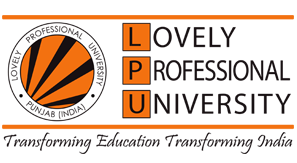
**INTRODUCTION TO DATA MANAGEMENT:**

**PROJECT REPORT**

(Project Semester August-December 2019)

******

***TOBACCO AND CANCER DATA ANALYSIS***

Submitted by

ASHISH KUMAR

Registration No: 11711039

Programme: B.Tech (CSE)

Section: KM067

Course Code: INT217

Under the Guidance of

**SANDEEP KAUR**

**Discipline of CSE/IT**

**Lovely School of Computer Science & Engineering**

**Lovely Professional University, Phagwara**

**CERTIFICATE**

This is to certify that Ashish Kumar bearing Registration no 11711039 has completed INT217 project titled, **“TOBACCO AND CANCER DATA ANALYSIS”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

**Miss. Sandeep Kaur**

**Assistant Professor**

**School of Computer Science & Engineering**

Lovely Professional University

Phagwara, Punjab.

Date:

**DECLARATION**

I, ASHISH KUMAR student of B.Tech. under CSE/~~IT~~ Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: Signature

Ashish Kumar

Registration No. 11711039

**ACKNOWLEDGEMENT**

I would like to express my special thanks of gratitude to my teacher Mrs. Sandeep Kaur who gave me the golden opportunity to do this wonderful project on the topic “TOBACCO AND CANCER DATA ANALYSIS” which also helped me in doing a lot of Research and I came to know about so many new things I am really thankful to them. Secondly is would also like to thank my parents and friends who helped me a lot in finalizing this project within the limited time frame.

**Table of Contents**

**1. Introduction**

**2. Scope of the Analysis**

**3. Source of dataset**

**4. ETL process**

**5. Analysis on dataset (for each analysis)**

1. **Introduction**
2. **General Description**
3. **Specific Requirements**
4. **Analysis results**
5. **Visualization**

**6. List of Analysis with results**

**7. References**

**8. Bibliography**

**INTRODUCTION**

**DATA SCIENCE**

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms, both structured and unstructured, similar to data mining.

Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data. It employs techniques and theories drawn from many fields within the context of mathematics, statistics, information science, and computer science. Data Science is a blend of various tools, algorithms, and machine learning principles with the goal to discover hidden patterns from the raw data

Data Science is a more forward-looking approach, an exploratory way with the focus on analyzing the past or current data and predicting the future outcomes with the aim of making informed decisions. It answers the open-ended questions as to “what” and “how” events occur.

Turing award winner Jim Gray imagined data science as a "fourth paradigm" of science (empirical, theoretical, computational and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the data deluge.

The term "data science" has appeared in various contexts over the past thirty years but did not become an established term until recently. In an early usage, it was used as a substitute for computer science by Peter Naur in 1960. Naur later introduced the term "datalogy".In 1974, Naur published Concise Survey of Computer Methods, which freely used the term data science in its survey of the contemporary data processing methods that are used in a wide range of applications.

In 1996, members of the International Federation of Classification Societies (IFCS) met in Kobe for their biennial conference. Here, for the first time, the term data science is included in the title of the conference ("Data Science, classification, and related methods"), after the term was introduced in a roundtable discussion by Chikio Hayashi.

**EXCEL**

Microsoft Excel is a spreadsheet developed by Microsoft for Windows, macOS, Android and iOS. It features calculation, graphing tools, pivot tables, and a macro programming language called Visual Basic for Applications. It has been a very widely applied spreadsheet for these platforms, especially since version 5 in 1993, and it has replaced Lotus 1-2-3 as the industry standard for spreadsheets. Excel forms part of Microsoft Office.

Microsoft Excel has the basic features of all spreadsheets, using a grid of cells arranged in numbered rows and letter-named columns to organize data manipulations like arithmetic operations. It has a battery of supplied functions to answer statistical, engineering and financial needs. In addition, it can display data as line graphs, histograms and charts, and with a very limited three-dimensional graphical display. It allows sectioning of data to view its dependencies on various factors for different perspectives (using pivot tables and the scenario manager). It has a programming aspect, Visual Basic for Applications, allowing the user to employ a wide variety of numerical methods, for example, for solving differential equations of mathematical physics, and then reporting the results back to the spreadsheet.

In a more elaborate realization, an Excel application can automatically poll external databases and measuring instruments using an update schedule, analyze the results, make a Word report or PowerPoint slide show, and e-mail these presentations on a regular basis to a list of participants. Excel was not designed to be used as a database.

Microsoft allows for a number of optional command-line switches to control the manner in which Excel starts.

**PROJECT**

Tobacco and Cancer Data Analysis

Tobacco use is a leading cause of cancer and of death from cancer. People who use tobacco products or who are regularly around [environmental tobacco smoke](https://www.cancer.gov/Common/PopUps/popDefinition.aspx?id=CDR0000046431&version=Patient&language=English) (also called secondhand smoke) have an increased risk of cancer because tobacco products and secondhand smoke have many chemicals that damage DNA.

Tobacco use causes many types of cancer, including cancer of the lung, larynx (voice box), mouth, esophagus, throat, bladder, kidney, liver, stomach, pancreas, colon and rectum, and cervix, as well as acute myeloid leukemia. People who use [smokeless tobacco](https://www.cancer.gov/Common/PopUps/popDefinition.aspx?id=CDR0000748220&version=Patient&language=English) (snuff or chewing tobacco) have increased risks of cancers of the mouth, esophagus, and pancreas.

There is no safe level of tobacco use. People who use any type of tobacco product are strongly urged to quit.  People who quit smoking, regardless of their age, have substantial gains in life expectancy compared with those who continue to smoke. Also, quitting smoking at the time of a cancer diagnosis reduces the risk of death.

Tobacco production has continued to shift from high- to low- and medium- Human Development Index (HDI) countries over the past 50 years. Many consider tobacco a cash crop, but studies conducted in multiple countries have found that tobacco farmers are often stuck in a cycle of debt that the tobacco industry perpetuates in its relationships with these farmers. In addition, as many as 16 countries use child labour in the production of tobacco. Tobacco farming is also bad for the environment, as tobacco depletes the soil of nutrients more than other crops and often requires the use of pesticides and chemical fertilizers.

**Objective/Scope of the Analysis**

The objective of the analysis is to implement the knowledge of excel learned throughout the semester in a practical manner to clearly test how good command we actually have over data management. Firstly, we need to clean data with the help of excel. Then our objective is to analyze the data for gaining insights and facts. We will also try to find top states on the basis of consumption of tobacco, mortality rate and cheap markets for tobacco.

**Sources**

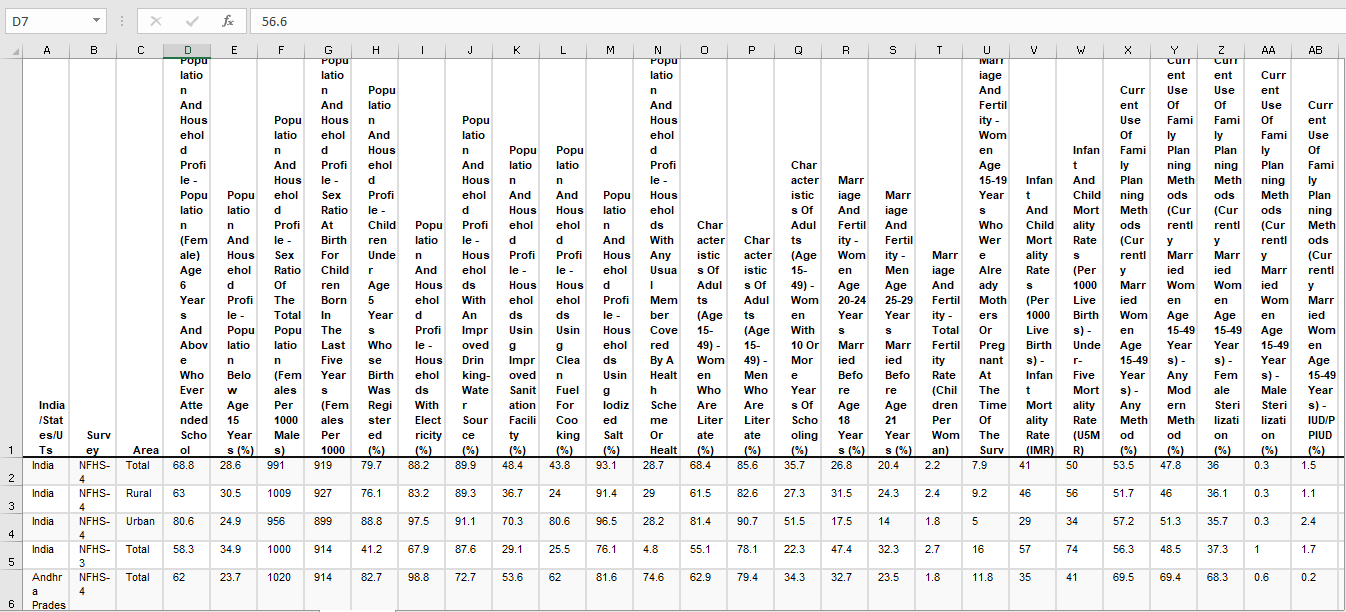
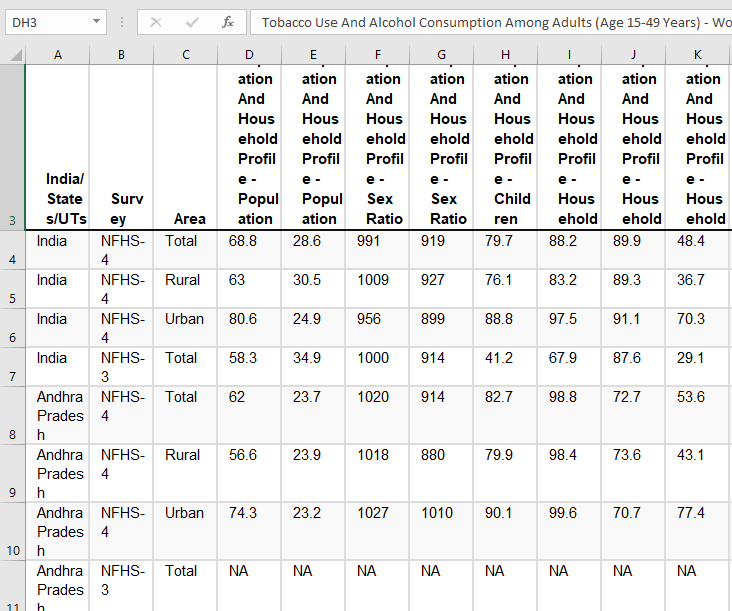
<www.data.gov.in>

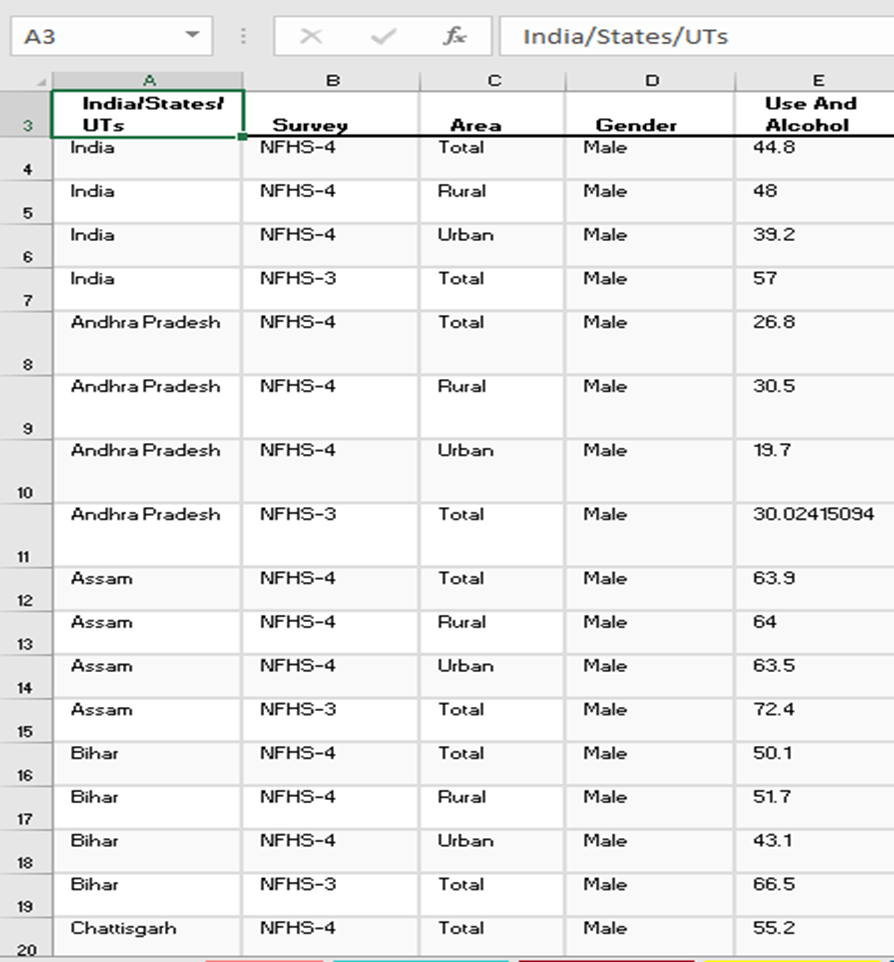
[www.catalog.data.gov](http://www.catalog.data.gov)

[www.wikipedia.com](http://www.wikipedia.com)

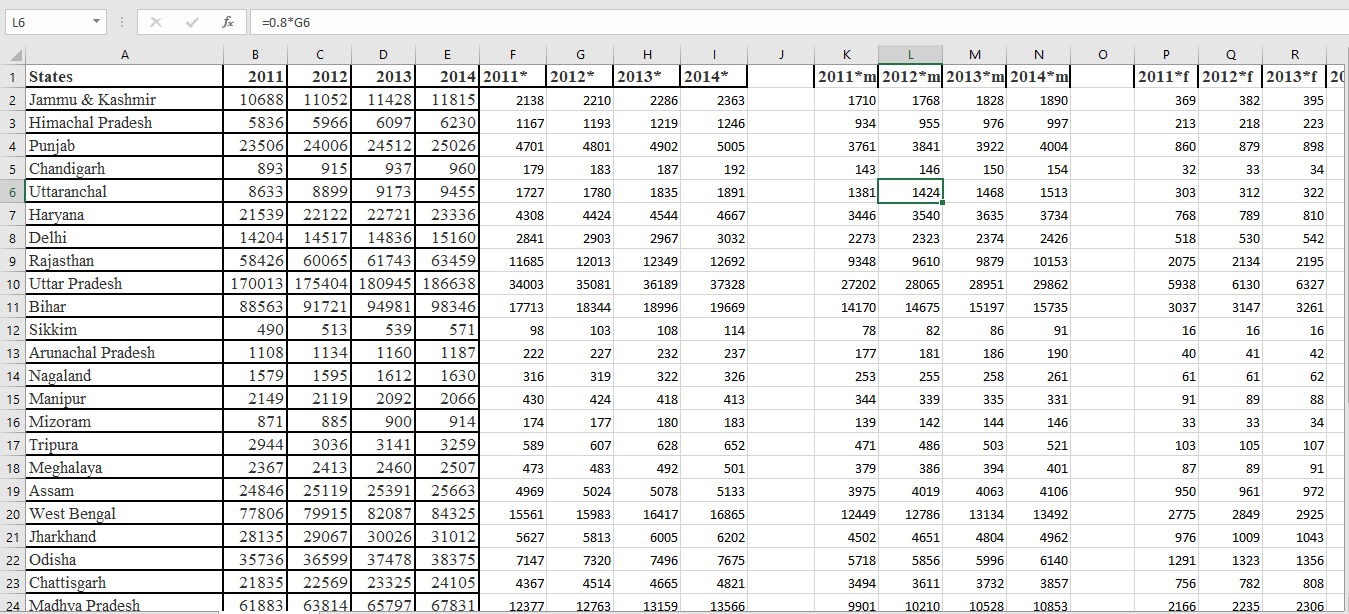
<www.cancer.gov>

**ETL Process**

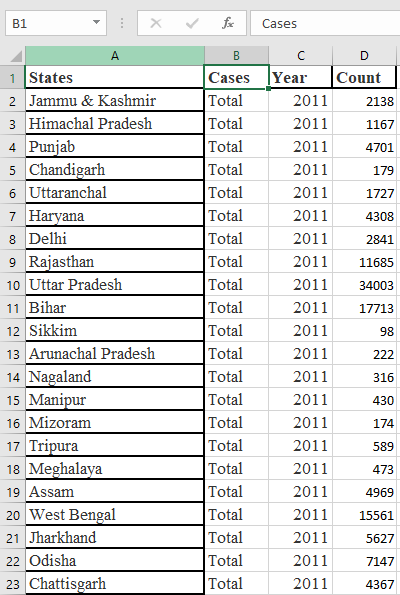
* ****Dataset was already in CSV so it could be easily load and cleaned in excel.
* We need to clean data and extract gender from data and remove some unnecessary columns.
* The Dataset below you can see is disintegrated and has 192 unnecessary column.
* The Data set also contains NULL values, therefore we either have to drop these columns or just taking the mean within the state data value.
* We need to clean data and extract gender from data and keep important columns.
* After removing the unnecessary columns and null values our data is left with only 4 columns including column gender which has been added explicitly.



* Similar steps for Mortality rate due to cancer.
* Dataset was already in CSV so it could be easily load and cleaned in excel.

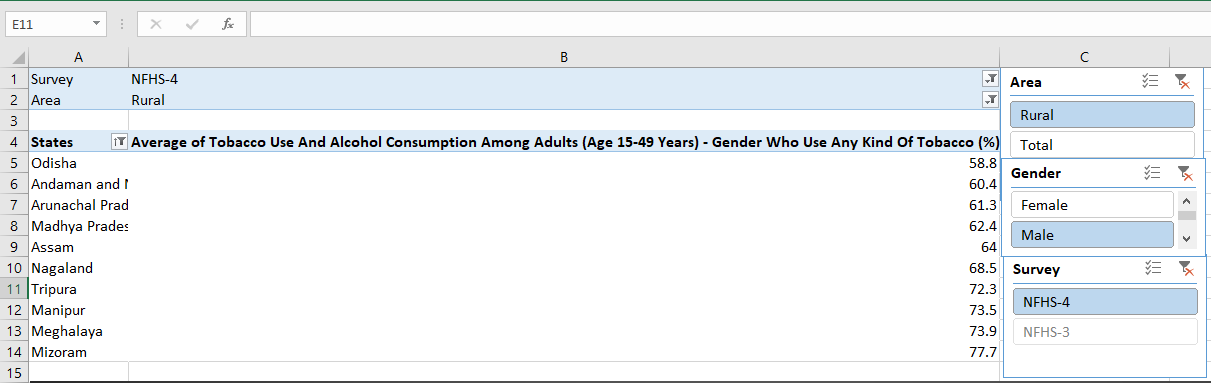


* We need to add columns like years which are mentioned: 2011, 2012, 2013,2014.
* We need to add column like Cases to specify the category of mortality rate and population effected by tobacco causing cancer.
* The data set below is cleaned and has been sorted on the bases of years.
* New column Cases has been added to specify the cause.

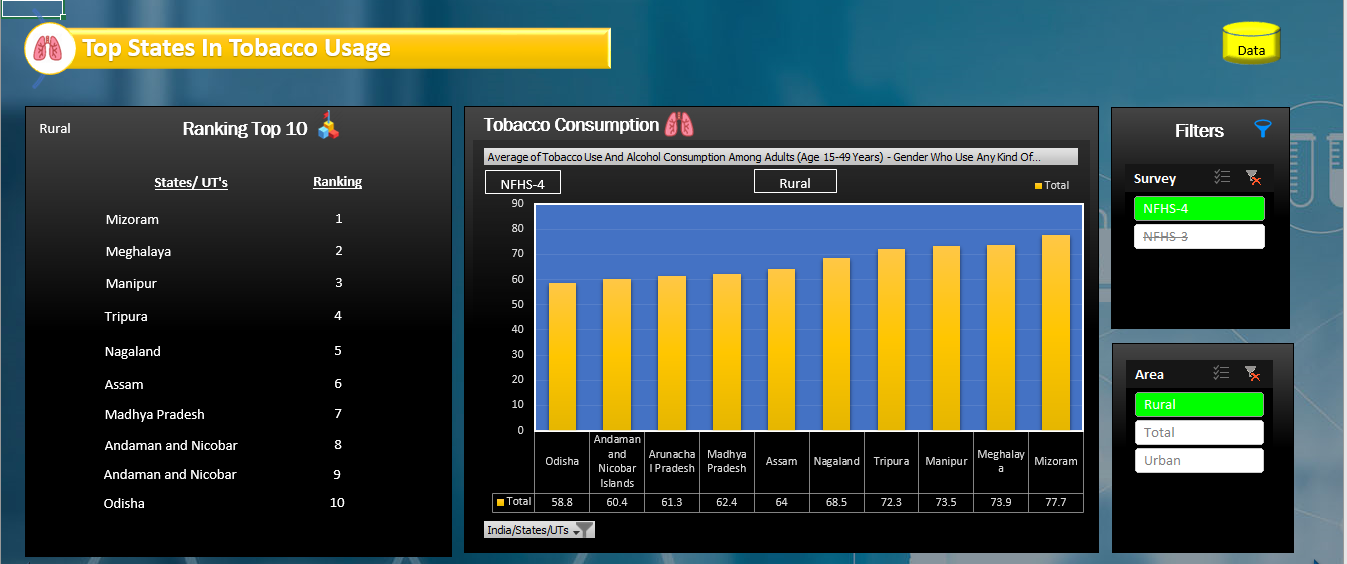


**Analysis of Datasets**

1. **Top 10 States with Highest Tobacco Consumption**

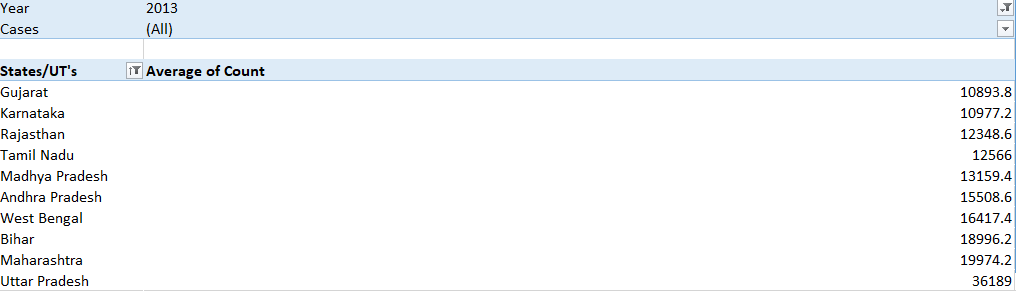
****Extracting data from test matches dataset using pivot tables.

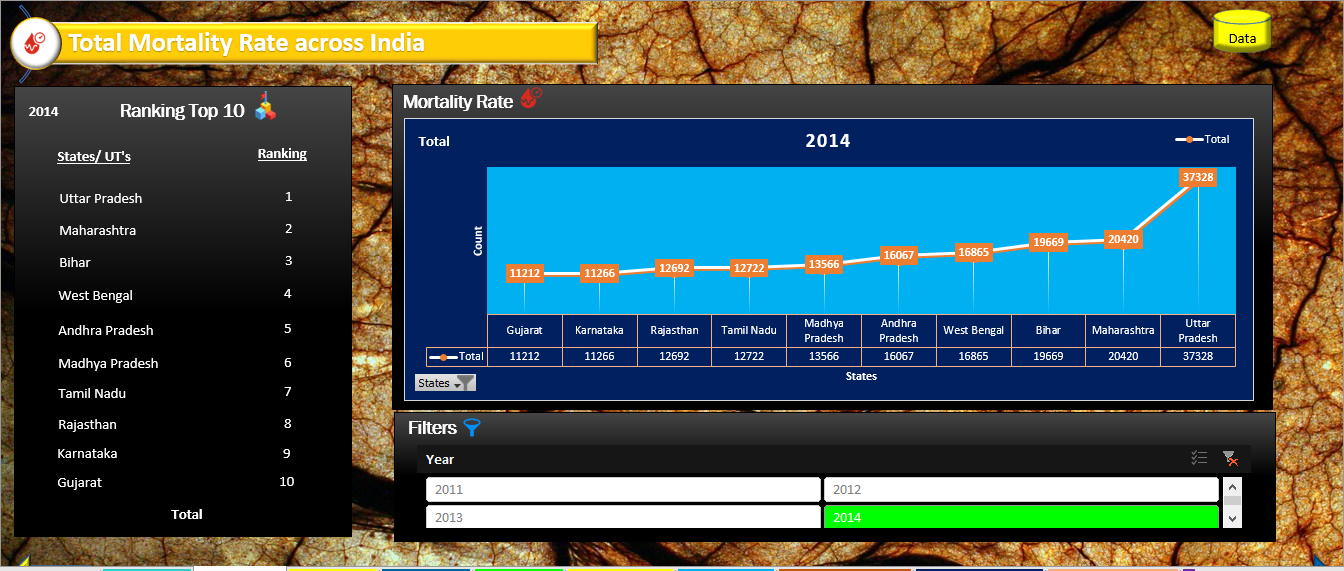
After copying data from pivot tables and making charts



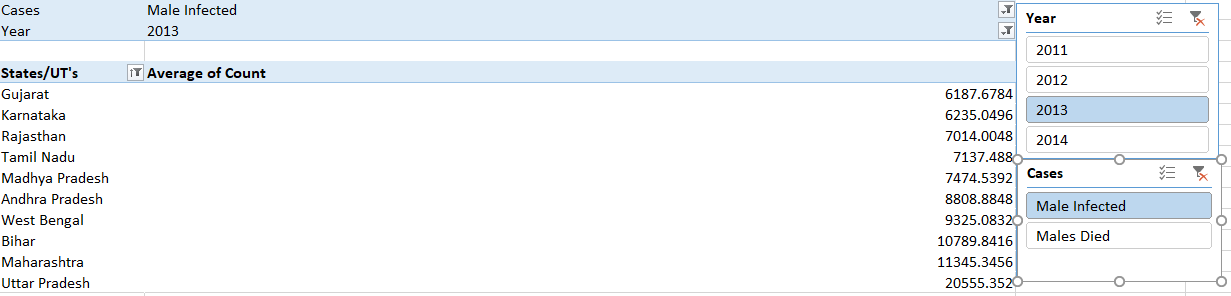
1. **Top 10 states with highest total Mortality Rate including Male and Female**

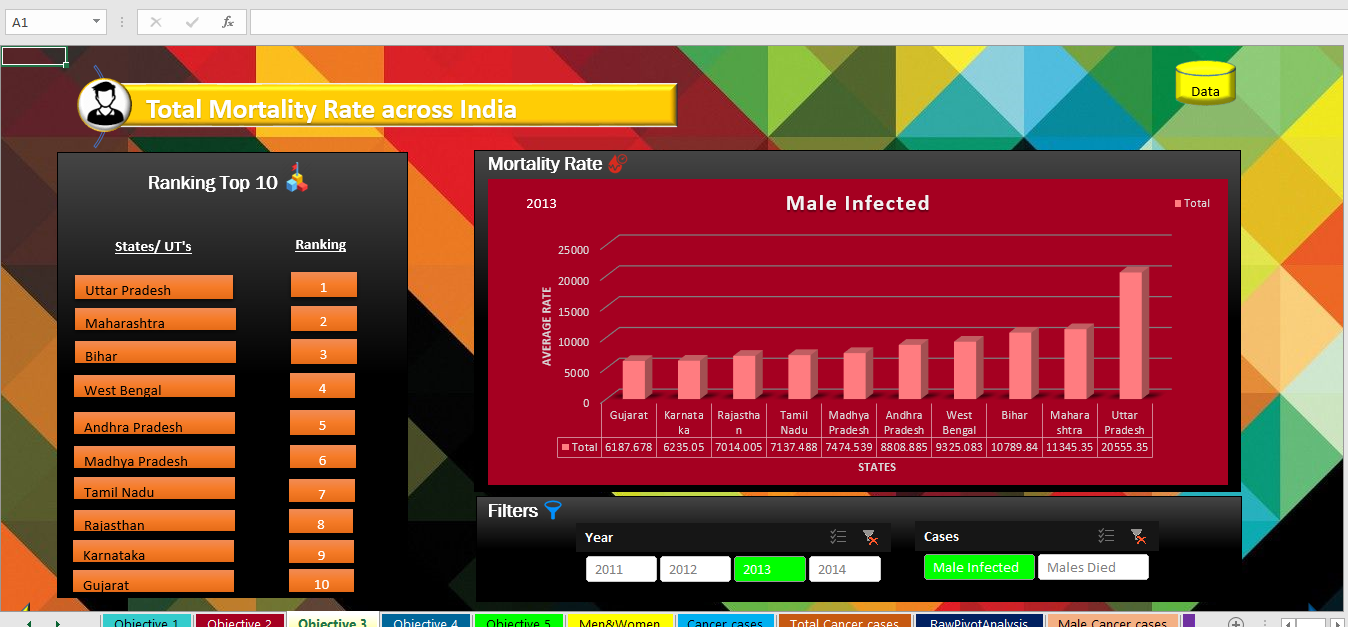
Use same pivot table for finding the Average Mortality Rate



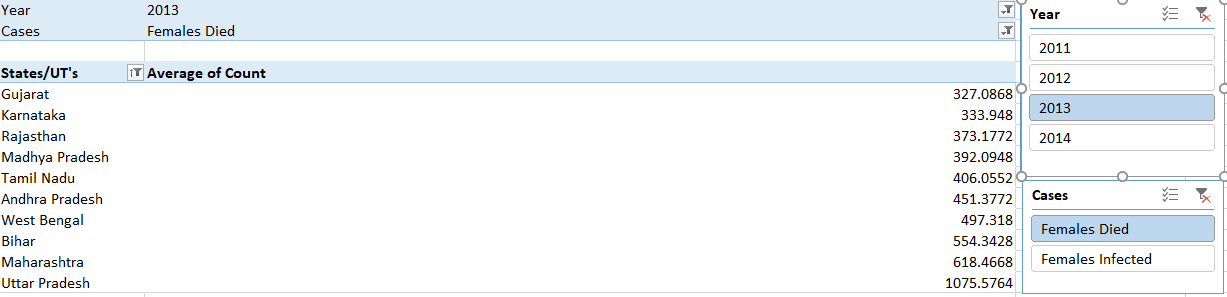
After copying Average Mortality Rate of Top States

1. **Top 10 States with Highest Male Mortality rate**

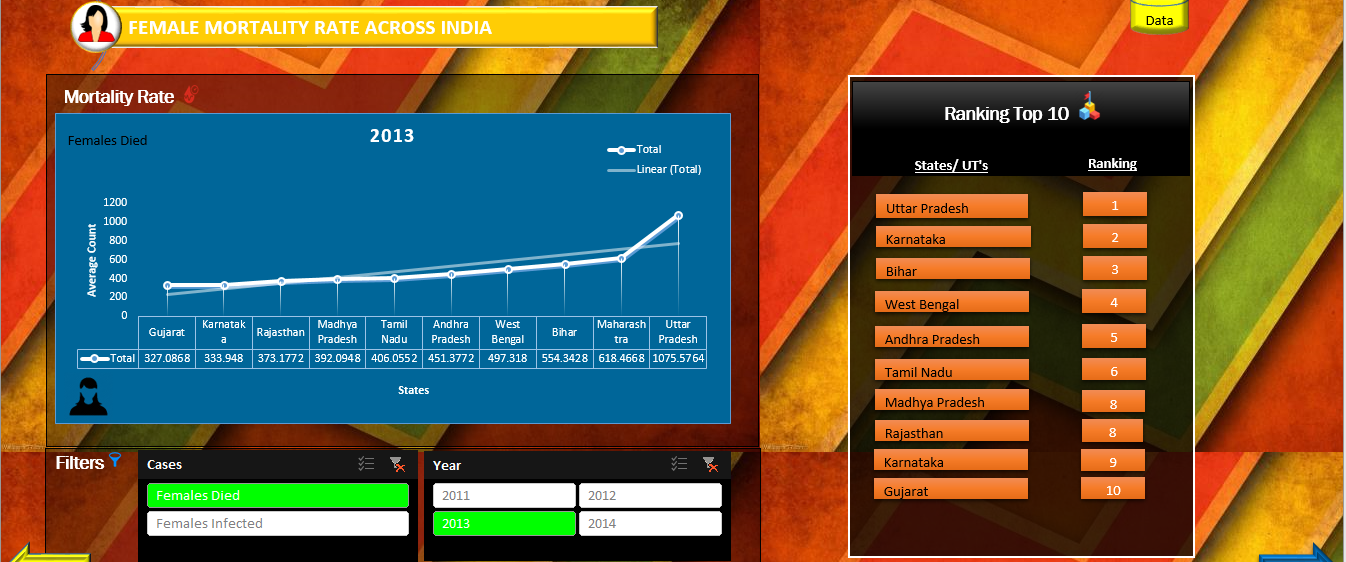
 Extracting data from Cancer cases dataset using pivot tables

After filtering, Top 10 states with high Male mortality rate, we get

1. **Top 10 states with Highest Female Mortality Rate**

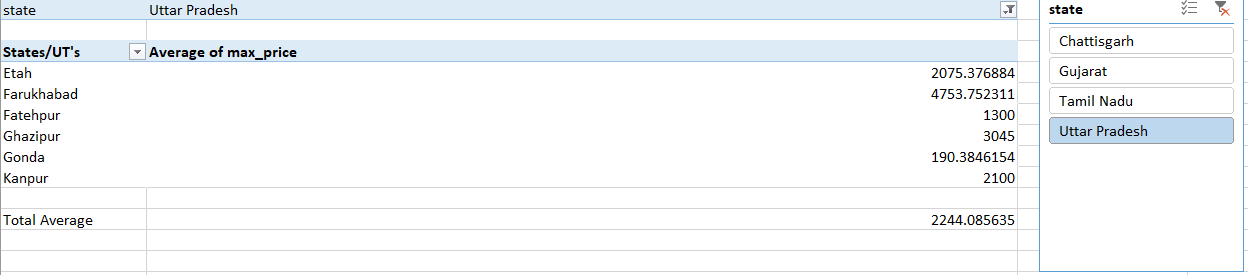
 Extract another table from batsman dataset

After filtering Top 10 states with High Female Mortality rate, we get

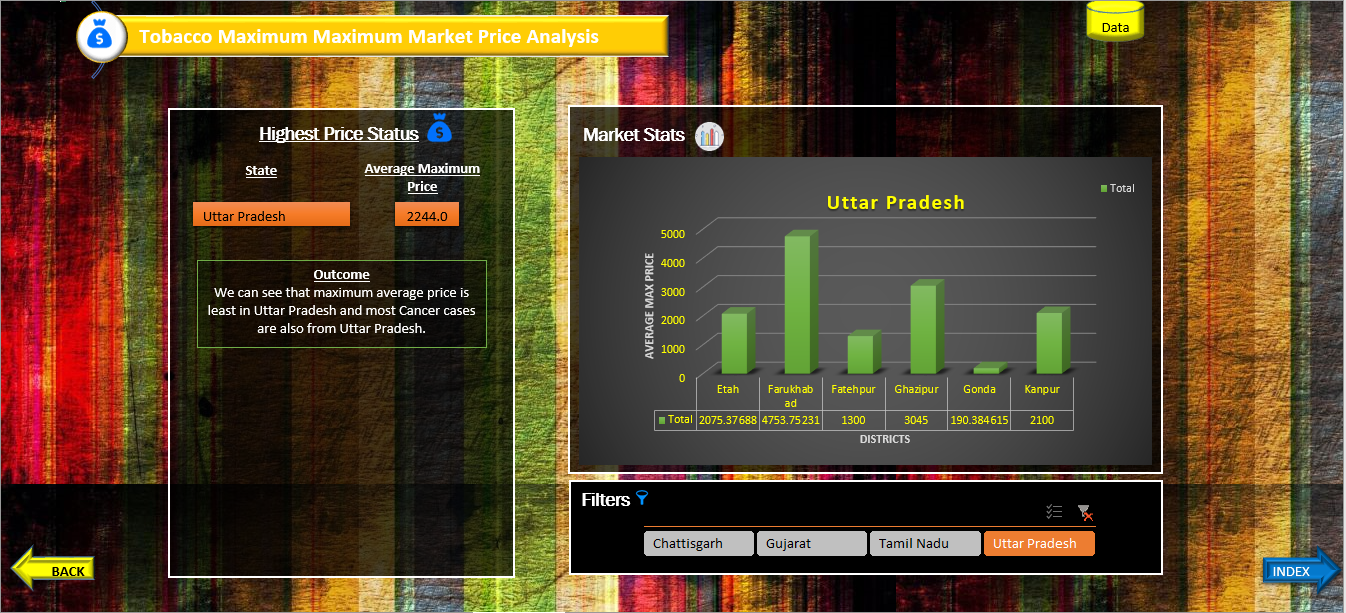


Note: There is a linear growth in cancer cases across India.

1. **Prominent states for the growth of Tobacco across India**

Similar analysis can for finding the districts with tobacco market also

After Filtering Top State with cheapest Tobacco Market



Note: Uttar Pradesh has top position in this list.

**Conclusion**

* India has 29 states, out of all these states has **Highest Tobacco Consumption**

Ranks (10-1)

|  |
| --- |
| Odisha |
| Andaman and Nicobar Islands |
| Arunachal Pradesh |
| Madhya Pradesh |
| Assam |
| Nagaland |
| Tripura |
| Manipur |
| Meghalaya |
| **Mizoram** |

* The Top 10 State with **Highest Total Mortality Rate** Ranks (10-1)

|  |
| --- |
| Gujarat |
| Karnataka |
| Rajasthan |
| Tamil Nadu |
| Madhya Pradesh |
| Andhra Pradesh |
| West Bengal |
| Bihar |
| Maharashtra |
| **Uttar Pradesh** |

* Top 10 Sates **with Highest Male Mortality Rate** Ranks (10 – 1).

|  |
| --- |
| Gujarat |
| Karnataka |
| Rajasthan |
| Tamil Nadu |
| Madhya Pradesh |
| Andhra Pradesh |
| West Bengal |
| Bihar |
| Maharashtra |
| **Uttar Pradesh** |

* Top 10 States with **Highest Female Mortality Rate**

|  |
| --- |
| Gujarat |
| Karnataka |
| Rajasthan |
| Madhya Pradesh |
| Tamil Nadu |
| Andhra Pradesh |
| West Bengal |
| Bihar |
| Maharashtra |
| **Uttar Pradesh** |
|  |

* Top State which has **Cheapest average maximum Price** rate for Tobacco market is **Uttar Pradesh.**

**Bibliography**

[www.gov.](http://www.gov.)in

[www.wikipedia.com](http://www.wikipedia.com)

MS EXCEL 2016

[www.catalog.gov.in](http://www.catalog.gov.in)