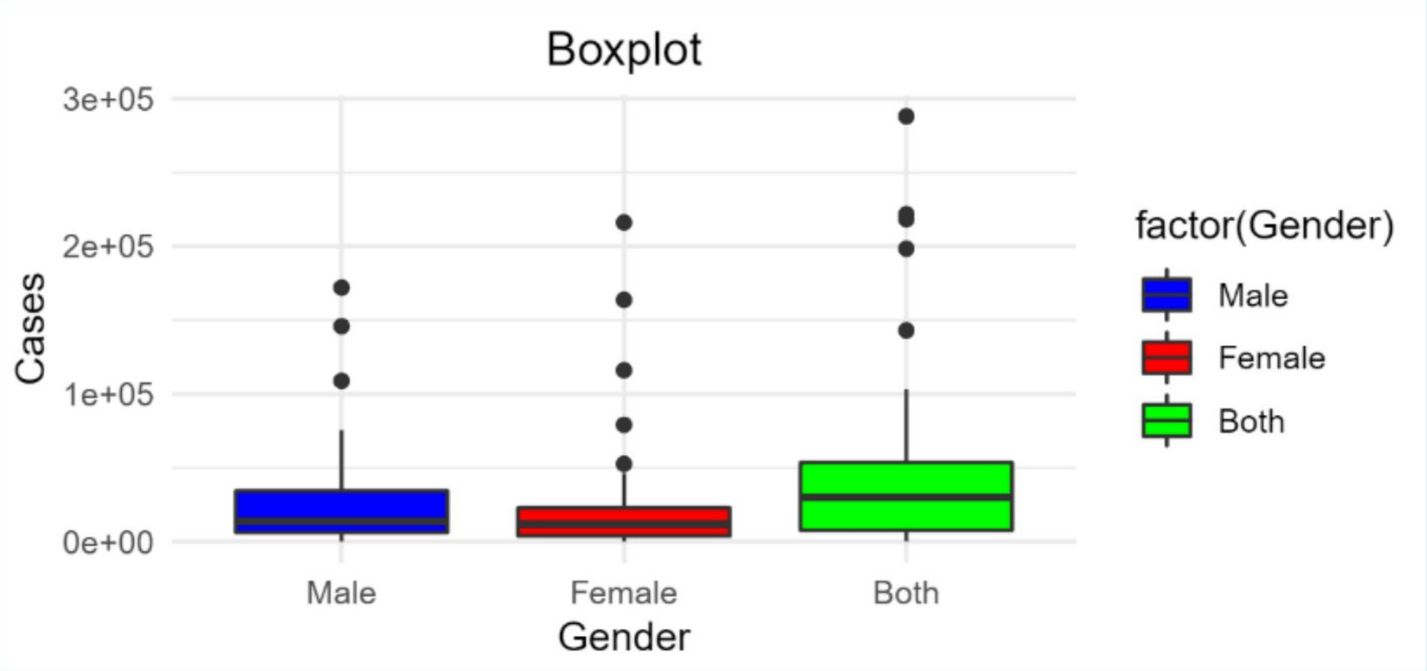
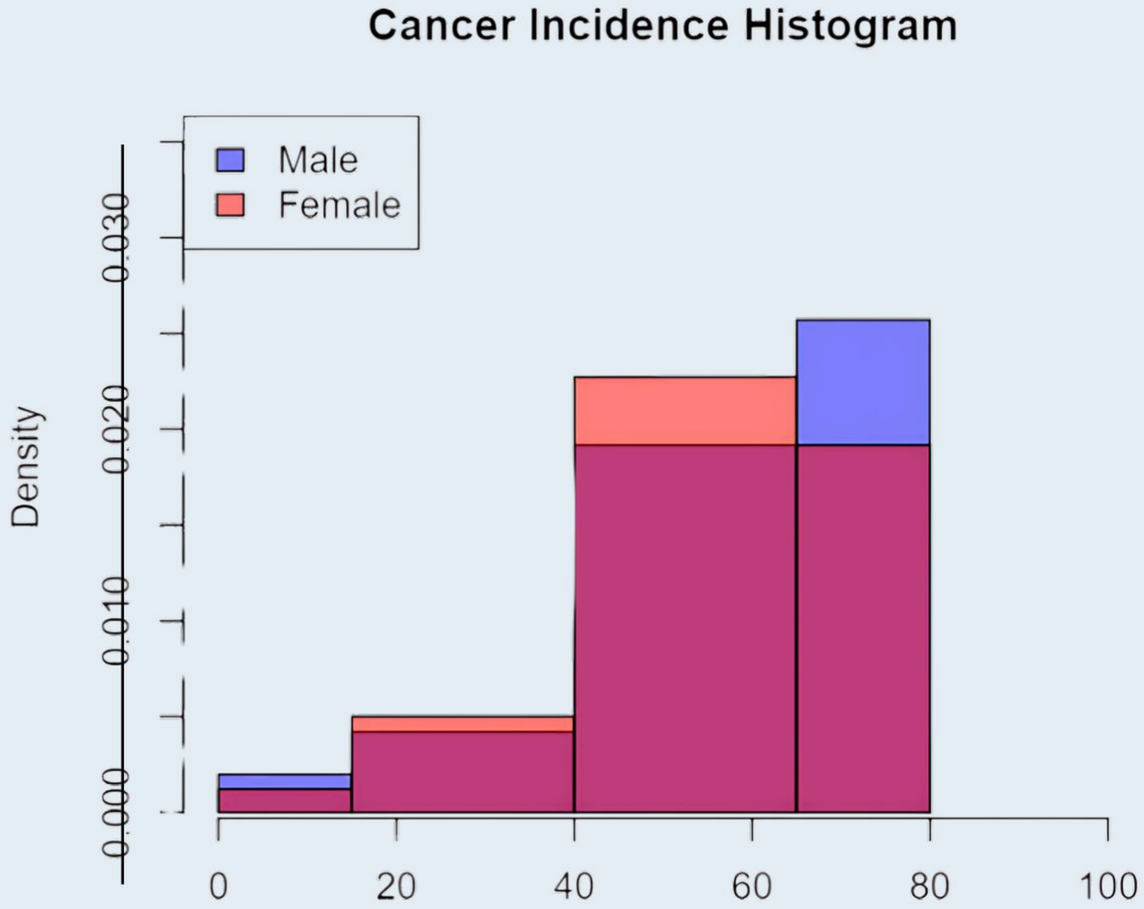


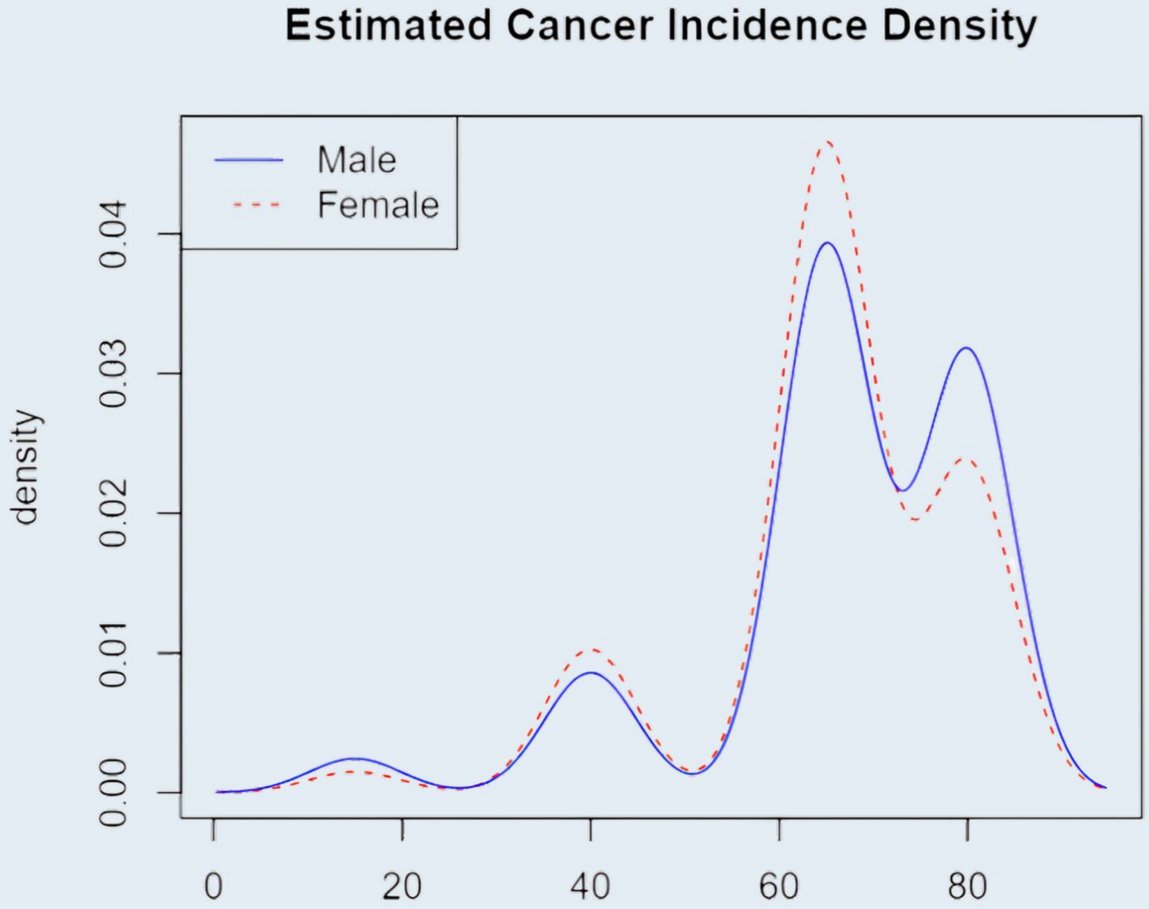
Figure 4: Cases by Gender

VISUALIZATIONS

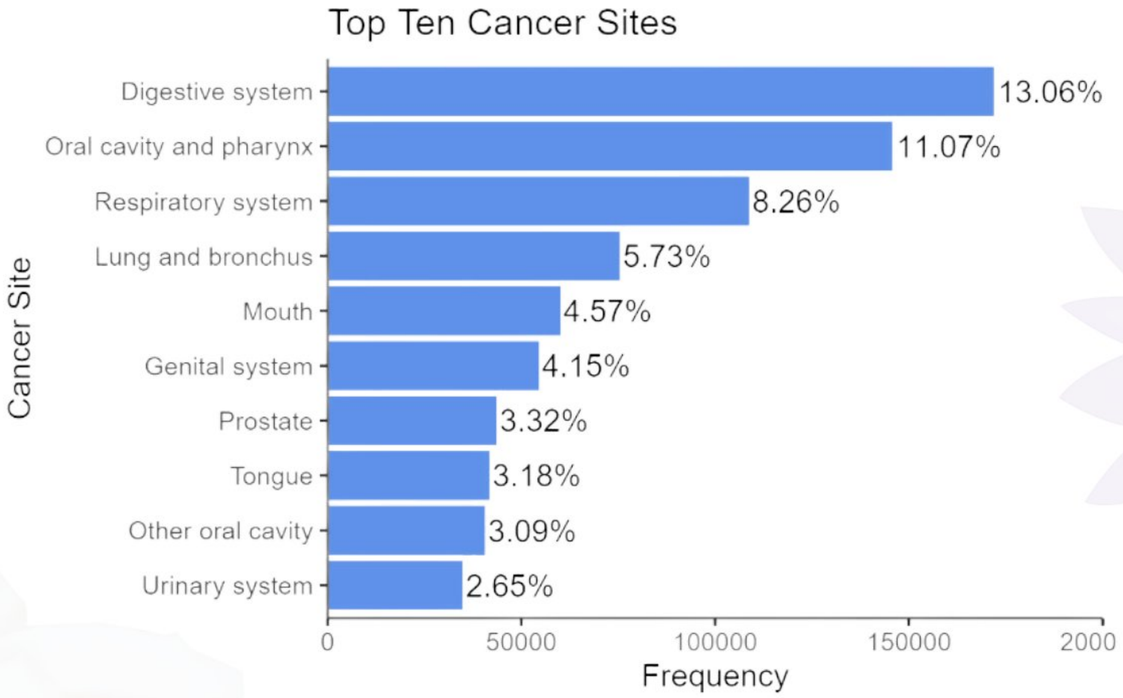
- The density curve for females exhibits a leptokurtic shape. In contrast, the density curve for males demonstrates a mesokurtic shape.

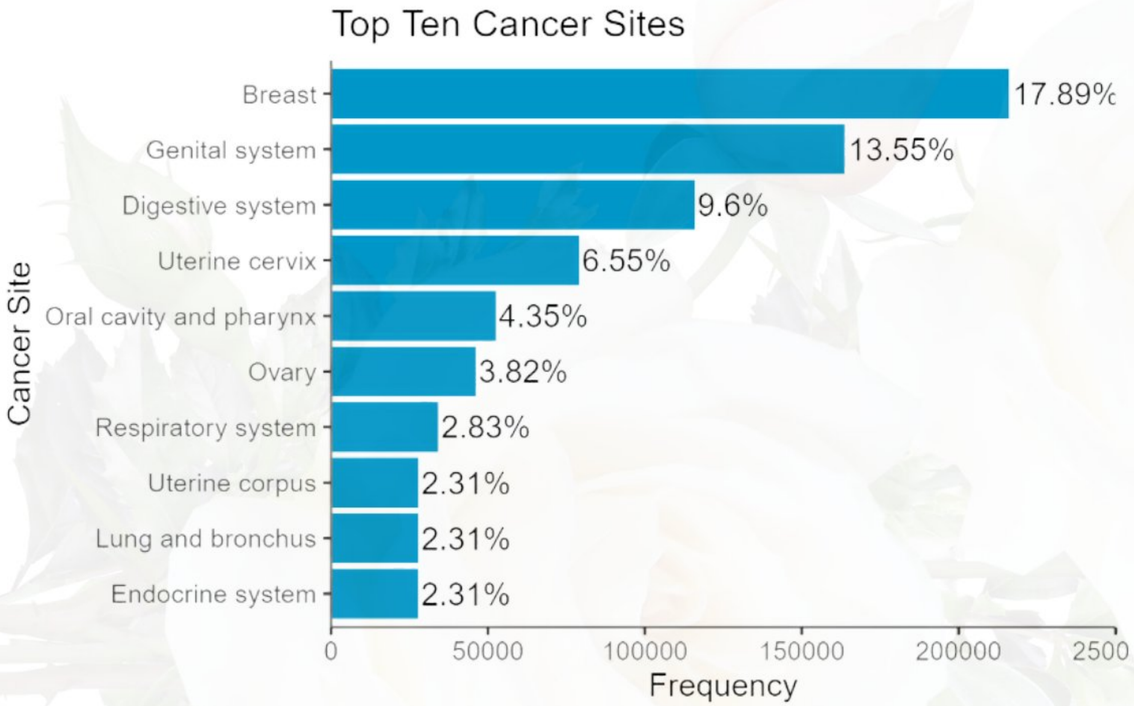
- Our examination of the age-wise distribution of cancer patients across genders revealed a common trend characterized by a negative skewness.





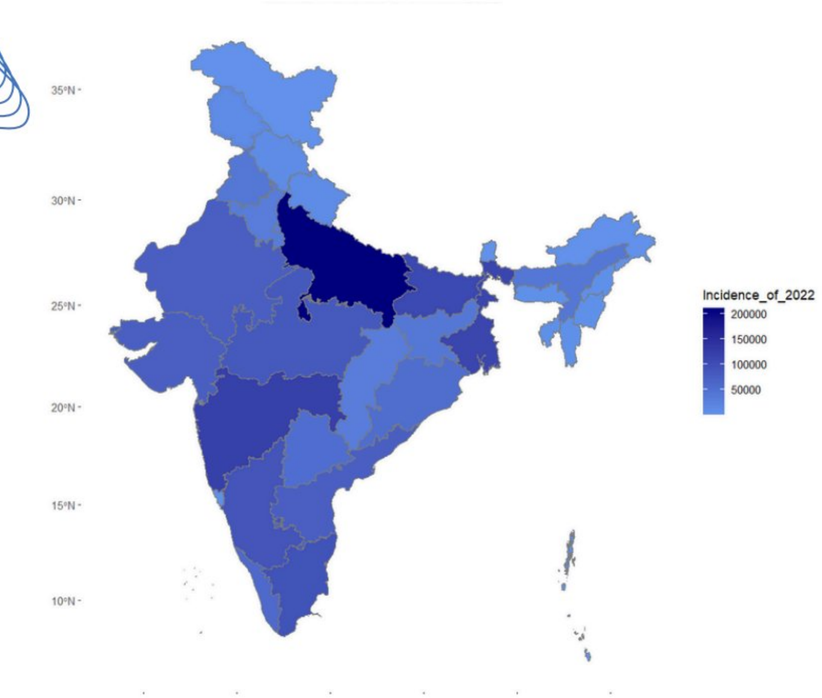
This figure presents the estimated top ten leading sites of cancer; among males these were Digestive System (13.06%), Oral Cavity and pharynx (11.07%), Respiratory System (8.26%), Lung and Bronchus (5.73%), Mouth (4.57%), Genital System (4.15%), Prostate (3.32%), Tongue (3.18%), Other Oral Cavity (3.09%) and Urinary System (2.65%).





The estimated top ten leading sites of cancer among females included Breast (17.89%), Genital System (13.55%), Digestive System (9.60%), Uterine Cervix (6.55%), Oral Cavity and Pharynx (4.35%), Ovary (3.82%), Respiratory System (2.83%), Uterine Corpus (2.31%), Lung and Bronchus (2.31%), Endocrine System (2.31%).

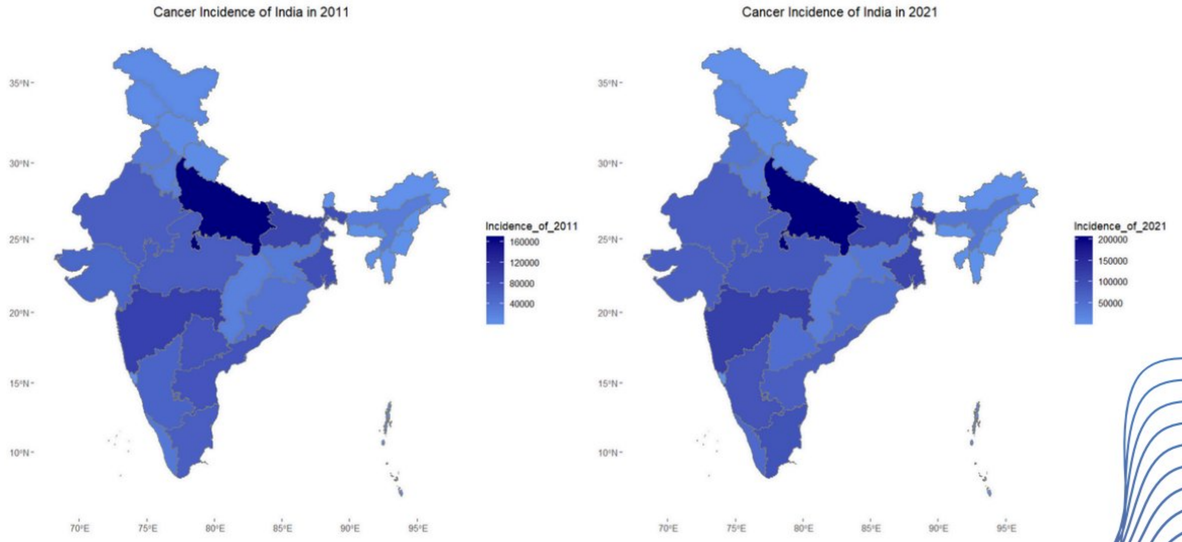
SPATIAL ANALYSIS



- 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.

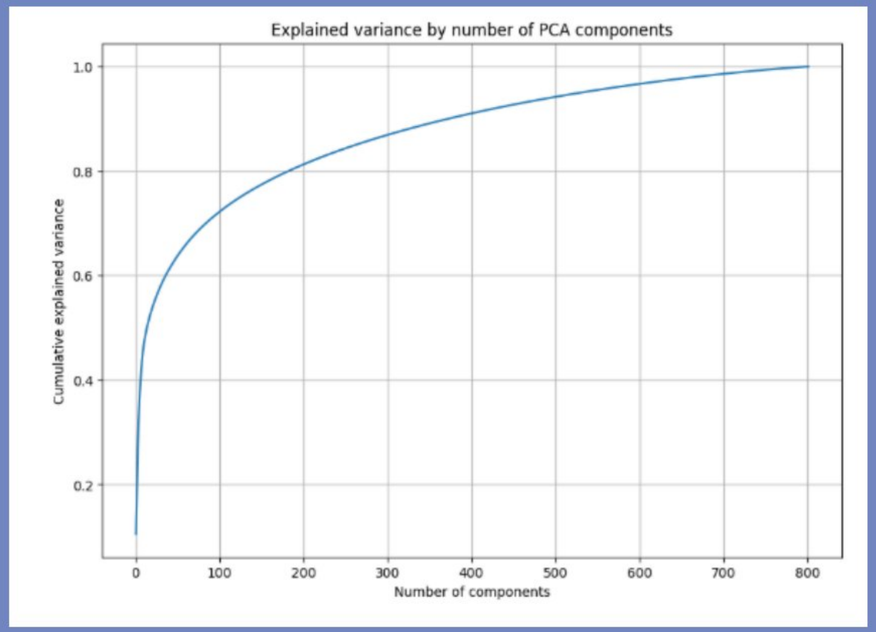
We performed a comparative 10 year gap analysis with the cancer incidences in years, 2011 and 2021 respectively.

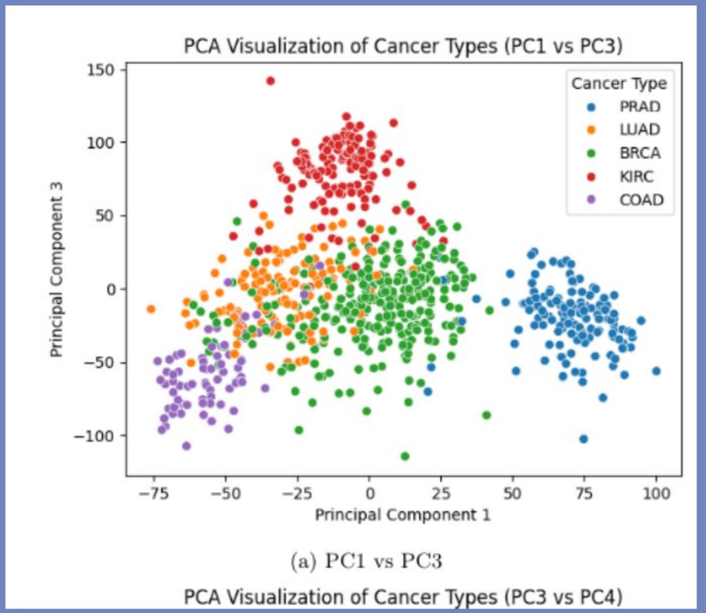
- We observed that Karnataka had improved over in cancer eradication whereas Hyderabad plunged to an increment in the number of cancer cases. The Telangana state, formed in 2 June, 2014, had less amount of cancer cases in 2021. Ladakh was formed as a Union Territory in 31 October, 2019. It showed few cancer patients in 2021, possibly due to its smaller population ratio.
- West Bengal displayed very slight changes in the cancer incidence ratio (i.e., Total no. of cancer patients in that state/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.



Principal Component Analysis

In our data each sample has expression values for around 20K genes. . Therefore, we need to identify a smaller set of attributes which would then be used to fit multiclass classification models. So, we have done dimensionality reduction using PCA.





(a) PC1 vs PC3

PCA Visualization of Cancer Types (PC3 vs PC4)

This plot typically shows the cumulative sum of explained variances explained on the Y-axis and the number of principal components on the X-axis.

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

In This plot PC1 and PC3 hold valuable information for distinguishing KIRC and PRAD from other cancer types.

We can see class PRAD is notably more separable than other classes, indicating effective feature differentiation.

Multinominal Logistic Regression

We want to fit Logistic Regression for predicting the probability of the occurrence of 5 different types of cancer for an individual.

Here we consider the different types of cancer as distinct labels, each associated with corresponding gene expression values. Our interest lies in fitting a Multinomial Logistic Regression Model. Let us examine the following table, which defines the probabilities of occurrence for each cancer type.

Table 6: Cancer Types with Corresponding Probabilities for j^{th} individual

Sl. No.	1	2	3	4	5	
Cancer Types	BRCA	COAD	KIRC	LUAD	PRAD	Total
Probability	π_{1j}	π_{2j}	π_{3j}	π_{4j}	π_{5j}	1

$$Y_j = \begin{cases} 1 & \text{if BRCA occurred} \\ 2 & \text{if COAD occurred} \\ 3 & \text{if KIRC occurred} \\ 4 & \text{if LUAD occurred} \\ 5 & \text{if PRAD occurred} \end{cases} \qquad Y_j = \begin{cases} 1 & \text{with probability } \pi_{1j} \\ 2 & \text{with probability } \pi_{2j} \\ 3 & \text{with probability } \pi_{3j} \\ 4 & \text{with probability } \pi_{4j} \\ 5 & \text{with probability } \pi_{5j} = 1 - \sum_{i=1}^{4} \pi_{ij} \end{cases}$$

Therefore, $P(Y_j = i) = \pi_{ij} = \frac{e^{X_j^T \beta_i}}{1+\sum_{i=1}^{4} e^{X_j^T \beta_i}}$; $i = 1, 2, 3, 4$

Now by Maximum Likelihood Estimation we try to maximize the likelihood function,

$$L(\beta) = L(\beta_1, \beta_2, \beta_3, \beta_4) = \prod_{j=1}^{n} \left[\left\{ \prod_{i=1}^{4} \left(\frac{e^{X_j^T \beta_i}}{1+\sum_{i=1}^{4} e^{X_j^T \beta_i}} \right)^{\mathbb{I}(Y_j=i)} \right\} \left(\frac{1}{1+\sum_{i=1}^{4} e^{X_j^T \beta_i}} \right)^{\mathbb{I}(Y_j=5)} \right]$$

in order to fit the model. We can maximize it by various numerical ways.

Here from the p-values, we can see that Gene -1 has significant effects in causing BRCA and KIRC, at level 0.05.

Now, let us fit the model and obtain the predicted probabilities of different types of cancer.

Table 7: Estimates and p-values of the regression coefficients

Coefficients	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept):1	2.345370	0.343819	6.822	9.01×10^{-12}
(Intercept):2	-0.547328	0.494891	-1.106	0.269
(Intercept):3	2.398823	0.365845	6.557	5.49×10^{-11}
(Intercept):4	0.296070	0.411232	0.720	0.472
$gene_1 : 1$	-0.492740	0.099693	-4.943	7.71×10^{-07}
$gene_1 : 2$	-0.002505	0.137841	-0.018	0.985
$gene_1 : 3$	-0.787602	0.112672	-6.990	2.75×10^{-12}
$gene_1 : 4$	-0.076462	0.115619	-0.661	0.508

Let us consider two individuals with gene expression values of Gene - 1 as 3.4678533 and 2.9411814. Then the fitted probabilites are as follows:

Table 8: Fitted Probabilities of the Specified Two Individiuals

Individual	BRCA	COAD	KIRC	LUAD	PRAD
1	0.3626322	0.11002989	0.1375935	0.1978839	0.1918605
2	0.3962039	0.09286059	0.1755878	0.1736390	0.1617087

Again, let us check it for Gene - 1 and Gene - 2 collectively
Now, let us fit the model and obtain the predicted probabilities of different types of cancer.
Let us consider two individuals with gene expression values of Gene - 1 and Gene - 2 as follows: 3.4678533, 2.9411814 and 2.9411814, 2.6632763 respectively. Then the fitted probabilites are:

Table 9: Fitted Probabilities of the Specified Two Individiual

Individual	BRCA	COAD	KIRC	LUAD	PRAD
1	0.3461035	0.1183316	0.1068651	0.2312489	0.19745091
2	0.4627836	0.1008106	0.2074741	0.1551230	0.07380868

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 $i = 1, 2, \dots, 5$)
- N total observations ($N = n_1 + n_2 + \dots + n_5$)
- X_{ij} is the observation in the i^{th} group and the j^{th} gene expression level ($j = 1, 2, \dots, 20532$)
- $\overline{X_n} \overset{asym}{\sim} \mathcal{N}(\mu, \frac{\sigma^2}{n})$ i.e., $\frac{\sqrt{n}(\overline{X_n} - \mu)}{\sigma} \xrightarrow{d} \mathcal{N}(0, 1)$ 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.

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_5$$

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.

$$H_1 : \text{At least one } \mu_i \text{ is different from the others.}$$

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

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

One vs. All t-test

- 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.

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

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

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

- Test Statistic: $t_{cal} \overset{H_0}{=} \dfrac{\bar{x_1} - \bar{x_2}}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$ 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. Else, we fail to reject H_0 and further data analysis is required.

Table 5: Assumptions

Independence	Gene expressions within the "*one*" group and "*all*" group are independent of each other.
Normality	Each group's data follow approximately the normal distribution (by Central Limit Theorem).
Random Sampling	The collective gene expressions data are obtained through a random sampling process from the population of cancer patients. This ensures that the sample is representative of the population.
Unequal Variances	The compared variances of the two groups are unequal.

- Results: We performed 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
```

Kolmogorov-Smirnov Test

10.2 Kolmogorov-Smirnov Test

10.2.1 One Sample Test

Previously, we performed Shapiro-Wilk test to check the normality of the age of cancer patients. WS test works better for checking whether the data is normal or not. But, to check about the other distribution we are using very popular Kolmogorov-Smirnov test.

We have,

- $X_{i_1}, X_{i_2}, ..., X_{i_{n_i}} \overset{iid}{\sim} F_i(x)$ where X_{i_j} is the j^{th} person of the i^{th} gender. Here $i = \{f, m\}$ and $j = 1, 2, ...n_i$ where $n_m = 712176$, $n_f = 749251$.
- The distribution functions F_i are assumed to be absolutely continuous $\forall i$.
- The test is a level $\alpha = 0.05$ both sided test.

The i^{th} null hypothesis H_{i0} is that the i^{th} age sample comes from Γ_i which is the gamma distribution function. The corresponding alternative hypotheses are opposite.

$$H_{i0} : F_i = \Gamma_i \text{ against } H_{ia} : F_i \neq \Gamma_i \ \forall i$$

The Kolmogorov-Smirnov test statistics are,

$$D_{n_m} = \max_x |S_{n_m}(x) - \Gamma_m(x)| \quad \text{and} \quad D_{n_f} = \max_x |S_{n_f}(x) - \Gamma_f(x)|$$

- $S_{n_m}(x)$ and $S_{n_f}(f)$ are respectively the empirical distribution function of male and female age sample.
- Rejection Criteria: If the p-value is less than $\alpha = 0.5$, we reject H_0.

```
    One-sample Kolmogorov-Smirnov test

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

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

Kolmogorov-Smirnov Test

10.2.2 Two Sample Test

Now we will test whether the age of male cancer patients and female cancer patients come from the same population or not by two sample KS Test.

We have,

- $X_{i_1}, X_{i_2}, ..., X_{i_{n_i}} \overset{iid}{\sim} F_i(x)$ where X_{i_j} is the j^{th} person of the i^{th} gender. Here $i = \{f, m\}$ and $j = 1, 2, ...n_i$ where $n_m = 712176$, $n_f = 749251$.
- The distribution functions F_m and F_f are assumed to be absolutely continuous.
- The test is a level $\alpha = 0.05$ both sided test.

The null hypothesis H_0 is that the two distribution functions F_m and F_f are same. While, the alternative hypothesis is they are not equal.

$$H_0 : F_f = F_m \quad \text{against} \quad H_a : F_f \neq F_m$$

The Kolmogorov-Smirnov test statistic can be defined as,

$$D_{m,f} = \max_x |S_{n_f}(x) - S_{n_m}(x)|$$

- $S_{n_m}(x)$ and $S_{n_f}(f)$ are respectively the empirical distribution function of male and female age sample.
- Rejection Criteria: If the p-value is less than $\alpha = 0.5$, we reject H_0.

```
    Two-sample Kolmogorov-Smirnov test

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

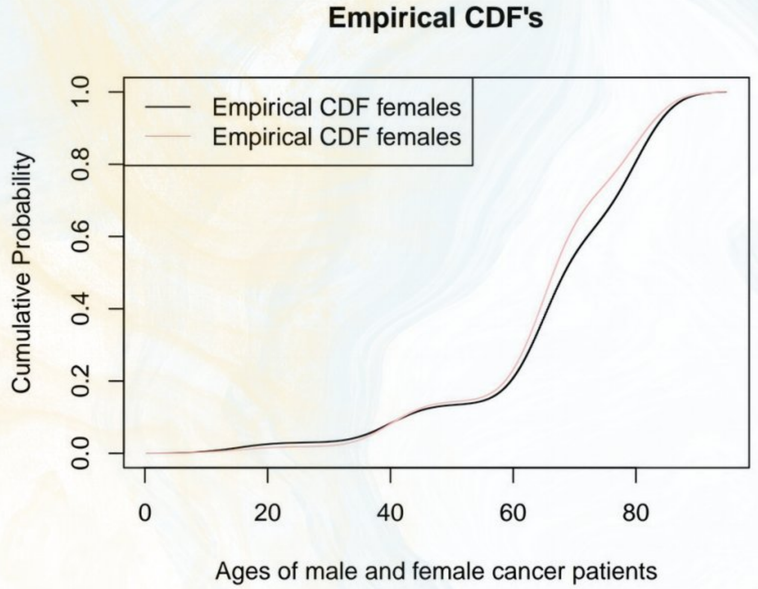


Figure 12: Plot of Empirical CDF's

CONCLUSION

Analysis of cancer incidence in India revealed significant findings:

1. Uttar Pradesh, Maharashtra, Bihar, West Bengal, Hyderabad, and Tamil Nadu were identified as high-risk regions & hilly areas & Andaman Nicobar Islands exhibited lower cancer rates compared to the plains.
2. the age-wise distribution of cancer patients, it indicated negative skewness, suggesting a common trend in cancer occurrence across genders and age groups
3. Through exploratory data analysis (EDA), we gained insights into the structure and content of the datasets. Breast Cancer (BRCA) contained the most data points, almost 38% of the entire dataset.
4. From the PCA, we found out that roughly 500 PCs were responsible for explaining about 95% of the explained variability.
5. In the multinomial logistic regression model, we inputted gene expression values of various genes –particularly those capturing maximum variability identified through PCA.
6. We used parametric tests, including multiple F-tests, to compare gene expression means across cancer types. Rejection of the null hypothesis for some genes suggests significant differences, warranting further analysis.
7. The one-vs-all t-test allowed us to compare specific cancer groups against the rest, contributing to a better understanding of inter-grouped differences

Thank you!