Data analysis on cancer prevalence using cloudera

Vaibhav Tripathi, Shubham Satvik Singh , Sahil Sankritya, Shubham Srivastav

Under the guidance of **Mr. S.N. RAJAN** Sir

Department of Information Technology ,

IMS ENGINEERING COLLEGE, GHAZIABAD, UTTAR PRADESH,INDIA.

Abstract: **Cancer is characterized by proliferation of cells that have managed to evade central endogenous control mechanisms. Cancers are grouped according to their organ or tissue of origin, but increasingly also based on molecular characteristics of the respective cancer cells. Due to the rapid technological advances of the last years, it is now possible to analyse the molecular makeup of different cancer types in detail within short time periods. The accumulating knowledge about development and progression of cancer can be used to develop more precise diagnostics and more effective and/or less toxic cancer therapies. In the long run, the goal is to offer to every cancer patient a therapeutic regimen that is tailored to his individual disease and situation in an optimal way. The aim of the project work includes the state wise, age wise and gender wise analysis of cancer data and to find the maximum individuals affected by cancer in each case (state, population, mean cases).**

INTRODUCTION

Cancer is a group of diseases involving abnormal cell growth with the potential to invade or spread to other parts of the body. It is the uncontrolled growth of abnormal cells in the body. In cancerous tumors or malignant tumors, the cells have lost the ability to stop growing. In other words, they have gone rogue and will not stop dividing. To better explain, nearly every cell in the body is able to grow and divide to make new cells, to a certain extent. This is important for all living organisms. When cells go rogue, however, there is a problem with the DNA. When mutations, which are changes in the DNA sequence, occur, they cause the cells to forget how to stop dividing. After some time, the mass of cells becomes a tumor. The tumor can either be malignant or it can be benign, which means it is not cancerous. Finally, the cancer cases are predicted by fitting a suitable trend equation and the results are presented in a graphical form and interpreted by means of Statistical analysis.

TECHNOLOGIES :

1.Cloudera for data modelling.

2.Convert unstructured data into structured data.

3.Python used for building models and visualization.

4. Pandas, seaborn modules used for plotting graphs.

5.Django framework for front end.

METHODOLOGY:

Data Preprocessing: The cancer cases data set used has over thousands rows and provides details regarding the cases that have occurred in the different states .The data set is first converted to a csv file. Using the Hive, as Hive fetch out the data from unstructured format to structured format and PIG LATIN is also used to fetch out required information from data sets.

Correlation matrix: A correlation matrix is a symmetric matrix showing the relation between various attributes. One can know which pairs have the highest correlation. There are various types of correlation matrices but in our project we are using graphical representation to analyse the dependencies among the attributes. Graphs depict the relationship using different colour shades. Higher the value, more dependency between the two attributes.

Correlation: The correlation coefficient is a statistical measure that calculates the strength of the relationship between the relative movements of two variables. The values range between -1.0 and 1.0. A calculated number greater than 1.0 or less than -1.0 means that there was an error in the correlation measurement. A correlation of -1.0 shows a perfect negative correlation, while a correlation of 1.0 shows a perfect positive correlation. A correlation of 0.0 shows no relationship between the movement of the two variables.

We found correlation among cancer cases of states from 2011-2017 are as follows:

Formula:

$$r = \frac{\sum (x - m_x)(y - m_y)}{\sqrt{\sum (x - mx)^2 \sum (y - my)^2}}$$

Correlation of Uttar Pradesh cancer cases:0.987

Correlation of Madhya Pradesh cancer cases:0.999

These values are showing perfect positive correlation.

CLUSTER ANALYSIS:

Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their understanding of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances between cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings (including parameters such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. It is often necessary to modify data pre-processing and model parameters until the result achieves the desired properties.

Analysis of Variance:

Analysis of variance (ANOVA), is the hypothesis testing used for more than two samples. The data has only one Independent variable so One way Anova (or) Complete randomised design is applicable. It test the hypothesis that there is a significant difference between the population (more than two) means.

The hypothesis is given by the following :

Ho:m1=m2=…=mn (1)

H1: At least two means are different (2)

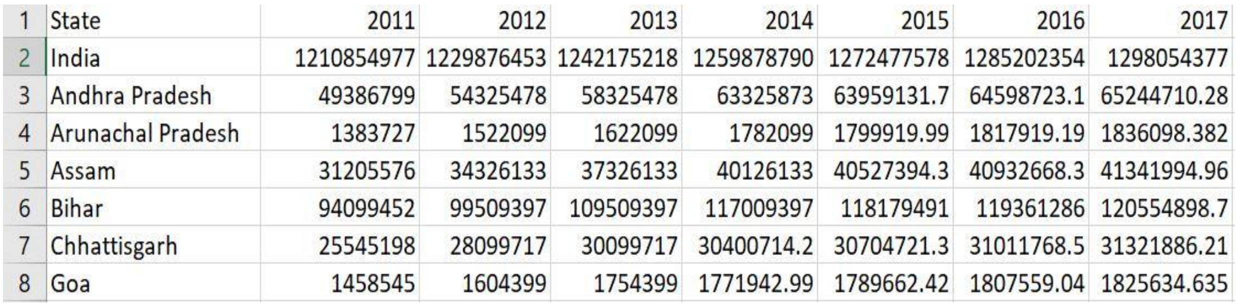
where, Ho is null hypothesis and it tells that there is no significant difference between the population means (means are same); H1 is alternative hypothesis and it tells that is a significant difference between the population. If the test statistics calculated is greater than the critical value then Ho is rejected and H1 is accepted else Ho is accepted.

Regression Analysis:

Regression analysis is used to predict the average value of 'y' for the given value of x, where x is independent variable and y is dependent variable. The fitted model assumes a linear relationship between

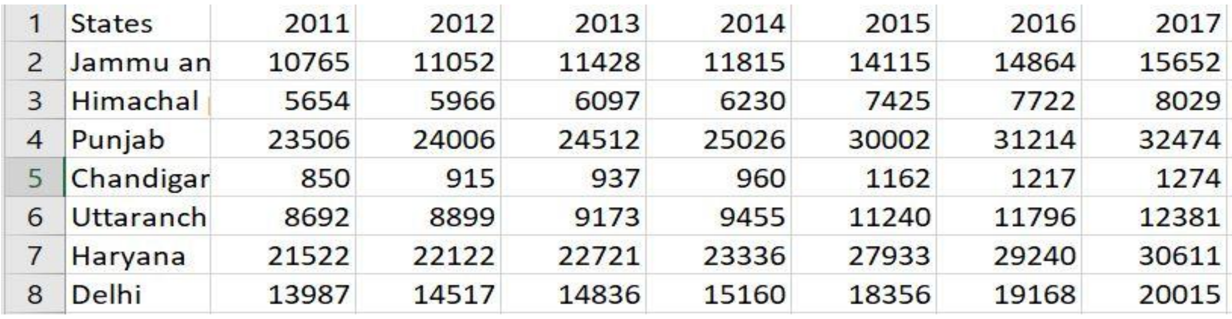
Regression equation is given as: y = (a + b x)

Sample Data of State wise population



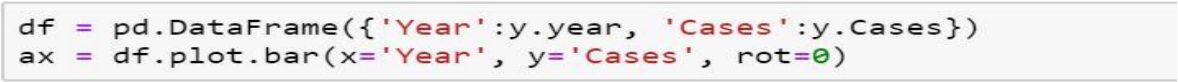
1	State	2011	2012	2013	2014	2015	2016	2017
2	India	1210854977	1229876453	1242175218	1259878790	1272477578	1285202354	1298054377
3	Andhra Pradesh	49386799	54325478	58325478	63325873	63959131.7	64598723.1	65244710.28
4	Arunachal Pradesh	1383727	1522099	1622099	1782099	1799919.99	1817919.19	1836098.382
5	Assam	31205576	34326133	37326133	40126133	40527394.3	40932668.3	41341994.96
6	Bihar	94099452	99509397	109509397	117009397	118179491	119361286	120554898.7
7	Chhattisgarh	25545198	28099717	30099717	30400714.2	30704721.3	31011768.5	31321886.21
8	Goa	1458545	1604399	1754399	1771942.99	1789662.42	1807559.04	1825634.635

Sample Data of cancer cases in INDIA

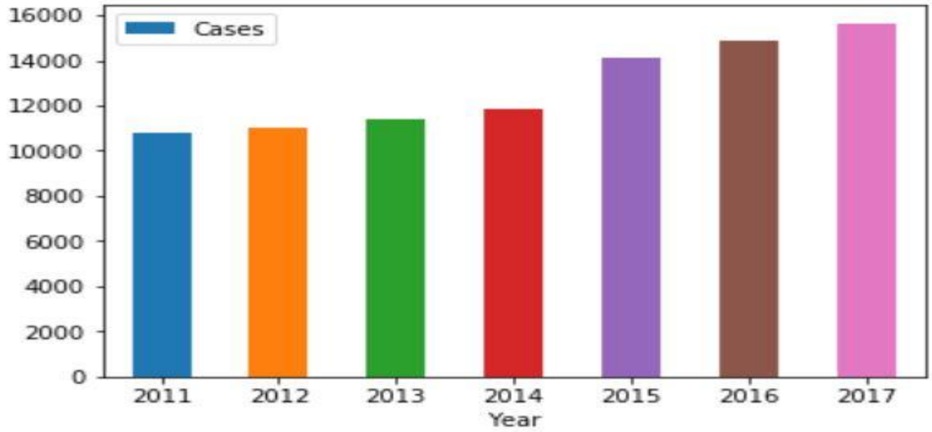


1	States	2011	2012	2013	2014	2015	2016	2017
2	Jammu an	10765	11052	11428	11815	14115	14864	15652
3	Himachal	5654	5966	6097	6230	7425	7722	8029
4	Punjab	23506	24006	24512	25026	30002	31214	32474
5	Chandigar	850	915	937	960	1162	1217	1274
6	Uttaranch	8692	8899	9173	9455	11240	11796	12381
7	Haryana	21522	22122	22721	23336	27933	29240	30611
8	Delhi	13987	14517	14836	15160	18356	19168	20015

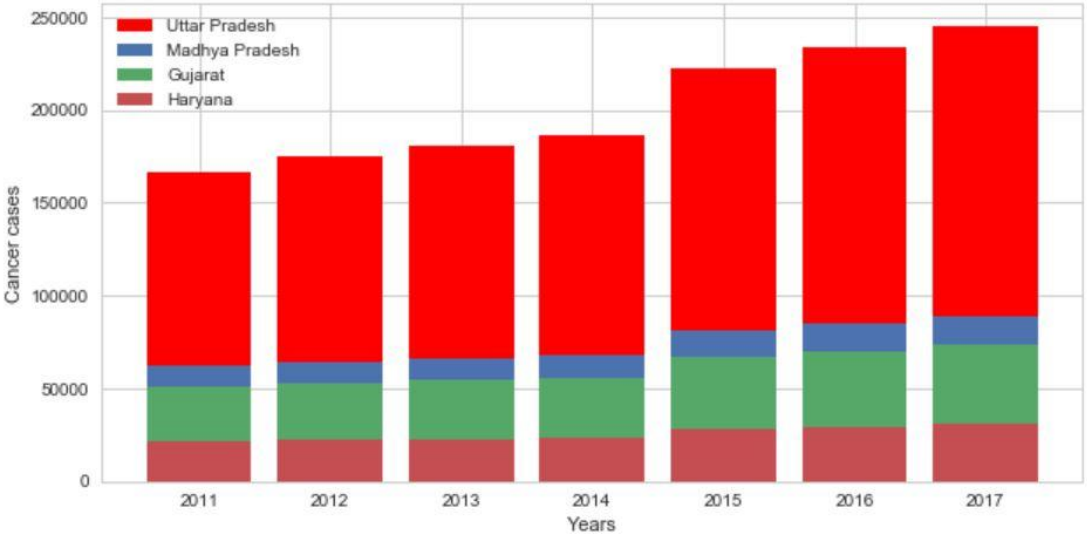
Visualization of cancer cases in Jammu & Kashmir



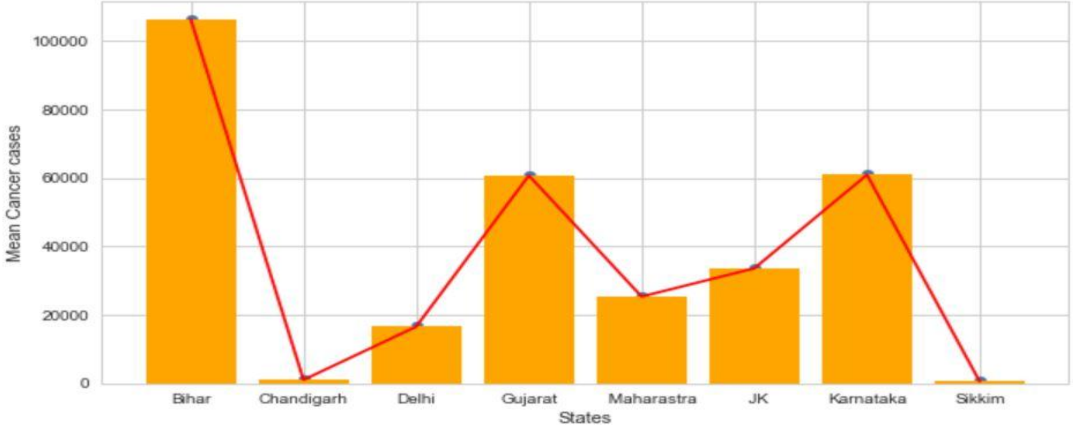
```
df = pd.DataFrame({'Year':y.year, 'Cases':y.Cases})
ax = df.plot.bar(x='Year', y='Cases', rot=0)
```



Cancer stats among states



Mean cases of some states from 2011-2017



CONCLUSION

Thus, by analysing the Cancer data from the year 2011 to 2017, by the above results and discussions we have reached a conclusion that the state which is maximum affected by Cancer is Uttar Pradesh. UP has to take some serious effects in their lifestyle to overcome this problem. Moreover, females are more affected by cancer(breast) when compared to males(prostate). The age group which has maximum infection is above sixty and above fifty age groups. This is because these people have poor immune system as they get older so, the tumor cells gets activated and spreads all over in the body. Finally, the prediction of cancer cases follows a linear pattern as increasing year by year. This research work can be further extended by analysing the cancer dataset by other statistical and machine learning techniques to discover the hidden and innovative results. Unlike more traditional forms of cancer treatment that directly target cancer cells—often with severe side-effects—**Allison and Honjo(Nobel Prize winners)** figured out how to help the patient's own immune system tackle the cancer more quickly.

REFERENCES

[1] www.cancerresearchuk.org/about-cancer/what-is-cancer

[2] https://en.wikipedia.org/wiki/Cancer

[3] https://www.cancer.org/

[4] https://www.cancer.net/cancer-types

[5] https://en.wikipedia.org/wiki/Analysis_of_variance

[6] https://en.wikipedia.org/wiki/K-means_clustering

[7] https://www.youtube.com

[8] https://www.data.gov.in