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)
    I. Introduction
    II. General Description
    III. Specific Requirements
    IV. Analysis results
    V. 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 (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 (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

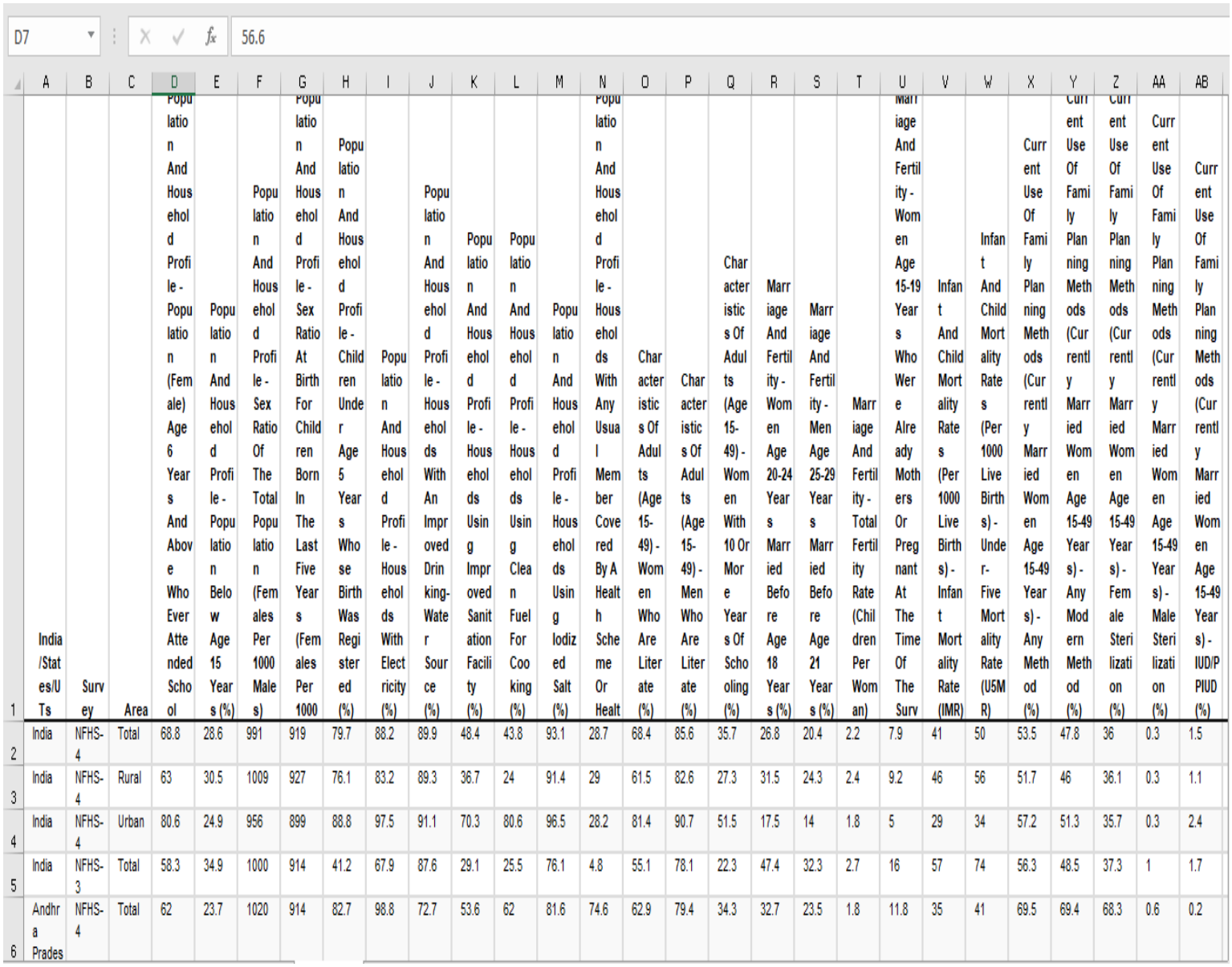
www.catalog.data.gov

www.wikipedia.com

www.cancer.gov

ETL Process

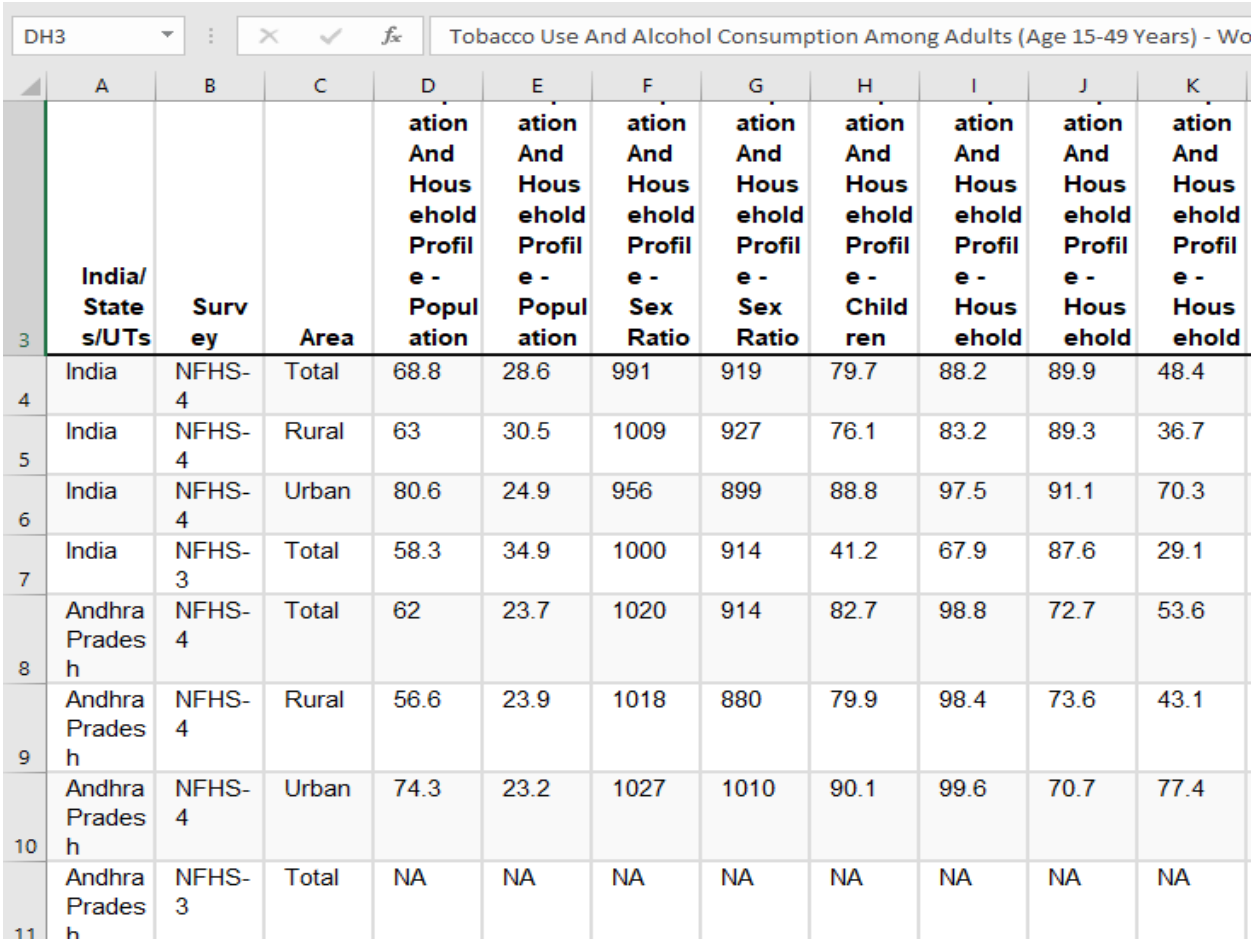
- Dataset was already in CSV so it could be easily load and cleaned in excel.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB
1	India/States/UTs	Survey	Area	Population And Household Profile - Population (Female) Age 6 Years And Above Who Ever Attended School	Population And Household Profile - Population Below Age 15 Years (%)	Population And Household Profile - Sex Ratio Of The Total Population (Females Per 1000 Males)	Population And Household Profile - Sex Ratio At Birth For Children Born In The Last Five Years (Females Per 1000	Population And Household Profile - Children Under Age 5 Years Whose Birth Was Registered (%)	Population And Household Profile - Households With Electricity (%)	Population And Household Profile - Households With An Improved Drinking-Water Source (%)	Population And Household Profile - Households Using Improved Sanitation Facility (%)	Population And Household Profile - Households Using Clean Fuel For Cooking (%)	Population And Household Profile - Households Using Iodized Salt (%)	Population And Household Profile - Households With Any Usual Member Covered By A Health Scheme Or Healt	Characteristics Of Adults (Age 15-49) - Women Who Are Literate (%)	Characteristics Of Adults (Age 15-49) - Men Who Are Literate (%)	Characteristics Of Adults (Age 15-49) - Women With 10 Or More Years Of Schooling (%)	Marriage And Fertility - Women Age 20-24 Years Married Before Age 18 Years (%)	Marriage And Fertility - Men Age 25-29 Years Married Before Age 21 Years (%)	Marriage And Fertility - Total Fertility Rate (Children Per Woman)	Marriage And Fertility - Women Age 15-19 Years Who Were Already Mothers Or Pregnant At The Time Of The Surv	Infant And Child Mortality Rates (Per 1000 Live Births) - Infant Mortality Rate (IMR)	Infant And Child Mortality Rates (Per 1000 Live Births) - Under-Five Mortality Rate (U5MR)	Current Use Of Family Planning Methods (Currently Married Women Age 15-49 Years) - Any Method (%)	Current Use Of Family Planning Methods (Currently Married Women Age 15-49 Years) - Any Modern Method (%)	Current Use Of Family Planning Methods (Currently Married Women Age 15-49 Years) - Female Sterilization (%)	Current Use Of Family Planning Methods (Currently Married Women Age 15-49 Years) - Male Sterilization (%)	Current Use Of Family Planning Methods (Currently Married Women Age 15-49 Years) - IUD/PPIUD (%)
2	India	NFHS-4	Total	68.8	28.6	991	919	79.7	88.2	89.9	48.4	43.8	93.1	28.7	68.4	85.6	35.7	26.8	20.4	2.2	7.9	41	50	53.5	47.8	36	0.3	1.5
3	India	NFHS-4	Rural	63	30.5	1009	927	76.1	83.2	89.3	36.7	24	91.4	29	61.5	82.6	27.3	31.5	24.3	2.4	9.2	46	56	51.7	46	36.1	0.3	1.1
4	India	NFHS-4	Urban	80.6	24.9	956	899	88.8	97.5	91.1	70.3	80.6	96.5	28.2	81.4	90.7	51.5	17.5	14	1.8	5	29	34	57.2	51.3	35.7	0.3	2.4
5	India	NFHS-3	Total	58.3	34.9	1000	914	41.2	67.9	87.6	29.1	25.5	76.1	4.8	55.1	78.1	22.3	47.4	32.3	2.7	16	57	74	56.3	48.5	37.3	1	1.7
6	Andhra Prades	NFHS-4	Total	62	23.7	1020	914	82.7	98.8	72.7	53.6	62	81.6	74.6	62.9	79.4	34.3	32.7	23.5	1.8	11.8	35	41	69.5	69.4	68.3	0.6	0.2

- 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.

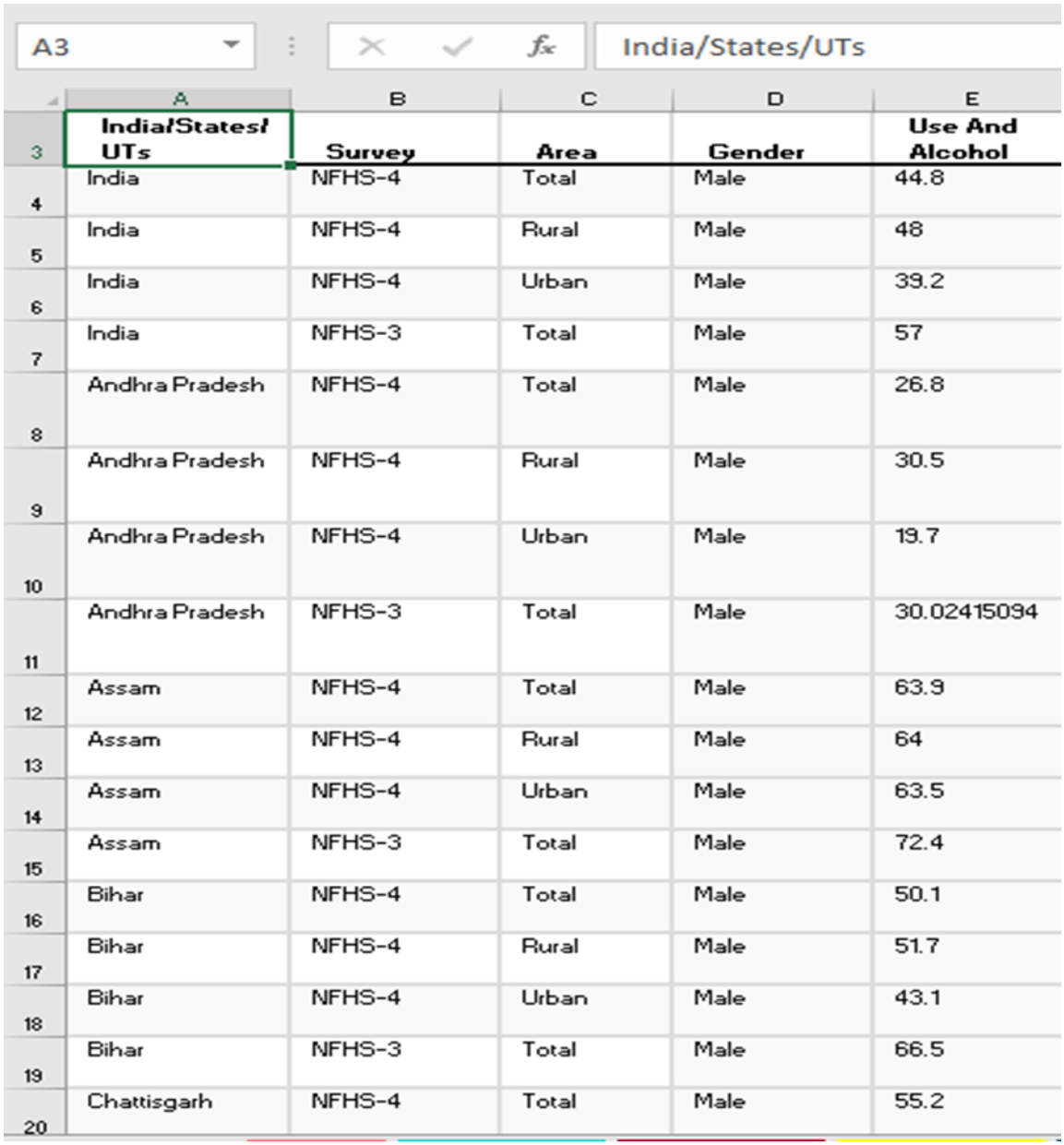


DH3 | Tobacco Use And Alcohol Consumption Among Adults (Age 15-49 Years) - Wo

	A	B	C	D	E	F	G	H	I	J	K
3	India/ States/UTs	Survey	Area	ation And Household Profile - Population	ation And Household Profile - Population	ation And Household Profile - Sex Ratio	ation And Household Profile - Sex Ratio	ation And Household Profile - Children	ation And Household Profile - Household	ation And Household Profile - Household	ation And Household Profile - Household
4	India	NFHS-4	Total	68.8	28.6	991	919	79.7	88.2	89.9	48.4
5	India	NFHS-4	Rural	63	30.5	1009	927	76.1	83.2	89.3	36.7
6	India	NFHS-4	Urban	80.6	24.9	956	899	88.8	97.5	91.1	70.3
7	India	NFHS-3	Total	58.3	34.9	1000	914	41.2	67.9	87.6	29.1
8	Andhra Pradesh	NFHS-4	Total	62	23.7	1020	914	82.7	98.8	72.7	53.6
9	Andhra Pradesh	NFHS-4	Rural	56.6	23.9	1018	880	79.9	98.4	73.6	43.1
10	Andhra Pradesh	NFHS-4	Urban	74.3	23.2	1027	1010	90.1	99.6	70.7	77.4
11	Andhra Pradesh	NFHS-3	Total	NA	NA	NA	NA	NA	NA	NA	NA

- 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.

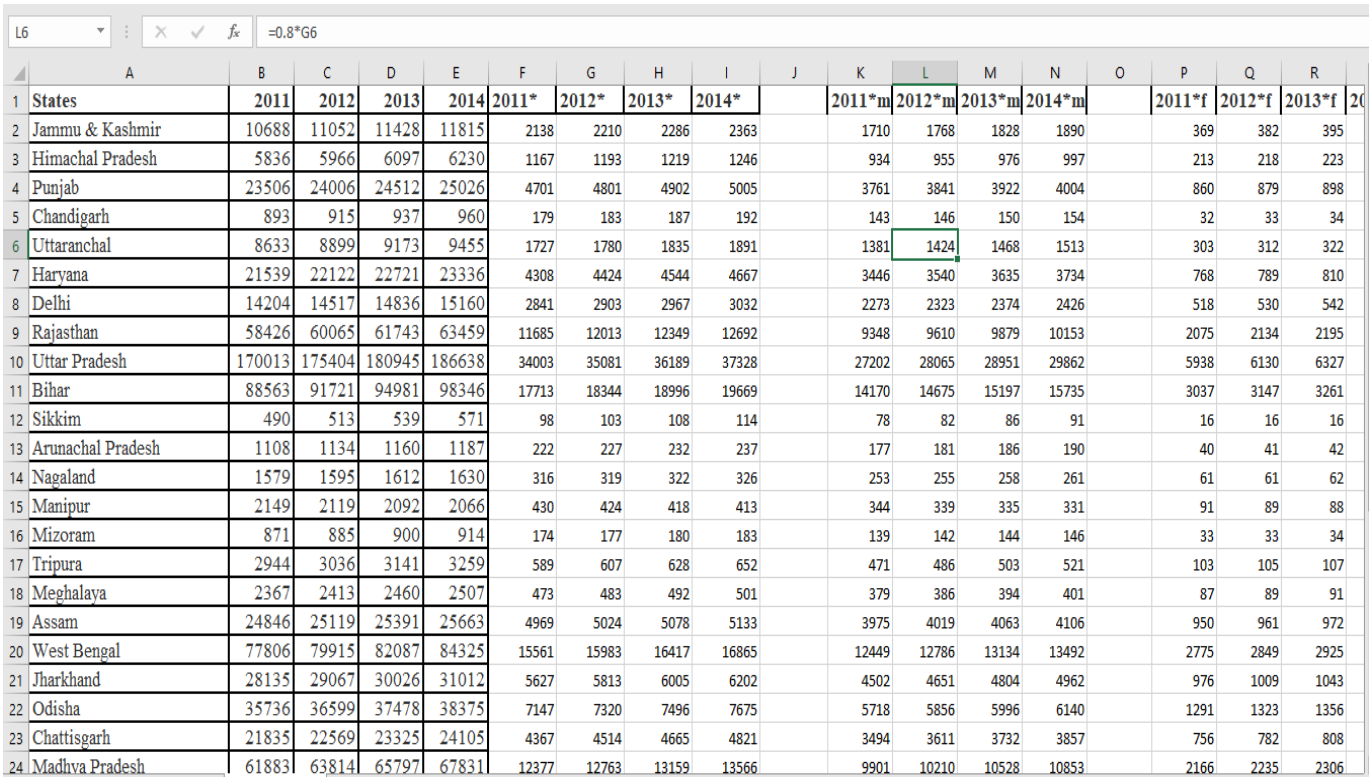
- After removing the unnecessary columns and null values our data is left with only 4 columns including column gender which has been added explicitly.



A3 | India/States/UTs

	A	B	C	D	E
3	India/States/ UTs	Survey	Area	Gender	Use And Alcohol
4	India	NFHS-4	Total	Male	44.8
5	India	NFHS-4	Rural	Male	48
6	India	NFHS-4	Urban	Male	39.2
7	India	NFHS-3	Total	Male	57
8	Andhra Pradesh	NFHS-4	Total	Male	26.8
9	Andhra Pradesh	NFHS-4	Rural	Male	30.5
10	Andhra Pradesh	NFHS-4	Urban	Male	19.7
11	Andhra Pradesh	NFHS-3	Total	Male	30.02415094
12	Assam	NFHS-4	Total	Male	63.9
13	Assam	NFHS-4	Rural	Male	64
14	Assam	NFHS-4	Urban	Male	63.5
15	Assam	NFHS-3	Total	Male	72.4
16	Bihar	NFHS-4	Total	Male	50.1
17	Bihar	NFHS-4	Rural	Male	51.7
18	Bihar	NFHS-4	Urban	Male	43.1
19	Bihar	NFHS-3	Total	Male	66.5
20	Chattisgarh	NFHS-4	Total	Male	55.2

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

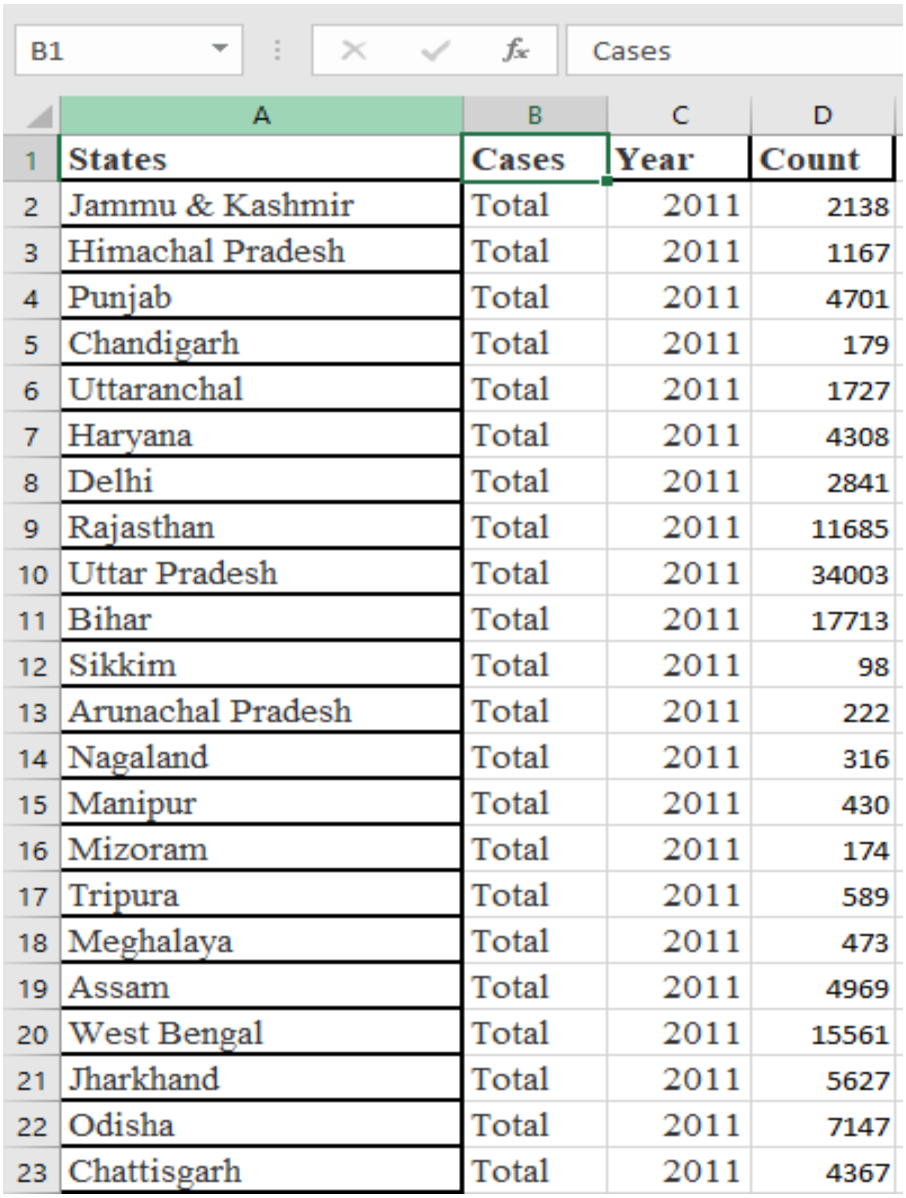


L6 =0.8*G6

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	States	2011	2012	2013	2014	2011*	2012*	2013*	2014*		2011*m	2012*m	2013*m	2014*m		2011*f	2012*f	2013*f
2	Jammu & Kashmir	10688	11052	11428	11815	2138	2210	2286	2363		1710	1768	1828	1890		369	382	395
3	Himachal Pradesh	5836	5966	6097	6230	1167	1193	1219	1246		934	955	976	997		213	218	223
4	Punjab	23506	24006	24512	25026	4701	4801	4902	5005		3761	3841	3922	4004		860	879	898
5	Chandigarh	893	915	937	960	179	183	187	192		143	146	150	154		32	33	34
6	Uttaranchal	8633	8899	9173	9455	1727	1780	1835	1891		1381	1424	1468	1513		303	312	322
7	Haryana	21539	22122	22721	23336	4308	4424	4544	4667		3446	3540	3635	3734		768	789	810
8	Delhi	14204	14517	14836	15160	2841	2903	2967	3032		2273	2323	2374	2426		518	530	542
9	Rajasthan	58426	60065	61743	63459	11685	12013	12349	12692		9348	9610	9879	10153		2075	2134	2195
10	Uttar Pradesh	170013	175404	180945	186638	34003	35081	36189	37328		27202	28065	28951	29862		5938	6130	6327
11	Bihar	88563	91721	94981	98346	17713	18344	18996	19669		14170	14675	15197	15735		3037	3147	3261
12	Sikkim	490	513	539	571	98	103	108	114		78	82	86	91		16	16	16
13	Arunachal Pradesh	1108	1134	1160	1187	222	227	232	237		177	181	186	190		40	41	42
14	Nagaland	1579	1595	1612	1630	316	319	322	326		253	255	258	261		61	61	62
15	Manipur	2149	2119	2092	2066	430	424	418	413		344	339	335	331		91	89	88
16	Mizoram	871	885	900	914	174	177	180	183		139	142	144	146		33	33	34
17	Tripura	2944	3036	3141	3259	589	607	628	652		471	486	503	521		103	105	107
18	Meghalaya	2367	2413	2460	2507	473	483	492	501		379	386	394	401		87	89	91
19	Assam	24846	25119	25391	25663	4969	5024	5078	5133		3975	4019	4063	4106		950	961	972
20	West Bengal	77806	79915	82087	84325	15561	15983	16417	16865		12449	12786	13134	13492		2775	2849	2925
21	Jharkhand	28135	29067	30026	31012	5627	5813	6005	6202		4502	4651	4804	4962		976	1009	1043
22	Odisha	35736	36599	37478	38375	7147	7320	7496	7675		5718	5856	5996	6140		1291	1323	1356
23	Chattisgarh	21835	22569	23325	24105	4367	4514	4665	4821		3494	3611	3732	3857		756	782	808
24	Madhya Pradesh	61883	63814	65797	67831	12377	12763	13159	13566		9901	10210	10528	10853		2166	2235	2306

- 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.



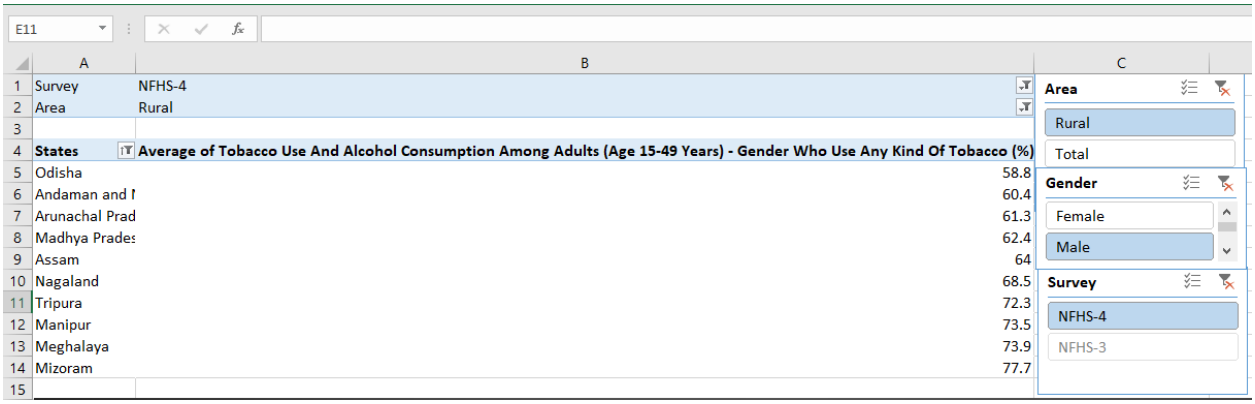
B1 | Cases

	A	B	C	D
1	**States**	**Cases**	**Year**	**Count**
2	Jammu & Kashmir	Total	2011	2138
3	Himachal Pradesh	Total	2011	1167
4	Punjab	Total	2011	4701
5	Chandigarh	Total	2011	179
6	Uttaranchal	Total	2011	1727
7	Haryana	Total	2011	4308
8	Delhi	Total	2011	2841
9	Rajasthan	Total	2011	11685
10	Uttar Pradesh	Total	2011	34003
11	Bihar	Total	2011	17713
12	Sikkim	Total	2011	98
13	Arunachal Pradesh	Total	2011	222
14	Nagaland	Total	2011	316
15	Manipur	Total	2011	430
16	Mizoram	Total	2011	174
17	Tripura	Total	2011	589
18	Meghalaya	Total	2011	473
19	Assam	Total	2011	4969
20	West Bengal	Total	2011	15561
21	Jharkhand	Total	2011	5627
22	Odisha	Total	2011	7147
23	Chattisgarh	Total	2011	4367

Analysis of Datasets

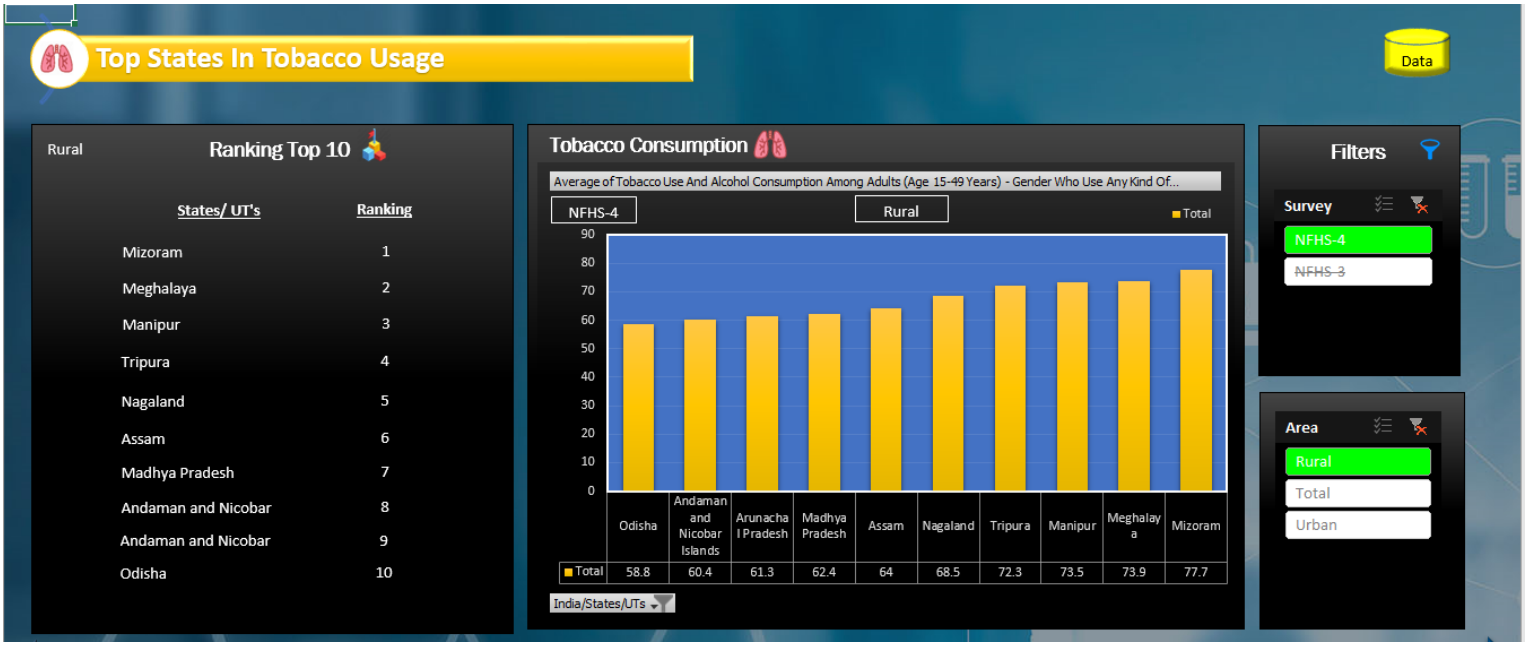
1. Top 10 States with Highest Tobacco Consumption

Extracting data from test matches dataset using pivot tables.



	A	B
1	Survey	NFHS-4
2	Area	Rural
3		
4	States	Average of Tobacco Use And Alcohol Consumption Among Adults (Age 15-49 Years) - Gender Who Use Any Kind Of Tobacco (%)
5	Odisha	58.8
6	Andaman and I	60.4
7	Arunachal Prad	61.3
8	Madhya Prades	62.4
9	Assam	64
10	Nagaland	68.5
11	Tripura	72.3
12	Manipur	73.5
13	Meghalaya	73.9
14	Mizoram	77.7
15		

After copying data from pivot tables and making charts

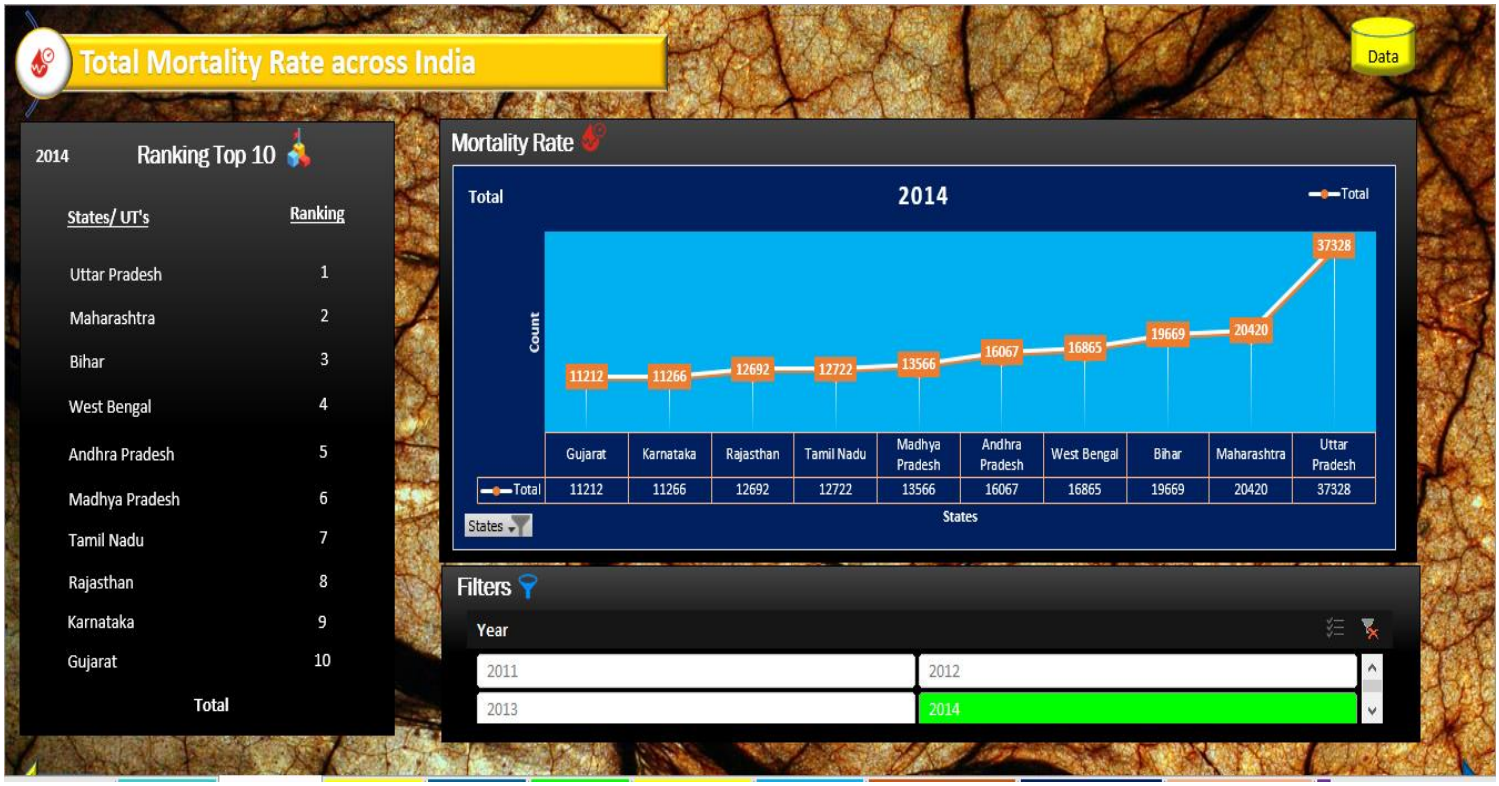


2. Top 10 states with highest total Mortality Rate including Male and Female

Use same pivot table for finding the Average Mortality Rate

Year	2013
Cases	(All)
States/UT's	**Average of Count**
Gujarat	10893.8
Karnataka	10977.2
Rajasthan	12348.6
Tamil Nadu	12566
Madhya Pradesh	13159.4
Andhra Pradesh	15508.6
West Bengal	16417.4
Bihar	18996.2
Maharashtra	19974.2
Uttar Pradesh	36189

After copying Average Mortality Rate of Top States

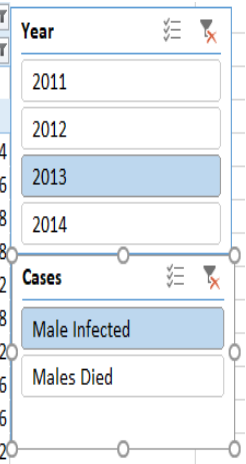


3. Top 10 States with Highest Male Mortality rate

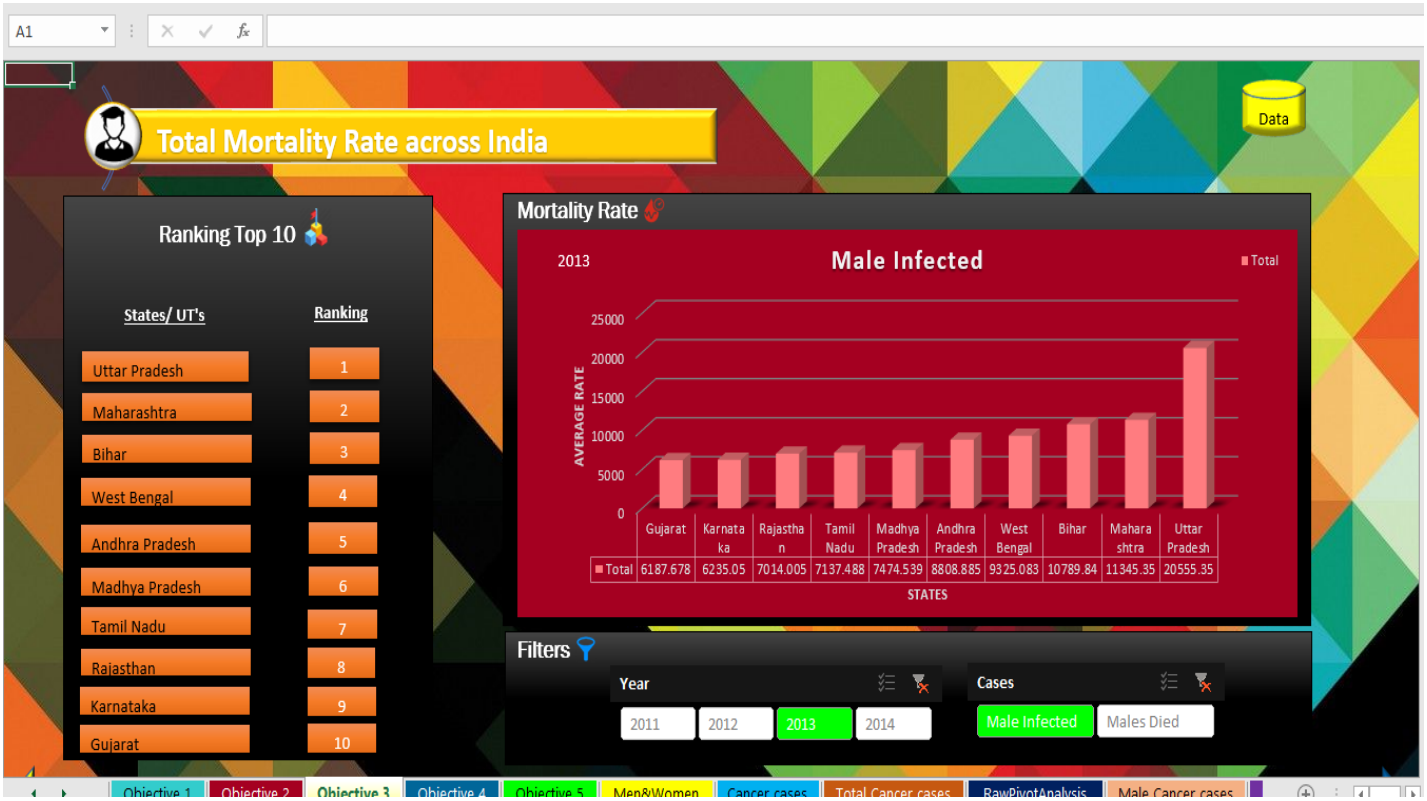
Extracting data from Cancer cases dataset using pivot tables

Cases	Male Infected
Year	2013

States/UT's	Average of Count
Gujarat	6187.6784
Karnataka	6235.0496
Rajasthan	7014.0048
Tamil Nadu	7137.488
Madhya Pradesh	7474.5392
Andhra Pradesh	8808.8848
West Bengal	9325.0832
Bihar	10789.8416
Maharashtra	11345.3456
Uttar Pradesh	20555.352



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



4. Top 10 states with Highest Female Mortality Rate

Extract another table from batsman dataset

Year	2013
Cases	Females Died

States/UT's	Average of Count
Gujarat	327.0868
Karnataka	333.948
Rajasthan	373.1772
Madhya Pradesh	392.0948
Tamil Nadu	406.0552
Andhra Pradesh	451.3772
West Bengal	497.318
Bihar	554.3428
Maharashtra	618.4668
Uttar Pradesh	1075.5764

Year: 2011, 2012, 2013, 2014

Cases: Females Died, Females Infected

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

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

5. Prominent states for the growth of Tobacco across India

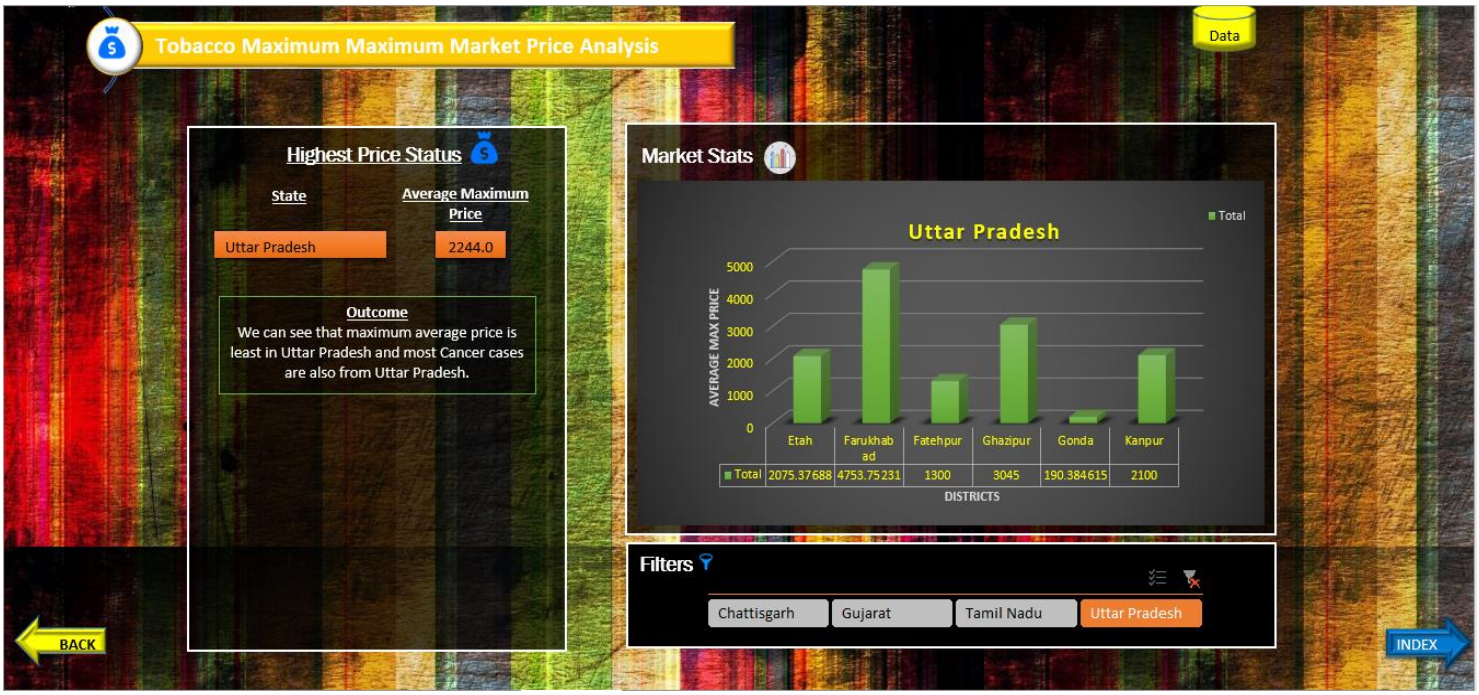
Similar analysis can for finding the districts with tobacco market also

state	Uttar Pradesh
States/UT's	**Average of max_price**
Etah	2075.376884
Farukhabad	4753.752311
Fatehpur	1300
Ghazipur	3045
Gonda	190.3846154
Kanpur	2100
Total Average	2244.085635

state

- Chattisgarh
- Gujarat
- Tamil Nadu
- Uttar Pradesh

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.in

www.wikipedia.com

MS EXCEL 2016

www.catalog.gov.in